HAW
HAMBURG

MASTERARBEIT

Synthetic Data Strategies for Clinical Named Entity Recognition

Improving Transformer-based Phenotype Detection in German Medical Text through Data Design

vorgelegt am 31. Dezember 2025
Cédric Uden

Erstprüferin: Prof. Dr. Marina Tropmann-Frick
Zweitprüfer: Prof. Dr. Kai von Luck

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG
Department Informatik
Berliner Tor 7
20099 Hamburg

Abstract

This thesis addresses the scarcity of annotated resources in German clinical text mining by developing synthetic data strategies for Named Entity Recognition (NER) of Human Phenotype Ontology (HPO) terms. Using Large Language Models (LLMs), the study evaluates the impact of generative request design on downstream model performance and compares commercial systems against open-weight models. Additionally, dataset shape alignment techniques are explored to align synthetic data distributions with a real-world Gold Standard and assess their impact on NER performance. Results demonstrate that models trained exclusively on synthetic data achieve competitive F1 scores, outperforming gazetteer-based baselines. While commercial models handle complex instructions more effectively, open-weight models prove to be viable alternatives when queries are appropriately decomposed. Ultimately, the research confirms that informed synthetic data generation is a robust solution for overcoming data limitations in non-English biomedical NER.

Zusammenfassung

Diese Arbeit befasst sich mit der Knappheit annotierter Ressourcen im Bereich des Deutschen Klinischen Text Mining, indem sie synthetische Datenstrategien für die Named Entity Recognition (NER) von Human Phenotype Ontology (HPO)-Begriffen entwickelt. Unter Verwendung von Large Language Models (LLMs) bewertet die Studie den Einfluss des generativen Anfrage-Designs auf die Modellleistung und vergleicht kommerzielle Systeme mit Open-Weight-Modellen. Darüber hinaus werden Dataset-Shape-Alignment-Techniken untersucht, um synthetische Datenverteilungen mit einem realen Goldstandard abzugleichen und deren Einfluss auf die NER-Leistung zu bewerten. Die Ergebnisse zeigen, dass Modelle, die ausschließlich mit synthetischen Daten trainiert wurden, konkurrenzfähige F1-Scores erzielen und Gazetteer-basierte Baselines übertreffen. Während kommerzielle Modelle komplexere Anweisungen effektiver handhaben, erweisen sich Open-Weight-Modelle als praktikable Alternativen, wenn die Anfragen entsprechend angepasst werden. Letztendlich bestätigt die Forschung, dass eine ausgereifte Generierung synthetischer Daten eine robuste Lösung zur Überwindung von Datenbeschränkungen in der nicht-englischen biomedizinischen NER darstellt.

Table of Contents

List of Figures

List of Tables

Assistive Technologies

Tool	Description
GitHub Copilot	Line completion & rephrasing
Perplexity Pro	Literature research
Google Gemini	Proofreading & translation
DeepL	Translation
Linguee.com	Translation
Microsoft Word	Grammar and spell checking

1 Introduction

The goal of this project is to research and develop the fundamentals for a named entity recognition (NER) pipeline for a clinical application that identifies candidate diseases based on phenotypic features. It aims to provide a modern replacement for the Phenomizer[1] presented by Köhler et al. (2009), a web-based tool that relies on rule-based NER and is now being superseded by contemporary machine learning (ML)-based approaches (Seow et al., 2025). Phenotypic information is expressed using the widely adopted Human Phenotype Ontology (HPO) (Gargano et al., 2024), whose entities are commonly referred to as *HPO terms* or *HPO concepts*.

The clinical motivation for such a system is rooted in the nature of medical data and current information extraction practices. More than 80% of data in healthcare is contained in unstructured text (Panahi, 2025; Rohanian et al., 2025), such as clinical notes, discharge summaries and electronic health records (EHRs). Extracting relevant information from these sources is therefore essential for downstream tasks like decision support, cohort identification and rare disease research. At the same time, the literature reports substantial growth in publications related to NER (Keraghel et al., 2024), reflecting the increasing importance of text mining in clinical and biomedical research. However, there remains a lack of annotated training data, especially in specialised and non-English medical domains (Frei & Kramer, 2023a; Keraghel et al., 2024). This data scarcity is a major barrier to high-quality NER models and motivates alternative strategies such as synthetic data generation and machine translation.

Against this background, the project seeks to automate the extraction of phenotypic features from unstructured text in EHRs using modern machine natural language processing (NLP) techniques. This work concentrates on the synthetic data generation used for training and development of the NER model responsible for identifying phenotypic concepts in EHRs. After a brief introduction to the medical application context, the methodology and implementation of the synthetic data creation process is presented, followed by a detailed description of the datasets used for training and evaluation.

Research on similar techniques to support text mining with NER in clinical research is well advanced in English-language settings and has produced several notable systems, including *PhenoTagger* by Luo et al. (2021), *PhenoBERT* by Feng et al. (2023) and *PhenoGPT* by Yang et al.

[1] https://hpo.jax.org/tools/phenomizer

(2024). These approaches demonstrate the potential of combining domain-specific resources with modern ML and NLP. In contrast, low-resource languages such as German require different strategies (Frei & Kramer, 2023b), as they lack both the breadth of annotated corpora and the ecosystem of tools available for English. This thesis therefore adapts established ideas to a non-English NLP context and investigates how machine translation and synthetic data generation can mitigate data scarcity and enable effective training of medical NER models in such settings.

1.1 Motivation

In clinical practice, a robust NER model for phenotypic features can help physicians link symptoms to diseases, especially in rare disease diagnostics. By automatically extracting and standardising phenotypic information from unstructured EHRs, it enables faster comparison of patient profiles and supports the detection of recurring patterns that may otherwise remain hidden in free-text documentation. This is particularly valuable for rare conditions, in which timely recognition of relevant symptom constellations is essential for patient outcomes.

The goal of the system is to link phenotypic features across patients in a consistent and reproducible way. To this end, the model should perform most of the labour-intensive extraction and normalisation steps so that physicians can focus on interpreting and validating the automatically generated suggestions. In this workflow, the model serves as a high-throughput pre-screening tool, while clinical experts retain responsibility for critical appraisal and decision-making.

1.2 Research Questions

- RQ1: What is the impact of data-synthesis generative request complexity on NER model quality?
- RQ2: To which extend can data-synthesis metrics be used to predict NER model performance?
- RQ3: Can open-weight models match top closed-source systems for data-synthesis in the non-English medical domain?
- RQ4: Does synthetic-only data suffice to train non-English medical NER?

2 Background

This chapter introduces the HPO and its structure as well as an overview of its key metrics. It also discusses the challenges of its structure, the data scarcity presented by the usage of the German language and the strategies to address these issues.

2.1 HPO Terms

Figure 2.1 is an excerpt of the phenotypes in the ontology, which can be represented as a directed acyclic graph. At the time of the start of this project, there is a total of 18 425 nodes in the ontology. The ontology is under ongoing development and subject to updates, therefore new nodes are to be expected in the near future. Currently, this project uses version 2025-03-03[1] of the ontology.

Each node is described by its label and its id. The bottom nodes are all leaves of the graph, while the top-most directly connected node is four nodes away from the root node ("Phenotypic abnormality / HP_0000118")[2].

Furthermore, the child nodes of the root represent key categories for downstream classification in this project. These special 23 nodes will be referred to as the parent category or superclass of the HPO term.

Figure 2.1 illustrates both the complexity and the advantages of using the Human Phenotype Ontology. It demonstrates how fine-grained the distinctions between concepts can be, with their interpretation possibly depending on external variables such as the age of the patient. At the same time, the networked nodes of the ontology facilitate systematic relationship analysis within its tree-based framework.

[1] https://github.com/obophenotype/human-phenotype-ontology/releases/tag/v2025-03-03

[2] Technically, the node "All" (HP_0000001) is the root, but this project focuses solely on phenotypic abnormalities, disregarding any nodes and their branches stemming from the clinical modifier, past medical history, biospecimen phenotypic feature, mode of inheritance, blood group and frequency nodes.

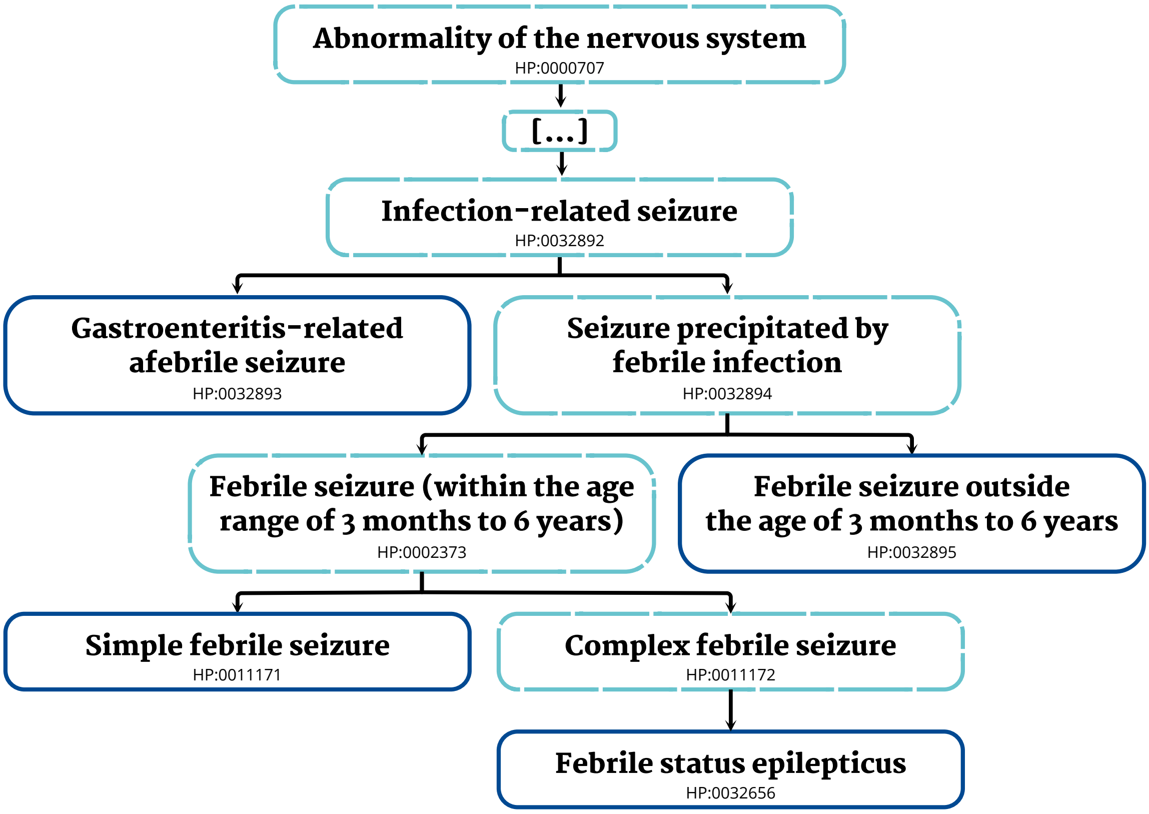

Abnormality of the nervous system
HP:0000707
[...]
Infection-related seizure
HP:0032892
Gastroenteritis-related afebrile seizure
HP:0032893
Seizure precipitated by febrile infection
HP:0032894
Febrile seizure (within the age range of 3 months to 6 years)
HP:0002373
Febrile seizure outside the age of 3 months to 6 years
HP:0032895
Simple febrile seizure
HP:0011171
Complex febrile seizure
HP:0011172
Febrile status epilepticus
HP:0032656

Figure 2.1: Excerpt of the HPO phenotypic abnormalities tree.

2.1.1 Data Scarcity and Translation Strategy

The HPO Internationalisation Effort (HPOIE) [3] is an existing translation effort for HPO terms. It features different languages, but the coverage of German data is lacking. Currently, there are only 3470 of 18 425 terms that have been translated.

To overcome this problem for this research, this project will compare the results of two machine translation strategies. One full dataset translation is based on the output of generative requests using Mixtral8x22b (Ollama, 2024). The model was chosen for its widespread availability and its strong performance in translation tasks. Additionally, DeepL, a widely adopted commercial service which specialises in machine translation tasks is also tested. This has been done to compare the quality of the translations from DeepL, which translates one property at a time, to a full translation of all properties enhancing its context.

[3] https://obophenotype.github.io/hpo-translations/

2.2 Gold Standard Dataset

For the success of this project, it is crucial to have a high-quality gold standard (GS) dataset to evaluate the performance of the NER models for real-world performance (IBM Think, 2024). The GS dataset has been developed by medical experts purposefully for this project. It is based on fictional medical cases based on their real-world experience. It has been designed with data protection in mind, so no real patient data has been used and the dataset can be shared in an academic context.

The dataset contains 26 EHRs in German. Each EHR contains the following sections: *Aktuelle Diagnose*, *Dauerdiagnosen* and *Körperlicher Untersuchungsbefund*.

Table 2.1 shows key statistics of the GS dataset.

	Count	Proportion
Total sentences	318	1.00
Total HPO terms	203	
Total unique HPO terms	150	0.74
Average labels per sentence	0.64	
Sentences without labels	172	0.54
Sentences with one label	115	0.36
Sentences with multiple labels	31	0.10

Table 2.1: GS dataset key characteristics.

The frequency of n-grams is an important aspect to consider when evaluating the quality of synthetic data (Mishra et al., 2020). The distribution of 3-grams (or trigrams) is analysed multiple times in this project. Figure 2.2 shows the trigrams of the original text data from the GS dataset. This analysis helps to understand the diversity and patterns in the text data to compare this to the synthetic data in downstream tasks.

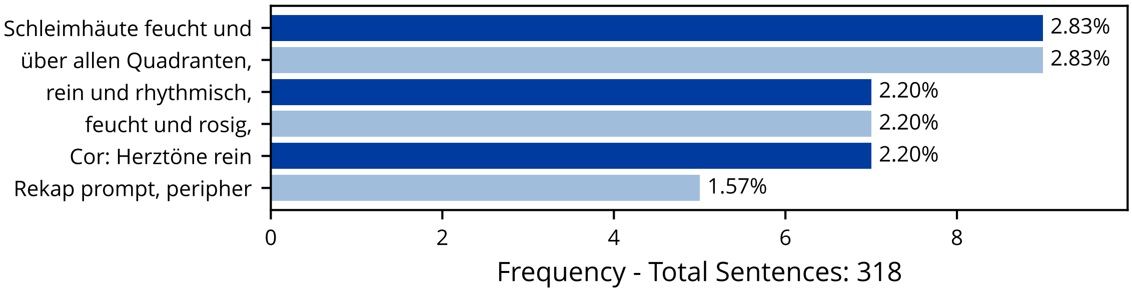


Figure 2.2: Top 6 trigrams in the gold standard dataset.

3 Related Work

This chapter presents an overview of related work in the fields of medical information extraction, synthetic data generation for medical unstructured text based on large language models (LLMs), and machine translation. Additionally, it examines specific studies on medical information extraction in German, with a focus on HPO entity recognition in English and clinical named entity recognition in German. The review highlights key findings, methodologies, and gaps in the existing literature, providing a foundation for the subsequent methodology and analysis presented in this thesis.

3.1 Large Language Models

Large language models have achieved strong performance across many general-domain benchmarks, yet their effectiveness remains limited for non-English languages, particularly in specialized domains such as medicine. Studies have consistently shown that multilingual and cross-lingual setups lag behind their English counterparts in both intrinsic and downstream tasks, with performance drops especially pronounced for languages like German and for domain-specific terminology (King, 2024; Rohanian et al., 2025; Seeha et al., 2025; Wang et al., 2025). These gaps are often attributed to imbalances in training data, reduced availability of high-quality domain corpora, and a lack of robust evaluation resources outside English.

The challenges are further exacerbated in clinical text, where language is highly specialized, context-dependent, and frequently noisy. Clinical documentation contains abbreviations, idiosyncratic formatting, institutional jargon, and evolving terminology that are not well represented in the general web-scale corpora used to train most LLMs (Seeha et al., 2025). For German clinical texts in particular, recent work has highlighted difficulties in reliably recognizing medical entities and capturing fine-grained clinical concepts, even when using state-of-the-art transformer-based architectures (Rohanian et al., 2025; Seeha et al., 2025). These studies emphasize that clinical NLP in non-English settings remains a challenging problem and underline the need for targeted resources, domain-adapted models, and evaluation frameworks tailored to clinical use cases.

3.2 Synthetic Data Generation for Medical Text

Synthetic data generation has emerged as a key strategy in biomedical research, particularly for addressing data scarcity, privacy concerns, and the challenges of working with sensitive clinical information. In a recent scoping review, Rao et al. (2025) report that 78% of applications of synthetic data in this domain focus on unstructured text, underscoring the central role of clinical narratives, reports, and other free-text documents in modern medical research. This trend reflects both the richness of information encoded in text and the difficulty of obtaining sufficiently large, shareable, and well-annotated corpora from real patient data.

Within this broader context, synthetic data has been highlighted as a promising tool for advancing research in rare diseases, where real-world data is particularly sparse. Mendes et al. (2025) argue that synthetic data can help bridge critical data gaps and has the potential to "revolutionise" rare disease research by enabling the development and evaluation of models that would otherwise be infeasible due to limited sample sizes. The quality of such synthetic data is crucial: as shown by Barr et al. (2025), LLMs can generate highly realistic data that respects appropriate value ranges, performs correct calculations (e.g., body mass index) without explicit formulas, and preserves definitional boundaries of clinical parameters even without detailed instructions. These findings suggest that, under suitable conditions, LLMs can produce synthetic clinical text and associated structured data that closely mirror real-world distributions.

Several studies emphasize that synthetic medical text is likely to play an increasingly important role in future downstream analyses. For instance, Alshaikhdeeb et al. (2025) explicitly argue that medical synthetic text generation is expected to become central to a range of clinical NLP applications. Similarly, Doshi and Bhattacharyya (2024) highlight the particular promise of synthetic data for multilingual NLP, where annotated resources are especially scarce and unevenly distributed across languages. In such settings, synthetic corpora can help mitigate resource imbalances, support the development of models for underrepresented languages, and enable more comprehensive evaluations across diverse linguistic contexts.

At the methodological level, generative request-based data augmentation (DA) with LLMs has gained attention as a flexible and scalable approach to synthetic data generation. Huang et al. (2025) note that these methods are becoming increasingly prominent, as they allow researchers to generate task-specific examples with controlled attributes using natural-language instructions. However, empirical work also cautions against over-reliance on synthetic data alone. For clinical named entity recognition in low-resource languages, Kamath and Vajjala (2025) show that, while LLM-generated data can substantially help train NER models from scratch, there remains a critical need for manually annotated gold-standard test sets to ensure robust and meaningful benchmarking.

Finally, the effectiveness of synthetic data is highly dependent on task, language, and model configurations. In a comprehensive study across Indian languages, Chitale et al. (2025) propose a framework for data generation, quality assessment, and downstream evaluation that explicitly moves beyond English-centric norms. Their results indicate that no single recipe for synthetic data generation works best across all tasks and models. Instead, strategies must be tailored to the linguistic, cultural, and domain-specific characteristics of each setting. This underscores the importance of context-sensitive design of synthetic data pipelines and suggests that future work should focus not only on improving the realism of generated text, but also on aligning generation methods with concrete clinical and linguistic requirements.

3.3 Machine Translation

Enis and Hopkins (2024) demonstrate that, for certain language pairs and under carefully designed generative query or fine-tuning setups, LLMs can occasionally outperform dedicated neural MT systems. This is particularly evident when translation tasks require leveraging broader contextual knowledge or handling ambiguous phrasing, where the capacity of the LLM for flexible reasoning and discourse-level understanding can compensate for weaker surface-level alignment.

At the same time, conventional machine translation (MT) pipelines remain highly valuable as tools for creating training data in multilingual NLP. Wang et al. (2025) successfully use MT to generate data for training multilingual models, showing that high-quality automatic translation can be an effective mechanism to transfer annotations and expand datasets into low-resource languages. Similarly, King (2024) employ MT to translate texts for tuning multilingual classifiers, illustrating how translation can serve as a bridge to adapt models to additional languages without requiring extensive language-specific annotation efforts.

Within the clinical domain, several studies highlight both the potential and the limitations of using MT and LLMs for medical translation. Kong et al. (2025) show that GPT-4 can outperform Google Translate for Spanish medical texts, supporting the view that advanced LLM-based systems can help mitigate existing gaps in the translation of clinical documents. However, a broader comparative analysis by Li et al. (2025) finds that traditional MT tools still offer stronger surface-level alignment, i.e., more literal, token-level correspondence between source and target, while LLMs provide greater contextual flexibility and better handling of implicit information. These findings underscore that neither paradigm is universally superior and highlight the need for domain-specific evaluation methods, as well as the continued importance of human oversight to ensure safe and responsible use of AI-based translation in healthcare.

In parallel, there is growing interest in using MT and LLMs to support domain-specific resources, such as medical ontologies. Noll et al. (2025) investigate the translation of HPO terms into German, comparing GPT-3.5 with DeepL. Their statistical analysis reveals no significant differences in mean human ratings between the two systems, indicating comparable translation quality for this specialized terminology. This suggests that both LLM-based and traditional MT systems can achieve robust performance on carefully scoped medical translation tasks, provided that appropriate quality control measures are in place.

Finally, the broader capabilities of modern LLMs used for translation are also reflected in their strong performance on general benchmarks. For example, the Mixtral family of models achieves high scores on the massive multitask language understanding (MMLU) benchmark (Jiang et al., 2024; Ollama, 2024), indicating substantial general-domain knowledge and reasoning abilities. While such results do not directly measure translation quality, they suggest that models with strong general capabilities may be particularly well-suited for complex multilingual and domain-specific applications, including medical MT, where both linguistic competence and background knowledge are essential.

3.4 Medical Information Extraction

Recent developments in medical information extraction have highlighted several key challenges and advances, particularly in the context of rapidly evolving medical knowledge and the adoption of transformer-based architectures. One notable example is the impact of the COVID-19 pandemic on clinical NER systems. As shown by Kühnel and Fluck (2022), the emergence of numerous novel entities, such as previously unseen disease names, treatments, and clinical concepts, posed significant difficulties for existing models. Systems trained on pre-pandemic corpora struggled to recognize and correctly classify these new terms, underscoring the limitations of static training data in a dynamic clinical environment. This phenomenon illustrates how shifts in medical practice and terminology can quickly degrade the performance of deployed models, emphasizing the need for continual model adaptation and domain-specific updating strategies.

Parallel to these challenges, recent work has demonstrated the strong benefits of Bidirectional Encoder Representations from Transformers (BERT)-based architectures for clinical NER. Studies such as Builtjes et al. (2025) and Seeha et al. (2025) show that transformer models, pre-trained on large general-domain corpora and subsequently adapted to medical text, can substantially outperform earlier approaches. Their contextualized representations enable more accurate disambiguation of technical terms and better handling of long-range dependencies, which are common in clinical documentation. Nevertheless, the COVID-19 case highlights that even these state-of-the-art models are vulnerable to domain drift and emerging terminology, reinforcing the importance of continuous domain adaptation, regular re-training on up-to-date data, and the integration of mechanisms to efficiently incorporate novel medical entities into existing information extraction pipelines.

3.4.1 Medical Information Extraction in German

Recent work on medical information extraction in German has consistently shown that the field lags behind English-language clinical NLP in terms of both resources and model performance. As Richter-Pechanski et al. (2023) emphasize, there is a notable scarcity of large, high-quality, and freely distributable German clinical corpora. Legal and ethical constraints on sharing patient data further exacerbate this situation, limiting the availability of training and evaluation datasets and making it difficult to develop and benchmark robust models. Compared to the extensive ecosystem of English clinical datasets and benchmarks, German clinical NLP therefore operates under structurally more restrictive conditions, which is reflected in slower progress and lower performance across many tasks.

Within these constraints, recent studies have nonetheless demonstrated substantial advances, particularly for clinical NER. Seeha et al. (2025) present a systematic evaluation of German medical NER models using existing datasets, showing that transformer-based architectures can achieve strong intrinsic performance when sufficient annotated data is available. Their work provides valuable insights into model design, pre-training choices, and domain adaptation strategies for German clinical text. However, their evaluation is largely limited to intrinsic metrics such as token- and entity-level precision, recall, and F1 on established benchmarks. There is currently little evidence on how these improvements translate to downstream clinical tasks, such as cohort identification or decision support in real-world hospital workflows. This gap between intrinsic evaluation and practical utility highlights an important direction for future research in German medical information extraction.

3.4.2 HPO Entity Recognition in English

Related literature on HPO entity recognition is currently focused on English, with relatively few studies addressing other languages. Recent work has focused on combining traditional ontology-based methods with modern deep learning and LLM-based approaches. Luo et al. (2021) introduce PhenoTagger, a hybrid system that integrates dictionary- and rule-based components with machine-learning techniques to recognize and normalize phenotype mentions to HPO concepts. By leveraging both curated lexical resources and statistical models, PhenoTagger achieves robust performance across heterogeneous clinical and biomedical texts, illustrating the benefits of hybrid architectures in this domain.

Building on transformer-based language models, Feng et al. (2023) propose PhenoBERT, which uses contextualized representations from BERT-style encoders to improve phenotype recognition. Their approach combines deep biomedical language modelling with ontology-aware post-processing, leading to substantial gains in recall and overall F1 compared to earlier systems. PhenoBERT demonstrates that pre-training on large biomedical corpora, followed by fine-tuning on phenotype annotation tasks, can capture subtle linguistic cues and domain-specific terminology that are crucial for accurate HPO tagging.

More recently, Yang et al. (2024) present a two-stage framework that integrates specialized encoders and large language models. Their system, PhenoBCBERT + PhenoGPT, first employs a domain-adapted BERT variant to perform high-precision phenotype recognition and then uses an LLM-based component to refine and disambiguate phenotype candidates. This design aims to combine the strong token-level tagging capabilities of transformer encoders with the broader reasoning and paraphrase understanding of LLMs, particularly for challenging or ambiguous phenotype mentions.

Complementing these approaches, Shlyk et al. (2024) propose REAL, a retrieval-augmented entity linking framework for phenotype recognition. REAL explicitly integrates retrieval-augmented knowledge from related biomedical texts into the entity linking process, allowing the model to access up-to-date and context-rich information when mapping text spans to HPO concepts. By coupling neural networks with retrieval-augmented ontology knowledge, REAL addresses limitations of purely local models and improves robustness for rare or previously unseen phenotypic expressions. Together, these methods illustrate a clear trend toward hybrid and retrieval-augmented architectures that tightly couple LLM-style representations with ontology-based resources for HPO entity recognition in English.

3.4.3 Clinical Named Entity Recognition in German

Taking one step back, related work on clinical named entity recognition in German is hereby reviewed, as HPO entity recognition is a specialized sub-task of clinical NER and not covered in detail in non-English contexts. Overall, existing studies consistently indicate that clinical NER in German remains considerably more challenging than in English, particularly in highly specialized technical domains such as medicine. As Bressem et al. (2024) note, the scarcity of large, high-quality German medical corpora and the relative under-representation of German in general-domain pre-trained models lead to noticeable performance gaps. Medical terminology further compounds these difficulties: complex compound nouns, domain-specific abbreviations, and idiosyncratic spelling conventions reduce the effectiveness of models that were primarily optimized for English or for non-clinical text. Consequently, many state-of-the-art approaches in German clinical NER still lag behind comparable English systems, both in terms of benchmark scores and robustness in real-world settings.

This performance gap is also reflected in concrete evaluation results. For instance, Diaz Ochoa et al. (2024) report entity-type F1 scores in the range of 0.38 for German ICD-10 NER, indicating that correctly identifying and classifying diagnostic codes remains a difficult task. Such figures underline that even for relatively well-defined label spaces like ICD-10, current models struggle with consistent recognition of clinically relevant entities in German texts. These limitations have practical implications, as accurate ICD-10 extraction is central to downstream tasks.

In addition to diagnostic coding, domain-specific sub-tasks such as drug NER have received growing attention. Frei and Kramer (2023b) contribute an annotated dataset for German drug NER, providing an important resource for training and evaluating models that aim to recognize medication names. Their work highlights both the complexity of drug terminology—encompassing brand names, generic names, dosages, and formulation details, as well as the importance of carefully curated gold-standard corpora for empirical progress. Nevertheless, the overall ecosystem of shared resources for German clinical NER remains relatively small, limiting the comparability of results and systematic benchmarking across institutions and domains.

Despite these challenges, recent studies demonstrate that strong performance is possible when models are carefully adapted and evaluated on available resources. Rohanian et al. (2025) present an extensive study of different NLP tasks in the oncology domain, reporting high F1 scores on existing benchmark datasets. Their cross-lingual specialist models show that domain-focused pre-training and fine-tuning can yield competitive results even for complex clinical tasks. However, as they themselves and related work such as NLPIE Research (2025) emphasize, many of these benchmarks are based on datasets that overlap substantially with the data used for model training or tuning. This raises concerns about overfitting and limits the conclusions that can be drawn regarding generalization to truly unseen clinical environments. Consequently, while these results are promising, they also underscore the need for more diverse, independently constructed evaluation sets and for rigorous experimental designs that better reflect real-world deployment scenarios in German clinical NER.

4 Methodology

This chapter outlines the methodological foundation underlying the conducted experiments and analyses. After a brief overview of the project, it describes the preparation of the data and the tagging scheme adopted for sequence labelling, the evaluation metrics used to quantitatively assess model performance, and the procedures employed to assess the quality of synthetic data. Together, these components establish a transparent and reproducible framework for investigating the research questions addressed in this thesis.

4.1 Project Overview

This section outlines the overall structure of the research and situates the individual components of the methodology in relation to one another. The work is organised around the goal of generating high-quality synthetic clinical text in German and leveraging it to improve NER performance for HPO term recognition and linking. By integrating ontology-driven resources, machine translation methods and generative models, the project seeks to address the data scarcity problem in German-language clinical narratives.

Figure 4.1 provides an overview of the methodology employed in this project. The primary objective is to develop and evaluate generative queries for creating synthetic data that enhances the training of NER models for HPO term recognition and linking. The process begins with preparing and analysing the GS dataset introduced in Section 2.1, which serves as the benchmark for downstream evaluation. In parallel, the HPO ontology described in Section 2.1 is processed and translated to provide structured terminology resources in German. These resources then form the foundation for designing generative requests and generative strategies used to create synthetic sentences that are both clinically plausible and rich in phenotypic detail.

To evaluate the performance, the various datasets generated through different strategies are used to train NER models. By comparing models trained on these different data configurations, it becomes possible to quantify the contribution of synthetic data to the recognition and linking of HPO terms. The evaluation setup is designed to be as close as possible to realistic clinical use cases, relying on the manually curated annotations of the GS dataset as the reference.

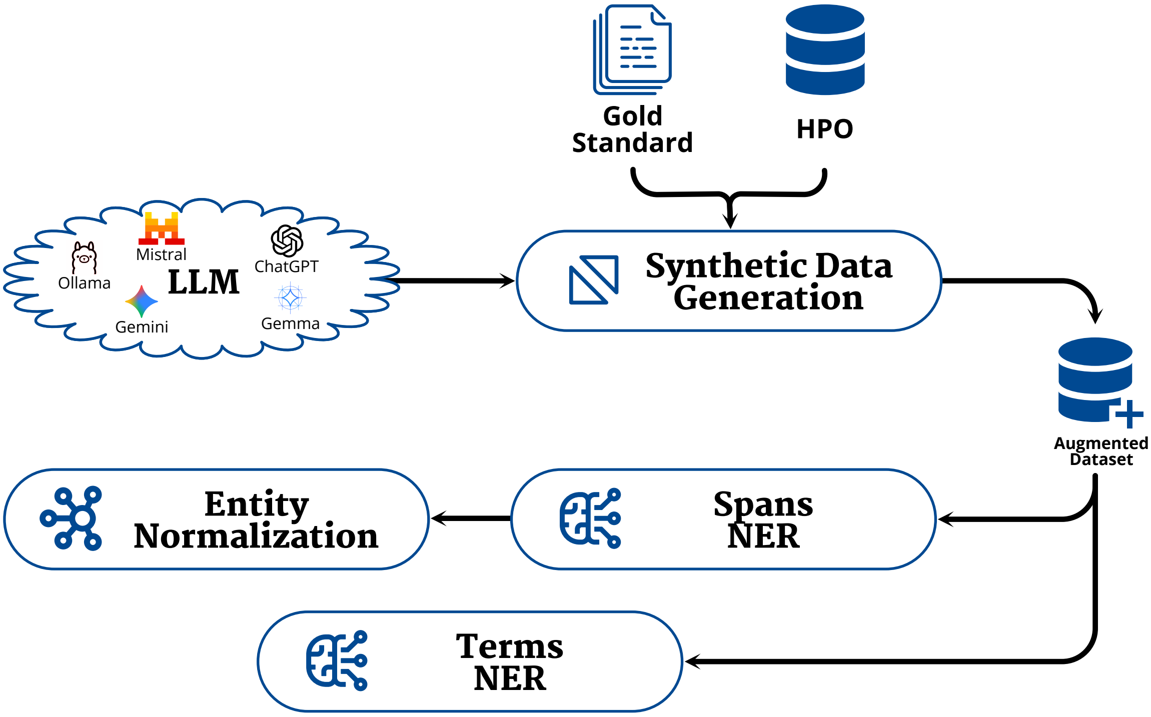


Figure 4.1: Methodology overview of the project.

The evaluation is split into two key approaches: the recognition of any phenotype in text, referred to as *Spans NER* or simply *spans*, and the recognition of the correct HPO term linked to the identified span, referred to as the *Terms NER* evaluation. *Spans NER* focuses on the ability of a model to detect relevant text segments in clinical documents that correspond to phenotypic abnormalities, regardless of the specific ontology entry. This perspective emphasises coverage and sensitivity in identifying all potentially relevant mentions in the text. For both evaluation types, the *medbert* model introduced by Bressem et al. (2024) is used as the base NER architecture.

In contrast, *Terms NER* addresses the correctness of mapping each identified span to a specific HPO concept. This requires both lexical and semantic understanding, as multiple closely related HPO terms may be plausible candidates for a given mention. The evaluation in this setting not only considers whether a phenotype has been detected but also whether the assigned concept identifier accurately reflects the intended clinical meaning. These two approaches are complementary, as they focus on different aspects of the NER task: whereas *Spans NER* aims to identify relevant text spans, *Terms NER* ensures that the identified spans are correctly linked to the appropriate HPO terms.

To link the identified spans to the correct HPO terms, the task of named entity normalization (NEN) is employed. Within this project, the predefined entities are provided by the HPO ontology. As the downstream systems that might be developed based on this project will require accessible human validation, it is important to consider cognitive and attentional characteristics of human reviewers. Empirical evidence suggests that humans find it easier to correct wrongly proposed entities (false positives) than to identify missing entities (false negatives) (Xu et al., 2023). Consequently, the methodological design and evaluation metrics in the later chapters explicitly take into account this asymmetry between error types. Emphasis is placed on configurations that reduce false negatives in both *Spans NER* and NEN, even at the cost of a moderate increase in false positives, as this trade-off better aligns with real-world validation workflows and the practical needs of clinicians and data curators.

4.2 Data Preparation and Tagging Scheme

Robust data preparation is a prerequisite for reliable sequence labelling and information extraction. In this work, the raw data are first cleaned and validated before being segmented into tokens. This includes steps such as removing irrelevant artefacts and handling formatting consistently.

To enable supervised learning for named entity recognition and related sequence labelling tasks, an IBO/IOB tagging scheme is adopted (Keraghel et al., 2024). In this scheme, each token in a sentence is assigned a label that indicates whether it begins an entity span (B-), lies inside an entity span (I-), or is outside any entity (O). This structured labelling allows models to learn not only which tokens correspond to entities, but also how entities extend across multiple tokens. The IBO/IOB scheme strikes a balance between expressive power and annotation complexity, providing a widely used and well-understood standard for encoding entity boundaries in natural language processing tasks.

To enable fair and efficient comparison across multiple LLM-generated datasets, all experiments in this project are conducted on a consistent subset of the HPO ontology. This subset comprises 300 HPO terms drawn from the full ontology: 150 terms that correspond to those appearing in the GS dataset and an additional 150 terms randomly selected from the remaining ontology. This design guarantees that all models are evaluated on an identical set of terms, thereby supporting systematic and comparable assessment of their performance.

4.3 Evaluation Metrics

The performance of the models is evaluated using standard metrics for classification and sequence labelling, namely precision, recall, and F1-score. Precision measures the proportion of predicted entities that are correct, thus reflecting the tendency of the system to avoid false positives. Recall quantifies the proportion of gold-standard entities that are successfully identified by the system, capturing its ability to avoid false negatives. The F1-score, defined as the harmonic mean of precision and recall, provides a single aggregate measure that balances these two aspects, which is particularly informative when dealing with imbalanced classes or differing error profiles across models.

For a more nuanced assessment of sequence labelling performance, the Nervaluate framework is employed (Segura-Bedmar et al., 2013). Nervaluate provides detailed evaluation of named entity recognition systems by supporting multiple matching strategies, such as exact, partial, strict, and generating comprehensive metrics across entity types (ETs). This enables a more fine-grained analysis of model behaviour beyond simple token-level accuracy, highlighting where models succeed or fail in correctly identifying entity spans.

In this work, the ET evaluation provided by Nervaluate is particularly valuable because it captures partial matches, which align well with the practical requirements of clinical phenotype recognition. It will be the only metric considered when referencing the Nervaluate results in the following chapters. In clinical contexts, identifying the core phenotypic concept is often more important than capturing the exact span boundaries. For example, both ”complex fever” and ”fever” successfully identify the underlying phenotype. This flexibility is especially beneficial for clinical applications, where partial matches can already provide substantial value through ontology-based disambiguation and downstream clinical decision support.

In addition, n-gram statistics are computed to compare local lexical patterns between real and synthetic text (Mishra et al., 2020). Differences in n-gram frequency distributions highlight whether common word sequences are reasonably preserved or whether implausible or overly repetitive patterns emerge. The goal is to generate diverse, novel sentences that still resemble the original data (Shen et al., 2023).

Finally, word embeddings are used to compare semantic representations derived from real and synthetic data (Manz et al., 2025). By projecting embeddings into a low-dimensional space using methods such as principal component analysis (PCA) one can examine whether semantically related terms occupy similar regions and whether the relative geometry of real and synthetic embedding spaces is preserved (Kim et al., 2024).

4.4 Gazetteer matching

Gazetteer-based NER systems utilize predefined lists of entities, known as gazetteers or dictionary based matching approaches, to identify and classify named entities within text (Anandika & Mishra, 2019). These systems operate by matching words or phrases in the input text against entries in the gazetteer. When a match is found, the corresponding ET from the gazetteer is assigned to the matched text. This simple yet effective approach is important to compare the merits of more complex models against a straightforward baseline that is hereby established.

In the context of clinical text and this HPO NER task, gazetteer-based systems can be particularly useful due to the specialized vocabulary. This is particularly relevant as this HPO NER task involves identifying medical terms and phenotypic abnormalities, which are well-represented in the HPO.

To assess the benefits of advanced generative query and LLM-based methods, this work employs a simple, knowledge-based baseline. Following Keraghel et al. (2024) and Weijers and Bloem (2025), the baseline uses a combination of dictionary lookup and string matching to identify clinical entities. Such approaches are comparatively easy to implement, interpretable, and efficient, but they often struggle with morphological variants, ambiguous terms, and context-dependent interpretations, all of which are common in clinical text.

Despite these limitations, dictionary lookup and string matching provide a useful lower bound for performance in clinical named entity recognition. By establishing a transparent and reproducible baseline, it becomes possible to quantify how much improvement is achieved by more sophisticated approaches, such as LLM-based generation, generative query extraction, or hybrid systems. Furthermore, this type of baseline is valuable for error analysis, because failure cases can often be traced back to missing synonyms, spelling variants, or context-dependent meanings, which can inform subsequent refinements of both the dictionary and the modelling pipeline.

Metric	**Score**
Precision	0.699
Recall	0.195
F1 Score	0.305
Nervaluate (ET) Precision	0.697
Nervaluate (ET) Recall	0.308
Nervaluate (ET) F1 Score	0.428

Table 4.1: Baseline evaluation of the gazetteer-based NER system on the gold standard dataset.

Table 4.1 summarizes the performance of this baseline system. The relatively high precision combined with low recall indicates that the method is conservative in its predictions: many predicted entities are correct, but a large fraction of true entities remains undetected. The still far from perfect precision can be attributed to non-pathological terms which are indeed correctly describing the HPO term but are not relevant in the clinical context, e.g., ”fever” in the sentence ”The patient has no fever.” These results confirm common expectations about dictionary-based systems and underline the potential for improvement through methods that can better exploit context and handle linguistic variability.

5 Generative Queries

This chapter examines generative query strategies for synthetic data generation in the context of clinical text and health-related named entity recognition. The chapter introduces a knowledge-based baseline method used for evaluation and comparison, before presenting the data augmentation setup, including the selection of foundation models. Together, these sections provide the conceptual and experimental foundation for the generative query approaches explored in this work.

Synthetic data generation has emerged as a central technique for addressing data scarcity, privacy constraints, and class imbalance in many machine learning scenarios. A recent scoping review by Kaabachi et al. (2025) provides an overview of different SDG methods, including classical statistical approaches such as GANs and VAEs, and more recent LLM-based approaches. The review emphasizes how SDG methods can be designed to preserve statistical properties of the original data while mitigating privacy risks, and it underscores the importance of rigorous evaluation of both utility and privacy leakage. Within this landscape, LLM-based SDG plays an increasingly important role, as it enables flexible, conditioned generation of domain-specific text with relatively low engineering overhead.

In parallel, Rao et al. (2025) survey synthetic data generation with a particular emphasis on open, reproducible workflows. Their review highlights that many existing SDG pipelines rely on proprietary systems, which poses challenges for transparency, long-term availability, and reproducibility of experiments. By contrast, the present work follows the direction outlined by Rao et al. (2025) and prioritizes non-proprietary models and reproducible setups wherever possible. This orientation is reflected in the choice of models and tooling in the data augmentation pipeline described later in this chapter.

5.1 Translation

High-quality translation is a crucial component when working with multilingual clinical corpora, especially where annotated resources are scarce in the target language. Recent studies have demonstrated that modern LLMs can deliver competitive or even state-of-the-art performance in translation tasks across a wide variety of language pairs. In particular, Thellmann et al. (2024) systematically evaluate multilingual LLMs and show that they can achieve strong performance not only on general-domain benchmarks but also on specialized and low-resource scenarios. Their work suggests that LLMs can serve as effective translation engines, often reducing the need for complex, task-specific machine translation pipelines.

In the context of this thesis, such findings motivate the use of LLM-based translation to create or align datasets across languages for downstream named entity recognition. Translation quality is not only relevant for human readability but also directly affects the consistency and coverage of clinical concepts, such as HPO terms, across languages. Careful evaluation of translation performance is therefore necessary to ensure that semantic content and medically relevant terminology are preserved.

Figure 5.1 illustrates an exemplary evaluation of translation quality in terms of its impact on downstream performance. The figure compares different translation settings and shows how LLM-based or translation via commercial APIs are comparable in terms of their impact on downstream entity recognition performance.

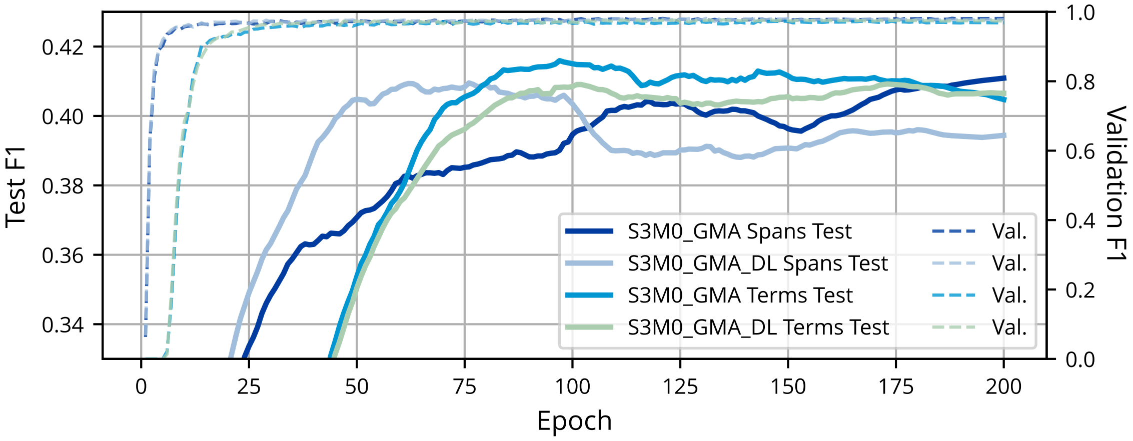


Figure 5.1: Results of downstream model evaluation using data generated based on differently translated texts.

The specific terminology of the datasets will be discussed in more detail in later sections. The DL suffix indicates that the data was translated using DeepL, while the other dataset was translated using an LLM-based approach. The results indicate that both translation methods yield similar downstream performance, suggesting that LLM-based translation is a viable alternative to commercial APIs in this context.

5.2 Data Augmentation

Data augmentation plays a central role in this work, both for enhancing the training signal of downstream models and for exploring how different generative query strategies affect the quality of synthetic examples. The augmentation pipeline relies on a diverse set of contemporary foundation models, chosen to balance openness, performance, and efficiency.

The set of models considered includes Gemini Pro 2.5 and Gemini Flash 2.0 (Comanici et al., 2025), which represent strong, general-purpose LLMs with different cost trade-offs. In addition, DeepSeek R1 (DeepSeek-AI et al., 2025) is designed to encourage explicit reasoning steps, which can be advantageous for generating structured clinical descriptions. Gemma 3 (Gemma Team et al., 2025) and Mistral Small (Mistral AI, 2025) provide lightweight, competitive open models that are suitable for on-premise or resource-constrained environments. The gpt-oss-20b model (OpenAI et al., 2025) is included as a strong open alternative in the 20B parameter range, while Phi-4 (Abdin et al., 2024) represents a compact model optimized for strong reasoning capabilities at a smaller scale. Together, this collection of models allows for systematic comparison of different architectures and training regimes in the context of synthetic clinical text generation and generative query.

Beyond the choice of models, the analysis of augmentation strategies can benefit from robust statistical techniques to smooth noisy metrics to reveal underlying trends. exponentially weighted moving average (EWMA), as discussed by Noor-ul-Amin et al. (2024), provide a simple yet effective tool for this purpose. EWMA methods place greater weight on more recent observations while still retaining information from earlier points, which is useful for monitoring evolving performance during iterative query refinement or model selection. By applying EWMA to evaluation scores over time or across query variants, it becomes easier to identify stable improvements and to avoid over-interpreting fluctuations caused by sampling noise or individual outlier runs.

5.2.1 Synthetic Sentences Containing One HPO Term

To begin the synthetic data generation (SDG) experiments, the generation process focuses on producing phrases that contain a single term only. Any dataset generated in this configuration is labelled as `S1M0`, where 'S' indicates a single term per sentence and the digit '1' denotes the version of the methodology used to generate the dataset. The 'M' in the label stands for multiple terms per sentence and foreshadows later experiments with more complex sentence configurations. In the current setup, `M0` indicates that no multiple terms per sentence are generated, making `S1M0` the most basic configuration. This simplest version serves as a starting point to assess how configuration complexity influences the results of the downstream evaluations.

`S1M0` - Initial Evaluation

The initial evaluation investigates whether synthetic-only data can be used to fine-tune clinical LLMs. For this purpose, a straightforward generative query configuration is selected to act as a proof of concept. The goal is to generate sentences that are syntactically and semantically plausible in a clinical context while containing exactly one annotated term per sentence. This configuration allows a controlled setup in which the impact of synthetic data on downstream NER performance can be isolated and studied systematically.

To establish a strong performance baseline, commercial models are included alongside open-source models. The commercial models are expected to have stronger generative capabilities in producing coherent clinical language, and therefore serve as a comparison basis for evaluating the quality of the synthetic data.

Although the underlying request is conceptually simple, it poses a considerable challenge for the LLM because of the extensive context required to handle the large number of distinct terms within a single query. As shown by Du et al. (2025), increasing context length alone can substantially degrade the generation quality of LLMs, particularly when the model must simultaneously track and integrate many separate items.

Codeblock 5.1 presents the generative request used to generate the synthetic data for the `S1M0` configuration. For rapid experimentation, the request contains the entire subset of the 300 HPO terms that were selected for this project.

For the `S1M0` configuration, several models are used to generate the synthetic datasets. They will be referenced as `S1M0`-based datasets. Each resulting dataset is additionally annotated with the abbreviation of the model used to generate the dataset. The following models are evaluated:

Codeblock 5.1: Generative query for `S1M0` dataset generation.

```
Generiere für jeden der folgenden Begriffe je einen kurzen, einen mittleren und
einen langen Satz, passend zum Kontext des Begriffs.
Nutze dafür klinische Sprache. Verwende für die Ausgabe die folgende Syntax:
- Packe jeden Satz zwischen exakt einem `<s>` Tag am Anfang und `</s>` Tag am Ende
- Packe jeden Begriff für jeden einzelnen Satz innerhalb des Satzes mit exakt
einem `<class>` Tag am Anfang und einen `</class>` Tag am Ende

Begriffe:
- Überdurchschnittliche Körpergröße
- Renale Agenesie
- Hydrocephalus
- <...>
```

- (P25) Gemini 2.5 Pro and (F20) Gemini 2.0 Flash (Comanici et al., 2025)
- (DSK) Deepseek R1 32b (Ollama, 2025a)
- (GMA) Gemma 3 27b (Ollama, 2025b)
- (MIS) Mistral Small 24B (Ollama, 2025d)
- (GPT) gpt-oss-20b (Ollama, 2025c)
- (PHI) Phi-4 14b (Ollama, 2025e)

To answer the research question 2 of how well automatic metrics can be used to gauge the quality of the generated data and its expected downstream performance, the Self-BLEU metric is first considered as a measure of diversity. Table 5.1 shows the Self-BLEU scores for the `S1M0` datasets generated by the different models.

Model	**Self-BLEU**
S1M0_P25	0.222
S1M0_F20	0.328
S1M0_DSK	DNF
S1M0_GMA	DNF
S1M0_MIS	DNF
S1M0_GPT	DNF
S1M0_PHI	DNF

Table 5.1: Self-BLEU scores of the `S1M0`-based datasets. Did not finish (DNF) indicates that the model failed to follow instructions to a degree that no valid evaluation was possible.

In this setting, the commercial models (P25, F20) clearly outperform the open-source models. They are the only ones that consistently generate syntactically correct sentences and adhere closely to the specified configuration, including the correct number of sentences and the proper use of tags. By contrast, the open-source models frequently struggle to follow the instruction format and cannot reliably handle the requested number of terms and sentences. As a result, all open-source models except the commercial ones receive a did not finish (DNF) for this configuration.

Furthermore, during the experiments the DSK and GPT models failed to accurately follow the requested output language for the synthetic sentences and therefore are excluded from further evaluations. Among the remaining models, the Self-BLEU score of the P25 model is significantly lower than that of the F20 model. Since lower Self-BLEU scores indicate higher variability between generated sentences, this suggests that P25 produces more diverse outputs. This behaviour aligns with the expectation that the larger and more capable P25 model should exhibit stronger generative diversity compared to F20.

Beyond Self-BLEU, a PCA analysis of sentence embeddings is performed to visualize the distribution of the generated sentences relative to the original sentences from the NER dataset. Figure 5.2 presents the resulting PCA projection for the S1M0 datasets.

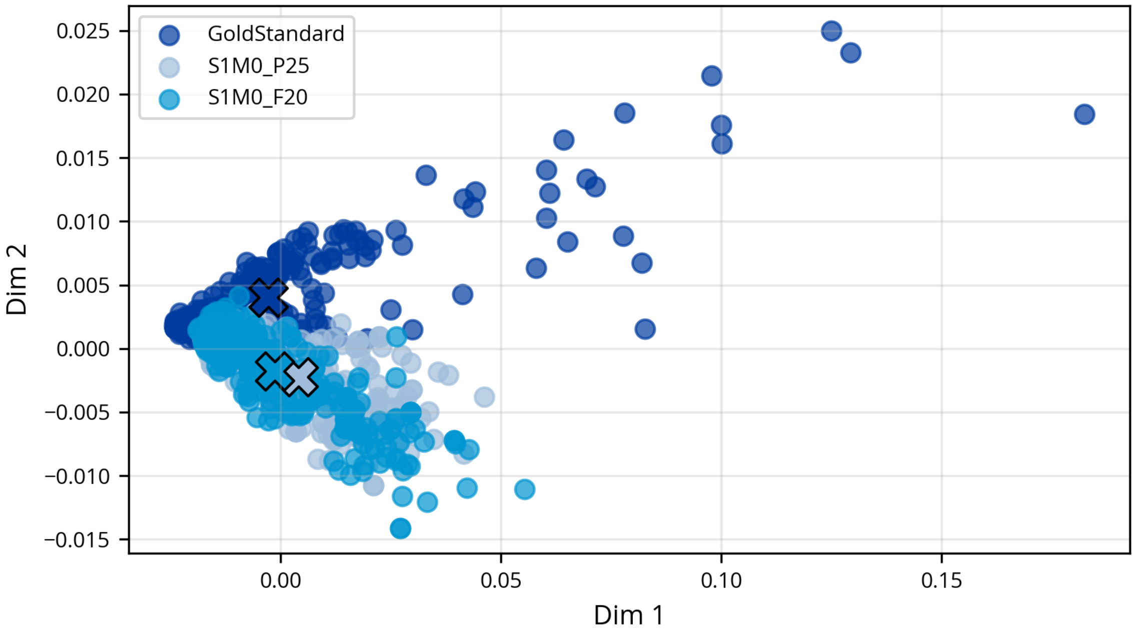


Figure 5.2: PCA visualization of the embedding representation of the S1M0-based datasets.

In this visualization, the cluster of sentences generated by P25 forms a broader and more dispersed region compared to that of F20. The centre of the P25 cluster, denoted by a cross on the plot, appears wider and more spread out, indicating that the generated sentences cover a larger portion of the embedding space. This again supports the conclusion that P25 generates more diverse synthetic sentences, not only in terms of lexical variation captured by Self-BLEU, but also in terms of semantic variability reflected in the embedding-based analysis.

To complement the Self-BLEU analysis, n-gram statistics are computed to further quantify diversity and redundancy in the generated text. Figure 5.3 shows the n-gram-based scores of the S1M0-based datasets.

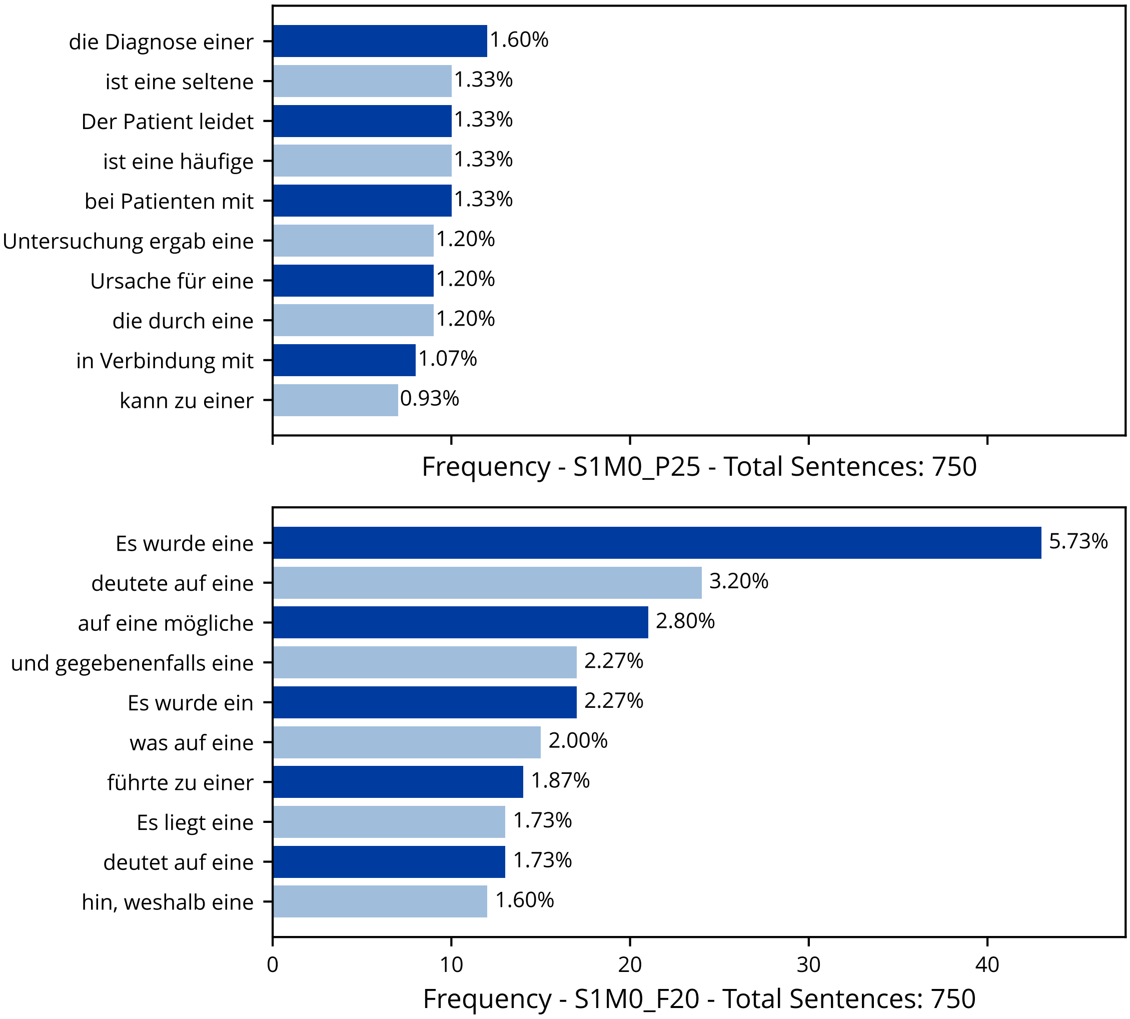


Figure 5.3: Top trigrams of the S1M0-based datasets.

The n-gram analysis confirms the findings from the Self-BLEU scores. Across all n-gram levels, the P25 model obtains lower scores than the F20 model, again indicating a higher degree of variation and lower repetition in its generated text. Taken together, the Self-BLEU, PCA, and n-gram analyses consistently show that P25 produces more diverse synthetic sentences than F20 while still maintaining the required syntactic structure and annotation format.

Following the intrinsic assessments of diversity and quality, the first downstream evaluation is carried out using a NER task. In this setup, the LLM is fine-tuned exclusively on the synthetic data generated under the S1M0 configuration. The objective is to understand whether such synthetic-only training data can already provide competitive performance on a real-world NER benchmark and how different generative models compare in this regard.

For the initial downstream analysis, F1 scores computed on entity spans are considered. Spans are the continuous text segments that are annotated as entity mentions in the NER task. At this stage, the evaluation disregards the specific ETs and focuses solely on whether entity boundaries are detected correctly. Figure 5.4 shows the span-level F1 scores for the models fine-tuned on the different S1M0 datasets.

The evaluation reveals that the P25-based synthetic dataset leads to better downstream performance than the F20-based dataset. This indicates that the greater diversity and syntactic reliability of P25's outputs translate into more effective learning signals for span detection. Additionally, the results demonstrate that synthetic data generated under the S1M0 configuration already outperforms the gazetteer baseline introduced in section 4.4. Even in this minimal configuration with single-term sentences, synthetic data thus provides a substantial advantage over a simple lookup-based baseline.

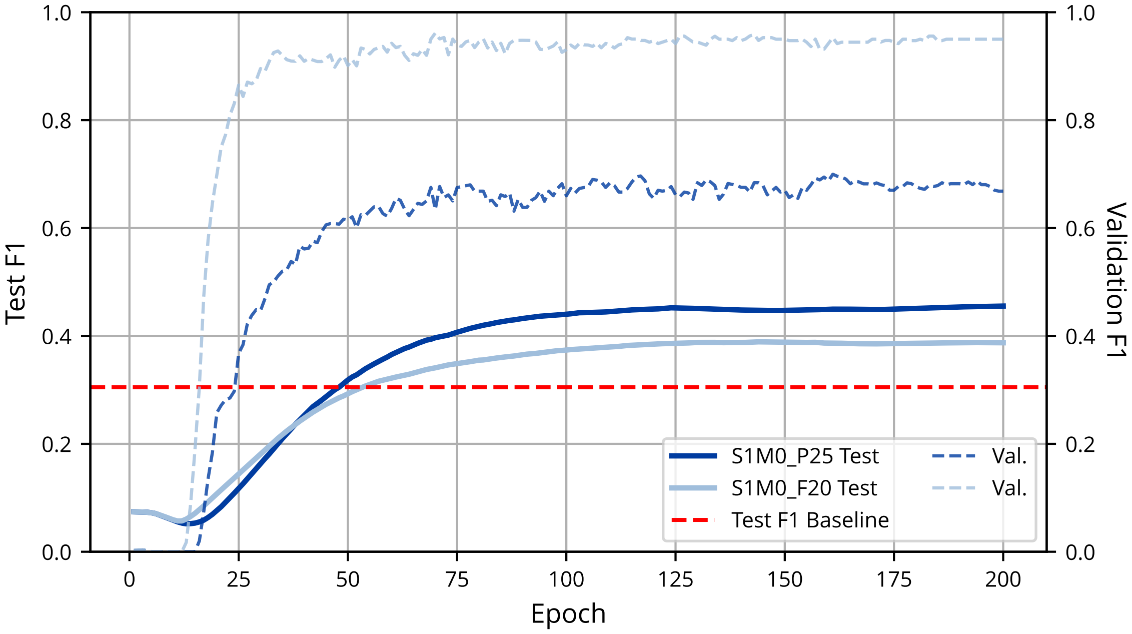


Figure 5.4: Spans F1 evaluation of the S1M0-based datasets.

In a second downstream evaluation, the same models and datasets are assessed using term-level metrics instead of span-level ones. Here, the focus shifts from correctly identifying the boundaries of entities to correctly predicting the specific entity terms. The terms correspond to the actual medical expressions or concepts annotated in the NER task.

The results of this term-based evaluation show that the number of generated sentences in the `S1M0` configuration is not yet sufficient to effectively train the models for robust term recognition. The corresponding F1 scores fall below the gazetteer baseline, indicating that the available synthetic data does not yet cover the necessary lexical variety to match or exceed the performance of a vocabulary-based approach. Figure 5.5 presents the term-level F1 scores for validation and test sets.

In both experiments, the performance of the validation set suggests that the less diverse and more repetitive sentences generated by F20 enable the fine-tuned model to memorize specific terms more easily and to overfit to the validation data. This suspected overfitting is further supported by the test set results, where both models show markedly lower performance than on the validation set. The gap between validation and test scores highlights that the current `S1M0` setup, while sufficient for improving span detection over the baseline, still lacks the breadth of term coverage required for more robust and generalizable term-level recognition.

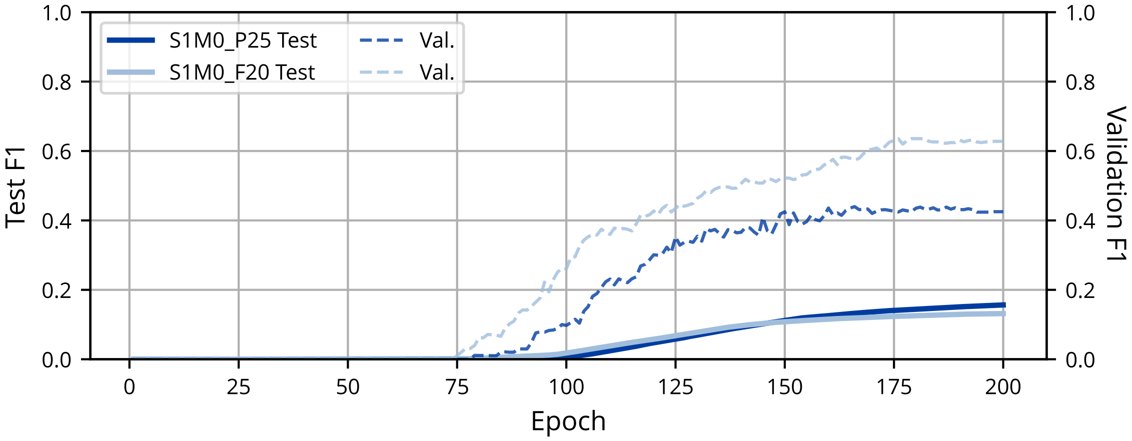


Figure 5.5: Terms F1 evaluation of the `S1M0`-based datasets.

From the perspective of span-level evaluation, the gazetteer-based baseline is already surpassed when using synthetic-only data generated under the `S1M0` configuration. This provides a first affirmative answer to the research question 4 of whether synthetic data alone can be sufficient for non-English clinical NER, at least with respect to reliably detecting entity boundaries. At the same time, the term-level evaluation highlights that this conclusion does not yet extend to the accurate recognition of specific medical expressions. The lower term-level F1 scores and the observed overfitting patterns indicate that additional improvements in the generation pipeline are required, in particular larger datasets, increased lexical variety, and more comprehensive coverage of rare and morphologically varied terms. These findings suggest that while span detection can already benefit substantially from basic synthetic data, robust term-level performance will likely require more sophisticated configurations with multiple terms per sentence and richer contextualization.

In summary, the `S1M0` configuration provides a controlled starting point for evaluating synthetic clinical text generation and its usefulness for downstream NER tasks. The analyses show that commercial models, in particular Gemini 2.5 Pro (P25) and Gemini 2.0 Flash (F20), are clearly superior to the open-source models in reliably following the generative request specification and generating syntactically valid clinical sentences. P25 consistently produces more diverse text than F20, as evidenced by lower Self-BLEU scores, broader PCA distributions, and lower n-gram scores. This increased diversity correlates with improved performance in span-level NER evaluation, where P25-based synthetic data outperforms both F20 and the gazetteer baseline.

At the same time, the term-level evaluation reveals the limitations of the `S1M0` setup. The amount and lexical coverage of synthetic data are not yet sufficient to surpass a simple gazetteer-based baseline in recognizing specific medical terms. The observed overfitting of the F20-based model on the validation set underscores the need for more varied and extensive synthetic data. Overall, the findings from the `S1M0` experiments motivate moving towards more complex configurations with multiple terms per sentence and larger datasets, in order to further enhance diversity, coverage, and downstream NER performance while maintaining strict control over annotation quality.

S2M0 - Requesting one HPO term per configuration

In this configuration, the generative request is simplified to request only a single HPO term per model invocation. By constraining the task in this way, the overall complexity of the generation process is reduced, which is particularly beneficial for less capable and smaller open-weight models. Instead of having to balance multiple competing objectives within a single request, the model can concentrate on producing one specific type of sentence at a time. This narrower focus is expected to improve both the faithfulness of the generated content to the requested HPO term and the overall linguistic quality of the output.

Codeblock 5.2 shows the exact request template used for the `S2M0` configuration. The request instructs the model to generate 40 unique sentences that describe the provided HPO label as it would appear in a clinical or medical report. Each sentence is encouraged to offer a different perspective or detail, similar to how various physicians might document their observations or diagnoses in clinical notes. Additionally, the definition and description of the HPO label are provided as reference material to aid the model in accurately capturing the intended meaning.

Codeblock 5.2: Generative query used for generating the synthetic `S2M0` dataset.

```
Generiere 40 einzigartige Sätze, die das bereitgestellte HPO-Label so beschreiben, wie
    sie in einem klinischen Bericht oder einem medizinischen Bericht erscheinen würden.
    Jeder Satz sollte eine andere Perspektive oder ein anderes Detail bieten, ähnlich wie
     verschiedene Ärzte ihre Beobachtungen oder Diagnosen in klinischen Notizen vermerken
     würden.
Bekannte klinische Abkürzungen sind erwünscht, welche ein Arzt beim Verfassen von
    Berichten verwenden würde. Vermeide es, in jedem Satz genau denselben Ausdruck zu
    verwenden, um Vielfalt zu gewährleisten und die Bandbreite klinischer Ausdrucksweisen
     widerzuspiegeln. Erstelle eine Aufzählung mit Aufzählungszeichen statt einer
    nummerierten Liste. Erstelle ausschließlich die Aufzählung und verzichte auf
    jeglichen Kommentar abseits der Liste. Jedes HPO-Label, auch in abgewandelter Form,
    ist im Text durch jeweils zwei Unterstriche vor und nach dem Label zu markieren,
    sodass es im Markdown-Format fett erscheint (Beispiel: __HPO-Label__). Zusätzlich
    sind als Gedächtnisstütze noch die Definition und die Beschreibung des HPO-Labels
    beigefügt.
HPO-Label: Leistenhernie
Definition: Vorwölbung des Inhalts der Bauchhöhle durch den Leistenkanal.
Beschreibung: Die Leistenhernie zeigt sich als Vorwölbung in der Leiste.
```

This design choice directly supports the overarching research question 3 of how well open-weight models can act as substitutes for closed, commercial systems. Any observed improvements or degradations in quality can be more confidently attributed to the model capabilities rather than to confounding request complexity. In addition, the simplified setup aligns better with real-world scenarios where model calls may need to be modularized and optimized for efficiency.

A further advantage of restricting the generation to one HPO term at a time lies in the improved traceability of the generation process. Since each model call is associated with exactly one target concept, it becomes easier to track which HPO terms are successfully generated in the output. This granularity supports more fine-grained error analysis and enables targeted adjustments to generative request templates. It also improves the interpretability of downstream evaluations, as the contribution of each generated sentence to the overall dataset can be more clearly mapped back to its originating term.

Comparing with previous version To maintain continuity with the earlier setup, the new `S2M0` configuration is systematically compared to the previous `S1M0` configuration, with both datasets being generated using the Gemini 2.0 Flash model. Whereas `S1M0` relied on more complex queries requesting multiple sentence types per HPO term, S2M0 relies on a more focused, single-term generative query strategy. Although the S2M0 query is more concise and places stronger emphasis on accurately realizing one specific term, it also reduces the effective context length. This shorter and more targeted context allows Gemini 2.0 Flash to better allocate its capacity to the core generation task, which is expected to yield more coherent and relevant outputs.

The impact of this change is quantitatively assessed using Self-BLEU scores, which measure the overlap between generations from S2M0 and those from S1M0. As shown in Table 5.2, S2M0 exhibits substantially higher Self-BLEU scores across all evaluated configurations. These gains indicate that the simplified querying strategy has a positive effect on both the stability and similarity of the generated sentences when compared to the previous setup.

Model	**Self-BLEU**
S1M0_F20	0.328
S2M0_F20	0.223

Table 5.2: Self-BLEU scores of the Gemini 2.0 Flash-based datasets comparing generative queries S2M0 and S1M0.

To complement the Self-BLEU analysis, a PCA-based visualization is used to inspect the geometric structure of the generated datasets. Figure 5.6 shows the distribution of S2M0 and S1M0 sentences in a reduced-dimensional space, constructed to highlight differences in semantic and lexical patterns with respect to a gold-standard reference. In the earlier analysis of S1M0, it was assumed that a larger distance between a cluster centre and the gold-standard cluster would necessarily reflect poorer data diversity. Under the new S2M0 configuration, this assumption no longer holds in such a straightforward manner.

Instead, the visualization reveals a more nuanced picture. The S2M0 data tends to form clusters that are more compact and located closer to the gold-standard distribution, whereas the S1M0 data appears to be spread further away in several regions of the space. This could suggest that the S1M0 dataset may have contained more sentences that diverged from the reference style and content, potentially indicating a less faithful coverage of the gold-standard diversity. In contrast, the S2M0 configuration appears to generate sentences that are both more consistent and better aligned with the target domain, while still preserving a sufficient level of variation.

A detailed comparison of the n-gram distributions provides further insight into the lexical properties of the generated datasets. As illustrated in Figure 5.7, the trigram frequencies of S2M0 and S1M0 are contrasted to determine how strongly the generation process tends to repeat or diversify specific word sequences. The trigram in particular reveals a notable improvement for S2M0: it exhibits fewer bins dominated by highly frequent, repetitive sequences compared to S1M0. This reduction indicates that S2M0 generations are less prone to overusing fixed

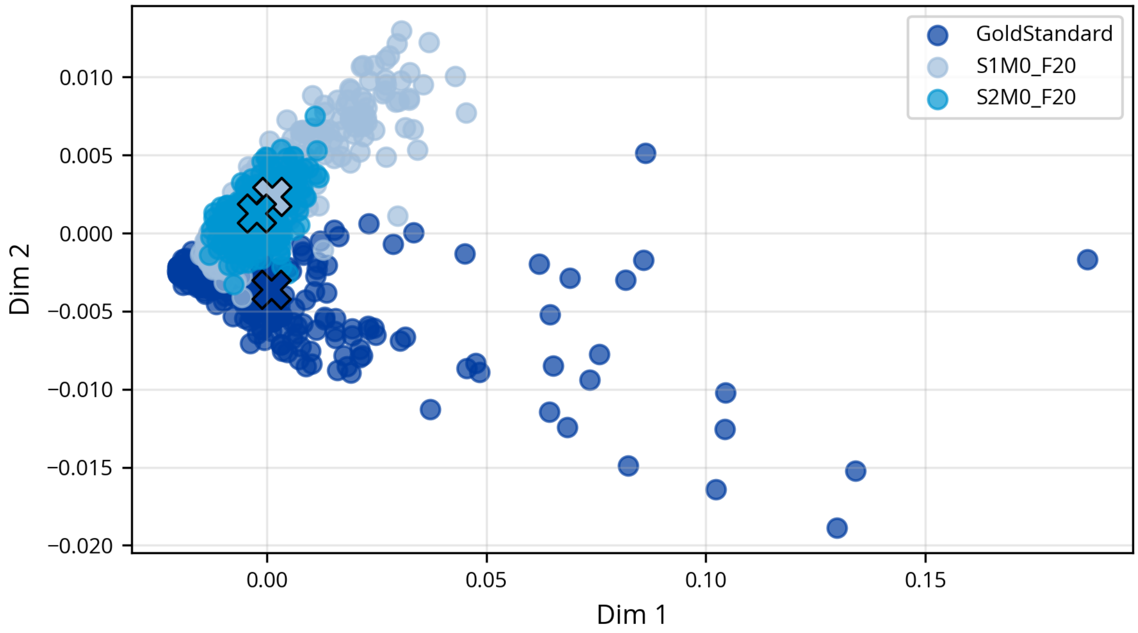


Figure 5.6: PCA visualization of the Gemini 2.0 Flash-based datasets comparing generative queries S2M0 and S1M0.

phrases or template-like constructions. Instead, the model produces a broader variety of local word patterns while still respecting the constraints imposed by the individual HPO terms. Such behaviour is desirable for downstream training, as it enriches the model with a more diverse set of lexical and syntactic realizations without compromising the semantic relevance of the sentences.

The downstream impact of the S2M0 dataset on model training is evaluated using F1 scores for both span-level and term-level predictions. Figure 5.8 compares the evaluation performance of models trained on S2M0 data with those trained on S1M0 data, again under the Gemini 2.0 Flash generation setup. Across the examined configurations, substantial improvements in F1 scores can be observed in favour of S2M0, for both span detection and term classification tasks.

These gains demonstrate that the higher-quality and better-aligned sentences in S2M0 translate directly into more effective supervision during training. The models are able to learn more robust patterns for recognizing and linking HPO terms. The training curves obtained with S2M0 shows faster convergence. This suggests that the simplified single-term querying strategy not only improves the intrinsic quality of the generated data, but also enhances the overall efficiency and reliability of the training process.

In summary, requesting a single HPO term per configuration in S2M0 leads to clear advantages over the earlier S1M0 setup. The simplified queries reduce task complexity and context length,

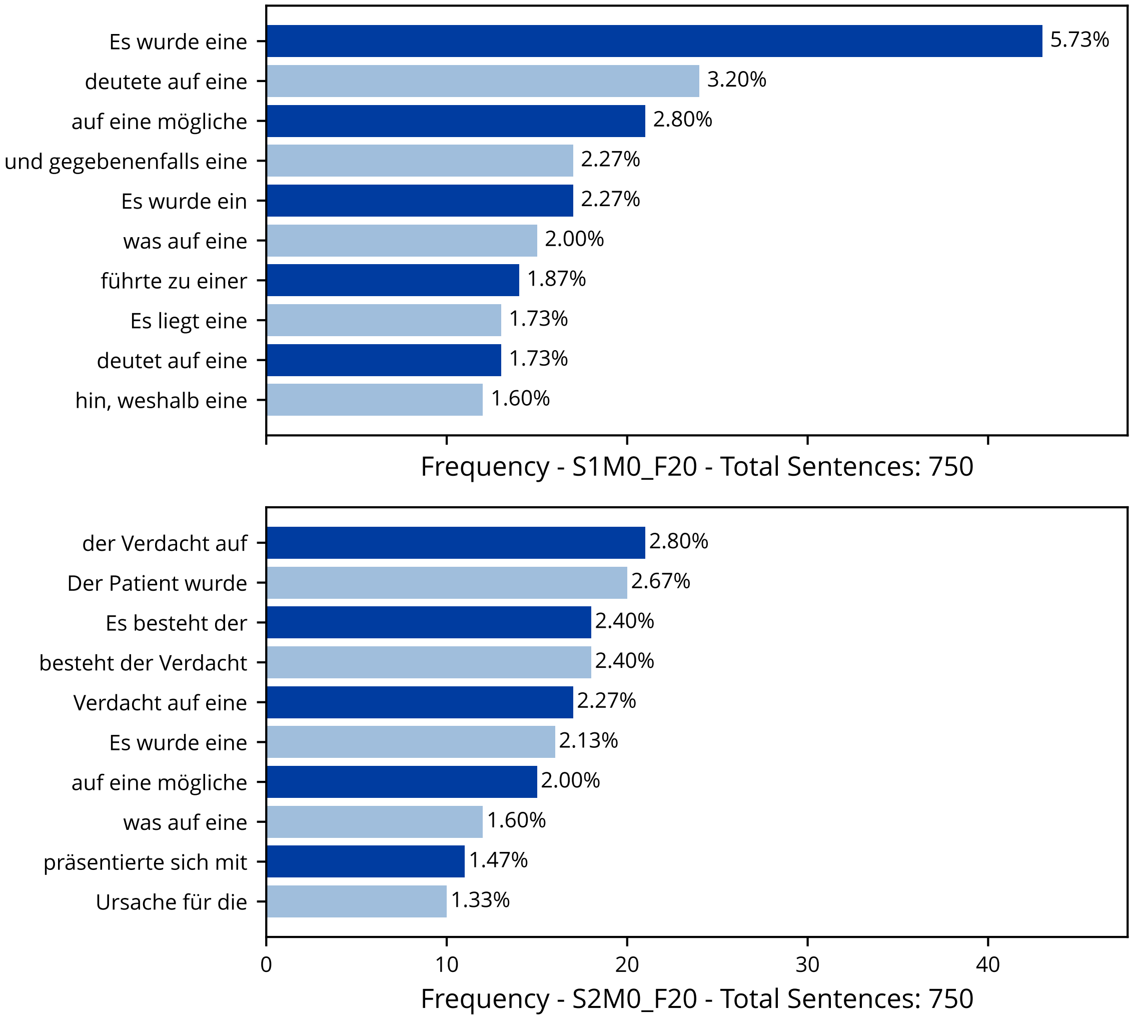


Figure 5.7: Top trigrams of the Gemini 2.0 Flash-based datasets comparing generative queries S2M0 and S1M0.

allowing the model to better focus on the intended generation objective. Self-BLEU, F1 scores, and n-gram analyses consistently point to higher data quality, improved alignment with the gold-standard distribution, and more effective downstream learning. These findings support the conclusion that carefully constrained, single-term querying is a promising strategy for leveraging open-weight models as viable alternatives to closed, commercial systems in HPO-focused text generation tasks.

Inspect training differences The previous evaluation revealed notable differences in training dynamics between the S2M0 and S1M0 datasets. While aggregate F1 scores already indicated an advantage for S2M0, they did not fully explain how these gains emerged over the course of training. To obtain a more detailed understanding, the evolution of F1 scores

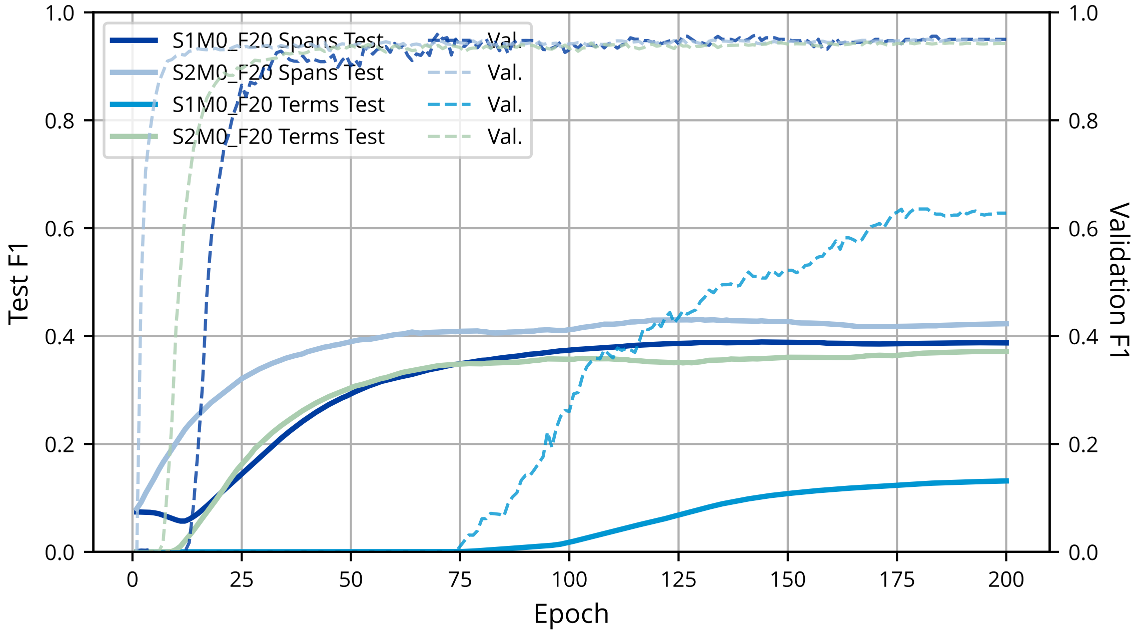


Figure 5.8: F1 evaluation of the Gemini 2.0 Flash-based datasets comparing generative queries `S2M0` and `S1M0`.

across epochs is examined for both span-level and term-level predictions. This temporal perspective helps to disentangle whether performance differences arise primarily from faster convergence, higher final scores, or both, and to what extent these effects can be attributed to dataset size versus query design.

A central assumption in the earlier analysis was that the smaller effective dataset size of `S1M0` contributes to slower training progress and weaker final test performance. Because `S1M0` contains fewer examples per HPO term and relies on more complex queries, it is hypothesized that models trained on this data would require more epochs to generalize effectively and would remain more sensitive to noise and idiosyncrasies in the synthetic sentences. At the same time, `S2M0` benefits from a larger set of examples generated with a more focused query design, making it difficult to separate the effects of quantity from those of query structure.

To test this hypothesis in a controlled manner, a new experiment is conducted in which the `S2M0` dataset is downsampled to match the size of `S1M0`. The resulting configuration, denoted as `S2M0_F20_SSA`, preserves the single-term query design of `S2M0` while enforcing the same number of training instances as in `S1M0`. This setup enables a direct comparison that isolates the impact of querying strategy and model choice from purely data-volume effects.

Figure 5.9 presents the span-level F1 trajectories for `S2M0_F20_SSA` and `S1M0` over the training epochs. The subsampled dataset shows that the convergence speed is depending on the dataset size. Also, the performance between S1M0 and S2M0 both using Gemini 2.0 Flash show that S2M0 achieves better performance.

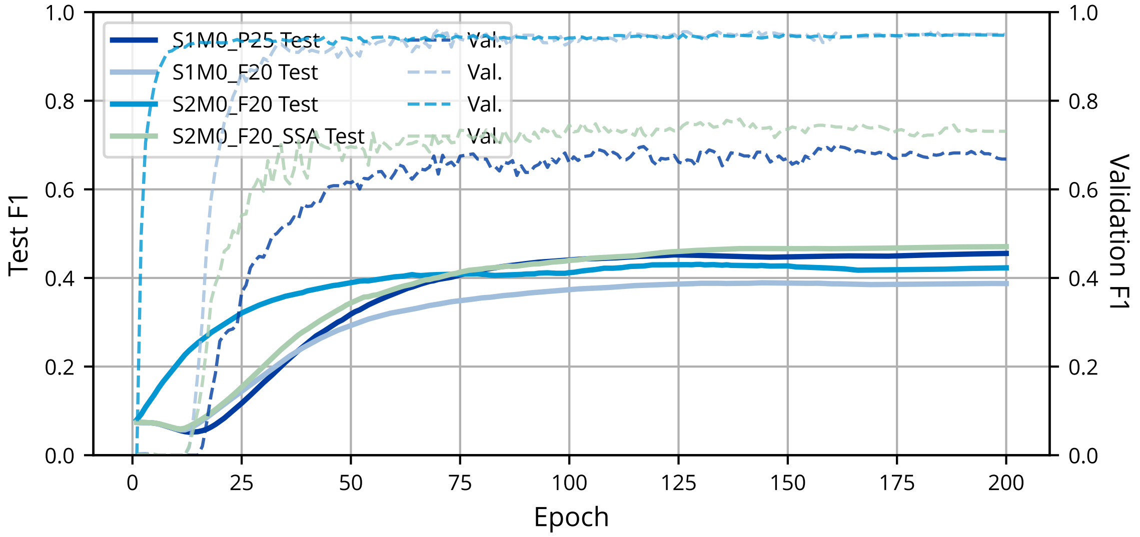


Figure 5.9: F1 spans evaluation of a subsampled part of S2M0 using the Gemini 2.0 Flash-based dataset. The SSA suffix denotes the subsampled dataset.

Figure 5.10 shows the corresponding term-level F1 trajectories for the same configurations. Here, reduced dataset size becomes more apparent.

Taken together, these controlled experiments demonstrate that the improved query design employed in S2M0, combined with Gemini 2.0 Flash as the generation back-end, achieves performance that is comparable to or better than the S1M0 setup, even though S1M0 relies on the more capable Gemini 2.5 Pro model. Moreover, the analysis confirms that the primary limitation of S1M0 lies not only in the querying strategy but also in the insufficient size and coverage of the synthetic dataset. When S2M0 is constrained to the same scale, it still exhibits more favourable training dynamics, underscoring the importance of both well-structured queries and adequate dataset coverage for effective HPO-focused model training.

Comparing models Building on the previous analysis of the S2M0 configuration, this section examines how different open-weight models compare to the commercial Gemini 2.0 Flash system when used as generators of synthetic training data. The goal is to assess to what extent open-weight models can serve as practical substitutes in HPO-focused tasks, not only in terms of overall performance, but also with respect to training dynamics and error characteristics.

Figure 5.11 shows the span-level F1 scores for models trained on the S2M0 dataset. No substantial differences can be observed between the various model configurations. In particular,

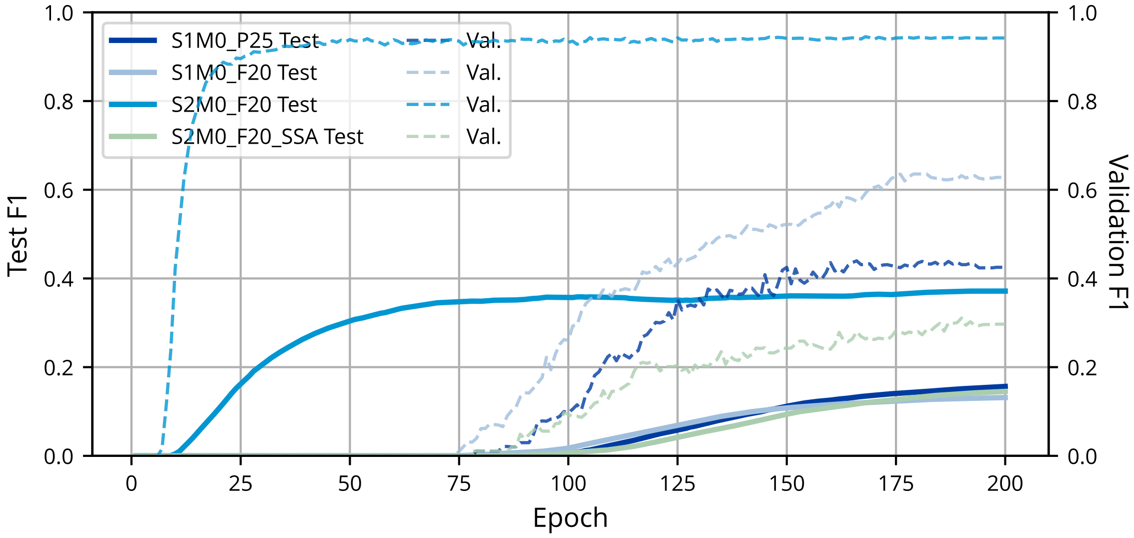


Figure 5.10: F1 terms evaluation of a subsampled part of S2M0 using the Gemini 2.0 Flash-based dataset.

compared to the commercial Gemini 2.0 Flash model, the open-weight models perform on a similar level in detecting mention boundaries. This close alignment indicates that, under the single-term S2M0 querying strategy, the choice of generator has only a limited impact on span-level performance.

This observation supports the conclusion that, for span detection, open-weight models can indeed match the performance of closed, commercial systems when trained on synthetic data generated with the S2M0 configuration. As a result, span-level predictions appear to be relatively robust to variations in the underlying generator model, provided that the query structure is sufficiently clear and constrained.

In contrast, Figure 5.12 presents the term-level F1 scores for models trained on the same S2M0 dataset, revealing more pronounced differences between the model configurations. While Gemini 2.0 Flash remains a strong baseline, it is no longer the dominant model in this setting. Instead, the Gemma3 configuration is able to outperform Gemini 2.0 Flash for term-level classification, suggesting that it generates synthetic data particularly well suited for the HPO prediction task.

Table 5.3 shows the Self-BLEU scores for the various model configurations trained on the S2M0 dataset. The results emphasize the weak performance of the Mistral model, which exhibits comparatively poor overlap with the reference generations, suggesting lower stability or greater divergence in style and content. By contrast, the Self-BLEU scores for the Gemma3 model are surprisingly high, indicating a strong similarity to the Gemini 2.0 Flash outputs and a relatively consistent generation behaviour across the dataset. These high scores were

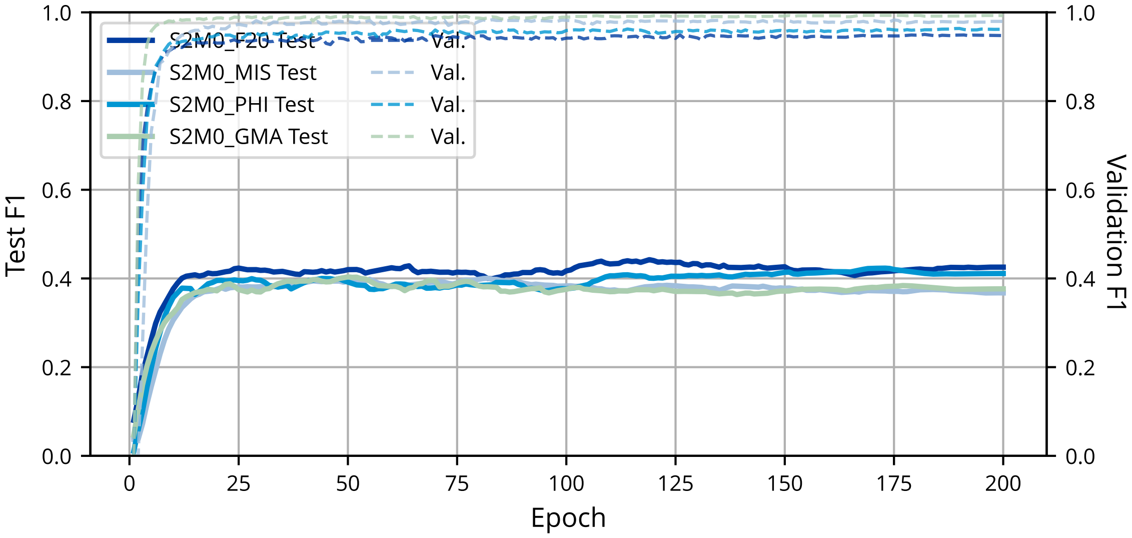


Figure 5.11: F1 spans evaluation of datasets resulting from the S2M0 generative query using different open-weight models.

not necessarily expected to translate as effectively into downstream evaluation performance as observed in the term-level F1 results.

Dataset	Self-BLEU
S2M0_F20	0.223
S2M0_PHI	0.233
S2M0_MIS	0.475
S2M0_GMA	0.297

Table 5.3: Self-BLEU score of datasets resulting from the S2M0 generative query using different open-weight models.

At the same time, the Self-BLEU analysis underlines how nuanced and incomplete such metrics are when interpreted in isolation. High overlap may reflect useful consistency, but it may also signal reduced diversity. Consequently, the combination of Self-BLEU, F1 scores, and qualitative inspections is necessary to obtain a more balanced view. On their own, the results in Table 5.3 cannot be used to draw definitive conclusions about model superiority.

The PCA visualization in Figure 5.13 further complements this picture by illustrating how the different model-generated datasets are distributed in a reduced-dimensional space relative to the gold-standard distribution. The stronger-performing Gemini 2.0 Flash and Gemma3 models form clusters that lie closer to the gold-standard data, reflecting their closer alignment in both semantic content and lexical style. In contrast, the weaker models appear more

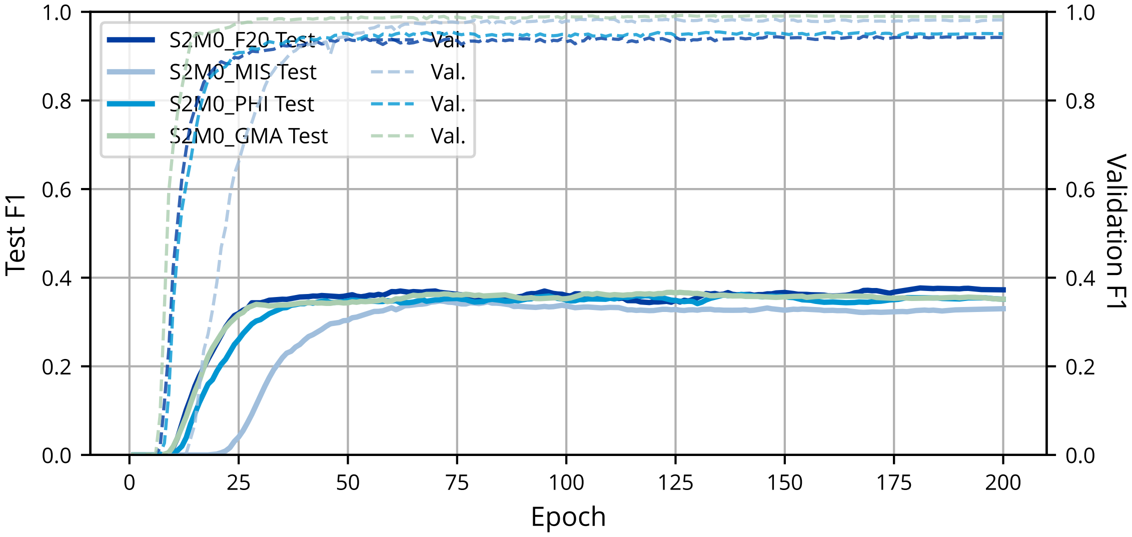


Figure 5.12: F1 terms evaluation of datasets resulting from the `S2M0` generative query using different open-weight models.

dispersed and further away in several regions, consistent with their less favourable evaluation scores.

Beyond geometric similarity, another important factor is the error rate during data generation with the various models. When generating the 300 HPO terms with 40 sentences each, the Phi and Gemma3 models, with 18 and 25 errors respectively, exhibited the lowest error rates. Meanwhile, the Mistral model produced the highest number of errors at 107 and even failed to generate data for one HPO term entirely, despite multiple attempts. These differences in generation robustness have practical implications for large-scale data construction workflows, where repeated failures or malformed outputs can introduce additional manual effort and potential biases.

Considering both the evaluation results and the close similarities between the Gemma3 and Gemini 2.0 Flash models in terms of performance, distributional properties, and error rates, the following experiments focus solely on the open-weight Gemma3 model for further analysis. This choice reflects a pragmatic trade-off between performance, reproducibility, and accessibility, while still maintaining comparability to a strong commercial baseline.

The n-gram distribution in Figure 5.14 adds another layer of evidence regarding the lexical characteristics of the generated datasets. However, interpreting these patterns in terms of overall model quality remains challenging. Higher n-gram diversity does not automatically guarantee better downstream performance, and more repetitive patterns can still be effective if they closely mirror the target domain style.

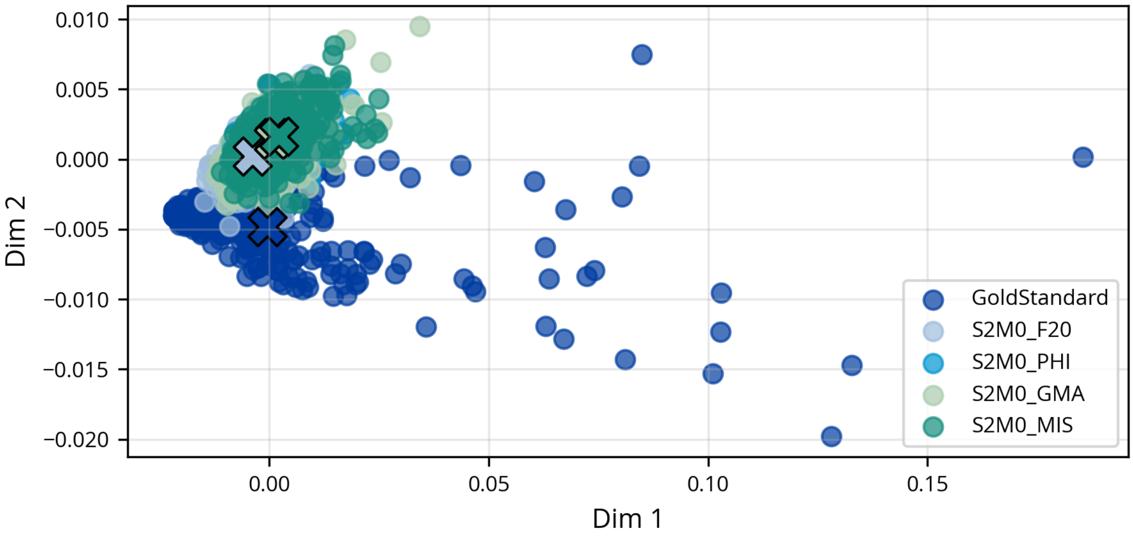


Figure 5.13: PCA visualization of datasets resulting from the S2M0 generative query using different open-weight models.

As a result, the n-gram analysis reinforces the notion that individual metrics and visualizations must be interpreted with care. Differences in n-gram distributions alone are insufficient to rank models or to fully explain the observed F1 and Self-BLEU scores.

In conclusion, the comparison of open-weight and commercial models under the S2M0 configuration shows that open-weight systems can reach, and in some cases exceed, the performance of a strong proprietary baseline when used for HPO-focused synthetic data generation. Span-level F1 scores suggest broad parity across models, while term-level results highlight specific advantages for Gemma3. Self-BLEU scores, PCA visualizations, n-gram distributions, and error-rate analyses together paint a nuanced and sometimes contradictory picture, underscoring that no single metric suffices to characterize model quality comprehensively. Nevertheless, the combined evidence supports the selection of Gemma3 as a competitive and practically viable open-weight generator for subsequent experiments, while also emphasizing the need for multi-faceted evaluation frameworks when assessing synthetic data pipelines in specialized biomedical domains.

S3M0 - Increased Complexity

In configuration S3M0, the overall generative query design is deliberately made more complex than in S2M0 in order to better guide the model and reduce ambiguity in the generation process. While previous configurations focused primarily on concise instructions, S3M0 introduces a richer structure that aims to provide the model with clearer expectations, more context, and

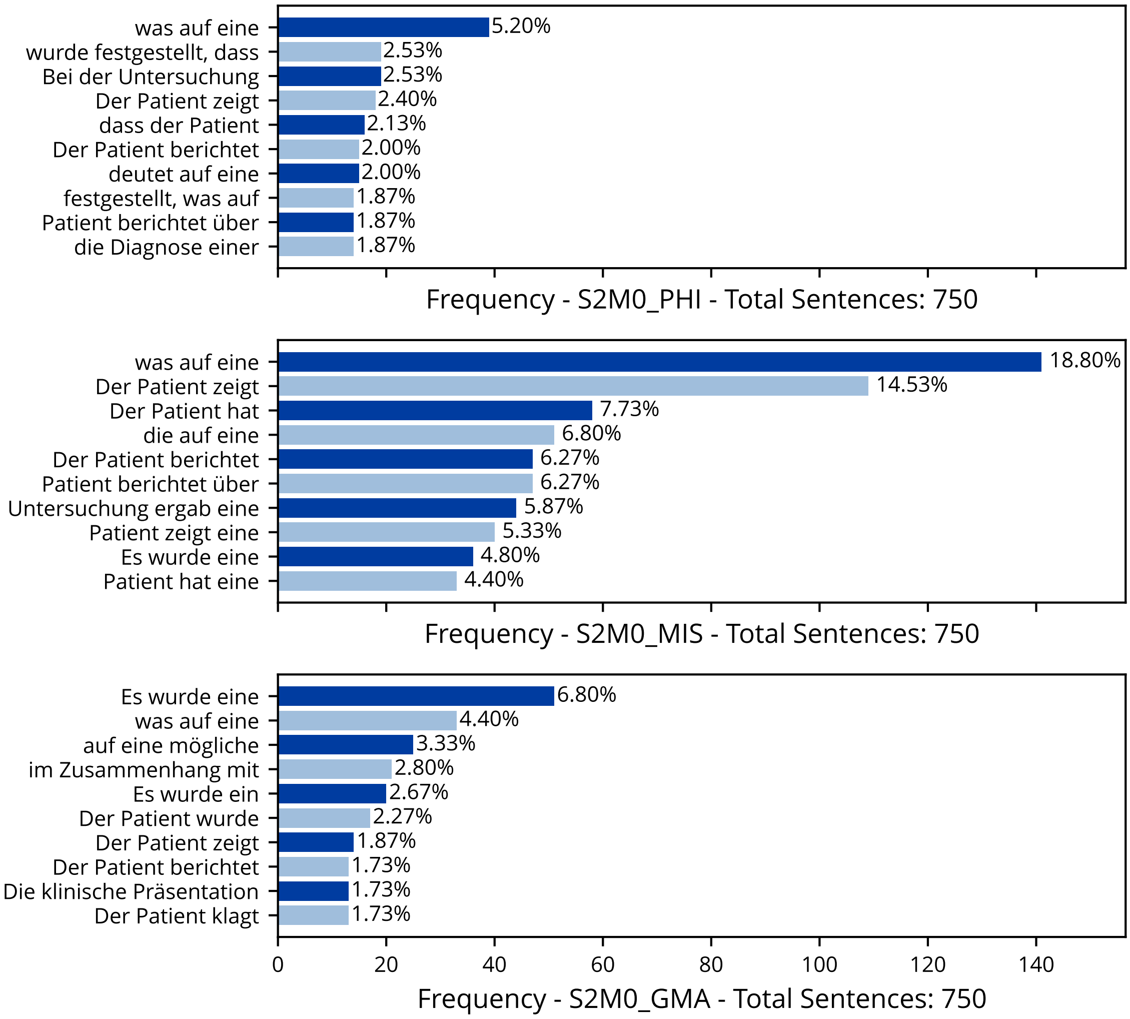


Figure 5.14: Top trigrams of datasets resulting from the S2M0 generative query using different open-weight models.

explicit examples. The configuration itself is too extensive to be reproduced in this section. A complete listing can be found in Appendix 8.1.1.

A central characteristic of S3M0 is the addition of explicit example inputs and outputs. Its structure can be summarized as follows:

- Instructions
- Example Input
- Example Output
- Template
- Final Instruction

The query is organized into several distinct components that follow a fixed order. First, the model receives a set of instructions specifying the task, constraints, and desired style and level of detail. This is followed by an *Example Input*, which illustrates the type and structure of the data that the model will process. Next, an *Example Output* is provided, showing a concrete instance of the expected response format and content quality. After these illustrative elements, the *Template* section provides the details of the phenotype to be described. Finally, a *Final Instruction* reiterates the most important requirements and attempts to ensure that the model adheres to the prescribed structure and objectives.

This structured design is intended to balance guidance and flexibility. The examples and template constrain the output format sufficiently to improve consistency, while still allowing the model to produce varied and informative content within those boundaries. As a result, `S3M0` aims to enhance both the quality and diversity of the generated outputs compared to `S2M0`.

Table 5.4 shows the Self-BLEU scores for `S3M0` compared to `S2M0`. A slight increase in diversity can be observed for `S3M0`, indicating that the outputs under the more complex configuration are somewhat less similar to each other than in the simpler setting. This suggests that the additional structure and examples do not lead to excessive convergence on a single response pattern, but instead encourage a broader range of expressions while maintaining relevance.

Model	**Self-BLEU**
S2M0_GMA	0.297
S3M0_GMA	0.278

Table 5.4: Self-BLEU scores of datasets resulting from the `S3M0` and `S2M0` generative queries.

The impact of `S3M0` on the representation space of the generated outputs is illustrated in Figure 5.15, which shows the PCA plot for `S3M0` compared to `S2M0`. The clusters remain overlapping, indicating that `S3M0` does not fundamentally change the nature of the outputs. It is not clear that the PCA plot can be used to gauge diversity differences in this case.

Figure 5.16 shows the n-gram overlap for `S3M0`. This time, no significant difference can be observed when compared to `S2M0`. The similarity of n-gram distributions indicates that, despite the increased complexity of the query and the inclusion of examples, the local lexical patterns remain largely unchanged. This suggests that `S3M0` primarily influences higher-level aspects such as structure, content selection, and semantic variation, rather than leading to a fundamentally different vocabulary or phrase usage.

The evaluation also shows slightly better results for `S3M0` as can be seen in Figure 5.17. Across the considered metrics, `S3M0` generally achieves marginal improvements over `S2M0`, indicating that the additional query structure and examples help the model align more closely

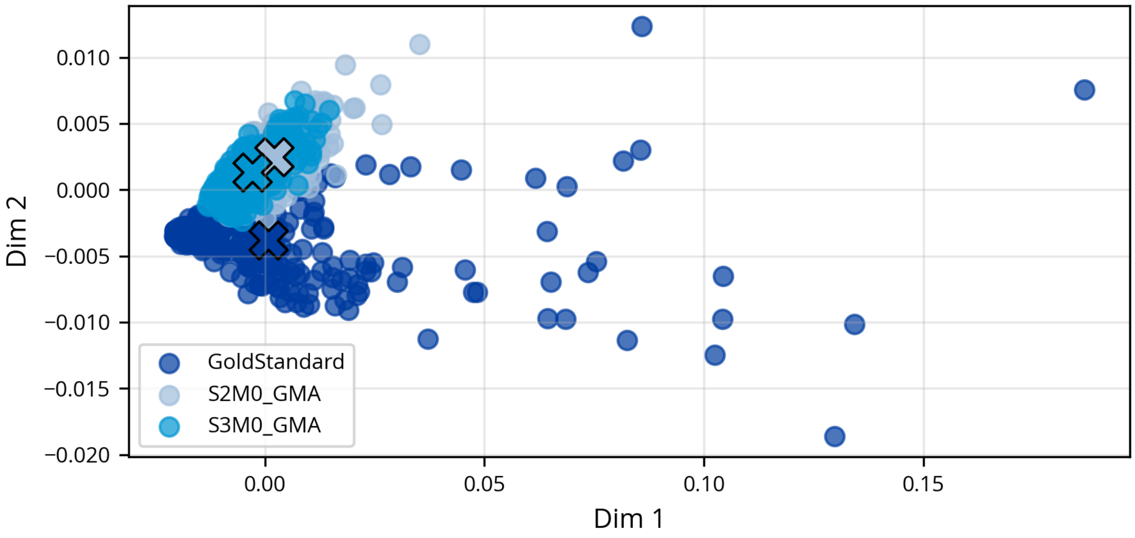


Figure 5.15: PCA visualization of datasets resulting from the S3M0 and S2M0 generative queries.

with the target outputs. What stands out is that the score for the term level evaluation is increased significantly. This improvement suggests that S3M0 enhances the ability of the model to identify and reproduce the relevant terms or concepts required by the task. The explicit example input and output, together with the template, appear to guide the model more effectively toward the correct terminology and content elements, thereby improving precision and recall at the term level.

In summary, the S3M0 configuration demonstrates that increasing query complexity through structured instructions, example inputs and outputs can yield measurable benefits. The results indicate a slight increase in overall diversity (Table 5.4), stable n-gram characteristics (Figure 5.16), and notably improved performance at the term level (Figure 5.17). These findings suggest that carefully designed, more complex queries can help large language models produce outputs that are both more diverse and more aligned with task-specific requirements, without sacrificing local linguistic consistency.

S4M0 - Add sampling to the data generation

In configuration S4M0, the original data generation setup from Section 5.2.1 is extended by introducing sampling into the construction of the generative request examples. Additionally, one section called *context* is added to the request, this is done to sample the context in which the sentence should be generated. This matches the real world sections *Aktuelle Diagnose*, *Dauerdiagnose* and *Körperlicher Untersuchungsbefund* as they are found in the gold standard test set. Otherwise, the core structure of the query is preserved, ensuring a fair and controlled

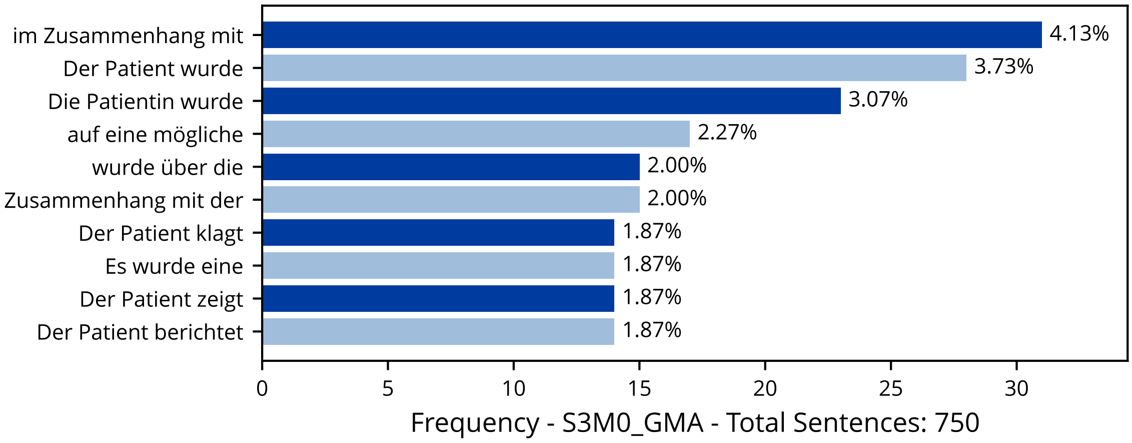


Figure 5.16: Top trigrams of datasets resulting from the S3M0 and S2M0 generative queries.

comparison with the earlier configuration S3M0, while specifically isolating the impact of sampling on diversity and downstream model performance.

To that end, the Example Input and Example Output sections in S4M0 are no longer populated by a single, immutable set of instances. Instead, they are drawn at random from a larger pool of manually validated, pre-prepared examples for each new generation request. As a result, every query presented to the model can contain a different combination of examples, even if the high-level instructions and templates remain unchanged. This dynamic sampling mechanism aims to reduce overfitting to a fixed context and to encourage the model to generalize better across the wider example space, thereby increasing the diversity of the generated outputs.

The effectiveness of this modification in terms of diversity can be observed quantitatively using Self-BLEU. Table 5.5 compares the Self-BLEU scores of S4M0 against those of S3M0. The scores for S4M0 are significantly lower, which indicates a notable reduction in similarity between generated samples. Since Self-BLEU captures the degree to which outputs resemble each other, a lower score reflects a higher degree of diversity. Consequently, the results in Table 5.5 provide clear evidence that the introduction of sampling into the generative request examples effectively increases the variability of the generated data without having to alter the underlying instructional content of the queries.

Model	**Self-BLEU**
S3M0_GMA	0.278
S4M0_GMA	0.180

Table 5.5: Self-BLEU scores of datasets resulting from the S4M0 and S3M0 generative queries.

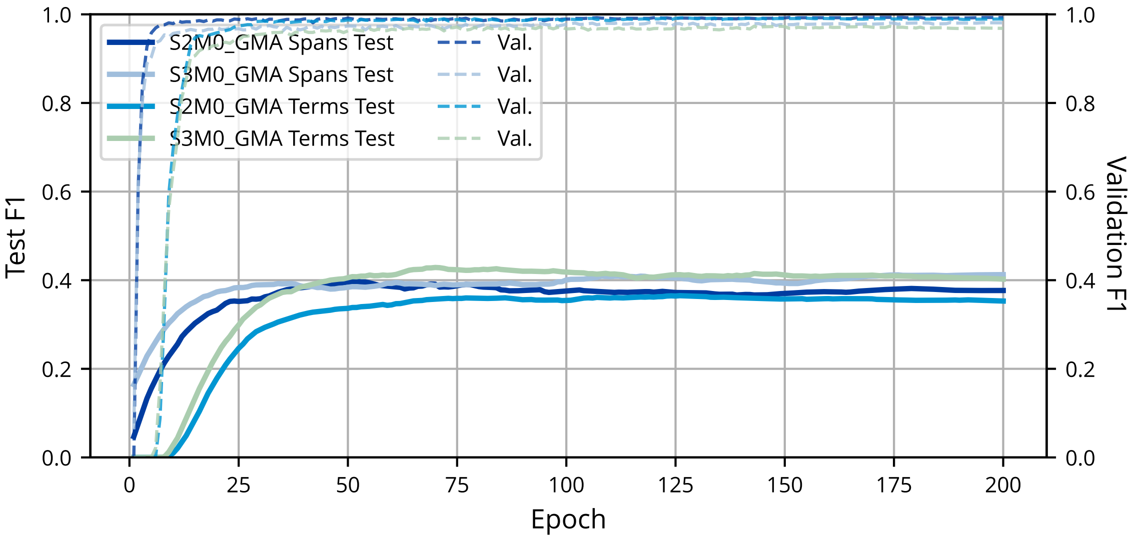


Figure 5.17: F1 evaluation of datasets resulting from the S3M0 and S2M0 generative queries.

A complementary view of the impact of sampling on the generated data is again provided by a PCA-based analysis of the embedding space. Figure 5.18 visualizes the distribution of generated samples for S4M0 in comparison to S3M0. In contrast to the Self-BLEU results, the PCA projection does not reveal pronounced structural differences in the arrangement of points between the two configurations. The clusters and overall spread appear largely similar, suggesting that, at the level of the lower-dimensional embedding used for visualization, the global structure of the generation space is preserved.

This observation indicates that the sampling strategy primarily introduces local variation among the outputs—reflected in reduced Self-BLEU—without fundamentally altering the broader semantic landscape captured by the embeddings. To further investigate these surface-level changes, the n-gram overlap statistics for S4M0 are analysed and compared to those of S3M0. Figure 5.19 presents the n-gram overlap scores for S4M0. Despite the marked difference in Self-BLEU, the n-gram profiles do not exhibit any substantial deviations from the S3M0 baseline. The distribution of common trigrams remains largely stable across the two configurations.

This suggests that sampling primarily affects how these n-grams are combined and arranged within individual outputs, rather than introducing entirely new lexical content. In other words, the vocabulary and typical phrase patterns remain similar, but the particular combinations in which they appear become more varied. Such a pattern is consistent with the design of S4M0, where the underlying corpus and generative request components are the same, yet the selection and ordering of examples changes from one generation to the next.

Ultimately, the impact of sampling must be assessed not only in terms of diversity metrics, but also with respect to downstream task performance. Figure 5.20 compares the evaluation

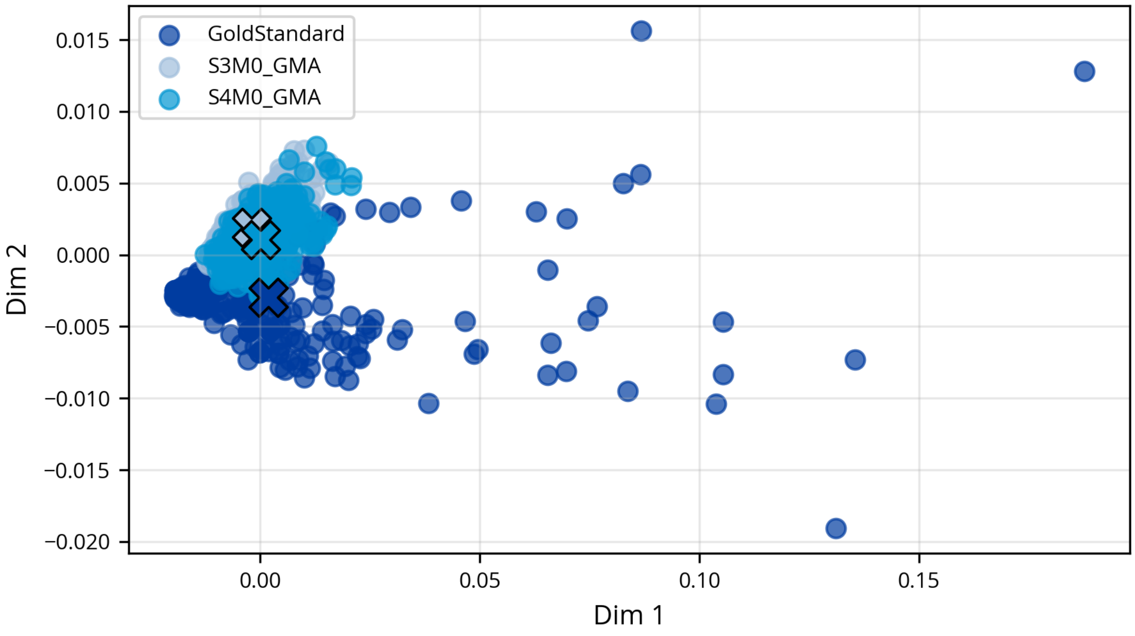


Figure 5.18: PCA visualization of datasets resulting from the S4M0 and S3M0 generative queries.

scores for S4M0 against those of S3M0 across all relevant metrics. The results show a consistent improvement in favour of S4M0. The F1 scores increase across the board, indicating that the more diverse training data produced by the sampled requests leads to better generalization and higher-quality predictions.

In summary, S4M0 demonstrates that incorporating sampling into the example selection process for query-based data generation can substantially increase output diversity, as evidenced by lower Self-BLEU scores in Table 5.5, while leaving the global embedding structure and n-gram statistics largely intact, as seen in Figures 5.18 and 5.19. Most importantly, the evaluation results in Figure 5.20 confirm that this increase in diversity translates into tangible performance gains. The configuration thus illustrates the practical benefits of replacing static in-context examples with dynamically sampled ones when generating synthetic training data under an otherwise unchanged request structure.

Interim Wrap-Up

The experimental configurations introduced up to this point (S1M0, S2M0, S3M0 and S4M0) provide a stepwise exploration of how generative request design, structural complexity, and example selection strategies influence both the quality and the utility of synthetically generated data for downstream NER tasks. Figure 5.21 summarizes the F1 evaluation results

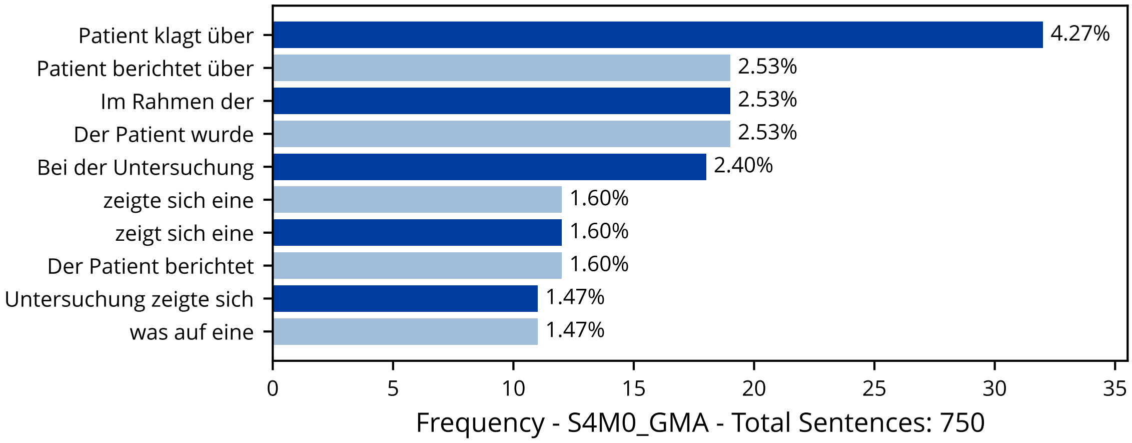


Figure 5.19: Top trigrams of datasets resulting from the S4M0 and S3M0 generative queries.

for all configurations and offers an integrative view of their comparative performance across span- and term-based metrics.

From the perspective of configuration complexity (RQ1), a clear positive correlation with downstream performance emerges. Moving from the simplest setups towards more elaborate ones. Particularly from S2M0 through to S4M0 results in consistent gains in F1 scores, as illustrated in Figure 5.21. Each incremental refinement of the generative request, such as providing more explicit instructions, stronger structural constraints, or richer contextual guidance, contributes to higher-quality synthetic outputs. These improvements in turn translate into more effective training data for the span- and term-level evaluation tasks. The evidence thus supports the view that increasing generative request complexity in a principled manner enhances the ability of the model to produce clinically meaningful and task-aligned sentences.

The conclusion based on the S1M0 dataset is not as straightforward. While the open-weight models failed to provide any meaningful output using this generative request configuration, the S1M0 dataset can be understood as a benchmark to compare the performance of state-of-the-art commercial models with open-weight models.

With respect to evaluation metrics (RQ2), Self-BLEU proves particularly informative for assessing diversity. Lower Self-BLEU scores correlate well with improved F1 performance, indicating that greater sample diversity benefits training. However, Self-BLEU captures only n-gram overlap patterns and misses other relevant dimensions such as semantic coverage, stylistic variation, and domain adherence. Therefore, Self-BLEU should be interpreted cautiously alongside complementary analyses like PCA-based embedding inspection and n-gram statistics to obtain a holistic view of data quality.

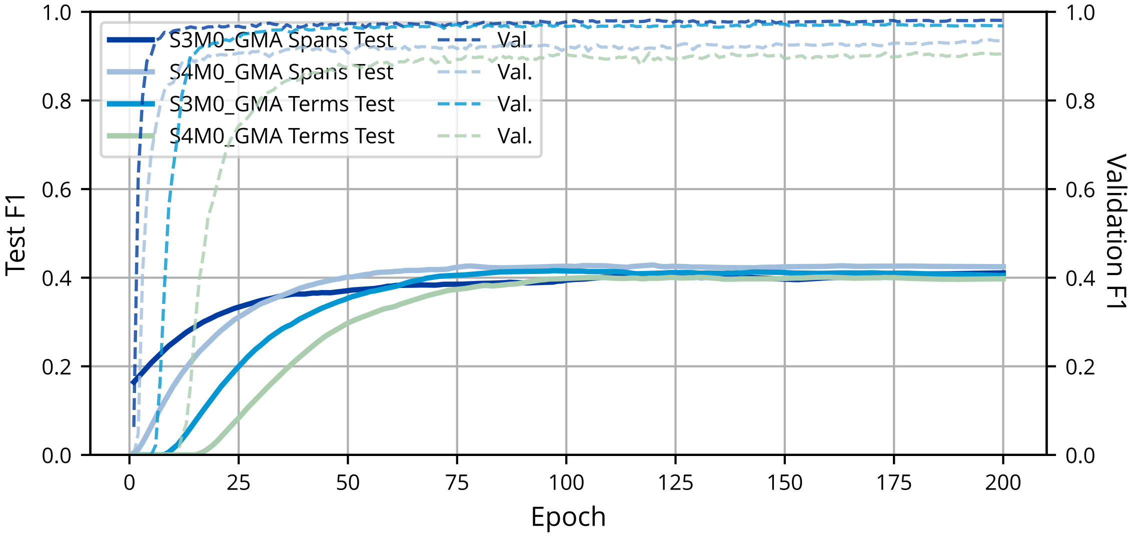
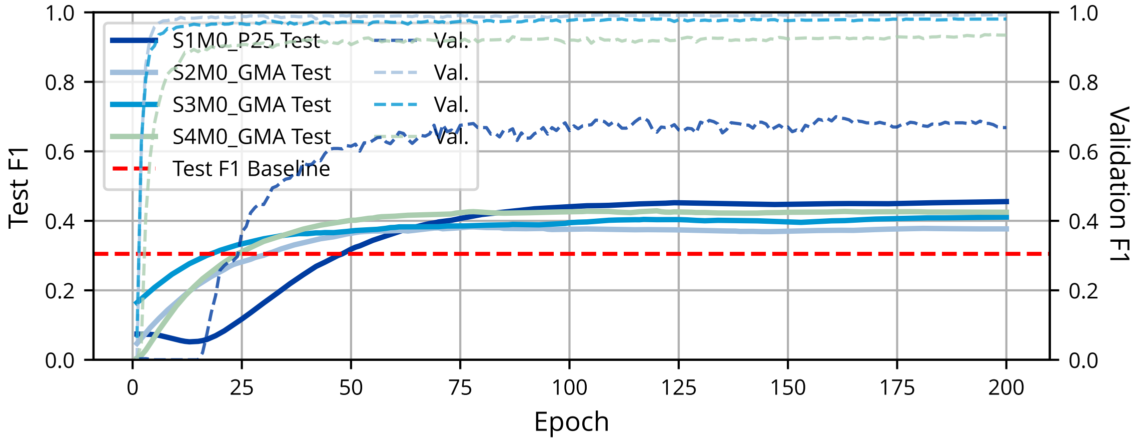


Figure 5.20: F1 evaluation of datasets resulting from the S4M0 and S3M0 generative queries.

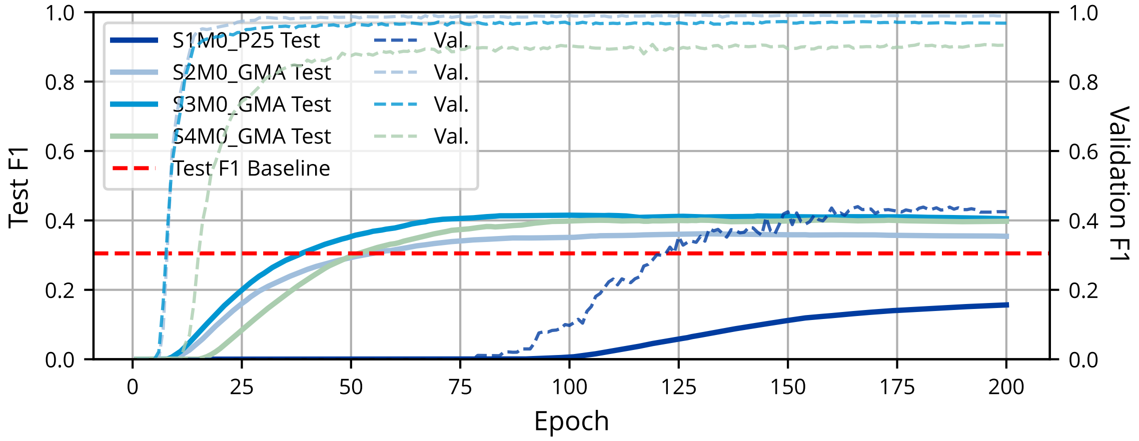
The comparison between open-weight models and commercial foundation models (RQ3) highlights a pronounced sensitivity to generative request design and task formulation. Configuration S1M0 employs a relatively simple request structure while requesting multiple distinct HPO terms within a single request. This setup imposes substantial demands on the model in terms of instruction following, contextual reasoning, and controlled generation. Under these conditions, commercial models demonstrate robust performance, whereas open-weight models struggle to meet the requirements, leading to a noticeable performance gap.

Subsequent configurations mitigate this challenge by simplifying and segmenting the requests, as seen in the progression towards S3M0 and S4M0. When the task is decomposed into smaller and more tightly guided components, open-weight models are able to generate substantially better synthetic data, narrowing the gap to their commercial counterparts. Nonetheless, a residual difference remains, particularly in terms of robustness and reliability under less constrained requests. Considering the ease of use of a straightforward design such as S1M0, commercial models retain a clear advantage: they achieve strong results even with minimal generative requests complexity, whereas open-weight models are more dependent on careful request construction and iterative refinement to reach comparable levels of performance.

Finally, regarding the question of whether synthetic-only data can suffice to train non-English medical NER models (RQ4), the empirical findings so far are encouraging. Across the configurations summarized in Figure 5.21, models trained exclusively on synthetic data achieve competitive F1 scores. The availability of a gold standard test set based on real-world data is crucial in this context, as it enables a rigorous assessment of how well the synthetic training material transfers to authentic clinical text.



(a) F1 Spans



(b) F1 Terms

Figure 5.21: Interim evaluation - generative queries for sentences with one HPO term.

In conclusion, the interim results across configurations S1M0 to S4M0 indicate that request complexity, diversity-oriented sampling strategies, and careful task decomposition are key levers for obtaining high-quality synthetic data for non-English medical NER. More complex and structured requests consistently lead to better downstream performance, open-weight models can approach the capabilities of commercial models when provided with sufficient guidance, and Self-BLEU emerges as a useful, albeit incomplete, proxy for diversity that correlates with F1 improvements. Most importantly, the experiments demonstrate that synthetic-only data, when generated under well-designed requests and validated against a real-world gold standard, can be sufficient to train effective medical NER models, thereby underscoring the practical utility of query-based synthetic data generation in resource-constrained settings.

5.2.2 Dataset Shape Alignment

After increasing the complexity of the configuration to generate sentences with a single HPO term, the next objective is to align the synthetic data more closely with the characteristics of the gold standard data, which represents real-world medical reports. While the previous configurations mainly focused on introducing and controlling the number of HPO terms per sentence, this step aims at better capturing the distribution of non-pathologic content and overall report structure observed in clinical practice. By doing so, the synthetic data is expected to provide a more realistic training signal for the downstream model.

To achieve this alignment, the configuration is extended to explicitly model sentences without any HPO terms. Real-world medical reports contain a substantial proportion of normal findings and neutral observations, and a synthetic dataset that neglects this aspect risks overemphasizing pathologic information. Including such no-terms sentences therefore serves both to mirror the natural class balance and to reduce the tendency of the model to overpredict pathological entities.

`S4M0_NT` - Generating no terms data

The first step towards better alignment with the gold standard data is the generation of a dedicated dataset that contains no HPO terms at all. This dataset is referred to as `S4M0_NT`. The design of this dataset is inspired by the strong performance observed with the Gemini 2.5 Pro configuration, as well as the high quality of the generated sentences when the configuration complexity was previously increased. The goal is to leverage these strengths to produce realistic, non-pathologic sentences that resemble the negative portions of real-world reports. The full configuration used for generating `S4M0_NT` is documented in Appendix 8.1.2.

To ensure that the no-terms dataset indeed contains no HPO-related entities, the baseline evaluation method introduced in Section 4.4 is applied. This method is used to systematically inspect the `S4M0_NT` data and verify that no entities are added inadvertently. In other words, it serves as a quality control step to help to confirm that the synthetic sentences truly represent non-pathologic content and do not encode hidden pathology.

The analysis of the 8500 sentences generated for `S4M0_NT` confirms that they are all non-pathologic. Typical examples include statements such as *Hautfarbe normal, keine Blässe.*, *Keine Rasselgeräusche, kein Giemen oder Brummen.*, *Kein Meningismus, Lasègue-Zeichen beidseits negativ.* or *Gangbild flüssig und sicher, keine Ataxie.*. These examples illustrate that the dataset consists of realistic clinical phrases describing normal findings and the absence of pathological signs, thereby closely matching the style and content found in actual medical documentation.

The impact of including the S4M0_NT data in the training process is first assessed in terms of diversity. Table 5.6 shows the Self-BLEU scores comparing the S4M0_GMA model trained with and without the no-terms data. A clear improvement in diversity can be observed when the S4M0_NT data is incorporated during training.

Model	Self-BLEU
S4M0_GMA	0.180
S4M0_GMA_NT	0.143

Table 5.6: Self-BLEU scores of the S4M0 generative query-based dataset resulting from the extension of no terms sentences.

Beyond diversity, the distributional similarity between synthetic and real data is examined using PCA. The PCA plots in Figure 5.22 compare the S4M0_GMA model with and without the inclusion of S4M0_NT. These visualizations reveal that adding no-terms data leads to a better coverage of the real-world data distribution. In particular, the synthetic samples spread more evenly across the embedding space occupied by the gold standard data, indicating that the model learns to generate a broader variety of clinically plausible sentences, including non-pathological findings.

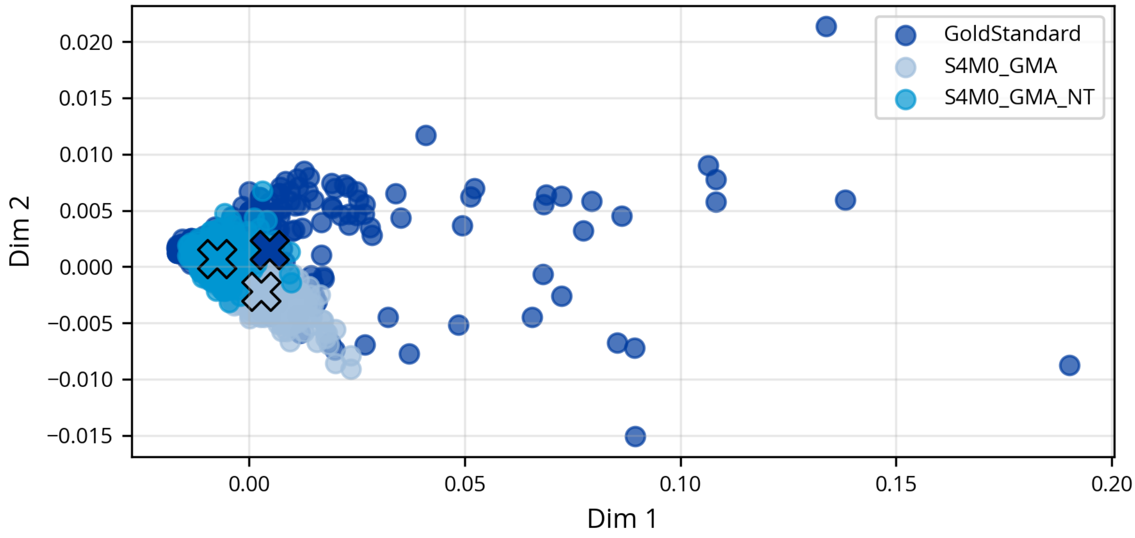


Figure 5.22: PCA visualization of the S4M0 generative query-based dataset resulting from the extension of no terms sentences.

Furthermore, the effect on local lexical structure is evaluated using n-gram statistics. The n-gram plots in Figure 5.23 show the distribution of short word sequences for the model trained with no-terms data. In comparison to the configuration without S4M0_NT (see Figure 5.19), a slight improvement in the n-gram distribution can be observed. This indicates that the inclusion of no-terms sentences helps the model more closely match the typical phrase patterns and collocations present in real clinical texts, without overemphasizing rare or highly specific pathologic expressions.

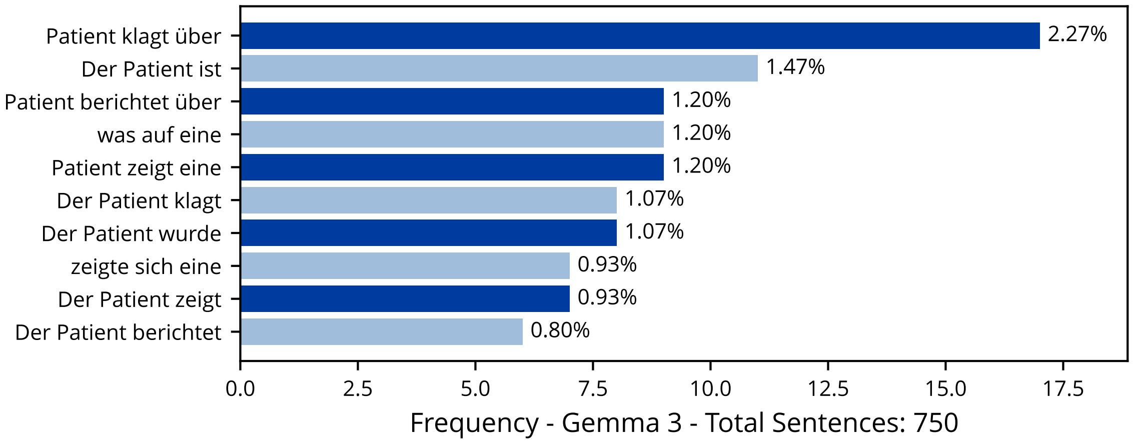


Figure 5.23: Top trigrams of the S4M0 generative query-based dataset resulting from the extension of no terms sentences.

However, the improved diversity and distributional alignment do not automatically translate into better overall extraction performance. When assessing the model using the evaluation setup described previously, a decrease in F1-score is observed when including the no-terms data, as illustrated in Figure 5.24. This indicates that the trade-off between precision and recall changes significantly once S4M0_NT is added, and that the model behaves more conservatively when predicting HPO entities.

A closer inspection of the individual components of the F1-score reveals the underlying dynamics more clearly. As shown in Figure 5.25, the recall drops notably when the no-terms data is included, while the precision improves substantially. This behaviour suggests that the model becomes more cautious in assigning HPO labels, preferring to miss some true entities rather than risk producing false positives. From an application perspective, this can be interpreted as a shift from a more sensitive towards a more specific model, which may be desirable or undesirable depending on the clinical use case.

To further understand this change in behaviour, confusion matrices for the two configurations are examined in Figure 5.26. These matrices provide a detailed view of the distribution of true positives, false positives, true negatives, and false negatives. It becomes apparent that the inclusion of the no-terms data leads to a reduction in the number of false positives, which directly explains the observed increase in precision. At the same time, the number of false negatives grows, which accounts for the decline in recall and thus the lower overall F1-score. This confirms that the model internalizes the presence of many normal, entity-free sentences and responds by raising its threshold for predicting HPO entities.

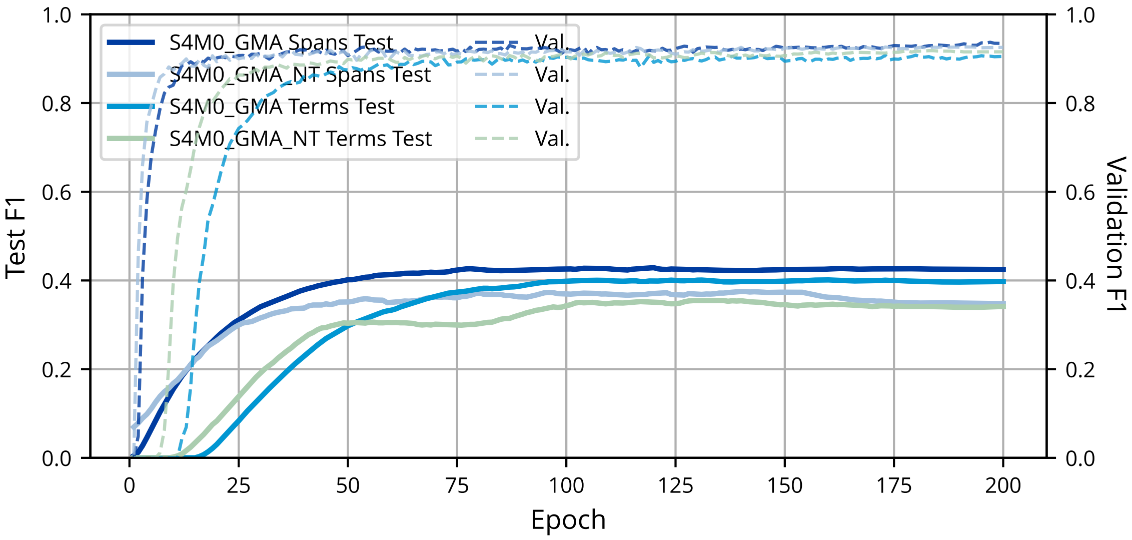


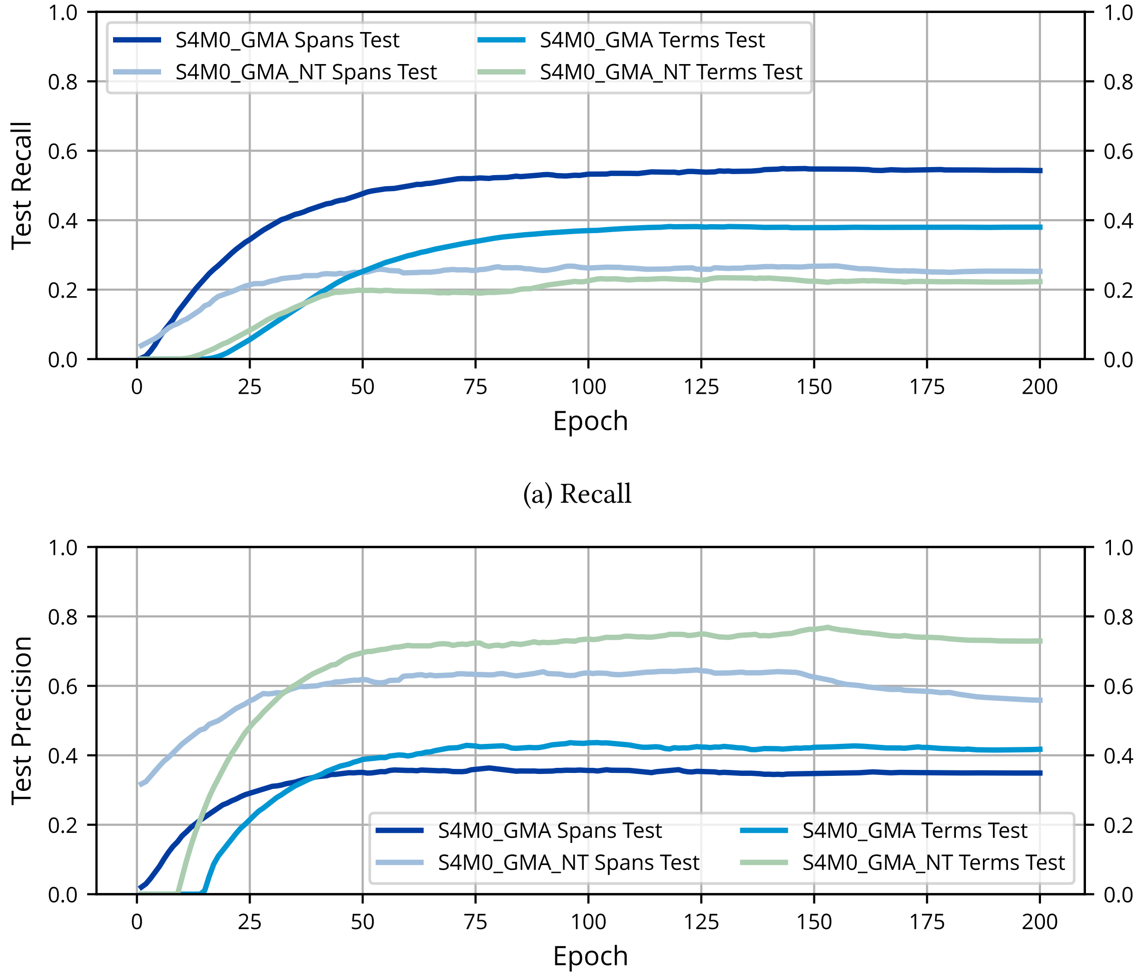
Figure 5.24: F1 evaluation of the S4M0 generative query-based dataset resulting from the extension of no terms sentences.

The analysis therefore highlights a central trade-off introduced by S4M0_NT: while the model becomes more reliable in terms of not over-predicting entities, it also becomes more conservative and misses a larger share of true annotations. The subsequent steps will therefore focus on rebalancing this trade-off, aiming to preserve the gains in precision and reduction of false positives while recovering as much recall as possible.

In summary, the introduction of the S4M0_NT no-terms dataset represents an important step towards aligning synthetic data with real-world medical reports. The additional data improves diversity, enhances coverage of the real data distribution, and yields more realistic n-gram patterns. At the same time, it induces a distinct shift in model behaviour: precision increases and false positives decrease, but recall is reduced and more false negatives occur. These findings underline the importance of carefully balancing normal and pathologic content in synthetic training data. The next stage of the work will therefore investigate strategies to mitigate the loss in recall while maintaining the benefits gained from the inclusion of no-terms data, with the overall aim of further improving the robustness and clinical usefulness of the model.

S4M0_EO - Adding entities only records

In this experiment, the focus lies on evaluating the impact of adding entity only (EO) records to the training data of the S4M0_GMA model. Unlike previous configurations, no additional contextual sentences or phrases are introduced. Instead, only the labels derived from the



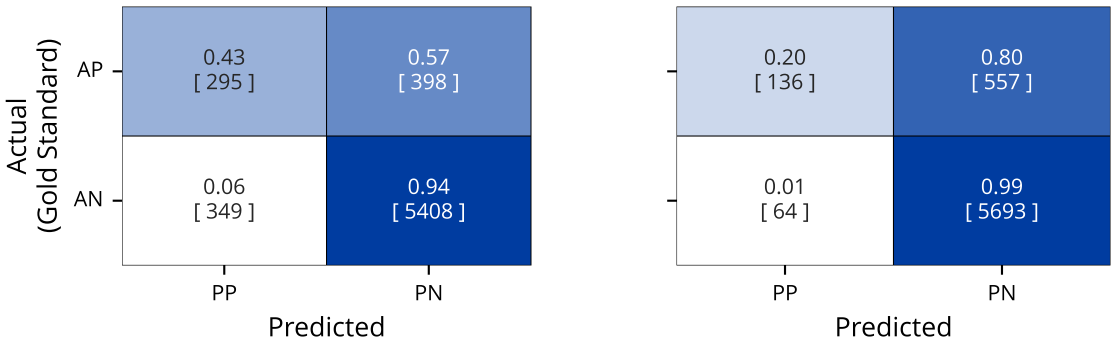
(a) Recall

(b) Precision

Figure 5.25: Precision and recall evaluation of the `S4M0` generative query-based dataset resulting from the extension of no terms sentences.

HPO terms are appended as isolated tokens. Consequently, the model is exposed exclusively to the target entities that should be learned, without any surrounding linguistic context that would normally embed these entities within natural language. This setup is conceptually similar to a gazetteer-style augmentation, where term lists are provided to the model with minimal structural information.

The underlying research question of this configuration is how strongly such gazetteer-like entity injection influences model performance. Specifically, it aims to assess whether exposing the model to a larger variety of raw entity labels, devoid of additional context, is sufficient to improve recognition capabilities, or whether this comes at the cost of degraded precision due to overgeneralisation.



(a) S4M0

(b) S4M0_NT

Figure 5.26: Confusion matrices of the S4M0 generative query-based dataset resulting from the extension of no terms sentences. AP: Actual Positives, AN: Actual Negatives, PP: Predicted Positives, PN: Predicted Negatives.

Figure 5.27 visualises the impact of the EO data on the latent space representation using a PCA projection, comparing S4M0 with and without EO augmentation. The distribution of samples exhibits a noticeable, yet relatively moderate, shift. This indicates that the EO records exert an effect on the learned representations. The embedding space is therefore influenced by the additional entities, but the overall geometry remains relatively close to previous versions.

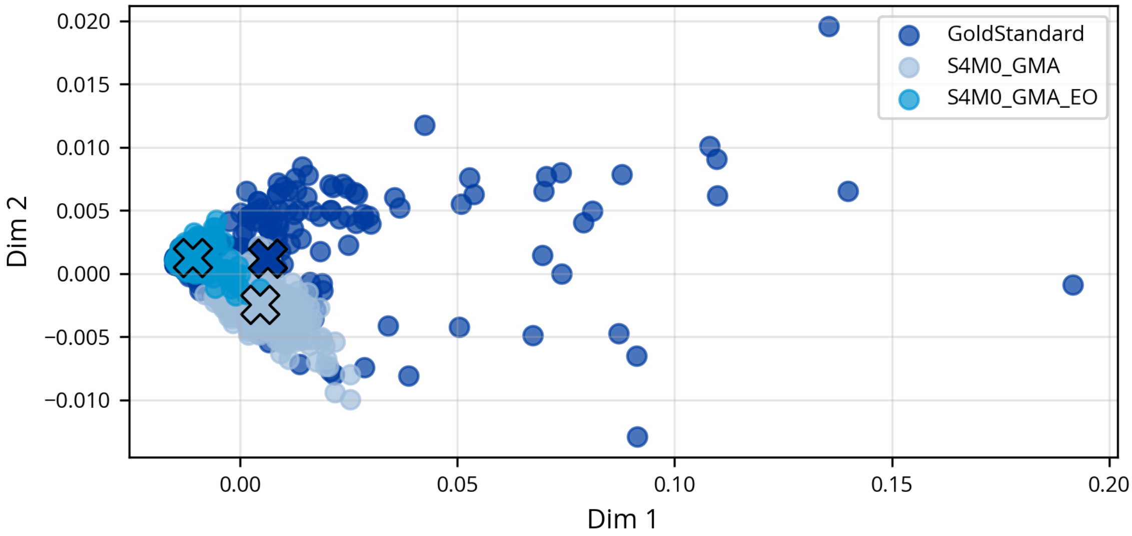


Figure 5.27: PCA visualization of the S4M0 generative query-based dataset resulting from the extension of entities only records.

The quantitative evaluation of this configuration reveals that the inclusion of EO data does not improve the overall span-level performance metrics. As shown in Figure 5.28, the F1 scores of the EO-augmented model are lower than those of the base S4M0_GMA model without additional data. This behaviour is consistent with the trends observed in the previous experiment that employed non-pathological, no terms (NT) sentences as augmentation. In both cases, the addition of synthetic or weakly contextualised data leads to a degradation in aggregate performance, suggesting that the model benefits more from high-quality contextual examples than from large quantities of minimally informative signals.

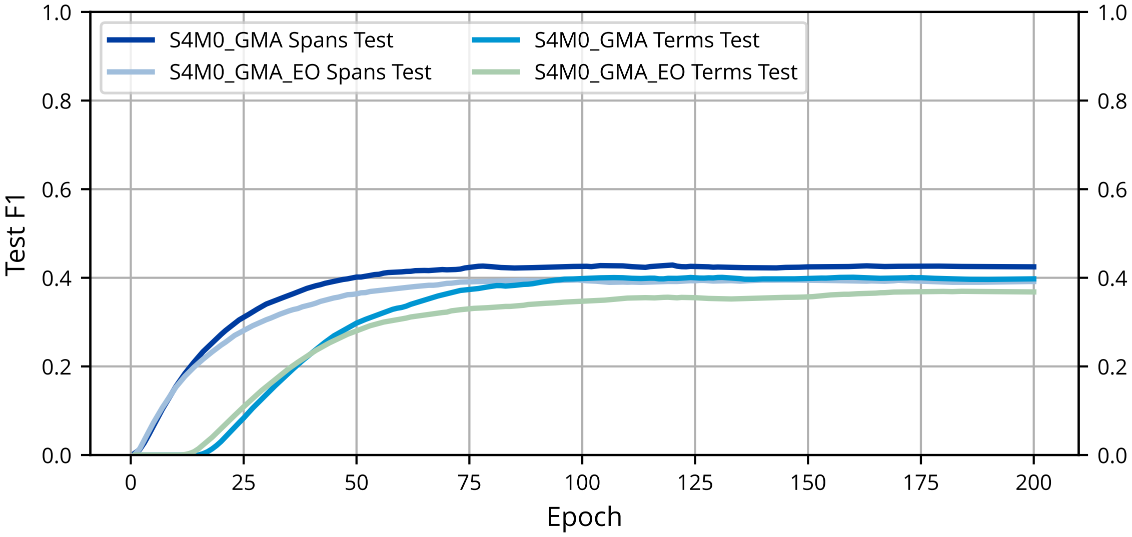
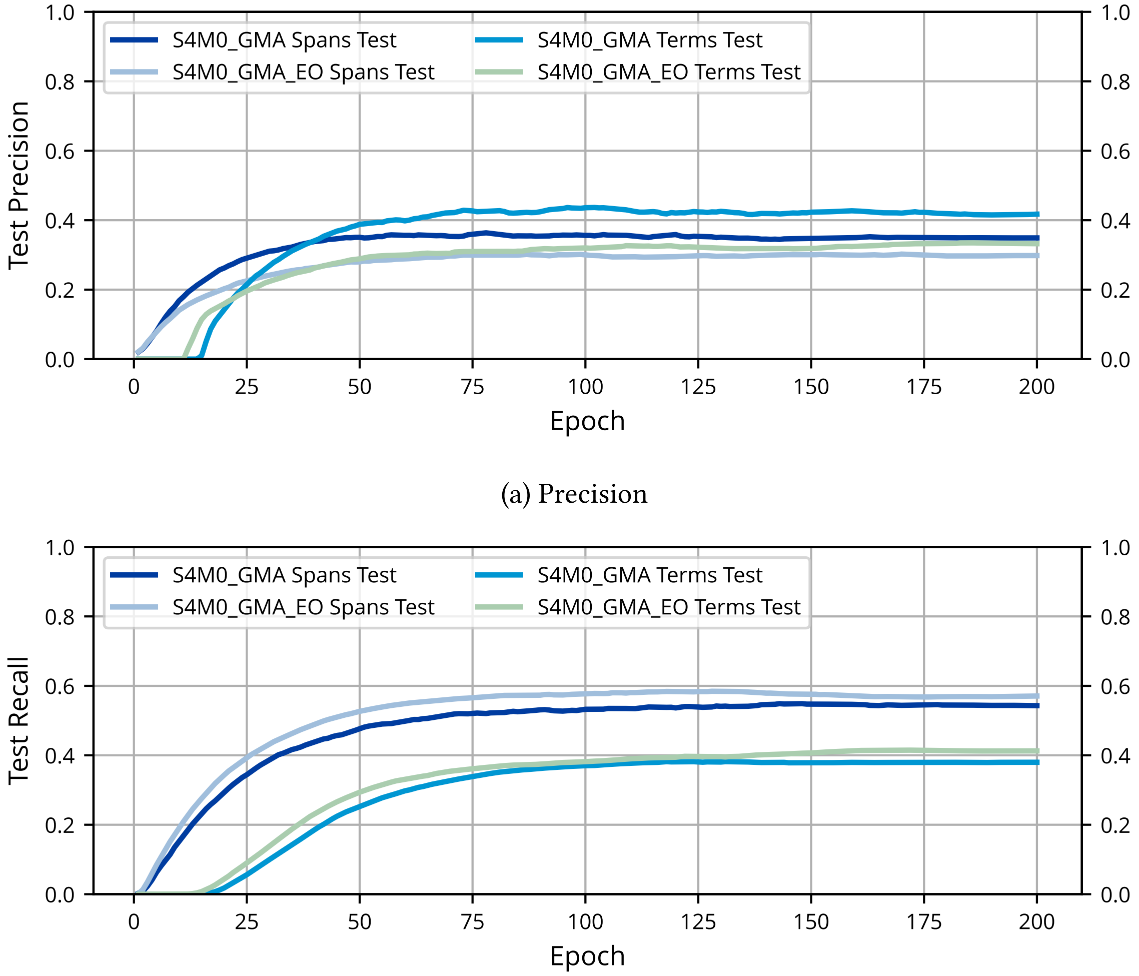


Figure 5.28: F1 evaluation of the S4M0 generative query-based dataset resulting from the extension of entities only records.

A more differentiated picture emerges when examining precision and recall separately. Figure 5.29 compares these metrics for S4M0_GMA with and without EO data. The recall is slightly improved when EO examples are added, indicating that the model becomes more capable of identifying a larger portion of the relevant entities present in the text. At the same time, precision drops markedly, which means that the model produces more incorrect or spurious entity predictions.

This trade-off stands in contrast to the previous NT-based experiment, where a slight improvement in precision was observed at the expense of recall. In other words, while NT augmentation encouraged the model to be more conservative and precise, EO augmentation encourages a more liberal prediction behaviour that favours coverage over correctness. This suggests that gazetteer-like exposure to many entity labels increases the tendency of the model to label spans as entities whenever they resemble known terms, even when the context does not fully support such a decision.

To substantiate these observations, the confusion matrices of both models are compared in Figure 5.30. The EO-augmented model exhibits a slight reduction in the number of false



(a) Precision

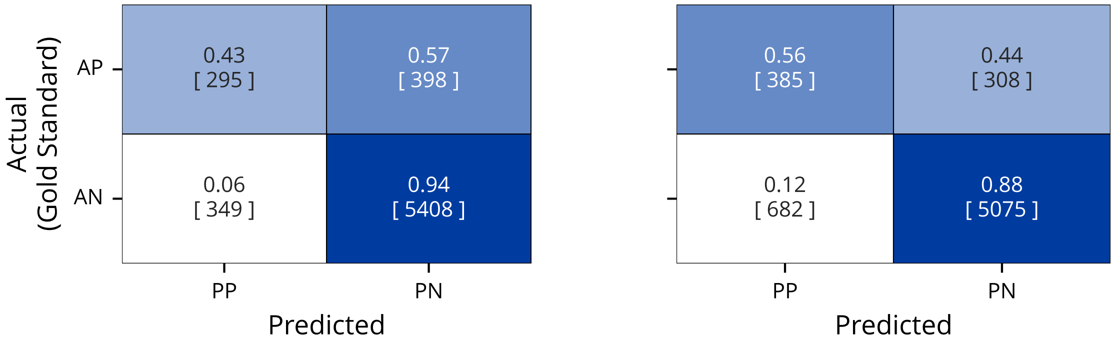
(b) Recall

Figure 5.29: Precision and recall evaluation of the `S4M0` generative query-based dataset resulting from the extension of entities only records.

negatives, which corresponds to the observed improvement in recall. More relevant entities are successfully identified, reducing the proportion of missed spans.

However, this benefit is accompanied by an increase in false positives, which explains the considerable loss in precision. The model more frequently assigns entity labels to spans that do not correspond to true entities, likely because the memorised entity labels from the EO records are applied too broadly. This behaviour is typical for systems exposed to large gazetteer lists without sufficient contextual constraints, as they tend to match surface forms while underutilising contextual cues required for precise disambiguation.

In summary, the EO experiment demonstrates that adding entities only records have a clear and systematic, but not uniformly beneficial, impact on model behaviour. The latent space



(a) S4M0

(b) S4M0_EO

Figure 5.30: Confusion matrices of the S4M0 generative query-based dataset resulting from the extension of entities only records.

and decision boundaries of the model are adjusted in such a way that more true entities are detected, improving recall to a limited extent. At the same time, the absence of contextual information in the EO examples encourages over-prediction, leading to a substantial increase in false positives and, consequently, a drop in precision and overall F1 score as seen in Figure 5.28.

These findings indicate that gazetteer-like augmentation alone is insufficient to yield robust performance gains in this setting. While it may be useful as a mechanism to enhance coverage of underrepresented entities, it needs to be combined with context-rich training data or additional regularisation strategies in order to mitigate the Precision loss observed in Figures 5.29 and 5.30. Overall, the results highlight the importance of contextual information for effective entity recognition and suggest that entity lists should be integrated carefully rather than in isolation.

Effects of subsampling and different subsampling strategies

Subsampling is a common strategy to reduce the computational cost of model training by limiting the amount of data used in each experiment. In the context of this thesis, subsampling is investigated with respect to its impact on model performance and the characteristics of the resulting training sets. Two subsampling strategies are compared: simple random subsampling and maximal marginal relevance (MMR)-based subsampling. While random subsampling selects instances without considering their content or redundancy, MMR-based subsampling aims to construct a more diverse subset of training examples by balancing relevance and dissimilarity between selected instances. This section analyses the influence of these strategies on the learned models and examines whether previously reported benefits of MMR-based selection can be reproduced.

The motivation for using MMR-based subsampling stems from prior work indicating that diversity-aware selection can improve model robustness and generalization. In particular, Ye et al. (2023) and Abadifard et al. (2023) report that selecting examples to maximize coverage of different patterns or perspectives can enhance downstream performance in various tasks. By analogy, it is expected that MMR-based subsampling could provide a more representative and heterogeneous training set than purely random selection, potentially preserving performance even when the overall number of training instances is reduced. The experiments conducted in this thesis therefore investigate whether these findings transfer to the examined setting and data.

An initial qualitative analysis of the subsampled datasets is based on n-gram distributions. Figure 5.31 compares the n-gram statistics of datasets obtained through random subsampling and MMR-based subsampling. If MMR-based selection indeed produced a more diverse training set, one might expect noticeable differences in the frequency and variety of observed n-grams, reflecting a broader coverage of linguistic patterns. However, the plotted distributions do not reveal a substantial deviation between the two strategies. The overall shape and relative frequencies of the most common n-grams appear similar, suggesting that, at least on the level of surface-form n-gram statistics, MMR-based subsampling does not induce a markedly different data distribution compared to random subsampling.

The effect of subsampling on downstream model performance is evaluated in terms of F1 score on spans, as shown in Figure 5.32. This figure contrasts models trained on the full training set with models trained on subsampled data using different strategies. Unsurprisingly, reducing the amount of available training data leads to a degradation in performance: models trained with any form of subsampling consistently underperform those trained on the complete dataset. This confirms the intuitive expectation that, all else being equal, more training examples enable the model to capture a wider range of patterns and edge cases, resulting in better overall performance.

When comparing the two subsampling strategies directly, the results in Figure 5.32 indicate no clear or systematic advantage of MMR-based subsampling over simple random subsampling. The observed F1 scores are closely aligned, and any differences fall within the range that can plausibly be attributed to experimental noise or minor variations in training. Thus, within the scope of these experiments and for the considered data and model configuration, MMR-based subsampling does not yield the expected improvement in performance relative to random selection. This stands in contrast to the positive effects reported by Ye et al. (2023) and Abadifard et al. (2023), suggesting that the benefits of MMR-based selection may be task-, model-, or dataset-dependent.

In summary, the analysis conducted in this section asserts that subsampling reduces training time but at the cost of lower predictive performance compared to using the full dataset. Furthermore, the comparison between random and MMR-based subsampling does not reveal

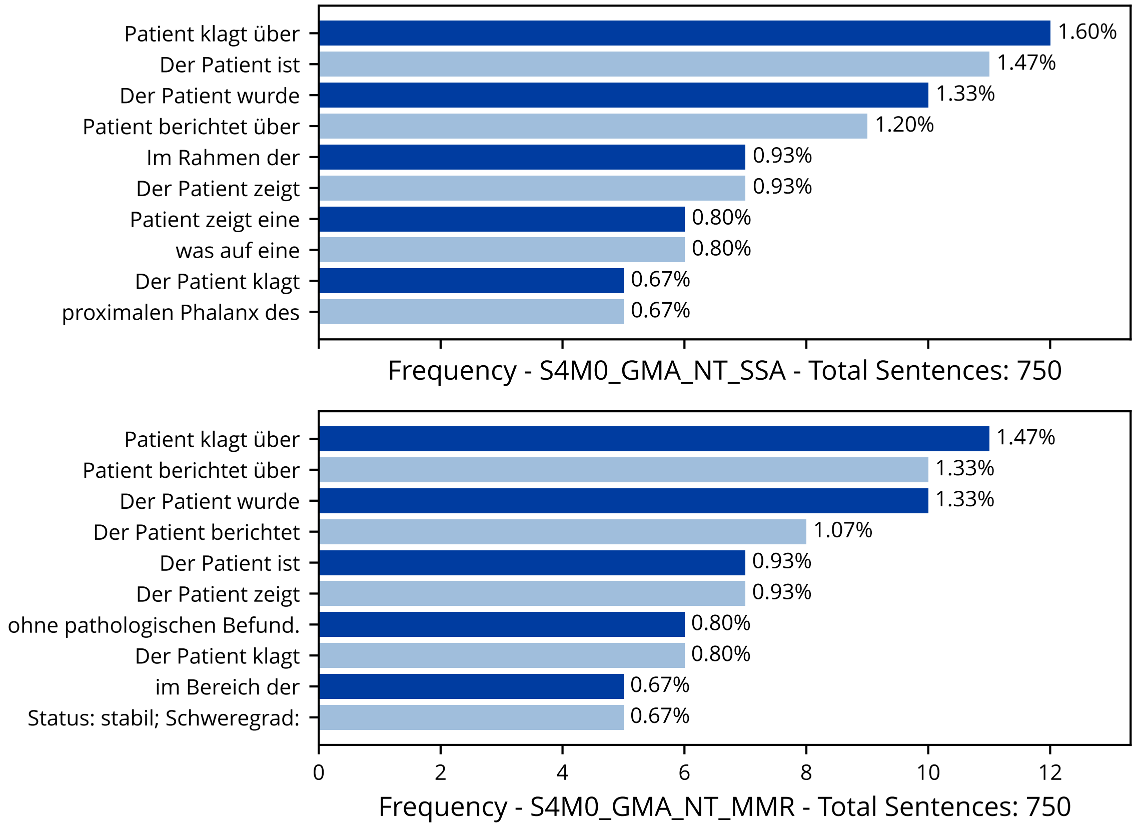


Figure 5.31: Top trigrams of the S4M0_NT dataset using different subsampling strategies.

substantial differences, neither in surface-level n-gram statistics (Figure 5.31) nor in span-level F1 scores (Figure 5.32). Consequently, although MMR-based subsampling is theoretically motivated as a diversity-enhancing strategy, its practical advantage over random subsampling is not confirmed in this experimental setup. Given the need to limit computational resources, subsampling will nevertheless be employed in subsequent experiments to reduce training time, with the understanding that this introduces a controlled and quantified decrease in performance relative to training on all available data.

Combining S4M0_NT and S4M0_EO

In order to further improve the quality and robustness of the embedding-based retrieval, the two methods S4M0_NT and S4M0_EO are combined into a single approach. For each document, the embeddings obtained from S4M0_NT and from S4M0_EO are first concatenated, and then MMR is applied to this joint representation. This combined method is denoted as S4M0_NT_EO_MMR. The underlying assumption is that S4M0_NT and S4M0_EO capture complementary aspects of

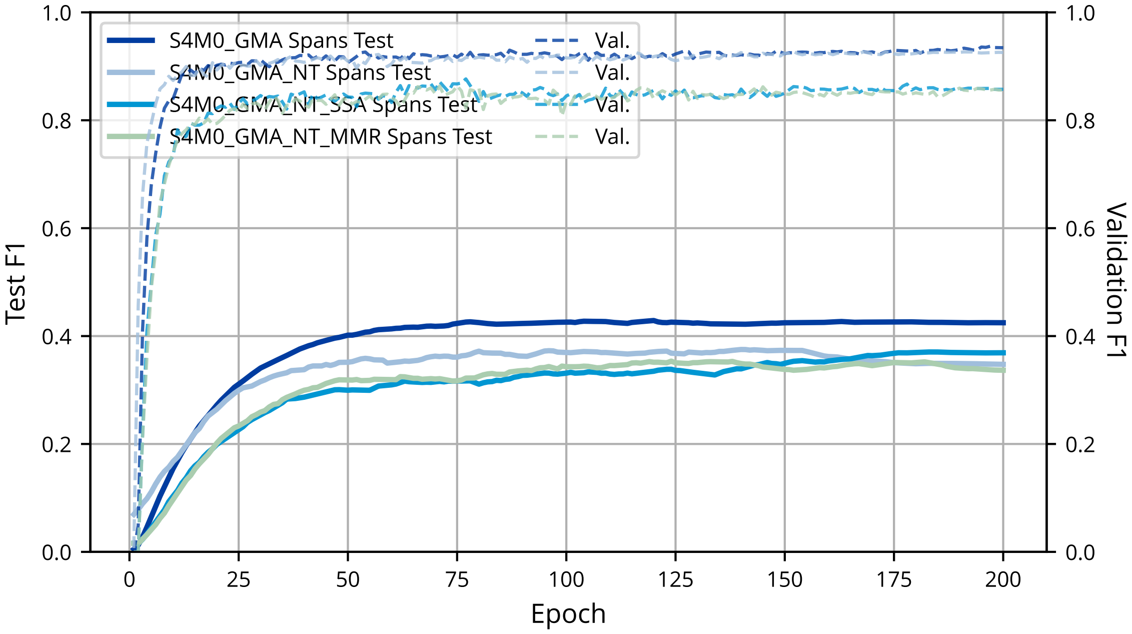


Figure 5.32: F1 evaluation of the S4M0_NT dataset using different subsampling strategies.

the documents, and that a joint embedding will provide a richer representation in the gold standard embedding space.

Figure 5.33 shows a PCA projection of the distinct components S4M0 as base dataset, S4M0_NT and S4M0_EO together with the reference embeddings. The documents in the combined embedding space appear more widely dispersed, while still covering a large portion of the area occupied by the gold standard embeddings. This spread indicates that the combined method can represent a wider variety of semantic nuances.

A closer inspection of Figure 5.33 also reveals clusters with distinct shapes. Within these clusters, documents originating from different methods are located in close proximity to one another. This behaviour suggests that, despite their different construction principles, S4M0_NT and S4M0_EO often map semantically similar documents into similar regions of the embedding space. The combined components therefore appear capable of aligning and integrating the strengths of both individual approaches, leading to more coherent and semantically meaningful document groupings.

The impact of this combination on retrieval quality can be seen in the evaluation results presented in Figure 5.34. When compared to the baseline S4M0_GMA setup, the combined method S4M0_NT_EO_MMR achieves substantially higher F1-scores in the spans evaluation. This strong improvement indicates that the joint embedding space more accurately captures the boundaries and content of relevant text spans.

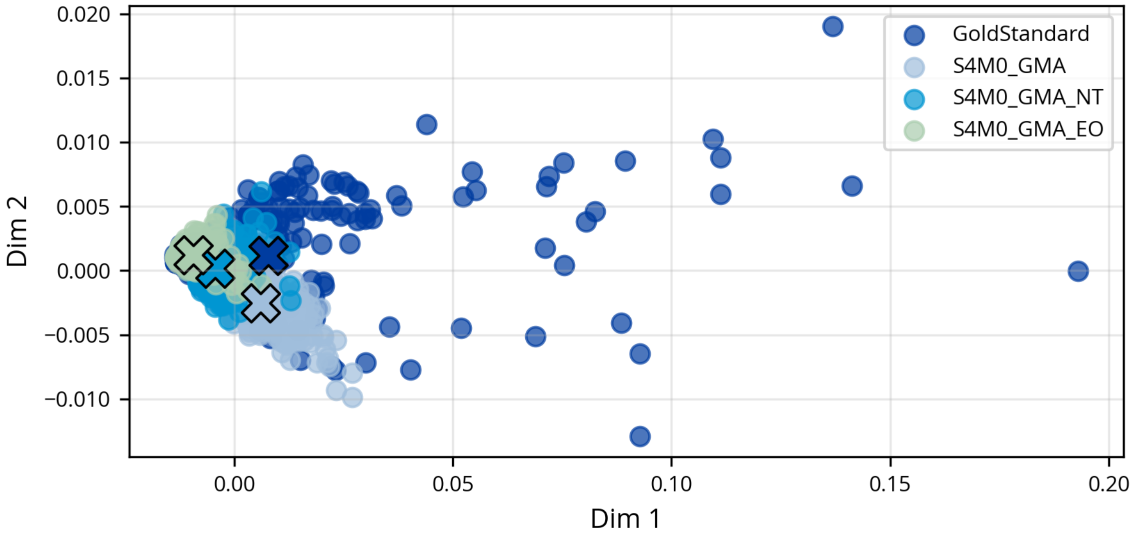


Figure 5.33: PCA visualization showing `S4M0` and its `NT` and `EO` extensions in the embedding space.

In contrast, the terms evaluation remains largely unaffected by the combination of `S4M0_NT` and `S4M0_EO`. The term-level performance being stable suggests that both individual methods already provide a sufficiently precise representation at the term granularity, and that the additional information introduced through concatenation does not translate into further measurable gains for this specific evaluation dimension.

Overall, these results indicate that combining `S4M0_NT` and `S4M0_EO` within a single framework is a promising strategy for improving span-level retrieval quality, while preserving the strong term-level performance of the individual methods.

Noise Injection

Noise injection is a technique used to improve the robustness and generalization capabilities of machine learning models by deliberately introducing perturbations, such as typos or character-level noise, into the input data during training. The central idea is that exposure to imperfect or noisy input better reflects real-world usage conditions, where user-generated text often contains spelling errors, informal language, or other inconsistencies. By simulating these imperfections during training, models can learn to maintain stable performance even when the input deviates from the clean, well-formed examples typically found in curated datasets.

Several studies have shown that noise injection can lead to improved performance when models are deployed in realistic settings that contain typos and other forms of input corruption

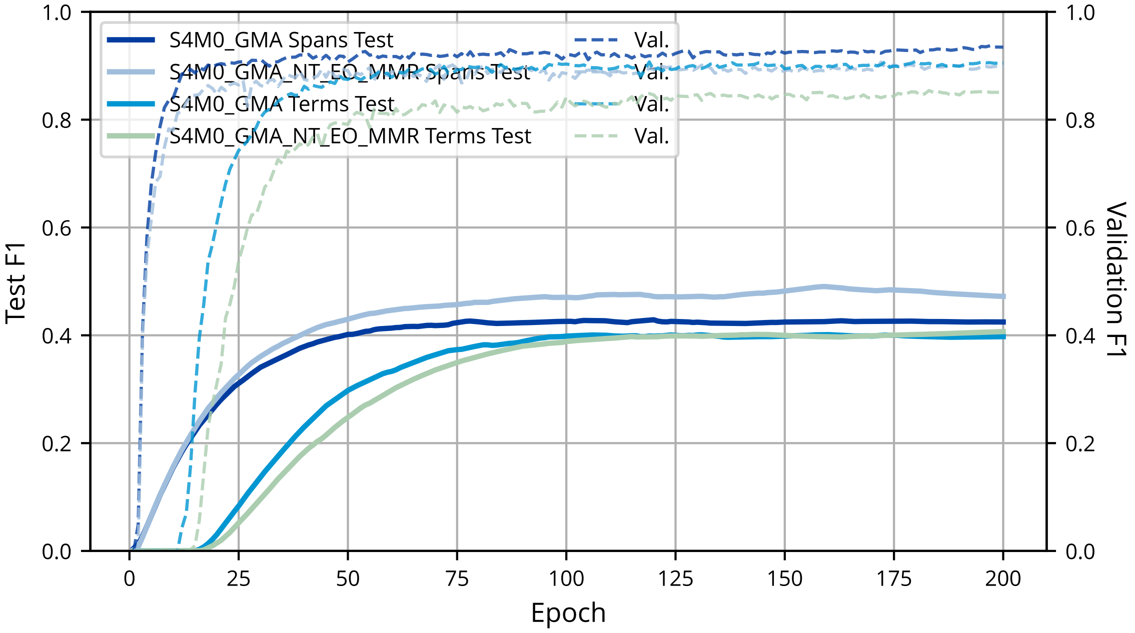


Figure 5.34: F1 evaluation of the `S4M0_NT_EO_MMR` dataset compared to `S4M0`.

(Hua et al., 2021; Liu et al., 2025). These works demonstrate that models trained exclusively on clean data tend to be brittle when confronted with noisy inputs, resulting in degraded prediction quality. In contrast, models trained with appropriate levels and types of noise are better able to tolerate such perturbations, thus yielding more reliable outputs in practical applications. Beyond mere robustness, some research even reports that carefully calibrated noise injection can enhance overall model performance, not only under noisy conditions but also on standard evaluation benchmarks (Duan et al., 2024). This effect is often attributed to a regularization-like behaviour, where noise prevents overfitting to idiosyncrasies in the training data and encourages the model to learn more generalizable patterns.

The empirical results in this work are consistent with these observations. Figure 5.35 presents a comparison of F1 scores between models trained with and without noise injection. The figure shows slight but consistent improvements when adding noise injection during training. Since the gold standard used for evaluation does not contain typos or other forms of textual noise, the measured gains represent a conservative estimate of the potential benefits. In a real-world scenario, where user input is more likely to contain spelling mistakes, abbreviations, or otherwise degraded text quality, the positive impact of noise injection can reasonably be expected to be marginally higher. Consequently, noise injection emerges as a practical and low-cost strategy to increase the robustness and applicability of text processing models in production environments.

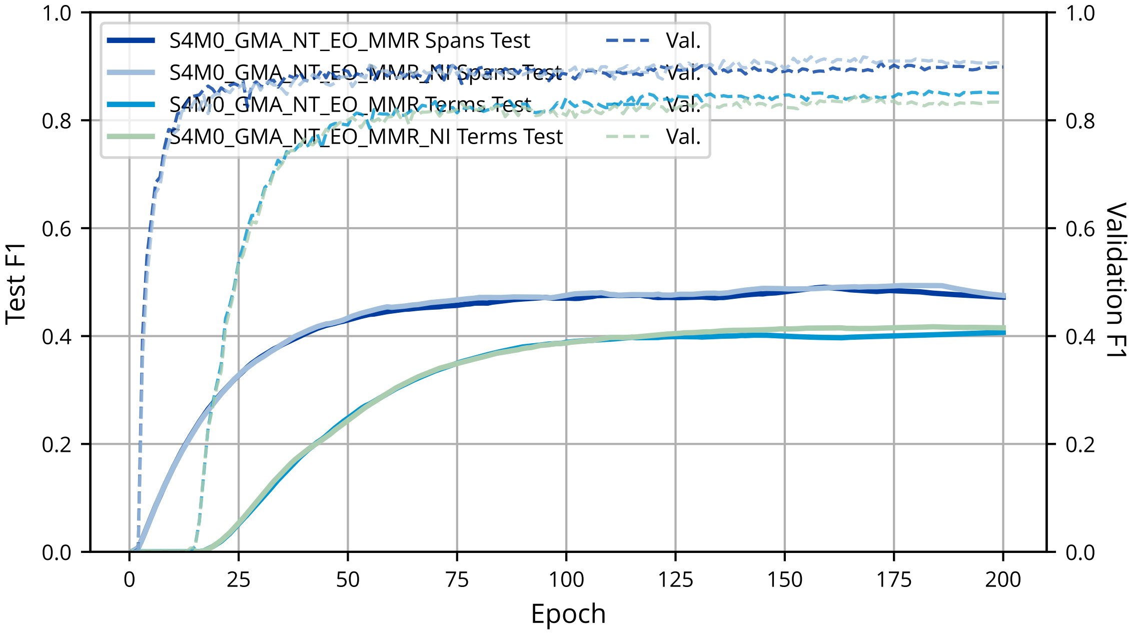


Figure 5.35: F1 evaluation comparing `S4M0_NT_EO_MMR` dataset with and without noise injection.

In conclusion, noise injection serves as an effective mechanism to align model training conditions with the noisy nature of real-world text input. The findings illustrated in Figure 5.35, together with previously reported results in the literature (Duan et al., 2024; Hua et al., 2021; Liu et al., 2025), indicate that incorporating noise during training can both mitigate sensitivity to input corruption and, in some settings, even enhance overall performance. Thus, noise injection represents a valuable addition to the set of techniques aimed at improving robustness, and it is particularly relevant for applications that must operate reliably on heterogeneous and error-prone user-generated text.

5.2.3 Synthetic Sentences Multiple HPO Terms

Now follows the last part of this research work, which investigates how to generate sentences containing multiple HPO terms within a single sentence. This extension of the data generation process aims to move closer to realistic clinical language, where multiple phenotypic concepts frequently co-occur in the same context. By doing so, the resulting synthetic data is expected to better reflect real-world sentence structures and semantic dependencies between terms, and thus to provide more informative training material for downstream models.

`S4M1` - Generate sentences with multiple HPO terms

To achieve the goal of generating sentences with multiple HPO terms, more in-depth configurations and stricter validation methods are required compared to the previous single-term setup. The overall objective is to design a generative request and verification pipeline that reliably produces sentences in which several HPO concepts are correctly mentioned, clearly marked, and consistently documented.

The core idea is as follows. Multiple HPO terms are proposed to the LLM, each accompanied, when applicable, by synonyms and definitions, and each identified by a unique identifier. The model is then instructed to generate sentences that include a random subset of these terms. This randomization introduces variety into the co-occurrence patterns of terms and helps to mitigate overly deterministic or repetitive combinations in the generated corpus. The full generative request is available in Appendix 8.1.3.

To keep the generation process manageable and to increase the likelihood of success, the model is asked to generate a single sentence at a time. Within each sentence, all occurrences of the selected HPO terms must be marked in bold syntax. In addition, the output must include a structured list of all terms that appear in the sentence, together with their corresponding identifiers. This output format is crucial, as it enables reliable downstream parsing of the synthetic text and facilitates automatic consistency checks.

The generated sentences then undergo a validation procedure. First, it is verified that every term that appears in the sentence in bold form is also present in the resulting list of included terms. Second, it is checked that the resulting list is a full subset of the originally proposed terms, i.e., that no spurious or unrequested terms have been added. Only sentences that satisfy these constraints are admitted into the final training set. Through this process, the `S4M1` configuration produces synthetic data that contains multiple, correctly annotated HPO terms in each sentence while maintaining internal consistency between text and metadata.

Figure 5.36 shows the PCA plot for the `S4M1` model compared to the `S4M0` model. To obtain a clearer picture of the newly generated data in `S4M1`, only sentences with multiple terms per sentence are displayed. The visualization highlights how the extended multi-term generation affects the distribution of the synthetic samples in the embedding space.

The newly added data points are well scattered outside the previously existing data points from S4M0, which indicates that the new data adds more variety to the training set. Overall, more overlap between the synthetic and real data points can be seen, suggesting that the multi-term generation in S4M1 shifts the synthetic examples closer to the space occupied by real clinical sentences.

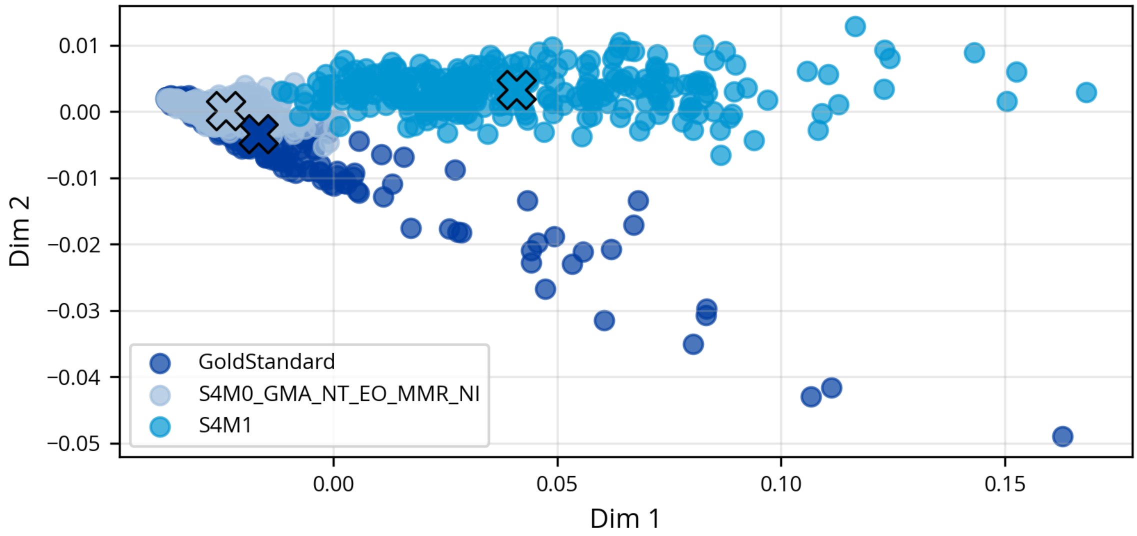


Figure 5.36: PCA visualization showing the previous S4M0_NT_EO_MMR_NI dataset and the newly generated S4M1 data points.

The models trained with S4M1 data therefore contain the newly generated multi-term sentences and extend the existing data that was generated with S4M0_NT_EO_MMR_NI. This enrichment of the training corpus is expected to benefit the learning of more complex contextual relationships and span boundaries, since the model is exposed to sentences where several phenotypic entities co-occur.

Figure 5.37 shows the evaluation F1 scores for the spans of the models trained with S4M1 data compared to the models trained with S4M0 data. The comparison allows an assessment of whether the added complexity and variety in S4M1 translate into measurable improvements in downstream sequence labelling performance.

The span-level evaluation shows no significant improvement when using the S4M1 data compared to S4M0. This indicates that the ability of the model to identify the correct text spans corresponding to phenotypic entities remains largely unchanged. In contrast, the term-level evaluation scores significantly improve when using the S4M1 data, demonstrating that the richer and more varied multi-term sentences help the model to better recognize and classify the underlying HPO concepts.

For the first time in this series of experiments, the term-level evaluation comes closer to the span-level evaluation. This narrowing of the performance gap suggests that the synthetic multi-term sentences enable the model to improve its understanding of the mapping between textual expressions and ontology terms. To further analyse how well the models perform under conditions that resemble real-world use cases, the Nervaluate results are considered to better gauge practical performance.

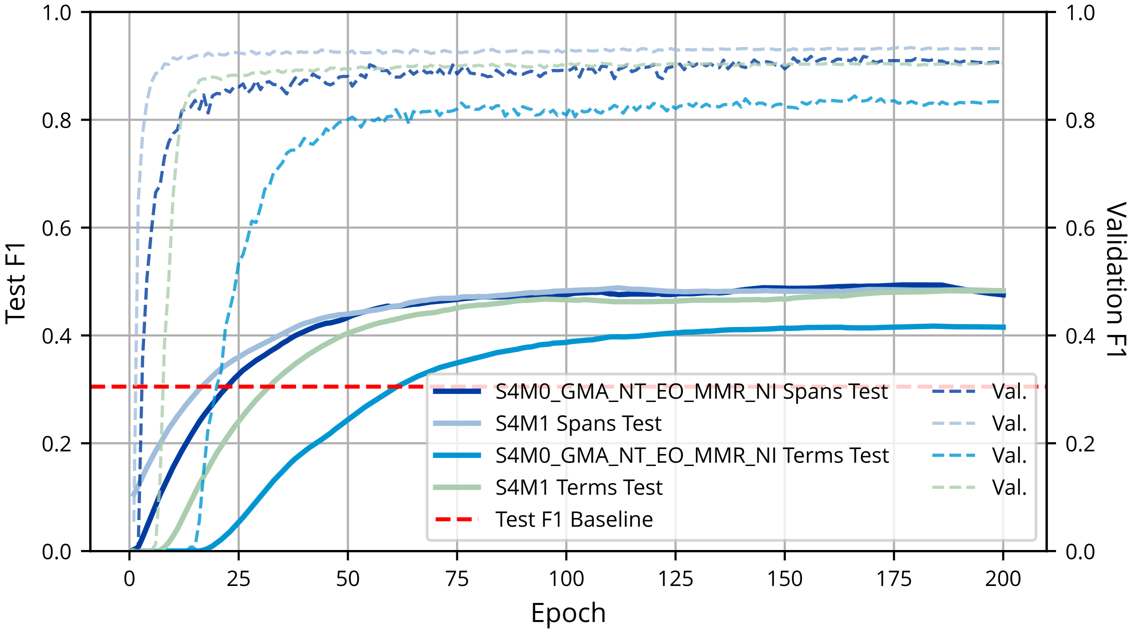


Figure 5.37: F1 evaluation of the previous S4M0_NT_EO_MMR_NI dataset with and without the extension of the S4M1 data points.

To test the MMR subsampling method once more in the context of multi-term sentences, the following experiment evaluates the Nervaluate F1 score for the S4M1 data using different subsampling strategies. In particular, the performance of models trained on data selected via MMR subsampling is compared to that of models trained on data obtained through random subsampling.

Again, no significant difference can be observed when using MMR subsampling compared to random subsampling, as illustrated in Figure 5.38. This suggests that, under the conditions studied here, the additional computational complexity of MMR-based selection does not yield clear benefits over simpler random selection for the Nervaluate metric, even when dealing with more diverse multi-term sentences.

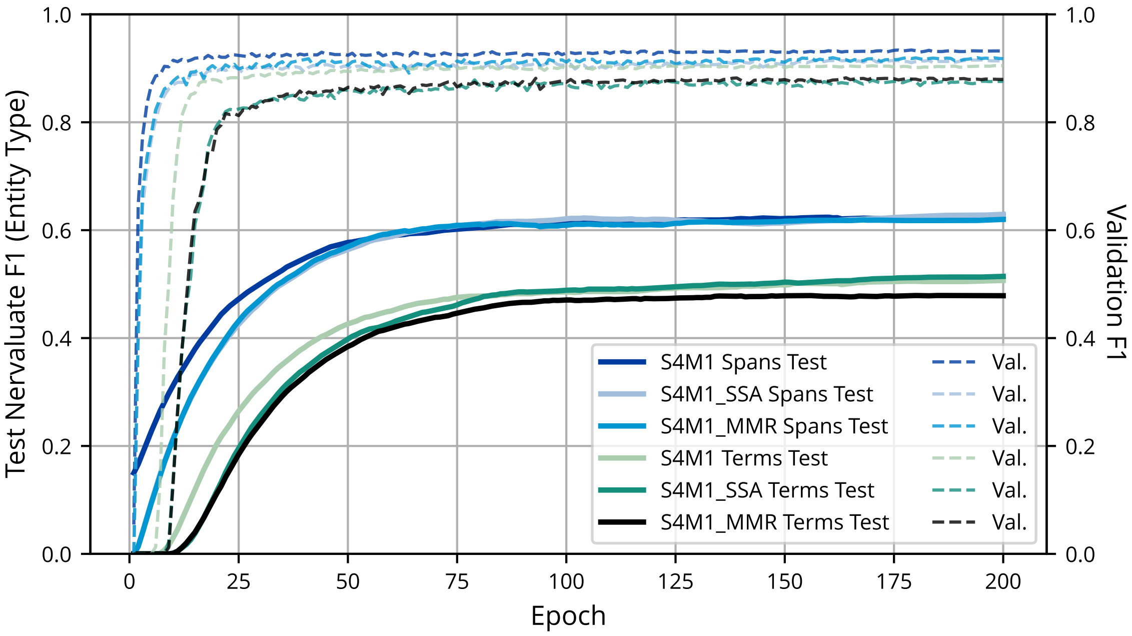


Figure 5.38: F1 Nervaluate evaluation of the previous S4M0_NT_EO_MMR_NI dataset with and without the extension of the S4M1 data points.

In summary, the S4M1 configuration extends the synthetic data generation process from single-term to multi-term sentences and introduces a validation scheme that ensures consistency between text and annotated terms. The resulting synthetic samples are more diverse and show better alignment with real data in the embedding space, as demonstrated by the PCA visualization in Figure 5.36. While span-level performance remains largely unchanged, term-level evaluation scores improve significantly (see Figure 5.37), indicating a clear benefit of multi-term generation for concept recognition. At the same time, the experiments with MMR subsampling (Figure 5.38) show no consistent advantage over random subsampling in this setting. Overall, the results suggest that multi-term synthetic sentences are a promising direction for enhancing ontology-based NER models, whereas more sophisticated subsampling strategies may require further refinement or alternative objectives to provide additional gains.

S4M2 - Add sampling to the generative request

In the preceding experiments, sampling during the generative request construction was introduced and evaluated in Section 5.2.1, followed by the integration of multiple term generation in Section 5.2.3. Building on these foundations, the configuration S4M2 aims to combine both techniques into a single experimental setup. The overarching objective of this configuration is to investigate whether the advantages of stochastic sampling in the requests, together with the generation of multiple candidate terms per input, can lead to further improvements in the quality and diversity of the suggested terminology.

The conceptual design of S4M2 extends the previous configurations by enriching the generative request with multiple sampled variants of the input examples, while simultaneously requesting multiple term candidates from the model for each instance. In principle, such a design is expected to better exploit the capacity of the model to generalise across different phrasings and contextualisations of the same underlying concept. By exposing the model to a broader range of example formulations in the generative request and then asking for several alternative term suggestions, the configuration is intended to encourage a richer exploration of the solution space while still being guided by domain-specific examples.

However, the practical realisation of this approach revealed substantial limitations related to the available context length of the underlying language model. Incorporating several sets of input examples, each with additional sampling-based variations, resulted in an excessively long query. As the total query length approached or exceeded the effective context window, the model was no longer able to process all components of the input reliably. This led to partial truncation of the query or to the model implicitly ignoring sections of the input, which, in turn, produced outputs that were inconsistent, incomplete, or semantically unrelated to the intended task. The degradation in output quality indicated that the combined use of extensive sampling and multiple term generation in a single query exceeded the practical capacity of the model in the current setup.

Due to these technical constraints, the S4M2 configuration did not yield results that could be meaningfully interpreted or compared with the previously evaluated setups. The outputs generated under this configuration were largely unusable for systematic analysis, as they did not reliably reflect the intended experimental conditions. Consequently, S4M2 is not further considered in the quantitative or qualitative evaluation presented in this work. The limitations encountered in S4M2 highlight the importance of careful query design under strict context constraints and point towards the need for more compact querying strategies or models with larger context windows for future research.

S4M3 - Next attempt to add sampling to the generative query

This section introduces configuration S4M3, which advances the generative query design explored in Section 5.2.3 by refining the use of sampling within the output examples. The overarching goal of this configuration is to enhance the ability of the model to generalize from a limited number of input examples by exposing it to a broader and more diverse space of plausible output sequences, while maintaining a controlled and interpretable query structure. In doing so, S4M3 aims to strike a balance between diversity and reliability in the examples provided to the model.

In contrast to previous configurations that either supplied multiple input/output pairs or relied on a smaller, more constrained set of examples, S4M3 is built around a single, fixed set of input HPO terms. Instead of varying the inputs, the configuration focuses on introducing variability in the outputs. For each input set, multiple sampled variations of possible output combinations are generated and included in the query. These variations represent different, but still clinically and semantically plausible, ways in which the model might associate or structure the given HPO terms. By doing so, the model is encouraged to internalize a richer representation of how such terms can appear in context, even when the underlying input remains constant.

The quality and diversity of these sampled outputs are central to the effectiveness of S4M3. The sentences used in the sampler for the query were carefully generated using commercial large language models and subsequently subjected to manual review. This review process ensured that only high-quality, contextually valid, and diverse sentence variants were retained. The emphasis on manual curation is particularly important, as it mitigates the risk of propagating noisy, ambiguous, or clinically implausible examples into the query, which could otherwise degrade performance. Given the limited number of distinct input examples available, this strategy of enriching the output space is intended to maximize the informational density for the downstream model.

Figure 5.39 presents the evaluation results of configuration S4M3 in comparison to the earlier S4M1 dataset. The figure summarizes the performance across the relevant evaluation metrics and allows a direct comparison of how the introduction of sampled output variations affects the ability of the model to identify and label HPO-related entities. Overall, the results indicate a minor yet consistent improvement of S4M3 over the S4M1 configuration. This suggests that the increased variety in the output examples provides incremental benefits, even though the overall query structure and the underlying input information remain unchanged.

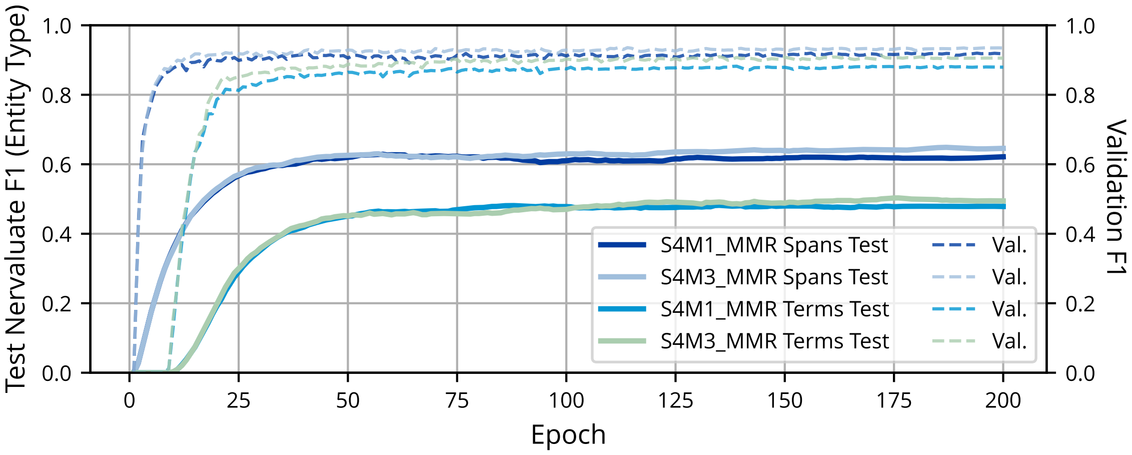


Figure 5.39: F1 evaluation comparing S4M3 and S4M1 datasets.

To obtain a more fine-grained view of these performance differences, a zoomed-in version of the evaluation is shown in Figure 5.40. In this closer view, the advantages of S4M3 in the terms-based evaluation become more apparent.

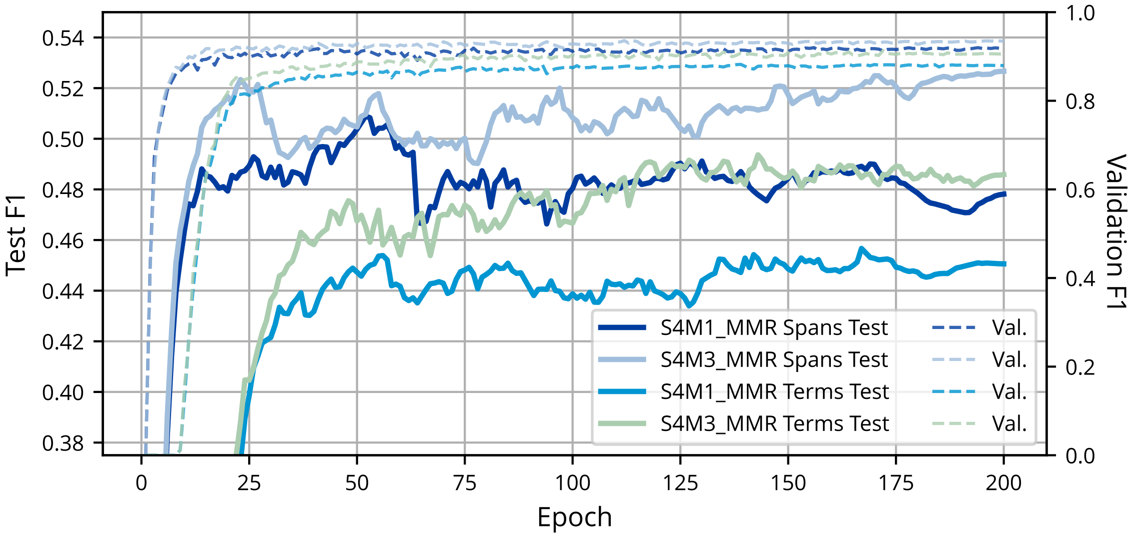


Figure 5.40: F1 evaluation comparing S4M3 and S4M1 datasets (zoomed in).

Taken together, the results for S4M3 demonstrate that enriching the query with multiple sampled output variations, derived from carefully curated LLM-generated sentences, can modestly improve model performance when input examples are scarce. While the observed gains over S4M1 are not dramatic, they are consistent and indicate that sampling-based diversification of the output space is a promising direction. This configuration therefore provides an intermediate step between simple, static generative query designs and more sophisticated sampling or augmentation strategies.

6 Refining NER Training

To obtain a more precise understanding of where the performance of the NER model can be improved, this chapter begins by taking a closer look at the characteristics of the gold standard data and their implications for model training. Particular attention is paid to structural aspects of the data, such as sentence length, the distribution of commas, and the composition of different document categories. These properties are examined both quantitatively, through descriptive statistics and evaluation tables, and qualitatively, using visualization techniques such as PCA. The goal of this analysis is to identify systematic patterns or irregularities in the data that may hinder effective learning or lead to unstable predictions.

Building on this data-centric view, explainable artificial intelligence (XAI) methods are employed to analyse and interpret the behaviour of the NER model. In particular, SHAP values are used to attribute model decisions to individual tokens and structural features within a sentence. By examining, for example, the influence of punctuation marks such as commas or the impact of very long sentences, it becomes possible to better understand which input characteristics the model relies on and where it may be overly sensitive or biased. This perspective supports the formulation of targeted refinements in preprocessing and training and provides a bridge between quantitative performance metrics and the underlying decision mechanisms of the model.

Afterwards, the chapter validates the best-performing NER model on different subsets of the gold standard data. These subsets reflect variations in sampling strategies and selection criteria that have been used throughout the training experiments. By comparing performance across these subsets, the analysis assesses the robustness and generalizability of the model and clarifies to what extent the observed improvements hold under changes in data distribution. This validation is essential to ensure that performance gains are not merely artifacts of a particular subset, but instead translate into more reliable behaviour across diverse clinical texts.

For completeness, the chapter then investigates whether further gains can be achieved by varying hyperparameters and by switching between different transformer-based language models as a backbone. Several German-language base models are compared with respect to their impact on NER performance under otherwise comparable training conditions. This systematic evaluation highlights the role of model capacity, pretraining domain, and architectural differences, and thus helps to position the chosen configuration within a broader landscape of possible approaches.

Finally, the chapter presents an initial approach to NEN, extending the scope from merely recognizing entity spans towards normalizing them to standardized identifiers in the HPO. This step bridges the gap between free-text mentions and structured phenotypic representations, enabling downstream applications such as cohort identification, phenotypic similarity search, and integration with other clinical and genomic resources. The preliminary experiments in this section explore different embedding-based methods and model architectures for mapping recognized entities to their corresponding ontology concepts and provide a first assessment of their suitability in the given domain.

6.1 Inspecting the gold standard data

To refine the training of the NER model in a targeted manner, the chapter first turns to a more detailed inspection of the GS data. The central assumption is that structural properties of the data, in particular sentence length and the distribution of commas, have a substantial impact on model behaviour and may explain systematic weaknesses observed in earlier experiments. Therefore, this section investigates which parts of the GS data are difficult for the model, how these difficulties manifest in the embedding space, and which preprocessing strategies can mitigate them without unduly altering the clinical content.

One key observation from the previous chapter was that aligning the GS data with synthetic training data in the PCA plots of the embedding representation is only partially successful. While large parts of the GS data can be matched reasonably well, a specific subset of sentences consistently remains distant from the synthetic data cluster in the PCA space. These sentences appear to constitute a structurally distinct group that the synthetic data generation process fails to replicate adequately. This subset becomes the starting point for a more systematic analysis of the GS data.

A first and intuitive hypothesis concerns the number of tokens per sentence. Long sentences are common in clinical documentation, where complex findings, diagnoses, and treatments are often compacted into a single narrative flow. Such sentences pose challenges not only for human readers, but also for transformer models with fixed context windows and position encodings. Consequently, the subsection below investigates how sentence length relates to the structure of the embedding space and to NER performance. Building on this, the following subsection then examines the role of commas as a proxy for complex sentence structure and clause chaining.

6.1.1 Number of tokens per sentence

The influence of sentence length on the embedding structure of the GS data is examined using a PCA-based visualization. Figure 6.1 shows the PCA plot, which divides the sentences into two groups: those with a number of tokens below the 90% quantile and those with a token count above this threshold. This split highlights the behaviour of the longest 10% of sentences in comparison to the rest of the data.

The visualization in Figure 6.1 reveals that sentences with an exceptionally high number of tokens tend to cluster separately in the embedding space. Instead of forming a smooth continuum with shorter sentences, they occupy a distinct region, which suggests that they share characteristics not present, or far less pronounced, in the majority of the data. Such characteristics may include nested clause structures or the bundling of multiple clinical events and relations into a single sentence. From a modelling perspective, these sentences are likely to contain patterns that the model sees only rarely and that are difficult to generalize from, particularly when the synthetic data used during training does not capture comparable complexity.

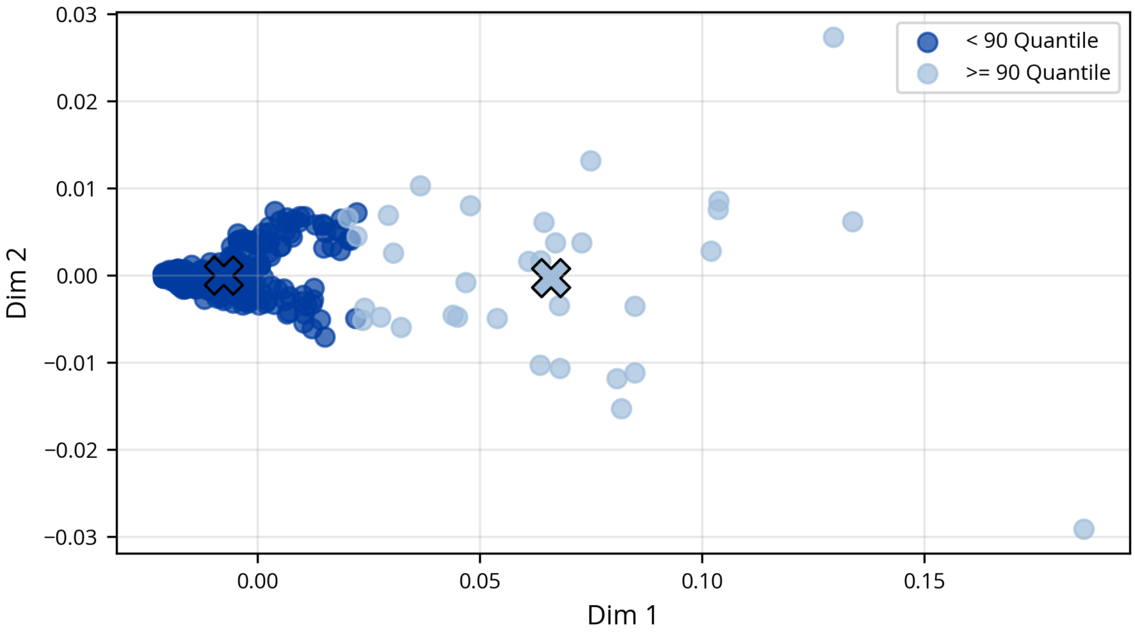


Figure 6.1: PCA visualization of the 10% longest sentences in the GS dataset.

To quantify the impact of these long sentences on NER performance, the evaluation on the full GS dataset using the most recent `S4M3_MMR` model is first considered as a reference. Table 6.1 reports the baseline scores when all sentences are included without any structural filtering. This reference serves as the benchmark against which subsequent manipulations of the dataset are assessed.

Metric	Value
F1 Score	0.494
Precision	0.451
Recall	0.547
Nervaluate (ET) F1 Score	0.602
Nervaluate (ET) Precision	0.516
Nervaluate (ET) Recall	0.721

Table 6.1: Reference GS evaluation results using the most recent `S4M3_MMR` model.

The next step is to omit sentences that exceed a predefined maximum length, thus removing a small fraction of extremely long sentences from the evaluation set. The value is chosen to correspond to the 90% quantile of sentence lengths in the GS. The corresponding results are shown in Table 6.2. This truncation leads to a strong increase in performance across all reported metrics. The magnitude of this improvement indicates that the removed sentences are indeed responsible for a disproportionate share of the errors made by the model. In other words, a relatively small subset of long sentences is capable of significantly depressing overall evaluation scores, even though it constitutes only a minor fraction of the data.

Metric	Value	Reference
F1 Score	0.582	0.494
Precision	0.465	0.451
Recall	0.780	0.547
Nervaluate (ET) F1 Score	0.675	0.602
Nervaluate (ET) Precision	0.558	0.516
Nervaluate (ET) Recall	0.854	0.721

Table 6.2: GS evaluation results for truncating GS to max length according to the 90% quantile of sentence lengths in the GS.

While this observation is useful to understand where the model struggles, simply discarding or truncating long sentences is not a viable strategy for practical NER systems, as it would selectively ignore precisely those complex clinical descriptions that often contain informative entities. Therefore, alternative approaches are explored that attempt to preserve the sentences while modifying their structure in a more controlled way. A first such attempt consists of splitting sentences at periods. The intuition is that very long sentences might actually encode multiple loosely connected propositions that could be separated into shorter, simpler sub-sentences without severely distorting the meaning.

The effect of this strategy on NER performance is shown in Table 6.3. Contrary to expectations, the results do not improve. Instead, they are slightly worse than the reference results in

Table 6.1. This suggests that splitting on periods introduces new boundary artifacts for the model, such as incomplete clauses or segments that lack sufficient context for reliable entity recognition. In clinical texts, periods may be used in abbreviations, measurement units, or structured list-like formats, so blindly splitting on this character can disrupt important local patterns.

Metric	Value	Reference
F1 Score	0.479	0.494
Precision	0.436	0.451
Recall	0.531	0.547
Nervaluate (ET) F1 Score	0.594	0.602
Nervaluate (ET) Precision	0.519	0.516
Nervaluate (ET) Recall	0.697	0.721

Table 6.3: GS evaluation results for splitting GS on period.

These findings motivate a more fine-grained view of sentence complexity. Instead of focusing solely on total length in tokens, the next subsection examines the internal punctuation structure of sentences. A closer inspection of the GS data reveals that many of the long and difficult sentences are characterized by a high density of commas, which serve to chain multiple clauses and enumerations. This observation leads to the question of how the presence and frequency of commas per sentence relate to the performance of the NER model.

6.1.2 How many commas per sentence

To investigate the role of commas, the same PCA-based visualization is repeated, but now the data is categorized by comma frequency rather than by simple sentence length. Specifically, Figure 6.2 distinguishes between sentences with more than five commas and those with five or fewer commas. This threshold targets sentences that exhibit particularly dense clause chaining and enumerated information, which are prevalent in clinical narratives such as medication lists, differential diagnoses, or combined findings and procedures.

The visualization in Figure 6.2 confirms that sentences with many commas behave similarly to the longest sentences in the previous analysis: they form a distinct region in the embedding space and appear as structural outliers. This alignment between length-based and comma-based outliers suggests that much of the difficulty associated with long sentences may stem from their internal punctuation structure rather than from length alone.

To quantify this insight, sentences containing more than five commas are truncated, and the resulting NER performance is reported in Table 6.4. These results are not as strong as those obtained by truncating based solely on maximum sentence length (Table 6.2), but they

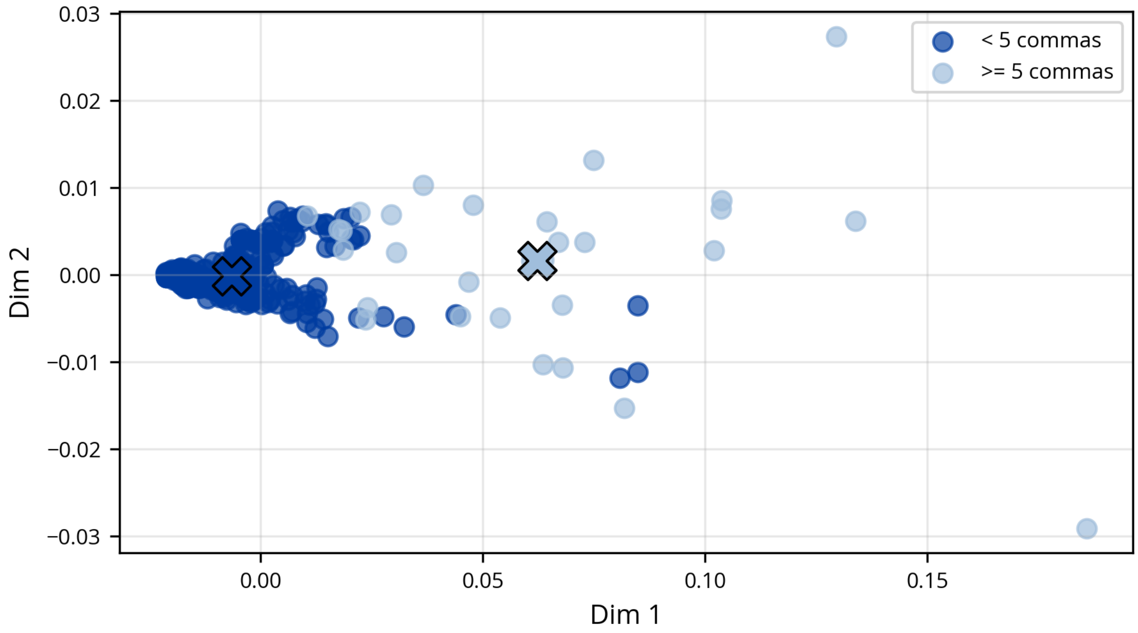


Figure 6.2: PCA visualization of GS sentences with more than 5 commas.

still exhibit a noticeable improvement compared to the reference evaluation in Table 6.1. This indicates that comma-rich sentences indeed contribute disproportionately to model errors. Moreover, a direct comparison of the affected data volume shows that 30 sentences are removed when truncating by comma count, compared to 33 sentences when truncating by the 90% length quantile. Thus, both strategies target a similarly small subset of complex sentences, but with slightly different selection criteria.

Metric	**Value**	**Reference**
F1 Score	0.530	0.494
Precision	0.450	0.451
Recall	0.644	0.547
Nervaluate (ET) F1 Score	0.634	0.602
Nervaluate (ET) Precision	0.536	0.516
Nervaluate (ET) Recall	0.776	0.721

Table 6.4: GS evaluation results for truncating GS at the 5th comma.

Again, truncation alone is not a satisfactory long-term solution, as it removes challenging yet clinically important text. Therefore, alternative manipulation strategies are explored that attempt to restructure comma-rich sentences while keeping their content accessible for the NER model. A straightforward idea is to split sentences at commas, thereby turning a single long sentence into multiple shorter units. However, blindly splitting at every comma would fragment a large portion of the corpus and severely distort the semantics of many sentences. Consequently, a more conservative approach is adopted: only sentences with four or more commas are split, and all commas in these selected sentences are treated as potential split points. The resulting performance is summarized in Table 6.5.

Metric	Value	Reference
F1 Score	0.416	0.494
Precision	0.303	0.451
Recall	0.664	0.547
Nervaluate (ET) F1 Score	0.498	0.602
Nervaluate (ET) Precision	0.360	0.516
Nervaluate (ET) Recall	0.811	0.721

Table 6.5: GS evaluation results for splitting the GS after the 4th comma.

The results in Table 6.5 show a clear performance degradation compared to the reference in Table 6.1. This indicates that the structural role of commas in clinical text is too important to be fully removed. By splitting sentences at every comma, the model is deprived of essential context, and entity spans may be cut into fragments that no longer align with the GS annotations. Thus, naive comma-based sentence splitting is not a viable solution.

To avoid this excessive fragmentation, a more targeted strategy is evaluated, where only the sixth comma in a sentence is replaced with a period. The underlying intuition is to keep the majority of the original comma structure intact, while introducing a single stronger boundary that divides extremely long clause chains into two more manageable segments. The results of this approach are reported in Table 6.6.

Metric	Value	Reference
F1 Score	0.485	0.494
Precision	0.455	0.451
Recall	0.519	0.547
Nervaluate (ET) F1 Score	0.604	0.602
Nervaluate (ET) Precision	0.530	0.516
Nervaluate (ET) Recall	0.701	0.721

Table 6.6: GS evaluation results for replacing commas with periods in sentences with more than 6 commas.

In contrast to the previous splitting experiments, this selective replacement yields results that are slightly better than the reference values in Table 6.1. It thus represents the first preprocessing strategy in this series that achieves any form of improvement by modifying the structure of long, comma-rich sentences. This suggests that long clause chains can be made more tractable for the model by inserting a small number of additional sentence boundaries, provided that the majority of the original comma-based structure is preserved.

Finally, an alternative approach is considered that aims to distribute the information content of comma-rich sentences over multiple sub-sentences while controlling the number of commas per segment. In this strategy, sentences are split at each *n*-th comma, resulting in an "*n*-comma gram" segmentation into shorter units with fewer commas. The evaluation of this method is shown in Table 6.7.

Metric	**Value**	**Reference**
F1 Score	0.498	0.494
Precision	0.451	0.451
Recall	0.556	0.547
Nervaluate (ET) F1 Score	0.601	0.602
Nervaluate (ET) Precision	0.518	0.516
Nervaluate (ET) Recall	0.716	0.721

Table 6.7: GS evaluation results for n-comma gram splitting.

The results in Table 6.7 are comparable to the reference evaluation and do not provide a meaningful improvement. This outcome suggests that the trade-off between reducing local complexity and preserving global sentence context is delicate. While *n*-comma gram splitting avoids the extreme fragmentation observed when splitting at all commas, it still introduces boundaries that may perturb entity spans and co-occurrence patterns in ways that are not beneficial for NER.

In summary, the analysis in this section shows that a small subset of structurally complex sentences, characterized by either extreme length, high comma density, or both, exerts a disproportionately negative influence on NER performance. Simple truncation of these sentences can substantially improve evaluation metrics, but at the cost of discarding precisely those instances that might be most informative in real-world clinical applications. More nuanced strategies, such as selective replacement of commas with periods, offer modest gains without severely altering the text, whereas more aggressive splitting schemes tend to harm performance. These findings highlight the importance of carefully designed preprocessing when dealing with highly structured clinical narratives and motivate the subsequent use of explainable AI methods in Section 6.4 and Section 6.5 to gain a deeper understanding of how the model reacts to such structural variations at the level of individual tokens and punctuation marks.

6.1.3 Split by categories

Finally, one more structural aspect of the GS data is considered: its distribution across different categories based on the affiliation of the sentence within the originating EHR. Clinical documentation is heterogeneous in nature, covering a wide range of communicative functions such as reporting physical findings, summarizing the history of a patient, or formulating diagnostic and therapeutic plans. Understanding how sentences are constructed in these different contexts can help to identify category-specific patterns or challenges that may systematically affect NER performance.

From a linguistic perspective, each category tends to follow its own conventionalized style, characterized by typical sentence templates and characteristic use of domain-specific terminology. For instance, some categories may rely on compact enumerations of findings or procedures, while others make more frequent use of narrative descriptions and temporal relations. For an NER model, such stylistic differences translate into varying distributions of entity types, token contexts, and punctuation patterns, all of which can influence how reliably entities are recognized. Therefore, inspecting the GS data through the lens of document category provides a complementary viewpoint to the sentence-level analyses of length and comma density discussed in the previous subsections.

Surprisingly, when visualizing the PCA plot by categories in Figure 6.3, distinct clusters emerge that correspond to different types of clinical narratives. In this visualization, sentences from the *Körperlicher Untersuchungsbefund* category tend to occupy similar regions in the embedding space, while sentences from different categories form into one clearly separable group. This indicates that the underlying sentence representations capture not only lexical-semantic content, but also stylistic and structural regularities that are characteristic for each category. The emergence of such clusters underscores the importance of considering sentence structure and document type during the preprocessing phase, as category-specific peculiarities may otherwise be obscured in aggregate analyses over the full dataset.

To assess how these structural differences translate into NER performance, the evaluation is next stratified by category. Instead of computing a single, global score over all GS sentences, the metrics are calculated separately for each document type. The resulting scores, summarized in Table 6.8, reveal substantial variation across categories. While the model achieves relatively strong performance in some types of clinical texts, it performs noticeably worse in the *Körperlicher Untersuchungsbefund* category, despite having been trained on the same underlying architecture. This category is typically composed of detailed descriptions of physical examination findings, often expressed in compact yet information-dense sentences. The pronounced drop in performance on this subset suggests that generic training and preprocessing strategies are insufficient to capture the particular way in which entities are embedded in these descriptions.

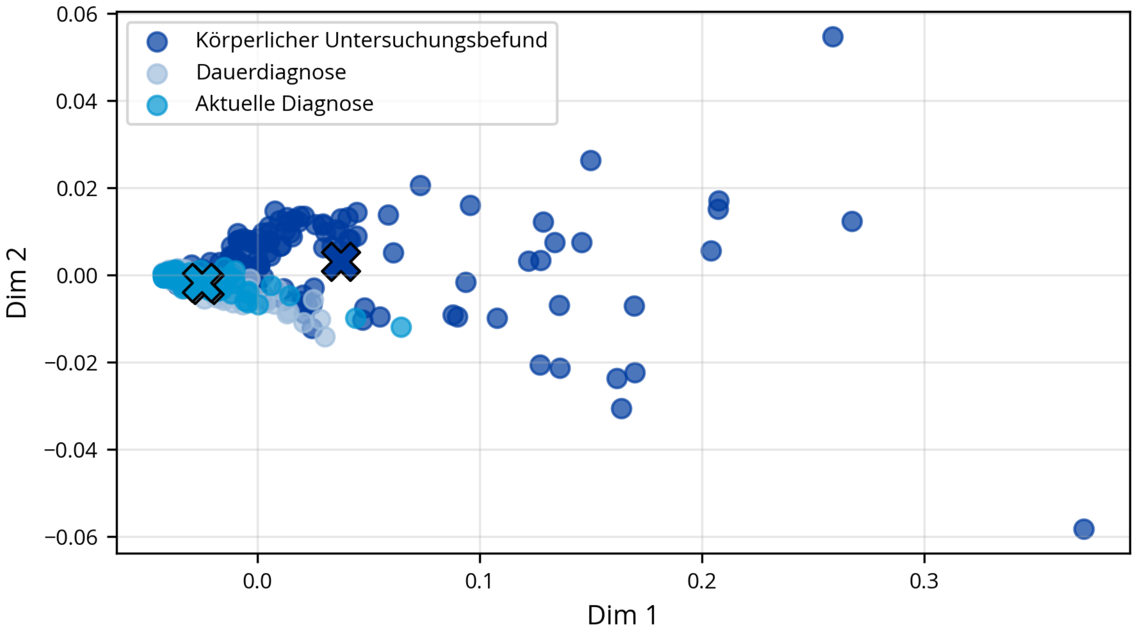


Figure 6.3: PCA visualization of sentences grouped by categories in the EHR.

This discrepancy in performance motivates a closer look at the data statistics for each category. Table 6.9 reports key characteristics, including the average number of tokens per sentence and several entity-related measures. It becomes evident that the category *Körperlicher Untersuchungsbefund* has the highest average number of tokens and therefore contains the longest sentences on average. This aligns with the broader finding from the earlier subsections that very long sentences tend to form structural outliers in the embedding space and are associated with lower NER performance.

In addition to sentence length, Table 6.9 shows that the average number of words until an entity is the highest in the category *Körperlicher Untersuchungsbefund*. This metric has been labeled as *Average Entities per Words*$^{-1}$. In other words, entities appear more sparsely in

	Aktuelle Diagnose	Dauerdiagnose	Körperlicher Untersuchungsbefund
F1 Score	0.60	0.60	0.28
Precision	0.48	0.48	0.35
Recall	0.82	0.80	0.23
Nervaluate (ET) F1 Score	0.67	0.74	0.38
Nervaluate (ET) Precision	0.55	0.65	0.32
Nervaluate (ET) Recall	0.84	0.86	0.45

Table 6.8: GS evaluation results after splitting by categories.

these texts, separated by long stretches of descriptive content. For the model, this sparsity implies that informative entity cues are embedded in extended sequences of non-entity tokens, making it harder to maintain focus on relevant spans and to accurately identify their boundaries. The combination of long sentences, dense descriptive content, and relatively infrequent entities thus creates a particularly challenging setting for NER.

	Aktuelle Diagnose	Dauerdiagnose	Körperlicher Untersuchungsbefund
Total Sentences	69	124	125
Sentences Relative	0.22	0.39	0.39
Total Entities	46	91	66
Relative Entities	0.23	0.45	0.33
Average Words per Sentence	4.7	4.0	18.6
Average Entities per Words^{-1}	7.1	5.4	35.2

Table 6.9: Gold standard data statistics by categories.

Taken together, the category-wise analysis corroborates the earlier observations on sentence length and comma density: structural complexity plays a central role in shaping model performance. Categories that favour shorter, more regularly structured sentences with frequent entities are handled comparatively well, whereas categories characterized by long, information-dense sentences and sparse entity distributions, such as *Körperlicher Untersuchungsbefund*, exhibit marked performance deficits. This suggests that the difficulties of the model are not purely semantic, but are closely tied to how clinical information is organized and encoded in different documentation contexts.

This analysis therefore indicates that the structural complexity of sentences in the category *Körperlicher Untersuchungsbefund* is a significant factor contributing to the poor performance on this subset. Future work could focus on developing category-specific preprocessing strategies, for example by introducing carefully designed segmentation rules that preserve clinically relevant context while reducing local complexity. In addition, targeted data augmentation for underperforming categories, such as generating synthetic examples that mimic the characteristic sentence structure and entity sparsity of physical examination reports, may help the model to better internalize the patterns encountered in these texts. More generally, the findings highlight the potential benefits of explicitly modelling document type and category information, either as an additional input signal during training or as a basis for specialized, category-aware NER components.

6.2 XAI: SHAP for NER

XAI methods provide a systematic way to analyse how complex models, such as transformer-based architectures for NER, arrive at their predictions. Among these methods, SHapley Additive exPlanations (SHAP) has gained prominence as a model-agnostic framework that assigns additive feature attributions to each input component, thereby quantifying its contribution to the model output (Zeng, 2024). In the context of clinical NER, this makes it possible to move beyond aggregate performance metrics and inspect which tokens and structural cues drive correct or incorrect entity predictions.

By computing token-level contribution scores, SHAP can reveal whether the model relies on linguistically and clinically plausible signals, such as medically meaningful terms and their modifiers, or whether it is overly influenced by superficial patterns like specific punctuation marks or position within the sentence. This aligns well with prior work that uses SHAP to interpret BERT-like models and to diagnose their strengths and weaknesses in sequence labelling tasks (Stockem Novo & Gedikli, 2023). In the present setting, SHAP is therefore applied with a dual objective: to confirm the structural phenomena identified in Section 6 and to provide a token-level explanation of how commas and sentence length affect model behaviour.

Figure 6.4 illustrates SHAP-based explanations for a representative comma token in a clinical sentence. In this visualization, positive attribution values indicate that a token pushes the model towards predicting an entity label, whereas negative scores suggest that the token suppresses entity predictions or supports the non-entity class. The comma token in question receives a clearly negative contribution, signalling that its presence reduces the confidence of the model in assigning an entity label to nearby tokens. This effect is consistent across multiple inspected examples and suggests that commas are systematically interpreted as boundaries or segmentation cues that delimit or interrupt potential entity spans.

These findings complement the quantitative analyses in Section 6, where comma-rich sentences were shown to be structurally distinct in the embedding space and to contribute disproportionately to model errors. While the earlier results demonstrated that sentences with many commas tend to form outlier clusters and depress evaluation scores, the SHAP analysis in Figure 6.4 clarifies how this effect arises at the level of individual tokens. Commas do not merely co-occur with difficult contexts; they actively steer the model away from entity predictions, which can be beneficial when they truly mark clause boundaries, but harmful when entities span across comma-separated segments or when important modifiers are attached after a comma. The strong negative attributions thus provide an explanation for why naive comma-based splitting strategies, such as those evaluated in Table 6.5, tend to degrade performance: they amplify a bias that is already present in the internal decision process of the model.

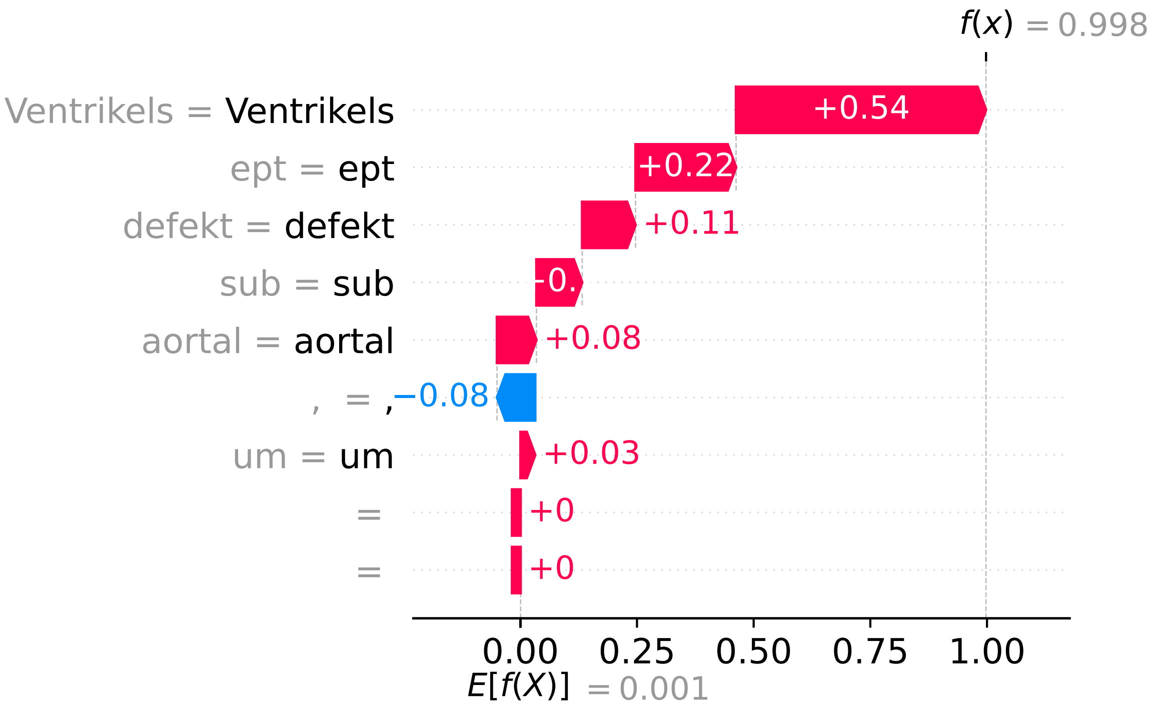


Figure 6.4: SHAP explanation for comma token.

A complementary perspective is obtained by applying SHAP to entire sentences with varying lengths and structural complexity. In Figure 6.5, the distribution of token-level SHAP values is visualized for long clinical sentences that are characteristic of the problematic regions identified in the PCA plots in Section 6. The SHAP attributions reveal that, in such contexts, only a small subset of tokens receives strongly positive contributions to entity predictions, whereas large stretches of the sentence exhibit near-zero or even slightly negative impact on the target entity class.

This pattern indicates that, for very long and structurally complex sentences, the model distributes its attention over many tokens that are only weakly informative for the entity classification task. As a result, the effective signal-to-noise ratio for entity-relevant cues decreases, and the model becomes less confident and more error-prone in identifying correct spans. In line with the empirical improvements observed when truncating or selectively segmenting long sentences (e.g., Tables 6.2 and 6.6), the SHAP analysis in Figure 6.5 suggests that reducing local complexity and limiting the amount of non-informative descriptive content within a single input window can help the model to focus its representational capacity on truly relevant tokens.

Taken together, the SHAP-based investigation corroborates and deepens the findings from the structural analyses of the gold standard data. Commas emerge as influential tokens with predominantly negative contributions to entity predictions, explaining why sentences

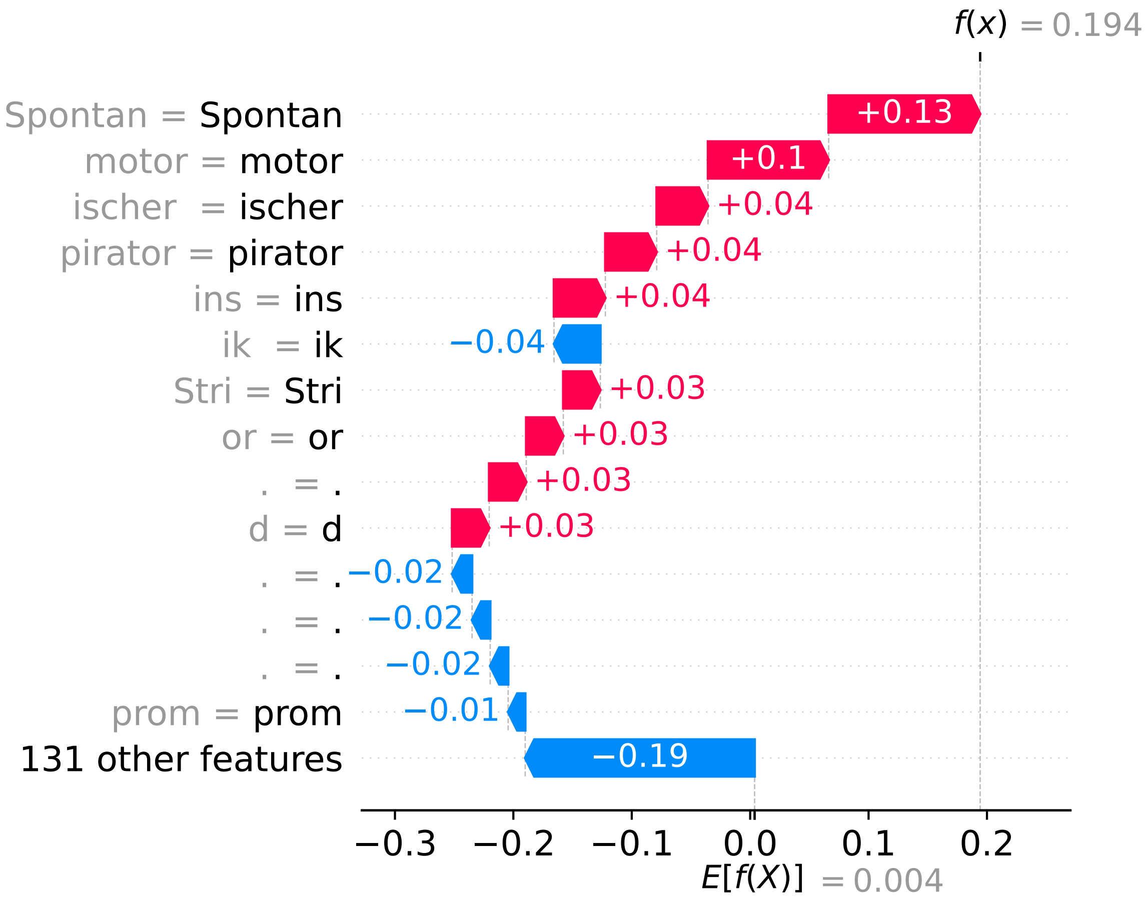


Figure 6.5: SHAP explanation for long sentences.

with dense comma usage disproportionately degrade performance. At the same time, long, information-dense sentences are shown to dilute the impact of entity-bearing tokens, making it more difficult for the model to learn stable decision boundaries. By exposing these mechanisms at the token level, SHAP provides actionable guidance for preprocessing and model refinement: segmentation strategies should be designed to introduce a small number of robust sentence boundaries, as in Table 6.6, while avoiding aggressive splitting schemes that overemphasize the already negative role of commas. More broadly, the SHAP analysis demonstrates how XAI can serve as a bridge between descriptive corpus statistics and concrete engineering decisions, informing the subsequent steps in refining, validating, and extending the NER system in the remainder of this chapter.

6.3 Validate best model using different subsets

After introducing the subset of 300 terms used across these experiments, as described in Section 4.2, the section now turns to validating the best-performing NER model on different slices of the HPO ontology. The central objective of this section is to examine how well the model generalizes when the evaluation focus shifts from a relatively small, carefully selected subset of ontology terms to a broader and more diverse set of phenotypic concepts. The size of the evaluation subset is increased from 300 to 500 terms in order to assess whether the performance trends observed earlier remain stable under more demanding conditions and whether the model continues to handle the expanded ontological coverage in a robust manner.

To this end, the best-performing configuration identified in the previous chapters is evaluated on two different HPO-based subsets: the original set of 300 terms and the extended set of 500 terms. Figure 6.6 compares the corresponding F1 scores, separately considering span-level and term-level evaluation metrics. For the span evaluation, the performance on the 500-term subset remains broadly comparable to the results obtained with 300 terms. This indicates that the model continues to identify entity boundaries reliably, even when confronted with a larger variety of phenotypic expressions. In other words, the structural aspects of entity recognition, such as detecting the correct start and end positions of mentions, are not substantially affected by the increased ontological coverage.

In contrast, the term-level evaluation shows a slight drop in performance when moving from 300 to 500 terms. This decline suggests that, while the model still locates entity spans with similar accuracy, it encounters more difficulty in assigning the correct phenotype labels when the number of possible target terms grows. The expanded subset necessarily introduces additional, and in many cases semantically related, concepts, thereby increasing the risk of confusion between closely neighbouring HPO entries. As a result, the term-level metrics in Figure 6.6 reflect the heightened classification complexity associated with a denser and more fine-grained ontology slice.

To investigate whether this term-level performance drop can be mitigated by providing the model with more training evidence for each concept, the evaluation is repeated with an increased number of samples per term in the training data. In this setting, the training corpus is enriched with additional instances covering the same 500-term subset, thereby aiming to improve the ability of the model to discriminate between closely related phenotypes and to internalize more robust lexical and contextual patterns for each term. The resulting performance is shown in Figure 6.7, again focusing on term-level F1 scores.

The results in Figure 6.7 indicate that increasing the number of samples per term largely compensates for the performance loss observed when expanding the ontology slice. With more training instances per concept, the model can better distinguish similar terms and

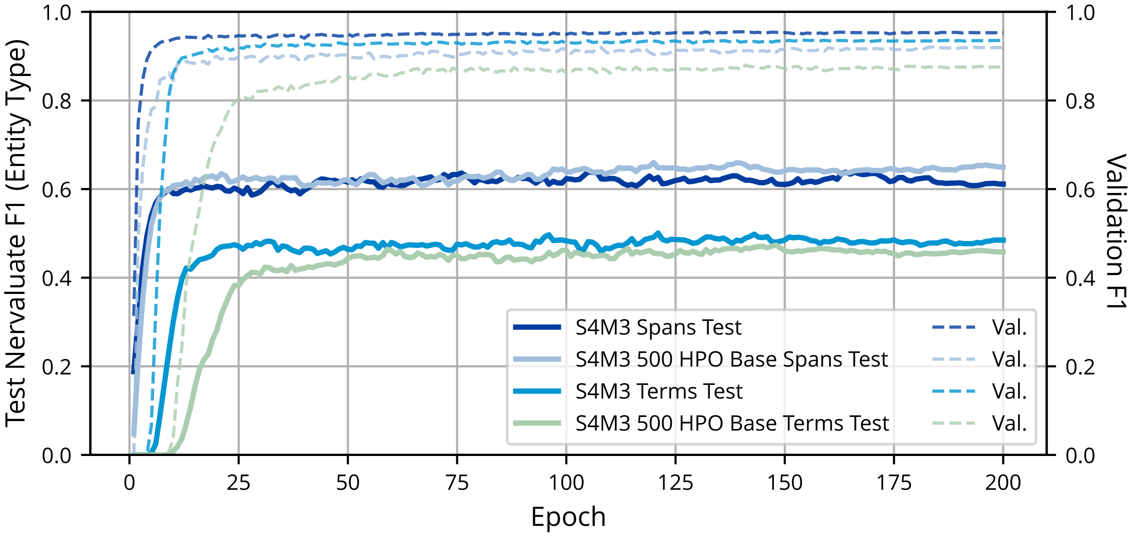


Figure 6.6: F1 Nervaluate evaluation after reshuffling the HPO terms (`S4M3 500 HPO Base`).

generalize more reliably across diverse clinical contexts. However, this gain in accuracy comes at a substantial computational cost. The accompanying training statistics highlight a marked increase in the number of training samples and in the average time per epoch, leading to longer total training times. Any future refinement of the system must therefore balance the benefits of adding more samples per term against the associated resource requirements, particularly when considering even larger subsets of the HPO or additional phenotype ontologies.

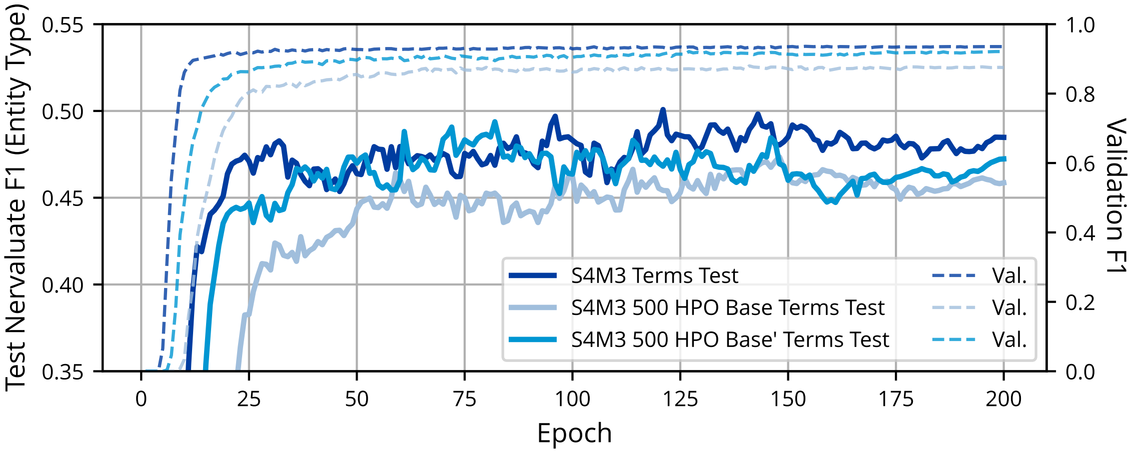


Figure 6.7: F1 Nervaluate evaluation after reshuffling the HPO terms (`S4M3 500 HPO Base`) and increasing the number of samples per term (`S4M3 500 HPO Base'`).

In summary, this validation on different HPO-based subsets demonstrates that the best-performing NER model maintains stable span-level performance when moving from 300 to 500 evaluation terms, but initially suffers a modest decline in term-level accuracy due to the increased label space and semantic granularity. This decline can be mitigated by augmenting the training data with more samples per term, at the cost of longer training times. These findings underscore the importance of carefully calibrating the trade-off between ontological coverage, data volume, and computational efficiency when scaling NER systems to broader subsets of clinical ontologies. They also provide a concrete foundation for the subsequent sections, which further explore how hyperparameter choices and different transformer backbones can influence the ability to cope with growing conceptual complexity in real-world clinical applications.

6.4 Hyperparameters & Different Base Models

This section investigates whether the NER system can be further improved by varying hyperparameters and exchanging the transformer backbone. The aim is to assess if the current base model and training setup form a local optimum or if alternative configurations yield better precision and recall on the clinical gold standard. For this purpose, several German transformer models are compared under similar training conditions, followed by experiments with key optimization parameters such as learning rate, batch size, and data splits.

On the model side, three representative German transformer backbones are considered: the cased German BERT-Base model (*BB*) (google-bert, 2019), the *GottBERT* model (*GOT*) (Scheible et al., 2024; „TUM/GottBERT_base_last · Hugging Face", n.d.), and the *GBERT*-Large architecture (*GB*) (Chan et al., 2020; „Deepset/Gbert-Large · Hugging Face", 2024). These models differ in terms of model capacity, pretraining corpora, and linguistic coverage, and thus provide a meaningful testbed for examining how sensitive the clinical NER task is to the choice of backbone.

The corresponding evaluation results are summarized in Figure 6.8, which reports the F1 scores for the different base models. Overall, no backbone clearly surpasses the previously selected configuration; most models perform within a narrow band, suggesting that, under the current data and task setup, backbone choice is not the main bottleneck for NER quality. A clear exception is the *GOT* model, which consistently underperforms, indicating that its pretraining or representational biases are less well aligned with the clinical texts in this work.

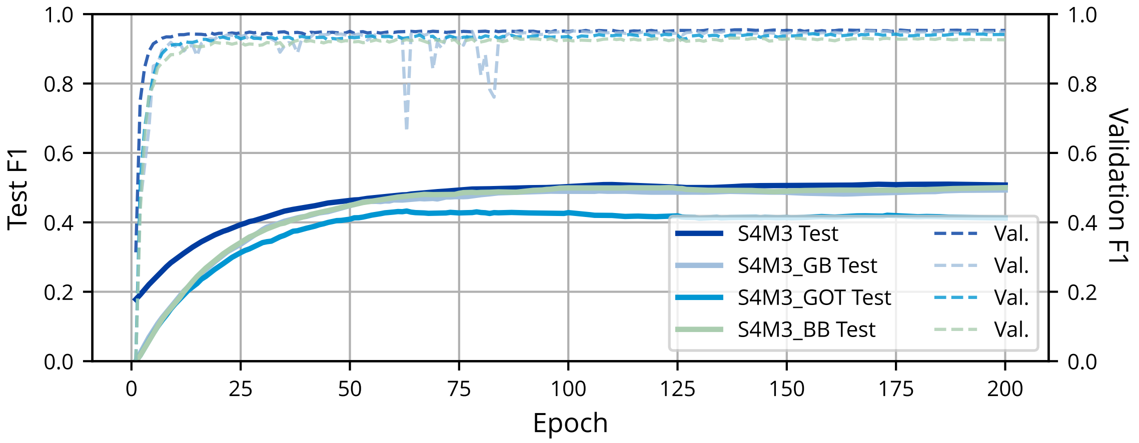


Figure 6.8: F1 spans evaluation of different NER base models.

Complementing the variation of base models, a series of experiments evaluates the impact of key training hyperparameters on NER performance. Different learning rates, batch sizes, and train–validation–test splits are systematically varied to probe the robustness of the best configuration and to identify promising settings. The goal is to avoid both underperforming setups that waste available information and overly aggressive tuning that produces unstable, non-generalizable results.

Figure 6.9 summarizes the F1 scores for the different hyperparameter combinations. Although some settings lead to small performance fluctuations, no configuration is clearly superior. Instead, the results lie within a narrow band, indicating that the NER model is robust to moderate changes in learning rate, batch size, and data splits and that the observed behaviour is driven mainly by task and data properties rather than by fine-tuned hyperparameters.

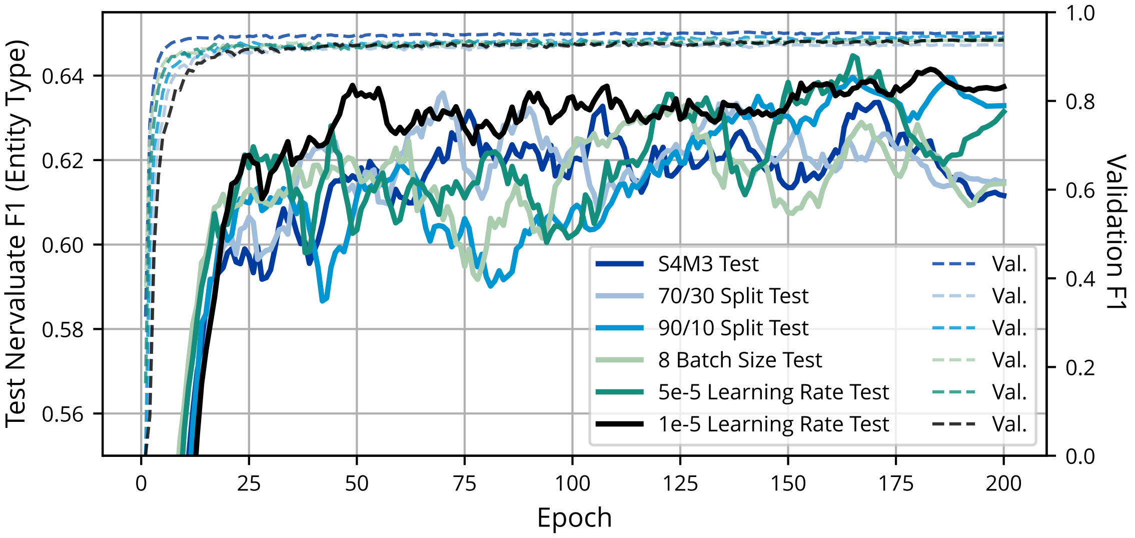


Figure 6.9: F1 spans evaluation of different training hyperparameters.

In summary, the experiments in this section show that neither switching the transformer backbone nor moderately tuning hyperparameters markedly improves NER performance on this clinical dataset. Apart from the clearly underperforming *GOT* backbone, all configurations behave similarly, and reasonable variations in learning rate, batch size, and data splits only yield small gains. This supports the conclusion that major improvements depend less on architecture or optimization and more on data-centric factors such as structural complexity, domain coverage, and ontological granularity.

6.5 NEN: Named Entity Normalization

In this section, the normalization of the entities recognized in the span evaluation of the NER model is addressed. The objective is to map the extracted entity mentions to their corresponding concepts in the HPO ontology and thereby bridge the gap between surface-level textual mentions and structured phenotypic representations. This normalization step is essential for downstream applications such as cohort identification, phenotype-based patient stratification, and interoperability with other clinical and genomic resources, all of which rely on stable, ontology-based identifiers rather than free-text strings.

To construct the embedding representations required for this mapping, several sentence and text embedding models are evaluated. In addition to the previously used embedding models derived from the NER base models, two newer architectures are considered for encoding both entity mentions and HPO terms:

- *multilingual-e5-large* („Intfloat/E5-Large · Hugging Face“, 2025)
- *ModernBERT-base* („Answerdotai/ModernBERT-base · Hugging Face“, 2024)

These models have been proposed as state-of-the-art embedding backbones in general-domain and multilingual scenarios and are therefore promising candidates for capturing nuanced semantic relations between clinical phrases and ontology entries. Using them alongside the earlier, NER-derived embeddings allows a direct comparison between task-specific and more general-purpose representation learning approaches in the context of NEN.

A key design choice in constructing embeddings for normalization concerns the amount of local context provided around each entity mention. To systematically analyse this effect, the entity spans are complemented with *n*-word context windows: for each mention, the *n* tokens preceding and following the span are concatenated with the entity text before it is fed into the embedding model. This strategy aims to enrich the representations with additional syntactic and semantic cues for correctly disambiguating between closely related HPO concepts. At the same time, it remains conservative enough to avoid overly long input sequences that might dilute the entity-specific signal.

For all evaluations, the top *n* column shows if the correct HPO term appears within the top *n* nearest neighbours in the embedding space. This metric reflects the ability to retrieve relevant ontology terms that could be used as a recommendation or candidate set for further disambiguation.

Figure 6.10 compares the different models for creating the embedding representation under a fixed context configuration. In all cases, two words of context are added before and after the entity mention, providing a short but informative local window. The results show that both the *medbert* and *multilingual-e5-large* models clearly outperform the previously used embeddings across the evaluated metrics. This suggests that specialized medical pretraining, as in the case of *medbert*, and strong multilingual semantic modelling capabilities, as provided by *multilingual-e5-large*, yield more discriminative representations for aligning clinical entity mentions with their corresponding HPO concepts.

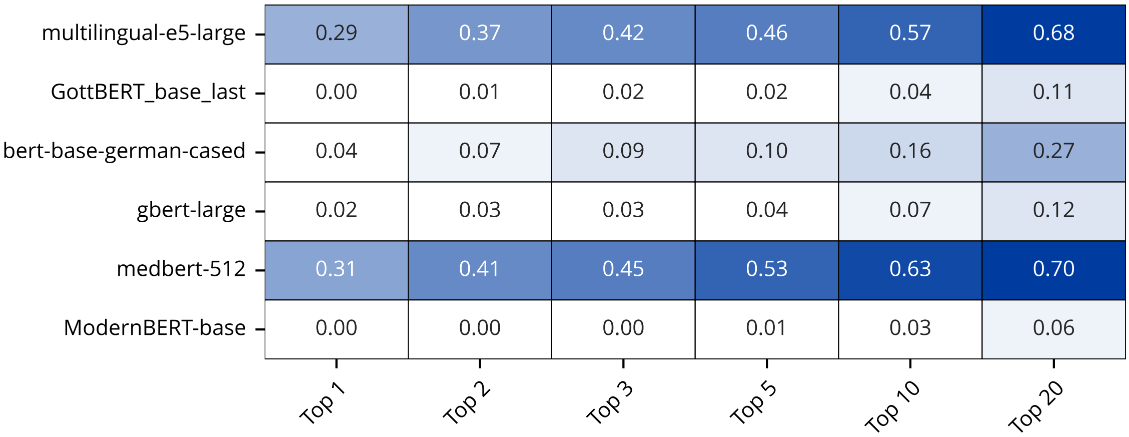
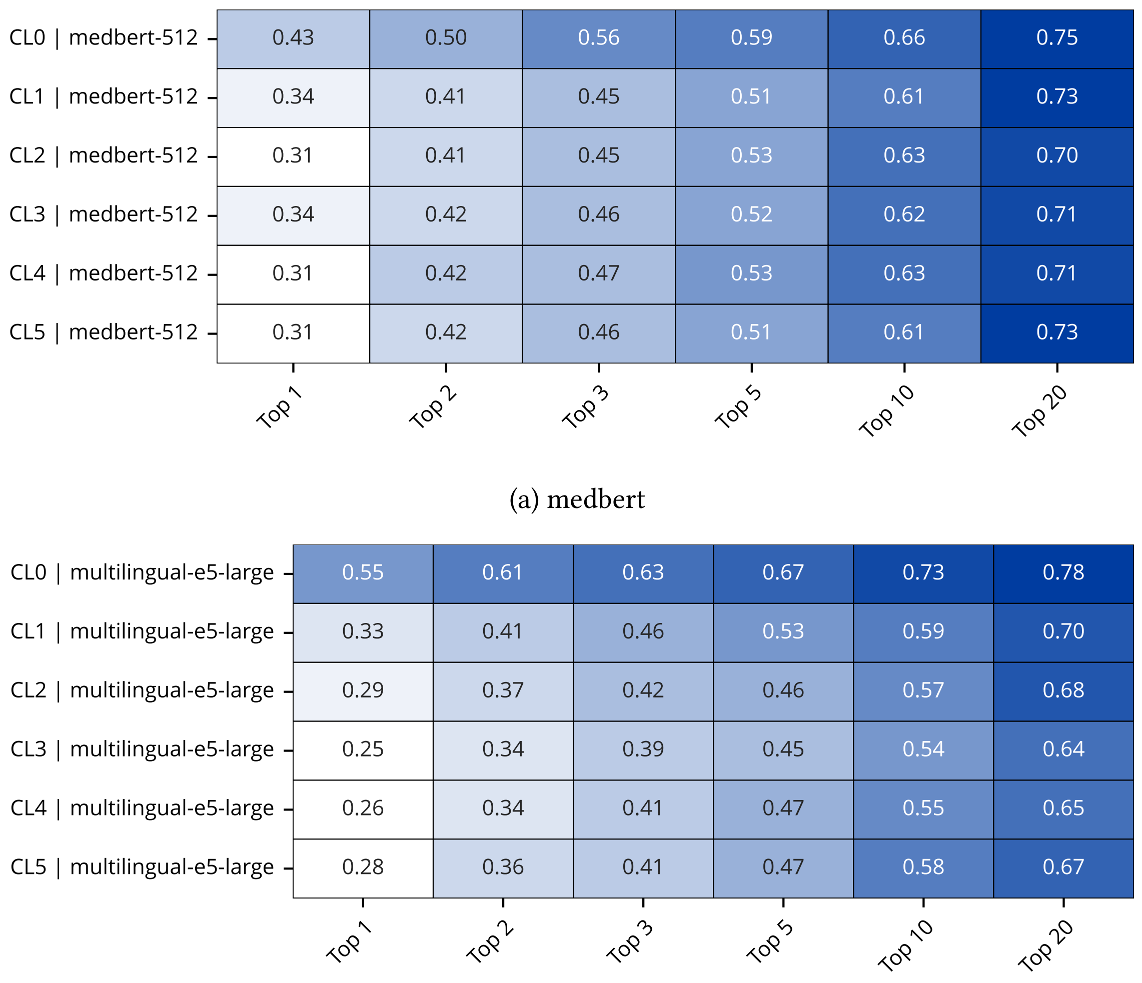


Figure 6.10: Comparing different models for creating the embedding representation.

Building on this observation, the next step is to investigate how the length of the context window interacts with the choice of embedding model. Intuitively, increasing the number of surrounding tokens could either improve normalization, by providing richer disambiguating cues, or harm it, by introducing additional noise and unrelated information. To assess this trade-off, different context sizes are evaluated for the two best-performing models identified in Figure 6.10, namely *medbert* and *multilingual-e5-large*. For each model, the score is computed across several *n*-word window configurations.

Figure 6.11 summarizes these results. The upper subfigure 6.11a reports the behaviour of *medbert* for increasing context lengths (CLs), while the lower subfigure 6.11b shows the corresponding values for *multilingual-e5-large*. The comparison highlights that the two models respond slightly differently to additional context. For some CL sizes, performance improves as more surrounding tokens are included. Beyond a certain point, however, the benefit plateaus suggesting that excessively long context windows may blur the focus on the central entity mention and make the representations less precise for normalization. Overall, the analysis confirms that adding context is not beneficial in this specific task.

Beyond model architecture and context size, the difficulty of the normalization task also depends strongly on the size and granularity of the candidate concept set. To explore this dimension, the final experiment in this section increases the number of possible HPO terms from 300 to 500, mirroring the expansion from a smaller, carefully curated ontology slice to a broader and more diverse phenotypic spectrum. The goal is to assess how well the



(a) medbert

(b) multilingual-e5-large

Figure 6.11: Comparing the score of different context lengths (CLs) for the corresponding embedding model.

current embedding-based approach scales when the search space of potential target concepts grows, as would be the case in realistic clinical deployments or future extensions to additional ontology branches.

Figure 6.12 presents the corresponding results. The comparison shows that performance declines when the number of candidate HPO terms is increased from 300 to 500, analogous to the term-level NER evaluation in Section 6. This reflects the intrinsic difficulty of discriminating among a larger set of, in many cases, semantically similar concepts, where subtle lexical and contextual differences become decisive. At the same time, the results demonstrate that the embedding-based normalization framework continues to function in principle, providing a basis for further refinements in candidate filtering, re-ranking, or ontology-aware postprocessing.

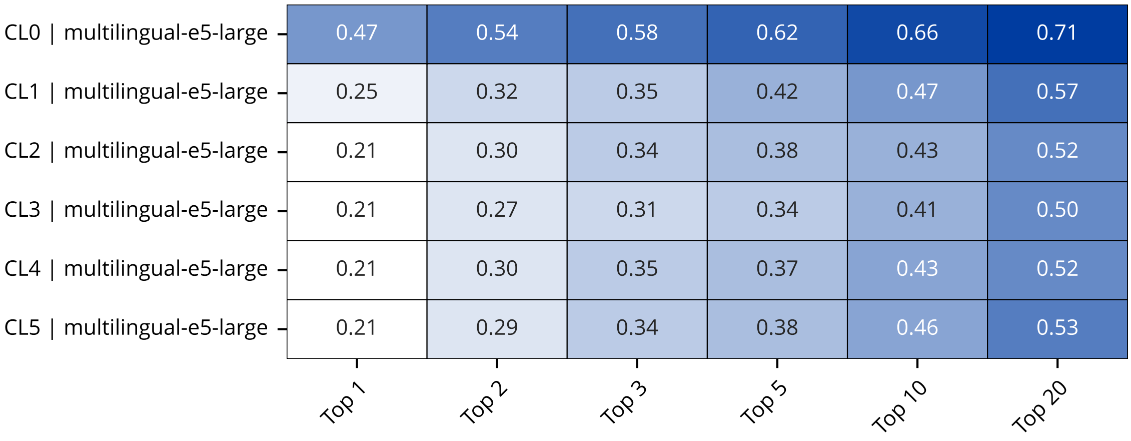


Figure 6.12: Comparing results when increasing the possible HPO terms from 300 to 500.

Overall, the experiments in this section highlight both the promise and the challenges of embedding-based NEN in the clinical phenotype domain. Modern embedding models such as *multilingual-e5-large* substantially improve the alignment between entity mentions and HPO concepts. At the same time, the results in Figure 6.12 underscore that normalization becomes increasingly difficult as the number of candidate terms grows and the ontology slice becomes denser. These findings suggest that future work should combine strong embedding backbones with more sophisticated candidate generation and disambiguation strategies, potentially incorporating ontology structure, hierarchical relations, and category-specific knowledge. In this way, the NEN component can be further strengthened to support robust, ontology-based downstream analyses on top of the NER system developed in the preceding sections.

7 Discussion

This chapter discusses the main findings of the thesis, their implications, and their place within NER research in specialized biomedical domains. It is structured into four parts: a summary of key experimental results, a discussion of known limitations, overarching conclusions on the research questions and proposed directions for future work.

7.1 Results Summary

The results of this thesis show that large language models can be used to build effective NER systems for extracting domain-specific concepts, such as HPO terms, from clinical or biomedical text. Systematic comparison of model configurations and data generation strategies indicates that carefully designed synthetic data can clearly improve performance over simple baselines.

Figure 7.1 compares the Nervaluate F1 scores of the main experimental setups from Chapter 6. A key contribution are the merits of the dataset shape alignment in Section 5.2.2 to analyse the gold standard distribution. Aligning the statistical properties of synthetic and target data reduces distributional mismatch and improves generalization, leading to clear improvements over earlier, less informed generation strategies.

However, the results on individual term recognition also underline the difficulty of the task. With finer-grained evaluation, extracting precise terms becomes harder and the margin over the strong baseline narrows. This suggests that further progress will require not only better synthetic data, but also more task-aware models and deeper use of domain structure.

Figure 7.2 shows the same comparison for the individual terms evaluations. Here, the improvements are even more pronounced, but are also due to the higher availability of data using open-weight models compared to costly commercial models.

But with the increased difficulty of the tasks in determining individual terms, the performance differences with the baseline show that it's hardly possible to surpass the baseline performance.

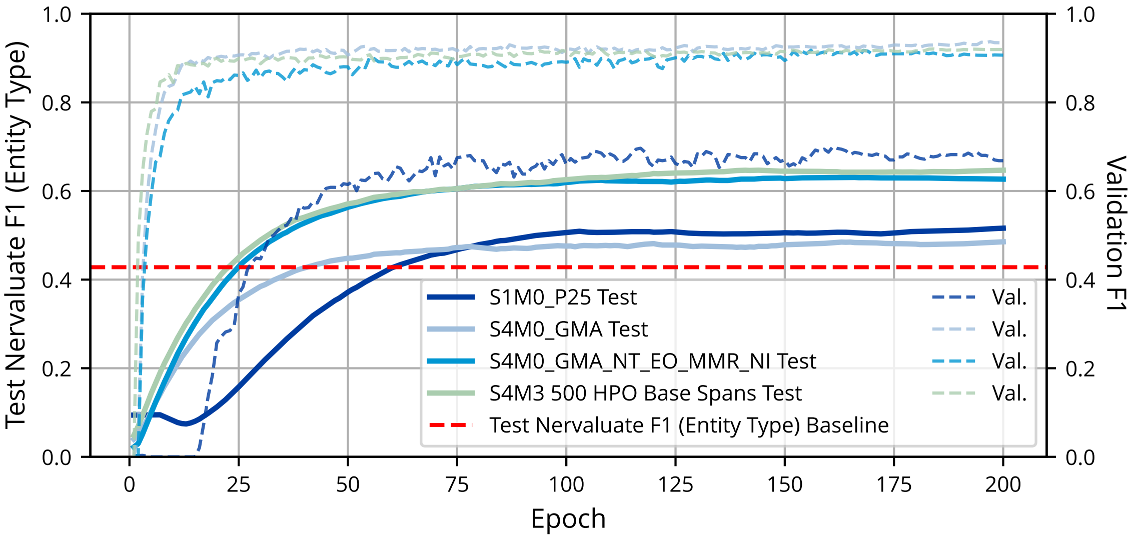


Figure 7.1: Final comparison of F1 spans Nervaluate evaluation scores.

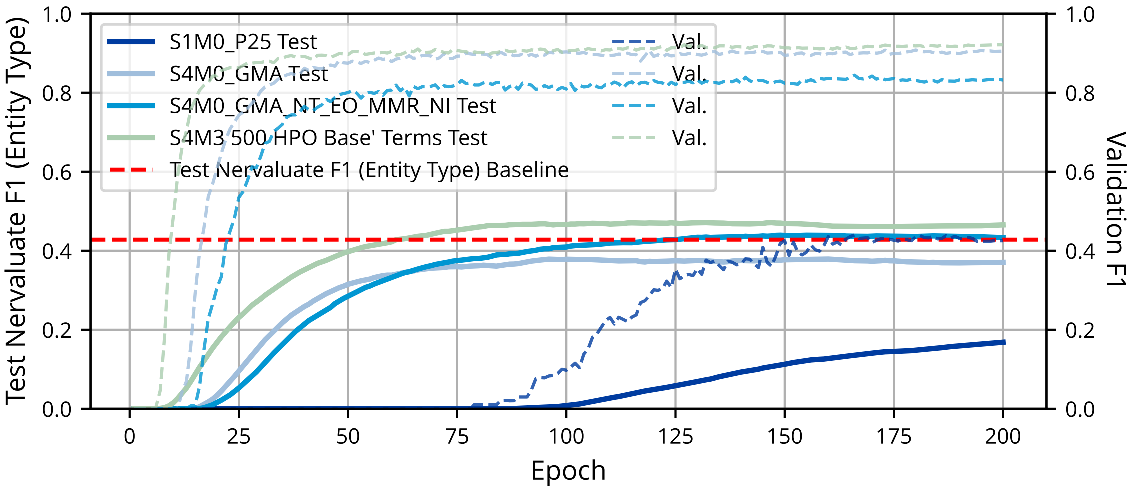


Figure 7.2: Final comparison of F1 terms Nervaluate evaluation scores.

7.2 Limitations

This thesis has several limitations that contextualize its findings and define the scope of its contributions. These relate mainly to the handling of negation, the lack of explainability analyses, and the open question of scalability to the full HPO ontology.

A first limitation is the absence of a dedicated negation detection pipeline. While the underlying language models can recognize negated expressions, the current system does not explicitly identify and filter out negated entities. Integrating a robust negation handling component, for example based on rule-based or supervised methods, could be crucial for further disambiguation between affirmed and negated phenotypes.

A second limitation is the lack of systematic explainability analysis of the behaviour of the model. Techniques such as SHAP, which attribute importance scores to input tokens or features, have not been applied to the experiments in Chapter 6. Consequently, design choices, such as the observation that splitting long sentences does not consistently improve prediction quality, remain insufficiently understood.

A third limitation concerns scalability with respect to the full HPO ontology. The current work focuses on a subset of HPO and does not yet exploit the complete hierarchical structure of the ontology. The scope of the experiments has been extended in Chapter 6, but only to a limited set of terms. It therefore remains open whether performance can be maintained when scaling to the substantially larger and more complex full ontology, with its increased label space and more fine-grained distinctions between related terms.

In summary, these limitations underline that the current system is a proof of concept rather than a mature clinical solution. The absence of explicit negation handling, the lack of detailed explainability studies, and the unproven scalability to the full HPO ontology constrain the generalizability and immediate applicability of the results, while at the same time indicating clear directions for future work.

7.3 Conclusion

The experiments show that commercial state-of-the-art language models with large context windows can rapidly prototype NER systems for specialized domains using simple generative requests. They enable quick construction of strong baselines without extensive engineering, which is especially useful where annotated data is scarce or costly.

At the same time, the results highlight that the configuration of the model requests plays a crucial role in determining performance. As suggested by the analysis related to RQ1, parameters such as generative request design, sampling strategy, context size, and instruction formulation substantially influence extraction quality. Careful tuning of these request configurations is therefore necessary to unlock the full potential of both commercial and open-weight models.

The findings for RQ3 indicate that open-weight models, despite their typically smaller context windows and the need for more complex setups, can serve as strong alternatives to commercial systems. When paired with well-designed synthetic data and alignment-based optimization, open-weight models can approach or even surpass the performance of more costly proprietary solutions in some settings. This makes them attractive for institutions with limited budgets or strict data governance requirements, as they offer greater control over data and deployment.

A key overarching conclusion, aligned with the investigations for RQ4, is that synthetic-only data can suffice to reach strong performance levels. When the generation process is guided by an understanding of the target domain and informed by dataset shape alignment, synthetic datasets can yield notable improvements over baseline metrics that rely solely on limited real-world annotations. This demonstrates that high-quality synthetic data is a viable resource for training NER models in specialized biomedical contexts.

The analysis associated with RQ2 shows that metrics derived from PCA plots, n-gram distributions, and self-BLEU scores provide useful signals about the similarity between generated and target data. These metrics can be employed to tune generation parameters and to identify synthetic datasets that better match the characteristics of the gold standard. However, the results also make it clear that such metrics are not yet a complete solution. Expert domain knowledge and further methodological research are required to fully interpret these indicators and to understand how they relate to downstream NER performance.

Overall, the work shows that the largest gains come from a deep understanding of the target domain and from aligning model configuration and data generation with that understanding. Instead of relying mainly on larger models or more data, domain insight, informed synthetic data generation, and careful evaluation form the most effective strategy for advancing NER in specialized biomedical domains.

7.4 Future Work

Several directions for future work emerge from the findings and limitations of this thesis. A first priority is closer collaboration with domain experts, such as clinicians and biomedical researchers, to refine the evaluation setup. Expert feedback could clarify whether span-level or term-level evaluations are more relevant for specific practical use cases, and which types of extraction errors are most critical in real-world workflows. This would enable task-oriented evaluation protocols that better reflect actual decision-making needs.

Another promising avenue concerns deeper integration of the HPO ontology into both modelling and downstream applications. Future research could investigate how the hierarchical and semantic structure of HPO can be used to generate recommendations, support differential diagnosis, or prioritize phenotypes in clinical decision support systems. Methods that explicitly exploit parent–child relationships, semantic similarity measures, or graph-based representations could improve both the robustness and interpretability of NER outputs.

In addition, extending the synthetic data generation pipeline with commercial models offers a potential path to further performance gains. Commercial models with larger capacity and richer training data may be able to produce more diverse and clinically realistic synthetic texts, which, when appropriately filtered and aligned, could further enhance model training. A hybrid strategy that combines open-weight models for scalable experimentation with commercial models for targeted data augmentation may prove particularly effective.

Finally, future work should address the currently missing components such as negation handling, more comprehensive explainability analyses, and large-scale scalability to the full HPO ontology. Addressing these aspects would bring the approach closer to deployment in real-world clinical environments and provide a more complete assessment of its strengths and limitations.

Bibliography

Abadifard, S., Bakhshi, S., Gheibuni, S., & Can, F. (2023). DynED: Dynamic Ensemble Diversification in Data Stream Classification. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 3707–3711. https://doi.org/10.1145/3583780.3615266

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., Lee, J. R., Lee, Y. T., Li, Y., Liu, W., Mendes, C. C. T., Nguyen, A., Price, E., de Rosa, G., Saarikivi, O., … Zhang, Y. (2024, December). Phi-4 Technical Report. https://doi.org/10.48550/arXiv.2412.08905

Alshaikhdeeb, B., Hemedan, A. A., Ghosh, S., Balaur, I., & Satagopam, V. (2025, July). Generation of Synthetic Clinical Text: A Systematic Review. https://doi.org/10.48550/arXiv.2507.18451

Anandika, A., & Mishra, S. P. (2019). A Study on Machine Learning Approaches for Named Entity Recognition. *2019 International Conference on Applied Machine Learning (ICAML)*, 153–159. https://doi.org/10.1109/ICAML48257.2019.00037

Answerdotai/ModernBERT-base · Hugging Face. (2024, December). Retrieved November 6, 2025, from https://huggingface.co/answerdotai/ModernBERT-base

Barr, A. A., Quan, J., Guo, E., & Sezgin, E. (2025). Large language models generating synthetic clinical datasets: A feasibility and comparative analysis with real-world perioperative data. *Frontiers in Artificial Intelligence*, *8*. https://doi.org/10.3389/frai.2025.1533508

Bressem, K. K., Papaioannou, J.-M., Grundmann, P., Borchert, F., Adams, L. C., Liu, L., Busch, F., Xu, L., Loyen, J. P., Niehues, S. M., Augustin, M., Grosser, L., Makowski, M. R., Aerts, H. J. W. L., & Löser, A. (2024). medBERT.de: A comprehensive German BERT model for the medical domain. *Expert Systems with Applications*, *237*, 121598. https://doi.org/10.1016/j.eswa.2023.121598

Builtjes, L., Bosma, J., Prokop, M., van Ginneken, B., & Hering, A. (2025, July). Leveraging Open-Source Large Language Models for Clinical Information Extraction in Resource-Constrained Settings. https://doi.org/10.48550/arXiv.2507.20859

Chan, B., Schweter, S., & Möller, T. (2020, December). German's Next Language Model. https://doi.org/10.48550/arXiv.2010.10906

Chitale, P. A., Gumma, V., Ahuja, S., Kodali, P., Uppadhyay, M., Sudharsan, D., & Sitaram, S. (2025, September). The role of synthetic data in Multilingual, Multi-cultural AI systems: Lessons from Indic Languages. https://doi.org/10.48550/arXiv.2509.21294

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., Marris, L., Petulla, S., Gaffney, C., Aharoni, A., Lintz, N., Pais, T. C., Jacobsson, H., Szpektor, I., Jiang, N.-J., … Bhumihar, N. K. (2025, July). Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. https://doi.org/10.48550/arXiv.2507.06261

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., … Zhang, Z. (2025, January). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. https://doi.org/10.48550/arXiv.2501.12948

Deepset/gbert-large · Hugging Face. (2024, March). Retrieved November 6, 2025, from https://huggingface.co/deepset/gbert-large

Diaz Ochoa, J. G., Mustafa, F. E., Weil, F., Wang, Y., Kama, K., & Knott, M. (2024). The aluminum standard: Using generative Artificial Intelligence tools to synthesize and annotate non-structured patient data. *BMC Medical Informatics and Decision Making*, *24*(1), 409. https://doi.org/10.1186/s12911-024-02825-4

Doshi, M., & Bhattacharyya, P. (2024). Synthetic Data for Multilingual NLP: A Survey.

Du, Y., Tian, M., Ronanki, S., Rongali, S., Bodapati, S., Galstyan, A., Wells, A., Schwartz, R., Huerta, E. A., & Peng, H. (2025). Context Length Alone Hurts LLM Performance Despite Perfect Retrieval.

Duan, F., Chapeau-Blondeau, F., & Abbott, D. (2024). Optimized injection of noise in activation functions to improve generalization of neural networks. *Chaos, Solitons & Fractals*, *178*, 114363. https://doi.org/10.1016/j.chaos.2023.114363

Enis, M., & Hopkins, M. (2024, April). From LLM to NMT: Advancing Low-Resource Machine Translation with Claude. https://doi.org/10.48550/arXiv.2404.13813

Feng, Y., Qi, L., & Tian, W. (2023). PhenoBERT: A Combined Deep Learning Method for Automated Recognition of Human Phenotype Ontology. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *20*(2), 1269–1277. https://doi.org/10.1109/TCBB.2022.3170301

Frei, J., & Kramer, F. (2023a). German Medical Named Entity Recognition Model and Data Set Creation Using Machine Translation and Word Alignment: Algorithm Development and Validation. *JMIR Formative Research*, *7*(1), e39077. https://doi.org/10.2196/39077

Frei, J., & Kramer, F. (2023b). Annotated dataset creation through large language models for non-english medical NLP. *Journal of Biomedical Informatics*, *145*, 104478. https://doi.org/10.1016/j.jbi.2023.104478

Gargano, M. A., Matentzoglu, N., Coleman, B., Addo-Lartey, E. B., Anagnostopoulos, A. V., Anderton, J., Avillach, P., Bagley, A. M., Bakštein, E., Balhoff, J. P., Baynam, G., Bello, S. M., Berk, M., Bertram, H., Bishop, S., Blau, H., Bodenstein, D. F., Botas, P., Boztug, K., … Robinson, P. N. (2024). The Human Phenotype Ontology in 2024: Phenotypes around the world. *Nucleic Acids Research*, *52*(D1), D1333–D1346. https://doi.org/10.1093/nar/gkad1005

Gemma Team, Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., Grill, J.-b., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., … Hussenot, L. (2025, March). Gemma 3 Technical Report. https://doi.org/10.48550/arXiv.2503.19786

google-bert. (2019). Google-bert/bert-base-german-cased - Hugging Face. Retrieved September 7, 2025, from https://huggingface.co/google-bert/bert-base-german-cased

Hua, H., Li, X., Dou, D., Xu, C.-Z., & Luo, J. (2021). Improving Pre-trained Language Model Fine-tuning with Noise Stability Regularization. *IEEE Transactions on Neural Networks and Learning Systems*, *36*(1), 1898–1910. https://doi.org/10.1109/TNNLS.2023.3330926

Huang, Y., Gao, Y., & Ren, C. (2025). A survey of data augmentation in named entity recognition. *Neurocomputing*, *651*, 130856. https://doi.org/10.1016/j.neucom.2025.130856

IBM Think. (2024, December). What Is Ground Truth in Machine Learning? | IBM. Retrieved October 10, 2025, from https://www.ibm.com/think/topics/ground-truth

Intfloat/e5-large · Hugging Face. (2025, February). Retrieved November 6, 2025, from https://huggingface.co/intfloat/e5-large

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., … Sayed, W. E. (2024, January). Mixtral of Experts. https://doi.org/10.48550/arXiv.2401.04088

Kaabachi, B., Despraz, J., Meurers, T., Otte, K., Halilovic, M., Kulynych, B., Prasser, F., & Raisaro, J. L. (2025). A scoping review of privacy and utility metrics in medical synthetic data. *NPJ Digital Medicine*, *8*, 60. https://doi.org/10.1038/s41746-024-01359-3

Kamath, G., & Vajjala, S. (2025, May). Does Synthetic Data Help Named Entity Recognition for Low-Resource Languages? https://doi.org/10.48550/arXiv.2505.16814

Keraghel, I., Morbieu, S., & Nadif, M. (2024, December). Recent Advances in Named Entity Recognition: A Comprehensive Survey and Comparative Study. https://doi.org/10.48550/arXiv.2401.10825

Kim, J., Choo, H., Shin, S.-Y., & Song, K. D. (2024). Synthesis and quality assessment of combined time-series and static medical data using a real-world time-series generative adversarial network. *Scientific Reports*, *14*(1), 19064. https://doi.org/10.1038/s41598-024-69812-7

King, A. (2024). Using Machine Translation to Augment Multilingual Classification.

Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., Mundlos, C., Horn, D., Mundlos, S., & Robinson, P. N. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *American Journal of Human Genetics*, *85*(4), 457–464. https://doi.org/10.1016/j.ajhg.2009.09.003

Kong, M., Fernandez, A., Bains, J., Milisavljevic, A., Brooks, K. C., Shanmugam, A., Avilez, L., Li, J., Honcharov, V., Yang, A., & Khoong, E. C. (2025). Evaluation of the accuracy and safety of machine translation of patient-specific discharge instructions: A comparative analysis. *BMJ Quality & Safety*. https://doi.org/10.1136/bmjqs-2024-018384

Kühnel, L., & Fluck, J. (2022). We are not ready yet: Limitations of state-of-the-art disease named entity recognizers. *Journal of Biomedical Semantics*, *13*(1), 26. https://doi.org/10.1186/s13326-022-00280-6

Li, A., Zhou, W., Hoda, R., Bain, C., & Poon, P. (2025, April). Comparing Large Language Models and Traditional Machine Translation Tools for Translating Medical Consultation Summaries: A Pilot Study. https://doi.org/10.48550/arXiv.2504.16601

Liu, Y., Zhao, R., Altinger, L., Schütze, H., & Hedderich, M. A. (2025, October). Evaluating Robustness of Large Language Models Against Multilingual Typographical Errors. https://doi.org/10.48550/arXiv.2510.09536

Luo, L., Yan, S., Lai, P.-T., Veltri, D., Oler, A., Xirasagar, S., Ghosh, R., Similuk, M., Robinson, P. N., & Lu, Z. (2021). PhenoTagger: A hybrid method for phenotype concept recognition using human phenotype ontology. *Bioinformatics*, *37*(13), 1884–1890. https://doi.org/10.1093/bioinformatics/btab019

Manz, T., Lekschas, F., Greene, E., Finak, G., & Gehlenborg, N. (2025). A General Framework for Comparing Embedding Visualizations Across Class-Label Hierarchies. *IEEE transactions on visualization and computer graphics*, *31*(1), 283–293. https://doi.org/10.1109/TVCG.2024.3456370

Mendes, J. M., Barbar, A., & Refaie, M. (2025). Synthetic data generation: A privacy-preserving approach to accelerate rare disease research. *Frontiers in Digital Health*, *7*, 1563991. https://doi.org/10.3389/fdgth.2025.1563991

Mishra, S., Arunkumar, A., Sachdeva, B., Bryan, C., & Baral, C. (2020, May). DQI: Measuring Data Quality in NLP. https://doi.org/10.48550/arXiv.2005.00816

Mistral AI. (2025, August). Mistral Small 3.2: 24B Multimodal LLM with Enhanced Instruction Following - mistral chat. Retrieved October 9, 2025, from https://mistral-ai.chat/models/small-3-2/

NLPIE Research. (2025, July). Nlpie/Llama2-MedTuned-Instructions · Datasets at Hugging Face. Retrieved October 10, 2025, from https://huggingface.co/datasets/nlpie/Llama2-MedTuned-Instructions

Noll, R., Berger, A., Kieu, D., Mueller, T., O. Bohmann, F., Müller, A., Holtz, S., Stoffers, P., Hoehl, S., Guengoeze, O., Eckardt, J.-N., Storf, H., & Schaaf, J. (2025). Assessing GPT and DeepL for terminology translation in the medical domain: A comparative study on the human phenotype ontology. *BMC Medical Informatics and Decision Making*, *25*, 237. https://doi.org/10.1186/s12911-025-03075-8

Noor-ul-Amin, M., Kazmi, M. W., Alkhalaf, S., Abdel-Khalek, S., & Nabi, M. (2024). Machine learning based parameter-free adaptive EWMA control chart to monitor process dispersion. *Scientific Reports*, *14*(1), 31271. https://doi.org/10.1038/s41598-024-82699-8

Ollama. (2024, April). Mixtral:8x22b. Retrieved August 13, 2025, from https://ollama.com/library/mixtral:8x22b

Ollama. (2025a). Deepseek-r1:32b. Retrieved October 15, 2025, from https://ollama.com/deepseek-r1:32b

Ollama. (2025b). Gemma3:27b. Retrieved September 10, 2025, from https://ollama.com/library/gemma3:27b

Ollama. (2025c). Gpt-oss:20b. Retrieved October 15, 2025, from https://ollama.com/library/gpt-oss:20b

Ollama. (2025d). Mistral-small:24b. Retrieved October 15, 2025, from https://ollama.com/library/mistral-small:24b

Ollama. (2025e). Phi4:14b. Retrieved October 15, 2025, from https://ollama.com/library/phi4:14b

OpenAI, Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R. K., Bai, Y., Baker, B., Bao, H., Barak, B., Bennett, A., Bertao, T., Brett, N., Brevdo, E., Brockman, G., Bubeck, S., Chang, C., … Zhao, S. (2025, August). Gpt-oss-120b & gpt-oss-20b Model Card. https://doi.org/10.48550/arXiv.2508.10925

Panahi, M. (2025, August). OpenMed NER: Open-Source, Domain-Adapted State-of-the-Art Transformers for Biomedical NER Across 12 Public Datasets. https://doi.org/10.48550/arXiv.2508.01630

Rao, H., Liu, W., Wang, H., Huang, I.-C., He, Z., & Huang, X. (2025, June). A Scoping Review of Synthetic Data Generation for Biomedical Research and Applications. https://doi.org/10.48550/arXiv.2506.16594

Richter-Pechanski, P., Wiesenbach, P., Schwab, D. M., Kiriakou, C., He, M., Allers, M. M., Tiefenbacher, A. S., Kunz, N., Martynova, A., Spiller, N., Mierisch, J., Borchert, F., Schwind, C., Frey, N., Dieterich, C., & Geis, N. A. (2023). A distributable German clinical corpus containing cardiovascular clinical routine doctor's letters. *Scientific Data*, *10*(1), 207. https://doi.org/10.1038/s41597-023-02128-9

Rohanian, M., Mehra, T., Miglino, N., Nooralahzadeh, F., Krauthammer, M., & Wicki, A. (2025). Towards scalable and cross-lingual specialist language models for oncology. *Scientific Reports*, *15*(1), 35480. https://doi.org/10.1038/s41598-025-19282-2

Scheible, R., Thomczyk, F., Tippmann, P., Jaravine, V., & Boeker, M. (2024). GottBERT: A pure German Language Model. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 21237–21250. https://doi.org/10.18653/v1/2024.emnlp-main.1183

Seeha, S., Wu, S., Hofenbitzer, J., Benzoni, C., Pallaoro, P., Scheible, R., Boeker, M., & Modersohn, L. (2025). German Medical NER with BERT and LLMs: The Impact of Training Data Size. *Studies in Health Technology and Informatics*, *327*, 798–802. https://doi.org/10.3233/SHTI250469

Segura-Bedmar, I., Martínez, P., & Herrero-Zazo, M. (2013, June). SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In S. Manandhar & D. Yuret (Eds.), *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (pp. 341–350). Association for Computational Linguistics. Retrieved September 12, 2025, from https://aclanthology.org/S13-2056/

Seow, W. L., Chaturvedi, I., Hogarth, A., Mao, R., & Cambria, E. (2025). A review of named entity recognition: From learning methods to modelling paradigms and tasks. *Artificial Intelligence Review*, *58*(10), 315. https://doi.org/10.1007/s10462-025-11321-8

Shen, T., Zhao, G., & You, S. (2023, April). A Study on Improving Realism of Synthetic Data for Machine Learning. https://doi.org/10.48550/arXiv.2304.12463

Shlyk, D., Groza, T., Mesiti, M., Montanelli, S., & Cavalleri, E. (2024). REAL: A Retrieval-Augmented Entity Linking Approach for Biomedical Concept Recognition. *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, 380–389. https://doi.org/10.18653/v1/2024.bionlp-1.29

Stockem Novo, A., & Gedikli, F. (2023). Explaining BERT model decisions for near-duplicate news article detection based on named entity recognition. https://doi.org/10.1109/ICSC56153.2023.00054

Thellmann, K., Stadler, B., Fromm, M., Buschhoff, J. S., Jude, A., Barth, F., Leveling, J., Flores-Herr, N., Köhler, J., Jäkel, R., & Ali, M. (2024, October). Towards Multilingual LLM Evaluation for European Languages. https://doi.org/10.48550/arXiv.2410.08928

TUM/GottBERT_base_last · Hugging Face. (n.d.). Retrieved November 6, 2025, from https://huggingface.co/TUM/GottBERT_base_last

Wang, J., Lu, Y., Weber, M., Ryabinin, M., Adelani, D., Chen, Y., Tang, R., & Stenetorp, P. (2025, February). Multilingual Language Model Pretraining using Machine-translated Data. https://doi.org/10.48550/arXiv.2502.13252

Weijers, R., & Bloem, J. (2025, May). An evaluation of Named Entity Recognition tools for detecting person names in philosophical text. In M. Hämäläinen, E. Öhman, Y. Bizzoni, S. Miyagawa, & K. Alnajjar (Eds.), *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities* (pp. 418–425). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.nlp4dh-1.36

Xu, L., Bing, L., & Lu, W. (2023). Better Sampling of Negatives for Distantly Supervised Named Entity Recognition.

Yang, J., Liu, C., Deng, W., Wu, D., Weng, C., Zhou, Y., & Wang, K. (2024). Enhancing phenotype recognition in clinical notes using large language models: PhenoBCBERT and PhenoGPT. *Patterns*, *5*(1), 100887. https://doi.org/10.1016/j.patter.2023.100887

Ye, X., Iyer, S., Celikyilmaz, A., Stoyanov, V., Durrett, G., & Pasunuru, R. (2023, June). Complementary Explanations for Effective In-Context Learning. https://doi.org/10.48550/arXiv.2211.13892

Zeng, X. (2024, August). Enhancing the Interpretability of SHAP Values Using Large Language Models. https://doi.org/10.48550/arXiv.2409.00079

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit mit dem Titel

Synthetic Data Strategies for Clinical Named Entity Recognition

selbstständig und nur mit den angegebenen Hilfsmitteln verfasst habe. Alle Passagen, die ich wörtlich aus der Literatur oder aus anderen Quellen wie z. B. Internetseiten übernommen habe, habe ich deutlich als Zitat mit Angabe der Quelle kenntlich gemacht.

Hamburg, 31. Dezember 2025

8 Appendix

8.1 Generative Request Templates

8.1.1 S3M0

```
# Vorlage für die Generierung synthetischer Daten (medizinischer Bereich)

Sie sind ein Experte für die Generierung medizinischer Daten. Sie erhalten einen
    bestimmten medizinischen Begriff, dessen Synonyme und eine Definition. Ihre Aufgabe
    ist es, realistische klinische Phrasen zu generieren, in denen dieser Begriff so
    verwendet wird, wie er in tatsächlicher medizinischer Dokumentation erscheint.

**Anweisungen:**

- Generieren Sie mehrere verschiedene klinische Phrasen (z. B. 40 - 50) für den
    angegebenen Begriff.
- Verwenden Sie in jeder Phrase entweder den Originalbegriff, ein Synonym oder eine
    gebräuchliche reale Schreibvariante.
- **Jedes Mal**, wenn der Begriff, ein Synonym oder eine Variante in einer Phrase
    vorkommt, muss dieser fett in Markdown-Syntax geschrieben werden (**so**).
- Stellen Sie sicher, dass die Phrasen natürlich, kontextuell korrekt und der realen
    klinischen Sprache entsprechend sind.
- Verwenden Sie verschiedene Satzstrukturen, Settings und Kontexte (z. B. Arztbriefe,
    Entlassungsberichte, Befundberichte usw.).
- Fügen Sie in Ihrer Ausgabe keine Definition, Synonyme oder Erklärungen hinzu - nur die
    Phrasen.
- Jede Phrase soll in einer neuen Zeile stehen.

---
```

```
## Beispiel-Eingabe

**Begriff:** Hypertonie
**Synonyme:** Bluthochdruck, HTN
**Definition:** Eine Erkrankung, bei der der Druck des Blutes auf die Arterienwände zu
    hoch ist.

---

## Beispiel-Ausgabe

- Die Patientin hat eine Vorgeschichte mit **Hypertonie**.
- Blutdruckmessungen deuten auf **Bluthochdruck** hin.
- In der Beurteilung zeigt sich eine unkontrollierte **HTN** trotz Medikation.
- Die Diagnose **Hypertonie** wurde nach mehreren Messungen bestätigt.
- Sie wurde letztes Jahr wegen **Bluthochdruck** medikamentös eingestellt.
- Die Familienanamnese ist positiv für **HTN**.
- **Hypertonie** wurde bei der Aufnahme festgestellt.
- Anhaltend **Bluthochdruck** trotz Lebensstiländerungen.
- Es gibt derzeit keine Hinweise auf Komplikationen durch **HTN**.
- Eine Überwachung auf **Hypertonie** wird empfohlen.

---

## Prompt-Vorlage

**Begriff:** Seeblaue Histiozytose
**Synonyme:** ['Meeresblauer Histiozyt']
**Definition:** Eine Anomalie der Histiozyten, bei der die Zellen aufgrund eines abnorm
    erhöhten Lipidgehalts ein meerblaues Aussehen annehmen. Histiozyten sind eine Art von
     Makrophagen. Seeblaue Histiozyten sind typischerweise große Makrophagen mit einem
    Durchmesser von 20 bis 60 Mikrometern und einem einzelnen exzentrischen Kern, dessen
    Zytoplasma bei einer Färbung mit Wright-Giemsa mit seeblauen oder blaugrünen Granula
    gefüllt ist.

---

**Generieren Sie 40 - 50 realistische klinische Phrasen wie oben beschrieben. Denken Sie
    daran: Jedes Vorkommen des Begriffs, seiner Synonyme oder Varianten muss fett in
    Markdown-Syntax geschrieben sein.**
```

8.1.2 No Terms Data Generation

```
# Vorlage für die Generierung synthetischer Daten (medizinischer Bereich)

Sie sind ein Experte für die Generierung medizinischer Daten.
Ihre Aufgabe ist es, realistische klinische Phrasen zu generieren, welche keinen Hinweis
    auf eine Krankheit (Pathologie) oder eine behandlungsbedürftige Veränderung geben.
Die Phrasen sollen beschrieben werden, wie sie in tatsächlicher medizinischer
    Dokumentation erscheint.

**Anweisungen:**

- Generieren Sie mehrere verschiedene klinische Phrasen (z. B. 1400 - 1500).
- Stellen Sie sicher, dass die Phrasen natürlich, kontextuell korrekt und der realen
    klinischen Sprache entsprechend sind.
- Verwenden Sie verschiedene Satzstrukturen, Settings und Kontexte (z. B. Arztbriefe,
    Entlassungsberichte, Befundberichte usw.).
- Fügen Sie in Ihrer Ausgabe keine Definition, Synonyme oder Erklärungen hinzu - nur die
    Phrasen.
- Jede Phrase soll in einer neuen Zeile stehen und als Liste formattiert sein.

---

## Beispiel-Ausgabe

- Herz: Rhythmisch, keine Geräusche, keine Stauungszeichen. Lunge: Vesikuläres
    Atemgeräusch, keine Nebengeräusche.
- Orientiert zu Person, Zeit und Ort. Pupillen isokor, reagieren prompt auf Licht. Keine
    Paresen, keine Sensibilitätsstörungen.
- Wundgebiet reizlos, keine Rötung, keine Schwellung, kein Austritt von Sekret.
- Äußere Genitale unauffällig. Vaginalschleimhaut rosig, keine Erosionen, keine Blutung.
    Portio glatt, geschlossen.
- Haut turgor normal, keine Effloreszenzen, keine Ulzerationen, keine Einblutungen.
- Trommelfelle beidseits intakt und lichtreflexreich, Gehörgang frei, Nasenschleimhaut
    rosig, Rachen unauffällig.

---

**Generieren Sie 1400 - 1500 realistische klinische Phrasen wie oben beschrieben.
Denken Sie daran: die Phrasen sollen keinen Hinweis auf eine Krankheit (Pathologie) geben.

Die Phrasen sollen sich an der Beispiel-Ausgabe orientieren.**
```

8.1.3 S4M1

```
# Vorlage für die Generierung synthetischer Daten im medizinischen Bereich

Sie sind ein Experte für die Generierung medizinischer Daten.
Sie erhalten eine Liste mit bestimmten medizinischen Phänotypen und ihrer ID.
    Gegebenenfalls sind zusätzlich Synonymen,  Definition und / oder einen Kommentar
    beigefügt.
Ihre Aufgabe ist es, eine realistische klinische Phrase zu generieren, in welcher
    wahlweise 2 bis 5 Phänotypen so verwendet werden, wie er in tatsächlicher
    medizinischer Dokumentation erscheint.

## Anweisungen

- Verwenden Sie in jeder Phrase entweder den originalen Phänotypen, ein Synonym, eine
    gebräuchliche reale Schreibvariante oder eine gebräuchliche Abkürzung.
- **Jedes Mal**, wenn der Phänotyp, ein Synonym oder eine Variante in einer Phrase
    vorkommt, muss dieser fett in Markdown-Syntax geschrieben werden (**so**).
- Stellen Sie sicher, dass die Phrasen natürlich, kontextuell korrekt und der realen
    klinischen Sprache entsprechend sind.
- Fügen Sie in Ihrer Ausgabe keine Definition, Synonyme oder Erklärungen hinzu - nur die
    Phrase.
- Nach jeder Phrase soll eine Liste mit allen ID's der tatsächlich genutzten Begriffen
    hinzugefügt werden.
- Die Reihenfolge der Liste ist sehr wichtig. Die Begriffe sollen in der identischen
    Reihenfolge wie diese in der generierten Phrase erscheinen aufgelistet werden.

### Kontext: Aktuelle Diagnose

Die Merkmale der Phänotypen sollen wie in einem Arztbrief in dem Abschnitt der aktuellen
    Diagnose beschrieben werden:
- Hauptdiagnose (Leitdiagnose) mit Status (bestätigt/Verdacht)
- Relevante Nebendiagnosen/Komorbiditäten und Differenzialdiagnosen
- Datum/Anlass der Diagnose, Stadium/Schweregrad oder Klassifikation
- Wichtigste Befundgrundlagen (klinisch, Labor, Bildgebung) und unmittelbare
    therapeutische Konsequenzen

## Beispiel Eingabe

### 34035: Pharyngealer Exsudat
Synonyme: Flüssigkeit aus dem Rachen, Rachenflüssigkeit
Definition: Aus der hinteren Wand des Rachens exudierte Flüssigkeit.
```

```
<...> 9 Weitere Phänotypen <...>

## Beispiel Ausgabe

**Epidurales Hämatom** im zervikothorakalem Übergang mit **Querschnittssymptomatik** der
    unteren Extremität und **Rückenschmerzen**
[100310, 10550, 3418]

# Ende der Beispiele. Start der tatsächlichen Liste von Phänotypen.

### 9412: Kegelförmige Epiphysen des 3. Fingers
Synonym: Kegelförmiges Ende des Mittelfingerknochens
Definition: Ein kegelförmiger Anblick der Epiphysen des 3. Fingers der Hand, der ein '
    Ball-in-a-socket'-Aussehen erzeugt. Die verwandte Entität "engelsförmige" Epiphyse
    bezieht sich auf eine ausgeprägte kegelförmige Epiphyse in Kombination mit einer
    Pseudoepiphyse am distalen Ende eines Phalanx.

### 30781: Erhöhter zirkulierender freier Fettsäurespiegel
Definition: Höhere als normale Werte der Fettsäuren, die im Plasma auftreten können, als
    Folge von Lipolyse in Fettgewebe oder wenn Plasmatriacylglycerole in Geweben
    aufgenommen werden.

<...> 8 Weitere Phänotypen <...>

# Finale Anweisungen

- Generieren Sie exakt eine klinische Phrase und wirklich nur die Phrase. Keine
    Kommentare oder Erklärungen.
- Die Reihenfolge und Anzahl der Phänotypen in der Phrase muss identisch mit der
    Reihenfolge der Phänotypen in der Liste sein.
- **Jedes Mal**, wenn der Phänotyp, ein Synonym oder eine Variante in einer Phrase
    vorkommt, muss dieser fett in Markdown-Syntax geschrieben werden (**so**).
```