Master Thesis

Evaluating Generalisation in Named Entity Recognition through Robustness Testing

Nikhil Mathews

Supervisor Antske Fokkens, Lucia Donatelli
2^{nd} reader Piek Vossen

a thesis submitted in fulfilment of the requirements for the degree of

MA Linguistics
(Text Mining)

Vrije Universiteit Amsterdam

Computational Linguistics and Text-Mining Lab
Department of Language and Communication
Faculty of Humanities

Date August 3rd, 2025
Student number 2844742
Word count 10,492

Abstract

Named Entity Recognition (NER) is a foundational task in natural language processing, yet its generalisation capacity across domains and linguistic variations remains underexplored. This thesis investigates the ability of BERT-based NER models to move beyond surface-level heuristics and generalise under distributional shift, ambiguous input, and adversarial perturbations. Four model configurations: BERT-base-cased and uncased, trained on CoNLL 2003 and OntoNotes 5.0, are evaluated on both in-domain and cross-domain settings.

Robustness is tested using seed sensitivity metrics, curated challenge sets, and adversarial examples. The results show that CoNLL-trained models perform better under controlled conditions but are more fragile when faced with syntactic ambiguity or unfamiliar entity distributions. OntoNotes-trained models generalise better across genres but struggle with precise boundary resolution. Notable error sources include overreliance on casing, misclassification of common noun names (e.g. 'Mark', 'Will'), and confusion with PER-ORG in discourse contexts.

These findings underscore the importance of data diversity, structured evaluation, and surface cue robustness to improve NER performance. This study contributes practical diagnostics for evaluating generalisation and offers challenge set templates to support future research in robust entity recognition.

Declaration of Authorship

I, Nikhil Mathews, declare that this thesis, titled *Evaluating Generalisation in Named Entity Recognition through Robustness Testing* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a Master's degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: 3 August 2025

Signed: Nikhil Mathews

Acknowledgments

I would like to express my gratitude to my supervisors, Antske Fokkens and Lucia Donatelli, for their guidance, encouragement, and feedback throughout the development of this thesis. Their knowledge and support were critical in shaping this work and in helping me navigate the many challenges that came with it.

I am also grateful to my friends and colleagues, Melina Paxinou, Elisabetta Dentico, and Cheryl Chen, for their ongoing support, thoughtful discussions, and for always being available to offer help when it was most needed. Their presence made this process not only more manageable but also fun.

To my partner, Sal, thank you for being a constant source of strength, inspiration, and clarity. Your help in brainstorming ideas, revising drafts, and keeping me grounded emotionally was invaluable. This thesis would not have been executed successfully without you.

I would also like to give an honourable mention to coffee, without which this work would not exist. Its unrelenting support in the form of caffeine-fueled nights and early mornings has powered almost every page of this document.

Lastly, I am grateful to all, named and unnamed, who contributed in small or large ways to this journey. Whether through a conversation, a shared paper, or simply being there, thank you.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Problem Definition

Named Entity Recognition (NER) is a foundational task in natural language processing (NLP) that involves identifying and classifying text spans into predefined categories such as persons, organisations, and locations. Although recent advances, particularly transformer-based architectures like BERT(Devlin et al., 2019), have led to substantial improvements on standard NER benchmarks, there is growing concern that these models do not truly generalise. Most evaluations remain restricted to in-domain test sets, raising the question of how well such models perform when faced with data that differ in style, domain, or linguistic composition from their training distribution.

Several studies have shown that high F1 scores on held-out subsets of the training distribution can mask a reliance on superficial patterns such as capitalisation, word shape, or memorised entity forms (Ribeiro et al., 2020). This is especially problematic in real-world settings where the input may be noisy, adversarially constructed, or drawn from unseen domains. Furthermore, biases in training data, such as over-representation of Western names or genre-specific conventions, can lead to models that systematically underperform on underrepresented groups or contexts (Zhao et al., 2018; Bender and Friedman, 2018). Despite these limitations, there has been limited work in systematically diagnosing generalisation failures in NER using both quantitative evaluation and controlled input variation.

Research Questions

This thesis addresses the overarching question:

> *Can a model trained in dataset X generalise effectively to dataset Y*

More specifically, it investigates the following sub-questions:

- How do differences in training data (e.g. CoNLL-2003 vs. OntoNotes 5.0) affect generalisation?
- How does the model performance vary across datasets, domains, and specific linguistic phenomena?

Approach

To answer these questions, I perform a structured empirical evaluation of four BERT-based NER models: `bert-base-cased` and `bert-base-uncased`, each trained on CoNLL-2003(Tjong Kim Sang and De Meulder, 2003) and OntoNotes 5.0(Hovy et al., 2006). I evaluate both similar distribution and different distribution performance using standard metrics, and also examine prediction consistency across multiple training seeds using agreement measures such as Krippendorff's alpha.

To probe deeper into model behaviour, I construct custom challenge sets designed to expose specific weaknesses. These include a set of adversarial perturbations (Stanford Challenge Set), a synthetically generated ethnic name dataset to evaluate performance on underrepresented name origins, and a common noun ambiguity set targeting names like "Mark" or "Will" that overlap with everyday vocabulary. Through these experiments, the aim is to diagnose not only when models fail, but also why.

Key Findings. The results reveal that models trained on OntoNotes generalise better to unseen domains, particularly due to its genre diversity and balanced tag distributions, whereas CoNLL-trained models rely more heavily on surface features such as capitalisation. Performance across seeds is broadly stable, thus I do not rerun multiple tests to confirm reproducability, but errors concentrate around ambiguous cases such as names that double as common nouns or alternate between person and organisation roles. Challenge set evaluations confirm that lexical ambiguity and contextual underspecification remain key limitations, especially in uncased models.

Thesis Structure

The remainder of this thesis is organised as follows:

- **Chapter 2** provides a detailed review of previous work on generalisation in NLP, contextual models such as BERT, evaluation frameworks, and relevant findings in NER.
- **Chapter 3** describes the CoNLL-2003 and OntoNotes 5.0 datasets used for model training and evaluation, including their design and structure.
- **Chapter 4** outlines the experimental methodology, including model configurations, training setup, evaluation protocols, and the design of the challenge sets.
- **Chapter 5** presents the empirical results, comparing performance between datasets, seeds, and challenge sets, followed by targeted analyses of bias and robustness.
- **Chapter 6** discusses the broader implications of the findings, including the limitations of current NER systems and suggestions to improve generalisation.
- **Chapter 7** concludes the thesis with a summary of key contributions and directions for future research.

Chapter 2

Background and Literature Review

2.1 Overview

This chapter provides a structured overview of generalisation in Natural Language Processing (NLP), beginning with early efforts to evaluate model performance beyond training data, such as Penn Treebank (Marcus et al., 1993), and progressing to more recent efforts such as the GenBench framework (Hupkes et al., 2023a). It outlines key challenges in assessing generalisation, including the variability introduced by different random seeds and the effects of distribution shifts between training and test data.

To ground the discussion, the chapter introduces contextual language models, especially BERT, which have significantly improved model performance across NLP tasks, including NER. Then it turns to methods for improving generalisation, including data augmentation, adversarial examples, and error analysis tools. The chapter concludes by returning to the specifics of Named Entity Recognition, where generalisation poses persistent challenges, and outlines how it will be addressed in this thesis using CoNLL-2003 and OntoNotes 5.0 datasets.

2.2 Defining Generalisation in NLP

Generalisation in NLP refers to the ability of a model trained on a specific type of data for a given task to maintain reasonable performance when the type or distribution of the data changes (Hupkes et al., 2023b). Although this concept has implicitly shaped NLP research for decades, one of the first large-scale efforts to formalise evaluation practices was the Penn Treebank project (Marcus et al., 1993), which introduced systematic training-test splits and helped establish held-out evaluation as a standard. Hupkes et al. (2023b) describes this as one of the first instances of what can be considered a form of generalisation evaluation, although generalisation was not an explicit focus of the work. Over time, researchers began to probe more targeted aspects of generalisation. For example, Lake and Baroni (2018) demonstrated that sequence-to-sequence models struggle with compositional generalisation, and McCoy et al. (2019) revealed how high in-domain accuracy could mask a reliance on shallow syntactic heuristics. These insights motivated a larger push toward interpretable and controlled evaluation setups.

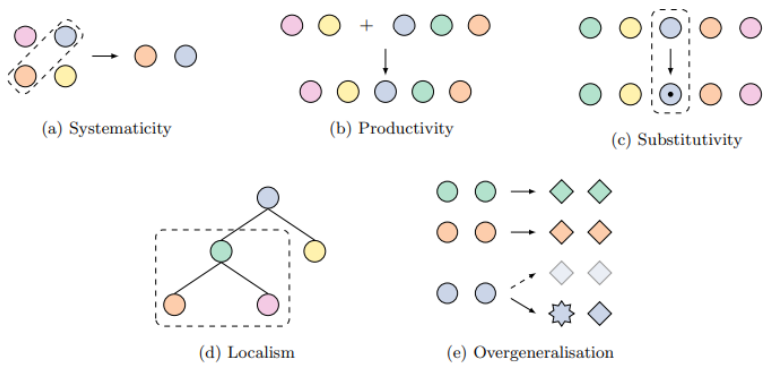

Figure 2.1: Reproduction of schematic depictions of the five tests proposed by Hupkes et al. (2020, Figure 1, p. 764) to evaluate the compositionality of neural network models. (a) Systematicity: recombining known parts into novel sequences. (b) Productivity: generalising to sequences longer than seen during training. (c) Substitutivity: assessing when words are treated as synonymous. (d) Localism: evaluating whether smaller constituents are composed before larger ones. (e) Overgeneralisation: testing a model's tendency to infer and apply rules vs. learning exceptions.

To help clarify and systematise this growing body of work, Hupkes et al. (2020) proposed a structured framework that decomposes generalisation into five dimensions: *diversity*, *specificity*, *compositionality*, *granularity*, and *difficulty*. These dimensions allow researchers to classify generalisation tasks more precisely and to design experiments that target specific types of generalisation failures. The framework is visually summarised in the original paper (see Figure 2.1), but in this thesis, I focus on three of the five dimensions: systematicity, substitution, and overgeneralisation.

Most recently, Hupkes et al. (2023b) surveyed more than 700 NLP experiments, identifying blind spots and inconsistencies in how generalisation is tested, and introduced **GenBench**, a benchmark suite and protocol designed to make generalisation evaluations more rigorous, transparent, and reproducible.

2.3 Contextual Models and BERT

One major advancement in addressing the limitations of traditional NER systems has been the development of contextual language models. Among these, **BERT** (Bidirectional Encoder Representations from Transformers) has become especially prominent. The introduction of the Transformer architecture by Vaswani et al. (2017) marked a breakthrough in NLP. Transformers use self-attention to model relationships between all words in a sentence, allowing for parallel processing and capturing long-range dependencies. Building on this architecture, **BERT** was introduced by Devlin et al. (2019) and quickly became a standard base model for a wide range of NLP tasks.

BERT is pretrained on large corpora such as Wikipedia and BooksCorpus using two self-supervised tasks: **Masked Language Modelling (MLM)** and **Next Sentence Prediction (NSP)**. In MLM, 15% of the tokens are randomly selected for prediction. Of these, 80% are replaced with a `[MASK]` token, 10% with a random word, and 10% remain unchanged. The model is trained to predict the original word based on both the

left and right context, encouraging deep semantic understanding. In NSP, the model is presented with two sentences and must predict whether the second sentence follows the first in the original text. This task trains the model to capture inter-sentence coherence.

These pretraining tasks equip BERT with a robust general-purpose language representation by encouraging the model to learn context-sensitive representations. This means that instead of assigning fixed meanings to words, BERT models each token to its surrounding context, enabling more accurate interpretations of ambiguous or polysemous terms. Such rich contextualisation is particularly valuable for downstream tasks like NER, where the meaning and role of a token often depend on its position in the sentence and its relationship to nearby entities.

In this thesis, I use `bert-base-uncased` and `bert-base-cased`, which differ in whether they retain case information, a potentially useful signal for recognising proper nouns in NER tasks. Although BERT's pretraining on large and diverse corpora like Wikipedia and BooksCorpus enables broad linguistic competence, effective generalisation to specific tasks still relies on the quality and diversity of the fine-tuning data. A model fine-tuned on a narrow or biased dataset may overfit to superficial patterns, such as specific name formats or domains, limiting its robustness to unseen examples.

Thus, although pretrained models like BERT provide a strong foundation, generalisation in NER remains a challenge, particularly when encountering rare entities, domain shifts, or variations in surface form. This thesis explores how such issues manifest in practice and examines whether targeted interventions in the fine-tuning data can improve the robustness of these models.

2.4 Generalisation through Data-Centric Interventions

A key challenge for generalisation lies in the sensitivity of models to seemingly minor changes in the input. Ribeiro et al. (2020) illustrates this problem: models trained on standard datasets were shown to fail on simple, human-intuitive variations. Their findings emphasise that high test accuracy does not guarantee true understanding or robust performance in the face of subtle perturbations.

One practical way to address this issue is by modifying the training data. Several categories of data-centric approaches have emerged:

Data augmentation techniques, such as synonym replacement, paraphrasing, or noise injection, aim to increase dataset diversity and improve model robustness. The survey by Feng et al. (2021) finds that while simple methods offer modest gains, particularly when fine-tuning large pretrained models, more sophisticated techniques (e.g. back-translation, paraphrasing) are more effective, especially in low-resource or out-of-domain settings. Although most studies focus on classification tasks, they note that augmentation can benefit span-based tasks like NER if label consistency is preserved.

Adversarial data generation methods go further by creating inputs designed to challenge model predictions in more targeted ways. For example, Wang et al. (2020) introduces CAT-Gen, a controllable adversarial text generator that creates minimally different inputs to test decision boundaries.

Contrastive data augmentation encourages models to learn finer-grained distinctions between similar examples. Qu et al. (2021) propose the CoDA framework, which combines multiple transformations with a contrastive objective to improve generalisation.

Complementing these generation methods are tools focused on testing and interpreting model behaviour in a more structured fashion. One such system is **Errudite** (Wu et al., 2019), which enables scalable, reproducible, and testable error analysis. Rather than relying solely on automatic perturbations, Errudite allows users to define hypotheses about model behaviour and test them by crafting controlled examples. These can be automatically applied to datasets and grouped by error type, making it easier to detect systematic weaknesses and identify undergeneralised decision rules. In their application to NER, the authors found that even strong models often make consistent errors related to entity boundaries and rare contexts, revealing brittle generalisation beyond surface-level patterns. This approach reinforces the importance of interpretability and structured diagnosis in understanding model robustness.

This thesis builds on these ideas by applying structured, interpretable training data manipulations within the NER context. I examine how altering the presence or frequency of specific entity types affects generalisation on held-out datasets. This contributes to a deeper understanding of how training data influences model behaviour beyond architecture alone.

2.5 Evaluating Generalisation

As described in the previous section, generalisation in NLP involves the model's ability to perform well across variations in the input space and task setting. The evaluation of NLP models has traditionally relied on static benchmarks and single-metric reporting, such as accuracy or F1 score on held-out test sets. However, recent research highlights that these metrics often overestimate model generalisation and robustness, especially under distribution shifts or linguistic perturbations (Hupkes et al., 2023a). To address these limitations, multiple evaluation paradigms have emerged, focussing on systematic robustness testing, generalisation across domains, and behavioural consistency.

One prominent line of work is the evaluation of *cross-dataset generalisation*. Yogatama et al. (2019) proposed a framework to assess "general linguistic intelligence" by evaluating pretrained models on multiple datasets within the same task family. For example, QA models trained on SQuAD were tested on TriviaQA and QuAC without fine-tuning. Their results showed substantial performance drops, revealing a gap between in-domain and cross-domain generalisation.

Complementing this, Miller et al. (2020) introduced a natural distribution shift benchmark for QA models. They systematically evaluated the *drop in F1 score* (ΔF1) when models trained on SQuAD were tested on newly constructed datasets from diverse domains, such as Reddit and Amazon QA. These domain shifts caused drops in the F1 score of up to 17 points, demonstrating that strong in-domain performance does not guarantee robustness.

Ribeiro et al. (2020) proposed **CheckList**, a framework for behavioural evaluation inspired by software testing. It defines three test types: *Minimum Functionality Tests* (MFTs), which check whether a model handles simple cases correctly; *Invariance Tests* (INVs), which ensure the model's predictions remain stable under meaning-preserving perturbations and *Directional Expectation Tests* (DIRs), which test whether model predictions change in expected ways under controlled input modifications. CheckList enables systematic behavioural probing in a variety of NLP tasks.

Furthermore, extending robustness evaluation, Belinkov and Bisk (2018) explored how character-level *synthetic and natural noise* (e.g., keyboard typos, swaps, and mis-

spellings) affects machine translation performance. They showed that standard models degrade significantly even under simple noise perturbations, and that robustness to one type of noise rarely generalises to others. Their study supports training with noise-injected data to improve stability.

Glockner et al. (2018) introduced a challenge set for Natural Language Inference (NLI), targeting models' failure to perform basic lexical inferences. Their test set altered existing SNLI examples via *synonym replacements, hypernym shifts, and antonyms*, revealing sharp accuracy drops and a reliance on superficial cues over semantic understanding.

Factors such as parameter initialisation, batch ordering, and dropout introduce variability across training runs, even under fixed hyperparameters. These stochastic elements are typically controlled using a random seed, which ensures reproducibility by fixing the outcome of the random number generators used during training. However, research has shown that the results can still vary substantially between runs with different seeds. In McCoy et al. (2020), the authors show that the performance of the model **can vary widely between random seeds**, even when the training conditions are kept constant. They trained 100 BERT models on the same dataset with different random seeds and found that while in-domain performance remained stable, performance on a challenge set (HANS) varied dramatically. This underscores the importance of considering randomness when evaluating generalisation.

Taking into account all of these, I apply two main evaluation criteria:

- **Seed variability analysis**, which involves running multiple training runs with different random seeds to measure performance stability.
- **ΔF1**, This metric has become standard in recent robustness and domain-shift research (Miller et al., 2020), where the drop in F1 between in-domain and out-of-domain evaluations is consistently reported to quantify performance degradation.

2.6 Generalisation in Named Entity Recognition

This thesis focusses specifically on the task of Named Entity Recognition (NER), which involves detecting predefined categories such as persons, organisations, and locations in text. I used two widely adopted NER datasets: CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003) and OntoNotes 5.0 (Hovy et al., 2006), to study how models generalise across domains. Although both datasets include newswire text, OntoNotes covers a wider range of genres, such as telephone conversations, weblogs, and broadcast media. This makes it particularly useful for evaluating how well a model trained in one context performs in another. Together, these datasets allow me to test both cross-domain generalisation (across different genres) and within-domain generalisation (between datasets with the same genre but different distributions).

Although generalisation is a key concern in NLP, few studies have explored it specifically within NER. One of the most well-known is by Augenstein et al. (2017), who tested older NER systems across six datasets and showed that performance drops significantly when the model sees new domains or entity types. They argued that this is because these systems rely too much on surface features and memorised forms.

More recent work by Fu and Liu (2020) showed that even neural models struggle with rare or unfamiliar entities. They proposed evaluating generalisation more carefully

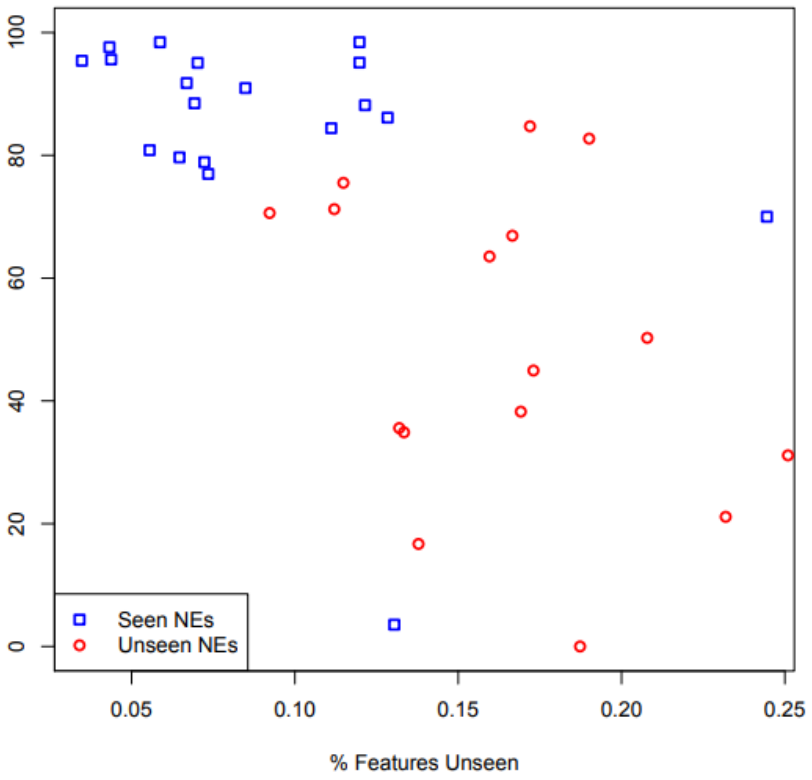

Figure 2.2: Reproduction of Figure 2 from (Augenstein et al., 2017, p. 22), showing F1 scores for seen vs. unseen NEs as a function of unseen features.

by looking at how models perform on specific types of entities, rather than just overall accuracy.

Other studies used challenge sets or adversarial examples to test robustness. For example, Reich et al. (2022) created perturbed test examples based on CoNLL and found that models often fail when entity contexts are slightly changed. Similarly, Lin et al. (2020) found that pretrained models often do not transfer well to new domains, especially when trained on biased data.

This thesis contributes to this body of work by combining robust evaluation methods with controlled training data manipulations to study how specific properties of the data influence NER generalisation. In particular, it explores how variation in entity distributions, proper/common noun ambiguity, and ethnic name origin affect the ability of BERT-based models to generalise across domains and perturbations.

2.7 State-of-the-Art Performance in NER

Recent developments in Named Entity Recognition (NER) have leveraged innovations in contextual representation learning and optimisation for data imbalance, leading to state-of-the-art performance on standard benchmarks. For the CoNLL-2003 dataset, which comprises newswire text annotated with four entity types, the model with the highest performance is based on automated encoding of embeds (ACE), achieving an F1 score of 94.6 Wang et al. (2021). This method automates the selection and con-

catenation of multiple pretrained embedding types, such as ELMo, BERT, and Flair, using a reinforcement learning controller trained to optimise downstream task performance. By dynamically learning which combination of embeddings is most effective for structured prediction, ACE achieves robust and transferable representations without manual tuning.

On the OntoNotes 5.0 corpus, which poses additional challenges due to its broader set of entity categories and multi-genre content, the best results are obtained by framing NER as a Machine Reading Comprehension (MRC) task combined with Dice loss optimisation, yielding an F1 score of 92.07 Li et al. (2020). In this setup, BERT is adapted to extract answer spans in response to queries representing each entity type, effectively casting entity recognition as span prediction. The model employs Dice loss, a function derived from the Sørensen-Dice coefficient, which better aligns with F1-oriented evaluation metrics and mitigates the negative effects of class imbalance by reducing the influence of abundant, easy-negative examples.

Although these methods represent the cutting edge in task-specific performance, they typically require complex architectures or computationally expensive reinforcement learning components. In contrast, vanilla BERT-based models offer an attractive balance between simplicity and performance. With minimal adaptation, BERT consistently achieves competitive scores above 94 F1 on CoNLL-2003 in ML4NLP settings and remains widely accessible, modular, and efficient. For this reason, BERT serves as a practical model to be used for this thesis.

Chapter 3

Datasets

3.1 CoNLL-2003 (English)

Curation Rationale The CoNLL-2003 dataset was introduced as part of the shared task on language-independent Named Entity Recognition (NER) at the Conference on Computational Natural Language Learning (CoNLL). The goal was to encourage research in multilingual NER by providing a standardised benchmark with consistent annotation guidelines across languages (Tjong Kim Sang and De Meulder, 2003). The dataset was created by repurposing and annotating existing newswire corpora for four named entity categories: persons, locations, organisations, and miscellaneous entities.

Language Variety The dataset includes two languages:

- **English:** British English as found in the Reuters RCV1 newswire collection.
- **German:** German-language newswire data from the ECI Multilingual Text Corpus.

The English portion is the most widely adopted in NER research. It serves as the basis for this thesis, and is therefore the primary focus of this data statement.

Annotator Demographics The dataset was annotated by employed annotators under the supervision of the CoNLL-2003 task organisers. Although detailed demographic information is not available, the annotation process followed clear guidelines to ensure consistency. The annotations were manually checked and corrected.

Text Characteristics The distribution of the English dataset is shown in 3.1:

Subset	Sentences	Tokens
Training	14,987	203,621
Development	3,466	51,362
Test	3,684	46,435

Table 3.1: Statistics for the English portion of the CoNLL-2003 dataset.

Each token is labelled with part-of-speech, chunking, and named entity tags. The named entity labels use the BIO tagging scheme.

Annotation Process The annotation followed the CoNLL-2003 guidelines, which define the named entity categories and the IOB2 tagging scheme (Tjong Kim Sang and De Meulder, 2003). The data was pre-processed with part-of-speech and chunk annotations using the MBT tagger, and then corrected manually. This annotation was also manually verified.

Inter-Annotator Agreement Inter-annotator agreement (IAA) scores were not formally published. However, annotation quality was monitored by manual revision, and inconsistencies were corrected by the task organisers. This absence of reported IAA is a known limitation of the dataset.

3.2 OntoNotes 5.0

Curation Rationale OntoNotes 5.0 was developed to provide a comprehensive, multilayered corpus to advance research in syntactic and semantic natural language processing. Created under the DARPA GALE programme (Olive et al., 2011), the dataset was designed to support tasks such as syntactic parsing, word sense disambiguation, coreference resolution, and semantic role labelling. It was curated to reflect various linguistic phenomena in multiple domains and languages (Hovy et al., 2006).

Language Variety OntoNotes 5.0 includes texts in three languages:

- **English:** American English, with some regional variation.
- **Chinese:** Mandarin Chinese (Simplified script).
- **Arabic:** Modern Standard Arabic.

For English, texts include both formal and conversational registers.

Annotator Demographics Annotation was conducted by trained linguists and computational linguists, most of whom were fluent in the language of the data. Specific demographic information about the annotators (e.g., nationality, gender, education) is not publicly available.

Speech Situation OntoNotes spans both spoken and written modalities. The English portion includes:

- **Spoken:** Broadcast conversation (interviews), broadcast news, and telephone conversations.
- **Written:** Newswire, magazine articles, weblogs, and conversational Web text (e.g. newsgroups).

Text genres were selected to ensure both domain diversity and coverage of formal/informal registers.

Text Characteristics The English corpus includes approximately 1.59 million words. Each document is annotated with multiple linguistic layers, including:

- Syntax (constituency and dependency)
- Proposition structures (predicate-argument structures)
- Named entities
- Coreference chains
- Word senses (linked to WordNet)

Genres and their proportions in the English portion (approximate) are provided in 3.2

Genre	**Percentage**	**Tokens (k)**	**Tokens (%)**
Newswire	25%	625	21.6%
Broadcast News	17%	200	6.9%
Broadcast Conversation	15%	200	6.9%
Weblog / Usenet (Web text)	20%	300	10.4%
Magazine (Tele / P2.5 mix)	8%	145	5.0%
Telephone Speech	15%	120	4.1%
Total (approx.)	100%	1,590	54.9%

Table 3.2: Genre distribution in the English portion of OntoNotes 5.0. Token counts (in thousands) and percentages are rounded.

Recording Quality The spoken data were transcribed from audio recordings with varying levels of recording quality. Broadcast speech is generally of high quality; telephone speech is of lower fidelity. The transcriptions were manually verified.

Annotation Process Each linguistic layer was annotated using task-specific guidelines and trained annotators. For example, coreference annotation followed the guidelines developed during the OntoNotes project. Multiple passes and quality control checks were used.

Inter-Annotator Agreement Reported agreement varies by annotation layer. For English:

- Part-of-speech tagging: 97.6% token-level agreement
- Syntactic constituents: 90.2% bracket agreement
- Word sense disambiguation: 80% sense agreement
- Coreference: agreement measured via MUC and B^3 scores over overlapping annotations

These figures reflect relatively high annotation consistency, although detailed agreement metrics are not uniformly reported across all layers.

3.2.1 Mapping OntoNotes 5.0 to CoNLL-2003

The OntoNotes 5.0 corpus defines a broad set of named entity categories, including conventional types such as `PERSON` and `ORG`, as well as finer-grained distinctions like `NORP` (nationalities, religious, or political groups), `FAC` (facilities), and `PRODUCT`. In contrast, the CoNLL-2003 dataset follows a more constrained annotation scheme, limited to four categories: `PER` (person), `ORG` (organisation), `LOC` (location), and `MISC` (miscellaneous).

To support consistent evaluation and enable cross-dataset experiments, I map OntoNotes entity labels to the CoNLL-2003 tag set. Categories with direct counterparts are mapped straightforwardly (e.g., `PERSON` to `PER`), while all others that do not fall under CoNLL's core categories are assigned to `MISC`. Categories such as `DATE`, `MONEY`, or `CARDINAL`, which are not annotated in CoNLL-2003, are excluded from consideration.

The complete mapping is summarised in Table 3.3.

Table 3.3: Mapping of OntoNotes-5.0 entity types to CoNLL-2003 categories

OntoNotes Category	**CoNLL Category**	**Notes**
PERSON	PER	Direct match (e.g., "John", "Mary")
ORG	ORG	Direct match (e.g., "Google", "UN")
GPE	LOC	Treated as location (e.g., "France")
LOC	LOC	Physical locations (e.g., "Nile River")
NORP	MISC	No direct equivalent (e.g., "Christian")
FAC	MISC	Facilities (e.g., "JFK Airport")
EVENT	MISC	Named events (e.g., "Olympics")
WORK_OF_ART	MISC	Books, songs, etc. (e.g., "Hamlet")
LAW	MISC	Legal documents (e.g., "First Amendment")
PRODUCT	MISC	Products (e.g., "iPhone")
LANGUAGE	MISC	Languages (e.g., "Mandarin")
DATE	(Excluded)	Not annotated in CoNLL-2003
TIME	(Excluded)	Not annotated in CoNLL-2003
MONEY	(Excluded)	Not annotated in CoNLL-2003
PERCENT	(Excluded)	Not annotated in CoNLL-2003
QUANTITY	(Excluded)	Not annotated in CoNLL-2003
ORDINAL	(Excluded)	Not annotated in CoNLL-2003
CARDINAL	(Excluded)	Not annotated in CoNLL-2003

Chapter 4

Methodology

4.1 Chapter Overview

This chapter presents the methodological foundation for investigating how Named Entity Recognition (NER) models generalise across different domains and linguistic variations. The core objective is to determine whether these models can move beyond surface-level heuristics and perform robustly when faced with unfamiliar or ambiguous input.

The chapter begins by detailing the datasets and model configurations used in the study, outlining the training procedures and evaluation protocols that form the experimental backbone. This includes an explanation of how the models were fine-tuned and evaluated on both in-domain and out-of-domain data.

To establish the reliability of the findings, I then examined the model's stability across multiple random seeds. This step enables us to quantify the degree to which observed patterns are robust to stochastic variation, providing a crucial consistency check before undertaking a deeper analysis (McCoy et al., 2020).

Building on this foundation, I explore potential biases within the training data itself. Through targeted analyses of tag frequency distributions, capitalisation patterns, and demographic representation, particularly in terms of ethnicity and gender, I identify structural imbalances that can shape model behaviour. These brief findings help motivate the design of targeted evaluation scenarios.

Finally, to investigate these issues in a more controlled and interpretable way, I turn to the construction of custom challenge sets. These datasets were designed to test generalisation under conditions that are either under-represented or inherently ambiguous, such as unfamiliar name origins, common noun overlap, and confusion between similar entity types like persons and organisations. Each challenge set is introduced with its underlying motivation and design logic.

4.2 Experimental Setup

This section outlines the datasets, model architectures, and training procedures used throughout this study. These components form the foundation for all subsequent analyses of model generalisation, bias, and robustness.

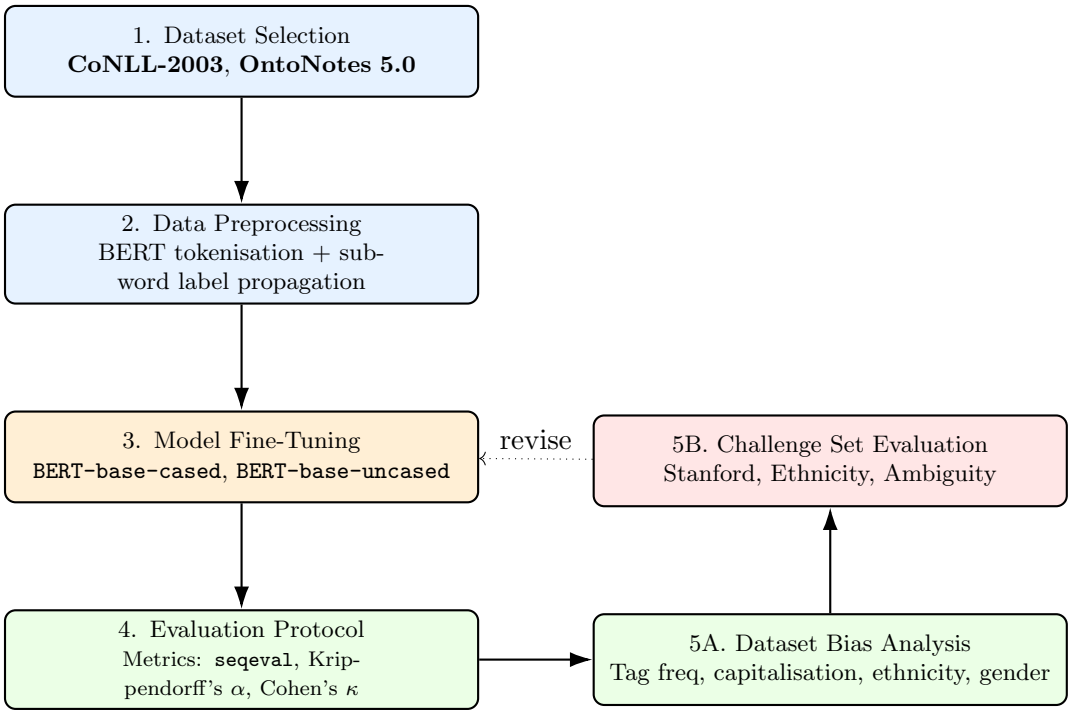

Figure 4.1: Overview of the experimental workflow used to assess model generalisation, robustness, and bias.

4.2.1 Datasets

I used two widely adopted English NER datasets: **CoNLL-2003** (Tjong Kim Sang and De Meulder, 2003) and **OntoNotes 5.0** (Hovy et al., 2006) as mentioned in Chapter 3; for convenience, I briefly summarise the key characteristics here. These datasets differ substantially in their domain coverage, annotation guidelines, and label distributions, providing a valuable basis for evaluating cross-domain generalisation.

CoNLL-2003 consists primarily of newswire text with four entity types: `PER` (person), `LOC` (location), `ORG` (organisation), and `MISC` (miscellaneous). In contrast, OntoNotes 5.0 combines data from multiple domains and provides annotations for a wider range of entity types. For comparability, I restrict OntoNotes to the same four entity classes as CoNLL, discarding other tags and sentences without named entities.

The sentences of each dataset are tokenised using the Hugging Face implementation of the BERT tokeniser to ensure consistency between the preprocessing and model input formats.

4.2.2 Model Architectures

I used BERT-based architectures for all experiments. BERT has been shown to perform strongly on token-level classification tasks such as NER, due to its ability to model rich contextual dependencies via bidirectional self-attention. Its pretrained representations capture both syntactic and semantic information, making it well-suited for identifying entity boundaries and disambiguating entity types in context.

Specifically, I fine-tune the following pretrained models:

- **BERT-base-cased**

- **BERT-base-uncased**

These variants were selected to examine the impact of casing information during pretraining, which can influence performance in case-sensitive tasks such as NER. Each model is fine-tuned independently on both CoNLL and OntoNotes, allowing for direct comparison of training dataset and architecture effects.

4.2.3 Training Procedure

The models are fine-tuned using the standard token classification head from the Hugging Face Transformers library. Each input sentence is tokenised using WordPiece tokenisation, and the gold label for each word is propagated to all its subtokens. This encourages the model to produce consistent predictions across subword units and reinforces entity boundaries within tokenised spans. During the evaluation, the predictions are aggregated at the word level using a majority vote on the subtokens to align with the original granularity of the annotation. The hyperparameters are kept constant throughout all runs: a learning rate of 2e-5, a batch size per device of 8, and a gradient accumulation over 2 steps to produce an effective batch size of 16. Each model was trained for 3 epochs with a weight decay of 0.01. To optimise GPU memory efficiency, mixed precision training (`fp16`) was enabled.

To ensure reproducibility, training is conducted using fixed seeds where specified, and GPU acceleration is enabled via PyTorch with CUDA support. In addition, I used the default token classification collator provided by the Hugging Face library, to ensure compatibility with the token classification task setup.

4.2.4 Data Preparation

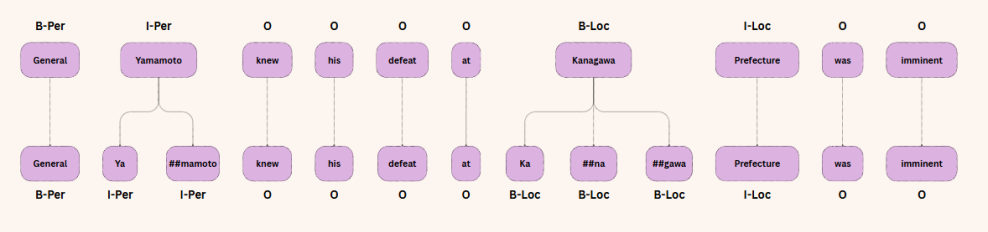

Figure 4.2: Subword tokenisation and entity label propagation in the `bert-base-cased` model. Each word is split into subword units using WordPiece tokenisation, and the corresponding NER label is propagated to all subwords.

Each sentence was first tokenised using a BERT-compatible tokeniser configured to treat the input as a sequence of individual words. This ensured accurate alignment between the original tokens and their associated entity labels. The tokeniser automatically applied truncation and padding as needed.

Given BERT's reliance on subword tokenisation (e.g., WordPiece), a single word could be split into multiple subword tokens. To maintain label consistency, each subword token inherited the label of its parent word, as illustrated in Figure 4.2. Special tokens such as `[CLS]`, `[SEP]`, and padding tokens were assigned a label of `-100`, following PyTorch's `ignore_index` convention, to exclude them from loss computation during training (Wolf et al., 2020).

4.2.5 Fine-Tuning

All models were fine-tuned using the standard token classification setup described in Section 4.2.3. Each configuration, defined by model variant and training data set, was independently trained using the same hyperparameters and training procedure.

To account for variability due to random initialisation, each configuration was trained with five different random seeds. The final evaluation scores are reported as the average in these five runs.

4.2.6 Evaluation Protocol

During the evaluation, the prediction of the model at the subword level was realigned with the original word-level annotations. This involved removing special tokens and grouping subword predictions based on word-ID mappings. When multiple sub-words corresponded to a single word, the most frequent predicted label among them was selected as the representative label. In cases of ties or no clear majority, the label of the first subword was used.

Once alignment was complete, all sequences were flattened and evaluated using the `seqeval` package. For each model, the precision, recall and F1 scores were reported at the entity level on the test set. All evaluations used the same tokeniser and label mapping to ensure comparability.

In addition to standard within-dataset evaluation, I also assess model performance on cross-dataset transfer: models trained on CoNLL are evaluated on the OntoNotes test set, and vice versa. These cross-dataset results provide a baseline for assessing robustness to partial domain shift, given the differences in annotation guidelines and domain coverage between the datasets. Beyond these benchmarks, further evaluation is conducted on custom challenge sets specifically designed to test generalisation under conditions such as ambiguous naming and demographic under-representation. These include variations in name origin and lexical ambiguity, which are not systematically covered in the original datasets.

4.3 Assessing Seed Sensitivity

To ensure that the findings are not the product of random initialisation (McCoy et al., 2020), I evaluated the consistency of the model predictions between multiple training seeds. BERT-based models introduce stochasticity through random weight initialisation, dropout, and data shuffling. As a result, small variations in training conditions can potentially lead to different learnt representations and outputs. To test the stability of the results, I train each model configuration using five distinct random seeds: 33, 42, 57, 106, and 812. All other hyperparameters are kept constant. Models are trained to completion using the same training procedure described in Section 4.2.3. After training, I evaluate each seed variant on the same test set and collect token-level predictions. These predictions are aligned with gold-standard annotations.

4.3.1 Agreement Metrics

To quantify the consistency of predictions across seeds, I compute two inter-model agreement metrics. First, I calculate Cohen's kappa (Cohen, 1960) between each pair of seed runs, treating each model as a separate annotator. This score adjusts for

expected agreement by chance and provides a standard measure of reliability in binary and multiclass settings. Second, I compute Krippendorff's alpha (Krippendorff, 2004), a more generalised metric that supports multiple raters and nominal data. This measure is particularly useful for aggregating the agreement between all five seeds in a single score, providing a robust summary of stability for each model configuration. These agreement metrics serve as a reliability check, ensuring that the patterns observed in subsequent analyses are not artifacts of a single training run. Although detailed scores are presented in Chapter 5, the high alpha values (0.950 - 0.968) observed in most models justify the use of single seed runs in exploratory experiments later in this chapter.

4.4 Entity Frequency and Dataset Bias

To understand the data-driven factors that can influence model behaviour, I performed a set of preliminary analyses on the CoNLL-2003 and OntoNotes 5.0 training sets. These analyses focus on tag frequency, capitalisation patterns, and demographic distributions, factors that can implicitly shape model representations and generalisation capacity. The findings of these experiments are discussed in depth in Chapter 5; here, I provide an overview of the analyses conducted as a basis for understanding potential sources of bias.

4.4.1 Tag Distribution

I begin by examining the relative frequency of the named entity types in each dataset. Entity tokens are grouped by their labels: `PER`, `ORG`, `LOC`, and `MISC`, excluding the non-entity `O` tag.

This analysis provides insight into which entity types are most frequent during training, and thus most likely to influence model behaviour. It also allows us to assess whether models may be disproportionately optimised for certain entity classes. A notable observation is the prominence of `PER` tags in both datasets, which guided the decision to use person entities as the focus of the construction of the challenge set in later experiments.

4.4.2 Capitalisation Patterns

Next, I analyse the use of capitalisation across entity tokens. For each entity class, I compute the proportion of tokens that begin with an uppercase letter. This allows us to assess how much casing acts as a surface-level cue for entity recognition.

Differences in capitalisation behaviour across datasets are used to inform subsequent experiments on casing robustness. In particular, I explore whether models trained on datasets with highly consistent casing patterns, such as CoNLL, may become overly dependent on orthographic signals. This motivates the inclusion of cased model variants in later chapters.

4.4.3 Ethnicity Analysis

To assess potential biases in the representation of person entities, the ethnic distribution of names in each dataset is estimated. Full person names are reconstructed from `B-PER` and `I-PER` spans, deduplicated, and passed to a name-based ethnicity classifier.

For this, I use the `ethnicseer` Python library, which uses character n-gram features and supervised classifiers such as LSTM and logistic regression to predict name ethnicity. These models are trained on data sets that include US voter records, Wikipedia biographies, and census data Treeratpituk (2018).

The predicted ethnicity codes are mapped to broader regional categories (e.g., European, East Asian, South Asian, etc.) for interpretability. This analysis provides a high-level view of the cultural and geographic diversity present in the training data.

The observation of a strong Western bias in both data sets informed the decision to construct a set of challenge sets that contained under-represented name types (Section 4.5.2).

4.4.4 Gender Representation

I also examine gender representation among person entities by estimating the gender of first names using a rule-based name-gender mapping tool. I use the `gender-guesser` Python library (B, 2022) that draws on a large database of international name-gender associations. The library outputs one of several labels: `male`, `female`, `mostly_male`, `mostly_female`, `andy` (androgynous), and `unknown`.

This analysis is intended to quantify the gender balance in the data and raise awareness of any under-representation. As with ethnicity, the findings here guided the sampling strategy for the challenge sets introduced later in this chapter.

4.5 Challenge Sets

To test the generalisation capacity of the models beyond the biases observed in the training data, I evaluate them on both an existing adversarial benchmark and a series of newly constructed challenge sets. These sets are designed to isolate specific linguistic or demographic variables and provide a more controlled way to evaluate the robustness of the model under variation and ambiguity.

Given the iterative nature of this process, some of the findings discussed in the next chapter are briefly introduced here to motivate the construction of the corresponding challenge sets.

4.5.1 Stanford Challenge Set

As a baseline for robustness testing, I use the Stanford Challenge Set, a curated subset of the CoNLL-2003 test data enhanced with expert-guided adversarial examples (Reich et al., 2022). After automatically generating adversarial variants, the authors **manually selected 1,000 high-quality examples** for evaluation. These examples were created using **385 distinct token/context manipulation patterns**: 44 token insertions and 42 context phrases for `ORG`, 82 token deletions and 16 context phrases for `LOC`, and 152 token insertions with 49 context phrases for `PER`.

The generation process involves: (i) an **eligibility check** to identify changeable entities (e.g., single-token "Brazil"); (ii) **entity token changes** such as adding "University" to convert a location into an organization; and (iii) **contextual modifications** like appending "and her team" to shift the entity type to a person. These steps are guided by expert-curated phrase sets and reinforced through manual filtering to ensure fluency and validity. Some examples of this are shown in Table 4.1

Transition	Count	Examples
Location or Person → Organization	510	**Original:** Every year, 500 new plastic surgeons graduate in Brazil and medical students from all over the world come to study there. **Augmented:** Every year, 500 new plastic surgeons graduate from Brazil University and medical students from all over the world come to study there.
Organization → Location	99	**Original:** Munich Re says to split stock. **Augmented:** Munich's largest corporation says to split stock.
Organization or Location → Person	391	**Original:** The Colts won despite the absence of injured starting defensive tackle Tony Siragusa, cornerback Ray Buchanan and linebacker Quentin Coryatt. **Augmented:** Colts Zardari and her team won despite the absence of injured starting defensive tackle Tony Siragusa, cornerback Ray Buchanan and linebacker Quentin Coryatt.

Table 4.1: Reproduction of Table 1 from (Reich et al., 2022, p. 1949), Expert-guided transition types for producing adversarial augmentations for NER. The original entity is colored in blue, the entity token change is colored in red, and the entity context change is colored in brown.

I apply the same evaluation protocol as in previous experiments, using the `seqeval` package to report tag-wise F1 scores. Four BERT-based configurations are tested: cased and uncased variants trained on CoNLL and OntoNotes, respectively. This benchmark provides a structured and externally validated basis for evaluating robustness prior to introducing my own challenge sets.

4.5.2 Custom Challenge Sets

Following the analysis of training data biases (see Section 4.4), I constructed several custom challenge sets to target specific failure modes and limitations. These sets were designed iteratively, using information from previous evaluations to motivate the inclusion of particular entity types, name origins, or linguistic ambiguities.

Each challenge set was built using hand-crafted templates and external lexical resources, with consistent annotation and tokenisation pipelines. The entity spans were tagged using BIO labels based on exact string matching, consistent with the tokenisation strategy used during model training.

To construct the ethnicity-based challenge set, I manually designed sentence templates to place names in various syntactic and discourse contexts, e.g., *"My name is {name}."*, *"The person you spoke to earlier, {name}, will send the report."*, or *"{name} and {name2} handled the presentation together."*. These examples were created from scratch to ensure a controlled and interpretable evaluation. The variety of templates ensures that the challenge examples reflect realistic usage while controlling for entity surface form and placement.

Ethnic Name Challenge Set

After identifying a Western-centric bias in both CoNLL and OntoNotes, I developed a challenge set containing names from underrepresented regions to assess how well models

generalise to unfamiliar name forms.

I used the `NameDataset` library, which compiles male and female names across a wide range of countries from sources such as the US Social Security Administration, the UK Office of National Statistics, and public datasets derived from the Facebook data leak Graham (2018). To address potential ethical concerns, the first and last names were randomly combined to avoid referencing real individuals.

The countries were grouped by continent using the `pycountry` Serafim and contributors (2023) and `pycountry-convert` libraries, enabling balanced sampling from six regions: Asia, Africa, South America, North America, Europe, and Oceania.

For each region, 1,000 sentences were generated using 28 predefined templates of varying syntactic structure, including both declarative and interrogative forms (e.g. "Have you met {name} from accounting?", "{name} left early today."). This yielded six region-specific datasets for evaluation.

Common Noun Ambiguity Challenge Set

Fortunately, the initial results showed strong performance in many non-Western names, especially South America and Asia (see chapter 5.4.1), despite my hypothesis that unfamiliar names would lead to higher error rates. Upon closer inspection, I found that most misclassifications stemmed not from ethnicity but from **semantic ambiguity**.

Names such as *"Mark"*, *"Will"*, or *"Hope"* were often misclassified because they double as common nouns or verbs. These errors were particularly frequent in contexts where surface cues like capitalisation or position in the sentence were unreliable, suggesting that the model struggled to use context to resolve ambiguity.

To test this hypothesis systematically, I developed a dedicated common noun challenge set. First, I identified ambiguous names by cross-referencing the most frequent US names from `NameDataset` with common English nouns from WordNet via `nltk.corpus.wordnet`. The names appearing on both lists were retained.

I then embedded these names into templates designed to obscure surface-level cues, similar to those used in the ethnicity challenge set. The resulting sentences were annotated with BIO tags and used to assess the model's ability to disambiguate named entities based solely on contextual understanding.

Chapter 5

Results

5.1 Overview

This chapter presents the results of a series of experiments designed to evaluate the generalisation and robustness of BERT-based Named Entity Recognition (NER) models. The evaluation begins with base performance comparisons between models trained on CoNLL-2003 and OntoNotes 5.0. Then, I examine sensitivity to random seed initialisation, assess robustness under adversarial conditions using the Stanford Challenge Set, and perform a series of targeted bias and diagnostic analyses.

Particular attention is paid to the influence of the dataset's characteristics, such as the frequency, capitalisation, gender, and ethnicity distributions of the entities. These analyses reveal structural imbalances that may affect model generalisation. To further investigate these limitations, I construct and evaluate challenge sets focused on ambiguous or under-represented phenomena, namely, ethnically diverse names, common noun ambiguity, and person-organisation confusion.

By systematically testing model behaviour across these settings, several recurring challenges are revealed, such as overreliance on superficial features (e.g., capitalisation) and difficulty resolving semantically ambiguous tokens. These findings inform both the interpretation of the current model's performance and the design of future improvements to build more robust NER systems.

5.2 Entity Frequency and Dataset Bias

5.2.1 Tag Distribution

A comparative analysis of entity type distributions in the training data reveals notable structural differences between CoNLL-2003 and OntoNotes 5.0 (Figure 5.1 and 5.2). In CoNLL, the `PER` (person) category dominates at 33.3%, followed by `ORG` (28.8%), `LOC` (24.4%), and `MISC` (13.5%). OntoNotes, on the contrary, shows a more even spread in aggregate: `ORG` leads with 31.7%, while `PER`, `LOC`, and `MISC` appear at 27. 0%, 23. 1% and 18. 2%, respectively.

However, a genre-level breakdown of OntoNotes reveals substantial variation. In the *newswire* portion, `ORG` and `LOC` are the most frequent types, comprising 20.7% and 15.2% of the tags, respectively, while `PER` accounts for just 11.7%. Conversely, the *non-news* genres show a higher relative frequency of person entities, with `PER` rising to 18.4% and `ORG` dropping to 9.4%. The `MISC` category also becomes more prominent,

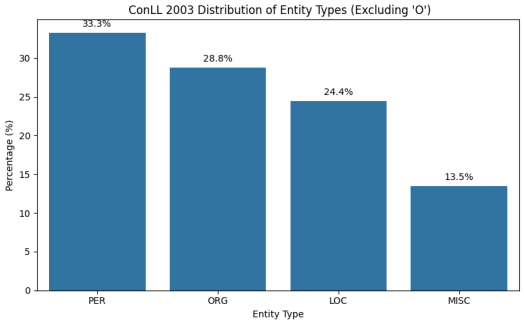

Figure 5.1: CoNLL 2003: Distribution of Entity Types (Excluding `O`)

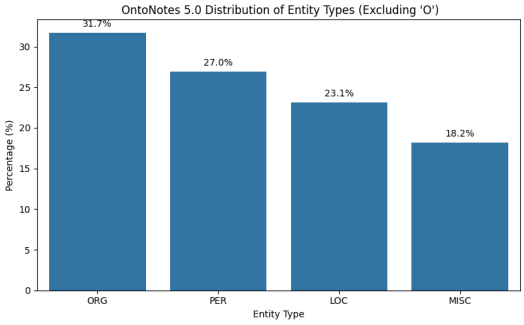

Figure 5.2: OntoNotes 5.0: Distribution of Entity Types (Excluding `O`)

increasing from 6.9% to 13.5%.

These distributions offer early clues about how training data might shape model behaviour. Models trained on CoNLL may become particularly attuned to person entities due to their prominence, which could help explain the strong performance seen in `PER` tags across evaluations. However, this also raises the risk of overfitting to a narrow class distribution, potentially limiting generalisation to other entity types.

The OntoNotes dataset, with its more balanced entity proportions and genre-specific shifts, may encourage broader generalisation while exposing models to more contextual variability. This difference becomes especially relevant under domain shift, where genre-specific biases may either hinder or aid transfer, depending on the match between training and test distributions.

The predominance of `PER` labels across both corpora also motivated my subsequent focus on ambiguity and demographic diversity within person entities. Since names are highly variable across languages and cultures and frequently overlap with common nouns, they serve as a productive site for exploring generalisation, ambiguity, and latent model biases.

5.2.2 Capitalisation Patterns

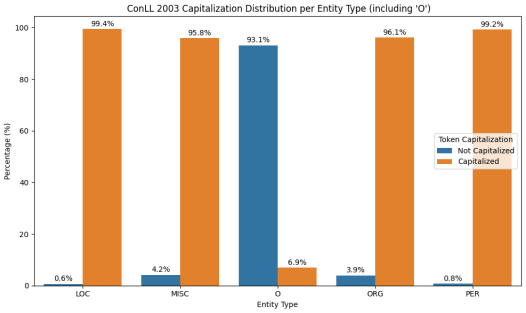

Figure 5.3: CoNLL 2003 Capitalisation Distribution per Entity Type

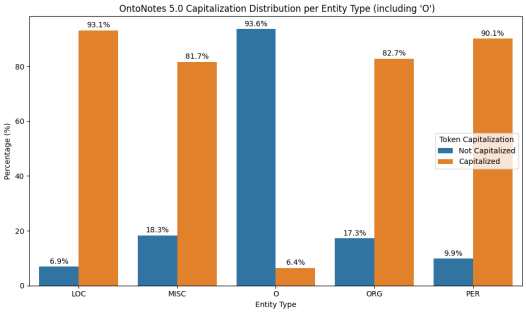

Figure 5.4: OntoNotes 5.0 Capitalisation Distribution per Entity Type

Capitalisation trends differ significantly between CoNLL-2003 and OntoNotes 5.0, revealing key surface-level biases. In CoNLL, almost all entity tokens are capitalised, with more than 95% for each entity type (`PER`: 99.2%, `LOC`: 99.4%, `ORG`: 96.1%, `MISC`:

95.8%). This consistency makes the case a highly reliable feature for entity detection, potentially encouraging over-reliance on orthographic signals.

OntoNotes exhibits substantial variation in the relationship between capitalisation and named entity status across its constituent genres. For example, in the newswire ("nw") domain, 66.9% of capitalised tokens are labelled as named entities. However, this proportion drops dramatically in other genres, such as telephone conversation ("pt"; 0.0%), web logs ("wb"; 22. 8%) and telephone conversation ("tc"; 16.2%). These figures highlight the domain diversity in OntoNotes and the corresponding variability in how capitalisation cues correlate with entity labels. In genres like "pt" and "tc", capitalisation is not a reliable signal of entityhood, probably due to a lack of fluency or transcription conventions. This diversity forces models trained on OntoNotes to rely less on superficial features like casing and more on contextual cues, which may help explain their better generalisation performance observed in cross-domain evaluations.

Models trained on CoNLL appear to rely heavily on surface features such as capitalisation when predicting named entities. This dependency is likely a consequence of the regular formatting and consistent casing of CoNLL, where most named entities are capitalised, and common nouns are not. As a result, models trained on CoNLL may implicitly treat capitalisation as a shortcut for named entity detection. To verify this, I analysed errors involving capitalised tokens labelled as 'O' (non-entity) in the gold annotations. In the OntoNotes test set, the CoNLL-trained model incorrectly predicted 6.7% of these capitalised 'O' tokens as named entities. In contrast, the OntoNotes-trained model made these errors only 4. 7% of the time on the CoNLL test set. This suggests that CoNLL-trained models are more likely to overpredict entity labels based on superficial cues such as casing, whereas OntoNotes-trained models, having been exposed to more variation and noise in the training data, are less dependent on casing and instead rely more on contextual information. This reduced sensitivity to casing in OntoNotes-trained models is consistent with a better generalisation to out-of-domain or noisy inputs.

5.2.3 Ethnicity

To assess demographic bias in the datasets, I examined the ethnic distribution of `PER` (person) entities in both CoNLL-2003 and OntoNotes 5.0. This analysis aimed to uncover whether certain ethnic or national name origins are disproportionately represented, which could influence the ability of models to generalise to names from under-represented groups.

Method The analysis was performed using a name-based classification approach. First, the full names of the people were reconstructed by combining contiguous `B-PER` and `I-PER` tokens in the data. Next, names were deduplicated and passed through the ethnicseer classifier, a tool that uses character n-gram models and demographic training data (e.g., US voter rolls, Wikipedia bios, census data) to predict name ethnicity (Treeratpituk, 2018; Ambekar et al., 2009). The predicted ethnicity codes were then assigned to broader nationality groupings for interpretability.

Findings Figure 5.5 and Figure 5.6 show the ethnicity distributions. CoNLL-2003 shows a strong Western bias, with English names making up 36.4%, followed by German (16.0%) and French (8.9%). Non-Western groups, such as Chinese, Korean, and Vietnamese, together account for under 3%. OntoNotes 5.0, while slightly more diverse, is still dominated by English (37.3%) and Chinese (25.7%) names. Names of Indian, Arabic, and other under-represented origins appear in much lower proportions.

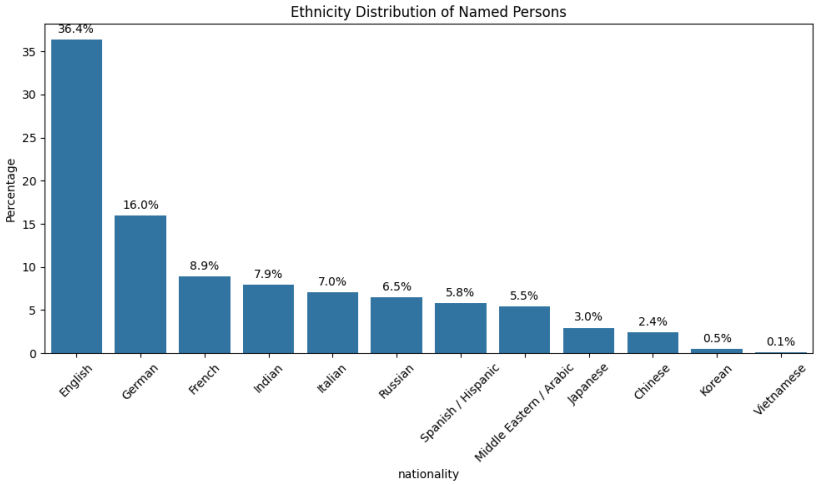

Figure 5.5: CoNLL 2003 Ethnicity Distribution of Named Persons

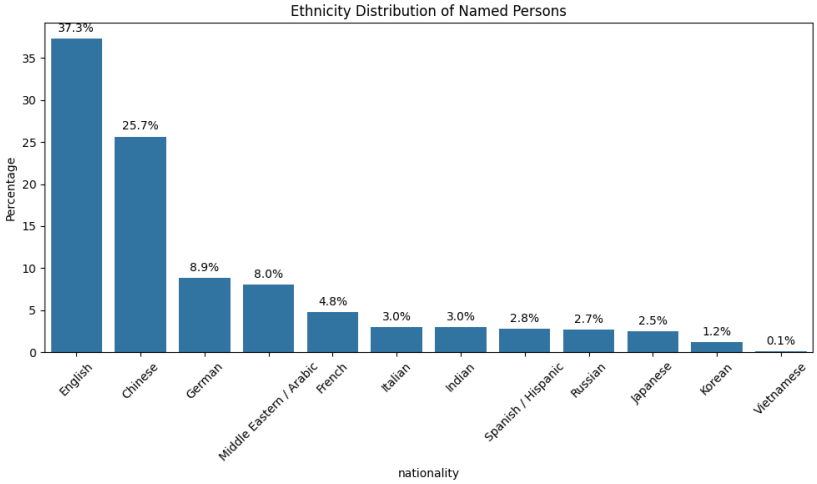

Figure 5.6: OntoNotes 5.0 Ethnicity Distribution of Named Persons

Implications for Generalisation The dominance of Western name forms, particularly in CoNLL, suggests that BERT-based NER models may be implicitly biased toward these patterns. As a result, models may struggle to identify or classify names from culturally or linguistically diverse backgrounds. This is especially relevant for real-world applications, where the diversity of names is much broader than what training corpora represent.

Such imbalances can lead to lower recall for non-Western names and decreased robustness in out-of-domain contexts. Based on these findings, I hypothesise that the model's performance is skewed in favour of Western entities. Following Mishra et al. (2020), who proposed an ethnicity-based evaluation for NER models, I constructed a target challenge set consisting of person names from various ethnic groups to empirically test this hypothesis in subsequent sections.

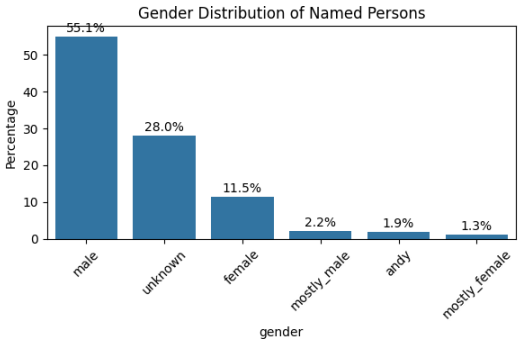

Figure 5.7: CoNLL 2003 Gender Distribution of Named Persons

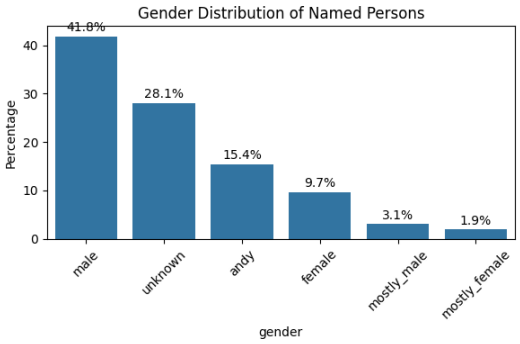

Figure 5.8: OntoNotes 5.0 Gender Distribution of Named Persons

5.2.4 Gender

The gender distribution of the `PER` entities across both datasets using a name-based classification method. The first names were extracted from the reconstructed named entities, and the gender was predicted using the `gender-guesser` library. This rule-based tool assigns labels such as `male`, `female`, `mostly_male`, `mostly_female`, `andy` (androgynous), and `unknown` based on international naming patterns.

Figure 5.7 shows that CoNLL 2003 is heavily skewed towards male names, with 55.1% classified as male and only 11.5% as female. A significant portion (28.0%) could not be confidently classified and was labelled `unknown`. The OntoNotes 5.0 dataset (Figure 5.8) shows a similar trend, though slightly more balanced: 41.8% male and 9.7% female, with a higher proportion of androgynous or ambiguous names (e.g. 15.4% `andy`).

These distributions reflect an underlying gender imbalance in both data sets, likely originating from bias in the original text sources. Such a skew can result in models disproportionately favouring male-associated patterns, potentially reducing performance or fairness when dealing with female or gender-neutral names. This reinforces concerns about representational bias in NER datasets and the importance of considering demographic diversity in training and evaluation pipelines. This also motivated the challenge sets to use more female names to test this issue.

5.3 Basic Experiments

5.3.1 Base Performance: CoNLL vs. OntoNotes

This section examines the generalisation ability of BERT-based Named Entity Recognition models across dataset boundaries. The primary aim is to assess how well models trained on one dataset (e.g., CoNLL-2003 or OntoNotes 5.0) perform when evaluated on another, highlighting the challenges of cross-domain robustness. As part of this analysis, I also compared cased and uncased variants of BERT to determine whether surface-form features, such as capitalisation, influence generalisation performance. Rather than selecting an optimal model for deployment, the focus here is on understanding how design choices affect robustness under distributional shifts.

To establish a reliable baseline, I begin by evaluating both cased and uncased BERT variants on the datasets on which they were trained. These initial comparisons are

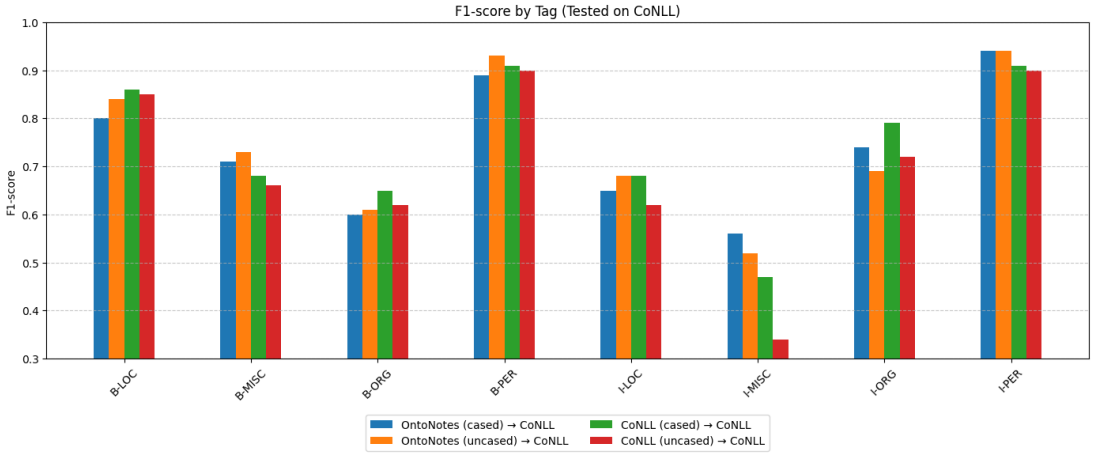

Figure 5.9: Tag-wise F1-scores for models evaluated on CoNLL-2003.

best understood in light of the properties of the training data discussed in Section 5.2. As shown in Figure 5.9, both models perform strongly in-domain on CoNLL-2003, achieving a macro-average F1 score of 0.94. The cased model slightly outperforms the uncased model on several tags, particularly those with orthographic cues such as *B-ORG*, *I-ORG*, and *I-PER*. Ambiguous tags such as *I-MISC* remain consistently harder across models. This aligns with the distributional irregularity of the `MISC` tag noted above, which appears less frequently and less consistently in both datasets.

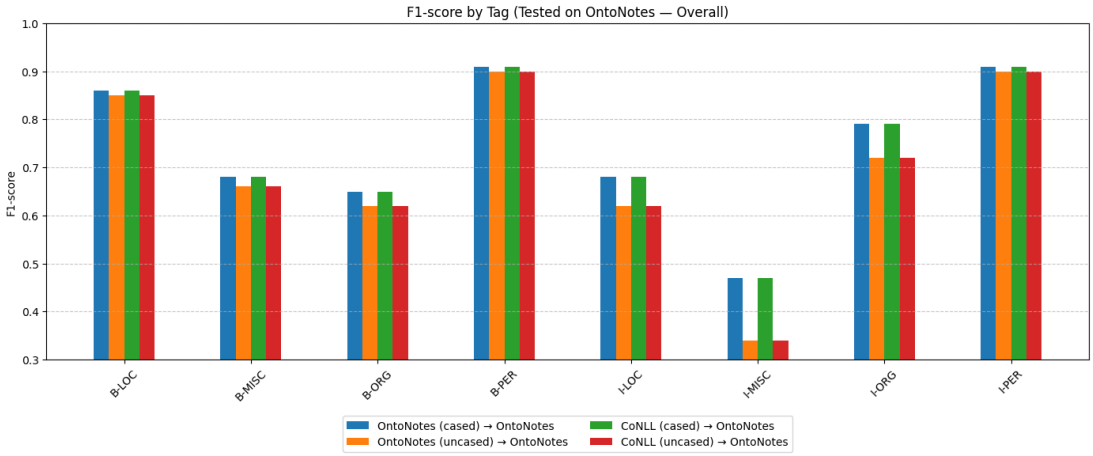

Figure 5.10: Tag-wise F1-scores for models evaluated on OntoNotes 5.0 (overall).

Figure 5.10 shows the in-domain performance on OntoNotes. Again, both models perform well, with the cased variant achieving a macro-average F1 of 0.94 and the uncased variant scoring 0.92. Despite strong overall performance, the tag variance is visible, especially for *I-MISC*, which suffers from inconsistent surface forms. The consistent advantage of the cased model supports the hypothesis that surface features contribute to boundary detection.

Further insight comes from separating OntoNotes into its newswire and non-newswire subsets, as shown in Figure 5.11. Here, the models perform better on the newswire text. The cased model reaches a macro F1 of 0.78 on newswire but only 0.65 on other genres;

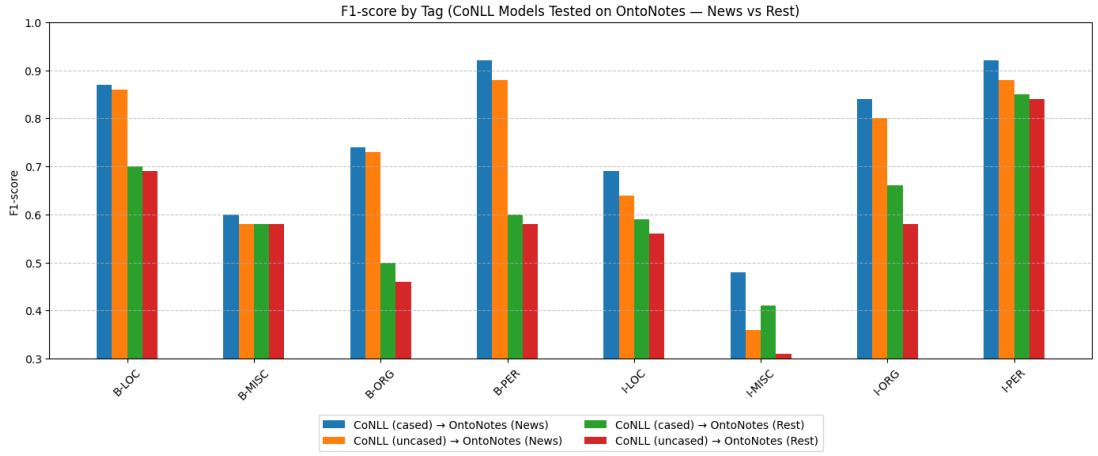

Figure 5.11: Tag-wise F1-scores for models evaluated on OntoNotes news vs. other genres.

for the uncased model, these scores are 0.75 and 0.62, respectively. This genre-specific gap highlights that OntoNotes exhibits internal domain diversity and that the similarity between train and test sets in terms of genre has a significant impact on performance. As shown in Section 5.2, genre-based shifts in both entity frequency and capitalisation may help explain these performance differences.

The results of the cross-dataset reveal a similar pattern. When trained on CoNLL and evaluated on OntoNotes, macro F1 drops to 0.77 (cased) and 0.73 (uncased), despite high in-domain scores. The decline is especially sharp on non-newswire genres (see Figure 5.11), reaffirming the impact of genre mismatch. Difficult tags such as *I-MISC* and *B-ORG* show the highest degradation, suggesting that models trained on narrow distributions struggle to generalise ambiguous boundaries.

In contrast, training on OntoNotes leads to more robust cross-domain performance. When tested on CoNLL, models trained on OntoNotes achieve macro F1 scores of 0.77 (cased) and 0.76 (uncased), nearly matching those of their CoNLL-trained counterparts. This result implies that training on a more diverse corpus helps models generalise better to more homogeneous test data. This is consistent with the broader and more balanced entity distributions reported in OntoNotes, which expose models to a wider variety of contexts and surface cues.

Overall, across all conditions, models that retain casing outperform their uncased variants, especially in cross-domain and mixed-genre settings. Capitalisation appears to aid generalisation, particularly for tags where surface-form patterns are semantically informative. These results, shown across Figures 5.9 - 5.11, suggest that surface signals, such as casing, contribute meaningfully to both label consistency and transfer robustness. All results are averaged over five random seeds to ensure stability across initialisations.

Interestingly, I observe that the I-MISC tag is consistently better learnt when training on OntoNotes, with F1 scores of 0.85 (cased) and 0.77 (uncased), compared to 0.47 and 0.34, respectively, when trained on CoNLL. This might reflect the higher frequency and broader representation of `I-MISC` in OntoNotes as seen in Fig. 5.2. This suggests that OntoNotes provides richer or more varied supervision for this particular label. However, models trained on CoNLL still perform reasonably well overall on the

OntoNotes test set, particularly for high-frequency entity types such as PER and LOC.

5.3.2 Seed Sensitivity and Prediction Stability

To determine the extent to which model predictions are influenced by random initialisation, I evaluated seed sensitivity across all four BERT configurations. Each model is trained five times using distinct random seeds (33, 42, 57, 106, 812), with fixed hyperparameters and identical training procedures. Evaluation is conducted under both in-domain and cross-domain conditions: specifically, models trained on CoNLL-2003 are evaluated on OntoNotes 5.0, and vice versa. The resulting predictions are aligned at the token level and compared against the gold standard annotations to assess variation in output due to random seed effects.

Although traditional performance metrics such as the F_1 score may appear stable across different seeds, they can obscure significant variability at the instance level. Previous work has shown that models with nearly identical overall performance can nevertheless produce divergent predictions on individual tokens or examples (Khurana et al., 2021). To capture this phenomenon, I calculate agreement-based metrics such as Krippendorff's α and pairwise overlap, offering a more fine-grained perspective on prediction stability.

Agreement Across Seeds

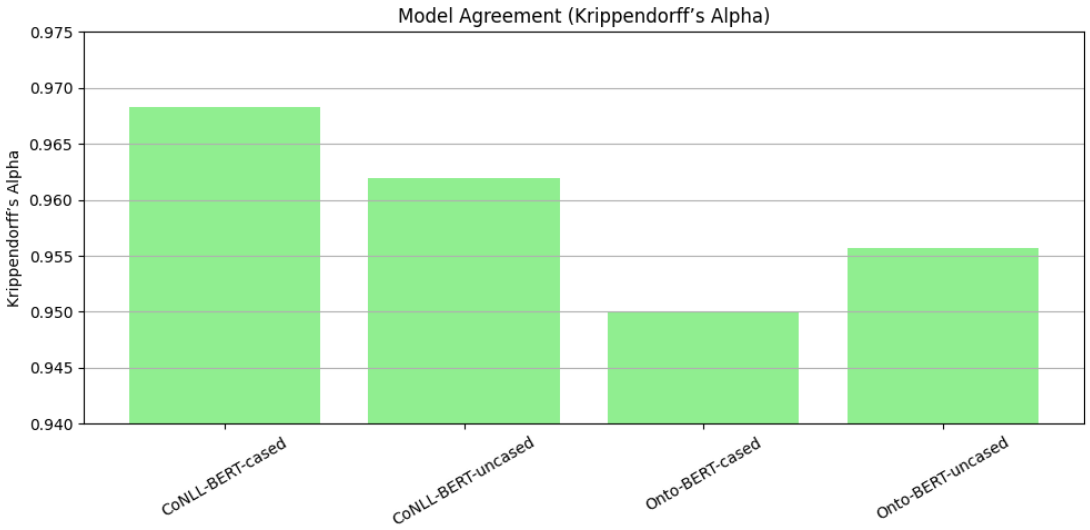

Figure 5.12: Model Agreement Across Seeds (Krippendorff's Alpha)

I quantify consistency using two standard agreement metrics: Cohen's kappa and Krippendorff's alpha. Cohen's kappa is computed in pairs between seed variants and reflects the agreement corrected for chance, while Krippendorff's alpha aggregates the agreement between the five runs and accommodates the nominal label data.

To allow for a direct comparison, I calculated Cohen's kappa for all possible model pairings and averaged the resulting values. These mean kappa scores were found to be closely aligned with the corresponding Krippendorff alpha values, reinforcing the overall stability of the predictions across seeds.

As shown in Figure 5.12, all models show high Krippendorff alpha values ($\alpha > 0.94$), indicating that predictions are broadly stable in seeds. The CoNLL-trained cased model achieves the highest agreement ($\alpha = 0.968$), closely followed by the CoNLL-uncased model ($\alpha = 0.962$). In contrast, models trained on OntoNotes show slightly reduced agreement, particularly the cased version ($\alpha = 0.950$). This would indicate that CoNLL-trained models, especially those using case information, produce slightly more consistent outputs under stochastic variation.

These findings reinforce the reliability of the results presented in this study. Although slight variability is present, especially in OntoNotes-trained models, the overall agreement between seeds is sufficiently high to support the use of single seed runs in downstream analyses.

5.3.3 Stanford Challenge Set Evaluation

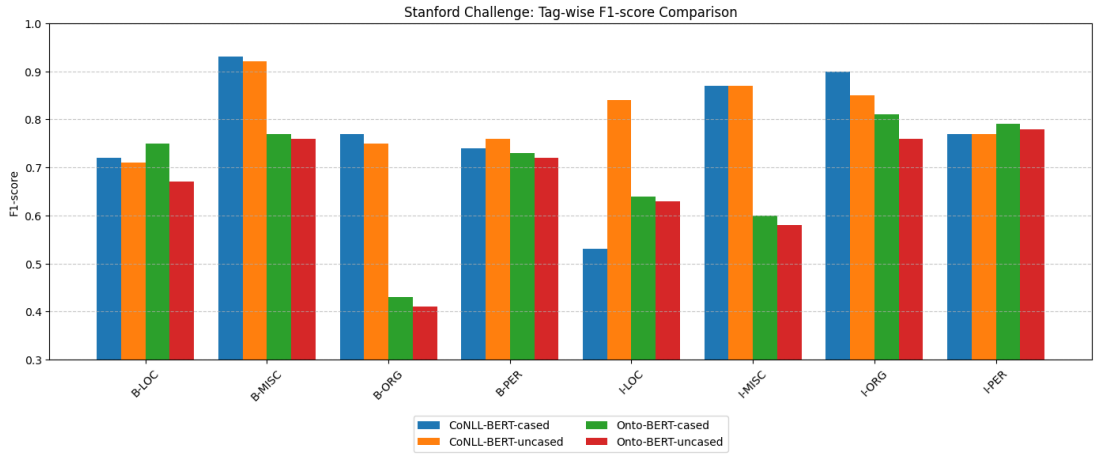

Figure 5.13: Stanford Challenge: Tag-wise F1-score Comparison

To evaluate the robustness of NER models under adversarial and contextually altered input, I tested the four BERT configurations on the Stanford Challenge Set (SALT_NLP, 2023). This dataset comprises carefully crafted perturbations applied to CoNLL-2003 examples, such as appending tokens that change the intended entity type (e.g., "Brazil" → "Brazil University") or modifying the surrounding context to introduce distractors or semantic ambiguity. These modifications are designed to test whether models rely on shallow heuristics (e.g., surface form or token identity) or genuinely capture deeper contextual dependencies.

Figure 5.13 presents tag-wise F1 scores across the four model variants. All models exhibit performance drops relative to in-domain evaluation, confirming that these perturbations challenge model generalisation. However, for most tags, the magnitude of the drop is relatively consistent across perturbation types. This indicates that the challenge set does not disproportionately affect any one category of transformation, suggesting that no single perturbation strategy is universally harder than others. One notable exception is the *I-LOC* tag, where performance is markedly lower: 0.53 for the CoNLL-cased model and 0.64 or below for all others. This likely reflects increased vulnerability to local boundary confusion when longer multi-token entities are disrupted or syntactically shifted.

The *CoNLL-trained cased model* (blue) retains relatively consistent performance across most tags, with high scores on 'B-MISC' (0.93), 'I-MISC' (0.87), and 'I-ORG' (0.90), despite challenging input. This shows a stable reliance on both contextual signals and surface-level cues. However, it still struggles with overlapping entity types, especially when multiple candidates appear nearby in perturbed contexts. The *uncased CoNLL model* (yellow) performs similarly in general, but shows slightly less precision on 'I-ORG' and 'B-ORG', suggesting casing plays a role in the disambiguation of overlapping or nested entities, particularly when surface form distinctions are subtle.

In contrast, the *OntoNotes-trained cased model* (green) shows a greater degradation of performance in several tags, especially on `B-ORG` (0.43), `I-MISC` (0.60), and `I-LOC` (0.64). Therefore, there is difficulty handling entity-type drift and restructuring that falls outside of the patterns seen during training. Given that the OntoNotes dataset spans a wider range of genres and syntactic constructions, one might expect it to produce more robust models. In fact, in standard in-domain and cross-domain evaluations (see Figure 5.9 and Figure 5.10), OntoNotes-trained models often outperform CoNLL-trained ones. However, more diffuse annotation patterns and contextual diversity in OntoNotes may lead to less consistent exposure to tightly coupled form-function mappings, making the model more vulnerable to adversarial perturbations that exploit such inconsistencies. The *uncased OntoNotes model* (red) performs slightly more consistently than its cased counterpart but remains vulnerable to semantic shifts and contextual ambiguity. In particular, its lower score on `B-ORG` (0.41) suggests difficulty generalising organisation types when surface cues are weak or misleading.

Taken together, these results demonstrate that while no single type of perturbation dominates the performance drop, certain entity types, particularly 'I-LOC', are disproportionately affected across models. The Stanford Challenge Set thus offers valuable diagnostic insight: It not only reveals a general degradation in accuracy under adversarial input, but also highlights specific weaknesses in boundary detection, entity-type resolution, and reliance on orthographic or syntactic signals.

5.4 Constructing Challenge Sets

5.4.1 Ethnic Name Challenge Sets

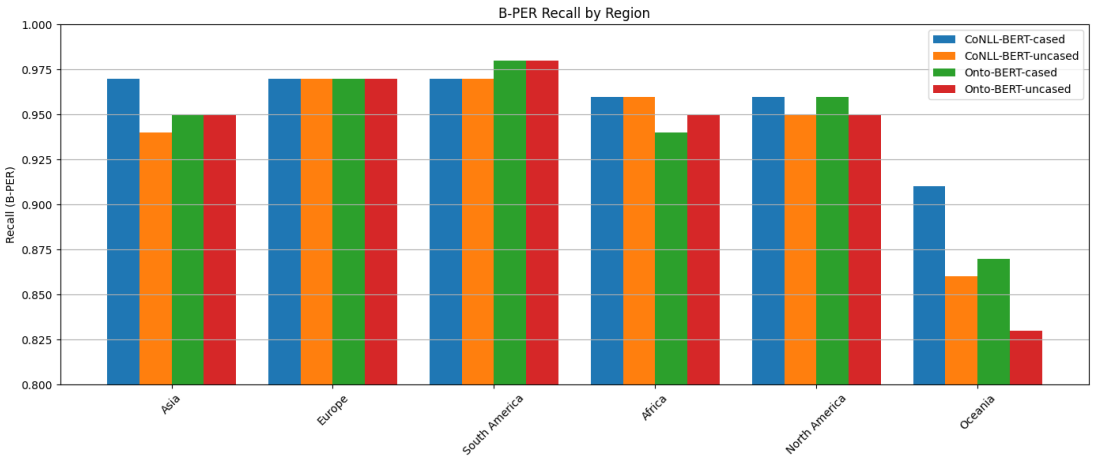

Figure 5.14: B-PER Recall by Region

To assess generalisation beyond training data, I evaluated model performance on the regionally diverse challenge sets described in Section 4.5. These sets were designed to test the model's ability to recognise names from underrepresented ethnic groups embedded in a variety of syntactic contexts.

Key Findings As shown in Figure 5.14, models performed consistently well on names from several regions, with a particularly high recall of South American and Asian names. However, performance dropped for names originating from North America and Oceania, an unexpected result given that these regions are well represented in both CoNLL-2003 and OntoNotes 5.0 training data. This would demonstrate that the issue is not one of ethnic unfamiliarity, but potentially of lexical properties specific to names from these regions.

Unexpected Source of Error: Lexical Ambiguity Closer inspection revealed that many of the misclassified North American names (e.g., *Will*, *Mark*, *Hope*) overlapped with common English nouns or verbs. These names are prone to ambiguity in context, and models often fail to correctly label them despite appropriate syntax. This indicates that the models rely heavily on surface-level heuristics such as capitalisation or expected position, rather than resolving semantic ambiguity through contextual cues.

Implication These findings suggest that, while the models exhibit strong cross-regional generalisation overall, their performance degrades in the presence of ambiguous tokens, even when those names come from regions present in the training data. To further investigate this vulnerability, I constructed a set of targeted challenges that focusses specifically on names that also function as common nouns or verbs. This is discussed in the following section.

5.4.2 Common Noun Ambiguity

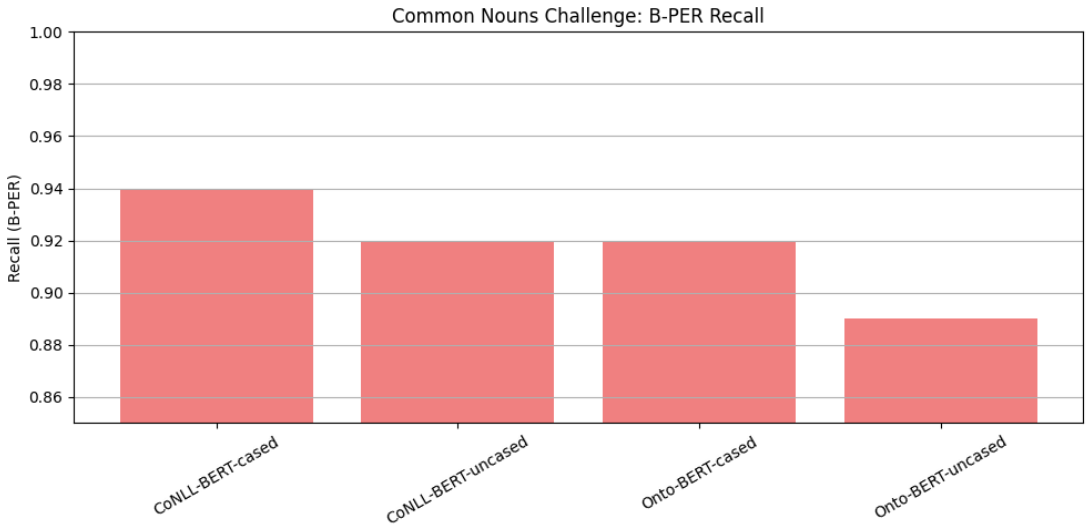

Figure 5.15: Common Nouns Challenge: B-PER Recall

Building on the observation that many misclassified names in the ethnic challenge set originated from North America and overlapped with common nouns (e.g., *Mark*, *Will*, *Hope*), I constructed a targeted challenge set to assess model performance under lexical ambiguity. Specifically, this evaluation focusses on names that also function as everyday English nouns, requiring models to rely on contextual understanding rather than superficial features such as capitalisation or token position. This set follows the same construction procedure and evaluation pattern as the Ethnicity Challenge Sets (see Section 5.2.3), but focusses specifically on names that also function as common English nouns or verbs. The syntactic templates and the sampling approach remain consistent, allowing a direct comparison of performance under different sources of ambiguity.

As shown in Figure 5.15, performance drops noticeably across all models compared to the ethnicity challenge. For direct comparison, the ethnicity challenge results are included in the same format. In that setting (see Figure 5.14), most of the models achieved above 95% recall for Asia, Europe, and South America, with lower scores for North America (see, e.g., 91% for CoNLL-cased) and Oceania (see, e.g., 83% for OntoNotes-uncased). The decline in the current task confirms the additional challenge posed by lexical ambiguity. Although in the common noun set, the CoNLL-cased model performs best at 94% recall, both OntoNotes-based models show a greater decline, particularly the uncased variant, which drops below 90%. This may reflect that OntoNotes-trained models may be more sensitive to lexical ambiguity and less robust when capitalisation cues are unavailable.

These results confirm my hypothesis that name/noun overlap presents a distinct challenge, particularly in contexts where models must distinguish entity types without relying on orthographic features. Such errors may contribute to the lower recall observed on North American names in the previous section and highlight the importance of contextual disambiguation in NER generalisation.

5.4.3 PER-ORG Confusion and Discourse Cues

In addition to the challenges posed by common noun ambiguity, a prominent source of model error emerged in the confusion between the `PER` and `ORG` entity types. To investigate this issue, an error analysis was conducted across all four model configurations. From the complete set of errors, 1,000 instances were randomly sampled, each consisting of the misclassified token, its gold and predicted tag, and the full sentence context. This allowed a qualitative examination of the kinds of ambiguity that contributed to systematic confusion.

In all models, confusion between `PER` and `ORG` was both frequent and bidirectional. The CoNLL-trained cased model yielded a total of 1,215 such errors, with many instances involving partial mislabeling of multi-token names (e.g., only the first token misclassified). The uncased variant produced 921 errors of this type, with similar patterns but slightly reduced frequency. The OntoNotes-trained cased model showed the highest total, 1,357 instances, driven primarily by a tendency to over-predict `PER` for organisational entities. The uncased OntoNotes model, although lower at 756, still displayed notable bidirectional confusion.

Many of these errors occurred in syntactically or semantically ambiguous contexts, where names such as “Kabariti” could plausibly refer to a person or an organisation. In simpler sentences, the absence of disambiguating cues led to unreliable tagging, whereas more complex sentences introduced long-range dependencies that the models

often failed to resolve. For example, in the sentence "Jordan's official state news agency Petra said Kabariti would hold discussions 'on the latest developments in the peace process and bilateral cooperation.' "The correct classification of "Kabariti" as `PER` depends on the resolution of the task, "would hold discussions" and understanding the broader structure of the discourse. Although some instances were labelled correctly when the supporting context was clear and local (e.g. coordinated mentions such as "CEO of Amazon"), such cases were less frequent, suggesting a general limitation in how current NER models integrate context beyond local lexical or positional cues.

5.4.4 Summary of Generalisation Challenges

Following the results of the challenge set experiments, two primary areas of weakness emerged as key targets for improving generalisation in these NER models:

- The **disambiguation of common nouns**, where tokens may function as both names and ordinary nouns or verbs.
- The issue of **PER-ORG disambiguation**, where tokens can denote either a person, an organisation, or in some cases even a location depending on context.

These findings reveal that improving generalisation in NER is not merely a matter of expanding coverage. Instead, it points to a deeper challenge: models must move beyond surface-level heuristics (e.g., capitalisation, token shape, or memorised entity lists) and develop a more robust understanding of context to correctly interpret ambiguous entities.

Common Noun Ambiguity In cases of lexical overlap between names and common nouns, model performance often deteriorates due to insufficient contextual grounding. This implies that existing NER models struggle to interpret syntactic roles or semantic expectations when capitalisation is unreliable or sentence structure is complex. Future solutions may include the integration of syntactic supervision, such as jointly training the model to also predict part-of-speech tags or syntactic dependencies. These auxiliary tasks can help the model better capture grammatical structure, making it more sensitive to whether a token is functioning as a name or a common noun based on its role in the sentence. Additional strategies could involve exposure to adversarial training examples with ambiguous tokens or embedding such cases more densely in the training corpora.

PER vs ORG Confusion The PER-ORG boundary remains a persistent challenge, especially for modern entities that blur traditional role distinctions (e.g. *'He works at Google'* frames Google as an organisation. In contrast, *'He was hired by Google'* can introduce ambiguity for the model, since the agent who performs the hiring could be interpreted as a company (organisation) or as an individual representing that company (person).) Standard NER systems often rely on shallow features or token context windows that are insufficient to resolve these differences. More accurate disambiguation may require models to incorporate document-level cues, such as entity co-reference chains, or access additional contextual knowledge about roles and relationships. Fine-tuning on datasets with richer inter-sentential context or integrating entity linking and coreference resolution modules may help address this limitation.

Chapter 6

Discussion

The results of this study confirm that BERT-based NER models, while highly effective under controlled conditions, exhibit notable weaknesses when faced with distributional shifts, syntactic variation, or ambiguous surface forms. The strong in-domain performance of CoNLL-trained models masks their vulnerability to overfitting on capitalisation patterns and genre-specific cues. In contrast, OntoNotes-trained models generalise better across genres but struggle with precise label boundaries and underrepresented name types.

The results of the challenge set underscore the value of targeted evaluation: Common noun ambiguity and PER-ORG confusion consistently triggered misclassification, particularly in cases where the local context was not sufficient. These failure modes reflect broader limitations in contextual disambiguation and discourse-level understanding in current transformer-based architectures.

Although this study provides a structured diagnostic framework for evaluating generalisation in NER, several limitations remain that open avenues for future research. First, the analysis was restricted to English and only two data sets, CoNLL-2003 and OntoNotes 5.0, both of which are news-centric and relatively formal in style. Extending the current methodology to multilingual NER settings would help uncover language-specific generalisation behaviours, particularly in morphologically rich or low-resource languages. Similarly, applying the evaluation framework to datasets with different annotation schemes, such as nested entities or fine-grained entity types, could test the adaptability of the models and the robustness of the analysis methods.

From a modelling perspective, this thesis focussed exclusively on BERT-based token classification models. Future work could examine whether alternative architectures, such as encoder-decoder models or prompt-based approaches (e.g., GPT-NER), exhibit similar vulnerabilities under perturbation. Since generation-based models may leverage global sentence-level representations differently, it would be particularly valuable to compare their generalisation patterns against the more local, token-based strategies employed by BERT.

In addition, there is considerable potential to expand the types of challenge set used. Although this study introduced controlled sets targeting ethnic name diversity and common noun ambiguity, future challenge sets could incorporate more aggressive or deceptive perturbations. For instance, it would be informative to construct a benchmark that tests whether CoNLL-trained models can be confused by random capitalisation of non-entity tokens, exploiting their overreliance on orthographic cues. Similarly, injecting noisy text, such as typos, keyboard errors, or context changes, could

help evaluate robustness under real-world user input conditions. These perturbations would approximate natural noise in social media, user-generated content, or noisy text.

Another promising direction involves probing the decision basis of the model. For example, analysing attribution patterns (e.g., using attention weights, gradient-based methods, or masking) could help identify which specific words or features the model attends to when assigning entity labels. This would reveal whether the model relies on semantically meaningful context or superficial anchors such as punctuation, affixes, or sentence position. Such findings could, in turn, inform the design of more effective adversarial examples or targeted training interventions.

Overall, expanding the evaluation framework in these directions would help to deepen understanding of NER generalisation and support the development of systems that are both more robust and more interpretable.

Overall, these findings suggest that improving generalisation in NER will require not just architectural changes, but also more representative datasets and principled evaluation strategies.

Chapter 7

Conclusion

This thesis examined the generalisation capacity of BERT-based models for Named Entity Recognition (NER), with a focus on performance across domain shifts, demographic variation, and controlled perturbations. I evaluated four BERT configurations: cased and uncased models trained on CoNLL 2003 and OntoNotes 5.0, and analysed their behaviour using both standard test sets and a series of targeted challenge sets.

The central research question guiding this work was: *Can a model trained in dataset X generalise effectively to dataset Y? To what extent do pre-trained NER models generalise beyond their training distributions, and what linguistic or dataset factors shape this generalisation?* To address this, I assessed the impact of dataset composition, casing, entity frequency, and capitalisation trends, and measured prediction stability across random seeds. I also introduced controlled evaluation sets targeting ethnicity, gender, and lexical ambiguity to isolate specific generalisation challenges.

Findings from these experiments show that models trained on OntoNotes tend to generalise better across domains due to broader genre and entity diversity. However, these same models can be more vulnerable to certain perturbations, such as common noun ambiguity, where inconsistent annotation patterns and genre-specific cues reduce robustness. CoNLL-trained models, by contrast, perform well on familiar surface patterns but generalise poorly to noisier or more diverse input. In both cases, overreliance on casing and positional heuristics was evident.

These findings underscore the limitations of current NER evaluation practices and highlight the need for fine-grained diagnostic benchmarks. The challenge sets introduced in this thesis provide a framework for testing models under realistic and systematically varied conditions. They also expose fairness concerns, particularly in the treatment of underrepresented name types and ambiguous tokens.

In a broader context, this work contributes to ongoing efforts to make NLP models more robust, equitable, and interpretable. As NER systems are increasingly deployed in socially sensitive applications such as content moderation, journalism, or legal analysis, ensuring that they perform reliably across different demographics and input types becomes critical. Future research can extend this work by exploring mitigation strategies such as multitask learning, adversarial training, or bias-aware pretraining objectives.

Ultimately, this thesis argues for a shift in how generalisation is evaluated in NER: not only by tracking what models get right under familiar conditions, but by systematically probing where, how, and why they fail when the context shifts.

Appendix A

Sentence Templates for Name Substitution

The following templates were used for evaluating the model's ability to generalise to proper names in diverse syntactic and discourse contexts. Each instance included either one or two substituted names.

- My name is {`name`}.
- This is {`name`}, our new team member.
- Have you met {`name`} from the accounting department?
- Everyone was waiting for {`name`} to arrive.
- Can you schedule a meeting with {`name`} next week?
- Do you remember when {`name`} joined the company?
- Isn't {`name`} presenting at the conference?
- What did {`name`} say about the proposal yesterday?
- {`name`}, the regional director, approved the budget.
- The person you spoke to earlier, {`name`}, will send the report.
- I met {`name`} and {`name2`} at the workshop in Jakarta.
- {`name`} said that {`name2`} might be late due to traffic.
- During yesterday's meeting, {`name`} brought up an interesting point.
- After the interview, we offered {`name`} the position.
- Although {`name`} was nervous, she performed well.
- At the annual summit in Kuala Lumpur, {`name`} gave a keynote speech.
- It was {`name`} who finalized the agreement.
- The report, as mentioned by {`name`} (our new data analyst), needs a revision.
- The final report was prepared by {`name`}.
- Nobody expected {`name`} to handle the crisis so well.

- The consultant who was hired by {`name`} recommended a new tool.
- The woman seen talking to {`name`} was the CEO.
- {`name`} and his colleague {`name2`} handled the presentation together.
- Both {`name`} and {`name2`} have extensive experience in logistics.
- {`name`} introduced {`name2`} as the lead consultant on the project.
- {`name`}, the new project manager, presented the budget.
- While the proposal was being finalized by the marketing team, {`name`}, who had just returned from Singapore, raised an objection.

Appendix B

Classification Reports

This appendix contains the complete classification reports for all evaluated model configurations. Each table reports precision, recall, F1-score, and support for individual entity tags, as well as macro-averaged and weighted metrics where applicable. Results are presented for in-domain evaluations, cross-domain generalisation, challenge sets (including demographic and lexical perturbations), and seed sensitivity analyses.

Label	**Precision**	**Recall**	**F1-score**	**Support**
B-LOC	0.79	0.90	0.84	8535
B-MISC	0.77	0.69	0.73	4062
B-ORG	0.73	0.52	0.61	7398
B-PER	0.92	0.93	0.93	7975
I-LOC	0.60	0.77	0.68	1356
I-MISC	0.42	0.69	0.52	1380
I-ORG	0.63	0.77	0.69	4251
I-PER	0.90	0.98	0.94	5503
O	0.99	0.98	0.99	201398
Accuracy			**0.95**	241858
Macro avg	0.75	0.80	0.77	241858
Weighted avg	0.95	0.95	0.95	241858

Table 1: Model: `bert-base-cased`, Train: `CoNLL-2003`, Test: `OntoNotes 5.0`

Label	Precision	Recall	F1-score	Support
B-LOC	0.56	0.82	0.67	539
B-MISC	0.77	0.75	0.76	319
B-ORG	0.40	0.42	0.41	1214
B-PER	0.80	0.66	0.72	1067
I-LOC	0.53	0.77	0.63	79
I-MISC	0.48	0.73	0.58	82
I-ORG	0.67	0.88	0.76	2026
I-PER	0.94	0.66	0.78	939
O	0.98	0.96	0.97	22466
Accuracy			**0.90**	28731
Macro avg	0.68	0.74	0.70	28731
Weighted avg	0.92	0.90	0.91	28731

Table 2: Model: `bert-base-cased`, Train: `CoNLL-2003`, Test: `Stanford Challenge Set`

Region	PER F1	I-PER F1	Accuracy	Macro F1	Support
Asia	0.96	1.00	0.98	0.42	10,884
Europe	0.97	—	0.98	0.39	10,818
Africa	0.96	1.00	0.98	0.49	10,854
South America	0.98	1.00	0.98	0.50	10,790
North America	0.97	0.99	0.98	0.42	10,789
Oceania	0.90	—	0.96	0.38	10,696

Table 3: Evaluation on name-origin challenge sets using a `bert-base-cased` model trained on CoNLL-2003. Performance on PER tags remains high, while other entity types are consistently missed, reflecting limitations in training diversity.

Label	Precision	Recall	F1-score	Support
B-PER	1.00	0.89	0.94	1168
O	1.00	1.00	1.00	9336
Accuracy			**0.99**	10504
Macro avg	0.40	0.38	0.39	10504
Weighted avg	1.00	0.99	0.99	10504

Table 4: Model: `bert-base-cased`, Train: `CoNLL-2003`, Test: `Common Noun Challenge Set`

Label	Precision	Recall	F1-score	Support
B-LOC	0.79	0.90	0.84	8535
B-MISC	0.77	0.69	0.73	4062
B-ORG	0.73	0.52	0.61	7398
B-PER	0.92	0.93	0.93	7975
I-LOC	0.60	0.77	0.68	1356
I-MISC	0.42	0.69	0.52	1380
I-ORG	0.63	0.77	0.69	4251
I-PER	0.90	0.98	0.94	5503
O	0.99	0.98	0.99	201398
Accuracy			**0.95**	241858
Macro avg	0.75	0.80	0.77	241858
Weighted avg	0.95	0.95	0.95	241858

Table 5: Model: `bert-base-uncased`, Train: `CoNLL-2003`, Test: `OntoNotes`

Label	Precision	Recall	F1-score	Support
B-LOC	0.56	0.82	0.67	539
B-MISC	0.77	0.75	0.76	319
B-ORG	0.40	0.42	0.41	1214
B-PER	0.80	0.66	0.72	1067
I-LOC	0.53	0.77	0.63	79
I-MISC	0.48	0.73	0.58	82
I-ORG	0.67	0.88	0.76	2026
I-PER	0.94	0.66	0.78	939
O	0.98	0.96	0.97	22466
Accuracy			**0.90**	28731
Macro avg	0.68	0.74	0.70	28731
Weighted avg	0.92	0.90	0.91	28731

Table 6: Model: `bert-base-uncased`, Train: `CoNLL-2003`, Test: `Stanford challenge set`

Region	Macro Precision	Macro Recall	Macro F1-score
Asia	0.43	0.42	0.42
Europe	0.40	0.39	0.39
Africa	0.50	0.49	0.49
South America	0.50	0.49	0.50
North America	0.42	0.42	0.42
Oceania	0.40	0.36	0.38
Average	**0.44**	**0.43**	**0.43**

Table 7: Model: `bert-base-uncased`, Train: CoNLL-2003, Test: Region-based Name Origin Challenge Set

Run	Macro Precision	Macro Recall	Macro F1-score
Run 1	0.74	0.80	0.76
Run 2	0.75	0.80	0.76
Run 3	0.75	0.80	0.77
Run 4	0.75	0.80	0.77
Run 5	0.75	0.80	0.77
Average	**0.75**	**0.80**	**0.77**

Table 8: Model: `bert-base-uncased`, Train: CoNLL-2003, Test: OntoNotes Seed Variability

Label	Precision	Recall	F1-score
B-LOC	0.80	0.80	0.80
B-MISC	0.76	0.68	0.71
B-ORG	0.75	0.50	0.60
B-PER	0.86	0.93	0.89
I-LOC	0.68	0.62	0.65
I-MISC	0.53	0.58	0.56
I-ORG	0.73	0.76	0.74
I-PER	0.89	0.99	0.94
O	0.98	0.99	0.99
Macro Avg	**0.78**	**0.76**	**0.77**
Weighted Avg	**0.95**	**0.95**	**0.95**

Table 9: Model: `bert-base-cased`, Train: OntoNotes, Test: CoNLL

Label	Precision	Recall	F1-score
B-LOC	0.69	0.83	0.75
B-MISC	0.78	0.76	0.77
B-ORG	0.43	0.43	0.43
B-PER	0.72	0.73	0.73
I-LOC	0.57	0.72	0.64
I-MISC	0.53	0.70	0.60
I-ORG	0.72	0.94	0.81
I-PER	0.90	0.70	0.79
O	0.99	0.96	0.97
Macro Avg	**0.70**	**0.75**	**0.72**
Weighted Avg	**0.92**	**0.91**	**0.92**

Table 10: Model: `bert-base-cased`, Train: OntoNotes, Test: Stanford

Region	Precision	Recall	F1-score
Asia	0.35	0.42	0.37
Africa	0.43	0.42	0.42
South America	0.50	0.49	0.50
North America	0.37	0.37	0.37
Europe	0.33	0.32	0.33
Oceania	0.33	0.31	0.32
Macro Avg	**0.39**	**0.39**	**0.39**

Table 11: Performance on Ethnic Name Challenge Sets. Model: `bert-base-cased`, Train: OntoNotes, Test: Challenge set

Entity Type	Precision	Recall	F1-score
B-PER	1.00	0.92	0.96
O	1.00	1.00	1.00
Macro Avg	**0.50**	**0.48**	**0.49**

Table 12: Performance on Common Noun Ambiguity Challenge Set. Model: `bert-base-cased`, Train: OntoNotes, Test: Common noun challenge

Seed	Precision	Recall	F1-score
Seed 1	0.78	0.76	0.77
Seed 2	0.78	0.76	0.77
Seed 3	0.78	0.76	0.77
Seed 4	0.78	0.75	0.77
Seed 5	0.78	0.76	0.77
Mean	**0.78**	**0.76**	**0.77**

Table 13: Seed variability results (macro average). Model: `bert-base-cased`, Train: OntoNotes, Test: CoNLL

Label	Precision	Recall	F1-score
B-LOC	0.82	0.89	0.85
B-MISC	0.64	0.69	0.66
B-ORG	0.61	0.63	0.62
B-PER	0.89	0.90	0.90
I-LOC	0.68	0.56	0.62
I-MISC	0.61	0.23	0.34
I-ORG	0.86	0.61	0.72
I-PER	0.93	0.87	0.90
O	0.98	0.99	0.99
Macro Avg	**0.78**	**0.71**	**0.73**
Weighted Avg	**0.97**	**0.97**	**0.97**

Table 14: Performance on OntoNotes (Test), Model: `bert-base-uncased`, Trained on: CoNLL

Label	Precision	Recall	F1-score
B-LOC	0.63	0.82	0.71
B-MISC	0.91	0.94	0.92
B-ORG	0.68	0.83	0.75
B-PER	0.89	0.66	0.76
I-LOC	0.83	0.86	0.84
I-MISC	0.86	0.88	0.87
I-ORG	0.81	0.90	0.85
I-PER	1.00	0.62	0.77
O	0.99	0.99	0.99
Macro Avg	**0.84**	**0.83**	**0.83**
Weighted Avg	**0.95**	**0.95**	**0.95**

Table 15: Performance on Stanford Challenge Set, Model: `bert-base-uncased`, Train: CoNLL

Region	Precision	Recall	F1-score
Asia	1.00	0.98	0.99
Europe	1.00	0.98	0.99
South America	1.00	0.98	0.99
Africa	1.00	0.98	0.99
North America	1.00	0.98	0.99
Oceania	1.00	0.97	0.98
Average	**1.00**	**0.98**	**0.99**

Table 16: NER performance on region-specific ethnic name challenge sets. Model: `bert-base-cased`, Trained on: CoNLL

Label	Precision	Recall	F1-score
B-LOC	0.00	0.00	0.00
B-ORG	0.00	0.00	0.00
B-PER	1.00	0.92	0.96
O	1.00	1.00	1.00
Macro avg	**0.50**	**0.48**	**0.49**
Weighted avg	**1.00**	**0.99**	**0.99**

Table 17: NER performance on the Common Noun Ambiguity Challenge Set. Model: `bert-base-cased`, Train: CoNLL

Label	Precision	Recall	F1-score
B-LOC	0.82	0.89	0.85
B-MISC	0.63	0.70	0.66
B-ORG	0.61	0.63	0.62
B-PER	0.90	0.90	0.90
I-LOC	0.68	0.57	0.62
I-MISC	0.60	0.24	0.34
I-ORG	0.87	0.61	0.72
I-PER	0.93	0.87	0.90
O	0.98	0.99	0.99
Macro avg	**0.78**	**0.71**	**0.73**
Weighted avg	**0.97**	**0.97**	**0.97**

Table 18: NER performance across five random seeds. Model: `bert-base-uncased`, Train: CoNLL, Test: OntoNotes

Label	Precision	Recall	F1-score	Support
B-LOC	0.80	0.80	0.80	8535
B-MISC	0.76	0.68	0.71	4062
B-ORG	0.75	0.50	0.60	7398
B-PER	0.86	0.93	0.89	7975
I-LOC	0.68	0.62	0.65	1356
I-MISC	0.53	0.58	0.56	1380
I-ORG	0.73	0.76	0.74	4251
I-PER	0.89	0.99	0.94	5503
O	0.98	0.99	0.99	201398
Accuracy			0.95	241858
Macro Avg	0.78	0.76	0.77	241858
Weighted Avg	0.95	0.95	0.95	241858

Table 19: Model: `bert-base-cased`, Train: OntoNotes, Test: CoNLL (Capitalised Tokens Only)

Label	Precision	Recall	F1-score	Support
B-LOC	0.82	0.89	0.85	17495
B-MISC	0.64	0.69	0.66	10657
B-ORG	0.61	0.63	0.62	13041
B-PER	0.89	0.90	0.90	15547
I-LOC	0.68	0.56	0.62	5367
I-MISC	0.61	0.23	0.34	7305
I-ORG	0.86	0.61	0.72	18313
I-PER	0.93	0.87	0.90	11086
O	0.98	0.99	0.99	1011383
Accuracy			0.97	1110194
Macro Avg	0.78	0.71	0.73	1110194
Weighted Avg	0.97	0.97	0.97	1110194

Table 20: Model: `bert-base-cased`, Train: CoNLL, Test: OntoNotes (Capitalised Tokens Only)

Label	Precision	Recall	F1-score	Support
B-LOC	0.96	0.97	0.96	2110
B-MISC	0.92	0.91	0.92	1000
B-ORG	0.94	0.95	0.95	1925
B-PER	0.98	0.98	0.98	2084
I-LOC	0.91	0.90	0.91	315
I-MISC	0.86	0.85	0.86	337
I-ORG	0.93	0.96	0.94	1039
I-PER	1.00	0.99	0.99	1488
O	1.00	1.00	1.00	49262
Accuracy			0.99	59560
Macro Avg	0.95	0.95	0.95	59560
Weighted Avg	0.99	0.99	0.99	59560

Table 21: Model: `bert-base-cased`, Train: CoNLL, Test: CoNLL (Capitalised Tokens Only)

Label	Precision	Recall	F1-score	Support
B-LOC	0.97	0.97	0.97	4315
B-MISC	0.91	0.91	0.91	2722
B-ORG	0.92	0.94	0.93	3314
B-PER	0.96	0.97	0.96	3890
I-LOC	0.93	0.92	0.93	1258
I-MISC	0.87	0.84	0.85	2067
I-ORG	0.94	0.96	0.95	4675
I-PER	0.97	0.97	0.97	2868
O	1.00	1.00	1.00	253580
Accuracy			1.00	276689
Macro Avg	0.95	0.94	0.94	276689
Weighted Avg	1.00	1.00	1.00	276689

Table 22: Model: `bert-base-cased`, Train: OntoNotes, Test: OntoNotes (Capitalized Tokens Only)

Label	Precision	Recall	F1-score	Support
B-LOC	0.82	0.90	0.86	16435
B-MISC	0.50	0.68	0.58	7455
B-ORG	0.74	0.73	0.73	22446
B-PER	0.89	0.88	0.88	12641
I-LOC	0.68	0.60	0.64	5154
I-MISC	0.52	0.27	0.36	5579
I-ORG	0.89	0.72	0.80	30834
I-PER	0.84	0.92	0.88	7761
O	0.99	0.99	0.99	965876
Accuracy			0.97	1074181
Macro Avg	0.76	0.74	0.75	1074181
Weighted Avg	0.97	0.97	0.97	1074181

Table 23: Model: `bert-base-uncased`, Train: CoNLL, Test: OntoNotes Newswire

Label	Precision	Recall	F1-score	Support
B-LOC	0.58	0.88	0.69	14389
B-MISC	0.50	0.70	0.58	10792
B-ORG	0.39	0.55	0.46	7517
B-PER	0.43	0.90	0.58	14691
I-LOC	0.53	0.59	0.56	4350
I-MISC	0.48	0.22	0.31	6920
I-ORG	0.68	0.51	0.58	9826
I-PER	0.84	0.85	0.84	11300
O	0.99	0.98	0.98	1568935
Accuracy			0.96	1648720
Macro Avg	0.60	0.69	0.62	1648720
Weighted Avg	0.97	0.96	0.97	1648720

Table 24: Model: `bert-base-uncased`, Train: CoNLL, Test: OntoNotes (Non-News Genres)

Label	Precision	Recall	F1-score	Support
B-LOC	0.83	0.91	0.87	16435
B-MISC	0.52	0.73	0.60	7455
B-ORG	0.72	0.76	0.74	22446
B-PER	0.91	0.94	0.92	12641
I-LOC	0.71	0.68	0.69	5154
I-MISC	0.55	0.43	0.48	5579
I-ORG	0.92	0.77	0.84	30834
I-PER	0.91	0.92	0.92	7761
O	0.99	0.99	0.99	965876
Accuracy			0.97	1074181
Macro Avg	0.78	0.79	0.78	1074181
Weighted Avg	0.97	0.97	0.97	1074181

Table 25: Model: `bert-base-cased`, Train: CoNLL, Test: OntoNotes (Newswire Genre)

Label	Precision	Recall	F1-score	Support
B-LOC	0.58	0.88	0.70	14389
B-MISC	0.48	0.72	0.58	10792
B-ORG	0.42	0.61	0.50	7517
B-PER	0.45	0.90	0.60	14691
I-LOC	0.54	0.66	0.59	4350
I-MISC	0.44	0.39	0.41	6920
I-ORG	0.70	0.62	0.66	9826
I-PER	0.85	0.85	0.85	11300
O	0.99	0.98	0.98	1568935
Accuracy			0.96	1648720
Macro Avg	0.61	0.73	0.65	1648720
Weighted Avg	0.97	0.96	0.97	1648720

Table 26: Model: `bert-base-cased`, Train: CoNLL, Test: OntoNotes (Non-News Genres)

Genre	Total	O Tokens	NE Tokens	O Rate (%)	NE Rate (%)
nw	153,291	50,702	102,589	33.08	66.92
pt	51,601	51,601	0	100.00	0.00
mz	23,866	8,348	15,518	34.98	65.02
wb	51,789	39,990	11,799	77.22	22.78
bn	41,325	14,008	27,317	33.90	66.10
bc	31,383	16,678	14,705	53.14	46.86
tc	13,061	10,940	2,121	83.76	16.24

Table 27: Distribution of capitalised tokens across OntoNotes genres

Label	Capitalised Count	Total Count	Capitalised Rate (%)
B-LOC	8,529	8,535	99.93
B-MISC	3,998	4,062	98.42
B-ORG	7,373	7,398	99.66
B-PER	7,963	7,975	99.85
I-LOC	1,300	1,356	95.87
I-MISC	1,217	1,380	88.19
I-ORG	3,820	4,251	89.86
I-PER	5,413	5,503	98.36
O	13,867	201,398	6.89

Table 28: Capitalised token counts and proportions by label in the CoNLL-2003 dataset

References

A. Ambekar, C. Ward, J. Mohammed, S. Male, and S. Skiena. Name-ethnicity classification from open sources. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 49–58, 06 2009. doi: 10.1145/1557019.1557032.

I. Augenstein, L. Derczynski, and K. Bontcheva. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech and Language*, 44, 01 2017. doi: 10.1016/j.csl.2017.01.012.

M. B. gender-guesser: Guess gender from first names using data from the us census and other sources. `https://pypi.org/project/gender-guesser/`, 2022. Accessed: 2025-07-08.

Y. Belinkov and Y. Bisk. Synthetic and natural noise both break neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018. URL `https://arxiv.org/abs/1711.02173`.

E. M. Bender and B. Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL `https://aclanthology.org/Q18-1041/`.

J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423/`.

S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy. A survey of data augmentation approaches for NLP. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.84. URL `https://aclanthology.org/2021.findings-acl.84/`.

J. Fu and P. Liu. Rethinking generalization of neural models: A named entity recognition case study. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34: 7732–7739, 04 2020. doi: 10.1609/aaai.v34i05.6276.

M. Glockner, V. Shwartz, and Y. Goldberg. Breaking nli systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-2103. URL `https://aclanthology.org/P18-2103/`.

P. Graham. names-dataset: A dataset of first and last names with associated countries. `https://pypi.org/project/names-dataset/`, 2018. Accessed: 2025-06-17.

E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA, June 2006. Association for Computational Linguistics. URL `https://aclanthology.org/N06-2015`.

D. Hupkes, V. Dankers, M. Mul, and E. Bruni. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795, 2020.

D. Hupkes, V. Dankers, K. Batsuren, K. Sinha, A. Kazemnejad, C. Christodoulopoulos, R. Cotterell, and E. Bruni, editors. *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, Singapore, Dec. 2023a. Association for Computational Linguistics. URL `https://aclanthology.org/2023.genbench-1.0/`.

D. Hupkes, M. Giulianelli, V. Dankers, M. Artetxe, Y. Elazar, T. Pimentel, C. Christodoulopoulos, K. Lasri, N. Saphra, A. Sinclair, D. Ulmer, F. Schottmann, K. Batsuren, K. Sun, K. Sinha, L. Khalatbari, M. Ryskina, R. Frieske, R. Cotterell, and Z. Jin. State-of-the-art generalisation research in nlp: A taxonomy and review. Working Paper 2210.03050, arXiv, Jan. 2023b. Preprint; published as Analysis in *Nature Machine Intelligence*.

U. Khurana, E. Nalisnick, and A. Fokkens. How emotionally stable is ALBERT? testing robustness with stochastic weight averaging on a sentiment analysis task. In Y. Gao, S. Eger, W. Zhao, P. Lertvittayakumjorn, and M. Fomicheva, editors, *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 16–31, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eval4nlp-1.3. URL `https://aclanthology.org/2021.eval4nlp-1.3/`.

K. Krippendorff. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, Thousand Oaks, CA, 2 edition, 2004.

B. M. Lake and M. Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2873–2882, 2018.

X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li. Dice loss for data-imbalanced NLP tasks. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.45. URL https://aclanthology.org/2020.acl-main.45/.

H. Lin, Y. Lu, J. Tang, X. Han, L. Sun, Z. Wei, and N. J. Yuan. A rigorous study on named entity recognition: Can fine-tuning pretrained model lead to the promised land? In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7291–7300, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.592. URL https://aclanthology.org/2020.emnlp-main.592/.

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993. URL https://aclanthology.org/J93-2004/.

R. T. McCoy, E. Pavlick, and T. Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1334.

R. T. McCoy, J. Min, and T. Linzen. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In A. Alishahi, Y. Belinkov, G. Chrupała, D. Hupkes, Y. Pinter, and H. Sajjad, editors, *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.21. URL https://aclanthology.org/2020.blackboxnlp-1.21/.

J. Miller, K. Krauth, B. Recht, and L. Schmidt. The effect of natural distribution shift on question answering models. In *The Effect of Natural Distribution Shift on Question Answering Models*, 2020. URL https://arxiv.org/abs/2004.14444.

S. A. Mishra, Y. He, and L. Belli. Assessing demographic bias in named entity recognition. *arXiv preprint arXiv:2008.03415*, 2020. URL https://arxiv.org/abs/2008.03415.

J. Olive, C. Christianson, and J. McCary, editors. *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer, New York, NY, 2011. doi: 10.1007/978-1-4419-7713-7.

Y. Qu, D. Shen, Y. Shen, S. Sajeev, J. Han, and W. Chen. Coda: Contrast-enhanced and diversity-promoting data augmentation for natural language understanding. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021. OpenReview preprint, available at https://openreview.net/forum?id=oxwACn3eOi.

A. Reich, J. Chen, A. Agrawal, Y. Zhang, and D. Yang. Leveraging expert guided adversarial augmentation for improving generalization in named entity recognition, May 2022. URL https://aclanthology.org/2022.findings-acl.154/.

M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

SALT_NLP. Stanford challenge set for ner. https://github.com/SALT-NLP/Guided-Adversarial-Augmentation, 2023. Accessed: 2025-06-16.

T. Serafim and contributors. pycountry: Iso country, subdivision, language, currency and script definitions. https://github.com/flyingcircusio/pycountry, 2023. https://pypi.org/project/pycountry/.

E. F. Tjong Kim Sang and F. De Meulder. 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, 2003. URL https://www.aclweb.org/anthology/W03-0419.

P. Treeratpituk. Name-ethnicity classification and ethnicity-sensitive name matching. In *Name-Ethnicity Classification and Ethnicity-Sensitive Name Matching.*, 09 2018.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Attention is All you Need*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

T. Wang, X. Wang, Y. Qin, B. Packer, K. Li, J. Chen, A. Beutel, and E. Chi. CAT-gen: Improving robustness in NLP models via controlled adversarial text generation. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5141–5146, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.417. URL https://aclanthology.org/2020.emnlp-main.417/.

X. Wang, Y. Jiang, N. Bach, T. Wang, Z. Huang, F. Huang, and K. Tu. Automated concatenation of embeddings for structured prediction. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2643–2660, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.206. URL https://aclanthology.org/2021.acl-long.206/.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.emnlp-demos.6.

T. Wu, M. T. Ribeiro, J. Heer, and D. Weld. Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 747–763, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1073. URL `https://aclanthology.org/P19-1073`.

D. Yogatama, C. Dyer, W. Ling, and P. Blunsom. Learning and evaluating general linguistic intelligence. In *International Conference on Learning Representations (ICLR)*, 2019. URL `https://arxiv.org/abs/1901.11373`.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In M. Walker, H. Ji, and A. Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL `https://aclanthology.org/N18-2003/`.