PortNLP at CRAC 2026: QLoRA Fine-Tuning with Bounded Entity Registry for Multilingual Coreference Resolution

Amber Shore, Russell Scheinberg, Malini Nagasundaram, Ameeta Agrawal
Portland State University
{ashore, rschein2, kamalan, ameeta}@pdx.edu

Abstract

We describe PortNLP's submission to the CRAC 2026 Shared Task on Multilingual Coreference Resolution (LLM track). Our system fine-tunes Qwen 3 14B with QLoRA on CorefUD 1.4 gold annotations across 27 corpora spanning 19 languages. Documents are processed in 500-700 character chunks with a bounded rolling context consisting of 500 characters of recent annotated text and a scored entity registry that tracks up to 30 active entities via a frequency×recency decay formula. We employ data augmentation and language-aware sampling strategies to handle typological and data-size diversity. Our system achieves 68.69 CoNLL F1 averaged across all 27 test corpora. We additionally present probing experiments on the LoRA adapter's internal representations, finding that coreference signal is concentrated in attention value projections rather than MLP modules, with the strongest readout at the earliest transformer layer.

1 Introduction

This paper describes PortNLP's system submitted to the CRAC 2026 Shared Task on Multilingual Coreference Resolution (Novák et al., 2026a). The system was built to annotate coreference in 19 languages, including pro-drop languages, across 27 corpora. Our system achieved 3rd place in the official ranking on the LLM track with an overall CoNLL F1 score of 68.69.

Our approach involved fine-tuning Qwen 3 14B to directly generate annotated text in a single pass. The core challenge is that many corpora contain long documents that make it difficult to maintain coherent entity tracking in a single pass. We address this through a rolling-window chunking strategy with a bounded entity registry for cross-chunk coherence.

Our contributions are: (1) a bounded context design informed by GLaRef@CRAC2025 (Seminck et al., 2025) that combines recent annotated text with a frequency×recency-scored entity registry; (2) language-specific few-shot examples and zero anaphora hints tailored to each of the 19 languages; (3) probing experiments revealing that coreference signal in the LoRA adapter is concentrated in attention value projections, particularly at the earliest transformer layer.

2 Background

The CRAC 2026 shared task uses the CorefUD 1.4 dataset (Novák et al., 2026b), which provides coreference annotations in the Universal Dependencies framework across 27 corpora and 19 languages. The LLM track permits in-context learning, fine-tuning, and prompt tuning.

Our initial strategy used the rolling-window chunking approach and DSPy-based agent annotation loop with API-served LLMs. Because of performance and efficiency issues with that approach, we pivoted to using supervised fine-tuning, a scored entity registry, and language-specific prompt engineering.

Our context design is directly informed by GLaRef@CRAC2025 (Seminck et al., 2025), which systematically evaluated context strategies for LLM-based coreference. Their key findings were: (1) providing full prior annotated context was "disastrously bad"; (2) 500 characters of recent context was optimal; and (3) shorter prompts consistently outperformed longer ones. We adopt their recommended 500-character budget for recent text and extend it with a scored entity registry for long-range entity tracking.

3 System Description

3.1 Overview

As required in the shared task guidelines, our system annotates documents with inline entity tags in a single generative pass. Single-word mentions

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 193–198

are tagged as `word|[eN,eN]`, multi-word spans as `first|[eN ... last eN]`, and zero anaphora as `##pronoun|[eN,eN]` after the parent token. Entity IDs are arbitrary sequential integers; only consistency within a document matters.

Documents are split into 500-700 character chunks at sentence boundaries, and each chunk is annotated independently with access to bounded context from prior chunks. Cross-chunk entity coherence is maintained through a scored entity registry.

3.2 Training Data Construction

Character-Based Chunking. Documents are chunked at the character level to ensure stable prompt sizes across typologically diverse languages. Boundary detection uses a tiered strategy: sentence-ending punctuation preferred, falling back to secondary punctuation then whitespace. This handles languages with sparse punctuation such as Ancient Greek and Old Church Slavonic.

Bounded Context. For each chunk $K > 0$, the model receives two forms of context from prior chunks:

1. **Recent annotated text**: the last 500 characters of concatenated gold output from prior chunks, trimmed at whitespace boundaries.

2. **Scored entity registry**: the top entities ranked by frequency$\times$recency, bounded to 30 entities and 400 characters. Each entry takes the form `eN: "representative"`, where the representative is the first non-pronoun surface form. The scoring formula is:

$$\text{score}(e) = \text{count}(e) \times 0.9^{\text{chunks\_since\_last\_seen}(e)} \tag{1}$$

This naturally retains protagonists (an entity with 15 mentions absent for 8 chunks scores $15 \times 0.9^8 = 6.4$) while discarding one-off mentions ($1 \times 0.9^2 = 0.81$ after just 2 chunks).

Training uses teacher forcing: gold annotations serve as context for subsequent chunks. This creates a train/test mismatch at inference, where the model's own predictions (potentially erroneous) become context. We mitigate this through entity ID permutation augmentation and through the bounded nature of the context, which limits error propagation.

Entity ID Permutation Augmentation. GLaRef identified the "ID restart problem" where models restart entity numbering from e1 at chunk boundaries. To address this, we generate one permuted copy of each training document where all entity IDs are randomly remapped via a bijection (e.g., e1$\rightarrow$e24, e2$\rightarrow$e7). This teaches the model to rely on the entity registry rather than memorizing fixed ID assignments.

Language Sampling. The raw dataset is heavily skewed: Czech accounts for 39% of training examples due to four constituent corpora. We apply temperature-scaled sampling at the document level with α=0.5 (square-root proportional), reducing Czech's share to approximately 18% while preserving all data from low-resource languages.

Prompt Design. The system prompt is brief ($\sim$200 characters), following GLaRef's finding that shorter prompts outperform elaborate descriptions. The user message includes: (1) full language name; (2) a zero anaphora hint for the 8 languages that annotate it (Catalan, Czech, Old Church Slavonic, Spanish, Ancient Greek, Hungarian, Polish, Turkish); (3) three language-specific few-shot examples extracted from gold data demonstrating single-word mentions, multi-word spans, and nested mentions; (4) the bounded context block; and (5) the text chunk to annotate.

3.3 Model and Training

Model Selection. We use Qwen 3 14B as our base model for its strong multilingual coverage and text-only architecture. Qwen 3's "thinking mode" is disabled throughout, as the training data contains no reasoning traces.

QLoRA Configuration. The model is loaded in 4-bit NF4 quantization with double quantization and bfloat16 compute. LoRA adapters (rank 32, α=64, dropout 0.05) target all attention modules (q/k/v/o_proj) and MLP modules (gate/up/down_proj), totaling approximately 128.5M trainable parameters.

Training. We train for 2 epochs with effective batch size 16 (per-device 8, gradient accumulation 2), learning rate 2×10^{-4} with cosine schedule and 5% warmup. Loss is computed only on annotation output tokens; prompt tokens are masked with label value -100. Training was performed on a single NVIDIA A100 40GB GPU using Unsloth for kernel-level optimization.

We noted that Unsloth's patched SFTTrainer produces unreliable evaluation loss due to silent sequence packing; we relied on downstream task evaluation instead.

3.4 Inference Pipeline

The LoRA adapter is merged into the base model weights, and the merged model is served with vLLM for high-throughput inference. Documents are processed in parallel "rounds": in each round, all active documents' current chunks are batched together for generation. After each round, each document's entity tracker is updated from the model's output and its rolling context is rebuilt. This achieves approximately 160× throughput improvement over sequential HuggingFace generation, reducing inference on 864 documents from an estimated 50 hours to under 1 hour.

Generated annotations are converted to CoNLL-U format using `text2text-coref` (Pražák, 2024) in two steps: a cleaning pass that aligns model output with the CoNLL-U skeleton tokenization, and a conversion pass that maps inline annotations onto the skeleton. Across all 27 test corpora, only 56 malformed tags were detected (orphaned opening or closing brackets), representing a 99.9% well-formedness rate.

3.5 Post-Processing

We applied minimal post-processing: (1) malformed tag removal using a validated fix script; and (2) stripping hallucinated zero anaphora tokens from `en_gum`, where the model produced hundreds of repetitive `##you` tokens in one document. This hallucination pattern was triggered by the rarity of zero anaphora in English training data (92 markers vs. 51,433 in Czech `cs_pdtsc`).

We additionally experimented with singleton merging using Claude as an LLM judge, but this yielded negligible improvement (+0.03 F1), suggesting most detected singletons were genuinely singleton entities.

4 Results

4.1 Main Results

Our final submission achieves 68.69 CoNLL F1 averaged across all 27 test corpora. Table 1 shows per-corpus results. For each corpus, we evaluated both epoch 2 and epoch 3 checkpoints independently and selected the higher-scoring result, yielding epoch 2 as optimal for 21 corpora and epoch 3 for the remaining 6.

Corpus	F1	Train docs
ru_rucor	81.24	145
en_litbank	78.37	80
pl_pcc	76.83	1463
hi_hdtb	75.58	142
es_ancora	75.36	1080
en_fantasycoref	74.80	171
cs_pdt	74.10	2533
ca_ancora	73.68	1011
nl_openboek	72.93	5
hbo_ptnk	72.69	10
no_bokmaalnarc	72.13	284
no_nynorsknarc	72.07	336
cs_pcedt	71.55	1875
grc_proiel	71.08	9
en_gum	70.60	177
fr_ancor	70.55	365
cs_pdtsc	69.77	1271
ko_ecmt	69.48	1204
lt_lcc	68.08	80
fr_democrat	68.08	50
de_potsdamcc	67.94	142
tr_itcc	62.97	19
hu_szegedkoref	62.53	320
hu_korkor	58.58	76
cu_proiel	57.91	11
fr_litbankfr	53.40	21
la_coreflat	44.79	8
Average	**68.69**	

Table 1: Per-corpus CoNLL F1 on the test set, sorted by score. The best submission uses epoch 2 results for 21 corpora and epoch 3 for 6 corpora where it improved (cherry-pick strategy).

4.2 Analysis

Strong and Weak Corpora. Russian (81.2) and English LitBank (78.4) score highest; both are narrative texts with clear, consistent entity naming patterns. The weakest corpora are Latin (44.8, only 8 training documents), French LitBank (53.4, very long documents of 60–72K characters), and Old Church Slavonic (57.9, 11 training documents). We find no strong correlation between document length and score; notably, Dutch `nl_openboek` has an average document length of 60K characters yet scores 72.9, while Latin at 4K characters per document scores only 44.8. The weak points are language-specific rather than architectural.

Zero Anaphora Challenge. Zero anaphora density strongly predicts difficulty. Turkish (`tr_itcc`) has the highest density at 19 zero anaphora markers per 1000 characters and scores only 63.0. Czech spoken dialogue (`cs_pdtsc`) has 11/1K and initially scored 52.1, though an additional training epoch dramatically improved this to 69.8, the sin-

gle largest per-corpus improvement across all experiments. In contrast, languages without zero anaphora annotations generally score above 70.

Training Epoch Comparison. A third training epoch yielded mixed results: 6 of 27 corpora improved (most notably `cs_pdtsc`: +17.7), while others remained unchanged or slightly degraded. This suggests the model had largely converged after 2 epochs for most languages, with spoken Czech as a notable exception requiring additional exposure.

Precision vs. Recall. On two corpora evaluated locally against gold annotations, the model showed distinct patterns:

- **English LitBank**: High recall (MUC R=91.2%) but lower precision (MUC P=83.8%), overpredicting mentions by ∼10%. CEAFe was the weakest metric (57.99), indicating spurious entity clusters.
- **German PotsdamCC**: More balanced precision and recall. CEAFe was much stronger (71.48), confirming fewer spurious entities.

5 Discussion

5.1 Design Decisions

Bounded vs. Unbounded Context. Our context budget of 900 characters (500 text + 400 registry) was informed by GLaRef's finding that more context hurts performance. The entity registry extends GLaRef's approach by providing long-range entity tracking without expanding the context window.

Teacher Forcing Trade-off. Training with gold context creates a distribution shift at inference, where the model must work with its own (possibly erroneous) predictions. We mitigate this through entity ID permutation augmentation (teaching the model that IDs are arbitrary), bounded context (limiting how far errors can propagate), and the registry's decay function (allowing recovery from early mistakes as new evidence accumulates).

6 Conclusion

We presented PortNLP's submission to CRAC 2026, achieving 68.69 CoNLL F1 across 27 corpora using QLoRA fine-tuned Qwen 3 14B with a bounded entity registry. The system demonstrates that a single fine-tuned 14B-parameter model can handle multilingual coreference across 19 typologically diverse languages, producing well-formed annotations at a 99.9% rate. The main performance bottlenecks are zero anaphora generation for high-density pro-drop languages and error accumulation in very long documents both of which suggest that hybrid approaches combining generative annotation with structured post-processing may be most promising for future work.

Probing experiments on the LoRA adapter's representations (Appendix A) reveal that coreference signal is concentrated in attention value projections (v_proj: 75.7% of learned weight), with the strongest readout at layer 0. We interpret this as evidence that the adapter primarily learns entity-aware lexical canonicalization at the earliest layer, with contextual disambiguation distributed across deeper layers. Causal ablation experiments are needed to confirm whether this readout pattern reflects the model's actual computation or a convenient probe shortcut (Hewitt and Liang, 2019).

Limitations

Our system has several limitations. Teacher forcing during training creates a distribution shift at inference where the model's own errors compound across chunks; this particularly affects long documents (60K+ characters split into 80+ chunks). The model hallucinates zero anaphora tokens for languages where this phenomenon was rare in training data (e.g., English GUM with only 92 training markers). We did not perform systematic hyperparameter tuning on the dev set, all hyperparameters were set based on prior work and standard defaults. Our probing experiments identify readout loci in the adapter's representations rather than causal mechanisms; the planned causal ablation experiments that would distinguish these were not completed. Finally, the completion-only loss still dilutes the training signal across "copy" tokens that dominate the output; tag-only loss masking would focus gradients entirely on annotation decisions.

Acknowledgments

We thank the CRAC 2026 shared task organizers for preparing the data and evaluation infrastructure.

References

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. 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).

Ondřej Pražák. 2024. text2text-coref: Converting inline coreference annotations to CoNLL-U. https://github.com/ondfa/text2text-coref.

Olga Seminck, Antoine Bourgois, Yoann Dupont, Mathieu Dehouck, and Marine Delaborde. 2025. GLaRef@CRAC2025: Should we transform coreference resolution into a text generation task? In *Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 119–129, Suzhou, China. Association for Computational Linguistics.

A Probing Experiments

To understand where coreference knowledge resides within the fine-tuned model, we conducted probing experiments on the LoRA adapter's internal representations.

A.1 Task and Setup

We formulate probing as online mention-to-entity classification, mirroring the SFT training regime exactly. For each mention in a chunk, a probe scores it against all K currently visible entities in the registry plus a learned NEW_ENTITY option, producing $K+1$ logits. The probe uses the same registry parameters as training (decay 0.9, max 30 entities).

Mention spans are represented as $[\text{first}; \text{last}; \text{mean}]$ pooled vectors from token-level features. Entity prototypes use exponential-decay running averages of prior mention vectors. The probe architecture computes $[\mathbf{m}; \mathbf{e}; \mathbf{m} \odot \mathbf{e}; |\mathbf{m} - \mathbf{e}|; \text{meta}]$ through a 2-layer MLP (hidden 512, projection 256), where meta consists of two scalar entity features (log mention count and recency in chunks), through a 2-layer MLP (hidden 512, projection 256).

We extract features from 200 gold training documents (10,108 mentions: 6,549 existing-entity, 3,559 new-entity) and train for 5–10 epochs with Adam (lr=10^{-3}).

Strategy	Overall	Existing	New
A (base)	14.3	19.0	5.6
B-mlp	42.2	33.0	59.1
B-gated	57.5	45.3	79.9
B-full	66.0	59.7	77.6
B-attn	75.7	69.9	86.5
B-layer	**76.5**	**69.9**	**88.6**

Table 2: Probing accuracy (%) for mention-to-entity classification. "Existing" = linking to a visible entity; "New" = predicting NEW_ENTITY.

A.2 Strategies

We compare six strategies:

- **A**: Base model hidden states (adapter disabled)
- **B-mlp**: LoRA Δy from MLP modules only
- **B-gated**: All 7 module types with learned softmax gating
- **B-full**: All LoRA modules averaged
- **B-attn**: Attention modules only (q/k/v/o_proj)
- **B-layer**: v_proj + k_proj with per-layer gating (80 keys)

For strategies B-mlp, B-full, and B-attn, LoRA deltas are averaged across all 40 layers before span pooling.

A.3 Results

The gap between Strategy A (14.3%) and any B variant confirms that the adapter learned genuine

coreference representations absent from the pretrained model. Attention modules alone (B-attn: 75.7%) outperform all modules averaged (B-full: 66.0%), indicating that MLP features add noise.

A.4 Module-Type and Layer Analysis

The B-gated experiment learns softmax weights over 7 module types, revealing a clear hierarchy: v_proj (0.757) $\gg$ k_proj (0.191) $\gg$ o_proj (0.041), with MLP modules near zero. Value projections carry the dominant signal, consistent with their architectural role of determining *what content* flows through attention.

The B-layer experiment provides finer granularity: **v_proj at layer 0 receives 70.1% of the total weight**, with v_proj.layer_1 (7.5%) and k_proj.layer_1 (3.1%) as distant runners-up. This surprising result suggests the adapter's primary mechanism is creating "entity-aware embeddings" at the earliest layer modifications that then propagate through all subsequent layers.

The probe identifies a readout locus, not a compute locus (Hewitt and Liang, 2019), a distinction we did not test with causal ablation. The extreme concentration at layer 0 likely reflects at least in part lexical canonicalization surface form matching and pronoun class identification rather than deep coreference reasoning. A mention-type-stratified analysis (proper names vs. pronouns vs. zero anaphora) would test this hypothesis but was not completed.

A.5 Zero Anaphora Probe

We additionally trained a per-token binary classifier predicting whether a zero anaphora token (##) should follow each word, using v_proj + k_proj features. On gold data from zero-anaphora languages, this achieved 95.6% accuracy with 98.4% recall but only 25.7% precision, indicating that the adapter features encode zero anaphora positions but that the class imbalance ($\sim$5% positive rate) makes the probe overly liberal.