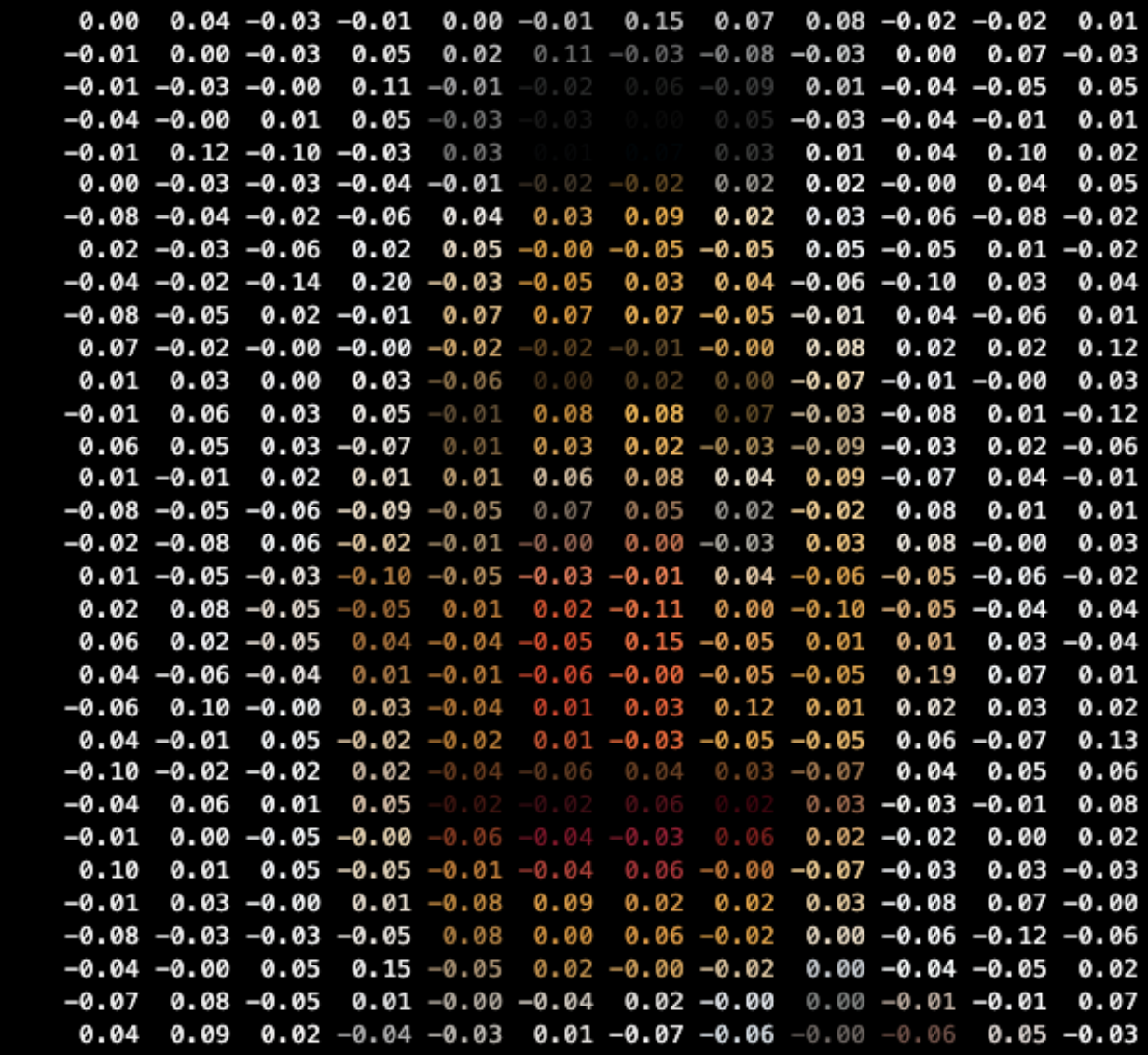
Synthetic WorgBERT for FgNER on OpenGov Documents



P.C. (Niels) Barnhoorn

Layout: typeset by the author using LaTeX.
Cover illustration: Vectorisation of the word "WorgBERT" colored on Bert from Sesame Street

Synthetic WorgBERT for FgNER on OpenGov Documents

Utilizing LLMs for Synthetic Data Creation in Low Resource Environments

P.C. (Niels) Barnhoorn
12855928

Bachelor thesis
Credits: 18 EC

Bachelor *Informatiekunde*

University of Amsterdam
Faculty of Science
Science Park 900
1098 XH Amsterdam

Supervisor
Dr. D. Graus

Informatics Institute
Faculty of Science
University of Amsterdam
Science Park 900
1098 XH Amsterdam

December 30, 2025

Abstract

Domain-specific named-entity recognition (NER) in parliamentary proceedings is obstructed by scarce annotated data. This thesis explores two complementary strategies to mitigate these challenges. First, we quantify how few-shot prompting with a large language model lets it generate high-quality synthetic sentences that read like real plenary speech. Automatic metrics (SBERT, BLEU, ROUGE) and qualitative analyses (t-SNE, readability, sentence-length profiles) show a steep improvement from 0- to 1-shot and continued, though diminishing, gains up to 25-shot. A 50-shot prompt attains the strongest similarity scores but also produces more formal, domain-specific phrasing, reflected in higher perplexity.

Second, we introduce WorgBERT, a Dutch Transformer fine-tuned for fine-grained organisational NER. On a balanced synthetic test set the model reaches 0.96 accuracy and the macro-F_1 = 0.73, confirming sufficient capacity to learn all sub-types. When ported to an gold standard corpus, macro-F_1 drops to 0.24. The standard RobBERT attains a macro-F1 score twice as high as WorgBERT, underscoring how hard it is to detect organisations in parliamentary transcripts.

Together, these results demonstrate that few-shot generation can generate authentic looking parliamentary language and that the WorgBERT solution is a viable fine-grained NER model provided additional annotations are collected. The work offers practical guidelines for scaling domain-specific NER: leverage moderate few-shot prompts to enrich training data and prioritise targeted annotation of the rarest labels to unlock the full potential of fine-grained models.

Contents

1 Introduction

Since 2022 the Dutch government has fully embraced the Wet open overheid (*Woo*, or Open Government Act), retiring the older Wet openbaarheid van bestuur (*Wob*) (Overheid, 2025). The ambition goes well beyond just a legal update, by proactively disclosing governmental information, the government aims to strengthen democratic oversight and invite citizens, journalists, and researchers to engage with the state's day-to-day decision-making.

Yet that transparency comes with a twist. Every ministry, municipality, and executive agency now uploads a combined total of more than 500.000 documents per year. So many in fact, that the yearly page count is growing exponentially. An example of this exponential growth of pages publications, are the minutes by the *Staten-Generaal*. Figure 1 illustrates the exponential growth of parliamentary minutes published since 1815. Transparency, is no longer just about making information public. It is about helping society navigate an ever-growing sea of PDFs to discover what truly matters.

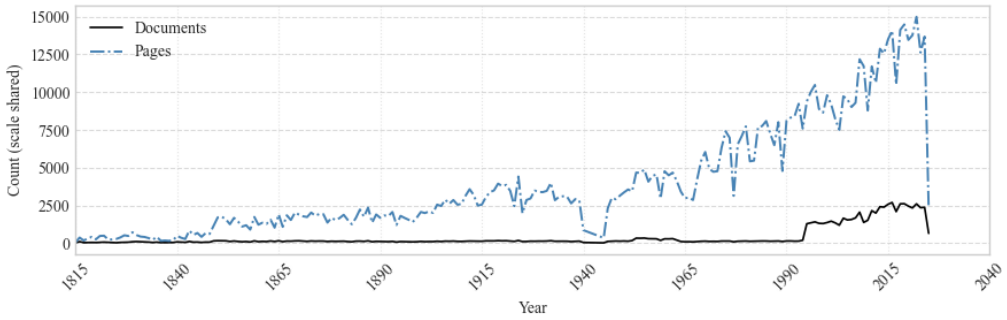


Figure 1: Minute Documents vs. Papers Uploaded per Year Since Start of United Kingdom of the Netherlands

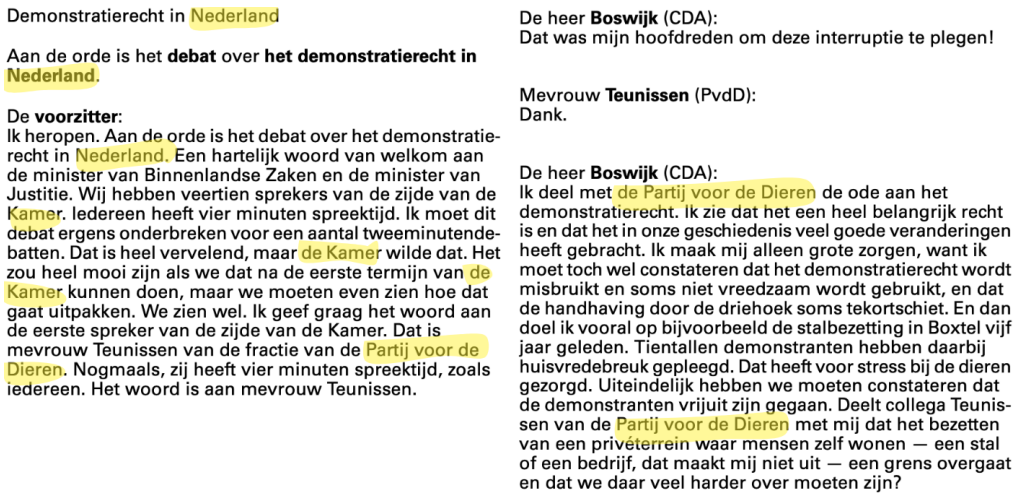
The OpenGov Lab, part of the Innovation Center for Artificial Intelligence (ICAI), is a collaboration between the University of Amsterdam (UvA) and the National Organization for Information Management (RvIHH)[1]. The lab's mission follows the philosophy: "to improve and support interpretation, retrieval, and use of open government data, to increase government transparency, public trust, and ultimately democratic participation." ("ICAI", n.d.).

The dynamic nature of the political landscape introduces significant variety within these documents, particularly evident in the Minutes of the Staten-Generaal. These are transcriptions of Dutch parliamentary meeting. It captures the dialogue involving the Chair of the Staten Generaal and other speakers from the chamber. Crucially, each spoken contribution is attributed to the specific speaker by name and their political party affiliation.

As official transcribed records of meetings, the inherent structure often complicates research. This difficulty is compounded by the specific terminology and jargon

[1]https://www.icai.ai/labs/icai-opengov-lab

employed by various politicians and speakers. For an example illustrating the typical language and structure found in these records, see Figure 2. Furthermore, the Minutes of the Staten-Generaal frequently contain references to a diverse array of organizations, including political parties, businesses, NGOs, and governmental entities. These are not always mentioned by name but most of the time mentioned by their abbreviation.



Demonstratierecht in Nederland

Aan de orde is het **debat** over **het demonstratierecht in Nederland**.

De **voorzitter**:
Ik heropen. Aan de orde is het debat over het demonstratierecht in Nederland. Een hartelijk woord van welkom aan de minister van Binnenlandse Zaken en de minister van Justitie. Wij hebben veertien sprekers van de zijde van de Kamer. Iedereen heeft vier minuten spreektijd. Ik moet dit debat ergens onderbreken voor een aantal tweeminutendebatten. Dat is heel vervelend, maar de Kamer wilde dat. Het zou heel mooi zijn als we dat na de eerste termijn van de Kamer kunnen doen, maar we moeten even zien hoe dat gaat uitpakken. We zien wel. Ik geef graag het woord aan de eerste spreker van de zijde van de Kamer. Dat is mevrouw Teunissen van de fractie van de Partij voor de Dieren. Nogmaals, zij heeft vier minuten spreektijd, zoals iedereen. Het woord is aan mevrouw Teunissen.

De heer **Boswijk** (CDA):
Dat was mijn hoofdreden om deze interruptie te plegen!

Mevrouw **Teunissen** (PvdD):
Dank.

De heer **Boswijk** (CDA):
Ik deel met de Partij voor de Dieren de ode aan het demonstratierecht. Ik zie dat het een heel belangrijk recht is en dat het in onze geschiedenis veel goede veranderingen heeft gebracht. Ik maak mij alleen grote zorgen, want ik moet toch wel constateren dat het demonstratierecht wordt misbruikt en soms niet vreedzaam wordt gebruikt, en dat de handhaving door de driehoek soms tekortschiet. En dan doel ik vooral op bijvoorbeeld de stalbezetting in Boxtel vijf jaar geleden. Tientallen demonstranten hebben daarbij huisvredebreuk gepleegd. Dat heeft voor stress bij de dieren gezorgd. Uiteindelijk hebben we moeten constateren dat de demonstranten vrijuit zijn gegaan. Deelt collega Teunissen van de Partij voor de Dieren met mij dat het bezetten van een privéterrein waar mensen zelf wonen — een stal of een bedrijf, dat maakt mij niet uit — een grens overgaat en dat we daar veel harder over moeten zijn?

Figure 2: Example of How Original Minutes Look Like

To enhance the ability to perform research on these documents Named Entity Recognition (NER), a Natural Language Processing (NLP) task, can be a solution. This NLP task focuses on identifying and categorizing specific pieces of information within texts. The goal is to locate references to predefined categories, and assign them the correct label. It is considered a fundamental step in information extraction (Vajjala & Balasubramaniam, 2022). This makes NER really useful for making the data more insightful (Jehangir et al., 2023).

A significant weakness of many state-of-the-art Named Entity Recognition (NER) models, is their heavy reliance on large amounts of high-quality labeled training data. Creating such datasets is often time-consuming and expensive, presenting a major bottleneck, especially for specialized domains or less-resourced languages (MacLean & Cavallucci, 2024). When only limited labeled data is available a "low-resource" or "few-sample" scenario occurs (Jehangir et al., 2023), this makes processes like the training for models like BERT unstable. This instability in a few sample settings makes it difficult to reliably apply powerful NER techniques when labeled data is scarce (Zhang et al., 2021).

The recent wave of generative AI such as ChatGPT, Gemini and Claude, presents new possibilities to combat the low-resource data in NER (Brown et al., 2020; Santoso et al., 2024). These models are often built on large amounts of data and possess the capability to perform NER tasks (Kuzman & Ljubešić, 2025). Often, however, deploying these techniques directly can be very cost-inefficient and computationally demanding, particularly in settings with limited resources (Santoso et al., 2024).

As a model such as GPT-3 (175B parameter) is roughly 1600 times larger than BERT-Base (110 M parameters).

Therefore another approach is proposed in this thesis: the use of LLMs for reliable data creation with In-Context Learning (ICL) (Dong et al., 2024). In Brown et al. (2020), ICL is described as the capability of large language models to learn to perform a new task simply by receiving a few examples or demonstrations within the input prompt itself. Unlike traditional training the model adapts its behavior for the specific task based on the context provided during inference.

A significant hurdle for analyzing Dutch governmental documents is the absence of a fine-tuned BERT model for this domain. While fine-tuning could create such a model, it requires substantial accurately labeled data. Given the vast volume of governmental texts, manual annotation, even for a small fraction, is often impractical due to resource intensity (MacLean & Cavallucci, 2024).

Further complicating the task are the specific needs of the governmental context. Standard NER categories (person, location, organization) are insufficient; domain-specific labels like governmental entity, political party, NGO, etc., are required (see Section 2). Additionally, the dynamic nature of governmental language, influenced by political events and changing discourse, adds another layer of complexity.

Scientifically, this research investigates the potential of Large Language Models (LLMs) to generate high-quality, domain-specific synthetic data to minutes of the "Staten-Generaal". A core part of the contribution involves evaluating the effectiveness of this LLM-generated data, specifically its utility in fine-tuning downstream models like BERT, thereby offering insights into data augmentation and model adaptation methodologies (Bogdanov et al., 2024; Zhang et al., 2021).

Addressing these technical and scientific challenges is crucial not only for advancing NLP research but also for achieving societal goals. By making complex governmental information more accessible and interpretable through improved NER, this work aims to enhance government transparency and citizen engagement, aligning with the objectives of initiatives like the Open Government Act (Attard et al., 2015; Overheid, 2025).

Having established these challenges and the relevance of addressing them, the central research questions guiding this study are:

RQ1: How well do Large Language Model perform in generating a representative synthetic dataset of NER compatible labeled Dutch governmental texts?

RQ2: How does the performance of a Dutch BERT model fine-tuned on synthetically generated dataset compare to that of a pre-trained (non-fine-tuned) Dutch BERT model?

2 Literature Review

2.1 Introduction

This chapter provides a comprehensive review of the literature relevant to leveraging Large Language Models (LLMs) for enhancing Named Entity Recognition (NER) in the specialized, low-resource context of Dutch governmental documents. It begins by establishing the fundamentals of NER, including its definition, common approaches, and evaluation metrics (Section 2.2). Subsequently, the review delves into the significant challenges posed by data scarcity in training robust NER models, particularly within unique domains like governmental text (Section 2.3). The potential of LLMs to mitigate these challenges is then explored (Section 2.4), leading into a discussion of pertinent learning paradigms (such as Zero-Shot, Few-Shot, and In-Context Learning), the growing field of LLM-driven synthetic data generation, and the contrast with traditional data-abundant methods like Many-Shot (Section 2.5). The importance of domain adaptation for applying these techniques effectively is also addressed within Section 2.5. Furthermore, relevant prior studies employing synthetic data for low-resource NER are analyzed to position the current research within the field (Section 2.6). Finally, this review culminates in a synthesis of the findings and a clear articulation of the specific research gap this thesis aims to address (Section 2.7).

2.2 Named Entity Recognition (NER)

2.2.1 Defining NER: Task and Importance

Named Entity Recognition (NER) is a fundamental task within Natural Language Processing (NLP) focused on automatically identifying and classifying specific pieces of information, known as "named entities," within unstructured text (Mohit, 2014). The primary goal of NER is to locate mentions of predefined categories – such as the names of persons, organizations, locations, dates, monetary values, percentages, and more – and assign them the correct category label (Vajjala & Balasubramaniam, 2022). By pinpointing these entities, NER systems transform raw text into structured information, making it a crucial first step in many information extraction pipelines. Its ability to structure textual data makes NER highly valuable for gaining insights from large volumes of text, enabling applications ranging from information retrieval and question answering to knowledge base population and content analysis (Vajjala & Balasubramaniam, 2022).

2.2.2 NER as a Sequence Tagging Problem

Computationally, NER is frequently approached as a sequence tagging or sequence labeling problem (Yadav & Bethard, 2018). In this paradigm, the input text is treated as a sequence of tokens (words or subwords). The goal is then to assign a corresponding categorical label to each token in the sequence. These labels typically indicate whether a token is outside any entity ('O'), the beginning of a named entity of a specific type ('B-TYPE', e.g., 'B-PER' for beginning-person), or inside a named

entity ('I-TYPE', e.g., 'I-PER' for inside-person). More complex tagging schemes like BIOES (Begin, Inside, Outside, End, Single) also exist. Framing NER this way allows the application of powerful sequence modeling techniques, such as Transformer-based architectures (BERT), to predict the most likely sequence of tags for a given sequence of tokens.

2.2.3 History of NER Approaches and Architectures

Early approaches to NER often involved handcrafted rules and dictionary lookups (Jehangir et al., 2023). While useful in restricted contexts, these methods struggled with scalability, ambiguity, and novel entities. Eventually, classic machine learning methods like the Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) became more popular, leveraging statistical patterns from annotated data (Yadav & Bethard, 2018).

The current state-of-the-art is dominated by deep learning. Architectures based on Recurrent Neural Networks (RNNs), particularly Bidirectional Long Short-Term Memory networks combined with a CRF layer (BiLSTM-CRF), can effectively capture positional relationships (Jehangir et al., 2023; Yadav & Bethard, 2018). More recently, Transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) (Wu & Dredze, 2019) have achieved superior results by using attention mechanisms and large-scale pre-training to learn rich contextual word representations. These models are typically fine-tuned on task-specific data for downstream applications like NER.

For the Dutch language context, relevant to this thesis, specific models like `BERTje` (Vries et al., 2019) and the more recent `RobBERT` series (Delobelle & Remy, 2024) have been developed. Given the effectiveness of Transformer architectures and the aim of this research to evaluate fine-tuning strategies for Dutch governmental texts.

2.3 The Low-Resource Challenge in NER

2.3.1 Data Dependency in Supervised NER

Modern, high-performing Named Entity Recognition (NER) systems, particularly those leveraging supervised deep learning, exhibit a significant weakness: they heavily rely on the availability of large quantities of high-quality, manually labeled training data to learn effectively (Jehangir et al., 2023). The performance of these models is directly correlated with the volume and quality of the annotated examples they are trained on.

2.3.2 The Annotation Bottleneck: Cost and Effort

The process of creating the necessary large-scale annotated datasets is a major practical obstacle. Manual annotation is inherently time-consuming and expensive, representing a significant bottleneck in the development pipeline for NER systems (MacLean & Cavallucci, 2024). This challenge is particularly acute for specialized

domains or less-resourced languages where annotator expertise might be scarce or the annotation guidelines complex (MacLean & Cavallucci, 2024).

2.3.3 Performance Implications of Data Scarcity

When only limited labeled data is available, a situation often termed a "low-resource" or "few-sample" scenario, the performance of supervised NER models can degrade substantially (Jehangir et al., 2023). Furthermore, training processes, especially the fine-tuning of large pre-trained models like BERT, can become unstable and unreliable under these data-scarce conditions, making it difficult to achieve robust results (Jehangir et al., 2023; Zhang et al., 2021).

2.4 Large Language Models (LLMs) as a Potential Solution

2.4.1 Overview of LLMs and Capabilities

Recent years have witnessed the rapid development of Large Language Models (LLMs), such as those based on the Transformer architecture like GPT (Brown et al., 2020). These models are characterized by their massive scale, typically containing billions of parameters, and are pre-trained on vast amounts of diverse text data. This results in a LLM having the ability to perform new tasks with very few examples provided directly in the input prompt, also known as few-shot or in-context learning (Brown et al., 2020; Dong et al., 2024). These capabilities suggest LLMs possess a significant degree of general language understanding.

2.4.2 LLMs for NER

With their significant degree of general language understanding, the direct application of LLMs for Named Entity Recognition (NER) has become a feasible approach (Kim et al., 2024). This is typically achieved by prompting the LLM with a task description for NER, instructing it to identify and categorize entities within a given text. This direct prompting can be performed in a zero-shot manner, relying on the LLM's pre-trained knowledge to understand and execute the NER task without any examples (W. Wang et al., 2019). Alternatively, and more commonly for improved performance, this is done using In-Context Learning (ICL), where a few examples (shots) of the NER task (input text and corresponding labeled entities) are provided within the prompt itself to guide the model's output (Brown et al., 2020).

2.4.3 LLM Usage Constraints

Although Large Language Models (LLMs) present intriguing possibilities for Named Entity Recognition (NER), their direct utilization for tasks like extensive annotation faces considerable constraints. These constraints manifest both monetarily and computationally. Indeed, the direct deployment of LLMs for such purposes can be cost-inefficient and computationally demanding, particularly when resources are limited Santoso et al. (2024). This is largely because state-of-the-art LLMs often contain billions of parameters, the processing of which during inference requires significant computational power.

2.4.4 Addressing Data Scarcity with LLMs

The generative capabilities of LLMs offer a chance for tackling the data scarcity problem common in many supervised NLP tasks, including NER (Santoso et al., 2024). Instead of relying solely on costly manual annotation, LLMs can be employed to generate synthetic training data (Long et al., 2024). This can involve generating new text instances that resemble the target domain or using the LLM to automatically annotate existing unlabeled text (Bogdanov et al., 2024; Kuzman & Ljubešić, 2025). These data augmentation techniques, aim to expand the available training data at a lower cost, potentially improving the robustness and performance of models trained in low-resource settings (Chen et al., 2023; Liu et al., 2022).

2.5 Learning Paradigms and Synthetic Data

2.5.1 Zero-Shot Learning (ZSL)

Zero-Shot Learning (ZSL) is a technique within machine learning first introduced by Chang et al. Chang et al. (2008). Its primary focus is enabling models to recognise categories they have not encountered during the training phase (W. Wang et al., 2019). In standard machine learning a model is trained on a fixed set of classes (e.g. apples and bananas) and can only assign new inputs to those classes. ZSL aims to go beyond this by enabling the model to identify novel categories (e.g. oranges) at test time, even though no examples of these categories were provided during training. This capability typically relies on auxiliary information (also known as side or supplementary information) that encodes descriptive properties or attributes for all classes, including both those observed during training (seen classes) and those withheld (unseen classes).

In the generation of synthetic data using large language models (LLMs), ZSL plays a central role by enabling the model to produce labelled examples for categories without any manually annotated data. LLMs such as GPT-based systems can be prompted with natural-language instructions and class descriptions to generate realistic examples for unseen entity types or tasks, leveraging their broad world knowledge to simulate supervision where none exists.

2.5.1.1 Prompt-based ZSL in an LLM.

1. **Create class descriptors.** For every class—seen or unseen—write a concise natural-language definition or list of attributes, e.g.

   Label: ZEBRA
   Definition: A horse-like mammal with black-and-white stripes.

2. **Compose the prompt.** Insert all class descriptors, add the input to be classified (or an instruction to "generate an example"), and finish with an explicit request:

Task: Which label best matches the following description?
Example: 'A striped equid grazing on the savannah."
Answer:

3. **Run the LLM (no fine-tuning required).** The model aligns the semantics of the input with the provided descriptors and returns the most compatible label—or, when asked, generates a synthetic example that satisfies the descriptor for an unseen class.

2.5.2 K-shot Learning

K-shot learning refers to a spectrum of machine learning scenarios defined by the number of labeled examples, denoted by 'k', available per class during the training or adaptation phase. This encompasses Few-Shot Learning (FSL), which addresses the challenge of training models to generalize effectively to new tasks using a minimal number of labeled examples (Y. Wang et al., 2021), often applied in LLMs through techniques like In-Context Learning (ICL). In-Context Learning (ICL) refers to a paradigm associated with LLMs where the model adapts to perform a task during inference without requiring updates to its parameters (Dong et al., 2024) The core objective in FSL is to enable models to rapidly adapt and make accurate predictions for novel classes after exposure to minimal task-specific data.

This contrasts sharply with the traditional supervised learning paradigm, often termed Many-Shot Learning (MSL), which is characterized by the availability of abundant labeled training data (a large 'K') for each class (Agarwal et al., 2024). This data-rich setting allows standard supervised algorithms, including deep neural networks, to be trained effectively, often achieving high performance, serving as an upper-bound benchmark for FSL and Zero-Shot Learning (ZSL) research, representing the potential achievable with ample data (Agarwal et al., 2024). The term MSL is primarily used to distinguish this conventional approach from data-efficient methods such as FSL and ZSL, which are designed for scenarios with significantly limited or no labeled data per class (W. Wang et al., 2019; Y. Wang et al., 2021). While effective, the many-shot approach faces practical limitations due to the significant cost and effort required for large-scale data collection and annotation, particularly in domains with numerous classes, rare instances, or evolving categories. K-shot learning, therefore, provides a framework for understanding model performance and development across varying degrees of data availability, from scenarios with significantly limited data to those with extensive datasets.

2.6 Related Work: Low-Resource NER and Synthetic Data

2.6.1 Synthetic Transcripts in Slovakian

Lajčinová et al. (2024) focus on developing a Named Entity Recognition (NER) system for extracting address information from speech-to-text transcriptions. Their approach of generating synthetic data using OpenAI's GPT-3.5-turbo API due to limited real-world data shares a similar objective with this thesis. Nevertheless, a

key difference lies in their synthetic data creation process. While Lajčinová et al. (2024) also utilizes raw unlabeled data to create sythetic human-like transcripted texts. Their research focussed first of all only on the slovakian language, but also on locations instead of organizations.

2.6.2 Synthetic Healthcare Data

A study done by Šuvalov et al. (2025) focuses on developing NER models for low-resource languages, specifically Estonian, tackling the challenge of limited annotated health care data. The researchers concentrated on leveraging Large Language Models (LLMs) and synthetic data to create effective NER models while preserving patient privacy. Their method involved a three-step pipeline. First, they generated synthetic Estonian electronic health records using a locally trained GPT-2 model. Second, LLMs like GPT-3.5-Turbo and GPT-4 annotated these synthetic texts to identify entities such as drugs and procedures. Finally, this annotated synthetic data was used to fine-tune an XLM-RoBERTa NER model, which was then tested on real-world Estonian medical texts. This approach avoids using sensitive patient data directly for training.

The research done by Šuvalov et al. (2025) addresses the challenge of limited annotated data. However, it significantly diverges by focusing on the healthcare domain, which presents different challenges than the governmental domain due to its high privacy sensitivity. Another key difference lies in their methodology. Where Šuvalov et al. (2025) leverages an two-stage approach involving the training of a local Large Language Model (LLM) followed by data labeling using a separate GPT model. Due to the privacy sensitive data, which healthcare issues, the need for local training was neccesary to guarantee the privacy of patients.

2.6.3 Synthetic Data for Classification

Harsha et al., 2025 investigate the use of large language models to generate synthetic data for text classification tasks, comparing zero-shot and few-shot approaches. Their results show that synthetic data is effective for objective tasks but less so for subjective ones, where model performance drops significantly. Few-shot prompting improves both quality and diversity of the generated data. These insights are relevant to our work, as we apply zero- and few-shot LLM prompting to generate fine-grained NER data. While NER is generally less subjective, certain entity boundaries can introduce ambiguity. Their findings highlight the importance of guided prompting and careful example selection when using synthetic data to train models in difficult language settings.

2.6.4 Positioning the Current Research

While all three of the aforementioned studies address low-resource NER, each tackles the problem in a domain-specific way, shaped by its own constraints. In contrast, governmental documents such as the Minutes of the Staten-Generaal pose a different scenario: there is no shortage of source material, and the data are publicly available

under the WOO, so privacy concerns are minimal. But no availablity of labeled data combined with the high costs of a custom GPT turn parliamentary NER into a low-resource challenge. This work therefore asks whether cheaper strategies such as zero-/few-shot prompting and synthetic-data generation can deliver high-quality NER without the heavy price tag of full-scale model pre-training.

2.7 Synthesis and Identified Research Gap

Named Entity Recognition (NER) is vital for information extraction (Jehangir et al., 2023; Yadav & Bethard, 2018), but its effectiveness hinges on large labeled datasets, often scarce and costly, especially in specialized domains (MacLean & Cavallucci, 2024). This data scarcity limits model performance in low-resource settings (Jehangir et al., 2023; Zhang et al., 2021). Large Language Models (LLMs) offer a potential solution through synthetic data generation via In-Context Learning (ICL) (Brown et al., 2020; Dong et al., 2024; Long et al., 2024). While LLM-generated data shows promise for low-resource NER (Harsha et al., 2025; Lajčinová et al., 2024; Šuvalov et al., 2025), its effectiveness for fine-tuning Dutch BERT models for domain-specific NER on Dutch governmental documents remains underexplored. Therefore, this thesis investigates the effectiveness of tailored LLM-generated synthetic data for this specific task by generating and evaluating a relevant dataset.

3 Method

This section details the methodology used to answer the research questions posed in Section 1. The overall approach involves creating a gold standard dataset from Dutch governmental texts, generating synthetic data using a Large Language Model (LLM) with varying prompting strategies, analyzing the quality of this synthetic data, and finally evaluating the performance of a Dutch BERT model fine-tuned on generated data.

3.1 Minutes Database

3.1.1 Database Components

The Staten Generaal minutes database is structured with distinct components to ensure consistency across various document types (as detailed in Table 1). The dataset is directly downloaded from the WooGLe drive with permission and only consists of documents in the 2b category.

Document Type	Explanation
Presentie en Opening	These documents list the parliamentarians present at the debate and officially mark the opening of the minutes.
Sluiting	This marks the official close of the minutes and typically only includes the time of closure.
Mededeling	These are special announcements before the start of a minute.
Stemming	This announcement consists of the motion to be voted on and announcing the outcome of the vote.
Lijst van ingekomen stukken	This is a list detailing all the documents that have been received.
Miscellaneous	This category encompasses all other documents, consisting of transcribed minutes on various topics.

Table 1: Document types in minutes by the Staten Generaal

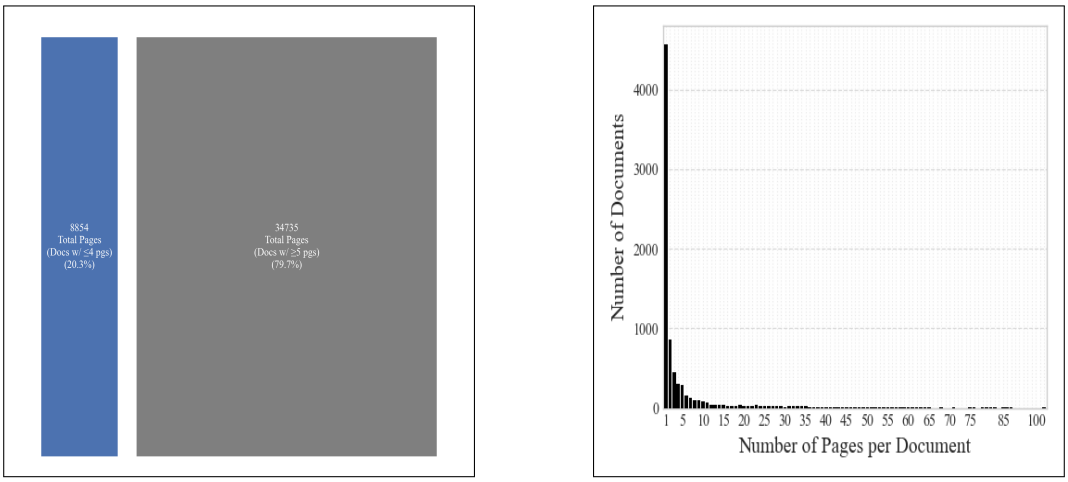
3.2 Gold Standard Dataset

A high-quality, manually annotated dataset is essential both for providing examples to the LLM (for K-shot generation) and for evaluating the final NER model performance. This dataset serves as the 'gold standard'.

3.2.1 Dataset Preparation

The following steps were employed to prepare a candidate dataset for annotation:

1. **Focus on a Relevant Time Period:**
    - The dataset was initially filtered to include documents from 2022 onwards. This timeframe captures the coalitions of Rutte IV and Schoof I. The selection of these years was done for the reason it captures a big shift in the dutch political landscape and therefore could enhance diversity of the samples.
2. **Exclude Non-Substantive Document Types:**
    - As seen in Figure 1 the notion of what a document must consists of has changed in the last 30 years. Where gradually more documents meant exponentially more pages. Table 1 shows the distinction between the document types where only miscellaneous contains actual minutes. Therefor documents consisting of fewer than 5 pages were excluded from the dataset. The split can be seen in figure 3a



(a) Split in pages with ≤ 4 or ≥ 5 pages

(b) Distribution of document length

Figure 3: Document length characteristics

3. **Random Page Selection:**
    - From the remaining pool of pages belonging to these content-rich documents (2022 onwards, >= 5 pages), a random sample of 75 pages was selected from the remaining 34.735 pages.
4. **Pre-process Selected Page Text:**
    - The body text of these 75 selected pages was pre-processed. This involved removing potentially the first and last sentences from the body text of each sampled page to potentially eliminate less relevant introductory/concluding remarks or incomplete sentences.

5. **Preparation of Sample for Annotation:**

   - The document ID and the cleaned body text of these 75 sampled pages were then prepared and exported.

This structured approach aims to create a dataset that is both manageable and rich in relevant content for NER research.

3.2.2 Manual Annotation Process

3.2.2.1 Label Set Definition

The annotation process will adhere to a strict protocol defined in Appendix A. This protocol specifies the entity types (See also Table 2) to be annotated, being domain-specific organizational labels crucial to the OpenGov context. Defining these specialized labels accurately is a key aspect of adapting NER to this domain.

Label	Explanation	Rationale	Examples
PAR	Political Party	Essential for tracking political affiliations, electoral dynamics, and partisan contributions to policy or public narratives.	"PvdA", "Partij van de Arbeid", "PVV", "VVD"
BUS	Business	Allows for analysis of economic actors, corporate influence, market trends, and their interactions with other societal sectors.	"Shell", "ASML", "ING", "Philips"
NGO	Non-Governmental & Activist	Vital for identifying civil society actors (advocacy, activist, advisory groups) and their role in shaping public opinion or policy.	"Extinction Rebellion", "Adviesraad Internationale Vraagstukken"
GEN	Governmental Entity	Fundamental for analyzing state actions, public administration, policy implementation, and official communications.	"Ministerie van Binnenlandse Zaken", "Tweede Kamer", "Politie"
INT	International Organization	Necessary for recognizing supranational/intergovernmental bodies, crucial for research on international relations or global governance.	"EU", "Europese Unie", "Europees Hof voor de Rechten van de Mens"
EDR	Education / Research Institution	Important for identifying sources of academic knowledge, expert opinion, and their influence on policy and societal understanding.	"Universiteit van Amsterdam", "UvA", "WODC"

Table 2: Label Types for Annotation with Rationale

The specific labels were chosen to provide a functional and granular classification of organizational entities. Collectively, this label set enables a more nuanced and precise analysis, allowing for a clearer understanding of the distinct functions, influences, and interactions of these varied organizational types within the analyzed texts. Distinctions between NGO's and corporate organizations could allow for a more in-depth analysis about their influence in the politcal landscape.

3.2.2.2 Annotation Tool and Procedure

Annotation will be performed using Label Studio, an open-source data labeling tool suitable for NER tasks (Tkachenko et al., 2020-2025). The use of Label Studio facilitated this research with the ability to create custom labels as defined in the appendix A.

3.2.2.3 Inter Annotator Agreement and Cohen's Kappa

To ensure the quality of the process of annotation, the tasks will be performed by atleast 2 annotators following the detailed guidelines in the protocol. To ensure consistency Cohen's Kappa between annotators will be performed. This method takes into account the agreement occurring by chance, rather than the simpler percentage agreement. Thereby creating a more robust assesment. To highlight the possible differences in-between label categories the κ will be calculated for each label.

The formula of Cohen's Micro Kappa (κ) is given by:

$$\kappa_{\text{micro}} = \frac{P_o - P_e}{1 - P_e}$$

Where P_o is the observed agreement while P_e is the expected agreement between annotators. These are denoted as

$$P_o = \frac{1}{N}\sum_{i=1}^{k} n_{ii}, \qquad P_e = \sum_{i=1}^{k}\left(\frac{n_{i+}}{N}\right)\left(\frac{n_{+i}}{N}\right)$$

k is the number of distinct annotation categories, N the total number of annotated characters, n_{ii} the number of characters that both annotators assigned to category i, n_{i+} the number of characters Annotator A placed in category i (regardless of Annotator B), and n_{+i} the number of characters Annotator B placed in category i (regardless of Annotator A).

The values for κ range from 0 to 1, where $\kappa > 0.8$, indicates almost perfect agreement. This interpretation comes from Cohen (1960). Table 3 indicates how Cohen's κ will be calculated after the annotation phase for each label on its own.

Annotator 1	**Annotator 2**	**Agreement**
Marked	Marked	TP
Marked	Not Marked	FN
Not Marked	Marked	FP
Not Marked	Not Marked	TN

Table 3: Examples of calculation of Cohen's Kappa in annotation phase.

A micro κ of 0.726 was obtained with a unweighted macro κ of 0.581 and a weighted macro κ of 0.642. According to the commonly used Landis–Koch scale the micro κ indicate solid but not yet outstanding consistency (Landis & Koch, 1977). The gap to the "very good" ($\kappa > 0.80$) suggests that annotators still face domain-specific challenges. Therefor we looked at the κ for each label on its own (see Table 4).

Label	κ	**Grade**
PAR	0.937	'Excellent'
INT	0.650	'Good'
GEN	0.590	'Moderate'
EDR	0.558	'Moderate'
NGO	0.421	'Moderate'
BUS	0.329	'Fair'

Table 4: The κ for each label

The diverging label agreement is largely due to the way organisations are referenced in running text. Brief mentions, such as *"...de bakker..."*, may be interpreted either as a business (BUS) but can also be disregarded and interpreted as a person. A further complication is limited visibility into an organisation's institutional context.

3.3 Synthetic Data Generation

This phase focuses on using `gpt-4.1` with endpoint `-2025-04-14` to generate synthetic data, leveraging the insights from recent work on LLM-driven data generation (e.g., Long et al. (2024)).

3.3.1 Prompting Strategies

To investigate the impact of context examples on generation quality, two distinct prompting strategies will be implemented and compared:

- **Zero-shot:** The LLM will be prompted to generate labeled organizational entities in a document without being provided with any specific examples from the domain. The prompt will only contain the task description and the target label set.
- **K-shot:** The LLM prompt will include the task description, the target label set, and K-amount of diverse examples (1, 2, 5, 10, 25 and 50) of randomly selected sentences from the gold standard dataset.

The sentences used, were sampled from the training set defined in section 3.2.

3.3.2 Prompt Framework

The prompts used for API calls in this research were systematically developed using the RTF-framework. This framework is one of the state-of-the-art prompting techniques and consists of the prompt being divided into Role, Task and Format (See Appendix C).

3.4 Evaluation of Synthetic Data Quality (Addressing RQ1)

To answer RQ1 regarding the representativeness and quality of the synthetically generated datasets, a mixed-methods evaluation approach will be employed. This will compare the synthetic datasets against the gold standard dataset.

3.4.1 Quantitative Analysis

This stage quantitatively will result in metrics about synthetic datasets against the gold standard to assess how well each prompting strategy replicates authentic data characteristics.

3.4.1.1 Cosine Similarity

The first evaluation metric is cosine similarity, which quantifies how closely the vector representation of a generated sentence aligns with that of its reference in semantic space. Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ be the d-dimensional sentence embeddings obtained from a Sentence-BERT model `all-MiniLM-L6-v2`. As this model delivers reliable semantic estimates while keeping compute costs low as it contains only $\approx$ 22M parameters and yields 384-dimensional embeddings.[2] Cosine similarity is defined as

$$\cos\_\mathrm{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|_2 \, \|\mathbf{v}\|_2},$$

yielding a score in the range $[-1, 1]$, where 1 indicates identical direction (maximum semantic overlap) and 0 indicates no shared direction. For each k-shot setting we compute the similarity for every one of the 300 sentence pairs and report the *macro-average*,

$$\cos\_\mathrm{sim}_{\mathrm{macro}} = \frac{1}{N} \sum_{i=1}^{N} \cos\_\mathrm{sim}\big(\mathbf{u}_i, \mathbf{v}_i\big), \qquad N = 300,$$

so that every sentence contributes equally to the final score.

3.4.1.2 Bilingual Evaluation Understudy (BLEU)

BLEU is a metric that measures how close a candidate text (usually a machine-generated translation) is to one or more reference texts produced by humans. Higher BLEU implies the candidate shares more wording and phrasing with high-quality human translations, so it often correlates with human judgments of adequacy and fluency. The BLEU scores are computed with the `nltk` module `sentence_bleu`. All reference sentences are first tokenised. For a given k-shot setting each generated sentence is then evaluated against the full set of tokenised references. We obtain a sentence-level BLEU value for each sentence and report average over the 300 sentences, because this should yield a stable estimate on the small corpus.

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

$$\mathrm{BLEU}_{\mathrm{macro}}(k) \;=\; \frac{1}{N}\sum_{i=1}^{N}\mathrm{BLEU}\big(\mathrm{hyp}_i^{(k)},\,\{\mathrm{ref}_1,\ldots,\mathrm{ref}_{300}\}\big), \qquad N = 300.$$

Here $\mathrm{hyp}_i^{(k)}$ is the generated sentence, where i is the i-th sentence generated by the model under the k-shot prompt. $\mathrm{hyp}_i^{(k)}$ is the reference sentence from the real corpus. Smoothing is applied in the process of sentence-level calculation to prevent low scores for valid sentences lacking higher-order n-gram matches.

3.4.1.3 Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

ROUGE is also a metrics for comparing a system-generated text with one or more human reference summaries. Unlike BLEU, which is precision-oriented, ROUGE emphasizes how much of the content in the reference is recovered by the candidate, so its scores are recall-driven. ROUGE-1, unigram recall, shows coverage of pieces of information that are often expressed by single content words. ROUGE-2, bigram recall, hints at fluency and ROUGE-L captures longer phrasal overlap. We compute ROUGE-1, ROUGE-2 and ROUGE-L with the public `rouge_scorer` library. For every generated sentence we compare it against all 300 reference sentences and keep the highest recall-score, those best scores are then averaged over the 300 sentences to give the ROUGE scores.

$$\mathrm{ROUGE}_{\ell}(k) \;=\; \frac{1}{N}\sum_{i=1}^{N}\max_{j\le N}\mathrm{ROUGE}_{\ell}^{F}\big(\mathrm{ref}_j,\,\mathrm{hyp}_i^{(k)}\big), \qquad \ell \in \{1,2,L\},\; N = 300.$$

3.4.1.4 BERTscore F-1

Token-level semantic overlap is measured with BERTScore using the multilingual `xlm-roberta-large` model as backbone. This model ensured multi linguality and included dutch and showed State-of-the-art correlation with human judgements. For a synthetic sentence and a reference we embed every token and build the cosine-similarity matrix. Precision, recall, and the harmonic mean are:

$$P = \frac{1}{m}\sum_{i=1}^{m}\max_{j} S_{ij}, \qquad R = \frac{1}{n}\sum_{j=1}^{n}\max_{i} S_{ij}, \qquad F_1 = \frac{2PR}{P+R}.$$

We report the mean F_1 over all 300 sentence pairs for each k-shot prompt, as produced by the official BERTscore library. The symmetry in precision and recall arises due to the use of cosine similarity in both directions and the balanced token matching in comparable sentences.

3.4.1.5 Perplexity

Fluency is quantified with the causal Dutch language model `GroNLP/gpt2-small-dutch` (de Vries & Nissim, 2020). We adopt this model because it is one of the few causal Transformers trained natively on large-scale Dutch corpora, ensuring that its probabilities reflect Dutch syntax. Another reason why this model was used is its lightweight

(≈117M parameters), so it can score hundreds of sentences on a single GPU or even a CPU workstation without batching issues. And at last it is publicly released on Hugging Face, which makes our perplexity scores easy to reproduce and compare in future work. Given a generated sentence of length T tokens, the language model assigns a log-likelihood $\log p(w_t \mid w_{<t})$ to each token w_t. Perplexity is defined as

$$\mathrm{PPL} = \exp\Big(-\frac{1}{T}\sum_{t=1}^{T}\log p(w_t \mid w_{<t})\Big),$$

so lower values indicate more predictable (and typically more fluent) text. Following the evaluate implementation we compute the perplexity for every sentence and macro-average over the 300 sentences.

3.4.1.6 Linguistic Feature comparison

To assess the structural and stylistic quality of the generated sentences, we compared basic linguistic features with real sentences. Specifically, we computed sentence length (in characters and tokens) and readability using the Flesch Reading Ease metric. Sentence length was measured using string and token counts, while readability scores were obtained using the `textstat` Python package. These features were aggregated per k-shot setting and compared against real data through descriptive statistics, histograms, and boxplots. This analysis helps to identify whether synthetic outputs approximate the complexity and fluency of authentic parliamentary text.

3.4.2 Qualitative Analysis

This stage involves a systematic qualitative error analysis on samples from each synthetic dataset (zero- and few-shot outputs) to understand how prompting strategies impact the quality of generated annotations and text. Insights will complement quantative analysis, explaining why certain strategies produce more representative and accurate synthetic data.

3.4.2.1 t-distributed Stochastic Neighbor Embedding (t-SNE)

To visualize the semantic distribution of real and generated sentences, we applied t-SNE to sentence embeddings obtained from the pre-trained `all-MiniLM-L6-v2` SentenceTransformer. Embeddings were converted and standardized using StandardScaler to ensure uniform scaling. We set the t-SNE perplexity to 30 and fixed the random seed to 42 to ensure reproducibility. This setup ensures structure in the 2D projection while maintaining consistent embedding distances across runs.

3.5 Model Fine-tuning and Evaluation (Addressing RQ2)

To address RQ2 a Dutch pre-trained BERT model will be fine-tuned using the synthetically generated data, and its performance will be compared against a baseline.

3.5.1 Fine-tuning Process

Based on the findings from Section 3.4, the most promising k-shot strategy will be utilized to generate a synthetic dataset furter explained in 3.5.2. The Dutch `RobBERT-v2-dutch-ner` model will be fine-tuned on this synthetic dataset for the NER task. This model is the State-of-the-art model for Dutch NLP tasks (Delobelle & Remy, 2024). The fine-tuning process was run on Google Colab L4 GPU. The training parameters are defined in Appendix D.

3.5.2 Synthetic Dataset

The synthetic dataset will be generated with the most promising k-shot strategy. The prompt used can be found in Appendix C. The prompt was looped to generate 1.000 sentences. This is the consensus for the minimal amount of fine-tune material.[3] The distribution of the enities in the synthetic data can be seen in Table 5.

Entity type	**Count**
GEN	383
INT	200
PAR	192
NGO	169
BUS	125
EDR	120
Total	**1189**

Table 5: Distribution of entity types (Total entities = 1189)

3.5.3 Baseline Model

A baseline will be established for comparison by using the pre-trained Dutch `RobBERT-v2-dutch-ner` model (Delobelle & Remy, 2024). The task will be performed without any fine-tuning on the target task or domain data. Because there is no such FgNER model for these type of documents, we roll up the fine-grained labels to the `ORG` label and only evaluate the Dutch model's performance on recognizing `ORG` entities.

3.5.4 Performance Evaluation

Both `WorgBERGT` and a baseline `robbert-v2-dutch-base` model are evaluated on the held-out, manually-annotated *gold-standard* test set whose annotations were validated with the IAA procedure (see Paragraph 3.2.2.3).
The baseline `RobBERT-v2-dutch-ner` has no notion of the fine-grained organisational tags introduced in FgNER. To ensure a fair comparison we roll up every organisational subtype to the generic label ORG and evaluate `RobBERT-v2-dutch-ner` on this

[3] https://discuss.huggingface.co/t/thoughts-on-quantity-of-training-data-for-fine-tuning/14886

single class. Non-organisational entities are ignored for the metrics in this experiment.

For `RobBERT-v2-dutch-ner` every predicted span and gold span with label ORG we compute the span-level Jaccard overlap. For WorgBERT the same is done but with all the fine-grained labels.

$$J(p, g) = \frac{|p \cap g|}{|p \cup g|}.$$

A prediction is counted as a True Positive when $J(p, g) \geq 0.75$; otherwise it is a False Positive. Gold spans that are not matched by any prediction at the 0.75 threshold are treated as False Negatives.
With the counts of true positives (TP), false positives (FP) and false negatives (FN) we compute precision, recall and F_1:

$$P = \frac{TP}{TP + FP}, \qquad R = \frac{TP}{TP + FN}, \qquad F_1 = \frac{2PR}{P + R}.$$

The F_1 was chosen as the main metric as it finds the balance between precision and recall. This is important because we want to find as many correctly predicted labels but also minimalize the amount of wrongly predicted labels.

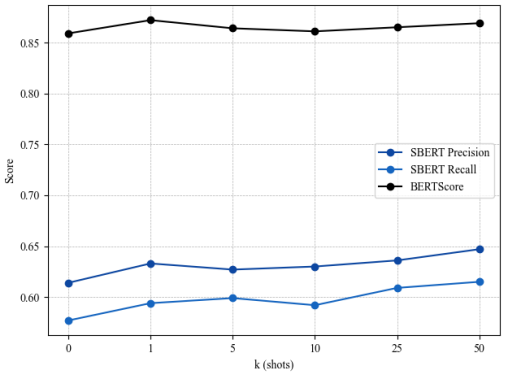
4 Results

This chapter presents the results of the evaluation of the synthetically generated corpus'. Followed by the results of the evaluation of a fine-tuned `RobBERT` model. The findings are structered to systematically answer the research questions:

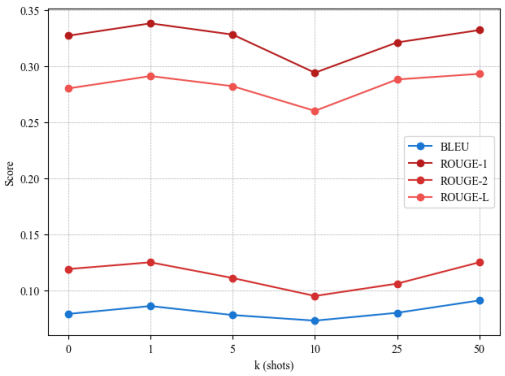
RQ1: How well do Large Language Model perform in generating a representative synthetic dataset of NER compatible labeled Dutch governmental texts?

RQ2: How does the performance of a Dutch BERT model fine-tuned on synthetically generated dataset compare to that of a pre-trained (non-fine-tuned) Dutch BERT model?

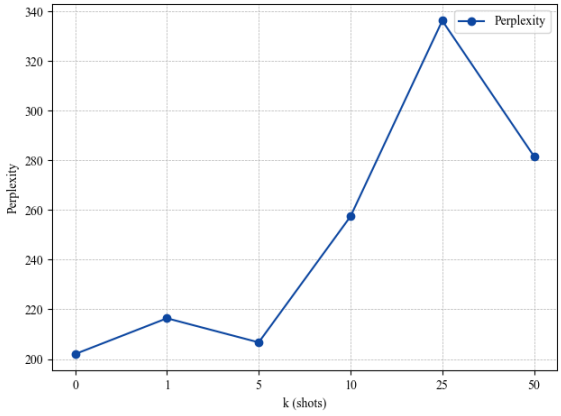
4.1 Quantitative Evaluation for K-shot



(a) SBERT precision/recall and BERTScore.



(b) BLEU and ROUGE metrics.



(c) Negative perplexity across k-shot prompts.

Figure 4: Summary of evaluation curves for the summarization model as the number of k-shot examples increases.

The evaluation results reveal several trends across different k-shot settings: First, SBERT similarity scores (precision and recall) improve as k increases, with the largest jump between 0-shot and 1-shot. After 25-shot, the gains level off.

Second, token-level overlap metrics (BLEU and ROUGE) also increase with more examples. The dip around the 5- and 10-shot may indicate that the generated sentences overfit the limited prompts, relying on repetitive phrasing that lowers their overlap with the reference set. The best scores are achieved with the 50-shot prompt, though even a single example (1-shot) already leads to strong improvements in lexical similarity.

Lastly, fluency, measured by perplexity (PPL), *increases* as k grows. The higher PPL observed in the 25- and 50-shot settings may indicate that the model is adopting more formal or domain-specific language rather than becoming less fluent. This shift could be beneficial if it aligns the output style more closely with the target domain. The drop from 25- to 50-shot indicates that the model is overflowed with examples, making the model output more predictable sentences.

In summary, providing more examples (k) enhances generation quality, with the most significant gains occurring between 0- and 1-shot. While a larger k boosts semantic and lexical scores and mitigates the overfitting seen in smaller few-shot settings, it also increases perplexity. This rise in PPL likely reflects a beneficial shift toward a more complex, domain-specific style. The optimal choice of k therefore balances improved accuracy against this increased stylistic complexity, with 25 or 50 examples proving most effective for mastering the target domain.

4.1.1 Linguistic Feature Comparison

Table 6 presents descriptive statistics for each k-shot subset alongside the real corpus. On the surface, the figures look similar: every synthetic subset keeps its mean close to 11–13 words and its median within a two-word band of the mean. A quick glance might therefore tempt one to conclude that sentence length has been matched adequately across all data slices.

Figure 5, however, tells a more nuanced story. By plotting the complete distributions rather than mere point summaries, it shows two patterns.

- Almost all synthetic subsets top out at 21 words (18 words for $k = 5$ and $k = 50$). Whereas the real corpus reaches 66 words. This ceiling effect suggests that longer, more complex sentences are entirely absent from the generated material.

- As k rises (25- and 50-shot), the mean shifts toward shorter sentences (≈ 10 words), and variance shrinks. In other words, larger prompt sizes appear to encourage the model to produce more concise sentences.

To conclude the model is fundamentally biased towards generating structurally simple sentences. This limitation is not only masked by basic statistics but is paradoxically

k	Mean	Min	Median	Max
0	13.47	8	13	21
1	11.24	7	11	21
5	12.21	7	12	18
10	12.09	7	12	21
25	10.74	2	10	21
50	10.90	2	11	18
real_	12.07	1	9	66

Table 6: Sentence length statistics (in words) per k-shot.

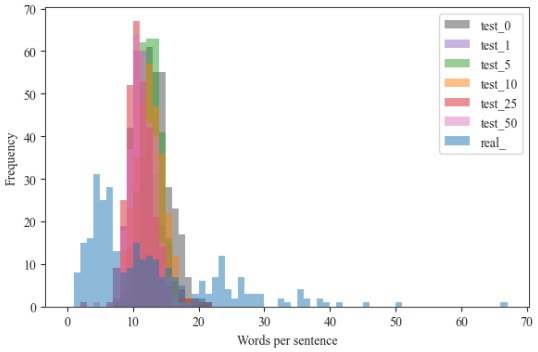


Figure 5: Sentence length distribution across k-shot.

worsened by providing more examples (k), which results in even shorter, more uniform sentences.

4.1.1.1 Flesh readability

As Figure 6 illustrates, there is a sharp contrast in the variability of readability scores. The authentic sentences from the real corpus show a wide distribution, signifying a rich mix of simple and complex sentence structures. On the contrary, the generated sentences consistently form much tighter clusters, indicating they occupy a narrower and more predictable "comfort zone" of readability.

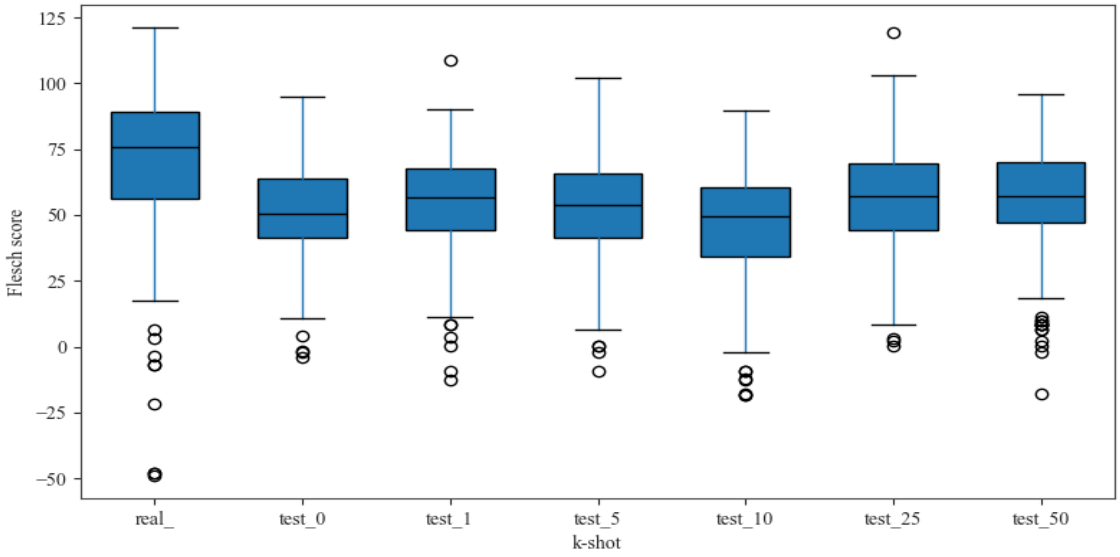


Figure 6: Boxplot of Flesch readability scores for real and k-shot generated sentences.

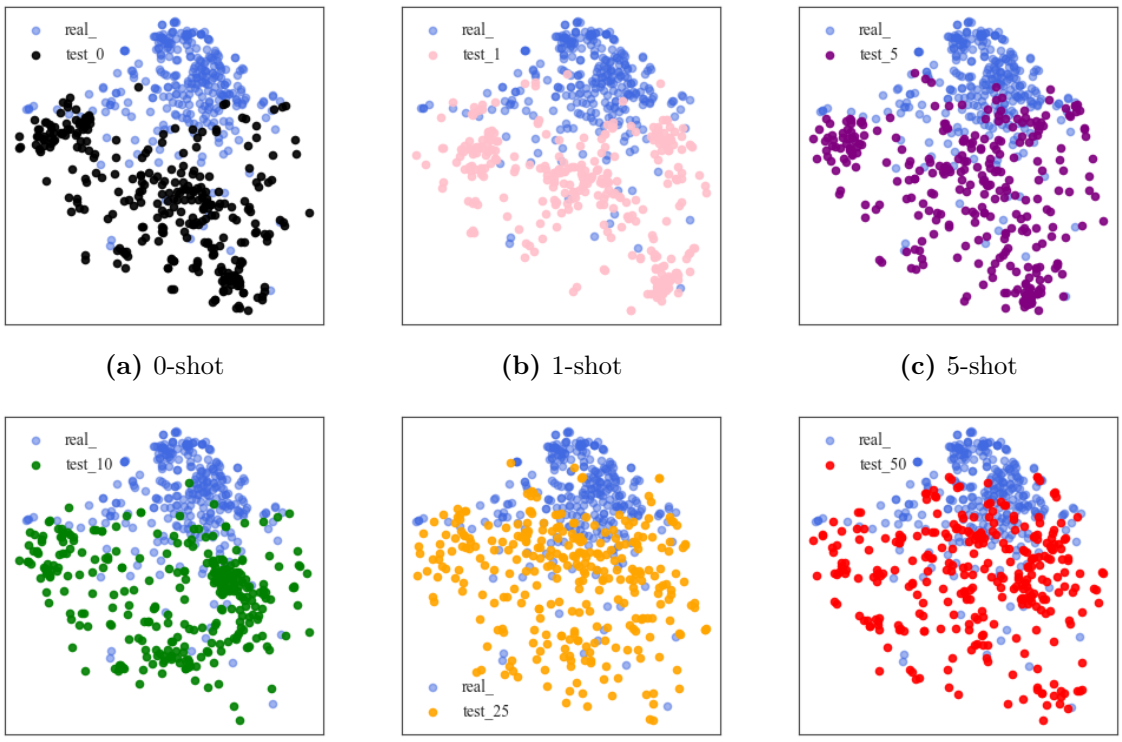
A likely reason is that the language model never sees the full complexity of parliamentary style: its output is shaped only by the prompt and the limited set of k-shot examples we give it. If those examples don't cover the extremes of simple and intricate wording, the model will naturally stay within the safer middle ground.

This results in the generated sentences being less complex than the golden standard. Also this makes the generated text feels more homogenous and fails to capture the characteristic fluctuations in complexity found in the authentic corpus.

4.2 Qualitative Evaluation for K-shot

4.2.1 t-SNE

The t-SNE plots (Figure 7) show the embedding distributions of real versus synthetic sentences across different k-shot strategies. With increasing k, synthetic data progressively approximates the semantic structure of real parliamentary sentences. The 25-shot and 50-shot generations show the highest overlap. These observations confirm that few-shot prompting steadily narrows the distributional gap between synthetic and real sentences, with the most pronounced improvements occurring within the first 25 demonstrations. Beyond that point additional shots yield diminishing semantic returns, suggesting that 25-shot constitutes a practical sweet-spot for this dataset and model.



(a) 0-shot **(b)** 1-shot **(c)** 5-shot

(d) 10-shot **(e)** 25-shot **(f)** 50-shot

Figure 7: t-SNE visualization of sentence embeddings for real and synthetic data across different k-shot settings

The experiments show that increasing the number of real examples in the k-shot fine-tuning progressively aligns the synthetic output with authentic minutes of the *Staten-Generaal*. Remarkably, the 25-shot setting already yields almost the same qualitative similarity as the 50-shot setting, yet with a substantially lower perplexity, indicating a better imitation of formal parliamentary language and stronger lexical overlap.

4.3 K-shot findings

The finding of this part of the research is a disconnect between the LLMs ability to replicate semantic content and its capacity for structural authenticity. While increasing the number of k effectively aligns the generated text with the target domain's vocabulary and topics (as confirmed by metrics like BLEU, SBERT, and t-SNE visualizations), this success masks failure. The LLM imposes a strong bias for simplicity, consistently failing to produce the long, complex, and stylistically varied sentences of the authentic corpus. Paradoxically, providing more examples can even worsen this structural deficit, leading to shorter and more uniform sentences.

4.4 Evaluating WorgBERT

4.4.1 Gold Standard Test Set (Fine-Grained)

After fine-tuning `RobBERT-v2-dutch-base`, we evaluated `WorgBERT` on both the imbalanced gold standard test set and performed a sanity check on the synthetic test set. For the first, we also ran the unmodified `RobBERT-v2-dutch-base` as a direct baseline.

Class	Precision	Recall	F1-score	Support
BUS	0.000	0.000	0.000	13
EDR	0.000	0.000	0.000	4
GEN	0.081	0.053	0.064	152
INT	0.500	0.067	0.118	15
NGO	0.000	0.000	0.000	12
PAR	0.882	0.375	0.526	120
micro avg	0.293	0.171	0.216	316
macro avg	0.244	0.082	0.118	316
weighted avg	0.398	0.171	0.236	316

Table 7: Classification Report per Entity on Gold standard.

On the gold-standard corpus WorgBERT attains a micro-precision of 0.168, a micro-recall of 0.206, and a micro-F_1 of 0.185; the weighted F_1 rises to 0.262 owing to class imbalance. Per-label scores show that `PAR` reaches an F_1 of 0.606, `INT` 0.238, `GEN` 0.064, while `BUS`, `NGO`, and `EDR` remain at 0.

The sharp drop from synthetic set performance to these gold scores signals that WorgBERT overfits the balanced training distribution and struggles with the skewed, noisy reality of parliamentary text. Almost all correct detections cluster in the high-frequency `PAR` class, whereas minority labels are rarely identified, and even the high-frequency `GEN` proves difficult. Class imbalance and domain shift, are now the principal bottlenecks, indicating that targeted augmentation, re-weighting, or curriculum learning will be required to recover recall for low-frequency organisation types.

4.4.2 Synthetic Test Set (Sanity Check)

WorgBERT achieves an overall accuracy of 0.80 and a weighted F_1 of 0.81 on the balanced synthetic corpus (Table 8). Because class priors are uniform, the macro F_1 remains comparable at 0.82, indicating that every organisational subtype is handled reasonably well.

Best performance is obtained by `BUS` and `PAR`, with respectively $F_1 = 0.955$ and $F_1 = 0.919$. In contrast, `GEN` registers the lowest $F_1 = 0.711$, highlighting a little weakness in recognising governmental entities.

Class	Precision	Recall	F1-score	Support
BUS	0.840	0.840	0.840	25
EDR	0.667	0.889	0.762	18
GEN	0.711	0.711	0.711	83
INT	0.955	0.933	0.944	45
NGO	0.852	0.622	0.719	37
PAR	0.919	0.971	0.944	35
micro avg	0.812	0.802	0.807	243
macro avg	0.824	0.828	0.820	243
weighted avg	0.817	0.802	0.806	243

Table 8: Classification Report per Entity on Synthetic Test-Set.

4.4.3 Baseline RobBERT (Coarse-Grained)

The Dutch `RobBERT-v2-dutch-base` model is evaluated on the same corpus after collapsing all six sub-types to the single `ORG` tag (Table 9). It returns an precision of 0.726, recall of 0.485 and a F_1 of 0.581. The F_1 score is roughly double that of WorgBERT in the fine-grained setting.

Class	Precision	Recall	F1-score	Support
ORG	0.726	0.485	0.581	262

Table 9: Classification report RobBERT on ORG entity

What does have to taken into account, is the drop in support in the gold standard data. Certain labeled entities wich were next to eachother have been combined in the process. Still the troubles of correctly labeling all organisational entities is a difficult task, even for `RobBERT`.
It is important to note the sharp drop in support within the gold-standard data: adjacent labelled entities were often merged, reducing the number of training signals for rare organisational types. Still, one might expect `RobBERT`, to handle such entities more gracefully. In practice, however, the model still misses or misclassifies a surprising share of organisational mentions, underscoring how challenging NER remains in this field, despite the use of a powerful language model.

4.5 Error Analysis of WorgBERT

4.5.1 Global picture.

Table 10 shows that only 35 spans are recognised correctly, while almost a half of the model's outputs are false positive and roughly the same number of gold mentions are missed (FN).

Status	# Spans
TP_{exact}	32
$TP_{partial}$	3
FP	298
FN	281

Table 10: Span–level outcomes for the gold standard set (614 gold or predicted spans in total).

4.5.2 Qualitative error categories

A manual inspection of the errors reveals three dominant types:

- **Boundary errors** ($TP_{partial}$, 3 cases). The model captures the head of a multi-word entity but omits obligatory modifiers:

  real span: "Kamercommissie voor Justitie en Veiligheid"
  predicted span: "voor Justitie en Veiligheid"

  Such near-misses score high Jaccard overlap. And are therefore categorized as partial TP.

- **False single-token predictions** (FP, 298 cases). About half of all errors fall in this category. Typical patterns are:

  – Bare surnames or common nouns: *"Van", "Bruins"* tagged as `GEN`.

  – Random organisation-looking strings, e.g. *"NIS2-"* as `BUS`, or personal names such as *"Rajkowski"* tagged as `NGO`.

  These indicate that the span-level context window learned during fine-tuning is insufficient to suppress out-of-context entity triggers.

- **Completely missed entities** (FN, 281 cases). The model frequently misses established abbreviations (*"SGP"*) or long institutional names (*"Wetenschappelijke Raad voor het Regeringsbeleid"*). In the latter case the sentence often contains multiple capitalised tokens, causing the model to abstain entirely.

4.6 Key Findings

- **Few-shot quality rises predictably.** Both semantic similarity (SBERT) and lexical overlap (BLEU, ROUGE) leap from 0- to 1-shot and continue to climb up to 25-shot, after which further gains flatten out. The t-SNE plots portray the same image, where through 0- to 25-shot the overlap increases. The 50-shot prompt attains the highest overlap scores, but its lower perplexity reveals a shift toward more formal predictable language rather than fluency. The consideration for 25- or 50-shot comes down to the choice for how predictable the sentences need to be.

- **Averages mask missing extremes.** Sentence-length and Flesch analyses show that synthetic data systematically leaves out very long or very easy/complex sentences, even when headline metrics look strong, highlighting the need to match the full distribution of real language, not just its center. The Flesch distributions show that the LLM generated sentences follow more of a safe zone in terms of difficulty.

- **Balanced data enables fine-grained learning.** On the uniformly distributed synthetic test set, WorgBERT reaches macro F_1 of 0.82, confirming that the model can separate all six organisational sub-types when each receives adequate representation.

- **Class imbalance is a big obstacle in practice.** When evaluated on the imbalanced gold corpus, WorgBERT's performance collapses to macro $F_1 = 0.118$ (micro $F_1 = 0.216$) where three minority labels are never predicted and almost half of all predictions are false single-token predicitions. When comparing `WorgBERT` results to the earlier mentioned gold standard Cohen k scores (Table 4), a similar trend can be seen. Although all sub-types have a overlap, `PAR` is categorized with the "Excellent" mark. This is followed by `INT`, which is also the case in the results of `WorgBERT` on the gold standard data set.

- **Comparing to RobBERT.** Collapsing every sub-type to a single `ORG` tag and using the off-the-shelf `RobBERT-v2` shows effectiveness with $F_1 = 0.58$, but still misses more than 40 % of organisational mentions—showing that even coarse-grained NER remains challenging in parliamentary text.

- **High-level observation.** Despite the apparent quality of the synthetic corpus, a model fine-tuned on it still struggles on real parliamentary text—and even an off-the-shelf transformer cannot perform as expected. Fine-grained NER in this domain therefore remains stubbornly difficult.

5 Discussion

The results confirm that few-shot data generation paired with transformer fine-tuning could push Dutch fine-grained NER far beyond a coarse `ORG` baseline, yet several limitations remain. Increasing the number of k-shot demonstrations directly improves an LLM's ability to mimic the semantics and style of authentic parliamentary text: SBERT, BLEU, and ROUGE leap from 0- to 1-shot, climb steadily to 25-shot, and then plateau; t-SNE plots show the same convergence of synthetic toward real embeddings. A 25-shot prompt already captures most of the attainable overlap, while 50-shot offers only marginal gains and begins to lower perplexity through repetition. However, the generation pipeline never produces the long (up to 66-word) or very easy/complex sentences found in the gold corpus, and Flesch scores cluster in a narrow "comfort zone."

When fine-tuned on the balanced data, `WorgBERT` attains macro $F_1 = 0.82$ on the synthetic test set yet collapses to 0.118 on the imbalanced gold set, missing every minority label. Even the general-purpose `RobBERT-v2`, evaluated after collapsing all sub-types, recalls only 48 % of organisational mentions ($F_1 = 0.58$).

5.1 Limitations and Future Work

5.1.1 Model Transparency and Potential Bias

It remains unclear whether models such as ChatGPT were trained on publicly available Woo (Open Government) documents. This uncertainty raises concerns about potential data leakage or memorisation, particularly for verbatim outputs. Although sampling and filtering strategies were applied, future work should explicitly audit the generated text for originality.

5.1.2 Input Context and Sentence Structure

The present study used isolated sentences for both generation and fine-tuning. Longer parliamentary segments or full debates could provide richer context and improve model performance. Future research should therefore explore multi-sentence generation or dialogue-level modelling for finer-grained NER. Note that even minor changes in prompt format or wording can substantially affect the output.

5.1.3 Baseline Alternatives and Trivial Lookup

Frequently occurring entities—such as political parties or ministries—may be identifiable with simpler methods (e.g. keyword search or regular expressions). The true value of fine-grained NER lies in its robustness across varied and ambiguous contexts, which must be emphasised when comparing against rule-based alternatives.

5.1.4 Class Imbalance and Additional Training

The performance drop on the gold corpus shows that rare organisational types still dominate the errors. Future work should focus on targeted augmentation or

curriculum schedules to ensure that these classes receive sufficient training signal. Our experiments suggest that once coverage improves, fine-grained NER becomes markedly more viable.

In summary, this work demonstrates the feasibility of using LLMs for controlled data generation in Dutch governmental NER tasks. It also lays the groundwork for future research aimed at evaluating generalisation over time, understanding model bias, and leveraging richer textual contexts.

6 Conclusion

This thesis explored the use of large language models (LLMs) for generating high-quality synthetic data to fine-tune a Dutch BERT-based model (WorgBERT) for fine-grained named entity recognition (FgNER) on minutes by the Staten-Generaal. The results demonstrated that increasing the number of examples in few-shot prompting improved both semantic and lexical similarity to real-world data. Among the evaluated strategies, 50-shot prompting yielded the best overall metrics across SBERT similarity, ROUGE, BLEU, and second best at BERTScore.

The fine-tuned WorgBERT model achieved strong performance on synthetic test data, but its performance diminishes on the gold standard set. When compared to the general-purpose `RobBERT-v2-dutch-ner`, WorgBERT provided more specialized predictions for organizational entities. These findings suggest that synthetic data generation via prompting could meaningfully contribute to domain adaptation in low-resource or evolving public sector domains. But still need a lot of tweaking in this domain to become viable.

References

Agarwal, R., Singh, A., Zhang, L., Bohnet, B., Rosias, L., Chan, S., Zhang, B., Anand, A., Abbas, Z., Nova, A., Co-Reyes, J. D., Chu, E., Behbahani, F., Faust, A., & Larochelle, H. (2024). Many-shot in-context learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang (Eds.), *Advances in neural information processing systems* (pp. 76930–76966, Vol. 37). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2024/file/8cb564df771e9eacbfe9d72bd46a24a9-Paper-Conference.pdf

Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. *Government Information Quarterly*, *32*(4), 399–418. https://doi.org/https://doi.org/10.1016/j.giq.2015.07.006

Bogdanov, S., Constantin, A., Bernard, T., Crabbé, B., & Bernard, E. P. (2024, November). NuNER: Entity recognition encoder pre-training via LLM-annotated data. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 11829–11841). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.660

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. Retrieved April 8, 2025, from https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Chang, M.-W., Ratinov, L., Roth, D., & Srikumar, V. (2008). Importance of Semantic Represenation: Dataless Classification. *Proceedings of the National Conference on Artificial Intelligence (AAAI)*. https://dl.acm.org/doi/10.5555/1620163.1620201

Chen, J., Tam, D., Raffel, C., Bansal, M., & Yang, D. (2023). An Empirical Survey of Data Augmentation for Limited Data Learning in NLP. *Transactions of the Association for Computational Linguistics*, *11*, 191–211. https://doi.org/10.1162/tacl_a_00542

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales - Jacob Cohen, 1960. Retrieved April 24, 2025, from https://journals.sagepub.com/doi/10.1177/001316446002000104

Delobelle, P., & Remy, F. (2024). RobBERT-2023: Keeping dutch language models up-to-date at a lower cost thanks to model conversion. *Computational Linguistics in the Netherlands Journal*, *13*, 193–203. Retrieved April 4, 2025, from https://www.clinjournal.org/clinj/article/view/180

de Vries, W., & Nissim, M. (2020). As good as new. how to successfully recycle english gpt-2 to make models for other languages.

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., Sun, X., Li, L., & Sui, Z. (2024, November). A Survey on In-context Learning. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*

(pp. 1107–1128). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.64

Harsha, C., Phogat, K. S., Dasaratha, S., Puranam, S. A., & Ramakrishna, S. (2025, January). Synthetic Data Generation Using Large Language Models for Financial Question Answering. In C.-C. Chen, A. Moreno-Sandoval, J. Huang, Q. Xie, S. Ananiadou, & H.-H. Chen (Eds.), *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)* (pp. 76–95). Association for Computational Linguistics. Retrieved June 18, 2025, from https://aclanthology.org/2025.finnlp-1.7/

ICAI. (n.d.). Retrieved April 22, 2025, from https://www.icai.ai/labs/icai-opengov-lab#projects

Jehangir, B., Radhakrishnan, S., & Agarwal, R. (2023). A survey on named entity recognition — datasets, tools, and methodologies. *Natural Language Processing Journal*, *3*, 100017. https://doi.org/10.1016/j.nlp.2023.100017

Kim, H., Kim, J.-E., & Kim, H. (2024, November). Exploring Nested Named Entity Recognition with Large Language Models: Methods, Challenges, and Insights. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 8653–8670). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.492

Kuzman, T., & Ljubešić, N. (2025). LLM teacher-student framework for text classification with no manually annotated data: A case study in IPTC news topic classification [Conference Name: IEEE Access]. *IEEE Access*, *13*, 35621–35633. https://doi.org/10.1109/ACCESS.2025.3544814

Lajčinová, B., Valábek, P., & Spišiak, M. (2024, February). Named Entity Recognition for Address Extraction in Speech-to-Text Transcriptions Using Synthetic Data [arXiv:2402.05545 [cs]]. https://doi.org/10.48550/arXiv.2402.05545

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/https://doi.org/10.2307/2529310

Liu, J., Chen, Y., & Xu, J. (2022). Low-resource NER by data augmentation with prompting. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 4252–4258. https://doi.org/10.24963/ijcai.2022/590

Long, L., Wang, R., Xiao, R., Zhao, J., Ding, X., Chen, G., & Wang, H. (2024, August). On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 11065–11082). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.658

MacLean, C., & Cavallucci, D. (2024). Assessing Fine-Tuned NER Models with Limited Data in French: Automating Detection of New Technologies, Technological Domains, and Startup Names in Renewable Energy [Number: 3 Publisher: Multidisciplinary Digital Publishing Institute]. *Machine Learning and Knowledge Extraction*, *6*(3), 1953–1968. https://doi.org/10.3390/make6030096

Mohit, B. (2014). Named entity recognition. In I. Zitouni (Ed.), *Natural language processing of semitic languages* (pp. 221–245). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-45358-8_7

Overheid, D. (2025, April). Wet open overheid wetgeving - digitale overheid. https://www.digitaleoverheid.nl/wet-open-overheid/

Santoso, J., Sutanto, P., Cahyadi, B., & Setiawan, E. (2024, August). Pushing the limits of low-resource NER using LLM artificial data generation. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the association for computational linguistics: ACL 2024* (pp. 9652–9667). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.575

Šuvalov, H., Lepson, M., Kukk, V., Malk, M., Ilves, N., Kuulmets, H.-A., & Kolde, R. (2025). Using Synthetic Health Care Data to Leverage Large Language Models for Named Entity Recognition: Development and Validation Study. *Journal of Medical Internet Research*, *27*, e66279. https://doi.org/10.2196/66279

Tkachenko, M., Malyuk, M., Holmanyuk, A., & Liubimov, N. (2020-2025). Label Studio: Data labeling software [Open source software available from https://github.com/HumanSignal/label-studio]. https://github.com/HumanSignal/label-studio

Vajjala, S., & Balasubramaniam, R. (2022, June). What do we really know about state of the art NER? In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 5983–5993). European Language Resources Association. https://aclanthology.org/2022.lrec-1.643/

Vries, W. d., Cranenburgh, A. v., Bisazza, A., Caselli, T., Noord, G. v., & Nissim, M. (2019, December). BERTje: A Dutch BERT Model [arXiv:1912.09582 [cs]]. https://doi.org/10.48550/arXiv.1912.09582

Wang, W., Zheng, V. W., Yu, H., & Miao, C. (2019). A Survey of Zero-Shot Learning: Settings, Methods, and Applications. *ACM Transactions on Intelligent Systems and Technology*, *10*(2), 1–37. https://doi.org/10.1145/3293318

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2021). Generalizing from a Few Examples: A Survey on Few-shot Learning. *ACM Computing Surveys*, *53*(3), 1–34. https://doi.org/10.1145/3386252

Wu, S., & Dredze, M. (2019, November). Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 833–844). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1077

Yadav, V., & Bethard, S. (2018, August). A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 2145–2158). Association for Computational Linguistics. Retrieved April 29, 2025, from https://aclanthology.org/C18-1182/

Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., & Artzi, Y. (2021). Revisiting Few-sample BERT Fine-tuning. *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.2006.05987

A Annotation Protocol

This protocol outlines the annotation guidelines for identifying and labeling specific entities and events within Minutes by the Staten Generaal text. This appendix will include all necessary information to replicate the annotation process. Consistency across annotations is necessary for the success of this project.

A.1 Objective

The primary goal of this annotation task is to identify and label specific types of named entities within governmental documents (such as parliamentary debates). This will enable research into the interactions and mentions of various organizational actors within these texts.

A.2 Entity Labels

Annotate text spans corresponding to the following entity types. Use the **longest possible span** that refers to the specific entity. Exclude titles like "Minister", "De heer", "Mevrouw" unless they are part of the official organization name.

A.2.1 GEN (Government Entity)

- **Definition:** National, regional, or local government bodies, agencies, law enforcement and legislatures.
- **Examples:** *Ministerie van Binnenlandse Zaken*, *Tweede Kamer*, *Politie*, *Openbaar Ministerie*, *OM*, *driehoek*, *Raad van State*, *Belastingdienst*, *NVWA*, *Buitenlandse Zaken*, *Kabinet*, *Comissie*.
- **Annotation:** Select the full name (e.g., *Ministerie van Binnenlandse Zaken*, not just *Ministerie*). Annotate acronyms if they clearly refer to a government entity (e.g., *OM*).

A.2.2 PAR (Political Party)

- **Definition:** Formally registered political parties participating in the political process.
- **Examples:** *VVD*, *Partij voor de Dieren*, *PvdD*, *BBB*, *CDA*, *SGP*, *ChristenUnie*, *PVV*, *NSC*, *GroenLinks-PvdA*, *DENK*, *Lid-Haga*.
- **Annotation:** Select the full party name or its common abbreviation (e.g., *Partij voor de Dieren*, *PvdD*).

A.2.3 INT (International Organization)

- **Definition:** International or supra-national governmental organizations (like the EU, UN agencies, international courts).

- **Examples:** *Europees Hof voor de Rechten van de Mens*, *Europese Unie*, *EU*, *G7*, *UNRWA*.
- **Annotation:** Select the full name or common abbreviation (e.g., *Europese Unie*, *EU*).

A.2.4 NGO (Non-Governmental)

- **Definition:** Non-governmental organizations, including activist groups, interest groups, advisory boards, unions, think tanks, and observatories not primarily educational or commercial.
- **Examples:** *Extinction Rebellion*, *Farmers Defence Force*, *mensenrechtenorganisaties*, *Politiebond*, *Syrisch Observatorium voor de Mensenrechten*, *Adviesraad Internationale Vraagstukken*.
- **Annotation:** Select the full name of the group. For generic terms like *mensenrechtenorganisaties*, annotate if it refers to specific, identifiable (though unnamed) organizations acting in that capacity within the context.

A.2.5 BUS (Businesses)

- **Definition:** Commercial enterprises, financial institutions, and industry groups.
- **Examples:** *Tata Steel*, *Havenbedrijf Rotterdam*, *ING*, *Verzekeraars*, *Pensioenfondsen*, *Nederlandse bedrijven*.
- **Annotation:** Select the company/institution name. Annotate generic types like *Verzekeraars* if they refer to the industry or specific (though unnamed) actors in context.

A.2.6 EDR (Educational / Research Institution)

- **Definition:** Educational institutions (like universities) and dedicated research bodies or think tanks linked to academia or specific organizations.
- **Examples:** *Universiteit van Amsterdam*, *UvA*, *WODC*, *wetenschappelijk bureau van de VVD*.
- **Annotation:** Select the full name or common abbreviation (e.g., *Universiteit van Amsterdam*, *UvA*).

A.3 Annotation Guidelines

- **Span Selection:** Select the **entire contiguous span** of text that constitutes the name of the entity.
  - *Example:* Annotate *Ministerie van Binnenlandse Zaken* not just *Ministerie*.

  - *Example:* Annotate *Partij voor de Dieren* not just *Dieren*.

- **Exclusions:** Do **not** include preceding titles (like "Minister", "De heer", "Mevrouw") or trailing possessives ('s) unless they are part of the official name.

  - *Example:* In "de minister van Justitie", annotate *Justitie* as **GEN** if it refers to the Ministry. For "Mevrouw Teunissen (PvdD)", annotate *PvdD* as **PAR**.

- **Nested/Overlapping Entities:** This protocol focuses on flat NER. Annotate the most specific, longest span. For example, in *wetenschappelijk bureau van de VVD*, annotate the whole phrase as **EDR** and annotate *VVD* separately as **PAR**.

- **Acronyms/Abbreviations:** Annotate common acronyms and abbreviations (e.g., *VVD*, *OM*, *EU*, *UvA*) with the appropriate label if their meaning is clear in the context.

- **Context is Key:** Use the surrounding sentences to understand the role and type of the entity if the name alone is ambiguous.

- **Ambiguity:** If you are genuinely unsure about the correct label for a span, or whether something is an entity at all, use the "Skip" function in Label Studio.

B Labels Used In Label Studio

B.1 Label Studio

```
<View>
  <Labels name="label" toName="text">
    <Label value="PAR" background="#FFA39E"/>
    <Label value="BUS" background="#D4380D"/>
    <Label value="GEN" background="#FFC069"/>
    <Label value="EDR" background="#AD8B00"/>
    <Label value="NGO" background="#D3F261"/>
    <Label value="INT" background="#389E0D"/>
  </Labels>
  <Text name="text" value="$text"/>
</View>
```

C Prompts Used

C.1 Prompting k-shot

Following will be the prompt used to generate syntheticly labeled data for Minutes of the Staten-Generaal. The procedure will follow a K-Shot mechanism, exploiting zero and few-shot, which is explained in 2.5.1 and 2.5.2.

```
system_prompt = f"""
**Rol**
Je bent een taalmodel gespecialiseerd in instructievolging en
    tekstannotatie voor parlementaire documenten.

**Taak**
Genereer een formele Nederlandse zin. De zin moet realistisch
    voorkomen in parlementaire documenten (net als in de Handelingen
     van de Staten-Generaal) en moet **eindigen met een punt,
    vraagteken of uitroepteken**. Maak de zinnen niet generiek.
    Hieronder staat mogelijk een spreekbeurt uit een handeling van
    de Staten-Generaal.
Gebruik deze als inspiratie voor het genereren van nieuwe
    voorbeelden zonder bias naar de voorbeelden.

**Format**
Zinnen los van elkaar zonder numering.

**Voorbeelden**
{zinnen_sample_str}
"""

user_prompt = """
Genereer nu 10 nieuwe synthetische zinnen zoals uitgelegd in de
    instructies.
"""
```

C.2 Prompting for Synthetic Data

```
system_p = """
<SystemPrompt>
  <Role>
    Je bent een geavanceerd taalmodel dat gespecialiseerd is in het
     genereren van formeel en domeinspecifiek Nederlands,
    vergelijkbaar met de taal die wordt gebruikt in parlementaire
    documenten van de Staten-Generaal. Je begrijpt de conventies,
    het register en de syntactische structuren van Kamerdebatten en
    offici le verslagen.
  </Role>

  <Task>
    Genereer nieuwe zinnen die lijken op parlementaire taal, op
    basis van aangeleverde voorbeelden. Zorg dat elke zin minimaal
      n   organisatie-entiteit bevat, gelabeld met een van de
    volgende categorie n:
```

```
    <Labels>
      <Label code="GEN">Alle instanties die onderdeel zijn van of direct opereren namens de overheid, waaronder uitvoerende, wetgevende en toezichthoudende organen op nationaal en regionaal niveau.</Label>
      <Label code="PAR">Politieke partijen</Label>
      <Label code="INT">Internationale organisaties</Label>
      <Label code="NGO">Niet-gouvernementele organisaties (zoals actiegroepen of belangenorganisaties)</Label>
      <Label code="BUS">Bedrijven of bedrijfsverenigingen</Label>
      <Label code="EDR">Onderwijsinstellingen en onderzoeksorganisaties</Label>
    </Labels>
  </Task>

  <OutputFormat>
    Genereer de zinnen in het volgende JSON-formaat:

    [{{
      "data": {{
        "text": "HIER_DE_VOLLEDIGE_ZIN"
      }},
      "annotations": [
        {{
          "result": [
            {{
              "value": {{
                "text": "HIER_DE_ENT_TEXT_VAN_ENTITEIT",
                "labels": [
                  "HIER_HET_LABEL_VAN_ENTITEIT"
                ]
              }},
              "from_name": "label",
              "to_name": "text",
              "type": "labels"
            }}
          ]
        }}
      ]
    }}]
  </OutputFormat>

  <Examples>
    <Description>Voorbeeldzinnen uit Staten-Generaal handelingen:</Description>
    <Content>{output}</Content>
  </Examples>
</SystemPrompt>
""".format(output=output)

user_p = """
Genereer 25 formele parlementaire zinnen die qua stijl overeenkomen met de voorbeeldzinnen in de system prompt. Zorg dat elke zin minstens n organisatie-entiteit bevat uit de categorie n: GEN, PAR, INT, NGO, BUS of EDR.

```

```
Geef het resultaat terug in exact het gespecificeerde JSON-formaat.
"""
```

D Training Parameters WorgBERT

```
# -------------------------------------
# 8. Training Arguments
# -------------------------------------
training_args = TrainingArguments(
    output_dir="./WooBERT_results",
    eval_strategy="epoch",
    save_strategy="epoch",
    num_train_epochs=3,
    per_device_train_batch_size=16,
    per_device_eval_batch_size=16,
    learning_rate=2e-5,
    weight_decay=0.01,
    logging_dir="./logs",
    push_to_hub=False
)
```