Fine-Tuned Llama for Multilingual Text-to-Text Coreference Resolution

Jakub Hejman and Ondřej Pražák and Miloslav Konopík

{hejmanj,ondfa,konopik}@kiv.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

This paper describes our approach to the CRAC 2025 Shared Task on Multilingual Coreference Resolution. We compete in the LLM track, where the systems are limited to generative text-to-text approaches. Our system is based on Llama 3.1-8B, fine-tuned to tag the document with coreference annotations. We have made one significant modification to the text format provided by the organizers: The model relies on the syntactic head for mention span representation. Additionally, we use joint pre-training, and we train the model to generate empty nodes. We provide an in-depth analysis of the performance of our models, which reveals several implementation problems. Although our system ended up in last place, we achieved the best performance on 10 datasets out of 22 within the track. By fixing the discovered problems in the post-evaluation phase, we improved our results substantially, outperforming all the systems in the LLM track and even some unconstrained track systems.

1 Introduction

Coreference resolution is the task of identifying mentions of entities and grouping the mentions of the same real-world entity together. It is a fundamental NLP task that is increasingly left to the implicit understanding of LLMs rather than being explicitly computed as an intermediate step of an NLP pipeline. As such, investigating the models' ability to accurately identify entities in real-world scenarios is a direct way of ensuring that their understanding of the material is robust. Additionally, coreference resolution is an unsolved task, and findings from it may well contribute to progress in related NLP problems. This task can be very challenging, especially in cases where coreferences span the whole document.

CorefUD (Nedoluzhko et al., 2022) is an extension of Universal Dependencies (Nivre et al., 2020) to include coreference harmonized across multiple languages. The recent version of CorefUD 1.3 (Novák et al., 2025b) contains 24 datasets in 17 languages. All data is stored in the CoNLL-U format, which stores the pretokenized text, dependency trees, and entity annotations within the miscellaneous column in a unified format. Basic statistics of individual datasets are shown in Table 1. CRAC shared task on multilingual coreference resolution is built upon this dataset, and 2025 is the fourth year this task has been running.

For generative LLMs, the coreference resolution task is still challenging, and standard benchmarks like SuperGLUE are mostly limited to the Winograd Schema Challenge (WSC) (Levesque et al., 2012). WSC was originally proposed as an alternative to the Turing test. It is a pronoun resolution problem that cannot easily be solved based on statistical patterns. General coreference resolution is typically not present in standard multi-task LLM benchmarks, yet there are many papers focusing on coreference resolution with LLMs. However, the experiments are often limited to a single dataset (Zhang et al., 2023; Stano and Horák, 2025).

As suggested last year (Novák et al., 2024), the CRAC 2025 coreference resolution shared task includes the LLM track, where the participants are asked to use a pure text-to-text approach to solve the task. The organizers also provide a recommended plaintext format of the CorefUD dataset together with the conversion tool. There are several other differences from previous years. As every year, several new datasets were added into CorefUD. The smallest datasets (en_parcorfull and de_parcorfull) were discarded due to very unstable results of all the systems across previous years.

This paper describes how we fine-tune Llama 3.1-8B in a text-to-text manner to participate in this track. Our approach relies on mention head prediction, joint pre-training, and empty node generation.

Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2025), pages 140–148
Suzhou, China, November 9, 2025. ©2025 Association for Computational Linguistics

dataset	total number of				entities				mentions			
					total	per 1k	length		total	per 1k	length	
	docs	sents	words	empty n.	count	words	max	avg.	count	words	max	avg.
ca_ancora	1,298	13,613	429,313	6,377	17,558	41	101	3.6	62,417	145	141	4.8
cs_pcedt	2,312	49,208	1,155,755	35,654	49,225	43	236	3.4	168,055	145	79	3.6
cs_pdt	3,165	49,419	834,707	21,092	46,460	56	173	3.3	154,437	185	99	3.1
cu_proiel	26	6,832	61,759	6,289	3,396	55	134	6.5	22,116	358	52	1.5
de_potsdam	176	2,238	33,222	0	880	26	15	2.9	2,519	76	34	2.6
en_gum	237	13,263	233,926	119	9,200	39	131	4.4	40,656	174	95	2.6
en_litbank	100	8,560	210,530	0	2,164	10	261	10.8	23,340	111	129	1.6
es_ancora	1,356	14,159	458,418	8,112	19,445	42	110	3.6	70,663	154	101	4.8
fr_ancor	455	31,761	454,577	0	13,204	29	103	4.3	56,459	124	17	1.9
fr_democrat	126	13,057	284,883	0	7,162	25	895	6.5	46,487	163	71	1.7
grc_proiel	19	6,475	64,111	6,283	3,215	50	332	6.6	21,354	333	52	1.7
hbo_ptnk	40	1,161	28,485	0	870	31	102	7.2	6,247	219	22	1.5
hi_hdtb	271	3,479	76,282	0	3,148	41	36	3.8	12,082	158	43	1.8
hu_korkor	94	1,351	24,568	1,569	1,122	46	41	3.6	4,091	167	42	2.2
hu_szegedkoref	400	8,820	123,968	4,857	4,769	38	36	3.2	15,165	122	36	1.6
ko_ecmt	1,470	30,784	482,986	0	16,536	34	55	3.4	56,538	117	12	1.3
lt_lcc	100	1,714	37,014	0	1,087	29	23	4.0	4,337	117	19	1.5
no_bokmaal	346	15,742	245,515	0	5,658	23	298	4.7	26,611	108	51	1.9
no_nynorsk	394	12,481	206,660	0	5,079	25	84	4.3	21,847	106	57	2.1
pl_pcc	1,828	35,874	538,885	18,615	22,143	41	135	3.7	82,706	153	108	1.9
ru_rucor	181	9,035	156,636	0	3,515	22	141	4.6	16,193	103	18	1.7
tr_itcc	24	4,732	55,358	11,584	4,019	73	369	5.4	21,569	390	31	1.1

Table 1: CorefUD 1.3 data sizes in terms of the total number of documents, sentences, words (i.e. non-empty nodes), empty nodes (empty words), coreference entities (total count, relative count per 1000 words, average and maximal length in number of mentions) and coreference mentions (total count, relative count per 1000 words, average and maximal length in number of words). All the counts are excluding singletons and for the concatenation of train+dev+test. Train/dev/test splits of these datasets roughly follow the 8/1/1 ratio. Taken from Novák et al. (2025a)

2 Related Work

Neural coreference resolution has traditionally been approached using encoder-only models (Joshi et al., 2020; Straka, 2023; Pražák et al., 2021; Pražák and Konopik, 2022) and Higher-Order Inference (HOI) (Xu and Choi, 2020). Recently, text-to-text models have gained popularity for this task (Zhang et al., 2023).

The most commonly used model for multilingual coreference resolution is mT5 (Raffel et al., 2020), which has been applied in both end-to-end (Straka, 2023) and text-to-text approaches (Bohnet et al., 2023; Stano and Horák, 2025; Skachkova, 2024). It was also utilized by the top system at CRAC 2024 (Novák et al., 2024).

A notable text-to-text approach is the Link-Append method proposed by Bohnet et al. (2023), which avoids an intermediate mention detection step by training a seq2seq model to predict actions that incrementally build coreference clusters.

Skachkova (2024) introduced a direct annotation scheme where the model generates document text along with brackets and cluster identifiers. Their system employs prompt tuning and incremental generation to label entities progressively, along with data augmentations to address common failure modes such as unchanged inputs, repeated outputs, and duplicate mentions.

Zhang et al. (2023) propose an output scheme which combines tag generation with a second operator that copies tokens from the input to avoid repetition.

An alternative direction to fine-tuning is prompting. Stano and Horák (2025) demonstrate this approach on the simpler anaphora resolution task. This result suggests that some LLMs possess in-context learning capabilities powerful enough to tackle coreference resolution without any specialized training.

Dobrovolskii (2021) suggested reducing the mention space by selecting a single word to represent each mention, using the syntactic head as the representative word. Their experiments were conducted on the English OntoNotes corpus. In the next step, after antecedent prediction, they employ a CNN-based span predictor to reconstruct the original mentions.

3 Model

We use the provided CoNLL-U-to-Text converter and train the model to generate document texts with entity tags inserted. Our model benefits from joint cross-lingual training, headword mention representation, and zero-mentions modeling.

Inspired by word-level coreference resolution and by previous CorefUD experiments (Pražák et al., 2024; Prazak and Konopík, 2024), we also evaluate the model with headword mention representation. Here, we represent mentions only by their syntactic heads (highest nodes in a dependency tree). The plaintext format suggested by the organizers does not include any syntactic information, so we modified the converter to extract syntactic heads of mentions from CoNLL-U. Considering that the official evaluation metric uses head-match, we do not need to reconstruct the original spans for evaluation. But this step would be fairly straightforward and can be done similarly to Dobrovolskii (2021).

We implement an optional document splitting pre-processing step to deal with datasets dominated by documents that are too long to train on in our setup. The documents are split hierarchically first by paragraphs, then by sentences, and then by words to fit into a limit of 250 words. We chose this limit empirically to fit all the datasets into our training context length. We manually enable this step for datasets that are problematic otherwise.

We train a joint model on a concatenation of all the datasets in the CorefUD 1.3 collection in the first step. In the second step, we fine-tune the joint model on each dataset separately.

Our model also predicts empty nodes and zero mentions. We fine-tune the model to insert empty nodes into the text, directly following its syntactic parent, as suggested by the provided CoNLL-U-to-Text converter.

4 Training & Inference

We fine-tune pre-trained Llama 3.1 8B (Grattafiori et al., 2024) using QLoRA (Dettmers et al., 2023) on a single NVIDIA A40 GPU. The frozen foundational model is quantized to 8 bits, and a LoRA adapter with a rank of 64 is optimized. We use completion-only training, which means that gradients are computed only on completion tokens and not on prompt tokens. This ensures that the model focuses on filling in the entity annotations instead of predicting the original document text.

Our models are trained with a maximum sequence length of 4096 tokens. Sequences that surpass the sequence length limit are filtered from the dataset before training starts. For some datasets, this leads to the removal of all documents from either the evaluation or training split. In these cases, we split the samples so that we effectively utilize the dataset as described in Section 3.

When generating the model's predictions, we use an increased sequence length. For most experiments and datasets, we allow up to 2048 tokens in the prompt and 4096 generated tokens because some datasets contain documents that are, on average, about 2 times longer with labels than without them (more in Section 5.3). For certain datasets, we increase the limits up to 8,192 for the prompt and 16,384 for generation. We do not observe issues with these implicit sequence length extensions between training and inference; scores continue to improve as inference context increases up to the maximum document length.

5 Results & Discussion

Table 2 shows the results of our system on development sets. It is split into two parts: submitted predictions and post-evaluation experiments. Since we did not have enough time to search a complete hyperparameter grid during the evaluation period, we evaluated just two variants of the model:

1. **standard model** – Full-span mention representation, zero mentions are ignored.
2. **heads_zeros model** – Headword mention representation, empty nodes generated, zero mention coreference predicted.

5.1 Submission-time Problems

We performed post-evaluation experiments to address the system's main shortcomings, since we could not resolve all the dataset-specific issues before the deadline. Our original submission exhibited the following problems:

1. **Improper training continuation for joint pre-training** – Due to a bug, joint pre-training

dataset	submitted		post-evaluation experiments		
	standard	heads_zeros	from joint	+ heads_zeros	+ long
ca_ancora	73.27	79.91	74.49	**82.19**	**82.19**
cs_pcedt	57.27	0	59	67.38	**68.89**
cs_pdt	68.75	0	71.24	**76.37**	**76.37**
cu_proiel	14	29	34.5	34.36	**42.95**
de_potsdam	74.4	77	78.95	80.14	**82.83**
en_gum	73.7	76.05	76.57	76.96	**77.16**
en_litbank	81.5	83	82.1	82.1	**84.75**
es_ancora	74.57	0	75.47	80.45	**81.68**
fr_ancor	25.5	26.06	30.7	35.5	**59.95**
fr_democrat	33.89	37.58	49.64	47.78	**57.65**
grc_proiel	50.33	0	54.26	51.61	**65.48**
hbo_ptnk	0	0	46.7	38.04	**69.45**
hi_hdtb	75.7	78.83	75.9	79.92	**80.95**
hu_korkor	40.94	0	46.91	64.72	**65.14**
hu_szegedkoref	62.88	68.52	62.92	67.83	**69.58**
ko_ecmt	66.46	62.02	**65.7**	63.75	**65.7**
lt_lcc	78.26	74.93	**79.33**	76.84	**79.33**
no_bokmaal	77.05	79.12	80.27	80.11	**80.69**
no_nynorsk	74.61	77.63	78.43	79.72	**82.06**
pl_pcc	61.85	0	60.27	**72.3**	**72.3**
ru_rucor	53.96	55.28	59.22	62.53	**63.71**
tr_itcc	24.72	30.76	-	-	**59.4**
avg	56.53	41.13	63.93	66.70	71.28
median	64.67	46.43	65.7	72.3	70.94

Table 2: Results on development splits. Best results are bold. The results on which the best submission is based are underlined. Results marked as '-' could not be evaluated due to massive overfitting and degradation of the output format.

did not improve performance and was therefore omitted from all dataset submissions.

2. **Conversion to CoNLL-U fails if there are more than nine subsequent empty nodes** – this is why there are many 0 scores for the *heads_zeros* model at evaluation time.

3. **Insufficient sequence length** – Causes 0 results for *hbo_ptnk* dataset and very low results for *tr_itcc*.

We solved all the above-mentioned problems later,[1] and the improvement achieved is shown in the second part of Table 2.

[1]Note that test data evaluation is still available only through CodaLab submission, so the post-evaluation entries have exactly the same conditions as the regular ones, except for the extended deadline. We made only 4 test submissions overall, when the limit is 10.

5.2 General Discussion

Table 2 shows that our baseline system achieves satisfactory performance (over 60%) on half of the evaluated datasets. For most of the remaining datasets, the main problem was insufficient maximum sequence length (for details, see Section 5.3).

Joint pre-training helps, but the improvements are somewhat modest (mostly 1-4%). This is a very different result compared to the participating systems from previous years. One factor is the difference in datasets. The two smallest datasets in CorefUD: `en_parcorfull` and `de_parcorfull` were removed from this year's CRAC competition. Such small datasets typically see the largest gain from joint pre-training, because the models tend to overfit more easily without it. The second factor is the difference in model architecture. Previous results make use of Transformers with task-specific heads, but our system trains only an adapter. The difference here comes from the ability to leverage the pre-trained models' representations. A randomly initialized head has no connection to the knowledge from pre-training, while the adapted transformer can quickly adjust by reusing its latent knowledge.

After fixing all the evaluation issues, we achieve reasonable performance (over 60%) for almost all the datasets with a few exceptions. For both French datasets, our performance is relatively low. We believe the main reason is still in long sequences and long-distance coreferences. The last problematic dataset is Turkish, where we achieve significantly better results on the test set than on the development set. We believe there is an issue with a document in the development set, which contains just two documents.

5.3 Sequence Lengths and Non-Latin Scripts

In our original submission, we had issues with documents or entire datasets surpassing our training context length limit. This limit was originally set to 4096 to compromise between the practical feasibility of the training and processing enough documents to efficiently train the models. More extensive analysis of the actual dataset sequence lengths and tokenization, whose main results are shown in Table 3, shows that this proves problematic for certain datasets.

The average sample length in a majority of datasets within CorefUD fits well into our original context length limit. In all cases except for `fr_democrat`, the median samples happen to fit exactly when the average sample length does too, guaranteeing a suitable amount of data to sufficiently train our models. In the case of fr_democrat, the average is swayed heavily by exceedingly long samples, and the dataset is, in principle, trainable under these conditions as well.

The datasets with training issues due to sequence length issues are `cu_proiel`, `en_litbank`, `grc_proiel`, `hbo_ptnk`, and `tr_itcc`. In the case of `en_litbank` and `tr_itcc`, this can be resolved either by increasing the training sequence length up to 8,192 or by splitting the documents for training.

For `cu_proiel`, `grc_proiel`, and `hbo_ptnk`, the excessive sequence lengths can be attributed to using non-Latin scripts and vocabulary that was not prevalent in the training data of the tokenizer. All three datasets suffer from high number of subword tokens per word, with Hebrew in `hbo_ptnk` reaching 7.7 tokens per word. This comes from the fact that some of the scripts' code points do not have a dedicated token and fall back to byte encoding.

Context length limitations cause issues during inference as well. Having some documents that are truncated by a small amount for inference does not lower model performance as drastically as having a large amount of unused training documents. Truncated documents during inference will decrease the maximum achievable score proportionally to the truncated length, but missing training documents may lead to drastic over-fitting and near-zero scores. In addition, increasing the inference sequence length is less memory-intensive than increasing the training sequence length, and we manage to run inference at up to 8,192 input tokens and 16,384 output tokens while still recovering additional score points. Because long-context inference is much more practical than long-context training, we settled on running inference for entire documents and invested our time in other optimizations.

5.4 Effective Context Length

To determine how much context is truly necessary for coreference resolution in the CorefUD datasets, we investigate the distances between entity mentions within documents. We compute the distance between all consecutive pairs of mentions of each entity within each document. To match our results with the application, we use the outer bounds, from the beginning of the first mention to the end of the second mention.[2] The distribution of these distances across all datasets is heavily right-skewed. The median distance is 16 words, with partial medians spanning between 6 (`tr_itcc`) and 25 (`es_ancora`). The 90%, 95%, and 99% quantiles are 118, 220, and 728 words, respectively. The longest distance in any dataset is 12,398 words in `fr_democrat`.

These values suggest that most mentions of an entity are close together, but there are some long-distance dependencies that require large context windows. Generally, a sliding context window of 4096 tokens should be sufficient for 95-99% of most datasets if implemented carefully. This way, just about all mentions would have at least one other mention within their context window. However, the remaining 1-5% of mentions would still need a larger context window. Without a method

[2]Our processing of discontinuous mentions is simplified. Zero mentions are counted as full words. Each part of a discontinuous mention counts as a separate mention.

dataset name	toks/word	word length	max text	max label	mean text	mean label
ca_ancora	1.60	5.18	5,404	8,152	528.4	782.5
cs_pcedt	1.78	5.90	7,255	9,831	888.5	1,230.8
cs_pdt	1.84	5.85	5,231	8,415	473.7	761.4
cu_proiel	3.56	5.55	**42,169**	**56,978**	**15,507.5**	**21,134.8**
de_potsdamcc	1.70	6.24	420	746	319.0	503.6
en_gum	1.10	5.02	2,152	5,403	1,103.4	2,629.0
en_litbank	1.09	4.86	3,624	5,958	2,301.0	3,747.6
es_ancora	1.43	5.35	2,471	3,765	485.7	755.4
fr_ancor	1.34	4.90	**20,768**	**41,700**	1,362.7	2,679.0
fr_democrat	1.45	4.98	**23,161**	**51,495**	6,619.8	**14,166.1**
grc_proiel	3.53	5.87	**53,486**	**71,886**	**22,042.8**	**29,976.9**
hbo_ptnk	7.70	5.55	**10,317**	**11,876**	5,951.6	6,918.5
hi_hdtb	2.53	4.83	1,682	2,286	742.2	1,004.9
hu_korkor	2.67	6.55	1,493	1,844	683.2	861.9
hu_szegedkoref	2.28	5.77	4,152	4,836	715.6	905.0
ko_ecmt	2.49	3.98	4,433	6,433	817.9	1,230.9
lt_lcc	2.70	6.37	2,217	2,773	1,016.5	1,244.5
no_bokmaalnarc	1.71	5.46	**10,989**	**21,353**	1,221.1	2,356.5
no_nynorsknarc	1.81	5.50	4,812	8,846	932.5	1,743.1
pl_pcc	2.13	5.85	5,784	11,327	629.2	1,126.3
ru_rucor	1.83	5.93	6,449	8,514	1,562.2	1,987.0
tr_itcc	1.86	6.38	4,920	7,181	4,411.8	6,634.8

Table 3: All statistics are computed on the train split of the dataset, using the meta-llama/Llama-3.1-8B tokenizer. Token counts above 10,000 tokens are highlighted in bold red, samples that fit into our initial training context length are colored blue and underlined. The "toks/word" column contains the average number of tokens per word in the data. Because of how the text is pre-tokenized, punctuation such as periods and commas count as words as well. The "word length" column contains the mean word length in Unicode code points. The last four columns contain the maximum number of tokens in a sample and the average sample length in tokens for both the model input and completion. Note that the training sequences actually consist of the concatenation of both sequences along with additional overhead for the completion marker.

to recover broken chains in long documents, these long-distance mentions could account for a disproportionately large portion of the final score.

The main factor in long context mentions and document length appears to be the type and source of the data. The longest distance comes from the short story "Sarrasine" by Honoré de Balzac, which is present in `fr_democrat`. The entity in question refers to the Lanty family and has many mentions throughout the story.

This analysis suggests that while most coreference relations occur within manageable context windows, a certain portion of datasets contain long-distance dependencies that prove challenging to our approach. These long-distance coreferences are especially prevalent in both French datasets. In contrast, other datasets with shorter average document length tend to have their mentions closer together. This raises the question of whether modeling long-distance mentions separately would improve efficiency and possibly performance, or whether simply scaling the context window is more practical.

5.5 Dataset Discrimination Capabilities

Our experiments included joint models trained on a mixture of all datasets without dataset-specific fine-tuning. We never invested the resources to fully evaluate these models. Partial results suggest that this general version of the model is typically weaker than the specialized models trained on each dataset individually. Investigating the joint approach gives insights into how a single model is able to generalize between datasets.

The originally employed prompt template does not explicitly contain information about which

system	ca_ancora	cs_pcedt	cs_pdt	cu_proiel	de_potsdam	en_gum	en_litbank	es_ancora	fr_democrat	fr_ancor	grc_proiel	hbo_ptnk	hi_hdtb	hu_korkor	hu_szeged	ko_ecmt	lt_lcc	no_bokmaal	no_nynorsk	pl_pcc	ru_rucor	tr_itcc	average
corpipe-best	**84.20**	**76.94**	80.64	62.63	**78.71**	77.38	80.91	84.64	**80.03**	**73.26**	**76.80**	67.27	81.90	**70.24**	**73.16**	69.21	**82.61**	**80.10**	**80.74**	**80.31**	79.71	67.51	76.77
corpipe-ens	84.09	76.92	**81.08**	**64.20**	77.89	**77.48**	80.04	**85.07**	79.64	72.51	76.14	66.75	**81.99**	69.72	73.09	**69.44**	81.62	80.09	79.66	80.29	80.05	65.50	76.51
corpipe-1	83.25	75.94	80.22	62.66	76.90	76.54	80.09	84.23	78.97	71.93	76.18	66.02	80.64	68.11	71.86	67.59	80.22	79.20	80.33	79.23	78.30	66.69	75.69
ours_post*	<u>81.35</u>	<u>72.12</u>	<u>74.97</u>	56.69	<u>69.78</u>	<u>75.76</u>	82.67	<u>82.01</u>	58.56	49.14	60.53	48.20	<u>77.42</u>	<u>65.76</u>	<u>69.81</u>	67.83	<u>69.17</u>	<u>76.78</u>	<u>72.06</u>	<u>76.56</u>	**84.41**	**69.09**	70.03
stanza	80.30	72.83	74.49	37.95	77.97	70.74	72.96	79.53	69.75	63.22	54.07	63.57	78.87	65.32	68.61	64.86	78.81	74.93	75.32	74.10	78.42	49.48	69.37
antoine.b	68.92	61.85	62.88	39.95	63.95	65.20	72.12	68.82	69.00	65.02	54.08	57.83	72.11	52.52	60.39	60.59	74.21	69.80	70.30	65.00	67.80	42.80	62.96
oseminck	73.45	<u>65.12</u>	<u>71.33</u>	58.25	59.60	58.73	69.01	<u>74.43</u>	66.74	<u>60.43</u>	<u>65.75</u>	43.96	56.36	<u>52.53</u>	59.82	63.04	62.55	64.74	61.63	<u>72.55</u>	68.79	<u>56.23</u>	62.96
moizsajid	60.87	51.36	54.30	<u>58.48</u>	48.74	69.78	70.38	61.75	<u>71.94</u>	57.59	57.85	**80.15**	71.32	43.49	52.27	66.05	59.16	72.76	68.86	70.83	71.40	39.00	61.74
PuxAI	68.01	56.94	62.96	43.74	57.41	61.71	69.12	70.52	63.77	61.54	47.86	45.31	66.85	50.58	61.61	50.32	<u>65.35</u>	65.18	63.00	66.55	67.59	56.06	60.09
ours	<u>79.17</u>	61.02	68.17	25.34	<u>67.63</u>	<u>73.64</u>	**84.05**	73.63	58.56	49.14	47.64	0.00	<u>75.84</u>	38.91	<u>67.32</u>	<u>68.30</u>	63.44	<u>73.77</u>	<u>71.96</u>	64.49	**80.12**	24.31	59.84
baseline-gz	70.53	68.00	67.43	27.69	57.90	64.97	66.59	71.71	65.37	56.27	29.78	23.77	69.86	49.86	59.05	63.04	69.32	66.11	66.76	65.63	63.39	47.14	58.64
baseline	69.94	57.32	63.20	24.10	57.90	64.96	66.59	71.32	65.37	56.27	26.98	23.77	69.86	46.61	58.34	48.34	69.32	66.11	66.76	64.08	63.39	40.06	56.39

Table 4: Results of all competing models in both tracks on the test set. Best overall scores are bold. Best scores within the LLM track are underlined (if they are not already bold). Row marked with * shows post-evaluation experiments. Post-evaluation results are also highlighted in the same manner, in addition to the official results. LLM track systems have names in bold.

dataset the current sample comes from. There are differences between how the individual datasets are annotated, and using a model trained on one while evaluating on another usually degrades model performance significantly. If the model did not know which annotation rule set to apply to each sample, it would be at a disadvantage. There are two options: either the fine-tuned LLMs already implicitly model the distinction between the datasets, or their performance can be further improved by giving them this information.

We hypothesize that it is possible that the different datasets are easily distinguishable due to factors like the length or domain of the document, or the tokenization used. Of the 22 datasets, only 5 pairs share language: `cs_pcedt` and `cs_pdt`, `en_gum` and `en_litbank`, `fr_ancor` and `fr_democrat`, `hu_korkor` and `hu_szegedkoref`, `no_bokmaalnarc` and `no_nynorsknarc`. We train a model to predict the dataset name before completing the annotations and find that it achieves 100% accuracy in classifying all datasets' evaluation splits. This result confirms that it is possible to distinguish all datasets based on the input text alone and that, when necessary, the LLM will implicitly utilize this information.

5.6 Final Results

Table 4 shows the final results on test sets. The column *ours_post** shows scores from our post-evaluation experiments, which are not a part of the official competition. From the results, we can see that though our system ended up in the last place, we achieved the best results within the LLM track for 10 datasets out of 22, which is the highest number of datasets won by a single system in this track. The reason for our low average score was in dataset-specific problems, which led to very low performance on these datasets. After fixing all issues in the post-evaluation phase, our system outperformed all other systems within the LLM track by a large margin. It would take the fourth place overall and become the second unique system (the first three *Corpipe* entries are variants of the same system by a single team).

6 Conclusion

We proposed a Llama-based text-to-text multilingual coreference resolution system with headword mention representation and joint pre-training for the CRAC 2025 shared task. We provide an extended analysis of different model configurations.

We found that generative tagging approaches struggle with large documents due to limited sequence length when running an open-weight model on a single machine. Languages with non-Latin scripts often tokenize inefficiently, leading to very long sequences.

Our system ended up in last place. However, we achieved the best results on 10 datasets out of 22. The main problem of our submission was the very low performance for a small subset of datasets, which was caused by some mistakes we did not manage to fix on time. After fixing all identified issues in the post-evaluation phase, we achieved the best results in the LLM track by a large margin, and we even outperformed some systems in the unconstrained track.

Considering the relatively small size of our model, we believe LLMs can achieve state-of-the-art results on CorefUD in the near future.

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.

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.

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

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.

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.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561, Rome, Italy. AAAI Press.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtskỳ, Amir Zeldes, and Daniel Zeman. 2022. Corefud 1.0: Coreference meets universal dependencies. In *Proceedings of LREC*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

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. 2025a. Findings of the fourth shared task on multilingual coreference resolution. In *Proceedings of the eighth Workshop on Computational Models of Reference, Anaphora and Coreference*, Suzhou. Association for Computational Linguistics.

Michal Novák, Martin Popel, Daniel Zeman, Zdeněk Žabokrtský, Anna Nedoluzhko, Kutay Acar, David Bamman, Peter Bourgonje, Silvie Cinková, Hanne Eckhoff, Gülşen Cebiroğlu Eryiğit, Jan Hajič, Christian Hardmeier, Dag Haug, Tollef Jørgensen, Andre Kåsen, Pauline Krielke, Frédéric Landragin, Ekaterina Lapshinova-Koltunski, and 23 others. 2025b. Coreference in universal dependencies 1.3 (CorefUD 1.3). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Ondřej Pražák and Miloslav Konopik. 2022. End-to-end multilingual coreference resolution with mention head prediction. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 23–27, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Ondrej Prazak and Miloslav Konopík. 2024. End-to-end multilingual coreference resolution with head-word mention representation. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 107–113, Miami. Association for Computational Linguistics.

Ondřej Pražák, Miloslav Konopík, and Pavel Král. 2024. Exploring multiple strategies to improve multilingual coreference resolution in corefud. *arXiv preprint arXiv:2408.16893*.

Ondřej Pražák, Miloslav Konopík, and Jakub Sido. 2021. Multilingual coreference resolution with harmonized annotations. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1119–1123.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

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.

Patrik Stano and Aleš Horák. 2025. Evaluating prompt-based and fine-tuned approaches to czech anaphora resolution. *Preprint*, arXiv:2506.18091.

Milan Straka. 2023. ÚFAL CorPipe at CRAC 2023: Larger context improves multilingual coreference resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 41–51, Singapore. Association for Computational Linguistics.

Liyan Xu and Jinho D. Choi. 2020. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8527–8533, Online. Association for Computational Linguistics.

Wenzheng Zhang, Sam Wiseman, and Karl Stratos. 2023. Seq2seq is all you need for coreference resolution. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11493–11504, Singapore. Association for Computational Linguistics.