# G2 - Evaluating Document Context and CRF Decoding in ModernBERT for CoNLL-2003 NER

Arda Unal
*University of Virginia*
*School of Data Science*
Charlottesville, Virginia
ene2qt@virginia.edu

Gregory Miller
*University of Virginia*
*School of Data Science*
Charlottesville, Virginia
aem2cz@virginia.edu

John Z. Karlovich
*University of Virginia*
*School of Data Science*
Charlottesville, Virginia
ddf4me@virginia.edu

Sean Hersee
*University of Virginia*
*School of Data Science*
Charlottesville, Virginia
cjf4xv@virginia.edu

***Abstract*—We evaluate whether document-level context and CRF decoding provide additive or synergistic gains over sentence-level ModernBERT for named entity recognition on CoNLL-2003. Using a $2 \times 2$ factorial ablation (context on/off, CRF on/off), we compare entity-level F1 across all configurations and analyze which entity types benefit most from each modification. Code and configurations are available on https://github.com/zach-karlovich/modernbert-ner-ablation.**

***Index Terms*—Named entity recognition, CoNLL-2003, ModernBERT, conditional random field, document-level context, ablation study**

## I. Introduction

Named Entity Recognition (NER) is a core task in natural language processing that involves identifying and classifying named entities in text, such as persons, organizations, and locations. Over the past two decades, NER systems have evolved from rule-based and dictionary lookup approaches through classical sequence models (CRFs), neural sequence models such as the Bidirectional Long Short-Term Memory with CRF decoding (BiLSTM-CRF) [1], and pretrained transformers (BERT [2]) to modern encoder architectures such as ModernBERT [3].

ModernBERT introduces several architectural advances to the BERT paradigm, including rotary positional embeddings, alternating global and local attention, and a native 8,192-token context window; however, it has not been widely benchmarked on CoNLL-2003. Two modifications remain underexplored for this architecture: (1) exploiting the long context window by feeding entire documents rather than isolated sentences, and (2) adding a CRF decoding layer to enforce structured label constraints at inference time.

Our central research question is: **Do document-level context and structured CRF decoding provide additive or synergistic gains over sentence-level ModernBERT on CoNLL-2003?** We answer this through a $2 \times 2$ factorial ablation, varying document context (on/off) and CRF head (on/off), and report entity-level and per-entity-type F1 across all four configurations.

## II. Dataset

We use the CoNLL-2003 English NER dataset [4], consisting of Reuters news articles annotated with four entity types: persons (PER), organizations (ORG), locations (LOC), and miscellaneous names (MISC), using Inside-Outside-Beginning (IOB2) tagging [5]. The dataset contains 946 / 216 / 231 documents (train/dev/test), 14,987 / 3,466 / 3,684 raw sentence-boundary counts, and 203,621 / 51,362 / 46,435 tokens. Table I reports span counts by entity type; notably, MISC spans occur at roughly half the frequency of the other three types.

We source the original-format data files (`eng.train`, `eng.testa`, `eng.testb`) via Kaggle[1] rather than the HuggingFace `datasets` library, which omits `-DOCSTART-` boundaries. Preserving these document markers is necessary for our document-context experiments. Our verification scripts yield parsed sentence counts of 14,041 / 3,250 / 3,453 and document counts of 946 / 216 / 231 (train/dev/test). For completeness, the larger 14,987 / 3,466 / 3,684 figures correspond to raw sentence-boundary counts in the source files, which include document-separator formatting boundaries and are not parsed training/evaluation units used by our models.

TABLE I: Named entity span counts by type and split [4].

| Entity | Train | Dev | Test |
|---|---|---|---|
| PER | 6,600 | 1,842 | 1,617 |
| ORG | 6,321 | 1,341 | 1,661 |
| LOC | 7,140 | 1,837 | 1,668 |
| MISC | 3,438 | 922 | 702 |

## III. Models

We evaluate five models that reflect how NER methods have evolved over time on the CoNLL-2003 dataset [4].

The simplest is a dictionary lookup method, which matches tokens against a predefined entity list without any learned features. This approach scored 59.61 F1 on the CoNLL-2003 test set [4], and serves as a lower bound for what a machine learning model should be expected to surpass.

Next is a vanilla Conditional Random Field (CRF), a classical sequence labeling model that uses hand-crafted features such as word shape, capitalization, and surrounding context to predict entity tags. We implement a CRF using sklearn-crfsuite with feature seat from Lample et al. [1].

[1]https://www.kaggle.com/datasets/juliangarratt/conll2003-dataset

The BiLSTM-CRF from Lample et al. [1] uses a bidirectional LSTM to capture context from both directions and a CRF layer to enforce valid tag sequences at inference time. This model achieved 90.94 F1 on CoNLL-2003 and was widely considered the standard neural NER approach before transformers became mainstream.

BERT [2], fine-tuned for token classification, achieves 92.4 F1 on CoNLL-2003. The jump from BiLSTM-CRF to BERT highlights how much pre-trained contextual representations improved NER performance.

Finally, ModernBERT [3] is a recently released encoder-only model that brings several years of architectural advances from the decoder-only LLM literature into the BERT paradigm. Notable changes include rotary positional embeddings, alternating global and local attention, and hardware-aware model design [6] that maximizes GPU utilization. ModernBERT-base has 22 layers and 149 million parameters, was trained on 2 trillion tokens with a native 8,192 sequence length, and surpassed DeBERTa-v3-base [7] on General Language Understanding Evaluation (GLUE) becoming the first masked language model (MLM) trained encoder to do so. We include it to explore how a cutting-edge model performs on an established NER benchmark, and whether recent architectural advances translate to meaningful gains over BERT.

## IV. Proposed Method

We evaluate two modifications to the sentence-level ModernBERT methodology on the CoNLL-2003 dataset: (1) document-level context via `-DOCSTART-` segmentation and (2) a CRF decoding head. We test whether these changes improve span-level F1 relative to the sentence-level softmax baseline. The standard approach for ModernBERT when handling the CoNLL-2003 dataset isolates and processes each sentence independently. In reviewing the dataset we noted that the 8,192-token sequence limit is significantly underutilized. We therefore incorporate same-document neighboring sentences into each training instance while preserving sentence-level supervision.

For document-context configurations, we segment the dataset using the `-DOCSTART-` boundaries and pack neighboring sentences into a shared context budget. This increases available cross-sentence context for entity disambiguation while keeping supervision anchored to the target sentence. When an article exceeds the token budget, we apply sliding-window chunking with overlap to preserve continuity across adjacent chunks.

We also evaluate a CRF head on top of ModernBERT to test whether explicit transition modeling improves sequence consistency beyond token-level softmax decoding. Although transformer representations already encode dependencies, CRF decoding may still help at entity boundaries and BIO transition points. We therefore include CRF-only and document-context+CRF settings to measure whether structured decoding provides incremental benefits under this pipeline.

### A. Subword Label Alignment

ModernBERT uses Byte Pair Encoding (BPE) tokenization [3] rather than BERT's WordPiece [2]. We use a first-subtoken alignment strategy [8]: for each word, only the first subword receives the BIO label, and continuation subwords are masked. Alignment is implemented with `word_ids()`, which is tokenizer-agnostic and supports both BPE and WordPiece.

To verify alignment empirically, we compared tokenization of the CoNLL-2003 training split under both tokenizers: `bert-base-cased` (WordPiece) splits approximately 16.8% of words into multiple subwords (mean 1.34 subwords per word), while `answerdotai/ModernBERT-base` (BPE) splits 30.9% (mean 1.48 subwords per word). Because supervision is applied only at first subtokens, tokenizer-specific split behavior can confound direct cross-tokenizer F1 comparisons; accordingly, our ModernBERT ablation varies context window and decoding head while keeping tokenizer/model family fixed.

For the softmax configurations (sentence-level and document-level without CRF), continuation subtokens are assigned a label of `-100` and ignored by the loss function.

CRF configurations require a valid tag at each unmasked token position; we therefore use a dense-label formulation and collapse predictions back to word-level labels for seqeval reporting.

### B. Sliding-Window

For document-context configurations, we apply word-aligned sliding-window packing with a target sequence length of 8,192 tokens and an effective content budget of 8,192 minus special tokens. We use 128-token overlap when sequence lengths exceed the budget; because windows are word-aligned, the realized overlap may be slightly smaller after word-aligning the window boundary. This preserves token-label alignment across window boundaries and maintains contextual continuity between adjacent chunks.

In CoNLL-2003, however, no training document exceeded the effective content budget in our verification runs, so document-context examples remained single-window in practice. As a result, this study evaluates in-budget document-context behavior and does not directly measure performance in true multi-window long-document regimes.

### C. CRF Loss Formulation

Following the CRF framework introduced by Lafferty et al. [9], we replace the standard per-token softmax classification head with a linear-chain CRF layer that models the conditional probability of an entire label sequence $\mathbf{y} = (y_1, \ldots, y_T)$ given an input sequence $\mathbf{x} = (x_1, \ldots, x_T)$.

For a given input-label pair, the CRF defines a sequence-level score as the sum of emission scores and transition scores along the label path:

$$S(\mathbf{x}, \mathbf{y}) = \sum_{t=1}^{T} \big[E(x_t, y_t) + A(y_{t-1}, y_t)\big], \tag{1}$$

where $E(x_t, y_t)$ is the emission score from the Transformer encoder and $A(y_{t-1}, y_t)$ is the learned transition score between tags.

The conditional probability of the label sequence is obtained by normalizing over all possible tag sequences $\tilde{\mathbf{y}} \in \mathcal{Y}^T$:

$$P(\mathbf{y} \mid \mathbf{x}) = \frac{\exp\big(S(\mathbf{x}, \mathbf{y})\big)}{Z(\mathbf{x})}, \tag{2}$$

where the partition function $Z(\mathbf{x})$ is defined as:

$$Z(\mathbf{x}) = \sum_{\tilde{\mathbf{y}} \in \mathcal{Y}^T} \exp\big(S(\mathbf{x}, \tilde{\mathbf{y}})\big). \tag{3}$$

Since enumerating all possible tag sequences $|\mathcal{Y}|^T$ is computationally infeasible, $Z(\mathbf{x})$ is computed by the forward algorithm in $\mathcal{O}(T \cdot |\mathcal{Y}|^2)$ time [9].

The model is trained by minimizing the negative log-likelihood of the correct tag sequence:

$$\mathcal{L}_{\text{CRF}} = -\log P(\mathbf{y} \mid \mathbf{x}) = -S(\mathbf{x}, \mathbf{y}) + \log Z(\mathbf{x}). \tag{4}$$

Minimizing this loss increases the gold label path score while decreasing the aggregate score of competing paths $\log Z(\mathbf{x})$.

### D. Attention Masking and Padding

Batched training requires padding shorter sequences to a uniform length. We use the tokenizer's attention mask to distinguish real tokens (mask = 1) from padding (mask = 0) throughout training and evaluation.

For softmax configurations, padding positions receive a label of `-100` and are automatically excluded from the cross-entropy loss. For CRF configurations, the attention mask is used as the CRF boolean sequence mask. In document-level runs, masking behavior is unchanged: every unmasked position has a valid tag, and every masked position is ignored by both the loss function and decoder. Consistent with our sliding-window verification, CoNLL-2003 documents generally remain in-budget, so document-context batches are typically single-window.

## V. Experiments

### A. Evaluation Plan

Our evaluation followed a structured ablation framework to isolate the impact of each proposed modification. All experiments were evaluated on the CoNLL-2003 validation and test sets using span-level F1 score (via seqeval) as the primary metric, consistent with prior NER literature. Each configuration was trained with three random seeds. We report entity-level F1 as mean $\pm$ standard deviation to quantify variance across runs.

#### 1. Baseline: Sentence-Level ModernBERT

We first reproduced a standard ModernBERT fine-tuning setup on CoNLL-2003:

- Each sentence processed independently
- Standard token classification head
- Cross-entropy loss
- Default sequence truncation (no document context)

**Goal:** Establish a controlled baseline F1 score for comparison.

#### 2. Document-Level Context (No CRF)

We then evaluated the impact of incorporating document-level context while keeping the prediction head unchanged.

**Modifications:**

- Concatenate sentences within each article using `-DOCSTART-` boundaries
- Utilize the full 8,192 token limit when possible
- Apply sliding window chunking for articles exceeding the limit
- Use overlapping windows to preserve context continuity

**Goal:** Measure the isolated impact of extended contextual input (Baseline vs. Document-Level ModernBERT).

#### 3. CRF Head (Sentence-Level)

Next, we evaluated whether adding structured label decoding improves performance independently of document context.

**Modifications:**

- Revert to sentence-level processing
- Replace standard token classification head with a CRF layer
- Train using negative log-likelihood from the CRF

**Goal:** Isolate the structural labeling effect of CRF decoding (Baseline vs. ModernBERT + CRF).

#### 4. Combined Model: Document Context + CRF

Finally, we evaluated the combined model with both modifications enabled:

- Document-level concatenation
- Sliding window chunking
- CRF decoding head

**Goal:** Measure whether gains are additive or synergistic (Document-Level Only vs. CRF Only vs. Combined Model).

### B. Data Loading

All experiments used the original-format CoNLL-2003 files (`eng.train`, `eng.testa`, `eng.testb`) with `-DOCSTART-` boundaries preserved. We validated dataset integrity before training: parsed sentence counts matched expected split sizes (14,041 / 3,250 / 3,453), `-DOCSTART-` counts matched expected document totals (946 / 216 / 231), and span counts matched canonical CoNLL-2003 statistics for PER / ORG / LOC / MISC. We also verified that document-level concatenation is lossless relative to sentence-level parsing, with no sentence drops or duplicates across splits. CRF-specific verification additionally confirmed correct padding-mask behavior in forward and decode passes, valid BIO transition constraints, and dense BIO label roundtrip consistency.

### C. Training Configuration

Table II summarizes training hyperparameters for each experimental condition. For the ModernBERT ablation, each cell uses the configuration with the highest test micro F1 for that variant; sentence-level and document-level conditions were

tuned independently due to different memory constraints and CRF head requirements. The CRF rows include a separate transition-matrix learning rate (CRF LR). Reference BERT is included as a comparison model and is not part of the $2 \times 2$ ModernBERT ablation.

TABLE II: Training hyperparameters by experimental condition. Each row uses the configuration reported in the corresponding run manifest. All runs use AdamW with warmup ratio 0.1. Document-level runs use gradient accumulation for effective batching at 8,192-token sequences.

| Condition | LR | CRF LR | Epochs | Max Len | Batch |
|---|---|---|---|---|---|
| Sent. | 5e-5 | — | 8 | 512 | 16 |
| Sent.+CRF | 6e-5 | 3e-4 | 10 | 512 | 32 |
| Doc. | 4e-5 | — | 5 | 8192 | 2 |
| Doc.+CRF | 5e-5 | 2.5e-4 | 5 | 8192 | 4 |
| BERT (ref.) | 2e-5 | — | 5 | 512 | 16 |

Weight decay is 0.05 for Sent., 0.01 for Sent.+CRF / Doc. / Doc.+CRF / BERT (ref.).

## VI. Results

Table III presents test micro F1 on CoNLL-2003 (mean $\pm$ std over three seeds) for each cell of the $2 \times 2$ factorial. Table IV provides the corresponding per-entity breakdown.

TABLE III: ModernBERT $2 \times 2$ factorial ablation: test micro F1 (%) on CoNLL-2003, mean $\pm$ std over 3 seeds. BERT is a reference encoder with a different tokenizer and is not part of the ablation.

| Config | Test micro F1 |
|---|---|
| Sent. | 90.12 $\pm$ 0.31 |
| Sent.+CRF | 90.15 $\pm$ 0.21 |
| Doc. | **91.61 $\pm$ 0.23** |
| Doc.+CRF | 90.12 $\pm$ 0.13 |
| *Reference (different tokenizer):* | |
| BERT | 91.37 $\pm$ 0.17 |

TABLE IV: Per-entity F1 (%) on CoNLL-2003 test set, mean $\pm$ std over 3 seeds. Bold indicates best value in each column across all rows.

| Config | PER | ORG | LOC | MISC |
|---|---|---|---|---|
| Sent. | 95.71±0.05 | 87.12±0.66 | 92.22±0.34 | 79.93±0.19 |
| Sent.+CRF | 95.59±0.16 | 86.82±0.43 | 92.33±0.22 | 80.72±0.59 |
| Doc. | **97.88±0.01** | 89.46±0.50 | 92.83±0.23 | 79.81±0.50 |
| Doc.+CRF | 97.07±0.43 | 87.34±0.04 | 91.70±0.08 | 77.59±0.99 |
| *Reference (different tokenizer):* | | | | |
| BERT | 96.12±0.14 | **89.60±0.36** | **93.09±0.03** | **80.93±0.30** |

Across the $2 \times 2$ ModernBERT ablation, document context without CRF achieved the strongest overall performance (91.61 $\pm$ 0.23 micro F1). Adding a CRF head provided a modest gain in the sentence-level setting (90.15 vs. 90.12) but reduced performance in the document-context setting (90.12 vs. 91.61), indicating no additive benefit from combining the two modifications in this setup. At the entity level, document-context ModernBERT was strongest on PER, while the BERT reference remained strongest on ORG and LOC. Unlike PER, MISC favored the BERT reference; sentence-level CRF showed a modest improvement in ModernBERT's MISC F1, but Doc.+CRF underperformed all other settings. As reported in the run manifests, these results are the best configurations per model family and are not strictly hyperparameter-matched across all cells.

## VII. Scope and Limitations

This study is scoped to CoNLL-2003 English NER and this training/evaluation pipeline. We intentionally used per-condition hyperparameter tuning and selected the best runs within each ablation condition rather than enforcing a single shared hyperparameter setting across all cells. This reflects practical optimization differences between sentence-level and document-level regimes and between the softmax and CRF heads. The reported comparison should be interpreted as performance under realistic condition-specific tuning, not as a strictly hyperparameter-matched causal estimate of additive or interaction effects.

A key limitation is that this corpus does not stress ModernBERT's context capacity: sliding-window verification finds zero over-budget training documents relative to the 8,190-subword content budget; sampled documents fit within a single window. Therefore, document-context findings should be interpreted as in-budget same-document context effects rather than evidence about true multi-window long-document behavior near or beyond ModernBERT's token capacity. This likely explains why a legacy reference model remains highly competitive overall and best on multiple entity types, despite ModernBERT's stronger long-context design.

Although document-context ModernBERT achieved the strongest overall micro-F1, the BERT reference remained the best on three of the four entity classes (ORG, LOC, and MISC), with ModernBERT leading only on PER. Therefore, gains should be interpreted as label-dependent rather than uniformly superior across entity types.

## VIII. Conclusion

ModernBERT is designed for longer-context inputs, where its larger token capacity can capture richer cross-sentence information than sentence-isolated processing. In our CoNLL-2003 setup, this design translated into the strongest ModernBERT performance when document-level context was enabled without CRF (91.61 $\pm$ 0.23 micro F1).

Across the $2 \times 2$ ablation, adding a CRF produced only a modest sentence-level gain and reduced performance when combined with document context, indicating no additive benefit from the combined configuration. At the entity level, document-context ModernBERT was the strongest on PER, while BERT reference remained strongest on ORG and LOC, showing that improvements were not uniform across the entity types.

Overall, for this task and pipeline, expanding contextual input was more effective than adding CRF decoding to ModernBERT.

In this benchmark regime, the result is less a replacement of BERT and more an incremental improvement with different strengths by metric and entity type.

One plausible explanation for the per-entity pattern is that document-level context primarily benefits person disambiguation in news-related text, where cross-sentence references are frequent, while gains in ORG/LOC/MISC are less consistent. Additionally, the tokenizer and segmentation differences between the reference BERT setup and ModernBERT may affect boundary-sensitive classes (especially MISC) more strongly than PER. Given that CoNLL-2003 does not stress the long-context capacity in this pipeline, these per-entity differences should be considered speculative rather than definitive causal evidence.

## IX. Future Research

Future work should evaluate this ablation on a larger, more contemporary NER corpus with substantially longer documents. In CoNLL-2003, inputs remain within a single window, limiting conclusions about true long-context behavior. A corpus that consistently exceeds standard encoder limits would force sliding-window processing in shorter-context encoders and provide a clearer comparison of document-context and CRF effects under long-document conditions. This would better characterize when ModernBERT's architecture yields meaningful gains relative to strong legacy references.

A hyperparameter-matched study that fixes a shared training protocol across all $2 \times 2$ cells would be a useful confirmatory experiment. This would complement the presented condition-tuned comparison by showing a clear interaction effect for document-context and CRF decoding under controlled optimization parameters.

## X. Declaration on Generative AI

During the preparation of this work, the authors utilized Claude (Anthropic) to assist with grammar checking, spelling, formatting, textual refinement, and code-level suggestions. All AI-assisted content was subsequently reviewed and edited by the authors, who bear full responsibility for the final publication.

## References


[1] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 260–270, 2016.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018. Version Number: 2.

[3] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference, 2024. Version Number: 2.

[4] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition, 2003. Version Number: 1.

[5] Lance A. Ramshaw and Mitchell P. Marcus. Text chunking using transformation-based learning. In *Proceedings of the Third ACL Workshop on Very Large Corpora*, pages 82–94, Cambridge, MA, USA, 1995.

[6] Quentin Anthony, Jacob Hatef, Deepak Narayanan, Stella Biderman, Stas Bekman, Junqi Yin, Aamir Shafi, Hari Subramoni, and Dhabaleswar Panda. The Case fconllor Co-Designing Model Architectures with Hardware, 2024. Version Number: 2.

[7] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention, 2020. Version Number: 6.

[8] Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. arXiv, 2019. Version Number: 2.

[9] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, pages 282–289, 2001.