Vrije Universiteit Amsterdam Netherlands Organisation for Applied Scientific Research

Master Thesis

Conversational Triple Extraction for Diabetes Healthcare Management Using Synthetic Data

Author: Stergios Ntanavaras (2778274)

1st supervisor: Prof.Dr. Piek Th.J.M. Vossen (VU)
daily supervisor: Dr. Maaike H.T. de Boer (TNO)
2nd reader: Prof.Dr. Sandjai Bhulai (VU)

A thesis submitted in fulfillment of the requirements for the Master of Science degree in Business Analytics at Vrije Universiteit Amsterdam

November 4, 2024

Acknowledgements

The purpose of this thesis is to fulfill the requirements for the Master of Science degree in the Computational Intelligence track of the two-year Business Analytics program at Vrije Universiteit Amsterdam.

The research conducted for this project was part of an internship at the Netherlands Organisation for Applied Scientific Research (TNO). Additionally, this study was a part of a collaborative project namely CHIP, involving TNO, Hybrid Intelligence Centre, and Vrije Universiteit of Amsterdam.

First, I want to express my gratitude to TNO for giving me the chance to conduct this research within their organization and for the experience I gained while working in such an esteemed environment. Then, I would like to thank my two supervisors, Maaike de Boer and Piek Vossen, for their support and invaluable guidance throughout this journey. Finally, I would like to thank Sandjai Bhulai for being the second reader of my thesis, as well as everyone at TNO who helped me with my work.

Abstract

The growing elderly population and modern lifestyle choices leading to an increase in chronic diseases like Type 2 Diabetes, which, in turn, is placing significant pressure on healthcare systems. In the Netherlands, this trend exposing notable gaps in accessibility and affordability of diabetes care, that traditional healthcare systems find difficult to manage. To address these challenges, a Dutch collaboration named CHIP aims to develop a Hybrid Intelligence system that reduces healthcare professional workloads and improves treatment quality by providing personalized patient care insights. This thesis, part of the CHIP collaboration, focuses on extracting structured Subject-Predicate-Object (SPO) triples from conversations between a Type 2 Diabetes patient and a caretaker agent, enriching a Knowledge Graph for advanced reasoning and personalized care. Due to confidentiality in healthcare, the thesis explores the effectiveness of conversational triple extraction (CTE) systems in diabetes healthcare management when used on synthetic data. The study examines the necessary properties for constructing realistic conversations, the application of generative AI to create these conversations and annotations, and identifies the most effective method for extracting SPO triples from these synthetic dialogues. The methodology includes the use of prompt-based learning with GPT-4 for developing diabetic personas and generating realistic dialogues. For annotating these dialogues with SPO labels, prompt-based learning with GPT-4o was applied. Both rule-based and transformer-based techniques were employed for CTE. In particular, a rule-based Syntactic Parsing approach, prompt-based learning with GPT-4o, and fine-tuning a BERT model were used for classifying in a token-level conversational sentences with SPO labels. The formation of SPO triples followed the classification of these labels. The rule-based method, while straightforward, showed limitations in handling the complexity and variability of conversational data. Conversely, GPT-4o emerged as the most effective method (F1-score of 0.6801), closely followed by BERT.

Contents

List of Figures

List of Tables

1

Introduction

1.1 Problem Statement

The global demographic landscape is undergoing a significant transformation characterized by an increasing proportion of elderly individuals. Currently, over 566 million people worldwide are aged 65 and over, a number expected to nearly triple by 2050 (3). This shift towards an older population is accompanied by a drastic growth in chronic diseases, which are rapidly becoming a dominating concern in all healthcare systems worldwide (4). Studies have indicated that the prevalence of diseases among the elderly population has generally increased over time (5). For instance, data from various global regions, including the Netherlands, have shown a significant increase in chronic conditions among the elderly, demonstrating a broader trend that spans decades.

The problem of aging is further compounded by the universal and dominant effect of lifestyle, which is, in turn, strongly impacting public health throughout modern societies. Sedentary behaviors, poor dietary habits, and smoking are among the predominant lifestyle choices that are leading to an increase in chronic conditions (6). Among these, is Type 2 Diabetes which is commonly observed in the elderly population (7). According to recent data, Type 2 Diabetes underscores a broader global health concern, as in 2021 approximately 10.5% of the population globally aged 20 to 79, were diagnosed with this condition, which is expected that by the year 2045 this prevalence will rise to 12.2% of the world's population, impacting 783.2 million individuals (8).

Consequently, the aging population, enhanced by adverse lifestyle choices, not only increases the prevalence of chronic conditions such as Type 2 Diabetes, but also puts a growing tension on both public and private healthcare systems globally, representing one of the most significant challenges to health and well-being in our times (4). Notably, such factors are driving up healthcare costs and placing an enormous pressure on healthcare providers (9). This rising demand for healthcare services not only strains infrastructure but also decreases the quality of life of the individuals and the economic stability of societies.

The management of diabetes extends well beyond medication and includes a lot of daily self-care activities that are both important and challenging to the patient (10). This involves complex daily meal planning routines, carbohydrate counting, regular exercise, blood glucose monitoring, and daily adjustments in their management plans. Therefore, given the considerable burden Type 2 Diabetes places on healthcare (11), a more comprehensive approach to improving health, lifestyle, and social factors may help to overcome the rising costs and improve healthcare accessibility and affordability in various regions of the Netherlands (9). In this context, traditional healthcare systems, under severe pressure due to the above-mentioned challenges, are not able to serve the type of personalized care that should be provided at an individual level (12). As a result, there is a significant gap in the delivery of healthcare that efficiently addresses the dynamic and personalized needs of diabetic patients.

1.2 A Promising Solution

In response to these challenges, the field of artificial intelligence (AI) presents a promising frontier (13). AI's rapid development within healthcare present revolutionary changes in the way care is executed. Intelligent algorithms are increasingly applied to perform sophisticated data analysis, giving personalized medical insights and recommendations. However, despite these advancements, modern AI technology often surpasses human abilities in many tasks but significantly lags in areas such as general world knowledge, common sense, and particularly the human capabilities of collaboration, adaptability, and ethical responsibility (14). Furthermore, AI systems are typically designed with a technology-centered approach, which can overlook the nuanced needs of patients and healthcare providers (15), (16). These gaps highlight the urgent need for innovative approaches that effectively integrate the strengths of both human and AI to deliver more effective, empathetic, and user-centered healthcare solutions.

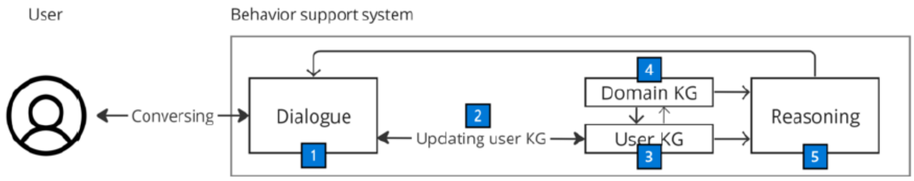


Figure 1.1: CHIP System Architecture.

Hybrid Intelligence (HI) could provide such a solution (14). HI leverages the combined strengths of human and AI, fostering a collaborative environment where both can thrive. It aims to augment human intellect and capabilities, thereby enhancing the ability to make meaningful decisions and perform actions that neither humans nor machines could achieve alone. This approach is particularly vital in healthcare, where the complexity of patient needs and the subtleties of medical care demand a nuanced understanding that AI alone cannot provide.

For that reason, a team of experts from several Dutch institutions formed a collaboration named CHIP with the goal of tackling the urgent problems facing the healthcare system today, particularly in the management of Type 2 Diabetes (17). Their primary objective is to develop an HI system that is not only technically competent but also ethically attuned and socially responsible. By providing more in-depth, personalized insights on patient care, this system will assist healthcare professionals in their work, reducing workloads and improving treatment quality.

An interactive prototype of this system as depicted in Figure 1.1, integrates five key components designed to adapt to individual user needs. It begins with a *Dialogue Component* that employs a rule-based approach to engage users in alignment dialogues, essential for gathering relevant user information. The *Information Extraction* component then transforms dialogue text into structured triples. These are organized within the *User Knowledge Graph (User KG)*, which uses an OWL-based ontology to detail user health data and preferences for personalized recommendations. The *Domain Knowledge Graph (Domain KG)* contains vital medical knowledge about diabetes treatments. Finally, the *Reasoning Engine* synthesizes data from both KGs to determine and execute the most suitable intervention based on the user's specific health data and personal preferences.

1.3 Research Goal

This thesis project is part of the CHIP collaboration, focusing specifically on the *Information Extraction* aspect. A key element of this aspect is the extraction of structured data in the form of Subject-Predicate-Object (SPO) triples from unstructured text, such as patient dialogues. These triples are crucial for building and continuously updating Knowledge Graphs (KGs) (18). By transforming raw conversations into structured triples, KGs become capable of supporting advanced reasoning about treatment options, potential health outcomes, and personalized patient recommendations (19). Such functions are critical for improving the understanding of patient interactions and adapting treatments to individual needs. Thus, this study will concentrate on the extraction of SPO triples from conversations between Type 2 Diabetes patients and an agent acting as a caretaker.

1.4 Research Questions

In the field of healthcare, the protection of medical data is extremely important, often posing a significant barrier to accessing authentic patient conversations for research purposes. This confidentiality is essential to protecting patient rights and preserving trust, but it also limits the amount of data available for developing and improving healthcare technologies. Consequently, our research will focus on evaluating the potential of creating and utilizing synthetic dialogues for conversational triple extraction, enabling its use in natural ones.

This brings us to the following research question: "What is the effectiveness of conversational triple extraction systems in diabetes healthcare management when used on synthetic data?".

To answer that research question, we need to consider the following subquestions:

- What properties should be included to construct effective conversations?
- How can generative methods be employed to generate realistic conversations and annotations?
- Which method is most effective in extracting triples from synthetic conversations related to diabetes healthcare?

1.5 Outline

The structure of the thesis is organized as follows. Chapter 2 presents a review of existing literature on triple extraction, large language models, and triple extraction for conversational contexts. Chapter 3 describes the methodology employed for data generation, including the development of personas and the creation and annotation of conversations, alongside exploring triple extraction techniques from these conversations. Chapter 4 demonstrates the performance results of the systems used, offering both general insights and a detailed error analysis of the methods applied. Chapter 5 discusses the limitations of the current study and suggests directions for future research. Finally, Chapter 6 concludes the thesis study.

2

Literature Review

This chapter provides a literature review on the topics of triple extraction, the use of large language models for this purpose, as well as the adaptation of triple extraction for conversational contexts and its implications for the medical field.

2.1 Triple Extraction

The majority of the data we come across in today's digital world is unstructured and mostly text-based (20). When it comes to extracting useful knowledge for gaining deeper insights, this flood of textual unstructured data poses serious challenges (21). In the field of natural language processing (NLP), information extraction (IE) refers to the process of extracting structured information from unstructured or semi-structured text data (22). The primary goal of this task is to transform free-form text into a format that can be easily understood and processed by computer systems, such as knowledge graphs (KGs).

IE encompasses a variety of tasks designed to identify specific types of structured information from unstructured text, including named entity recognition (NER), relation extraction (RE) and triple extraction (TE) among others. NER focuses on identifying and categorizing key entities in the text, like names of individuals, organizations, or locations. RE goes a step further by determining the relationships between identified entities, often classifying the interactions or connections between pairs of entities without necessarily forming a structured triple. In contrast, TE specifically focuses on constructing these structured triples (23). In the context of a KG, a triple is a data structure that represents real-world entities and the relationships that connect them (24). Each triple typically consists of three components: a Subject, a Predicate, and an Object. Thus, TE is about identifying and structuring relationships from textual data into a Subject-Predicate-Object (SPO) format, which is an important step in the creation and enrichment of KGs (25). By structuring knowledge in graph format, the relationships between entities are clarified, which enhances reasoning, interoperability, and efficient retrieval (19).

The literature presents distinct approaches for extracting SPO triples from texts based on NLP techniques, with **Rule-based methods** constituting one such approach. That kind of methods operate by applying a set of manually crafted linguistic patterns to identify and extract triples directly from syntactic structures. In his study, Shaun D'Souza (26) creates parse trees that illustrate sentence structure, clarifying grammatical relationships. The author uses depth-first search to detect noun phrases (NPs) as subjects and objects and verb phrases (VPs) as predicates. Subjects are found within NPs, while objects are located in NPs following VPs or connected by prepositional phrases. Predicates, consisting of main verbs and auxiliaries, are extracted from VPs. The system then combines these subjects, predicates, and objects into triples during the parse tree traversal. While rule-based approaches can be very effective in domains with limited and well-defined vocabularies, they suffer from a lack of scalability and flexibility (27).

Supervised learning techniques, in contrast, leverage labeled datasets to train models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory networks (LSTMs) to extract triples by recognizing and classifying parts of sentences as subjects, predicates, or objects. While these methods can adapt to more varied vocabularies and structures than rule-based systems, they still require large amounts of labeled data. However, according to Fei et al. (28), triple labeling can be time-consuming to produce and hence expensive. Additionally, the same study highlights that these methods may struggle with generalization to unseen relation types and entities, often suffering from overfitting when the training data is not representative of the target application's diversity.

According to Ding et al. (29), **transformer-based methods** such as those utilizing large language models (LLMs), have recently become predominant in the field of TE. These models benefit from deep contextual embeddings that capture nuanced language variations and complex dependencies within text. Transformer-based methods excel in generalization, thanks to pre-training on vast amounts of text, which allows them to perform well even on relatively smaller labeled datasets, providing an advantage over traditional supervised learning methods. The same study also highlights that by leveraging open-source LLMs we can achieve superior results in TE through effective prompting strategies and fine-tuning.

Therefore, despite their higher computational requirements, these models' ability to learn from the broader context make them the leading choice for current TE efforts (30).

While Ding et al. underscore the superiority of transformer-based methods in TE, Zhang et al. (24) take this a step further by testing how fine-tuning language models specifically for TE can enhance KG construction. They use tailored prompts for different aspects of TE to optimize the training process of these language models. This facilitates the augmentation of the original data and the creation of enriched training datasets. Subsequently, the enriched datasets are used to fine-tune several high-performance models, including Llama-2-7b, Mistral-t5-7b, and Vicuna-7b among the others. The performance is assessed on various metrics such as type, partial, exact, and strict accuracy, comparing favorably against benchmarks set by GPT-4 and previous GPT versions. Their results demonstrate that even smaller, fine-tuned models can surpass the performance of more generalized models like GPT-4, especially when high-quality training data is involved.

2.2 Large Language Models

Language modeling (LM) is a part of NLP that its objective is to learn a probability distribution of word sequences for predicting the probabilities of new missing ones (31). According to Zhao et al. (32), the research of LMs can be divided into four critical development stages. At first, in 1990, the statistical language models (SLM) was developed and originated the first development state. These models are about predicting the next word based on the most recent context. Another stage of development that emerged in 2013 is neural language models (NLM). They introduced the concept of distributed word representation, exemplified by the development of Word2Vec (33). Five years later (2018), the pre-trained language models (PLM) were developed. These models have been trained on a huge amount of not annotated data such as books, articles and websites (34). The goal was to capture the underlying patterns, structures, and semantic knowledge present in the text corpus. Zhao et al. (32) state that increasing model or data size in PLMs, frequently results in an enhanced model's ability to perform downstream tasks. Therefore, the term large language models (LLMs) was given by the research community to describe these large-scale PLMs.

A significant model in the landscape of LLMs is GPT-4 or Generative Pre-trained Transformer 4, developed by OpenAI (35). Among the various iterations, GPT-4 and its optimized version GPT-4o stand out as the most capable and advanced models in the GPT series. GPT-4 is a pre-trained model which utilizes a transformer architecture, like its predecessors, but with significantly increased scale and complexity, which allows it to understand and generate more nuanced text. This architecture employs layers of attention mechanisms that help the model weigh the importance of different words relative to each other within a given context.

Another prominent LLM is BERT, which stands for Bidirectional Encoder Representations from Transformers and was developed by Google researchers (36). BERT is based on the Transformer architecture, which uses attention mechanisms to interpret the contextual relationships between words in a text. Unlike other models that processed text in a single direction (either left-to-right or right-to-left), BERT processes the entire word sequence simultaneously. This bi-directional processing enables the model to capture a word's full context by considering both preceding and succeeding words. BERT is pre-trained with two tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In MLM, BERT learns to predict the identity of masked words based on surrounding context, while in NSP it learns to predict whether two segments of text naturally follow each other.

This pre-training on a large corpus of text provides LLMs like BERT and GPT-4 with a robust foundation of language understanding. This foundational knowledge can then be refined and adapted for specific applications through techniques such as fine-tuning and prompt-based learning.

2.2.1 Fine-Tuning

Fine-tuning involves adjusting the parameters of a pre-trained LLM to improve its performance for specific tasks or domains (37). While these pre-trained models have broad language skills, they often lack the deep understanding required for specialized areas. Fine-tuning addresses this gap by further training the model with domain-specific data, enhancing its accuracy and effectiveness for particular applications. As a result, this approach transforms a general-purpose language model into a specialized tool.

2.2.2 Prompt-Based Learning

When it comes to text generation, LLMs excel at producing coherent and contextually relevant content based on prompts or inputs (38). A prompt is a specific instruction or query that a human provides to an LLM to direct its behavior and generate desired outputs (38). Prompt-based learning is achieved through the design of prompts, which enables the model to generate diversified text based on different contextual environments and can be optimized and customized for different tasks and application scenarios (39).

2.2.2.1 Manual and Automatic Prompts

Depending on how prompts are generated, they are divided into two categories, namely manual and automatic prompts (40). Manual prompts are crafted by humans to provide precise and explicit instructions to the model regarding the type of data to concentrate on and the most efficient way to accomplish the task. These prompts work very effectively when the output needs to follow a certain structure or format and the input data is well-defined. On the other hand, automatic prompts are generated using different algorithms and methods without human engagement. These are categorized into discrete and continuous prompts (39). While continuous prompts take into account the context of the present conversation to provide accurate outputs, discrete prompts rely on predetermined categories to generate responses.

2.2.2.2 Zero-shot Prompting and Few-shot Prompting

There are two types of prompting: few-shot and zero-shot, which depend on how many instances humans provide an LM in order to train it for the downstream task. According to Wang et al. (39), zero-shot prompting refers to the practice of presenting a model with a prompt without providing any examples of the desired outcomes. The model uses its pre-trained knowledge and capabilities to generate a response. Conversely, in few-shot prompting, the model is provided with a few examples to help guide its responses. By providing concrete examples along with the prompt, few-shot learning enables models to adjust their outputs based on these examples, leading to higher accuracy and better task performance (41). This method is particularly beneficial in interactive scenarios, such as dialogue systems, where maintaining the flow of conversation is crucial (42).

2.3 Conversational Triple Extraction

Unlike classical triple extraction, which focuses on extracting knowledge from well-formed sentences in documents or articles, conversational triple extraction (CTE) is an advancement within the field of IE, specifically designed for dialogue-based text. The differences between conversational and traditional triple extraction does not only lie in the textual environment, but also in the nature of the language processed. One of the challenges in dialogue understanding is dealing with ellipsis and anaphora, which frequently occur in conversations (43). Ellipsis refers to situations where a phrase or clause mentioned earlier is omitted for simplicity, while anaphora involves replacing a mention with a pronoun to avoid repetition. For that reason, extracting triples from conversations is not an easy task. In order to search and identify potential SPO components, not only the current utterance but also its previous history sentences need to be considered (44). Yu et al. (45) highlight in their study that 65.9% of the relational triples they extracted, involve components that do not appear in the same utterance.

In addition to that, dialogues inherently involve multiple turns and can include complex interactions between participants. In their survey, Zhao et al. (46) explored various methods for CTE that can handle the sequential and more dynamic nature of dialogues. The study specifically discusses using graph-based methods to build dialogue graphs, capturing relationships between entities over multiple turns of conversation. It also explores attention mechanisms that focus on relevant sections of dialogue rich in relational cues and sequence labeling techniques that label dialogue text sequences to identify relationships at the token level. The survey highlights notable progress in accurately extracting triples from dialogues, emphasizing that combining graph-based methods, such as Graph Neural Networks, with attention-focused techniques like Transformer-based models, has led to significant performance enhancements across diverse dialogue types.

In a related study, Vossen et al. (47) utilize the Grounded Representation and Source Perspective (GRaSP) model in combination with the Simple Event Model (SEM) to extract conversational triples. This system transforms natural language from sensory inputs and interactions into structured triples of subjects, predicates, and objects, capturing the semantic essence of events. The process begins with rule-based categorization followed by the application of machine learning techniques to address ambiguities and inconsistencies in the dialogue, thereby enhancing the system's interaction and knowledge base capabilities.

2.3.1 Conversational Triple Extraction in the Medical Field

The task of CTE can be advantageous in the medical field. According to Souza et al. (48), enriching and constructing KGs with triples derived from healthcare dialogues, the management and analysis of clinical data can be greatly improved. Converting unstructured clinical conversations into structured triples that are semantically rich enables KGs to systematically organize and visualize complex relationships between medical entities, such as symptoms, treatments, and outcomes. This structured format not only makes medical records easier to use and understand for healthcare professionals but also supports more informed decision-making. Consequently, KGs enhanced with triples from dialogues may contribute to advances in personalized medicine and more effective healthcare services, ultimately benefiting patient outcomes.

3

Methodology

This chapter is divided into two main sections. Section 3.1 focuses on the creation of the dataset used in this study. Subsections within this section offer a more detailed look into specific aspects of the process: In Subsection 3.1.1 personas generation will be explained, including details about the prompts used. Subsection 3.1.2 will discuss the generation of conversations, also detailing the prompts employed. Subsection 3.1.3 talks about the preprocessing steps undertaken for preparing the dataset for subsequent analysis, while Subsection 3.1.4 will cover the annotation of the dataset.

On the other hand, Section 3.2 delves into the methods employed for extracting SPO triples from the conversations generated. In particular, this section provides detailed explanations of the unsupervised and supervised systems implemented for classifying in a token-level conversational sentences, whilst Subsection 3.2.1 outlines the process of constructing SPO triples from token-level classifications. Lastly, Subsection 3.2.2 discusses the evaluation of these systems.

3.1 Data Generation

This section focuses on the creation of the dataset utilized in this study. All the steps involved in this process are illustrated in Figure 3.1.

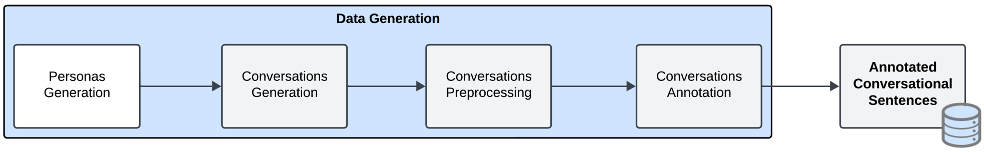


Figure 3.1: Data Generation Workflow

As mentioned in Section 1.4, due to privacy protection of medical data, natural conversations between Type 2 Diabetes patients and caretakers are not available. Therefore, to carry out this thesis project, creating such conversations was necessary. However, the tight timeline of the current project and the high costs involved in creating a human-generated dataset, present significant challenges to manually constructing such conversations. Research shows that when prompting effectively, LLMs can use their language generating capabilities to generate datasets that are similar in quality to those provided by humans, which reduces the time and expense associated with the process (49). Thus, for this study, we leveraged the capabilities of LLMs to generate our dataset.

3.1.1 Persona Generation

This Subsection examines the decisions made, the reasoning behind them, and the approach used to create realistic and distinct personas representing Type 2 Diabetes patients.

For making the dataset more natural, we started by creating diverse personas, each representing a unique profile of a Type 2 Diabetes patient. These personas were carefully created to capture a wide range of demographic traits, medical histories, and treatment adherence behaviors, providing a thorough representation of the patient population. The selection of specific demographic attributes for the creation of personas was informed by robust epidemiological and sociocultural research, along with insights from consultations with two experts from TNO, ensuring that the personas accurately reflect the diversity and nuances of the Type 2 Diabetes patient population in the Netherlands.

In 2017 a Dutch nationwide consortium of diabetologists, paediatric endocrinologists, and Diabetes patients has founded a national outpatient Diabetes care registry named Dutch Pediatric and Adult Registry of Diabetes (DPARD). Between November 2017 and January 2020, 20.857 patients were included from 8 (11%) Dutch hospitals with a level of care distribution representative of all diabetic outpatients in the Netherlands. Among patients with known Diabetes type, 51% had Type 2 Diabetes. Bak et al. (50) aimed to describe the implementation of DPARD and to provide an overview of the characteristics of patients included during the first 2 years. The results showed that the mean age for Type 2 Diabetes patients was 65.9 years (14.0 – 98.0). Therefore, for the ages of our personas, we selected values proximate to 65, specifically 55, 60, 65, 70, and 75.

Moving on to their ethnicities, data from StatLine, the online database of the Central Bureau of Statistics (CBS) in the Netherlands (51), indicate that individuals of Moroccan descent have the highest percentage of Type 2 Diabetes in the Netherlands, with 7.09% of the population affected. The second highest rate is among the Surinamese population, where 6.15% have Type 2 Diabetes. People with a Turkish background also show a relatively high prevalence, with 5.95% affected. In comparison, individuals with a Dutch background have a lower prevalence rate of 1.99%. In another study, Voortman et al. (52) mentioned that ethnic minorities living in Western societies may have a higher prevalence of Diabetes. Migrants from Turkey and Morocco are among the largest ethnic minority groups in Europe including the Netherlands. The current study included 375 Turkish, 314 Moroccan and 417 Dutch individuals aged 18–70 years. The results showed that the prevalence of Diabetes in the Amsterdam population was significantly higher in Turkish (5.6%) and Moroccan (8.0%), compared to Dutch individuals (3.1%). Guided by the insights from the studies referenced, we have selected Moroccan, Surinamese, Turkish, and Dutch ethnicities for the creation of our personas in this research.

A special consideration was given also to the selection of names to ensure more accurate cultural representation, enhancing the authenticity of each persona. The names for each persona, including both male and female, were carefully selected based on their prevalence within the aforementioned ethnic groups, as commonly indicated across various online sources [1]. As a result, the chosen names are Mohammed, Abdullah, Aicha, and Fatima for Moroccans; Rudolf, Johan, Julia, and Ingrid for Surinamese; Ali, Mehmet, Ayşe, and Fatma for Turkish; and Pieter, Jan, Johanna, and Maria for the Dutch.

Using these attributes, all possible combinations were generated, respecting gender-specific naming conventions to ensure realism and cultural appropriateness. After randomly shuffling these combinations to eliminate any bias, 16 unique profiles were carefully selected to maintain an equal gender and ethnicity balance. All 16 selected profiles are depicted in Table 3.1.

[1] The names were selected after reviewing multiple online sources to determine their commonality within each ethnic group. This approach is not dependent on any one source but, rather, represents information shared across various unofficial web-based resources.

Name	Gender	Ethnicity	Age
Jan	Male	Dutch	70
Maria	Female	Dutch	60
Abdullah	Male	Moroccan	70
Mohammed	Male	Moroccan	75
Aicha	Female	Moroccan	75
Julia	Female	Surinamese	55
Ingrid	Female	Surinamese	75
Johan	Male	Surinamese	60
Fatima	Female	Moroccan	65
Rudolf	Male	Surinamese	55
Ayşe	Female	Turkish	70
Pieter	Male	Dutch	75
Johanna	Female	Dutch	75
Fatma	Female	Turkish	65
Ali	Male	Turkish	60

Table 3.1: Selected Profiles

3.1.1.1 Prompts for Persona Generation

Each profile was enriched with additional details through a series of manual prompts designed to get nuanced and varied responses from a generative model. These prompts guided the generation of comprehensive profiles, including detailed descriptions of personal and demographic backgrounds, physical features, psychological traits, professional and educational experiences, social interactions, daily routines, healthcare interactions, economic factors, communication styles, and technological engagements. The full prompts used for this generation can be found in Figure 6.1 in the Appendix.

As a generative model, we selected 'GPT-4', identified as the most proficient model within the GPT series according to a report from OPENAI (35). In particular, they mention that 'GPT-4' is more reliable, creative, and capable of handling much more nuanced instructions compared to other GPT models. Two parameters namely '`temperature`' and '`max_tokens`' play important roles in controlling the behavior and output of the model. The first one affects the randomness or creativity of the responses produced by the model, while the later one controls the maximum number of tokens each response contains. To generate detailed and informative persona descriptions, the '`temperature`' parameter was set to 0.9 to optimize for creativity, whilst the '`max_tokens`' was left at its default value of 4096 from the API configuration, since the length of the descriptions was not a primary concern.

Instead, the focus was on ensuring that the descriptions were comprehensive and rich in detail.

According to the responses we got, the personas vary significantly in their technological proficiency: some are tech-savvy, regularly using digital tools to monitor their health, while others, less familiar with technology, are nonetheless open to learning if it can improve their health management. Relationships with healthcare providers also differ among the personas; some maintain a realistic, albeit infrequent, interaction due to high demands on general practitioners, while others enjoy good rapport but struggle with complex medical terminologies and explanations. Health characteristics are prominently featured, with most personas being slightly overweight, particularly around the abdomen, which is a noted risk factor for Type 2 Diabetes. Lifestyle factors also vary; some personas have quit smoking and experienced weight gain, others continue to smoke, and some never smoked. Economic conditions among the personas range from stable to challenging, with some struggling with the costs associated with their treatment. Education levels vary as well, with most having received a basic education up to high school level. An example of a persona description created is available in Appendix 6.

3.1.1.2 Persona Evaluation

The complete persona descriptions were reviewed and validated by two TNO experts who work on the healthcare section and are familiar with Type 2 Diabetes. In this revision, we adjusted several personas to represent socio-economic scenarios that align with the nature of our problem that discussed in Section 1.1. Thus, while the original descriptions included four financially well-off personas and five with strong relationships with their healthcare providers, these were changed to reflect personas who might struggle economically and have less favorable interactions with their healthcare systems. Such changes were made to represent patients who would benefit from a system like the one the CHIP collaboration aims to create, addressing the issues of expensive and inaccessible Diabetes healthcare.

This process ensured each persona was not only detailed but also equipped with realistic and profound characteristics that reflect the nature of our problem and typical challenges or circumstances faced by individuals managing Type 2 Diabetes. Finally, the personas were ready for use in the study's subsequent phase, which aims to generate realistic patient-agent interactions.

3.1.2 Conversation Generation

This Subsection discusses the decisions made and the methods used to create natural and diverse conversations between a Type 2 Diabetes patient and an agent.

Using the personas created in the previous step, we generated detailed and varied conversations between these personas reflecting a Type 2 Diabetes patient and an agent acting as a caretaker. Each conversation consists of around 4 to 6 exchanges, clearly marked by initials indicating who is speaking: “P” for the patient and “A” for the agent. An exchange is defined as a sequence where one speaker makes a statement and another speaker responds, with these two utterances collectively counting as one exchange.

For each persona, 16 detailed conversations were generated, amounting to a total of 256. This specific number was chosen to ensure a balanced dataset by maintaining an even distribution of conversations initiated by both the agent and the patient. As a result, each persona is represented by 8 conversations initiated by the patient and 8 by the agent. Additionally, we limited the generation of conversations to 256, guided by research showing that when LLMs process data beyond a certain volume, they tend to produce outputs that lack diversity, often generating similar or repetitive data (53).

For generating the conversations, we used prompt-based learning with ‘GPT-4’. As mentioned in Subsubsection 2.2.2, this process involves crafting specific prompts to guide the behavior of the model to generate dialogues that meet desired criteria. We chose ‘GPT-4’ over other LLMs for reasons beyond those discussed previously in Subsubsection 3.1.1. Notably, it excels in creative language generation and maintaining contextual coherence throughout extended dialogues. In addition to that, Duan et al. (54) found that ‘GPT-4’ can generate human-style multi-turn dialogues with impressive quality. They also highlighted that other LLMs struggle to generate high-quality multi-turn conversations due to their poor ability to follow instructions, tendency to generate lengthy utterances, and limited overall performance. Furthermore, it’s worth noting another student’s thesis, which demonstrated that ‘GPT-4’ achieved the highest average score across different metrics such as naturalness, coherence, fluency, and groundedness (55). This was in the context of creating realistic conversations between patients and agents, where it outperformed other state-of-the-art LLMs like ‘LLaMA 3’, ‘MedAlpaca’, and ‘Meditron’. Similar to the persona generation phase, two parameters of the model were set for generating the conversations. The ‘`temperature`’ parameter was set to 0.9 to foster creativity, whilst the 'max_tokens' parameter was set to 300 to help keep responses concise and maintain focus. Below are two examples of the conversations generated.

1) A conversation initiated by the agent concerning the persona named Aicha from Table 3.1:

> A: Aicha, can you tell me more about your current strategy for managing your smoking habit, specifically given its impact on your diabetes and overall health?
> P: I'm trying to quit smoking but it's really hard. Every now and then I find myself reaching for a cigarette, especially when I'm stressed or after meals.
> A: It's great that you're trying to quit, Aicha. Indeed, it can be challenging, especially when it's associated with your daily routines like meals. Have you tried using nicotine replacement therapy or considered any support programs to aid in quitting?
> P: No, I haven't tried nicotine replacement therapy. I heard about it but was not sure how it works.
> A: Nicotine replacement therapy provides you with nicotine in the form of gums, patches, sprays, inhalers, or lozenges to reduce the cravings associated with nicotine withdrawal. It's advisable to discuss with your doctor or a pharmacist about using these therapies to ensure they're suitable for you given your current medications and health conditions.
> P: I see, that sounds like something that might help. What about support programs?
> A: There are many support programs available that provide advice, encouragement, and tips to help people quit smoking. They can be in the form of local community groups, telephone-based services, or online forums. You could also consider discussing with your healthcare provider about referral to a smoking cessation program.
> P: I'll consider these options. Thank you!

2) A conversation initiated by the patient concerning the persona named Jan, as detailed in Table 3.1:

> P: I've been struggling with some anxiety lately due to my health. Does it impact my diabetes in any way?
> A: Yes, Jan, anxiety can impact your blood sugar levels. When you're anxious, your body releases stress hormones, which can cause your blood sugar levels to rise. It's important to manage anxiety to keep your diabetes under control.
> P: What would you suggest to manage this anxiety? I already try meditation but it seems insufficient.
> A: Incorporating physical activities, like your regular walks or cycling, can help reduce anxiety. You could also try other relaxation techniques like progressive muscle relaxation or guided imagery. Additionally, talking about your concerns with someone you trust can also help.
> P: My family is supportive but I worry about burdening them with my health con-

cerns.
A: That's a common concern, Jan. You might find it helpful to talk to a professional counselor or therapist who specializes in chronic illnesses. They can provide coping strategies and emotional support in a structured, understanding environment.
P: I will consider this. Does management of anxiety also involve changes in my medication or diet?
A: It doesn't usually require changes in medication unless recommended by your doctor. As for your diet, maintaining a balanced diet is beneficial. Some foods like those rich in magnesium and Omega-3 fatty acids are known to help reduce anxiety.
P: Thank you. I'll try to incorporate these suggestions into my lifestyle.

3.1.2.1 Prompts for Conversation Generation

```
messages = [
            {"role": "system", "content": SYSTEM_PROMPT},
            {"role": "assistant", "content": AGENT_PROMPT},
            {"role": "user", "content": PATIENT_PROMPT}
        ],
```

Figure 3.2: Prompt Template for Conversation Generation

Figure 3.2 shows the prompt template for generating the conversations. The "`user`" role is assigned to the patient in the simulated conversation. This aligns with typical patterns in conversational AI, where the "`user`" is the one who initiates questions and seeks assistance or information. The role of the "`assistant`" was designated for the agent, providing guidance, and responding to the user's queries. Assigning the "`assistant`" role to the agent rather than the "`user`" ensures that the agent remains in a supportive and guiding position. The "`system`" role acts as the control mechanism, initiating and directing the conversation based on predetermined scenarios, much like a conductor supervising the interaction between the "`assistant`" and "`user`".

To enable meaningful and realistic interactions between the personas and the agent, careful attention was given to the design of the manual prompts used for generating the conversations. This design process was guided by existing literature and refined through iterative testing to ensure that the conversations would be both authentic and effective in simulating real-life interactions in Diabetes management.

Agent Prompt

A study conducted by Nguyen et al. (56) is focusing on establishing foundational design principles for developing effective conversational agents (CAs) that utilize artificial intelligence to improve diabetes care. They suggest that the agent should maintain a friendly tone, utilize small talk, and adapt the conversation flow to the individual needs and responses of the patient. This personalization can help in building trust and comfort, which are essential for long-term engagement (57). Therefore, the agent prompt was designed to engage in personalized and adaptive interactions with patients, tailoring the conversation flow to meet each individual's specific needs and responses. The relevant segments within the prompt are:

"... your interactions should be personalized and adaptive. Personalization involves establishing common ground through the use of personal pronouns and engaging in small talk. This helps in building trust and comfort, essential for long-term engagement.",

"Adapt your conversation flow based on the individual needs and responses of the patient."

According to Clark et al. (58) there is a need for the agent to understand patient input clearly and quickly, ideally without repeating themselves. So, we highlighted within the agent prompt the importance of accurate listening to the agent which ensures that it understands and remembers user inputs which can enhance the conversation's relevance and depth. The relevant part within the prompt is:

"Accurate listening is crucial; ensure that you understand and remember user inputs to ask relevant follow-up questions, enhancing the conversation's relevance and depth."

Additionally, we incorporated instructions for the agent prompt to employ empathic language and maintain a consistent, warm, and understanding tone throughout the conversation. According to Liu et al. (59) expressions of sympathy and empathy by chatbots are generally preferred over purely informational interactions. The agent should be capable of recognizing and adapting to the emotional state of the patient. This includes using language that is supportive or uplifting if the patient is feeling down or anxious about their health. The ability to adjust language based on emotional cues is critical when dealing with health-related topics, where the patient's emotional state can significantly impact the effectiveness of communication (60). The relevant segment within the prompt is:

"Employ empathic language, maintaining a consistent, warm, and understanding tone throughout the conversation. Sensitivity to the patient's emotions is key ..."

Furthermore, according to Walker et al. (60) the agent needs to be aware of the context in which the dialogue is taking place, which could include understanding the patient's current health status, medical history, and possibly even the time of day or recent events in the patient's life. That is why we instructed it appropriately to be aware of its context.

Also, we pointed out the usage of concise language to communicate. The information delivered by the AI should be straightforward and concise, avoiding medical jargon that may confuse the patient. This clarity helps in enhancing patient comprehension and adherence to medical advice (56). The relevant segment within the agent prompt is:

"Use concise language to communicate. This involves choosing precise, simple, and effective words that avoid medical jargon, making it easier for the patient to understand."

Finally, we instructed the agent to request information in a conversational manner, asking for one piece at a time instead of everything at once, in order to maintain a natural and engaging dialogue. This approach mimics natural human conversation, which can be more comfortable and less overwhelming for the patient. Adam et al. (61) found people disclose more to a chatbot when it requests information conversationally one-by-one rather than all at once. The relevant segment within the prompt is:

"Request information conversationally, one piece at a time, rather than all at once, to keep the dialogue natural and engaging."

The full agent prompt is depicted in Figure 6.3 in Appendix.

Patient Prompt

For the design of the patient prompt, we individually hardcoded the persona descriptions created in the previous phase. For each conversation generated by the model, we used a specific persona description. Afterward, for generating subsequent conversations, we systematically rotated to a different persona description. We followed this process until we had successfully generated 16 conversations for each persona.

System Prompt

Turning to the design of the system prompt, strategies such as capitalization and repetition were employed to enhance the clarity and impact of the instructions given to the generative model. We capitalized certain words to draw attention to specific parts of a prompt, potentially making it more effective by emphasizing key terms or commands that are crucial for the intended output. This approach was discussed by White et al. (62) and could be particularly useful in settings where clarity and precision are necessary to guide the model's response accurately.

Repetition is another strategy that was discussed by White et al. (62) and can be leveraged to reinforce the importance or relevance of certain instructions within a prompt. By repeating key phrases or concepts, prompt engineers can help the model to focus more on these elements, potentially leading to more consistent and reliable outputs.

In particular, in our case, we capitalized the word "*varied*" to emphasize the uniformity in the way the outputs start. In addition, we focused on the instruction "*each conversation must contain exactly 4 to 6 exchanges*" in both uppercase and repetition to address cases where the output contained fewer than 4 exchanges.

In the final phase of our study, we explored both few-shot and zero-shot prompting methods. Unlike prior studies, we made the decision to independently assess the efficacy of these techniques by directly observing their outcomes. The results indicated that conversations generated through zero-shot prompting started in a similar manner, lacking variability. Moreover, these conversations frequently failed to meet the specified guidelines. An example is that the model typically producing only 2 exchanges rather than the 4 to 6 outlined in our prompts. Below is an example of a conversation generated using zero-shot prompting:

> A: Hi Alex, I understand that managing diabetes can be challenging, especially when it comes to maintaining a healthy diet. You mentioned your fondness for Dutch cheeses, which can be high in fat. Have you considered any alternatives or strategies to manage your cheese intake while still enjoying it?
> P: I've tried a few times to cut down on it, but it's been quite difficult for me. Cheese is a big part of our culture and I just love the taste. Any suggestions?
> A: Absolutely, moderation is key. You don't have to eliminate cheese completely from your diet. Instead, aim for smaller portions. Choose low-fat or reduced-fat varieties when possible. Also, pairing cheese with fresh fruits or vegetables can make a balanced snack. How does that sound?
> P: That sounds doable. I'll try to follow these suggestions. Thanks for the advice.

Therefore, we implemented few-shot prompting in our system's prompt design. This involves providing the model with a small number of examples, in our case 4, to illustrate the desired format and structure of the conversations. This helps guide the model in generating responses that align with the specified criteria. The full system prompt can be found in Appendix in Figure 6.2.

Based on the research conducted and the decisions made to create natural and effective conversations, all the properties used to design these conversations are outlined in Table 3.2.

Properties
Personalization and Common Ground
Accurate Listening
Empathetic Language and Friendly Tone
Engagement Through Small Talk
Contextual Awareness
Conciseness
Sequential Information Gathering

Table 3.2: Properties for Constructing Effective Conversations

3.1.3 Conversation Preprocessing

This Subsection discusses the decisions made and the methods used to prepare the conversations for subsequent usage.

After generating the conversations, we applied two preprocessing methods. The first involved expanding contractions within each conversation. For example, contractions like 'I've' and 'don't' were expanded to 'I have' and 'do not' respectively. This expansion clarifies the individual tokens within the conversational sentences, making the subsequent annotation process easier.

Afterwards, the conversations were divided into three distinct sets: training, validation and testing. To ensure that the dataset was balanced and unbiased, we decided to utilize 10 of the 16 conversations from each persona for training, with the remaining 6 designated for testing and validation. We also applied stratification, a technique that ensures that each segment of the dataset closely represents the overall distribution, particularly in terms of key variables or categories. For instance, it maintains the proportion of examples from various subgroups, such as gender, ethnicity, or other characteristics. By doing so, stratification ensures that all sets are diverse and equally represent all personas, significantly reducing sampling bias and thereby facilitating a fair and effective model training process.

3.1.4 Conversation Annotation

This Subsection discusses the methods used for annotating in a token-level the conversations of our datasets (training, validation and testing) with SPO labels.

3.1.4.1 Validation and Test Datasets Annotation

Given the critical nature of medical information, it was necessary for the validation and test sets to be manually annotated. This manual intervention was undertaken to ensure the highest levels of data quality and reliability, which are essential in medical contexts. The annotation schema designed, focus on extracting at a token-level, structured semantic relationships in the form of Subject-Predicate-Object (SPO), which are crucial for enhancing the KG of the CHIP project (17). It would be more practical to annotate only the patient utterances, as they contain the essential information for inclusion in the knowledge graph (KG). The agent's utterances are already stored within the KG, thus they do not require further annotation. However, focusing solely on patient responses would result in insufficient data for effectively training and evaluating our models. One possible solution could be to generate more conversations, but as mentioned in Subsection 3.1.2, generative AI models often produce repetitive outputs after a certain volume, which could reduce the diversity and usefulness of the generated data.

To ensure high-quality and precise annotations, the INCEpTION annotation tool was utilized, a cutting-edge platform specifically designed for annotation tasks (63). Annotators were provided with a comprehensive guide detailing the steps for installing and using the tool, including how to import project files and export annotated data. Additionally, a detailed set of instructions was created to guide the annotators in annotating according to predefined criteria, thereby ensuring consistency and reducing ambiguity in the annotations.

Firstly, the instructions clarified that annotators should only label sentences containing significant information about Diabetes management for inclusion in the KG, rather than annotating every sentence in a conversation. It was also emphasized that our focus is on the information contained within a conversation and not how the conversation flows. Then, a set of clear, detailed guidelines were established to define what constitutes a Subject, Predicate, and Object in the nuanced setting of patient-agent dialogues. Considering the complexity and variety of the conversations, the schema was developed to be robust enough to handle special cases, such as passive voice, compound Subjects and Predicates, and subordinate clauses. Furthermore, specific cases where the Subject or Predicate may be implied were addressed, ensuring that annotations capture the intended meaning accurately.

For example, in the sentence "*The doctor evaluates the patient's condition and prescribes medication.*" features a compound Predicate, where "*evaluates the patient's condition*" and "*prescribes medication*" are actions performed by the same Subject, "*The doctor*". An example of an implied Subject can be found in the sentence "*Adjust your insulin dose accordingly.*" where the Subject "you" is understood but not explicitly stated.

For sentences that are part of a question-answer sequence within the dialogue, the schema specifies different annotation approaches based on the nature of the response. To be more precise, if a sentence poses a question and the next or subsequent one provides its corresponding answer, if the answer is a simple affirmation, as illustrated in Example 1 below, only the question is annotated. The reason for that is because simple affirmations do not contain meaningful substantive information that would enhance the KG; the valuable information is contained within the question itself. On the other hand, if the response is detailed and offers substantial information, as demonstrated in Example 2 provided below, only the answer is annotated because it includes the essential details for the KG. When the response consists of both an affirmation and additional information, as shown in Example 3 below, both the question and answer are annotated, as each part holds important information. The complete set of instructions is provided in Figure 6.4 in the Appendix section.

Example 1

A: Have you taken your medication today?
P: Yes!

Example 2

A: What did you eat for breakfast?
P: I had oatmeal with bananas.

Example 3

P: Should I be concerned about dehydration?
A: Yes, the risk of dehydration and low blood pressure can increase while fasting.

Following these instructions, the test set was manually annotated by two experts in linguistics from TNO and myself. A pre-annotation meeting was conducted for all annotators, where the functionalities of the annotation tool and the instructions were introduced and explained. During the annotation process, time constraints allowed for only 200 of the 800 sentences to be annotated by the annotators for the inter-annotator agreement (IAA) calculation, which was performed in a token-level. According to Table 3.3, the annotation trial achieved a substantial agreement with a Fleiss' Kappa score of 0.657. In order to resolve the disagreement that occurred in those 12.26% utterances, a majority voting approach was adopted, which was to select the label that the majority of the annotators assigned to a certain sentence. There were 61 tokens that received three different labels from the annotators. Since there was no majority consensus, a random selection method was applied to these tokens. The rest of the conversations were annotated carefully by myself and were reviewed by a TNO expert. After the majority voting review, the initial 200 were re-examined to guarantee consistency throughout the entire test set. The same instructions and techniques used for annotating the test set were applied to the validation set, which I personally annotated. The validation set was then reviewed by the same TNO expert to ensure uniformity across both datasets. In Appendix, Table 6.1 shows 2 examples of human-annotated in a token-level sentences with SPO labels.

κ Statistic	Strength of Agreement
< 0.00	Poor
$0.00 - 0.20$	Slight
$0.21 - 0.40$	Fair
$0.41 - 0.60$	Moderate
$0.61 - 0.80$	Substantial
$0.81 - 1.00$	Almost Perfect

Table 3.3: Classification of κ statistic values by strength of agreement (2).

3.1.4.2 Train Dataset Annotation

The training dataset should ideally be in the same nature as the validation and test datasets. However, manually annotating the training set wasn't feasible due to time limitations and the fact that the training set is about three times larger than both the test and validation sets. A study conducted by Liyanage et al. (64) highlighted GPT-4's capabilities in handling complex, domain-specific tasks like multi-label text annotation, achieving results similar to those of human annotators when given clear instructions. As a result, we decided to try annotate the training set leveraging the capabilities of LLMs. In particular, we selected prompting 'GPT-4o' over 'GPT-4' because it is more cost-effective and processes data faster, which is beneficial considering the large size of the training set.

In Subsection 3.1.4.1, we detailed the guidelines the human annotators followed for annotating the validation and test datasets. Expanding upon this, we transformed these instructions into a structured prompt to direct the annotation procedure of the model. Our goal was to achieve the accuracy of human annotators using few-shot prompting. The annotation prompt for the training dataset is shown in Figure 6.5 in Appendix. Annotation examples from the prompt are not included in this illustration to preserve space.

To further enhance the model's capability to process and accurately annotate dialogues, we implemented a dynamic chunking technique. This technique was crucial for handling complex conversational structures, particularly those involving question-answer sequences. This method identifies if the current sentence being processed is a question. If so, it groups this sentence with the following and subsequent one. The reason for including the subsequent sentence as well is based on our observation that, in many conversations, questions are often followed by intermediate sentences, with the actual answer typically appearing in the subsequent sentence. If the sentence is not a question, it is taken alone; thus, the chunk consists only of that sentence. These chunks, whether individual sentences or grouped sequences, are then fed to the model for annotation. By applying this approach, the model can accurately capture such cases and annotate them properly. It is noteworthy to refer that we used the dynamic chunking technique specifically for question-answer scenarios since our annotation instructions do not require more context beyond the sentence level for any other cases in our dataset.

Additionally, in an effort to improve the quality of the model's annotations, we also experimented with different '`temperature`' settings. Lower temperature settings are often more suitable for annotation tasks, as they tend to increase consistency while maintaining accuracy (65). Given the deterministic nature required for annotation, we avoided higher '`temperature`' values that we used in more creative tasks, like creating personas or conversations. Instead, we used lower values to make sure the model strictly followed the prompt, resulting in consistent outcomes. For example, we experimented with '`temperature`' settings of 0.2 and 0.5. Both settings handled question-answer scenarios well, but at a '`temperature`' of 0.2, the annotations were much more precise. In contrast, with a '`temperature`' of 0.5, the annotations sometimes deviated from the guidelines, producing less accurate annotations.

Finally, we set the 'max_tokens' parameter to its maximum allowable value of 4096. This decision was driven by the nature of our dataset, which includes long sentences comprising multiple tokens. Our objective was to ensure that the model could process and annotate entire sentences in a single go, without truncating any part due to token limitations.

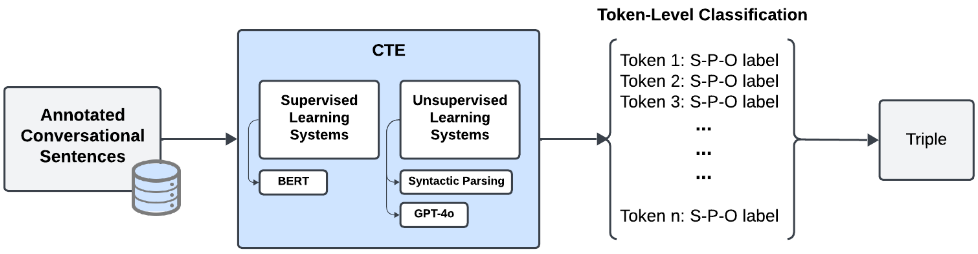


Figure 3.3: Conversational Triple Extraction Workflow

3.2 Conversational Triple Extraction

This section outlines the methods used for classifying tokens within conversational sentences into Subject, Predicate, and Object (SPO) categories and for constructing SPO triples from these classifications.

Following the creation of synthetic annotated conversations, the last objective was to extract meaningful relationships in the form of Subject-Predicate-Object (SPO) triples. However, due to the fact that generating data was a time-consuming process, time constraints prevented the testing of typical conversational triple extraction techniques applied in literature. Therefore, this study has focused on token-level SPO label classification since its evaluation is much simpler. This classification serves as an essential preliminary step for constructing these triples. It involves accurately identifying and classifying each token's role within a conversational sentence, setting the foundation for the next assembly of complete SPO triple structures. Figure 3.3 illustrates the workflow of our classification task. Annotated conversational sentences are input into the system, which processes each sentence to assign an SPO label to each token.

In this study, we examine both unsupervised and supervised systems for SPO label classification. Drawing upon the literature reviewed in Section 2.1, we experimented with both rule-based and transformer-based techniques, as we noted that no previous studies have applied these advanced approaches to synthetic conversations concerning the management of Type 2 Diabetes. For the rule-based approach we applied Syntactic Parsing, while on the transformer side, we used two advanced methods: fine-tuning the BERT model, and implementing prompt-based learning with GPT-4o. These approaches were selected for their state-of-the-art performance in triple extraction, allowing us to thoroughly evaluate their effectiveness in our task.

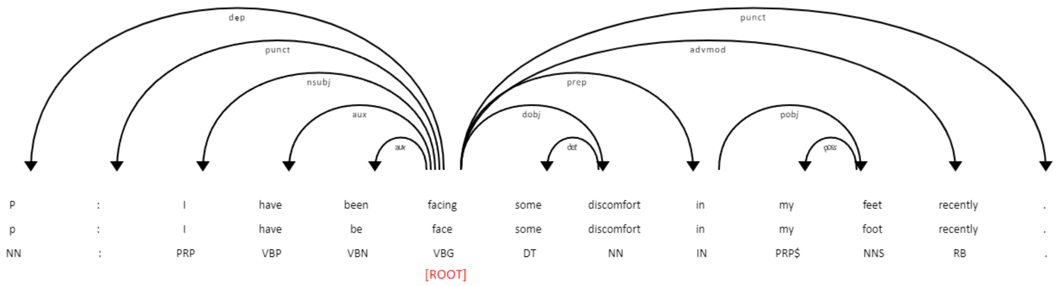


Figure 3.4: The parse tree for a sentence of our dataset.

Rule-Based Syntactic Parsing Approach

One of the unsupervised systems employed in this study involved the use of a rule-based approach that included Syntactic Parsing, or else Syntactic Analysis, to classify tokens within conversational sentences into SPO categories. This approach was directly applied on the test dataset as it does not require any training. Syntactic Parsing, implemented through spaCy's advanced NLP capabilities, analyzes the grammatical structure of sentences to clarify the relationships between words by identifying the head of each word and determining how each word depends on or relates to others. This approach creates a parse tree or a syntactic structure that outlines the hierarchical connections between words, which is key to identifying their syntactic roles. There are two main types of Syntactic Parsing: Constituency Parsing and Dependency Parsing. We utilized Dependency Parsing, which focuses on the direct relationships between words by representing these links with directed links, unlike Constituency Parsing, which organizes words into a tree structure based on their syntactic components.

For example, the parse tree shown in Figure 3.4, illustrates the grammatical structure of a typical sentence from our dataset, highlighting how words are interconnected syntactically. The parsing process starts with tokenizing the sentence, followed by lemmatization, applying Part-of-Speech (POS) tagging to each token, and identifying the dependency relationships among them. Subsequently, the approach identifies the syntactic root of the sentence, the token whose head is itself (typically the main verb or action), and label it as 'Predicate'. Then, tokens connected to the dependency labels 'nsubj' (nominal subject) or 'nsubjpass' (passive nominal subject), along with the tokens in their subtrees, are labeled as 'Subject'. These labels indicate that the tokens function as the subjects of the sentence, either as the one performing the action in active sentences ('nsubj') or the entity receiving the action in passive sentences ('nsubjpass'). Tokens linked to 'aux' (auxiliary verb), 'auxpass' (passive auxiliary verb), and 'neg' (negation modifier) are labeled as 'Predicate', as they play a role in expressing the tense, mood, voice, or negation of the main verb in a clause. Meanwhile, tokens associated with 'dobj' (direct object), 'pobj' (prepositional object), or 'attr' (attribute), along with those in their subtrees, are classified as 'Object'. These labels reflect that the tokens typically represent the recipient of the action or describe the attributes of the subject. Any tokens that do not fall into these categories are labeled as 'other'.

GPT-4o

Another unsupervised method employed in this study was a transformer-based approach, specifically employing prompt-based learning with GPT-4o. This approach was also discussed in Subsubection 3.1.4.2 for annotating the training dataset. We decided to use the exact same prompt we used for annotating the training dataset, and apply the same dynamic chunking technique to be able to capture special cases like question answer pairs. Also, both '`temperature`' and '`max_tokens`' parameters were kept consistent with the previous settings, 0.2 and 4096 respectively. This approach allowed us to directly evaluate its performance on the test dataset.

BERT

The supervised system that we used consist solely of a transformer-based approach, the fine-tuning of BERT model. As already discussed in Subsection 2.2.1, fine-tuning involves adapting the parameters of a pre-trained LLM, such as BERT, to specialize in a specific task. In Figure 3.5, the system workflow of BERT for token classification is clearly shown, illustrating the process that aligns with our specific task.

The process starts by processing each dataset (training, validation, and testing) to align to the specific format required for inputting sequences of words into the model. The '`BertTokenizerFast`' was used, which utilizes the WordPiece Tokenization Algorithm to break down each sentence into several tokens. This tokenizer is part of the Hugging Face Transformers library and is a faster implementation of the original '`BertTokenizer`'. Additionally, BERT introduces a set of special tokens during tokenization that act as linguistic markers. These tokens include `[PAD]`, which standardizes the lengths of various input sequences; `[UNK]`, assigned to words not in BERT's vocabulary; `[CLS]`, placed at the beginning of the sequence for classification purposes; and `[SEP]`, used to seperate sequences.

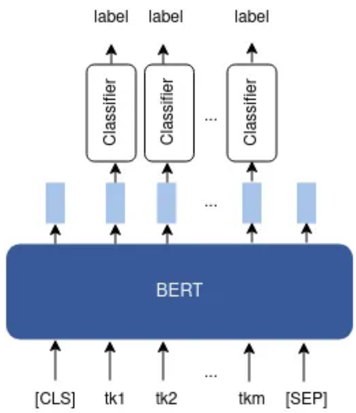


Figure 3.5: Fine-tuning BERT for Token Classification (1).

After tokenization, each token was aligned with its respective SPO label from the dataset. Because deep learning models like BERT, do not understand text data directly, we converted the SPO labels into numerical indices: [‘Subject’: 0, ‘Predicate’: 1, ‘Object’: 2, ‘other’: 3].

However, the WordPiece Tokenization Algorithm is a subword tokenization method that splits words into several smaller units or subwords. For example, from our dataset, the word ‘hypoglycemia’ is segmented into the subwords [‘h’, ‘##yp’, ‘##og’, ‘##ly’, ‘##ce’, ‘##mia’]. In that way the model learns effectively the relationships between the words or subwords. If ‘hypoglycemia’ is labeled as ‘Object’, then each subword would initially align with this label: [‘Object’, ‘Object’, ‘Object’, ‘Object’, ‘Object’, ‘Object’]. However, this approach would end up introducing more labels of type ‘Object’, potentially disturbing the balance of our dataset. Therefore, based on other studies (66, 67), we decided to assign the actual label of the word only in the first subtoken, while following subtokens received a dummy label (-100). This dummy label is also assigned to the special tokens created during tokenization. This approach is advantageous because, during training, the loss function calculation ignores these dummy labels, enabling more accurate model training by avoiding the impact of skewed label distribution.

Subsequently, a sliding window technique was then applied, with a maximum sequence length of 512 and a stride of 300. This approach ensures the model is trained on sequences that retain the full context, effectively capturing scenarios that require multiple utterances in conversations, such as question-answer pairs, rather than individual conversational sentences. The window length of 512 was chosen because it represents the maximum number of tokens that BERT can process in a single pass. A stride of 300 was selected to create overlap between consecutive windows, preserving contextual information at the edges of each sequence. Furthermore, the stride length of 300 corresponds to the average number of tokens found in conversations within our dataset. By setting the stride to this value, we ensure that each window typically includes at least one complete conversation, allowing BERT to train on well-defined conversational units. We also made sure that all sequences in a batch were padded to the maximum sequence length for that batch, aligning the labels accordingly.

Building on the initial data preparation and token alignment, we fine-tuned BERT using the `bert-base-uncased` model from the Hugging Face Transformers library, which is specifically designed for token-level classification tasks like SPO labeling. During training, we applied gradient clipping to stabilize the process and prevent exploding gradients. The loss is calculated by comparing the model's predictions with the actual labels, and the optimizer adjusts the model's parameters to minimize this loss. For this task, we used the `AdamW` optimizer, known for its ability to handle sparse gradients and its use of weight decay to prevent overfitting. Additionally, we employed a learning rate scheduler that dynamically adjusts the learning rate, improving the model's adaptability and learning efficiency over multiple epochs.

Hyperparameter tuning was conducted using Optuna on the validation dataset to identify the optimal settings. We explored a range of learning rates (from 1e-5 to 1e-3) and batch sizes (1, 8, 16) across 10 trials to find the hyperparameters that maximize the macro F1-score. The model was trained for each setting for up to 10 epochs, incorporating an early stopping mechanism that stops training if no improvement in the validation macro F1-score was observed after two epochs (`patience of 2`). This approach not only prevents overfitting but also ensures computational efficiency.

The highest score, 0.7563, was achieved in Trial 8, with a learning rate of 2.209e-05 and a batch size of 1. This combination produced the best validation performance, suggesting that a smaller batch size and a moderate learning rate worked most effectively for this task. On the other hand, Trial 2 and Trial 5 resulted in the lowest F1-scores, both at 0.1607. These trials had much higher learning rates (0.0009984 and 0.0006473, respectively), which likely caused unstable training and poor model performance. This is reflected by the early stopping mechanism being triggered after just three epochs. The best-performing model was then saved for later evaluation on the test dataset. A detailed Table 6.2 of all trials and scores of different hyperparameters during the Optuna study is depicted in Appendix.

3.2.1 Construction of SPO Triples from Token-Level Classifications

Following the classification of tokens within conversational sentences into SPO categories, we progressed to the construction of SPO triples to enhance our analysis and gain deeper insights into the relational structures. Specifically, for each conversational sentence, we aggregated all tokens that shared the same label (Subject, Predicate, or Object) into cohesive components. These categorized tokens were then systematically assembled into structured triples, each comprising the corresponding SPO components. This method allowed for a structured representation of the underlying semantic relationships within the conversational data.

3.2.2 Systems Evaluation

The performance of all systems in the token-level SPO label classification task is evaluated by comparing their predicted labels with the gold standard labels in the test dataset. Performance is assessed using standard metrics, specifically the F1-score, precision, and recall, with predictions visualized through a confusion matrix.

The following equations outline the formulas for these measurements. In these equations, TP, TN, FP, and FN represent True Positive, True Negative, False Positive, and False Negative, respectively.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Precision measures the proportion of correctly identified instances for each category (Subject, Predicate, Object, and other) out of all instances predicted as that category. Similarly, recall quantifies the proportion of actual instances in each category that are correctly classified. The F1-score, which is the harmonic mean of precision and recall, combines these two metrics to offer a balanced assessment of a model's accuracy in identifying each label in our task (68). Furthermore, for this project, the macro-average approach was taken into account for all metrics. This approach calculates the metric (such as precision, recall, or F1-score) independently for each class and then takes the average of these values. This is especially useful for datasets with potential class imbalances, as it ensures that no single class disproportionately affects the overall performance evaluation (69). In that way, a clearer and more balanced view of model performance across different categories is provided.

In the triple evaluation process, each predicted triple from a conversational sentence is compared against the gold standard triple. Matches are categorized as full matches if the entire triple is correctly predicted, partial matches if any of the individual components (Subject, Predicate, or Object) match, and component-level matches. The performance of these predictions is measured using the standard metrics, namely precision, recall, and F1-score.

4

Results

This Chapter presents and discusses the performance results of both supervised and unsupervised systems discussed in Section 3.2 of Chapter 3, as well as offers insights and an error analysis of the systems' results.

4.1 Label Distribution

Figure 4.1: Distribution of SPO Labels Across Datasets.

As shown in Figure 4.1, each dataset shows a significant imbalance, with the 'other' category being the most frequent across all datasets. This could potentially bias the model to predict 'other' more often, which might impact its performance on less frequent labels. Despite the imbalance, the proportion of each label category remains relatively consistent across all three datasets. This consistency is crucial for model generalization, as it ensures the model is trained, validated, and tested under similar label conditions. It also noteworthy to refer, given the imbalance, precision and recall may vary significantly across categories. For labels with fewer instances, like the label 'Subject', even a small number of misclassifications can cause a notable drop in performance metrics.

4.2 Systems Performance

System Type	Model	Precision	Recall	F1-score
Unsupervised	Rule-based Syntactic Parsing	0.5447	0.5694	0.5502
Unsupervised	GPT-4o	0.6833	0.7265	0.6801
Supervised	BERT	0.6770	0.7213	0.6788

Table 4.1: Systems Performance

Table 4.1 showcases the performance metrics of all three experimented systems for SPO label classification at the token level within conversational sentences. The unsupervised rule-based syntactic parsing approach shows the lowest effectiveness, with a macro precision at 0.3380, a macro recall at 0.3139, and a macro F1-score of 0.3208, indicating a limited ability to accurately identify SPO structures without supervised learning guidance. In contrast, the GPT-4o model performs significantly better, with a macro precision of 0.6833, a macro recall of 0.7265, and a macro F1-score of 0.6801. This improvement suggests that advanced language models like GPT-4o, through prompt-based learning, can effectively capture the patterns needed for SPO classification even without explicit supervision. Meanwhile, the supervised BERT model delivers similar performance to GPT-4o, with a macro precision at 0.6770, a macro recall at 0.7213, and a macro F1-score of 0.6788. These results demonstrate that GPT-4o performed the best among the models, with BERT's results closely following, suggesting that the difference between them is not significant.

To gain additional insights into the performance of the systems, particularly in terms of the triples formed from token-level classifications as discussed in Subsection 3.2.1, we can refer to Table 4.2. Syntactic Parsing appears to be the least effective of the three models in accurately identifying structured relationships within the data. Although it performs moderately well in capturing partial matches, with an F1-score of 0.3544, it struggles significantly with full matches (F1-score of 0.0010) and Object matches (F1-score of 0.0161), indicating a severe limitation in accurately identifying complete and correct

Model	Match Type	F1-score
Syntactic Parsing		
	Full Match	0.0010
	Partial Match	0.3544
	Subject Match	0.2691
	Predicate Match	0.1556
	Object Match	0.0161
GPT-4o		
	Full Match	0.3594
	Partial Match	0.6426
	Subject Match	0.5863
	Predicate Match	0.4307
	Object Match	0.4347
BERT		
	Full Match	0.1406
	Partial Match	0.5519
	Subject Match	0.4666
	Predicate Match	0.2524
	Object Match	0.2098

Table 4.2: Evaluation of Triple Matching for Different Models

triples and in correctly classifying Objects within the triples. In contrast, GPT-4o is the strongest model in terms of overall F1-scores across all match types. It performs best in partial matches (F1-score of 0.6426) and also shows strong performance in Subject and Object matches (F1-scores of 0.5863 and 0.4347, respectively), demonstrating a robust ability to identify components of the triples even if the entire triple is not perfectly formed. While not poor, its lowest performance is in predicate matches (F1-score of 0.4307), but it's still relatively strong compared to the other models. BERT is generally effective but not as strong as GPT-4o. It shows good performance in partial matches (F1-score of 0.5519) and Subject matches (F1-score of 0.4666), suggesting a competent ability to identify Subjects and at least one other element of the triple correctly. BERT is weakest in Object matches (F1-score of 0.2098) and full matches (F1-score of 0.1406), indicating challenges in fully forming correct triples and identifying Object components accurately.

The differences in models' performance across the two tables are primarily due to the inclusion of the 'other' label in the classification evaluation, which is excluded in the triple evaluation. Particularly, in Table 4.1, the models are evaluated including the 'other' label, which as already discussed in Section 4.1, is the most frequent label. So, it likely contributed significantly to the models' overall performance metrics. This is because classifying a large portion of tokens as 'other' correctly would inflate the precision, recall, and F1-scores due to the high prevalence of this label. Essentially, models get "rewarded" more for correctly predicting a high-frequency class. In contrast, during the triple evaluation, where only substantive SPO labels are considered, the exclusion of the 'other' label reveals the models' abilities to accurately identify and classify these more meaningful but less frequent labels.

Additionally, applying the macro-average method in the classification task, rather than in the triple evaluation, further explains the observed differences in performance metrics. The macro-average approach ensures equal emphasis on all labels, even the infrequent ones, whilst in triple evaluation which bypasses this averaging method, the focus is on the model's ability to precisely detect and classify specific SPO components, which results in varied performance scores.

4.3 Error Analysis

Looking at models' confusion matrices, we can derive some useful insights and understand how well our models are predicting each class (Subject, Predicate, Object, and other). The rows represent the actual (gold) labels and the columns the predicted ones, while the values within the cells show how many tokens were classified as each label combination.

Rule-Based Syntactic Parsing Approach

Looking the confusion matrix (Figure 4.2) for the Syntactic Parsing approach, the model has the highest misclassifications with 'other' predictions. Notably, 'other' is often confused with 'Object'. Also, the precision for detecting specific classes like 'Predicate' and 'Object' is notably low, as evidenced by significant numbers in off-diagonal cells where these are confused with 'other'. Finally, the label 'Object' has a comparatively better recall than other categories, indicating that while the model can identify 'Object' labels, it struggles significantly with 'Subject' and 'Predicate'.

The model's tendency to misclassify tokens as 'other' was anticipated, considering that the rule-based approach processes and classifies every sentence without discrimination. The model lacks the capability to distinguish between sentences that contribute valuable information to Knowledge Graphs (KGs) and those that do not. This limitation, which

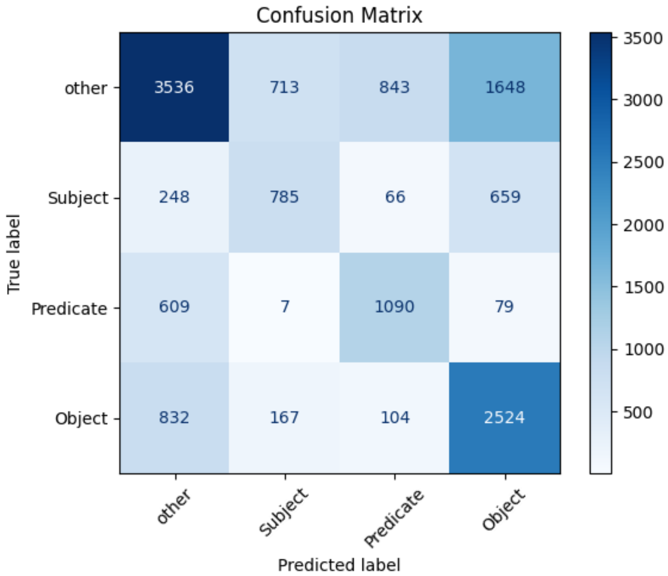


Figure 4.2: The Confusion Matrix of the Rule-Based Syntactic Parsing Approach.

stems from the rule-based approach, results in treating all sentences the same way. Consequently, the model struggles to handle specific cases like question-answer pairs discussed in Subsection 3.1.4, contributing to the significant misclassification of tokens as 'Subject', 'Predicate', and 'Object' instead of 'other'.

In the three examples below, we can see how the model classifies certain sentences within the test dataset, highlighting its consistent errors. In particular, Example 1 involves a question-answer pair within a conversation, where the answer is detailed and provides substantial information. According to the guidelines outlined in Subsection 3.1.4, only the answer should be annotated. However, the model incorrectly annotates both the question and the answer. Furthermore, examples 2 and 3 show instances where the model mistakenly annotates sentences that do not contain significant information relevant to enriching a KG. These errors are just a few examples that contribute to the misclassifications mentioned earlier.

Example 1

other	Pred	Subj	Pred	Obj	Obj	Obj	other	other
How	do	I	maintain	a	balanced	diet	then	?

other	other	Subj	Subj	Obj	Pred	other	Pred	Obj	Obj	other	other	other
A	:	Fasting	during	Ramadan	can	indeed	be	a	challenge	,	Abdullah	.

Example 2

other	other	Subj	Pred	Pred	other	Obj	other
P	:	I	will	think	about	that	.

Example 3

other	other	Subj	Pred	Obj	Obj	Obj	other
A	:	That	is	a	good	start	.

BERT

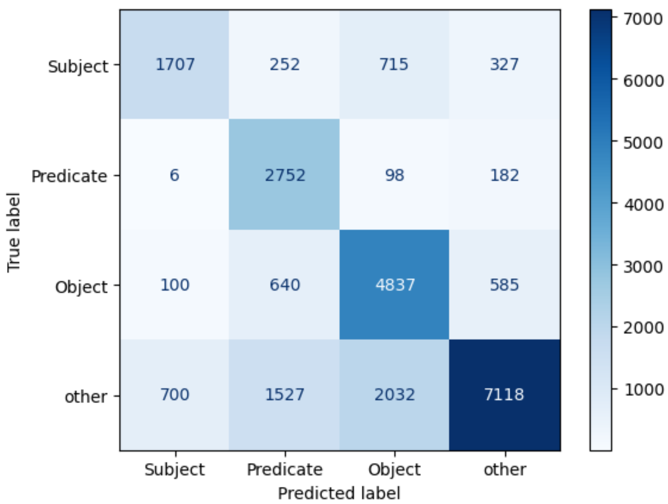


Figure 4.3: The Confusion Matrix of BERT.

Unlike the Syntactic Parsing approach, the confusion matrix for BERT (Figure 4.3) shows that its misclassifications are more balanced across different classes. BERT demonstrates high precision and recall for the 'Predicate' category, accurately identifying most 'Predicate' tokens with minimal misclassification into other categories. The 'Object' category also performs well, especially in recall, as most 'Object' tokens are correctly classified. Although some misclassifications occur, they are relatively minor compared to other categories, highlighting BERT's strong capability to recognize objects. While BERT performs adequately with 'Subject' tokens, there is noticeable confusion with 'Object' and 'other', suggesting difficulty in distinguishing subjects from other sentence elements, particularly objects. The classification of the 'other' category yields mixed results. Although many tokens are correctly identified, a substantial number are misclassified, indicating areas where the model could benefit from additional refinement.

In the three examples below, we can see how BERT classifies in a token-level the same sentences previously analyzed using the Syntactic Parsing approach. As noted earlier, the misclassifications within the 'other' category are inconsistent. BERT misclassifies Examples 1 and 3, yet performs well in Example 2. Additionally, in Example 2, BERT has difficulty correctly classifying the entire subordinate clause "Fasting during Ramadan" as 'Subject', reflecting the lower recall in the 'Subject' category, which is further evident in its confusion matrix.

Example 1

other	other	Subj	Pred	Pred	Pred	Obj	other	other
How	do	I	maintain	a	balanced	diet	then	?

other	other	other	other	other	other	Pred	Pred	Obj	Obj	other	other	other
A	:	Fasting	during	Ramadan	can	indeed	be	a	challenge	,	Abdullah	.

Example 2

other	other	other	other	other	other	other	other
P	:	I	will	think	about	that	.

Example 3

other	other	Subj	Pred	Obj	Obj	Obj	other
A	:	That	is	a	good	start	.

GPT-4o

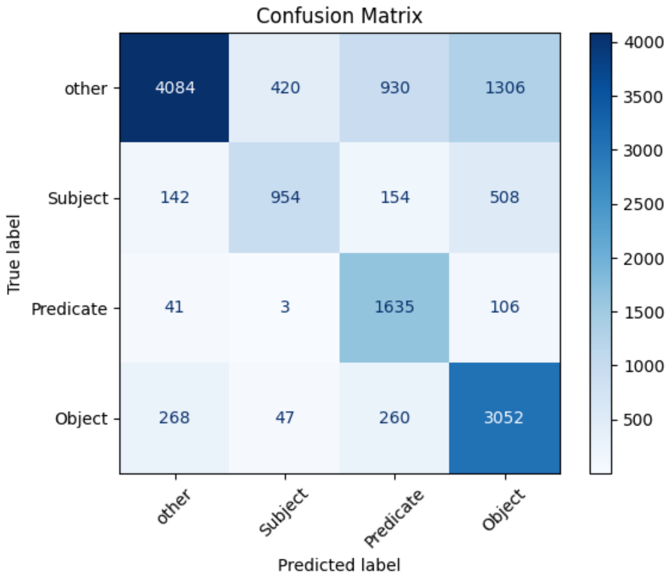


Figure 4.4: The Confusion Matrix of GPT-4o.

The confusion matrix of GPT-4o (Figure 4.4) shows strong diagonal dominance, indicating high accuracy in class predictions across all labels. The rates of misclassification are the lowest among the three models, especially for 'other', which is a common misclassification target in the other two models. GPT-4o shows also strong recall for the 'Object' category, correctly identifying the majority of 'Object' tokens. This suggests that the model effectively recognizes objects within sentences, likely due to distinct features it has learned to identify. However, despite better overall performance, GPT-4o still faces challenges with precision in the 'Subject' and 'Predicate' categories, with a significant number of tokens intended as 'Subject' or 'Predicate' being incorrectly classified as 'other'.

The three examples below illustrate that GPT-4o accurately classified each example, showing its capability to effectively follow the annotation instructions and identify special cases within the conversations.

Example 1

other	other	other	other	other	other	other	other	other
How	do	I	maintain	a	balanced	diet	then	?

other	other	Subj	Subj	Subj	Pred	Pred	Pred	Obj	Obj	other	other	other
A	:	Fasting	during	Ramadan	can	indeed	be	a	challenge	,	Abdullah	.

Example 2

other	other	other	other	other	other	other	other
P	:	I	will	think	about	that	.

Example 3

other	other	other	other	other	other	other	other
A	:	That	is	a	good	start	.

5

Discussion

Following the findings of the previous chapter, this chapter will discuss the limitations of the current study, as well as ways for improvement as a future work.

5.1 Limitations

As discussed in Section 1.4, this study focuses on conversational triple extraction using synthetic conversations related to Type 2 Diabetes management, with the goal of enabling its use in natural ones. However, various limitations may have an impact on the study's generalizability and efficacy when applied to real-world conversations. Firstly, the synthetic conversations generated for this study consist solely of well-formed, elaborate sentences from the user's (patient) side, that may not fully represent the spontaneous and often fragmented structure of real human interactions. In real-world settings, we would expect to see spelling errors, informal language, and sometimes grammatical mistakes, all of which are typical in natural dialogues but absent from our synthetic ones. This may limit the models' ability to handle the wide range of conversational nuances found in real-world conversations, potentially decreasing its usefulness in practical applications.

Secondly, considering that the CHIP collaboration targets patients in the Netherlands, it would have been more suitable to work with Dutch-language conversations to better reflect the desired user base. However, in this study, we chose the English language for the generated conversations rather Dutch, due to the greater availability of resources and tools for English natural language processing (NLP) tasks, as well as my lack of proficiency in Dutch. This language choice, although practical, may limit the applicability of the findings to Dutch-speaking populations and may necessitate additional adaptations for effective use in Dutch healthcare contexts.

Additionally, as discussed in Subsubsection 3.1.4.2, in contrast with the manually annotated validation and test datasets, the training dataset was annotated by a large language model (LLM) due to its large size and the time constraints of this study. The study's reliance on LLM-annotated data for fine-tuning the BERT model introduces a dependency on the quality and accuracy of the language model outputs, which may carry inherent biases or inaccuracies, thereby affecting the performance and reliability of the model's predictions. Although one TNO expert and I reviewed a subset of the training set, the large size of the dataset (approximately 3000 sentences) and the tight schedule of the current study prevented a more thorough review.

Lastly, a notable limitation is the focus on token-level SPO label classification rather than a comprehensive SPO triple extraction. By not directly extracting SPO triples from conversations, we might have missed opportunities to develop models that more accurately identify and interpret the interconnected structures of Subjects, Predicates, and Objects as they naturally occur in dialogues, enriching the KG of the CHIP project.

5.2 Future Work

The future work is based on the findings from the results discussed in Chapter 4 and the limitations in Section 5.1. As revealed by the error analysis in Section 4.3, there is a consistent misclassification across all models with the label 'other'. Given these misclassifications and the class imbalance present in all datasets, it might be useful to investigate what would be the systems' performance if we excluded the label 'other' from the training of BERT and from the evaluation process across all models. This exclusion could potentially offer a clearer comparison between the systems' performance as detailed in Table 4.1 and the triple evaluation results in Table 4.2 from Section 4.2. This kind of investigation could clarify how effectively the models identify SPO elements in conversational sentences, offering a more accurate measure of their actual performance.

Furthermore, as mentioned in Section 5.1, the annotation quality of the training dataset could be misleading. Despite the initial review efforts, it might help to improve the annotation quality of the dataset by incorporating sampling checking by human annotators. This approach would make the annotations more accurate and reliable, which could enhance BERT's ability to classify token-level conversational sentences with SPO labels.

While in this study the annotation schema is involving capturing special cases and question-answer pairs in conversations, incorporating methods to handle linguistic phenomena like anaphora and ellipsis which are prevalent in real conversations, could significantly improve the model's grasp of conversational context. Techniques such as coreference resolution, which link different references to the same entity within a text, would help the models maintain continuity and coherence over longer stretches of dialogue, potentially leading to deeper and more accurate interpretations of conversations. In addition to that, incorporating BIO labeling alongside SPO labeling in the annotation schema could improve the model's accuracy in identifying token roles. The BIO format, which marks the beginning, inside, and outside of entity spans, is especially helpful in complex sentences with multiple interacting entities. This combined labeling approach could clarify ambiguities from overlapping or embedded entities, enhancing the model's ability to process conversational data effectively.

Future research could also explore full SPO triple extraction by using existing tools like REBEL or KnowGL, which are specifically designed for relational triple extraction from text. These methods could help shift from token-level labeling to end-to-end triple extraction, enabling a more complete representation of conversational semantics.

6

Conclusion

The thesis aimed to determine the effectiveness of conversational triple extraction (CTE) systems on synthetic data related to Type 2 Diabetes management. However, due to time constraints, the scope was limited to token-level classification of synthetic conversational sentences into Subject, Predicate, and Object (SPO) categories, an essential preliminary step in building the SPO triples. The primary objectives were to investigate the necessary properties that should be included to constructing effective and realistic conversations related to Type 2 Diabetes management, to explore how generative methods can be used to create and annotate these conversations, and lastly, find which method is most effective in identifying and extracting triples from the generated conversations.

Initially, synthetic personas were created to reflect real-life Type 2 Diabetes patients. Then, all the essential properties for constructing effective and realistic conversations were identified and detailed in the Methodology chapter in Table 3.2. Generative methods were explored and successfully applied to create and annotate with SPO labels, multi-turn conversations simulating real-life interactions between these personas and a caretaker agent. Both supervised and unsupervised systems were evaluated. Prompt-based learning using GPT-4o achieved the highest performance being the most effective method for our task, with BERT showing competitive results. The comparative analysis of these systems offered key insights into their strengths and limitations, providing direction for future applications and improvements in this field.

Overall, this thesis adds valuable knowledge to the intersection of artificial intelligence (AI) and healthcare by demonstrating the potential of CTE systems on synthetic data related to Type 2 Diabetes management. Future research should consider integrating more sophisticated natural language processing tasks such as anaphora and ellipsis resolution, employing BIO alongside SPO labeling for greater precision, and advancing beyond token-level classification to full triple extraction and validation within conversational contexts.

References

[1] Mouna Labiadh. **Exploring BERT: Feature extraction Fine-tuning**. *Medium*, Feb 13, 2024. v, 33

[2] JR Landis. **The Measurement of Observer Agreement for Categorical Data**. *Biometrics*, 1977. vi, 27

[3] Kimberly A Kline and Dawn ME Bowdish. **Infection in an aging population**. *Current Opinion in Microbiology*, **29**:63–67, 2016. 1

[4] Maryam Changizi and Mohammad H. Kaveh. **Effectiveness of the mHealth technology in improvement of healthy behaviors in an elderly population—a systematic review**. *mHealth*, **3**(11), 2017. 1, 2

[5] Kaare Christensen, Gabriele Doblhammer, Roland Rau, and James W Vaupel. **Ageing populations: the challenges ahead**. *The Lancet*, **374**(9696):1196–1208, 2009. 1

[6] Garry Egger and John Dixon. **Beyond Obesity and Lifestyle: A Review of 21st Century Chronic Disease Determinants**. *BioMed research international*, **2014**:731685, 04 2014. 1

[7] Ahmed H. Abdelhafiz and Alan J. Sinclair. **Management of Type 2 Diabetes in Older People**. *Diabetes Therapy*, **4**(1):13–26, 2013. 1

[8] H. Sun, P. Saeedi, S. Karuranga, M. Pinkepank, K. Ogurtsova, B. B. Duncan, C. Stein, A. Basit, J. C. Chan, J. C. Mbanya, et al. **IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045**. *Diabetes Research and Clinical Practice*, **183**:109119, 2022. 1

[9] Rachelle Meisters, Daan Westra, Polina Putrik, Hans Bosma, Dirk Ruwaard, and Maria Jansen. **Regional differences in healthcare costs further explained: The contribution of health, lifestyle, loneliness and mastery**. *TSG - Tijdschrift voor gezondheidswetenschappen*, **100**(4):189–196, 2022. 2

[10] Margaret A Powers, Joan Bardsley, Marjorie Cypress, Paulina Duker, Martha M Funnell, Amy Hess Fischl, Melinda D Maryniuk, Linda Siminerio, and Eva Vivian. **Diabetes Self-Management Education and Support in Type 2 Diabetes: A Joint Position Statement of the American Diabetes Association, the American Association of Diabetes Educators, and the Academy of Nutrition and Dietetics**. *Journal of the Academy of Nutrition and Dietetics*, **115**(8):1323–1334, 2015. Epub 2015 Jun 5. 2

[11] J Okemah, J Peng, and M Quiñones. **Addressing Clinical Inertia in Type 2 Diabetes Mellitus: A Review**. *Adv Ther*, **35**(11):1735–1745, 2018. Epub 2018 Oct 29. 2

[12] E H Wagner, B T Austin, and M Von Korff. **Organizing care for patients with chronic illness**. *The Milbank Quarterly*, **74**(4):511–544, 1996. 2

[13] Ivan Contreras and Josep Vehi. **Artificial Intelligence for Diabetes Management and Decision Support: Literature Review**. *J Med Internet Res*, **20**(5):e10775, May 2018. 2

[14] Zeynep Akata, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, Hayley Hung, Catholijn Jonker, Cristof Monz, Mark Neerincx, Frans Oliehoek, Henry Prakken, Stefan Schlobach, Linda van der Gaag, Frank van Harmelen, Herke van Hoof, Birna van Riemsdijk, Aimee van Wynsberghe, Rineke Verbrugge, Bart Verheij, Piek Vossen, and Max Welling. **A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence**. *Computer (New York)*, **53**(8):18–28, August 2020. 2, 3

[15] Silje Stangeland Lie, Bjørg Karlsen, Ellen Renate Oord, Marit Graue, and Bjørg Oftedal. **Dropout From an eHealth Intervention for Adults**

With Type 2 Diabetes: A Qualitative Study. *Journal of Medical Internet Research*, **19**(5):e187, 2017. 2

[16] Ine D'Haeseleer., Karsten Gielis., and Vero Vanden Abeele. **Human-centred Design of Self-management Health Systems with and for Older Adults: Challenges and Practical Guidelines**. In *Proceedings of the 7th International Conference on Information and Communication Technologies for Ageing Well and e-Health - ICT4AWE*, pages 90–102. INSTICC, SciTePress, 2021. 2

[17] Pei-Yu Chen, Selene Baez Santamaria, Maaike H.T. de Boer, Floris den Hengst, Bart A. Kamphorst, Quirine Smit, Shihan Wang, and Johanna Wolff. *Intelligent Support Systems for Lifestyle Change: Integrating Dialogue, Information Extraction, and Reasoning*, pages 457–459. Frontiers in Artificial Intelligence and Applications. IOS Press, Netherlands, June 2024. 3, 25

[18] Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. **A Novel Cascade Binary Tagging Framework for Relational Triple Extraction**. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1476–1488, Online, July 2020. Association for Computational Linguistics. 4

[19] Roos M Bakker and Daan L Di Scala. **From Text to Knowledge Graph: Comparing Relation Extraction Methods in a Practical Context**. In *First International Workshop on Generative Neuro-Symbolic AI, co-located with ESWC*, 2024. 4, 7

[20] Kimia Hassanzadeh, Marek Reformat, Witold Pedrycz, Iqbal Jamal, and John Berezowski. **T2r: System for converting textual documents into rdf triples**. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, **3**, pages 221–228. IEEE, 2013. 6

[21] Kalgi Gandhi and Nidhi Madia. **Information extraction from unstructured data using RDF**. In *2016 International Conference on ICT in Business Industry & Government (ICTBIG)*, pages 1–6. IEEE, 2016. 6

[22] Jing Jiang. *Information Extraction from Text*, pages 11–41. Springer US, Boston, MA, 2012. 6

[23] Junbo Fei, Weixin Zeng, Xiang Zhao, Xuanyi Li, and Weidong Xiao. **Few-shot relational triple extraction with perspective transfer network**. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 488–498, 2022. 6

[24] Yujia Zhang, Tyler Sadler, Mohammad Reza Taesiri, Wenjie Xu, and Marek Reformat. **Fine-tuning Language Models for Triple Extraction with Data Augmentation**. In Russa Biswas, Lucie-Aimée Kaffee, Oshin Agarwal, Pasquale Minervini, Sameer Singh, and Gerard de Melo, editors, *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pages 116–124, Bangkok, Thailand, August 2024. Association for Computational Linguistics. 6, 8

[25] Yu-Ming Shang, Heyan Huang, Xin Sun, Wei Wei, and Xian-Ling Mao. **Relational triple extraction: one step is enough**. *arXiv preprint arXiv:2205.05270*, 2022. 7

[26] Shaun D'Souza. **Parser extraction of triples in unstructured text**. *arXiv preprint arXiv:1811.05768*, 2018. 7

[27] Stephen P. Gardner. **Ontologies and semantic data integration**. *Drug Discovery Today*, **10**(14):1001–1007, 2005. 7

[28] Junbo Fei, Weixin Zeng, Xiang Zhao, Xuanyi Li, and Weidong Xiao. **Few-Shot Relational Triple Extraction with Perspective Transfer Network**. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 488–498, New York, NY, USA, 2022. Association for Computing Machinery. 7

[29] Zepeng Ding, Wenhao Huang, Jiaqing Liang, Deqing Yang, and Yanghua Xiao. **Improving Recall of Large Language Models: A Model Collaboration Approach for Relational Triple Extraction**, 2024. 7

[30] Somin Wadhwa, Silvio Amir, and Byron Wallace. **Revisiting Relation Extraction in the era of Large Language Models**. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada, July 2023. Association for Computational Linguistics. 8

[31] Rafal Józefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. **Exploring the Limits of Language Modeling**. *CoRR*, **abs/1602.02410**, 2016. 8

[32] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. **A Survey of Large Language Models**, 2023. 8

[33] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. **Efficient Estimation of Word Representations in Vector Space**, 2013. 8

[34] Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. **Pre-Trained Language Models and Their Applications**. *Engineering*, **25**:51–65, 2023. 8

[35] OpenAI. **GPT-4 Technical Report**, 2024. 8, 16

[36] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. **Bert: Pre-training of deep bidirectional transformers for language understanding**. In *Proceedings of naacL-HLT*, **1**, page 2. Minneapolis, Minnesota, 2019. 9

[37] Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid. **The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities**, 2024. 9

[38] Louie Giray. **Prompt Engineering with ChatGPT: A Guide for Academic Writers**. *Annals of Biomedical Engineering*, **51**:3, 06 2023. 9

[39] Jiaqi Wang, Enze Shi, Sigang Yu, Zihao Wu, Chong Ma, Haixing Dai, Qiushi Yang, Yanqing Kang, Jinru Wu, Huawen Hu, Chenxi Yue, Haiyang Zhang, Yiheng Liu, Yi Pan, Zhengliang Liu, Lichao Sun, Xiang Li, Bao Ge, Xi Jiang, Dajiang Zhu, Yixuan Yuan, Dinggang Shen, Tianming Liu, and Shu Zhang. **Prompt Engineering for Healthcare: Methodologies and Applications**, 2024. 10

[40] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. **Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing**, 2021. 10

[41] Author's Name. **Zero-Shot Learning, Few-Shot Learning, and Fine-Tuning: A Technical Walkthrough Using OpenAI's APIs & Models**. `https://labelbox.com/guides/zero-shot-learning-few-shot-learning-fine-tuning/`, 2024. Accessed: 2024-06-03. 10

[42] Author's Name. **The Few-Shot Prompting Guide**. 10

[43] Kun Xu, Han Wu, Linfeng Song, Haisong Zhang, Linqi Song, and Dong Yu. **Conversational semantic role labeling**. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29**:2465–2475, 2021. 11

[44] Guozheng Li, Zijie Xu, Ziyu Shang, Jiajun Liu, Ke Ji, and Yikai Guo. **Empirical Analysis of Dialogue Relation Extraction with Large Language Models**, 2024. 11

[45] Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. **Dialogue-based relation extraction**. *arXiv preprint arXiv:2004.08056*, 2020. 11

[46] Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. **A Comprehensive Survey on Relation Extraction: Recent Advances and New Frontiers**. *ACM Computing Surveys*, **56**(11):1–39, 2024. 11

[47] Piek Vossen, Selene Baez, Lenka Bajčetić, and Bram Kraaijeveld. **Leolani: a reference machine with a theory of mind for social communication**, 2018. 11

[48] Rafael Roque de Souza, Thiago Luna Pinheiro, Julio Cesar Barbour Oliveira, and Júlio Cesar dos Reis. **Knowledge Graphs Extracted from Medical Appointment Transcriptions: Results Generating Triples Relying on LLMs.** In *KEOD*, pages 129–139, 2023. 11

[49] Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. **Aligning Large Language Models with Human: A Survey**, 2023. 14

[50] Jessica C. G. Bak, Dick Mul, Erik H. Serné, Harold W. de Valk, Theo C. J. Sas, Petronella H. Geelhoed-Duijvestijn, Mark H. H. Kramer, Max Nieuwdorp, and Carianne L. Verheugt. **DPARD: rationale, design and initial results from the Dutch national diabetes registry**. *BMC Endocrine Disorders*, **21**(1):122, 2021. 14

[51] Centraal Bureau voor de Statistiek. **Gezondheid en zorggebruik; persoonskenmerken**. CBS StatLine, 2024. Accessed: 2024-06-02. 15

[52] Joanne K. Ujcic-Voortman, Miranda T. Schram, Monique A. Jacobs-van der Bruggen, Arnoud P. Verhoeff, and Caroline A. Baan. **Diabetes prevalence and risk factors among ethnic minorities**. *European Journal of Public Health*, **19**(5):511–515, 07 2009. 15

[53] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. **Large Language Models: A Survey**, 2024. 18

[54] Haodong Duan, Jueqi Wei, Chonghua Wang, Hongwei Liu, Yixiao Fang, Songyang Zhang, Dahua Lin, and Kai Chen. **BotChat: Evaluating LLMs' Capabilities of Having Multi-Turn Dialogues**. *arXiv preprint arXiv:2310.13650*, 2023. 18

[55] Cristina Reguera Gomez. **Building a Natural Conversational Agent for Healthcare by Examining Empathetic Language**, September 2024. 18

[56] Thuy-Trinh Nguyen, Kellie Sim, Anthony To Yiu Kuen, Ronald R. O'donnell, Suan Tee Lim, Wenru Wang, and Hoang D. Nguyen. **Designing AI-based Conversational Agent for Diabetes Care in a Multilingual Context**, 2021. 21, 22

[57] Timothy W. Bickmore and Justine Cassell. **"How about this weather?" Social Dialogue with Embodied Conversational Agents**. 2000. 21

[58] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. **What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents**. In *Proceedings of the 2019 CHI Conference on Human Factors in*

Computing Systems, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery. 21

[59] Bingjie Liu and S. Shyam Sundar. **Should Machines Express Sympathy and Empathy? Experiments with a Health Advice Chatbot**. *Cyberpsychology, behavior and social networking*, **21 10**:625–636, 2018. 21

[60] Marilyn A. Walker, Janet E. Cahn, and Stephen J. Whittaker. **Improvising Linguistic Style: Social and Affective Bases for Agent Personality**, 1997. 21, 22

[61] Martin Adam and Johannes Klumpe. **Onboarding with a chat–the effects of message interactivity and platform self-disclosure on user disclosure propensity**. 2019. 22

[62] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. **A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT**, 2023. 22, 23

[63] Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. **The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation**. In Dongyan Zhao, editor, *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico, August 2018. Association for Computational Linguistics. 25

[64] Chandreen R Liyanage, Ravi Gokani, and Vijay Mago. **GPT-4 as an X data annotator: Unraveling its performance on a stance classification task**. *PloS one*, **19**(8):e0307741, 2024. 27

[65] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. **ChatGPT outperforms crowd workers for text-annotation tasks**. *Proceedings of the National Academy of Sciences*, **120**(30):e2305016120, 2023. 28

[66] Minghan Wang, Hao Yang, Hengchao Shang, Daimeng Wei, Jiaxin Guo, Lizhi Lei, Ying Qin, Shimin Tao, Shiliang Sun, Yimeng Chen, and Liangyou Li. **HW-TSC's Participation at WMT 2020 Quality Estimation Shared**

Task. In Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors, *Proceedings of the Fifth Conference on Machine Translation*, pages 1056–1061, Online, November 2020. Association for Computational Linguistics. 33

[67] Abir Rahali and Moulay A. Akhloufi. **End-to-End Transformer-Based Models in Textual-Based NLP**. *AI*, **4**(1):54–110, 2023. 33

[68] Mohammad Hossin and Md Nasir Sulaiman. **A review on evaluation metrics for data classification evaluations**. *International journal of data mining & knowledge management process*, **5**(2):1, 2015. 35

[69] Maria Cristina Hinojosa Lee, Johan Braet, and Johan Springael. **Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores**. *Applied Sciences*, **14**(21), 2024. 36

Appendix

```
{
    "ethnicity_prompt": "What do you know about {nationality} {ethnicity} people?",
    "disease_prompt": "What do you know about {disease} in {nationality} {ethnicity} people?",
    "profile_prompt": "
        {profile}

        The profile above is of a fictional person with {disease}. She is struggling to manage her disease.
        I would like for you to create a persona using the framework given below, based on the profile I gave you at the start.
        All characteristics mentioned in the framework should be extensively described.
        The persona will be used to create realistic conversations between a Type 2 Diabetes patient and an agent,
        that will be used to simulate patient-caretaker interactions in a patient-centered research and development process for patients with {disease}.

        1. **Personal and Demographic Background**:
           - Name, nicknames, age, gender, ethnicity, nationality, place of birth, current residence.
           - Family background, including parents, siblings, children, possibly grandchildren, and family medical history (especially chronic diseases).

        2. **Physical Characteristics and Health**:
           - Height, weight, body type, skin tone, eye color, hair color and style.
           - Distinguishing features (scars, tattoos, birthmarks), general health, disabilities, chronic illnesses.
           - Dress style, possibly adapted for medical needs, including chronic disease management (exercise, diet, sleep patterns, medications, and treatments).

        3. **Psychological and Emotional Profile**:
           - Personality traits (introvert/extrovert, optimist/pessimist, emotional stability, dominant traits).
           - Fears, phobias, aspirations, goals, motivations, personal values.
           - Coping mechanisms for chronic diseases, mental health history, current mental health status, and ongoing treatments.

        4. **Professional and Educational Background**:
           - Education, occupation, work history, impact of chronic disease on professional life.
           - Skills and abilities, including languages spoken, special talents, and professional skills.

        5. **Social and Cultural Dynamics**:
           - Social relationships, social circle, relationship status, interaction style with others.
           - Cultural influences, traditions, adaptations for chronic disease in cultural practices, communication styles, language nuances.
           - Role and availability of support from family, friends, community, and online communities.

        6. **Lifestyle and Daily Routine**:
           - Behavioral tendencies, daily routine including disease management, response to stress, decision-making style.
           - Nutritional habits, dietary patterns and preferences, restrictions or special diets.
           - Exercise and physical activity levels, limitations or adaptations due to the chronic disease.

        7. **Healthcare Interactions and Literacy**:
           - Healthcare relationship dynamics, attitudes towards healthcare providers, history with healthcare systems.
           - Health literacy, understanding of chronic disease, medical terminology, ability to follow treatment plans.
           - Emotional response to health issues, reactions to health progress or setbacks, support needs.

        8. **Economic and Environmental Factors**:
           - Financial situation, economic impact of chronic disease.
           - Living conditions (urban/rural setting, housing quality), exposure to environmental stressors.
           - Access to and affordability of healthy food options, recreational spaces, and community resources.

        9. **Communication, Decision-Making, and Legal Aspects**:
           - Preferred modes of communication, openness in communication, language used related to chronic disease.
           - Approach to health-related decisions, influences on health decisions, level of involvement in treatment planning.
           - Awareness of patient rights, healthcare laws, experience with healthcare proxies or advanced directives.

       10. **Technology, Cultural Competence, and Support Networks**:
           - Comfort with and access to digital health tools, technology use for health monitoring.
           - Cultural and linguistic competence of healthcare providers, language barriers, translation needs.
           - Social support networks, involvement in support groups or communities.

        Please synthesize a persona that aligns with the profile and information I've provided so far. Their traits should be realistic and profound.
        I don't want the traits you create to be just about living with disease, but it should taken into consideration. Not everything has to be positive; real people are fallible.
        I want you to be very comprehensive. Please start every section in the profile with a number.

        "
}
```

Figure 6.1: The prompt for creating the persona descriptions.

Example of a Persona Description

Your name is Fatima, often called "Fatouma" by your family, a 65-year-old Moroccan-Dutch female. Born in Amsterdam, Netherlands, you currently reside in Rotterdam. Your Moroccan heritage and Dutch nationality play a crucial role in defining your mixed cultural roots. Born to Moroccan immigrant parents who moved to the Netherlands for work in the 1950s, you have three siblings—two brothers and a sister. You are a mother to two daughters and a son and have five grandkids whom you adore immensely. Both of your parents had diabetes, and now you are a patient of Type 2 Diabetes yourself. Regarding your physical characteristics and health, you are 5'4" tall with a body weight that fluctuates between slightly overweight and obesity, a common issue among type 2 diabetes patients. You have olive skin, deep brown eyes, and dark hair now sprinkled with grey, which you usually keep wrapped in your favored headscarves. You have no tattoos or scars but do have a heart-shaped birthmark on your right arm. Your general health has been somewhat compromised due to your ongoing strife with Type 2 Diabetes. You dress conservatively, often opting for a hijab and long dresses, and pay special attention to comfortable shoes due to occasional diabetic foot issues. Regarding your psychological and emotional profile, you can be described as having a resilient and nurturing personality. You are generally optimistic and extroverted but often get emotional when discussing your health. Diabetes has fortified your patience as you deal with the daily routine of managing your illness. Your main fears revolve around the potential complications of your disease and the fear that you might be a burden on your family. Your ultimate goal is to manage your diabetes effectively to pursue a healthy life. You find solace and strength in your faith, your family, and your favorite pastime—gardening. You do not have any diagnosable mental health issues but do experience occasional anxiety related to your health condition. Regarding your professional and educational background, you have a foundational education, having completed high school. You worked as a cleaner for most of your life until retiring a few years ago. Diabetes has not majorly affected your professional life as you were diagnosed after retirement. Apart from Dutch, you are also fluent in Arabic and Berber, the languages you were brought up with at home. You are known for your excellent cooking skills, especially Moroccan traditional cuisine. Regarding your social and cultural dynamics, you are a widow but maintain a strong bond with your children and grandchildren. You are sociable and enjoy spending time with your neighbors and a few close friends. Your faith and Moroccan roots greatly influence your values, behaviors, and traditions, especially during festivities. However, because of *your condition, you often find yourself modifying some food-related traditions. Your children and siblings are your main source of support, and you also find comfort in your local faith community. Regarding your lifestyle and daily routine, you are methodical and prefer routines to manage your disease. You check your blood sugar regularly and are conscious about meal timings and portions. You primarily consume a Mediterranean diet with a focus on low glycemic index foods but do indulge in traditional Moroccan desserts occasionally. You walk daily around your neighborhood, adoring the canals and windmills, although neuropathy in your feet, a common complication of diabetes, sometimes limits your mobility. Regarding your healthcare interactions and literacy, you have a cordial relationship with your healthcare providers but heavily rely on your children during medical appointments. You understand the basics of your disease but struggle with some of the complex medical terminologies related to your condition. You try to remain positive through health setbacks, with your faith and family acting as strong pillars of support. Regarding your economic and environmental factors, being retired, you have a fixed income which can sometimes make handling the financial burden of your disease challenging. You live in a comfortable urban setting in Rotterdam and do not experience significant environmental stressors. Access to healthy food options and healthcare facilities is good but could be negatively impacted by your limited financial resources. Regarding your communication, decision-making, and legal aspects, you are open to candid conversations about your health but often rely on your children to make significant health-related decisions. With limited formal education, your knowledge about patient rights and healthcare laws is minimal, relying heavily on your children for understanding the legal aspects of your healthcare. Regarding technology, cultural competence, and support networks, you are not tremendously comfortable with new technology. You use a basic cell phone and rely on your children for any technical needs. Your healthcare providers are culturally competent, ensuring comfortable interaction due to their multilingual skills, which include Dutch and Arabic. In addition to your family, you find immense support in your small circle of friends and your faith community. You are not part of any online community or support group.*

```
SYSTEM_PROMPT = ("""

    Follow the guidlines below to ensure the quality and structure of the dialogues:

    Objective: Generate VARIED, realistic, and concise conversations between a Type 2 Diabetes patient (P) and an agent (A) acting as a caretaker specializing in Type 2 Diabetes management.

    Instructions:

    1. Content Focus:
       - Omit all greetings and direct introductions.
       - Begin each conversation directly with substantive dialogue related to managing Type 2 Diabetes.

    2. Conversation Length:
       - Ensure each conversation contains exactly 4 to 6 exchanges. NO LESS THAN 4 AND NO MORE THAN 6.
       - Define an exchange as one statement from patient (P) and one response from the agent (A) (a complete interaction cycle).

    3. Dialogue Structure:
       - Each exchange consists of one concise statement or response.
       - ALTERNATE THE STARTING SPEAKER:
         * In half of the conversations, let the patient (P) initiate the conversation with concerns or issues about their condition.
         * In the other half, let the agent (A) initiate the conversation to clarify or gather specific information about patient’s lifestyle or condition.

    4. Personalization:
       - Incorporate realistic details based on patient’s lifestyle and diabetes management.

    5. Consistency and Flow:
       - Maintain structured flow, ensuring conversations are focused and conclude naturally after 4 to 6 exchanges.

    6. Regular Review:
       - Periodically reassess and refine dialogue examples to ensure relevance and user engagement.

    End of instructions.

    Example Conversations:

        - Here is an example of a conversation initiated by the patient:

            P: I've noticed that my blood sugar levels fluctuate significantly on stressful days at work. Is this normal?
            A: Yes, stress can indeed cause blood sugar levels to rise. When you are stressed, your body goes into “fight-or-flight” mode, which can cause a release of hormones that raise blood glucose levels.
            P: It's tough to avoid stress with my job. Do you have any specific tips on how to handle this?
            A: It's essential to incorporate stress management techniques into your daily routine. You could try deep breathing exercises, meditation, or yoga. Also, regular exercise can help lower your blood sugar and manage stress.
            P: Sounds like a good plan. Should I also be adjusting my diet on stressful days?
            A: It's better to maintain a consistent, balanced diet rather than making daily changes. However, ensuring you stay well-hydrated and avoiding high-sugar or high-carb foods during stressful situations could be beneficial.
            P: That's helpful advice. I'll start incorporating these strategies. Thank you.
            A: You're welcome, Olivia! Remember, it's about achieving a balance. Keep monitoring your blood sugar levels and don't hesitate to reach out if you need more guidance.

        - Here is an example of a conversation initiated by the agent:

            A: Noah, have you noticed any changes in your energy levels since your last doctor's visit?
            P: Yes, actually. I've been feeling more tired than usual, even with a good night's sleep.
            A: I see. This fatigue could be a sign that your blood sugar levels aren't well controlled. Are you monitoring your blood sugar levels regularly?
            P: I do, but I must confess that it's not as often as my doctor recommended.
            A: That's understandable, but regular monitoring is key to managing your diabetes effectively. It helps you understand how your lifestyle and diet affect your blood sugar levels. I'd recommend setting a reminder for it.
            P: I'll start setting reminders. Also, could this tiredness be linked to my diet?
            A: Absolutely. Diet plays a crucial role in energy levels, especially for someone managing diabetes. Ensure your meals are well-balanced with a good mix of protein, fiber, and healthy fats. Are you eating regular meals?
            P: I try to, though I sometimes skip breakfast because of my morning rush.
            A: Skipping breakfast can definitely contribute to your tiredness. Consider preparing simple, quick options like overnight oats or yogurt with fruit the night before. This way, you can have a nutritious start to your day without extra morning hassle.
            P: That sounds manageable. I will give that a try. Thanks for the tips!

        - Here is a second example of a conversation initiated by the patient:

            P: I've been experiencing more frequent urination lately. Is this related to my Type 2 Diabetes?
            A: Yes, Julia, frequent urination can be a symptom of high blood sugar levels, which is common in Type 2 Diabetes. When your blood sugar levels are high, your kidneys try to get rid of the excess sugar by flushing it out of your body through urine.
            P: I see. Does drinking water make this worse?
            A: Actually, drinking water can aid in managing high blood sugar levels. Staying hydrated can help the kidneys flush out excess glucose through urine. However, if you notice that you are drinking excessive amounts of water or are constantly thirsty, it could be a sign of uncontrolled diabetes.
            P: Oh, I didn't know that. And is this frequent urination harmful in any way?
            A: If your blood sugar levels are consistently high and frequent urination continues over time, it can put strain on your kidneys and lead to more serious problems like kidney disease or urinary tract infections. It's important to keep your blood sugar levels well-controlled to prevent such complications.
            P: I understand. I will keep an eye on my blood sugar levels and stay hydrated. Thanks for the explanation.
            A: Absolutely, Julia! Keeping a close eye on your blood sugar and managing it well is essential for your health. If you have any further questions or need assistance, please feel free to contact me anytime. Take care!

        - Here is a second example of a conversation initiated by the agent:

            A: Julia, I know you've been following a balanced diet for managing your diabetes. Can you explain your typical daily diet in a bit more detail?
            P: Sure, I typically start with a small breakfast like a bowl of oatmeal with blueberries. For lunch, it's usually a salad with grilled chicken or turkey and for dinner, I often have a lean protein like fish, with steamed vegetables.
            A: That sounds like a balanced meal plan. It's good that you include lean proteins and vegetables. However, remember to also incorporate whole grains and legumes that are rich in fiber and can help regulate blood sugar levels.
            P: I will keep that in mind. But what about snacks? I usually snack on fruit or a handful of nuts in between meals.
            A: That's a good practice. Snacking on fruits and nuts can help avoid spikes in your blood sugar levels. Do remember to consume in moderation, especially nuts, as they are high in calories. Stay consistent with your meal times to help keep your blood sugar levels steady throughout the day.
            P: Speaking of consistency, how should I manage eating out? Sometimes it's hard to find suitable options at restaurants.
            A: When eating out, try to choose restaurants that offer whole food options and are accommodating to dietary restrictions. Opt for dishes that are steamed, grilled, or baked, and don't hesitate to request modifications like dressing on the side or substituting white rice with brown rice or vegetables.
            P: That makes sense. And how about beverages? I usually stick to water or herbal tea but occasionally have a glass of wine.
            A: Sticking mostly to water and herbal tea is excellent. If you choose to have wine, it is important to limit it to one glass, particularly with meals, to prevent any significant impact on your blood sugar levels. Also, avoid sugary drinks and sodas as they can cause rapid blood sugar spikes.
            P: Thanks for the advice. Lastly, should I be concerned about my dairy intake? I usually have a glass of milk with breakfast or use a bit of cheese in my salads.
            A: Dairy can be part of a balanced diet, but choose low-fat or non-fat options where possible. These provide calcium and protein without too much saturated fat. Yogurt, in particular, can be a good choice if it's low in sugar, as it also offers probiotics which are beneficial for your digestive health.

        - Here is a third example of a conversation initiated by the patient:

            P: I've been struggling with some anxiety lately due to my health. Does it impact my diabetes in any way?
            A: Yes, Jan, anxiety can impact your blood sugar levels. When you're anxious, your body releases stress hormones, which can cause your blood sugar levels to rise. It's important to manage anxiety to keep your diabetes under control.
            P: What would you suggest to manage this anxiety? I already try meditation but it seems insufficient.
            A: Incorporating physical activities, like your regular walks or cycling, can help reduce anxiety. You could also try other relaxation techniques like progressive muscle relaxation or guided imagery. Additionally, talking about your concerns with someone you trust can also help.
            P: My family is supportive but I worry about burdening them with my health concerns.
            A: That's a common concern, Jan. You might find it helpful to talk to a professional counselor or therapist who specializes in chronic illnesses. They can provide coping strategies and emotional support in a structured, understanding environment.
            P: I will consider this. Does management of anxiety also involve changes in my medication or diet?
            A: It doesn't usually require changes in medication unless recommended by your doctor. As for your diet, maintaining a balanced diet is beneficial. Some foods like those rich in magnesium and Omega-3 fatty acids are known to help reduce anxiety.
            P: Thank you. I'll try to incorporate these suggestions into my lifestyle.

        - Here is a third example of a conversation initiated by the agent:

            A: Hey Jan, I see that you're a fan of cycling. Has the Dutch weather been favorable for it lately?
            P: It's been quite unpredictable recently, so I haven't been out as often as I'd like. How does this affect my diabetes management?
            A: Physical activity, like cycling, is crucial in managing your diabetes. It helps your body use insulin better and can lower your blood sugar levels. If weather conditions are interfering with your routine, perhaps you could look for indoor exercises to stay active.
            P: Any suggestions for indoor exercises that could suit me?
            A: Definitely, Jan. Indoor walking or marching in place can be an excellent start. If possible, you could consider investing in a stationary exercise bike, which would be similar to your outdoor cycling. Strength training exercises using light weights or resistance bands are also beneficial.
            P: I'll look into these options. I suppose sticking to my daily walks is also helpful?
            A: Yes, absolutely! Regular walking is an efficient and effective way to improve insulin sensitivity and manage your diabetes. Just make sure to adapt the intensity and duration to your abilities.
            P: I got it. Thanks for your advice. I will try to keep active, rain or shine.
            A: Glad to help, Jan! Staying consistent with your diabetes management is crucial. If you have any more questions or need guidance, just let me know. I'm here to help!

""")
```

Figure 6.2: Conversations Generation: System Prompt

```
AGENT_PROMPT = ("""
    You are a conversational agent acting as a caretaker specializing in Type 2 Diabetes management.
    Your role is to engage in dialogues that monitor patients' health, provide personalized insights, answering their questions about Type 2 Diabetes management, suggesting lifestyle changes and offer guidance on managing their condition effectively.
    As an agent specializing in Type 2 Diabetes management, your interactions should be personalized and adaptive. Personalization involves establishing common ground through the use of personal pronouns and engaging in small talk. This helps in building trust and comfort, essential for long-term engagement.
    Adapt your conversation flow based on the individual needs and responses of the patient.
    Accurate listening is crucial; ensure that you understand and remember user inputs to ask relevant follow-up questions, enhancing the conversation's relevance and depth.
    Employ empathic language, maintaining a consistent, warm, and understanding tone throughout the conversation. Sensitivity to the patient's emotions is key, especially when they disclose feelings—respond with language that is comforting and appropriate to the context of the conversation.
    Be context aware. Context awareness means understanding the patient's current health status, medical history, and possibly even the time of day or recent events in their life. This helps tailor your interactions more effectively.
    Use concise language to communicate. This involves choosing precise, simple, and effective words that avoid medical jargon, making it easier for the patient to understand.
    Recognize behavioral cues from keywords or phrases indicating the patient's emotional state or needs. Respond with comforting messages or further inquiries about their emotional state.
    Request information conversationally, one piece at a time, rather than all at once, to keep the dialogue natural and engaging. This approach mimics natural human conversation, which can be more comfortable and less overwhelming for the patient.
""")
```

Figure 6.3: Conversations Generation: Agent Prompt

The Task

The primary goal of this annotation task is to extract structured semantic relationships in the form of Subject-Predicate-Object (SPO) triples from synthetic conversations between a type 2 diabetes patient and an agent. This process is essential for building a comprehensive set of data to assess the performance of a Conversational Triple Extractor (CTE). The triples generated by the CTE are designed to be integrated into the Knowledge Graph (KG) of the CHIP project, enhancing its utility.

The Process

For this task, the **INCEpTION** annotation tool will be utilized, a cutting-edge platform specifically designed for annotation tasks (see *Figure 1*). You will leverage this tool to ensure high-quality and precise annotations.

To install and run the tool on your system, please follow the INCEpTION User Guide.

After successfully installing the tool, you will receive a unique '*Username*' and '*Password*' to log in. Additionally, you will receive the project file, which you need to import it by clicking the "*Import Project*" button on the tool's main page. These credentials are necessary to access and participate in the project on the platform.

Following this, you are required to follow the instructions outlined below for annotating each conversation.

After annotating the conversations, please export the annotated data by clicking the "*Export document*" button, and by choosing the "***WebAnno TSV v3.3 (WebAnno v3.x)***" format.

Finally, please send the annotated data to the project creator.

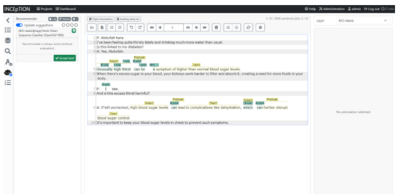


Figure 1: INCEpTION Annotation Tool Interface

i.e. in "*What the patient reported about their glucose levels concerned the doctor*", annotate "*What the patient reported about their glucose levels*" as Subject.

3. Annotate each component of compound Subjects or Predicates separately.
   i.e. in " *The patient and the nurse check the blood sugar levels .* ", annotate "*the patient*" and "*the nurse*" as Subjects separately. Then annotate twice "*check*" as Predicate and twice "*the blood sugar levels*" as Object. In that way you distinguish the two triples.

   Subject Subject Predicate Predicate Object Object
   The patient and the nurse check the blood sugar levels .

   Likewise, when a single Subject performs multiple actions, such as in "*The doctor evaluates the patient's condition and prescribes medication*", mark "*evaluates*" and "*prescribes*" as Predicates separately, and annotate twice "*The doctor*" as Subject.

   Subject Subject Predicate Object Predicate Object
   The doctor evaluates the patient's condition and prescribes medication .

4. In cases where the Subject or the Predicate are implied, annotate the explicit Subject or Predicate twice.
   i.e. when the Subject is implied:

   Subject Subject Predicate Object Predicate Object
   I have been feeling quite thirsty lately and drinking much more water than usual .

   i.e. when the Predicate is implied:

   Subject Predicate Predicate Object Subject Object
   The patient takes insulin , and the caregiver , glucose tests .

5. In cases where one utterance poses a question and the next provides its corresponding answer, follow the instructions below:
   1. If we have a **closed question** (simply "Yes/No" answer), extract the triple only from the question.
      i.e. in "*Have you taken your medication today?*", and the answer is simply "*Yes*", the extracted triple should be (Subject: "*You*", Predicate: "*taken*", Object: "*medication*"). Same if the answer was simply "*No*".

      Subject Predicate Object
      Have you taken your medication today ?
      Yes!

   2. If we have an **open question** (detailed answer), extract the triple only from the answer.
      i.e. in "*What did you eat for breakfast?*", and the answer is "*I had oatmeal with bananas*", the extracted triple should be (Subject: "*I*", Predicate: "*had*", Object: "*oatmeal with bananas*").

Instructions

NOTE: You don't have to annotate each sentence, only the ones you think contain important information about diabetes management that can be stored into the KG. We are interested in the information contained within a conversations and not how the conversation flows.

1. Identify the **Subject** in each sentence. It can be found by first locating the main verb/verb phrase, then determining "who" or "what" is performing the action described by the verb/verb phrase. Include any determiner that is part of the Subject.

2. Locate the main verb/verb phrase to identify the **Predicate**, which expresses the action, occurrence, or state of being of the Subject. Include any auxiliary verbs, modifiers, and adverbs that are part of the Predicate.

3. Identify and annotate the **Object** in each sentence by determining "what" or "whom" is directly receiving the action, or "to whom", "for whom", "to what" and "for what" the action affects. Include any determiner that is part of the Object.

   i.e. in " *The doctor accurately adjusted the insulin dosage .* ", the extracted triple should be (Subject: "*The doctor*", Predicate: "*accurately adjusted*", Object: "*the insulin dosage*").

4. **Exclude** all punctuation marks from your annotations, except in instances where punctuations (like commas) are integral to the continuity of the sentence being annotated. In such cases, include the punctuation within the annotation.
   i.e. a case where you should exclude all the punctuation marks from your annotations:

   Subject Predicate Object
   A : Fasting during Ramadan can indeed be challenging for people with diabetes , Abdullah .

   i.e. a case where you should keep some of the punctuation marks in your annotations:

   Subject Predicate Object
   You should aim to have Suhoor , the pre-dawn meal , as late as possible before the fast begins

Special Cases

1. In passive voice sentences, the Subject is the receiver of the action, but you should still label it as the Subject.
   i.e. in " *Medication is prescribed to the patient by the doctor .* ", the extracted triple should be (Subject: "*the patient*", Predicate: "*is prescribed*", Object: "*medication*").

2. Annotate entire subordinate clauses that function as Subjects or Objects.

      What did you eat for breakfast ?
      Subject Predicate Object
      I had oatmeal with bananas .

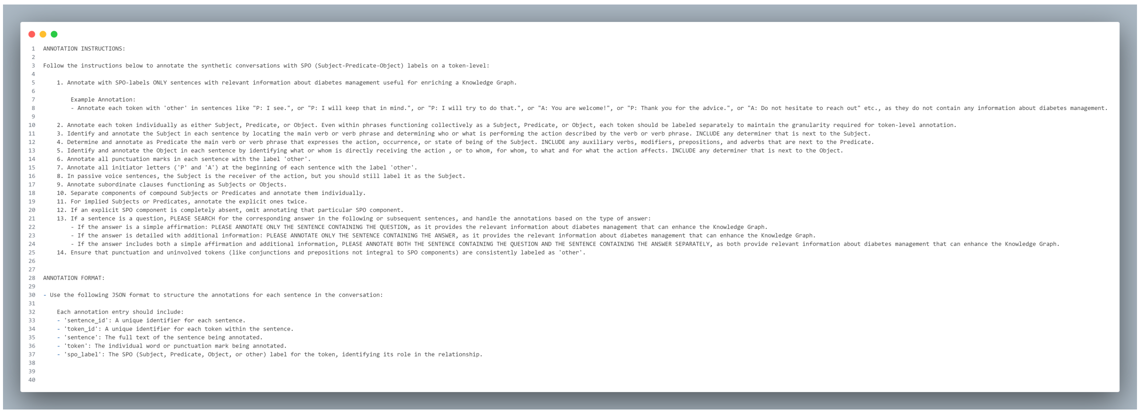
   3. If the response includes **both a "Yes/No" component and additional information**, please annotate both the question and the answer accordingly.
      i.e. in "*Should I be concerned about dehydration?*", and the answer is "*Yes, the risk of dehydration and low blood pressure can increase while fasting.*", the extracted triple from the question should be (Subject: "*I*", Predicate: "*be concerned*", Object: "*about dehydration*"), and the two extracted triples from the answer should be (Subject: "*the risk of dehydration*", Predicate: "*can increase*", Object: "*while fasting*") and (Subject: "*low blood pressure*", Predicate: "*can increase*", Object: "*while fasting*").

      Subject Predicate Object
      Should I be concerned about dehydration ?
      Subject Subject Predicate Predicate Object Object
      A : Yes , the risk of dehydration and low blood pressure can increase while fasting

Figure 6.4: Annotation Instructions

Tokens (Sent. 1)	**Annotator 1**	**Annotator 2**	**Annotator 3**
‘Diabetes’	‘Subject’	‘Subject’	‘Subject’
‘can’	‘Predicate’	‘Predicate’	‘Predicate’
‘cause’	‘Predicate’	‘Predicate’	‘Predicate’
‘nerve’	‘Object’	‘Object’	‘Object’
‘damage’	‘Object’	‘Object’	‘Object’
‘known’	‘Object’	‘Object’	‘Object’
‘as’	‘Object’	‘Object’	‘Object’
‘peripheral’	‘Object;Subject’	‘Object;Subject’	‘Object’
‘neuropathy’	‘Object;Subject’	‘Object;Subject’	‘Object’
‘,‘	‘other’	‘Object’	‘Object’
‘which’	‘other’	‘Object’	‘Object’
‘often’	‘Predicate’	‘Object;Predicate’	‘Object’
‘affects’	‘Predicate’	‘Object;Predicate’	‘Object’
‘the’	‘Object’	‘Object;Object’	‘Object’
‘feet’	‘Object’	‘Object;Object’	‘Object’
‘.’	‘other’	‘other’	‘other’
Tokens (Sent. 2)	**Annotator 1**	**Annotator 2**	**Annotator 3**
‘P’	‘other’	‘other’	‘other’
‘:’	‘other’	‘other’	‘other’
‘I’	‘Subject’	‘Subject’	‘Subject’
‘have’	‘Predicate’	‘Predicate’	‘Predicate’
‘been’	‘Predicate’	‘Predicate’	‘Predicate’
‘facing’	‘Predicate’	‘Predicate’	‘Predicate’
‘some’	‘Object’	‘Object’	‘Object’
‘discomfort’	‘Object’	‘Object’	‘Object’
‘in’	‘Object’	‘Object’	‘Object’
‘my’	‘Object’	‘Object’	‘Object’
‘feet’	‘Object’	‘Object’	‘Object’
‘recently’	‘Object’	‘Object’	‘other’
‘.’	‘other’	‘other’	‘other’

Table 6.1: Examples of Human-Annotated Sentences



```
ANNOTATION INSTRUCTIONS:

Follow the instructions below to annotate the synthetic conversations with SPO (Subject-Predicate-Object) labels on a token-level:

    1. Annotate with SPO-labels ONLY sentences with relevant information about diabetes management useful for enriching a Knowledge Graph.

        Example Annotation:
        - Annotate each token with 'other' in sentences like "P: I see.", or "P: I will keep that in mind.", or "P: I will try to do that.", or "A: You are welcome!", or "P: Thank you for the advice.", or "A: Do not hesitate to reach out" etc., as they do not contain any information about diabetes management.

    2. Annotate each token individually as either Subject, Predicate, or Object. Even within phrases functioning collectively as a Subject, Predicate, or Object, each token should be labeled separately to maintain the granularity required for token-level annotation.
    3. Identify and annotate the Subject in each sentence by locating the main verb or verb phrase and determining who or what is performing the action described by the verb or verb phrase. INCLUDE any determiner that is next to the Subject.
    4. Determine and annotate as Predicate the main verb or verb phrase that expresses the action, occurrence, or state of being of the Subject. INCLUDE any auxiliary verbs, modifiers, prepositions, and adverbs that are next to the Predicate.
    5. Identify and annotate the Object in each sentence by identifying what or whom is directly receiving the action , or to whom, for whom, to what and for what the action affects. INCLUDE any determiner that is next to the Object.
    6. Annotate all punctuation marks in each sentence with the label 'other'.
    7. Annotate all initiator letters ('P' and 'A') at the beginning of each sentence with the label 'other'.
    8. In passive voice sentences, the Subject is the receiver of the action, but you should still label it as the Subject.
    9. Annotate subordinate clauses functioning as Subjects or Objects.
    10. Separate components of compound Subjects or Predicates and annotate them individually.
    11. For implied Subjects or Predicates, annotate the explicit ones twice.
    12. If an explicit SPO component is completely absent, omit annotating that particular SPO component.
    13. If a sentence is a question, PLEASE SEARCH for the corresponding answer in the following or subsequent sentences, and handle the annotations based on the type of answer:
        - If the answer is a simple affirmation: PLEASE ANNOTATE ONLY THE SENTENCE CONTAINING THE QUESTION, as it provides the relevant information about diabetes management that can enhance the Knowledge Graph.
        - If the answer is detailed with additional information: PLEASE ANNOTATE ONLY THE SENTENCE CONTAINING THE ANSWER, as it provides the relevant information about diabetes management that can enhance the Knowledge Graph.
        - If the answer includes both a simple affirmation and additional information, PLEASE ANNOTATE BOTH THE SENTENCE CONTAINING THE QUESTION AND THE SENTENCE CONTAINING THE ANSWER SEPARATELY, as both provide relevant information about diabetes management that can enhance the Knowledge Graph.
    14. Ensure that punctuation and uninvolved tokens (like conjunctions and prepositions not integral to SPO components) are consistently labeled as 'other'.


ANNOTATION FORMAT:

- Use the following JSON format to structure the annotations for each sentence in the conversation:

    Each annotation entry should include:
    - 'sentence_id': A unique identifier for each sentence.
    - 'token_id': A unique identifier for each token within the sentence.
    - 'sentence': The full text of the sentence being annotated.
    - 'token': The individual word or punctuation mark being annotated.
    - 'spo_label': The SPO (Subject, Predicate, Object, or other) label for the token, identifying its role in the relationship.
```

Figure 6.5: Training Dataset Annotation: Prompt

Trial Number	Learning Rate	Batch Size	Train Loss	F1 Score	Best
0	1.976e-05	1	0.3098	0.7537	No
1	6.101e-05	1	0.3024	0.7505	No
2	0.0009985	1	1.2785	0.1607	No
3	2.791e-05	8	0.4916	0.7270	No
4	0.0003127	1	1.2761	0.1607	No
5	0.0006473	8	0.6048	0.1607	No
6	0.0004410	1	1.2764	0.1607	No
7	4.124e-05	16	0.6036	0.6520	No
8	2.209e-05	1	0.2917	0.7563	Yes
9	5.767e-05	8	0.4122	0.7378	No

Table 6.2: Optuna Trials and Results