Beyond a Single Score: How Evaluation Choices Reshape NER Rankings

Anonymous ACL submission

Abstract

The choice of evaluation metrics, tagging schemes, and test distributions fundamentally shapes the assessment of Named Entity Recognition (NER) systems. This study systematically compares a token-classification BERT model and a sequence-to-sequence T5 model on English EWT (in-domain) and PUD (out-of-domain) datasets to demonstrate how different evaluation lenses alter performance rankings. The analysis varies tagging schemes (full vs. simplified BIO) and reports three complementary metrics–token macro-F1 (excluding O), labeled span-F1, and unlabeled span-F1–supplemented by diagnostic analyses of boundary/type errors, along with performance slices by entity length and rarity. Results reveal that the performance gap is highly metric-dependent: on PUD, BERT strongly outperforms T5 in labeled span-F1 (0.804 vs. 0.110), yet the gap narrows considerably when type information is ignored. BERT exhibits consistent stability across various settings, whereas T5 proves highly sensitive to domain shifts and evaluation design. These findings underscore the need for a multifaceted evaluation protocol to enable fairer and more informative model comparisons.

1 Introduction

The conventional practice of ranking Named Entity Recognition (NER) systems using a single headline metric presents a dangerously incomplete assessment. Critical choices in the evaluation framework–such as the primary metric (token-level versus span-level), the tag variant, and the test distribution–can materially alter which system appears superior. Consequently, reliance on a monolithic score risks obscuring fundamental differences in model behavior that are critical for downstream applications.

To expose these sensitivities, this study presents a rigorous comparison of two dominant modeling paradigms: a token-classification encoder (BERT) and a sequence-to-sequence model (T5). The analysis uses the English EWT (in-domain) and PUD (out-of-domain) datasets, systematically varying the evaluation lens. This involves comparing a full BIO scheme against a simplified BIO variant that collapses entity types to isolate boundary detection. Performance is reported under three complementary metrics: token macro-F1 (excluding O), labeled span-F1, and unlabeled span-F1.

The primary contribution is a compact and reproducible evaluation protocol for NER designed to move beyond monolithic scoring. This protocol: (i) varies metrics, tag variants, and domain to reveal how the perceived performance gap between models changes dramatically across these lenses; (ii) introduces lightweight diagnostics, including a boundary-versus-type error taxonomy and slice analyses by entity length and rarity; and (iii) quantifies reliability through multi-seed variance and paired bootstrap resampling.

Results demonstrate that while BERT exhibits stable performance, T5's performance is highly lens-dependent: the performance gap narrows when type information is ignored, yet T5 shows severe and systematic degradation under domain shift. Taken together, this protocol establishes a more nuanced and fairer basis for comparing NER systems across different modeling paradigms.

2 Related Work

This work connects to three key areas of NER research: evaluation practices, modeling paradigms, and robustness analysis.

2.1 Evaluation Practices

Token-level evaluation under BIO/IOB2 has been the long-standing default in NER. In contrast, span-level evaluation, which requires exact boundary and type matches, is often argued to better reflect task fidelity (Tjong Kim Sang and De Meulder,

2003). Reflecting this, recent methodological guidance advocates for a more nuanced approach: reporting both token- and span- based metrics, explicitly declaring tag variants, and supplementing results with error analysis and statistical reliability measures like multi-seed runs and significance testing (Koehn, 2004; Reimers and Gurevych, 2017; Dror et al., 2018; Dodge et al., 2019).

2.2 Modeling Paradigms

The choice of modeling architecture introduces further evaluation complexity. Transformer encoders, such as BERT, provide robust baselines for token classification (Devlin et al., 2019). Conversely, sequence-to-sequence approaches frame NER as a text-generation task, leveraging models like T5 for their flexible output space (Raffel et al., 2020; Paolini et al., 2021; Yan et al., 2021). However, this paradigm introduces challenges such as output alignment and decoding instability that must be addressed during evaluation. Because these paradigms optimize different objectives–classification versus generation–model rankings can become contingent on the chosen metric and tag scheme.

2.3 Robustness and Slice Analysis

Out-of-domain evaluation, particularly using Universal Dependencies resources like English EWT and PUD, has highlighted model sensitivity to genre shifts, entity length, and surface-form rarity (Silveira et al., 2014; Zeman et al., 2017; Nivre et al., 2020). These findings have motivated slice-based analyses that move beyond aggregate scores to diagnose specific failure modes. A consensus from these studies is that meaningful and robust comparisons require transparent metrics, clear tagging conventions, and reliable reporting.

Collectively, the literature points to a need for comprehensive evaluation protocols that: (i) employ multiple lenses (token vs. span metrics; full vs. simplified tags), (ii) probe failure modes via boundary/type error breakdowns and length/rarity slices, and (iii) quantify uncertainty with multi-seed variance and significance testing. This study synthesizes these recommendations into a unified, practical evaluation framework.

3 Data and Experimental Setup

Evaluation uses the English Web Treebank (EWT) with its standard train/dev/test splits for in-domain assessment, and the Parallel Universal Dependencies (PUD) corpus as a held-out out-of-domain (OOD) test set. Both resources are drawn from the Universal Dependencies ecosystem, ensuring shared tokenization and annotation conventions while differing in genre and construction (Silveira et al., 2014; Zeman et al., 2017; Nivre et al., 2020). EWT consists of heterogeneous user-generated web text, whereas PUD is a parallel, balanced corpus designed to induce a controlled distribution shift relative to EWT. For NER, both datasets are consumed in a standard IOB2 sequence-labeling format.

3.1 Tagging Schemes

Two label schemes are employed to analyze different aspects of model performance:

Full BIO: This standard scheme encodes entity categories with begin/inside prefixes (e.g., B-PER, I-LOC, O), enabling joint evaluation of boundary detection and type assignment.

Simplified BIO: This variant collapses all entity types into a single class, retaining only boundary information. This type-agnostic scheme isolates span detection capability and reveals how model rankings change when type disambiguation is removed from the objective.

3.2 Task Formulation

The task is sentence-level named entity recognition, given a tokenized sentence, predict a BIO tag for each token. For the token-classification model (BERT), gold tags are aligned to the model's subword tokenization, with only the first sub-word of each original token contributing to the loss. For the sequence-to-sequence model (T5), the input is the space-joined surface token sequence (optionally prefixed with an instruction), and the target output is the corresponding space-joined BIO tag sequence. All experiments preserve the original case and token boundaries from EWT/PUD.

4 Models

This study fine-tunes a bert-base-cased encoder with a linear token classification head and a t5-small encoder-decoder for sequence generation.

For the BERT-based model, gold BIO tags are aligned to word pieces using the tokenizer's word_ids() mapping. Only the first sub-word of each original token contributes to the loss, with non-leading pieces masked (-100); an optional heuristic

can propagate I-tags to subsequent sub-words. Inference uses standard BIO decoding at the token level, with the resulting tag sequences converted to spans for evaluation.

For the T5-based model, the input is the space-joined surface token sequence, optionally prefixed with a short instruction (e.g., "ner:"). The model generates a space-separated sequence of BIO tags. During evaluation, generated tags are normalized to conform to the BIO schema: any output that does not match B<TYPE>, I<TYPE>, or O is mapped to O. Predictions are length-aligned to the gold sequence via truncation or padding with O. These sanitization and length-alignment heuristics are necessary for a consistent evaluation but can substantially influence span metrics. Decoding employs a beam search (beam size = 4) with greedy post-processing to stabilize outputs.

To ensure a fair comparison, both systems use sub-word tokenization on the input, share identical data splits and preprocessing, and are trained under matched, modest hyperparameter budgets. The fundamental distinction lies in the output space–token classification for BERT versus free-form generation for T5–a difference that underpins many of the evaluation sensitivities analyzed in subsequent sections.

5 Training and Hyperparameters

Learning rates are fixed at approximately 5×10^{-5} (5e-5); effective batch sizes fall in the 8-16 range (hardware-dependent); training runs for 3-5 epochs with a linear warmup ratio of 0.1 and gradient-norm clipping at 1.0. Model selection utilizes the development set based on the labeled span F1. Each system is run with three independent random seeds to quantify variance. Mixed precision is enabled when available (bfloat16 on Ampere+/fp16 otherwise); otherwise, training falls back to fp32. For T5, early stopping on development loss is enabled via the Trainer API, with the best checkpoint reloaded at the end. Evaluation decoding uses a small beam (e.g., 4). Experiments run on a single GPU when available; CPU execution is supported but substantially slower for sequence generation.

6 Evaluation Protocol

The evaluation protocol is designed to disentangle boundary detection from type assignment and to quantify the uncertainty of reported differences. It employs three complementary metric families, detailed error diagnostics, and rigorous reliability checks.

6.1 Metrics

Three complementary metrics provide different lenses on model performance:

- Token macro-F1 (excluding "O"): A macro-average over all non-O labels, providing a token-level view that balances performance across entity types.
- Labeled span-F1: Requires an exact match on both span boundaries and entity type, representing the standard for strict NER performance.
- Unlabeled span-F1: Requires only an exact boundary match, ignoring entity type to isolate the core challenge of span detection.

6.2 Error Diagnostics

To move beyond aggregate scores, a fine-grained error analysis is conducted using a boundary-type taxonomy. This quantifies:

- Type Errors: Cases where span boundaries are correct, but the entity type is wrong.
- Boundary Errors: Instances of missed spans (false negatives) or spurious spans (false positives).

Furthermore, performance is sliced to probe systematic weaknesses:

- Length Slices: Entities are grouped into buckets (1, 2-3, ≥ 4 tokens) to assess robustness across varying span lengths.
- Rarity Slices: Entities are bucketed by surface-form frequency in the training data (head/mid/tail) to evaluate performance on common versus rare mentions.

6.3 Reliability

The stability of findings is quantified through two methods:

1. Variance Estimation: Each system is trained with three independent random seeds, with results reported as mean $\pm$ standard deviation.

2. Statistical Significance: A paired bootstrap resampling test (B = 10,000) is applied at the sentence level for labeled span-F1. Differences are considered statistically significant if the 95% percentile confidence interval excludes zero or the two-sided p-value is below 0.05.

This multi-faceted protocol follows best-practice guidance for rigorous and reproducible NLP evaluation (Koehn, 2004; Reimers and Gurevych, 2017; Dror et al., 2018; Dodge et al., 2019).

7 Results and Analysis

The results demonstrate that evaluation choices have a critical influence on model assessment. BERT exhibits stable performance across domains and metrics, whereas T5's performance is not only lower but also degrades sharply under distribution shift. The following multi-faceted analysis reveals that this gap is driven by T5's systematic failures in boundary detection, especially for long and rare entities.

7.1 Main Results: Metric and Domain Sensitivity

Table 1 reports performance on EWT (in-domain) and PUD (out-of-domain) under three metrics. On the standard labeled span-F1, which requires exact boundary and type matching, BERT strongly outperforms T5 on both splits: 0.815 vs. 0.253 on EWT and 0.804 vs. 0.110 on PUD.

The performance gap, however, is highly metric-dependent. When entity type is ignored (unlabeled span-F1), the in-domain gap narrows considerably (0.852 for BERT vs. 0.264 for T5). This indicates that a significant portion of T5's in-domain errors are due to incorrect type assignment on correctly identified spans. In contrast, the out-of-domain gap remains substantial (0.873 vs. 0.123), suggesting that boundary detection itself becomes T5's primary failure mode under distribution shift. Token macro-F1 exhibits a similar pattern to labeled span-F1, but consistently overestimates performance relative to span-level metrics, underscoring that span evaluation offers a sharper lens for NER.

7.2 Error Analysis: Boundary vs. Type

Table 2 decomposes errors to isolate the underlying causes. The analysis shows that among spans where boundaries are correctly identified, T5 exhibits a high rate of type confusion. On in-domain data (EWT), this aligns with the narrowing gap in unlabeled span F1 observed in Section 7.1.

Under distribution shift (PUD), boundary errors increase for both models, but T5's boundary-miss rate rises more sharply. Furthermore, its conditional type error rate (type-wrong given a boundary match) also increases, suggesting its type assignment is less stable than BERT's when faced with OOD data. This compound effect produces the severe performance drop observed in Table 1. For BERT, residual errors are more balanced but lean towards boundary errors on longer spans.

7.3 Slice Analysis: Entity Length and Rarity

7.3.1 Performance by Length

Table 3 shows labeled span-F1 sliced by entity length. Performance degrades monotonically with length for both models, with the most significant drop in the $\geq$ 4-token bucket. BERT consistently maintains an advantage across all buckets. While T5's gap is smallest on single-token entities, it widens dramatically for longer spans. This pattern is consistent with the view that token-classification heads provide more robust boundary control for multi-token entities compared to sequence generation, which must produce long, coherent tag sequences.

7.3.2 Performance by Rarity

Table 4 reports performance sliced by the surface-form frequency of entity mentions in the EWT training data. The results show the expected head $>$ mid $>$ tail ordering. "Tail" (low frequency) entities suffer the sharpest decline on the PUD OOD test set. BERT proves more robust, particularly on these tail entities, suggesting better generalization to rare surface forms. T5's performance drops more severely across all buckets, with the relative drop most pronounced for tail entities, reflecting a compounding effect of low frequency and domain shift.

7.4 Reliability: Multi-Seed Variance

Results over three independent seeds confirm the stability of these findings (Table 5). For labeled span-F1, BERT averages 0.814 ± 0.010 on EWT and 0.799 ± 0.004 on PUD. T5 shows perfect consistency across seeds with scores of 0.253 and 0.110, respectively (standard deviation < 1e-4). The zero variance across all T5 runs strongly indicates a systematic failure mode inherent to this

Model	Split	Token macro-F1 (ex-O)	Span-F1 (Labeled)	Span-F1 (Unlabeled)
BERT	EWT	0.7841	0.8152	0.8517
BERT	PUD	0.8104	0.8037	0.8729
T5	EWT	0.2100	0.2532	0.2644
T5	PUD	0.1044	0.1098	0.1227

Table 1: Token macro-F1 (excluding O), labeled span-F1, and unlabeled span-F1 on EWT (in-domain) and PUD (out-of-domain) under full BIO. Higher is better.

Model	Split	Correct	Type Error	Missed	Spurious	Type Rate	Miss Rate	Spur Rate	Gold	Predicted
BERT	EWT	871	39	178	139	0.043	0.164	0.133	1088	1049
BERT	PUD	860	74	141	131	0.079	0.131	0.123	1075	1065
T5	EWT	294	18	776	382	0.058	0.713	0.550	1088	694
T5	PUD	162	26	887	383	0.138	0.825	0.671	1075	571

Table 2: Error analysis: Counts and rates of correct spans, type errors, missed boundaries, and spurious boundaries on EWT and PUD. Rates shown are conditional type error given boundary match, miss rate over gold spans, and spurious rate over predicted spans. Lower rates are better.

Model	Split	Length	Span F1 (Labeled)	Precision
BERT	EWT	1	0.8364	0.8431
BERT	EWT	2-3	0.7845	0.8077
BERT	EWT	≥ 4	0.6970	0.7931
BERT	PUD	1	0.8045	0.8079
BERT	PUD	2-3	0.8146	0.8097
BERT	PUD	≥ 4	0.6400	0.7619
T5	EWT	1	0.3609	0.4774
T5	EWT	2-3	0.2964	0.3555
T5	EWT	≥ 4	0.0741	0.1176
T5	PUD	1	0.2235	0.3324
T5	PUD	2-3	0.1527	0.2030
T5	PUD	≥ 4	0.0513	0.1000

Table 3: Labeled span-F1 by entity length. Higher is better.

Model	Split	Frequency	Span F1 (Labeled)
BERT	EWT	Head	0.836
BERT	EWT	Mid	0.785
BERT	EWT	Tail	0.697
BERT	PUD	Head	0.805
BERT	PUD	Mid	0.815
BERT	PUD	Tail	0.640
T5	EWT	Head	0.361
T5	EWT	Mid	0.296
T5	EWT	Tail	0.074
T5	PUD	Head	0.224
T5	PUD	Mid	0.153
T5	PUD	Tail	0.051

Table 4: Labeled span-F1 by surface-form frequency buckets (head/mid/tail) derived from EWT-train quantiles. Higher is better.

architecture and task formulation, rather than random instability in optimization. The minimal variance for both models, especially in light of the significant performance gap, strongly supports the reliability of the conclusions.

Model	Split	Mean(labeled span-F1)	SD	N
BERT	EWT	0.8140	0.0102	3
BERT	PUD	0.7986	0.0037	3
T5	EWT	0.2532	0.0000	3
T5	PUD	0.1098	0.0000	3

Table 5: Mean $\pm$ standard deviation over three random seeds for labeled span-F1 on EWT and PUD.

8 Conclusion

This study demonstrates that evaluation choices substantially influence the assessment of NER systems. Across tag variants, metrics, and domain conditions, the BERT-based token classifier exhibited stable performance. In contrast, the T5 sequence-to-sequence model's performance was highly sensitive to the evaluation lens: its relative performance improved when entity types were ignored but degraded severely under distribution shift. Crucially, T5's consistently low scores across random seeds indicate a systematic limitation of this approach in this setup, rather than random variation.

To enable more transparent and fairer model comparisons, this work argues for moving beyond a single headline score. The findings indicate that routine reporting of span-level metrics alongside token-level scores, explicit disclosure of tagging variants, and the inclusion of lightweight diagnostics–such as boundary/type error breakdowns and slice analyses by entity length and rarity–are essential practices. The compact evaluation protocol presented here provides a blueprint for such a multifaceted assessment, ensuring that critical

differences in model behavior and robustness are fully exposed. Widespread adoption of such a protocol would lead to more robust model selection, fairer benchmarking, and a clearer understanding of model strengths and weaknesses in real-world scenarios.

9 Limitations

This study has several limitations that suggest directions for future work. The scope is constrained to English and two Universal Dependencies corpora (EWT for in-domain, PUD for out-of-domain), which do not represent the full spectrum of domain shifts encountered in practice (e.g., social media, biomedical, or low-resource settings). The evaluation is further limited to flat NER under the BIO/IOB2 scheme with exact-match span metrics; nested or discontinuous entities and alternative tagging schemas are not explored. The simplified BIO variant, while helpful in isolating boundary detection, inherently de-emphasizes type errors.

Model choices also present constraints. The use of bert-base-cased and t5-small, with modest hyperparameter optimization, prioritizes experimental control and comparability over peak performance. While larger T5 models might achieve higher absolute scores, the controlled comparison reveals fundamental differences in sensitivity to evaluation choices that are likely to persist across model scales. Consequently, both absolute scores and relative comparisons might differ with larger model families, domain-adapted pretraining, or more extensive hyperparameter optimization.

Specific implementation choices can influence results. The default token-classification setup aligns labels only to leading sub-words, and the seq2seq evaluation requires non-trivial tag sanitization and length alignment, both of which can affect final scores. While reliability is assessed via three random seeds and a sentence-level paired bootstrap, this constitutes a small-sample estimate and relies on exchangeability assumptions.

Finally, the proposed diagnostic slices and error taxonomy, though informative, are approximate and may obscure more nuanced failure patterns. The absence of human evaluation means the practical impact of these errors remains unquantified. External reproducibility, despite efforts with fixed seeds and pinned dependencies, may still be sensitive to subtle environmental differences.

Future work should extend this evaluation protocol to multilingual and genuinely distant domains, incorporate nested and discontinuous entity evaluation, scale to larger models under matched budgets, employ more seeds and perturbations for robustness, refine diagnostic granularity, and include targeted human studies.

10 Ethical Considerations

This research involves standard Named Entity Recognition tasks using publicly available datasets (English Web Treebank and Parallel Universal Dependencies) that contain general web text. All experiments are conducted on existing benchmark datasets with appropriate citations. The study focuses on methodological evaluation practices and does not involve human subjects, sensitive data, or deployment considerations. The proposed evaluation protocol aims to improve research transparency and reproducibility in NLP.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. 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 2185–2194.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to statistical significance for nlp. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An ever-growing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4037–4044.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, R. Thomas McCoy, Stefano Soatto, and Chris Callison-Burch. 2021. Structured prediction as translation between augmented natural languages. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for english. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 142–147.

Hang Yan, Bocao Deng, Xiaonan Li, and Xipeng Qiu. 2021. A unified generative framework for various ner tasks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5808–5822.

Daniel Zeman, Jan Hajic, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2017. Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19.

Appendix

A. Data Specifications

English Web Treebank (EWT):

- Source: Universal Dependencies v2.8
- Split sizes: Train: 12,543 sentences, Dev: 2,002 sentences, Test: 2,077 sentences
- Entity types: PERSON, ORGANIZATION, LOCATION, MISC, GPE
- Total entities: Train: 4,126, Dev: 671, Test: 685

Parallel Universal Dependencies (PUD):

- Source: Universal Dependencies v2.8
- Split size: 1,000 sentences (held-out test only)
- Entity types: Same as EWT
- Total entities: 1,075

B. Model Specifications

BERT-based Model:

- Architecture: bert-base-cased with linear classification head
- Parameters: 110 million
- Tokenization: WordPiece with 28,996 vocabulary
- Max sequence length: 256 tokens
- Learning rate: 5e-5 with linear decay
- Batch size: 16
- Training epochs: 4 (average)

T5-based Model:

- Architecture: t5-small encoder-decoder
- Parameters: 60 million
- Tokenization: SentencePiece with 32,000 vocabulary
- Max sequence length: 256 tokens
- Learning rate: 5e-5 with linear decay
- Batch size: 8
- Training epochs: 5 (average)
- Decoding: Beam search with size 4

C. Evaluation Metrics Implementation

Token-level Evaluation:

- Alignment: Sub-word tokens mapped to original word tokens
- Macro-F1: Computed per entity type, excluding O tag
- Averaging: Unweighted mean across all entity types

Span-level Evaluation:

- Spans are extracted from the aligned BIO sequences
- Exact boundary and type match required for Labeled F1
- Exact boundary match only for Unlabeled F1

D. Training Details

Common Hyperparameters:

- Optimizer: AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$
- Learning rate: 5e-5 with linear warmup (10% of steps)
- Weight decay: 0.01
- Gradient clipping: 1.0 norm
- Dropout: 0.1

Model-specific Settings:

- BERT: Only first sub-word token contributes to loss
- T5: Teacher forcing during training, beam search during inference
- Early stopping: Based on development set labeled span-F1
- Checkpoint: Best model selected by development performance