NERQual: Evaluating the Robustness of Named Entity Recognition Models to Data Quality Issues

Anonymous ACL submission

Abstract

Transformer-based models for Named Entity Recognition (NER) achieve strong performance on clean benchmarks, yet their reliability in real-world settings remains insufficiently understood. In practical applications, NER systems are frequently exposed to degraded input originating from optical character recognition, automatic speech recognition, and user-generated text. We address this gap by deriving a taxonomy of data quality perturbations from a systematic literature review, consolidating fragmented prior work into a unified framework. Guided by this taxonomy, we conduct controlled experiments evaluating six widely used transformer architectures under varying perturbation types, intensities, and training configurations. Our results show that syntactic perturbations cause the most severe performance degradation across models, while architectural features such as character-level processing and disentangled attention confer specific robustness advantages. Furthermore, we demonstrate that targeted training with perturbed data can recover more than half of the lost performance with minimal impact on clean-data accuracy.

1 Introduction

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing (NLP) and a cornerstone of downstream applications such as information extraction and semantic search (Yadav and Bethard, 2018; Rogers et al., 2020). While transformer-based models have set new performance standards on established benchmarks like CoNLL-2003 (Tjong Kim Sang and de Meulder, 2003; Devlin et al., 2019; Akbik et al., 2018), these controlled environments offer only a narrow view of model reliability. Consequently, high benchmark scores provide limited insight into how these architectures behave when deployed in volatile real-world settings (Ribeiro et al., 2020; Belinkov and Bisk, 2018).

In production environments, NER systems are frequently exposed to noisy text originating from Optical Character Recognition (OCR), Automatic Speech Recognition (ASR), or user-generated content, which introduces spelling errors, syntactic irregularities, semantic distortions, and annotation inconsistencies (Wang et al., 2021; Belinkov and Bisk, 2018; Hamdi et al., 2023). Such data quality issues can cause severe performance degradation, with reported drops of up to one third compared to clean benchmark data (Dirkson et al., 2022; Bhadauria et al., 2024).

While robustness has attracted increasing attention, existing studies typically address only limited aspects of the problem. Perturbation intensity is often operationalized and measured in different ways across studies, which renders results difficult to compare. Furthermore, previous analyses commonly focus on a restricted set of perturbation types, model architectures, or experimental objectives (Li et al., 2020; Bhadauria et al., 2024; Das and Paik, 2022; Simoncini and Spanakis, 2021; Lin et al., 2021; Moradi and Samwald, 2021). This fragmentation is further amplified by a lack of uniformity in evaluation settings, with studies focusing either on test-time robustness or robustness via training-time perturbations, rarely evaluating both within a single framework.

Altogether, these factors result in a fragmented body of robustness evidence, where findings are difficult to align across perturbation types, intensity definitions, model architectures, and evaluation settings. To address these limitations, we conduct a systematic experimental study that examines how different types and intensities of data perturbations affect transformer-based NER models under different perturbation settings, including perturbations applied during training, at inference time, and in combination. By evaluating multiple architectures within a unified experimental framework, our approach enables direct comparison across perturba-

tion types, intensity levels, and perturbation settings. To support further research in this area and facilitate model auditing, we provide our experimental framework as an open-source library, designed for easy integration and reproducibility[1].

2 Related Work and Taxonomy

The robustness of NLP models under perturbed input has been investigated through various lenses, including adversarial evaluation, data augmentation, and analysis of naturally occurring noise. Early research established that neural sequence labeling models are highly sensitive to small input perturbations, such as character-level noise or lexical substitutions, which can lead to substantial performance degradation (Belinkov and Bisk, 2018; Ebrahimi et al., 2018). Subsequent studies confirmed that these robustness issues persist for transformer-based architectures despite their improved contextual modeling capabilities (Dirkson et al., 2022; Náplava et al., 2021; Lin et al., 2021).

While a broad range of perturbation methods has been proposed, existing studies typically examine only limited subsets of perturbations (Simoncini and Spanakis, 2021; Li et al., 2020; Moradi and Samwald, 2021), which makes comparisons across experimental settings and model architectures difficult or incomplete. To obtain a more comprehensive picture of how data quality can be degraded in NER, we conducted a systematic literature review. From an initial pool of 400 publications, we identified 68 relevant studies on robustness and data quality in NLP and NER, including adversarial attacks, robustness evaluations, data augmentation, dataset analysis, and reported limitations. The systematic literature review was conducted following the guidelines proposed by Kitchenham and Charters (2007). Further methodological details are provided in Appendix A.

Our analysis revealed that several studies adopt a higher level of abstraction, grouping perturbations by their underlying linguistic characteristics rather than specific implementation details (Goyal et al., 2023; Bodapati et al., 2019; Simoncini and Spanakis, 2021; Qiao et al., 2024; Morris et al., 2020; Li et al., 2019). Following this paradigm, we categorize perturbations according to the linguistic level they affect: the orthographic surface form, the syntactic structure, or the semantic content. This hierarchical organization is summarized in Table. 1

[1] https://github.com/blinded_for_review

Perturbation type	Key References
Orthographic	
General spelling	(Ai et al., 2024; Al Sharou et al., 2021; Aliakbarzadeh et al., 2024; Bagla et al., 2021; Bhadauria et al., 2024; Chen et al., 2024; Esmaail et al., 2024; Li et al., 2020; Malykh, 2019; Malykh and Lyalin, 2018; Moradi and Samwald, 2021; Moradi et al., 2021; Namysl et al., 2020; Náplava et al., 2021; Nguyen and Chen, 2023; Qiao et al., 2024)
Keystroke errors	(Araujo et al., 2020; Bagla et al., 2023, 2021; Moradi et al., 2021)
Phonetic errors	(Nguyen and Chen, 2023; Qiao et al., 2024)
Morphological	(Náplava et al., 2021)
Elongation & char repetition	(Al Sharou et al., 2021; Moradi and Samwald, 2021; Moradi et al., 2021)
OCR errors	(Ai et al., 2024; Bhadauria et al., 2024; Boros et al., 2020; Das and Paik, 2022; Ehrmann et al., 2024; Hamdi et al., 2023; Namysl et al., 2020)
Capitalization	(Ahmed et al., 2024; Al Sharou et al., 2021; Bodapati et al., 2019; Das and Paik, 2022; Esmaail et al., 2024; Moradi and Samwald, 2021; Náplava et al., 2021; Zhu et al., 2025)
Character manipulation	(Araujo et al., 2020; Ebrahimi et al., 2018; Garg and Ramakrishnan, 2020; Goyal et al., 2023; Li et al., 2020; Nguyen and Chen, 2023)
Diacritic errors	(Al Sharou et al., 2021; Náplava et al., 2021)
Syntactic	
Grammatical	(Al Sharou et al., 2021; Chen et al., 2024; Malykh, 2019; Malykh and Lyalin, 2018; Moradi and Samwald, 2021; Qiao et al., 2024; Srinivasan and Vajjala, 2023)
Word order	(Liang et al., 2024; Moradi and Samwald, 2021; Moradi et al., 2021; Náplava et al., 2021)
Sentence struct.	(Bhadauria et al., 2024; Garg and Ramakrishnan, 2020; Li et al., 2020; Shi et al., 2024; Yuan et al., 2019; Zhu et al., 2025)
Tokenization & Segmentation	(Bernier-Colborne and Vajjala, 2024; Boros et al., 2020; Hamdi et al., 2023; Szymański et al., 2023)
Punctuation	(Al Sharou et al., 2021; Esmaail et al., 2024; Goyal et al., 2023; Náplava et al., 2021)
Semantic	
Entity replacement	(Dirkson et al., 2022; Lin et al., 2021; Lothritz et al., 2023; Morris et al., 2020; Qiao et al., 2024; Srinivasan and Vajjala, 2023; Wang and Li, 2024; Yan et al., 2022)
Context word replacement	(Araujo et al., 2020; Chen et al., 2024; Das and Paik, 2022; Dirkson et al., 2022; Garg and Ramakrishnan, 2020; Jia et al., 2019; Jin et al., 2021; Li et al., 2020; Lothritz et al., 2023; Moradi and Samwald, 2021; Morris et al., 2020; Qiang et al., 2024; Wang and Li, 2024; Wu et al., 2024; Ying et al., 2022; Zeng et al., 2021; Zhu et al., 2025)
Paraphrasing	(Goyal et al., 2023; Qiang et al., 2024; Srinivasan and Vajjala, 2023; Wang et al., 2021; Zeng et al., 2021)

Table 1: Perturbation taxonomy derived from the systematic literature review, grouping text-level perturbations into orthographic, syntactic, and semantic categories with representative perturbation types and corresponding studies. The taxonomy structures the perturbation design used in the experimental evaluation.

As illustrated in Table 1, **orthographic perturbations** are the most documented category, encompassing deviations in spelling and visual representation that do not necessarily alter grammar or meaning. These include general typos (Esmaail et al., 2024; Malykh and Lyalin, 2018; Náplava et al., 2021), keyboard errors (Araujo et al., 2020; Bagla et al., 2023; Moradi et al., 2021), and phonetic misspellings (Nguyen and Chen, 2023; Qiao et al., 2024). Technical noise often arises from OCR errors, such as misidentifying 'I' as 'l' (Hamdi et al., 2023; Namysl et al., 2020), or diacritic inconsistencies (Al Sharou et al., 2021; Náplava et al., 2021), while regional and historical variants further challenge models (Boros et al., 2020). Capitalization issues range from inconsistent casing to full "Case Ablation" (Bodapati et al., 2019; Das and Paik, 2022). Additionally, stylistic elongations (e.g., "soooo nice") and targeted character manipulations like insertion or swapping are frequently used for synthetic robustness testing (Ebrahimi et al., 2018; Goyal et al., 2023; Li et al., 2020; Moradi and Samwald, 2021).

Syntactic perturbations concern sentence structure and grammatical correctness, testing a model's ability to process structural dependencies. Disruptions include general grammatical errors (Chen et al., 2024; Malykh and Lyalin, 2018) and word order permutations (Liang et al., 2024; Náplava et al., 2021). Structural changes, such as shortening, expansion, or word merging, further challenge contextual inference (Bhadauria et al., 2024; Li et al., 2020; Shi et al., 2024). Particularly critical for NER are tokenization and segmentation errors (Ehrmann et al., 2024; Hamdi et al., 2023; Szymański et al., 2023), where incorrect splits or lost sentence boundaries hinder entity identification and disrupt the transformer's contextual window.

Semantic perturbations disrupt meaning while maintaining a plausible surface form. A common method is replacing context words with synonyms to determine if a model recognizes an entity when familiar lexical cues change (Garg and Ramakrishnan, 2020; Li et al., 2020; Morris et al., 2020; Qiang et al., 2024). More intensive methods involve entity replacement,swapping names with others from the same category (Dirkson et al., 2022) or using distributionally mismatched terms (Lin et al., 2021; Ying et al., 2022). At the sentence level, paraphrasing is used to alter surface structure while preserving underlying semantics (Srinivasan and Vajjala, 2023; Wang et al., 2021).

3 Experimental Setup

To evaluate model robustness, we follow the established methodology of constructing challenge datasets that target specific linguistic phenomena (Belinkov and Glass, 2019). By generating multiple challenge datasets through controlled variations, we facilitate a rigorous comparison between clean baseline performance and noise-induced variants. In contrast to real-world datasets that often contain multiple, interacting sources of noise, our approach allows us to isolate individual degradation types and systematically vary their intensity. Following the taxonomy introduced in Section 2, we implement a suite of orthographic, syntactic, and semantic perturbations.

To quantify how different transformer architectures adapt to these degradations, we evaluate performance across four training and testing configurations:

- **Clean-Clean:** A baseline setting where models are trained, validated, and evaluated on the original, unmodified data to establish reference performance under ideal conditions
- **Clean-Perturbed:** Models are trained and validated on clean data but evaluated on a perturbed test set, allowing measurement of immediate sensitivity to data degradation at inference time
- **Perturbed-Perturbed:** Matching perturbation types are applied during training, validation, and testing. This assesses to which extend training with degraded data can mitigate performance losses
- **Perturbed–Clean:** Models are trained and validated on perturbed data and evaluated on a clean test set to assess the model's ability to generalize to unperturbed inputs

By evaluating multiple architectures within this unified framework, our approach enables a direct comparison across different perturbation types and evaluation settings while ensuring that observed performance changes are attributable to specific degradation factors.

3.1 Data

All experiments are conducted on the CoNLL-2003 English dataset (Tjong Kim Sang and de Meulder, 2003), a widely adopted benchmark for NER.

The widespread use and clean manual annotations of this dataset make it well-suited for controlled robustness experiments, as it enables a direct comparison between clean and perturbed data while ensuring that findings are relatable to the broader body of NER research.

Our framework is implemented using the Hugging Face Datasets API (HuggingFace, 2025). We utilize the standard dataset splits to ensure reproducibility and comparability with prior work. Although this study focuses on CoNLL-2003, the framework is modular and can be readily applied to other NER datasets supported by the same API without changes to the perturbation or evaluation pipeline.

3.2 Models

To investigate the relationship between architectural design and robustness, we evaluate six transformer-based models that differ in architecture and input representation.We use pre-trained versions of these models and fine-tune them for NER on the CoNLL-2003 dataset. This setup reflects how models are commonly deployed in practical applications. The selection includes:

BERT (Devlin et al., 2019) serves as the baseline architecture. We use BERT-base, which consists of 12 transformer encoder layers with 12 self-attention heads and approximately 110M parameters. The model is pretrained using masked language modeling and next sentence prediction and relies on WordPiece subword tokenization.

RoBERTa (Liu et al., 2019) follows the same encoder architecture as BERT-base (12 layers, 12 heads) but is pretrained on substantially larger and more diverse text corpora. In addition, it removes the next sentence prediction objective and applies dynamic masking, resulting in a model with approximately 125M parameters. This design isolates the effects of large-scale pretraining from architectural changes.

XLNet (Yang et al., 2019) replaces masked language modeling with permutation-based language modeling to capture bidirectional context without masking. Comparable in size to BERT-base (12 layers, 12 heads, $\tilde{1}$10M parameters), it employs a two-stream self-attention mechanism and relative positional encodings.

DeBERTa (He et al., 2021) extends the BERT architecture by disentangling content and positional information within the attention mechanism. We use DeBERTa-base, which maintains a similar depth and attention configuration (12 layers, 12 heads) but increases model capacity to approximately 139M parameters.

DistilBERT (Sanh et al., 2019) is a compressed variant of BERT trained via knowledge distillation. It reduces the number of encoder layers from twelve to six, resulting in approximately 66M parameters while retaining the same hidden dimensionality and attention layout.

CANINE (Clark et al., 2022) differs fundamentally from the other models by operating directly on character-level input without explicit tokenization. It combines convolutional downsampling with Transformer layers to process long character sequences and contains approximately 104M parameters.

3.3 Perturbation Design and Implementation

To ensure a systematic and reproducible evaluation of data quality, we developed a custom perturbation framework. Our analysis of existing libraries like TextAttack (Morris et al., 2020) and NLPAug (Ma, 2019) showed that they are not specifically designed for NER and do not support the BIO labeling scheme, which leads to ground truth corruption when token boundaries change. Furthermore, SeqAttack (Simoncini and Spanakis, 2021) does not implement the full range of perturbations identified in our taxonomy and lacks support for character-based models like CANINE (Clark et al., 2022). To address this, we developed a custom framework that ensures BIO-integrity through automated label re-alignment and by maintaining architectural neutrality across subword and character-level models.

The framework is profile-based and organizes perturbations into orthographic, semantic, and syntactic classes. Each class can be configured to execute specific methods or the entire category. A dedicated control method manages candidate selection and equally distributes selected tokens among the active perturbation methods while enforcing specific selection constraints. A detailed overview of all implemented methods and examples is available in Appendix B.

A central focus of our implementation is scientific reproducibility. We initialize the random number generators of Python, NumPy, PyTorch, and CUDA with a shared global seed to guarantee reproducible experiments. To ensure statistical significance, every experiment is executed across five different seeds and three distinct learning rates (1e-5, 3e-5, 5e-5). Furthermore, our framework in-

corporates entity-protection, allowing us to isolate perturbations to either the context or the entities themselves. To simulate increasing degrees of real-world noise, the perturbations are applied at three distinct intensity levels of 10%, 20%, and 30%. This combination of variables results in a total of 2,700 training runs.

While the framework is a custom development, the individual transformation methods are inspired by and adapted from established literature. For example, within the semantic category, we implement three primary methods for lexical and contextual replacement. Following the approaches of (Chen et al., 2024; Srinivasan and Vajjala, 2023), we use WordNet-based synonyms and antonyms to introduce lexical variation. Inspired by Li et al. (2021) Li et al. (2020), we utilize static GloVe (Pennington et al., 2014) embeddings to identify semantically similar neighbors for replacement. Finally, drawing on the logic of Lin et al. (2021) and Ma (2019), we employ Masked Language Modeling (MLM) via an independent ALBERT (Lan et al., 2020) model to generate high-quality, contextually fitting perturbations.

4 Results

This section presents the experimental findings across all four evaluation settings. Performance is assessed using the entity-level micro-F1 score, necessitating an exact match for both the entity boundary and the predicted category.

4.1 Baseline

The `Clean-Clean` setting serves as performance ceiling, establishing a reference for model behavior under ideal conditions. As summarized in Table 2, RoBERTa and DeBERTa achieve the highest scores, followed by BERT and XLNet. DistilBERT and CANINE show slightly lower results. The low standard deviations across all models confirm the stability of the fine-tuning process.

Model	Micro-F1 (%)	Std. Dev. σ
BERT	91.8	$\pm$ 0.3
RoBERTa	92.9	$\pm$ 0.2
DeBERTa	92.7	$\pm$ 0.3
DistilBERT	90.3	$\pm$ 0.4
CANINE	89.4	$\pm$ 0.5
XLNet	91.8	$\pm$ 0.3

Table 2: Baseline Performance on Clean CoNLL-2003 Test Data (clean-clean setting).

4.2 Perturbed Test Data

This section evaluates model sensitivity to inference-time noise by testing architectures trained on clean data against perturbed test sets. As shown in Figure 1, performance correlates negatively with the proportion of modified tokens. Syntactic perturbations cause the most severe degradation, dropping the mean F1-score below 76% at 30% intensity. This category also exhibits the highest volatility with a standard deviation of $\sigma = \pm 1.5$ at 30%. In contrast, orthographic modifications are best compensated by the models, remaining above 80% at the same noise level with lower volatility ($\sigma = \pm 0.9$). Semantic perturbations occupy an intermediate position, with a slowing degradation rate that leads semantic and orthographic performance to converge at maximum intensity.

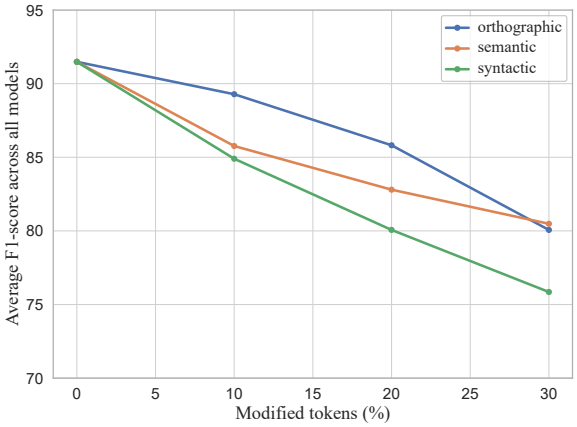

Figure 1: Mean Micro-F1 across models under increasing levels of orthographic, semantic, and syntactic test-set perturbations.

The progression of performance degradation is non-linear. Syntactic disturbances show the sharpest initial drop at only 10% noise intensity. Conversely, orthographic perturbations follow a flatter trajectory initially and only exhibit accelerated decline between 20% and 30%. This suggests the existence of a tipping point where the volume of erroneous tokens overwhelms the contextual compensation capabilities of the models.

Detailed results per model are reported in Table 3. RoBERTa and DeBERTa achieve the strongest performance across all perturbation types, maintaining F1 scores above 85% at the initial perturbation level of 10%. This stability is reflected in their lower variance ($\sigma \approx 0.8$) compared to the model average ($\sigma \approx 1.2$), indicating more resilient internal representations.

Type	Noise	BERT	RoBERTa	DeBERTa	DistilBERT	CANINE	XLNet
Baseline	0%	91.8 ± 0.3	**92.9** ± 0.2	92.7 ± 0.3	90.3 ± 0.4	89.4 ± 0.5	91.8 ± 0.3
Orthographic	10%	89.7 ± 0.6	**91.1** ± 0.5	90.7 ± 0.5	86.0 ± 0.8	88.6 ± 0.4	89.6 ± 0.7
	20%	85.9 ± 0.9	**88.0** ± 0.6	86.9 ± 0.7	81.3 ± 1.2	86.5 ± 0.4	86.3 ± 0.8
	30%	79.8 ± 1.2	**82.7** ± 0.9	82.2 ± 1.0	76.1 ± 1.6	**82.7** ± 0.6	80.0 ± 1.2
Syntactic	10%	86.1 ± 0.8	**88.0** ± 0.7	87.7 ± 0.7	83.2 ± 1.0	83.4 ± 0.9	86.2 ± 0.8
	20%	83.3 ± 1.0	**85.4** ± 0.8	84.8 ± 0.9	78.5 ± 1.2	80.9 ± 1.1	83.9 ± 1.0
	30%	81.3 ± 1.2	83.6 ± 1.0	**83.7** ± 1.1	74.9 ± 1.5	78.1 ± 1.3	81.3 ± 1.2
Semantic	10%	84.9 ± 1.0	86.7 ± 0.8	**86.9** ± 0.9	81.8 ± 1.3	82.7 ± 1.1	86.4 ± 1.0
	20%	79.9 ± 1.3	82.1 ± 1.0	**82.7** ± 1.1	76.6 ± 1.6	77.8 ± 1.3	81.3 ± 1.3
	30%	76.1 ± 1.5	77.7 ± 1.2	**78.8** ± 1.2	72.2 ± 2.2	72.6 ± 1.6	77.7 ± 1.5

Table 3: Model performance with a perturbed test dataset (entity-level Micro-F1, %) $\pm\sigma$ over five random seeds and three different learning rates on CoNLL-2003 (clean-perturbed setting).

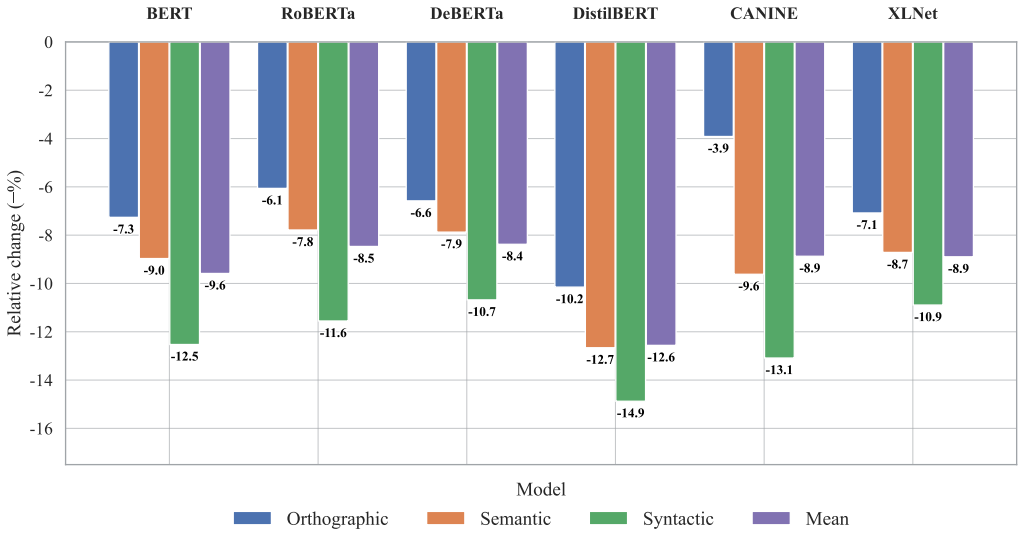

Figure 2: Relative performance change per model averaged across all intensity levels and compared to clean baselines. These averages may obscure significant failures at higher intensities; for example, stability at 10% orthographic perturbation partially offsets the sharp performance degradation seen at 30%.

Architectural advantages become evident under specific types of degradation. CANINE demonstrates exceptional robustness to orthographic perturbations: despite a lower clean-data baseline (89.4%), its character-level processing matches RoBERTa's absolute performance (82.6%) at maximum perturbation levels with high consistency ($\sigma = \pm 0.6$). This approach avoids subword segmentation errors triggered by typos, which severely impact models like DistilBERT, which falls to 76.1%. As illustrated in Figure 2, this translates to a minimal relative drop of only -3.3% for CANINE, identifying it as an orthographic specialist. However, this advantage is category-specific, the relative analysis shows that under semantic (-9.6%) and syntactic (-13.1%) perturbations, CANINE loses its architectural edge and exhibits a higher relative decline than the top-tier models.

Syntactic perturbations remain the most significant challenge across all models, with even the strongest architectures dropping below 79% F1 at maximum intensity. DeBERTa achieves the highest syntactic resilience (78.8%), likely benefiting from its disentangled attention mechanism, which preserves structural relations by treating relative positions and content separately. The vulnerability of compressed architectures is highlighted in Figure 2 by the significant instability of DistilBERT, which shows the greatest total average loss of -12.6%. This suggests that parameter distillation reduces the redundancy required to reconstruct contextual information from sequences where structural or semantic cues are fragmented.

Type	Noise	BERT	RoBERTa	DeBERTa	DistilBERT	CANINE	XLNet
Baseline	0%	91.8 ± 0.3	**92.9** ± 0.2	92.7 ± 0.3	90.3 ± 0.4	89.4 ± 0.5	91.8 ± 0.3
Orthographic	10%	90.1 ± 0.4	91.5 ± 0.3	**91.6** ± 0.4	87.8 ± 0.6	88.7 ± 0.5	90.7 ± 0.5
	20%	88.7 ± 0.6	**90.1** ± 0.5	90.0 ± 0.5	83.1 ± 0.8	87.9 ± 0.5	89.2 ± 0.6
	30%	86.4 ± 0.8	**89.0** ± 0.8	**89.0** ± 0.7	81.6 ± 1.2	86.9 ± 0.6	86.6 ± 0.6
Semantic	10%	89.6 ± 0.6	90.3 ± 0.5	**91.0** ± 0.5	87.7 ± 0.8	87.6 ± 0.7	89.9 ± 0.5
	20%	86.7 ± 0.8	88.3 ± 0.7	**89.1** ± 0.7	84.6 ± 1.0	84.5 ± 0.9	86.9 ± 0.7
	30%	85.3 ± 0.9	87.5 ± 0.8	**88.4** ± 0.7	81.9 ± 1.1	82.2 ± 1.0	85.2 ± 0.9
Syntactic	10%	88.4 ± 0.7	90.0 ± 0.6	**90.5** ± 0.6	86.6 ± 0.9	85.8 ± 0.9	89.0 ± 0.8
	20%	86.3 ± 0.9	88.6 ± 0.8	**89.2** ± 0.7	83.2 ± 1.2	82.5 ± 1.1	87.9 ± 1.0
	30%	84.1 ± 1.1	86.0 ± 0.9	**87.7** ± 0.9	81.1 ± 1.5	81.2 ± 1.2	86.4 ± 1.0

Table 4: Model performance with perturbed training, validation, and test dataset (entity-level Micro-F1, %) $\pm\sigma$ over five random seeds and three different learning rates on CoNLL-2003 (perturbed-perturbed setting).

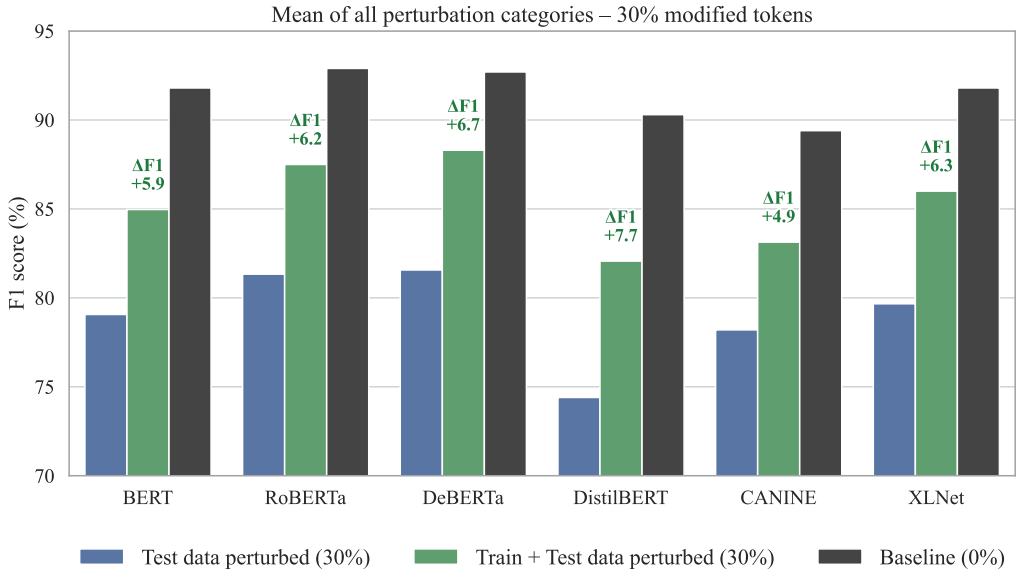

Figure 3: Mean Micro-F1 scores across all perturbation categories at 30% modified tokens, comparing clean data, perturbed test data, and perturbed training–test data for each model. ΔF1 denotes the absolute improvement in Micro-F1 (in percentage points) achieved by training with perturbed data compared to perturbing the test set only.

4.3 Perturbed Training & Test Data

This section examines whether incorporating perturbations during training mitigates the performance degradation identified in Section 4.2.

For all architectures a noticeable performance recovery is visible in Table 4. The degradation curves seem to follow a flatter trajectory, suggesting that training with degraded data eliminates the sharp tipping points seen in previous experiments. This improvement indicates that training on perturbed data stabilizes model behavior across various intensities. Average standard deviations decrease from $\sigma \approx 1.2$ to $\sigma \approx 0.8$.

To quantify this recovery, we examine the maximum load limit at 30% modified tokens in Figure 3. At this intensity, models recover on average 58.1% of the initial performance loss for orthographic, 52.1% for syntactic, and 41.9% for semantic perturbations.

Models that are already robust in one category seem to benefit more from training on degraded data. In the syntactic domain, DeBERTa and XLNet demonstrate strong adaptability with high absolute gains of +6.7 and +6.3 percentage points (pp) respectively. Their specific architectural features, such as disentangled attention or permutation-based pre-training, might enable them to integrate novel structural patterns more efficiently than models with absolute positional encodings.

DistilBERT records the highest relative gain of +7.7 pp, which could imply that models with lower baseline stability possess a higher relative adaptation potential when specifically exposed to the

noise distribution during training. Conversely, the gain for CANINE is lower at +4.9 pp. This might be explained by its high intrinsic character-level robustness, which potentially leaves less room for additional optimization.

An important observation is the consistent advantage of DeBERTa over RoBERTa across all categories after robust training. While both models perform similarly on clean data, DeBERTa achieves higher gains, such as +6.7 pp in the semantic domain compared to +6.2 pp for RoBERTa. This suggests that architectural features like the disentangled attention mechanism might allow for a more effective integration of perturbed patterns than architectures relying primarily on scaling.

Despite this adaptability, absolute F1-scores in the syntactic area do not reach the levels seen in orthographic experiments after robust training. This underlines that a disrupted sentence structure remains fundamentally more difficult to compensate for than local spelling errors.

4.4 Perturbed Training Data

This section examines the impact of degraded training data on model performance under ideal (clean) test conditions. Evaluating against the original error-free test set allows for an empirical assessment of the "Cost of Robustness" and simulates practical constraints like web-scraped training sources. Comprehensive results are documented in Appendix C.

All architectures demonstrate resilience to training-phase perturbations. Unlike the sharp declines observed during noisy inference, even a 30% perturbation rate during training results in minimal performance loss, with standard deviations remaining close to baseline levels. Orthographic changes have the least impact; RoBERTa remains nearly stable at 92.6% (-0.3 pp), while DeBERTa marginally exceeds its baseline with 92.8% (+0.1 pp) at the 20% level. This non-linear behavior suggests that moderate orthographic variance may function as data augmentation, potentially preventing overfitting and improving model robustness.

While semantic and syntactic perturbations lead to slightly higher declines, the overall losses remain marginal. Even at 30% syntactic modification, the most affected top-tier models like RoBERTa and DeBERTa only drop to 90.5% (-2.4 pp) and 91.0% (-1.7 pp) respectively.

These results underscore that while compromised grammatical structures during training make learning position-dependent patterns more difficult, the models retain a high degree of their baseline efficacy on clean data.

5 Conclusion

In this paper, we evaluated the robustness of six transformer architectures for NER using a newly developed, NER-specific perturbation framework. The framework is grounded in a structured taxonomy of perturbation types and enables controlled, reproducible evaluation of data quality degradation across different intensities and training settings.

Our results reveal pronounced architectural differences in robustness. While syntactic perturbations cause the strongest performance degradation overall, DeBERTa and XLNet demonstrate superior resilience in this setting, likely due to their more sophisticated handling of positional information. Across all perturbation types, RoBERTa and DeBERTa emerge as the most robust general-purpose models, maintaining higher stability as the proportion of modified tokens increases. CANINE exhibits strong robustness to orthographic perturbations due to character-level processing, while DistilBERT shows consistently larger performance drops, highlighting the vulnerability of reduced model capacity.

Furthermore, we demonstrate that noise-aware training on perturbed data substantially mitigates performance degradation at inference time. This strategy recovers much of the lost performance with only minor impact on clean evaluation. This suggests that robustness can be improved at relatively low cost through targeted training interventions.

Future research should extend this analysis to additional datasets, domains, and languages to assess the generalizability of the observed patterns, particularly in multilingual and domain-specific settings. Incorporating empirically observed real-world error patterns into perturbation design represents a promising direction for more realistic robustness evaluation. Overall, the presented framework and findings provide actionable insights for selecting and designing NER architectures that remain reliable under imperfect real-world conditions.

Limitations

This study is subject to several limitations that should be considered when interpreting the results.

First, while the perturbation taxonomy is

grounded in prior work, it may not capture the full spectrum of data quality issues encountered in practice. Although the taxonomy was derived from a broad literature base, it is possible that certain perturbation types or variants were not covered, which may limit the completeness of the robustness analysis.

Second, the experimental evaluation is restricted to a fixed set of six transformer-based architectures. While these models represent widely used and diverse design choices, the findings may not directly generalize to multilingual models, domain-specific pretrained models, or more recent architectural variants that were not included in this study.

Third, all experiments are conducted on the CoNLL-2003 dataset, which consists of relatively clean and homogeneous newswire text. Robustness patterns may differ in more domain-specific or informal settings, such as medical, legal, or social media text, where error distributions and linguistic characteristics can vary substantially.

Finally, the applied perturbations are synthetically generated and uniformly distributed, whereas real-world data often exhibits more complex and context-dependent error patterns. While the perturbation methods are linguistically motivated and grounded in prior work, we do not include a human evaluation of their realism or plausibility. Human validation and the incorporation of empirically observed error distributions could further improve the realism of robustness evaluations and represent important directions for future work.

Acknowledgments

Blinded for review

References

Anil Ahmed, Degen Huang, Syed Yasser Arafat, and Imran Hameed. 2024. Enriching urdu ner with bert embedding, data augmentation, and hybrid encoder-cnn architecture. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(4):1–38.

Chaoyi Ai, Yong Jiang, Shen Huang, Pengjun Xie, and Kewei Tu. 2024. Learning robust named entity recognizers from noisy data with retrieval augmentation. *Preprint*, arXiv:2407.18562.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Khetam Al Sharou, Zhenhao Li, and Lucia Specia. 2021. Towards a better understanding of noise in natural language processing. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 53–62.

Amirhossein Aliakbarzadeh, Lucie Flek, and Akbar Karimi. 2024. Exploring robustness of multilingual LLMs on real-world noisy data. In *Eighth Widening NLP Workshop (WiNLP 2024) Phase II*.

Vladimir Araujo, Andres Carvallo, Carlos Aspillaga, and Denis Parra. 2020. On adversarial examples for biomedical nlp tasks. *Preprint*, arXiv:2004.11157.

Kartikay Bagla, Shivam Gupta, Ankit Kumar, and Anuj Gupta. 2023. Noisy text data: foible of popular transformer based nlp models. In *The Third International Conference on Artificial Intelligence and Machine Learning Systems*, pages 1–5, New York, NY, USA. ACM.

Kartikay Bagla, Ankit Kumar, Shivam Gupta, and Anuj Gupta. 2021. Noisy text data: Achilles' heel of popular transformer based nlp models. *Preprint*, arXiv:2110.03353.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Gabriel Bernier-Colborne and Sowmya Vajjala. 2024. Annotation errors and ner: A study with ontonotes 5.0. *Preprint*, arXiv:2406.19172.

Divya Bhadauria, Alejandro Sierra-Múnera, and Ralf Krestel. 2024. The effects of data quality on named entity recognition. *Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024)*, pages 79–88.

Sravan Bodapati, Hyokun Yun, and Yaser Al-Onaizan. 2019. Robustness to capitalization errors in named entity recognition. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 237–242, Hong Kong, China. Association for Computational Linguistics.

Emanuela Boros, Ahmed Hamdi, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G. Moreno, Nicolas Sidere, and Antoine Doucet. 2020. Alleviating digitization errors in named entity recognition for historical documents. *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 431–441.

Shuguang Chen, Leonardo Neves, and Thamar Solorio. 2024. Context-aware adversarial attack on named entity recognition. In *Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024)*, pages 11–16, San Ġiljan, Malta. Association for Computational Linguistics.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine : Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Sudeshna Das and Jiaul Paik. 2022. Resilience of named entity recognition models under adversarial attack. *Proceedings of the First Workshop on Dynamic Adversarial Data Collection*, pages 1–6.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *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.

Anne Dirkson, Suzan Verberne, and Wessel Kraaij. 2022. Breaking bert: Understanding its vulnerabilities for named entity recognition through adversarial attack. *Preprint*, arXiv:2109.11308.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2024. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, 56(2):1–47.

Naji Esmaail, Nazlia Omar, Masnizah Mohd, Fariza Fauzi, and Zainab Mansur. 2024. Named entity recognition in user-generated text: A systematic literature review. *IEEE Access*, 12:136330–136353.

Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: Bert-based adversarial examples for text classification. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181.

Shreya Goyal, Sumanth Doddapaneni, Mitesh M. Khapra, and Balaraman Ravindran. 2023. A survey of adversarial defenses and robustness in nlp. *ACM Computing Surveys*, 55(14s):1–39.

Ahmed Hamdi, Elvys Linhares Pontes, Nicolas Sidere, Mickaël Coustaty, and Antoine Doucet. 2023. In-depth analysis of the impact of ocr errors on named entity recognition and linking. *Natural Language Engineering*, 29(2):425–448.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

HuggingFace. 2025. Datasets - library.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. *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 4127–4140.

Lifeng Jin, Linfeng Song, Kun Xu, and Dong Yu. 2021. Instance-adaptive training with noise-robust losses against noisy labels. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5647–5663.

Barbara Kitchenham and Stuart Charters. 2007. Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *Eighth International Conference on Learning Representations*.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2021. Contextualized perturbation for textual adversarial attack. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5053–5069.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2019. Textbugger: Generating adversarial text against real-world applications. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. Bert-attack: Adversarial attack against bert using bert. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202.

Xiaobo Liang, Runze Mao, Lijun Wu, Juntao Li, Min Zhang, and Qing Li. 2024. Enhancing low-resource nlp by consistency training with data and model perturbations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:189–199.

Bill Yuchen Lin, Wenyang Gao, Jun Yan, Ryan Moreno, and Xiang Ren. 2021. RockNER: A simple method

to create adversarial examples for evaluating the robustness of named entity recognition models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3728–3737, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Du Jingfei, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Cedric Lothritz, Saad Ezzini, Christoph Purschke, Tegawendé Bissyandé, Jacques Klein, Isabella Olariu, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2023. Comparing pre-training schemes for luxembourgish bert models. *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 17–27.

Edward Ma. 2019. Nlpaug - nlp augmentation framework.

Valentin Malykh. 2019. Robust to noise models in natural language processing tasks. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 10–16.

Valentin Malykh and Vladislav Lyalin. 2018. Named entity recognition in noisy domains. In *2018 International Conference on Artificial Intelligence Applications and Innovations (IC-AIAI)*, pages 60–65. IEEE.

Milad Moradi, Kathrin Blagec, and Matthias Samwald. 2021. Deep learning models are not robust against noise in clinical text. *Preprint*, arXiv:2108.12242.

Milad Moradi and Matthias Samwald. 2021. Evaluating the robustness of neural language models to input perturbations. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1558–1570.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.

Marcin Namysl, Sven Behnke, and Joachim Köhler. 2020. NAT: Noise-aware training for robust neural sequence labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1501–1517, Online. Association for Computational Linguistics.

Jakub Náplava, Martin Popel, Milan Straka, and Jana Straková. 2021. Understanding model robustness to user-generated noisy texts. *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 340–350.

Minh Nguyen and Nancy Chen. 2023. Firo: Finite-context indexing of restricted output space for nlp models facing noisy input. *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 834–845.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Yao Qiang, Subhrangshu Nandi, Ninareh Mehrabi, Greg Ver Steeg, Anoop Kumar, Anna Rumshisky, and Aram Galstyan. 2024. Prompt perturbation consistency learning for robust language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1357–1370, St. Julian's, Malta. Association for Computational Linguistics.

Haishan Qiao, Yuliang Zhu, Meiying San, Zeyang Chen, Sheqiang Zhao, Yan Chen, Zhangyan Li, and Shenghui Shi. 2024. A named entity recognition method for fields including electrical power based on data enhancement. In *2024 2nd International Conference on Artificial Intelligence and Power Engineering (AIPE)*, pages 11–16. IEEE.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Shenghui Shi, Zhangyan Li, Lin Yin, Zhaoying Chai, Kechen Fan, Daguang Jiang, Liangfei Zheng, and Dongqi Huang. 2024. Research and data analysis on relationship extraction methods based on multi-domain texts. In *2024 9th International Conference on Big Data Analytics (ICBDA)*, pages 197–202. IEEE.

Walter Simoncini and Gerasimos Spanakis. 2021. Seqattack: On adversarial attacks for named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 308–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Akshay Srinivasan and Sowmya Vajjala. 2023. A multilingual evaluation of NER robustness to adversarial inputs. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 40–53, Toronto, Canada. Association for Computational Linguistics.

Piotr Szymański, Lukasz Augustyniak, Mikolaj Morzy, Adrian Szymczak, Krzysztof Surdyk, and Piotr Żelasko. 2023. Why aren't we ner yet? artifacts of asr errors in named entity recognition in spontaneous speech transcripts. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1746–1761.

Erik Tjong Kim Sang and Fien de Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Hongtao Wang and Ang Li. 2024. Are deep neural networks robust to named entities? an adversarial attack and defense perspective. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12556–12560. IEEE.

Sihao Wu, Xiaoling Wu, Rui Xie, and Wentao Zhan. 2024. Chinese ner based on adversarial training and interactive attention. In *Proceedings of the 2024 16th International Conference on Machine Learning and Computing*, pages 557–561, New York, NY, USA. ACM.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158.

Jun Yan, Yang Xiao, Sagnik Mukherjee, Bill Yuchen Lin, Robin Jia, and Xiang Ren. 2022. On the robustness of reading comprehension models to entity renaming. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 508–520, Seattle, United States. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zheyu Ying, Jinglei Zhang, Rui Xie, Guochang Wen, Feng Xiao, Xueyang Liu, and Shikun Zhang. 2022. 3rs: Data augmentation techniques using document contexts for low-resource chinese named entity recognition. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. 2019. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30(9):2805–2824.

Guoyang Zeng, Fanchao Qi, Qianrui Zhou, Tingji Zhang, Zixian Ma, Bairu Hou, Yuan Zang, Zhiyuan Liu, and Maosong Sun. 2021. OpenAttack: An open-source textual adversarial attack toolkit. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 363–371, Online. Association for Computational Linguistics.

Jizhao Zhu, Akang Shi, Zixuan Li, Long Bai, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. 2025. Towards robust universal information extraction: Dataset, evaluation, and solution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28052–28070, Vienna, Austria. Association for Computational Linguistics.

A Systematic Literature Review

The systematic literature review was conducted following the guidelines of Kitchenham and Charters (2007). The goal was to identify and structure prior work on robustness and data quality issues in Natural Language Processing (NLP), with a focus on Named Entity Recognition (NER) and transformer-based models.

To ensure a consistent and transparent selection process, predefined inclusion and exclusion criteria were applied throughout the review. These criteria are summarized in Table 5.

Category	Include	Exclude
Topic	Robustness or data quality issues in NLP or NER	No clear NLP focus or no substantive robustness analysis
Model Type	Transformer-based encoder architectures	Non-transformer models
Quality Issue	Explicit and reproducible perturbations	Unspecified or opaque input changes
Publication	Peer-reviewed conference or journal papers; preprints	Books or non-scientific literature
Time Frame	2017–2025	Before 2017
Language	English or German	Other languages
Availability	Full text accessible	No full-text access
Citations	More than 10 citations	10 or fewer citations

Table 5: Inclusion and exclusion criteria used in the systematic literature review.

To refine the search strategy, a pilot search evaluated candidate keywords related to robustness and noise. High-yield terms (e.g., *noise*, *robust*, *data quality*, *adversarial*) were retained, while low-yield terms were discarded.

Searches were conducted in the ACL Anthology, IEEE Xplore, ACM Digital Library, and arXiv as these showed as the most promising during pilot search. For each database, the first 100 results ranked by relevance were screened. Studies were filtered through deduplication, title and abstract screening, and full-text assessment.

The final search query was:

```
("natural language" OR "NLP" OR
"named entity recognition" OR
"NER")
AND
("data quality" OR robust
OR nois* OR perturbation OR
adversarial)
```

The screening process is summarized in Figure 4. Out of the 400 initially identified studies, 68 met all inclusion criteria and were included in the final review. These studies informed the perturbation taxonomy and experimental design used in this work.

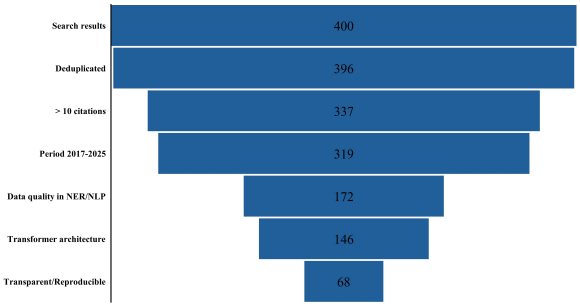

Figure 4: Study selection and screening process for the systematic literature review. Numbers indicate the remaining studies after each filtering step.

B Implemented Methods & Examples

This appendix provides qualitative examples of the perturbation methods implemented in this work, covering orthographic, syntactic, and semantic categories. The examples illustrate how controlled data quality perturbations are applied to the CoNLL-2003 corpus and correspond to the perturbation taxonomy introduced in Section 2. Orthographic, syntactic, and semantic perturbations are shown in Tables 6, 7, and 8, respectively. All perturbation implementations are publicly available in our GitHub repository`https://github.com/blinded_for_review` .

C Results - Perturbed Training Data

Detailed results for experiments with perturbed training and validation data evaluated on the clean test set are reported in Table 9. Across models, performance changes remain small even at higher perturbation levels.

Method	Original Text	Perturbed Text
Insertion	The company said it expects higher profits next year.	The company said it expects higbher profits next year.
Deletion	Peter Blackburn reported from London.	Peter Blackburn reported from Londo~~n~~.
Substitution	France and Germany agreed on the proposal.	France and Germany agr~~r~~ed on the proposal.
Case change	IBM said the market was stable after recent reports.	Ibm said the market was stable after recent reports.
Character swap	. . . would continue to support economic reforms in Europe.	. . . would continue to uspprot economic reforms in Europe.
Diacritics	José Mourinho joined the club after talks with Chelsea officials.	Jose Mourinho joined the club after talks with Chelsea officials.
Homoglyphs	Google announced a new partnership with Apple on Monday.	Google announced a new partnership with App1e on Monday.

Table 6: Overview of implemented orthographic perturbation methods with examples from the CoNLL-2003 corpus before and after applying orthographic transformations that simulate typical spelling and typing errors. Modified tokens are highlighted in red.

Method	Original Text	Perturbed Text
Insert punctuation	The United Nations called for new peace talks in Geneva.	The United , Nations called for new peace talks in Geneva.
Delete punctuation	. . . from Bank of America, New York, and the United States Treasury met in London.	. . . from Bank of America ~~,~~ New York, and the United States Treasury met in London.
Remove spaces	France and Germany signed the agreement yesterday.	Franceand Germany signed the agreement yesterday.
Split words	Microsoft announced new software on Monday.	Micro␣soft announced new software on Monday.
Delete words	Tony Blair met President Bush in London.	Tony Blair met ~~President~~ Bush in London.
Repetition	IBM reported strong results in the first quarter.	IBM IBM reported strong results in the first quarter.
Swap	Google announced a new partnership with Apple.	Google announced a new partnership Apple with.

Table 7: Overview of implemented syntactic perturbation methods with examples from the CoNLL-2003 corpus before and after applying syntactic transformation operations that simulate structural deviations in sentence organization and token segmentation. Modified tokens are highlighted in red.

Method	Original Text	Perturbed Text
WordNet-based synonym substitution	The foreign ministry's Shen told Reuters Television . . .	The foreign ministry's Shen told Haaretz Video . . .
WordNet-based antonym substitution	German first-time registrations of motor vehicles jumped . . .	German unwary registrations of motor vehicles jumped . . .
Embedding-based substitution (GloVe)	The meeting was attended by Tony Blair in London.	The meeting was operated by Tony Blair in London.
Contextual substitution (MLM via ALBERT)	Talks between France and Germany continued in Brussels.	Talks between France and Germany collapsed in Brussels.

Table 8: Overview of implemented semantic perturbation methods with examples from the CoNLL-2003 corpus before and after applying WordNet-, GloVe-, and MLM-based substitution techniques (here using ALBERT) to generate semantic variations. Modified tokens are highlighted in red.

Type	Noise	BERT	RoBERTa	DeBERTa	DistilBERT	CANINE	XLNet
Baseline	0%	91.8 ± 0.3	**92.9** ± 0.2	92.7 ± 0.3	90.3 ± 0.4	89.4 ± 0.5	91.8 ± 0.3
Orthographic	10%	91.5 ± 0.4	**92.5** ± 0.3	92.6 ± 0.3	89.9 ± 0.4	89.3 ± 0.4	91.6 ± 0.4
	20%	91.3 ± 0.3	92.7 ± 0.3	**92.8** ± 0.4	89.5 ± 0.3	89.0 ± 0.5	91.6 ± 0.3
	30%	90.9 ± 0.4	92.6 ± 0.4	**92.5** ± 0.4	88.8 ± 0.4	88.8 ± 0.5	91.3 ± 0.4
Semantic	10%	90.7 ± 0.4	91.7 ± 0.3	**92.2** ± 0.4	89.5 ± 0.5	88.9 ± 0.5	91.2 ± 0.4
	20%	90.8 ± 0.4	91.8 ± 0.4	**92.3** ± 0.4	88.9 ± 0.4	89.0 ± 0.6	91.2 ± 0.4
	30%	90.1 ± 0.3	91.6 ± 0.4	**92.0** ± 0.5	88.1 ± 0.5	88.7 ± 0.6	90.8 ± 0.5
Syntactic	10%	90.1 ± 0.3	91.7 ± 0.4	**91.8** ± 0.3	88.8 ± 0.5	88.5 ± 0.4	90.1 ± 0.4
	20%	89.7 ± 0.3	91.2 ± 0.3	**91.3** ± 0.4	87.9 ± 0.5	87.8 ± 0.5	89.4 ± 0.5
	30%	89.1 ± 0.5	90.5 ± 0.4	**91.0** ± 0.4	87.7 ± 0.6	87.1 ± 0.5	88.7 ± 0.5

Table 9: Model performance with a perturbed training & validation dataset (entity-level Micro-F1, %) $\pm\sigma$ over five random seeds and three different learning rates (1e-5, 3e-5, 5e-5) on CoNLL-2003 (perturbed-clean setting).