A Synthetic Conversational Dataset for Type 2 Diabetes Management

Stergios Ntanavaras[1], Maaike H. T. de Boer[2], Piek Vossen[1]
[1] Vrije Universiteit Amsterdam, 1081 HV Amsterdam, The Netherlands
[2] TNO, 3769 DE Soesterberg, The Netherlands
s.ntanavaras@student.vu.nl, maaike.deboer@tno.nl, piek.vossen@vu.nl

Abstract

Access to real patient-doctor conversations in the medical domain is often restricted due to privacy concerns, making it difficult to build robust conversational AI systems. To address this, we present a novel methodology for generating a high-quality synthetic dataset designed for conversational triple extraction in Type 2 Diabetes management. Using structured prompting with GPT-4, we generated 16 demographically and medically diverse diabetic personas, and 256 multi-turn conversations between these personas and a caretaker agent, simulating realistic and context-rich interactions. The conversations incorporate critical properties such as personalization, empathy, contextual awareness, and medically grounded advice, as validated through both LLM-based and human expert evaluations. These synthetic conversations are further annotated with Subject-Predicate-Object (SPO) labels at the token level, integrating both manual and LLM-automated methods, forming the foundation for downstream tasks like triple extraction. Our work demonstrates the feasibility of using generative AI to simulate healthcare conversations at scale, offering a solution for data-scarce domains.

Keywords: Synthetic Data Generation, Conversational AI, Dialogue Annotation, Conversational Triple Extraction, Type 2 Diabetes Management

1. Introduction

The increasing burden of chronic diseases such as Type 2 Diabetes poses a significant challenge to healthcare systems worldwide, particularly in aging populations (Changizi and Kaveh, 2017). Traditional healthcare infrastructures often struggle to provide personalized and accessible care on a scale (Wagner et al., 1996), creating an urgent need for innovative solutions that can support both patients and healthcare professionals in managing these complex conditions.

Hybrid Intelligence (HI) and conversational agents offer a promising way to support chronic disease management, including Type 2 Diabetes, by making care more personalized and accessible (Laranjo et al., 2018; Dudzik et al., 2024). They can provide around-the-clock availability, consistent information delivery, and scalable patient engagement, while reducing the burden on healthcare professionals. However, developing and evaluating such agents requires access to high-quality domain-specific datasets that reflect the nuances of real patient-agent interactions.

Unfortunately, several critical gaps exist in the current landscape of healthcare dialogue datasets. First, especially for chronic diseases like Type 2 Diabetes, data privacy regulations restrict access to authentic patient dialogues, resulting in a scarcity of annotated conversational data. Second, most existing datasets are general-purpose and do not address the specific informational and behavioral needs associated with managing Type 2 Diabetes (Peerbolte et al., 2025). Lastly, most of the available dialogue resources are based on human–human interactions, offering limited support for training and evaluating agent-driven communication. Together, these limitations present a critical bottleneck in the advancement of language technologies for medical settings.

To overcome this challenge, this paper presents a methodology for generating a high-quality synthetic conversational dataset tailored to the context of Type 2 Diabetes management. These synthetic conversations simulate realistic interactions between patients and a caretaker agent, providing the necessary foundation for downstream tasks such as triple extraction and knowledge graph enrichment. More specifically, the dataset is designed to train and benchmark models that extract Subject–Predicate–Object (SPO) triples from patient utterances, which are then used to populate and maintain a User Knowledge Graph within a broader HI system for personalized lifestyle support.

We propose a methodology for constructing such a dataset using structured prompting with detailed task instructions and large language models (LLMs). Our approach combines persona-driven scenario creation with dialogue generation, guided by design principles from healthcare communication and conversational AI. We describe how we created culturally grounded patient personas, generated multi-turn dialogues, and ensured diversity and realism across interactions.

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 171–181

This work is situated within the CHIP initiative (den Hengst et al., 2026), a collaborative effort by Dutch institutions to develop an HI system for Type 2 Diabetes management. CHIP is a modular, microservice-based research platform that integrates dialogue processing, information extraction, knowledge graphs, and reasoning to support lifestyle changes in patients. Within CHIP's architecture, conversational interactions are first processed by a knowledge extractor, which converts free-form dialogue text into SPO triples. These triples are stored in a User Knowledge Graph alongside a Domain Knowledge Graph of medical knowledge. A reasoning component then consults both graphs to tailor interventions to individual patient contexts and goals. The synthetic dialogues and conversational triple extraction methods presented in this paper directly support the knowledge extraction step in this pipeline, providing annotated training and evaluation data for models that must operate on realistic, domain-specific patient–agent exchanges. By generating this resource synthetically, we circumvent the privacy constraints that would otherwise prevent the use of real patient dialogues for model development.

To support transparency and reproducibility, all diabetes personas, synthetic conversations, and their corresponding annotations are available in the accompanying repository [1].

2. Related Work

Synthetic dataset generation is increasingly seen as a practical way to address data scarcity in conversational AI, especially in domains with privacy or access constraints (Chavda and Bhattacharyya; Ramesh et al., 2024). Early studies explored data augmentation techniques, using rule-based transformations or back-translation to expand existing corpora, but these approaches produced limited linguistic and contextual diversity (Soudani et al., 2023).

With the advent of large language models (LLMs), researchers have shifted toward generating full dialogues synthetically. For instance, Honovich et al. (Honovich et al., 2022) demonstrated how prompt-based models can be leveraged to automatically generate diverse synthetic training conversations that capture richer context, intent, and diversity while reducing dependency on costly or restricted real-world data.

In clinical and healthcare dialogue modeling, a smaller but growing body of work has begun producing synthetic resources. SynDial (Das et al., 2024) generates synthetic patient–physician dialogues from clinical notes and uses iterative refinement to maintain factuality. NoteChat (Wang et al., 2024) produces doctor-patient conversations conditioned on clinical notes for better documentation and note generation performance. MedSynth (Mianroodi et al., 2025) provides over 10,000 dialogue-note pairs across many disease codes, facilitating both Dialogue-to-Note and Note-to-Dialogue tasks. Bedrick et al. (Bedrick et al., 2025) offer a typology of existing synthetic datasets for clinical dialogues, highlighting how synthetic methods vary in realism, domain specificity, and use case.

Beyond task-oriented generation, recent research has emphasized the role of social and personal dimensions in synthetic dialogues. Santamaría et al. (Santamarıa et al., 2023) argue that open-domain dialogue datasets should not only encode factual content but also capture personal viewpoints and shared memories over time. Their work highlights the need for structured representations that track factoids and perspectives across long-term interactions, which can inspire the development of more transparent and socially attuned conversational agents. This perspective resonates with the challenges in healthcare, where patient narratives are inherently personal and context-dependent.

Despite these advances, synthetic datasets tailored for downstream tasks such as conversational triple extraction remain scarce. This is especially true for resources that include Subject-Predicate-Object (SPO) annotations, multi-turn patient–agent dialogues, and domain grounding for Type 2 Diabetes. Our dataset and study address this gap by combining synthetic dialogue generation, persona grounding, comprehensive SPO annotation, and evaluation in a knowledge-graph-oriented pipeline for CHIP.

3. Methodology

To ensure practical utility and align with the needs identified in Section 2, we target three constraints: limited access to real dialogues due to privacy, lack of domain-specific resources for Type 2 Diabetes, and the need to model agent-driven interactions. We define design requirements to guide the creation of the synthetic conversational dataset and to align with the CHIP initiative.

The dataset was designed to meet the following requirements:

- All conversations focus on Type 2 Diabetes management, covering topics such as symptoms, treatment options, lifestyle changes, and patient concerns.
- Dialogues should reflect *patient–agent* setting rather than *human–human* interaction.

[1] https://github.com/stergios97/synthetic-conversational-data-diabetes-management

- Each dialogue contains 4–6 exchanges with natural turn-taking; openings are roughly balanced between patient-start and agent-start to avoid initiation bias.
- Diversity across intents, scenarios, and personas, with coverage of all defined profiles.
- Only fully synthetic content, with no real patient data involved, in order to comply with privacy standards and ethical research practices.

The data generation process consists of three steps: creating diverse patient personas, generating multi-turn dialogues with a caretaker agent, and annotating conversations with Subject–Predicate–Object (SPO) roles. The result is a dataset of fully annotated conversational sentences, with each step described in the following sections.

3.1. Persona Generation

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 designed to capture a wide range of demographic traits, medical histories, and treatment adherence behaviors, providing a detailed 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 domain experts, ensuring that the personas accurately reflect the diversity and nuances of the population of patients with Type 2 diabetes in the Netherlands.

To construct realistic patient personas, we grounded our trait selection in epidemiological and demographic data from Dutch public health sources. Age values were chosen around the national mean of 65.9 years for Type 2 Diabetes outpatients, as reported by the Dutch Pediatric and Adult Registry of Diabetes (DPARD) (Bak et al., 2021). We selected five representative age groups (55, 60, 65, 70, 75) to reflect the elderly population most impacted by the disease. The ethnicity was chosen based on prevalence statistics from StatLine, the Dutch Central Bureau of Statistics (CBS) (Centraal Bureau voor de Statistiek, 2024) and supporting studies (Ujcic-Voortman et al., 2009), which show elevated Type 2 Diabetes rates among Moroccan (7.09%), Surinamese (6.15%), and Turkish (5.95%) populations, compared to Dutch individuals (1.99%). Therefore, these ethnic groups were selected for inclusion, given both their medical relevance and their demographic presence in the Netherlands. To ensure cultural authenticity, we assigned each persona a name commonly associated with their ethnic background, as verified through multiple public sources [2]. Moroccan personas were named Mohammed, Abdullah, Aicha, and Fatima; Surinamese: Rudolf, Johan, Julia, and Ingrid; Turkish: Ali, Mehmet, Ayşe, and Fatma; Dutch: Pieter, Jan, Johanna, and Maria.

Using these attributes, all possible combinations were generated, respecting gender-specific naming conventions to ensure realism and cultural appropriateness. From these, 16 unique profiles were carefully selected to achieve a balanced and manageable set that covers the full range of age, gender, education, and cultural background while ensuring equal representation across gender and ethnicity. The number 16 was chosen as it provides sufficient diversity without overwhelming the annotation process. All selected profiles are presented in Table 1.

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
Mehmet	Male	Turkish	65

Table 1: Selected Profiles

3.2. Prompts for Persona Generation

Once the core demographic attributes were selected, we generated detailed biographies for each persona using structured prompting with detailed task instructions. To guide the model in producing realistic and context-rich outputs, we designed manual prompts that elicited narrative descriptions that include key dimensions of diabetes care. These included 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 generation was carried out using GPT-4 [3], identified as the most proficient model within the GPT series according to a report from OPENAI (OpenAI, 2024). To generate detailed and informative persona descriptions, the `temperature`

[2]Names were chosen based on their common usage within each ethnic group, as verified across multiple public online sources.

[3]OpenAI `gpt-4` (June 2023 release, `gpt-4-0613`)

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, ensuring that the descriptions were comprehensive and rich in detail.

The resulting descriptions captured interrelated aspects of each persona's life that could influence their interaction with a conversational healthcare agent. For instance, some personas were digitally literate and health-conscious, while others had limited access to care or expressed mistrust in medical institutions.

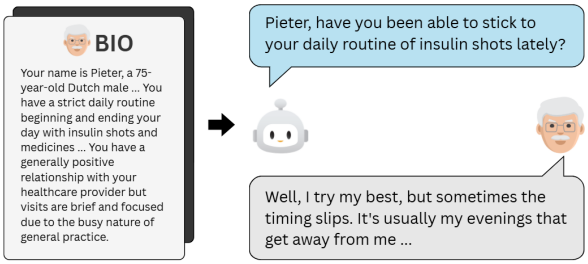

Figure 1: Example of conversation generation from a persona profile. The biography of "Pieter" (75-year-old Dutch male with Type 2 Diabetes) is used to generate a sample dialogue with the caretaker agent.

3.3. Conversation Generation

Building on the finalized personas, we generated synthetic patient-agent dialogues to simulate realistic Type 2 Diabetes management scenarios. This transition from persona definition to conversation generation is illustrated in Figure 1. For each of the 16 personas, we produced 16 conversations. This number provided adequate topical diversity per persona while keeping manual checks and subsequent annotation/evaluation feasible and balanced across the corpus (equal gender/ethnicity coverage and manageable size). Conversations were evenly divided between patient- and agent-initiated openings, and each dialogue was restricted to 4 to 6 exchanges with natural turn-taking and internal coherence.

The final dataset consists of 256 dialogues covering a wide range of diabetes-related topics, including medication adherence, lifestyle adjustments, emotional support, and access to care. These conversations provide a robust foundation for training and evaluating conversational systems focused on personalized health communication. Details of the prompting setup are given in Section 3.4. Below are two examples of the conversations generated.

Patient-initiated example ("Abdullah", Table 1):

P: I'm worried about fasting for Ramadan with my diabetes. Is it possible without harming my health?
A: Yes, Abdullah, fasting during Ramadan is possible with diabetes, but it requires careful planning. You should consult your healthcare provider to adjust your medication and meal times according to your fasting schedule.

[conversation continues...]

Agent-initiated example ("Maria", Table 1):

A: Maria, how has your blood sugar control been over the past week? Do you notice any fluctuations?
P: Well, lately I've noticed that my blood sugar tends to increase significantly after my dinner meal.
A: I see. Can you share what a typical dinner looks like for you?
P: Usually, it's a lean protein like chicken, and I try to fill half my plate with vegetables.

[conversation continues...]

3.4. Prompts for Conversation Generation

The generation of realistic and varied dialogues required carefully designed prompts tailored to the roles of the patient and the agent, as well as a coordinating system prompt. Each component was constructed to ensure coherent, context-sensitive exchanges grounded in the persona characteristics.

The *system prompt* defined the scope and structure of each dialogue, instructing the model to produce exactly 4 to 6 exchanges per conversation. To enforce this constraint and improve adherence, we applied strategies such as capitalization and repetition of critical instructions. These techniques enhanced the model's compliance with formatting and interaction rules during generation.

The *agent prompt* was crafted to encourage empathetic, adaptive, and informative behavior. Drawing on healthcare communication research, we included directives for using friendly and supportive language, engaging in small talk, and adapting responses to the patient's emotional and informational needs. The agent was also instructed to avoid medical jargon, maintain conversational tone, and gather information incrementally, asking one question at a time to avoid overwhelming the patient.

For the *patient prompt*, we injected each persona biography directly into the input to simulate realistic behavior aligned with their background, beliefs, and

Overview		Dialogue Acts	
Number of dialogues	256	Statements	77.8%
Total utterances	2,027	Comments	7.5%
Total tokens	52,019	Opinions	7.3%
Unique word types	5,337	Positive answers	2.8%
Avg. utt./dialogue	7.9	Commands	2.8%
		Back-channeling	0.7%
Emotions			
Neutral	785	Curiosity	432
Gratitude	211	Approval	142
Fear	33	Disappointment	11
Sadness	19		

Table 2: Dataset statistics, dialogue acts, and emotions.

communication style. This allowed each dialogue to reflect the unique characteristics of the persona, including their health concerns, digital literacy, or trust in medical advice.

We experimented with both zero-shot and few-shot prompting approaches. Zero-shot prompts often failed to produce the required dialogue length or diversity. In contrast, few-shot prompting significantly improved both structure and variety. As a result, we adopted a few-shot format using four curated examples per prompt to guide the model toward generating coherent, contextually rich conversations.

All conversations were generated via the OpenAI GPT-4 model (same version as noted earlier); the generation parameters were set to a `temperature` of 0.9 to encourage linguistic variation and a `max_tokens` of 300 to keep outputs concise.

All prompts used in the generation process are made available in the accompanying repository [1].

4. Dataset Analysis and Evaluation

4.1. Statistical Analysis

Table 2 summarizes key dataset statistics, dialogue act distributions, and emotion annotations. The synthetic dataset comprises 256 multi-turn patient–agent conversations related to Type 2 Diabetes management, totaling 2,027 utterances and 52,019 tokens, with a vocabulary of 5,337 unique word types. On average, each dialogue contains 7.9 utterances, demonstrating a moderately complex interaction structure.

Dialogue acts were annotated using a method consistent with the approach described by Santamaría et al. (Santamaria et al., 2023), implemented with the MIDAS dialogue act classifier (Yu and Yu, 2019). Specifically, each utterance was passed through a Transformers text classification pipeline loading a MIDAS-based XLM-RoBERTa model, and the top-ranked dialogue act label was assigned. Analysis of dialogue acts reveals that statements dominate (77.8%), followed by comments (7.5%), opinions (7.3%), and positive answers (2.8%). Commands and back-channeling are less frequent, at 2.8% and 0.7% respectively. This distribution reflects a dataset focused on information exchange and patient education, with limited small-talk or filler content.

Emotion annotations were derived using the same pipeline-based approach, this time loading the GoEmotions BERT model (Demszky et al., 2020) and selecting the top-ranked emotion label per utterance. The results indicate that neutral utterances are the most prevalent (785 instances), followed by curiosity (432), gratitude (211), and approval (142). Negative emotions such as fear (33), disappointment (11), and sadness (19) appear, but are relatively rare. These results show that while patient concerns are captured, the conversational tone remains largely informative and supportive.

The dataset demonstrates high lexical diversity with a type–token ratio (TTR) of 0.10 (5,337 unique types over 52,019 tokens). Subjects are mostly first- and second-person pronouns (“I”, “you”), aligning with the interactive nature of patient–agent dialogues. Frequent predicates include “be”, “try”, “feel”, and “make”, indicating a strong focus on advice, recommendations, and patient states.

Utterance lengths vary substantially, with character counts ranging from 37 to 352. However, most responses fall within 80–150 characters, balancing naturalness and informativeness. Additionally, there is minimal reliance on very short utterances (e.g., “Yes”, “No”), which supports richer patient-agent interactions.

Compared to the open-domain dialogue datasets analyzed by Santamaría et al. (Santamaria et al., 2023) and discussed in Section 2, our dataset differs substantially in domain specificity, conversational structure, and interactional goals. While open-domain resources often emphasize personal perspectives, emotional variability, and human–human social bonding, our dataset is designed to simulate patient–agent interactions focused on fact-based, medically accurate exchanges. Conversations are shorter than narrative-rich corpora like PEDC (Eisenberg and Sheriff, 2020) but exhibit greater linguistic variation and topical specificity than most task-oriented datasets. This unique combination of healthcare-focused, multi-turn dialogues with annotated emotional and linguistic features makes the dataset well-suited for training and evaluating conversational agents in Type 2 Diabetes management, addressing limitations in existing resources.

4.2. Persona Evaluation

To ensure the generated personas were both realistic and relevant for the diabetes care domain, all profiles underwent a manual evaluation and refinement process in collaboration with domain experts—researchers working on diabetes lifestyle support with a background in health communication and conversational AI. This review focused on validating the coherence, plausibility, and representativeness of each profile in relation to the challenges faced by patients managing Type 2 Diabetes.

During the review, particular attention was given to aligning personas with the broader healthcare challenges motivating CHIP. For instance, initial versions of the generated biographies included several well-off individuals with strong healthcare support networks. To better reflect the systemic gaps that CHIP aims to address, these profiles were revised to incorporate traits such as financial hardship, limited access to care, and lower health or digital literacy.

This expert-guided refinement ensured that each persona embodied characteristics that would meaningfully test the capacity of conversational agents to respond to real-world complexities in patient communication and care personalization. The final set of personas thus captures a wide spectrum of needs, barriers, and behaviors, making them a robust foundation for generating diverse and clinically relevant synthetic dialogues.

4.3. Conversation Evaluation

To assess the linguistic quality of the conversations generated, we performed an LLM-based and a human-based evaluation using a custom questionnaire. Both approaches focused on four core dimensions of language quality: fluency, naturalness, clarity, and structure. These dimensions were selected to reflect how effectively the dialogues simulate natural and coherent patient–agent communication.

4.3.1. LLM-based Evaluation

All 256 generated dialogues were included in the LLM-based evaluation process following the G-Eval framework (Liu et al., 2023). Each dialogue was holistically assessed on each of the four criteria mentioned above using GPT-4o [4]. The model was instructed to assign a probability distribution on a 5-point Likert scale (Likert, 1932) based on a detailed rubric per dimension. Then, we computed two key metrics from the outputs: 1. the expected score: a weighted average representing the most likely rating, and 2. the entropy: a measure of confidence (lower entropy indicating higher confidence). To ensure reproducibility, the evaluation was conducted with `temperature` set to 0.0, making the process deterministic. Table 3 summarizes the average scores and their variability (standard deviations) across dialogues.

Dimension	Avg. Score (±SD)	Avg. Entropy (±SD)
Fluency	4.70 ± 0.14	0.64 ± 0.19
Naturalness	4.45 ± 0.09	0.91 ± 0.04
Clarity	4.47 ± 0.05	0.91 ± 0.02
Structure	4.14 ± 0.12	0.86 ± 0.10

Table 3: GPT-4o evaluation results showing mean $\pm$ SD of scores (1–5 scale) and entropy across 256 dialogues.

As shown in Table 3, the synthetic dialogues achieved consistently high scores across all four dimensions, with particularly strong performance in fluency (4.70 ± 0.14) and clarity (4.47 ± 0.05). These results suggest that the model consistently judged the conversations as grammatically correct, clear, and easy to follow. The small standard deviations indicate that these high scores are uniform across the 256 dialogues, especially for clarity, where the dispersion is minimal (SD $= 0.05$). Naturalness (4.45 ± 0.09) was also highly rated, indicating that dialogues generally sounded realistic and conversational, though with slightly more variation compared to fluency. The relatively low entropy values across all dimensions reflect stable and confident model judgments. Structure received the lowest score (4.14 ± 0.12), pointing to occasional weaknesses in how model responses were organized or sequenced, which may offer room for future refinement in prompt design or dialogue planning.

Dimension	Agent Starter	Patient Starter
Fluency (±SD)	4.70 ± 0.14	4.70 ± 0.14
Naturalness (±SD)	4.46 ± 0.10	4.45 ± 0.09
Clarity (±SD)	4.46 ± 0.05	4.48 ± 0.04
Structure (±SD)	4.13 ± 0.13	4.15 ± 0.09

Table 4: Evaluation scores (mean $\pm$ SD) by dialogue starter type.

Looking at dialogue initiation at Table 4, we found no substantial difference in evaluation scores between agent- and patient-starting dialogues. Both groups scored almost identically across fluency, naturalness, and clarity. Patient-initiated dialogues had slightly higher clarity and structure, though the differences are marginal ($\leq$0.02). Entropy values were also nearly identical, suggesting GPT-4o's confidence did not vary depending on the first speaker. This indicates that our prompting setup yielded consistent dialogue quality regardless of turn-taking structure at the opening.

[4] OpenAI `gpt-4o` (August 2025 release)

A methodological consideration of this evaluation setup is the potential for self-preference bias: since the dialogues were generated by GPT-4 and evaluated by GPT-4o (a model from the same family), the evaluator may exhibit systematic leniency toward outputs that conform to its own generation patterns (Panickssery et al., 2024). To mitigate the impact of this concern, we complement the LLM-based evaluation with an independent human expert evaluation (4.3.2), which may capture quality dimensions that the LLM-based evaluation alone cannot.

4.3.2. Human-based Evaluation

To complement the LLM-based analysis, we conducted a human study using our Synthetic Patient-Agent Conversation Evaluation Questionnaire (SPACEQ), inspired by the Chatbot Usability Questionnaire (CUQ) (Holmes et al., 2019). The full questionnaire is available in the accompanying repository [1]. Although the LLM-based evaluation was conducted over all 256 generated conversations, for the human evaluation, we selected 20 dialogues (covering all 16 persona profiles from Table 1) to limit cognitive load while ensuring representativeness.

Ten domain experts—researchers working on diabetes lifestyle support mostly affiliated with CHIP or related projects in the medical field—were instructed to read each dialogue carefully and rate it on a 5-point Likert scale (Likert, 1932) (1 = Very Poor, 2 = Poor, 3 = Acceptable, 4 = Good, 5 = Excellent), with optional free-text remarks after each item to capture explanations and qualitative feedback. The questionnaire content and rating rubric mirror the language-level criteria used in the LLM-based evaluation in Section 4.3.1, enabling a direct, dimension-wise comparison between human judgments and model-based scores.

Dimension	Avg. Human Score ($\pm$SD)	α
Fluency	4.65 $\pm$ 0.54	-0.035
Naturalness	3.95 $\pm$ 0.89	0.020
Clarity	4.48 $\pm$ 0.77	0.021
Structure	4.54 $\pm$ 0.66	0.050

Table 5: Human evaluation over 20 dialogues and 10 expert raters. For each dimension, we report the mean and standard deviation (SD) across all ratings, alongside inter-rater agreement using Krippendorff's α (ordinal).

As shown in Table 5, raters judged dialogues as highly fluent (4.65 ± 0.54) and clear (4.48 ± 0.77), but naturalness received the lowest score (3.95 ± 0.89). Higher standard deviations for naturalness and clarity suggest greater variability, with some dialogues considered realistic while others were described as "too rigid" or "repetitive". In contrast, fluency showed lower variability, indicating stronger agreement on grammatical correctness.

Inter-rater agreement was estimated with Krippendorff's α (ordinal) (Krippendorff, 2011) across the 10 raters for the 20 conversations. The coefficients were close to zero, which at face value would suggest poor agreement. However, this result should be interpreted in the context of a well-documented statistical artifact: when ratings are concentrated within a narrow range, reliability coefficients such as Krippendorff's α lose discriminative power due to restricted variance (Gwet, 2014). In our case, the vast majority of ratings fell between 4 and 5 on the 5-point scale, meaning that raters broadly converged in judging dialogues as "Good" or "Excellent". The low α values therefore do not indicate disagreement but rather reflect insufficient rating dispersion for the coefficient to detect the convergence that is evident in the raw score distributions. Future evaluations could consider using a finer-grained scale (e.g., 7- or 10-point) or percent agreement metrics to better capture agreement under ceiling conditions.

Rater comments provide additional context for these results. For fluency, many notes recognized positively the language as "*smooth and grammatically correct*" and the explanations as "*easy to follow*". On the other hand, for naturalness, several dialogues were described as templated or slightly rigid. Specifically, raters described parts of the dialogue as "*robotic*", "*scripted*" or "*less natural*" due to frequent name repetition, abrupt topic shifts, or judgmental phrasing in single turns (e.g., "*Frequent name repetition makes it sound less natural.*"). For clarity and structure, participants consistently appreciated stepwise guidance and succinct, practical advice, with positive remarks highlighting clear sequencing such as "*first... then... next...*" and appropriate escalation when needed, as in "*suggesting an appointment with a specialist when symptoms persist*". A few comments also pointed to tone, endorsing supportive phrasing in many cases while suggesting that some questions could be softened. Overall, the human evaluation confirmed strong linguistic quality, but identified naturalness as the main area where improvement is needed.

5. Triple Extraction

Following conversation creation, we extracted Subject–Predicate–Object (SPO) triples [5] to provide structured representations for enriching CHIP's Domain Knowledge Graph.

[5] Here Subject-Predicate-Object do not represent syntactic dependencies but rather semantic relations in the Semantic Web paradigm.

5.1. Conversation Annotation

We frame conversational triple extraction (CTE) as token-level sequence labeling followed by triple assembly. Each token is annotated as *Subject*, *Predicate*, *Object*, or *other*, allowing multiple tokens per role and tokens participating in multiple triples.

Given medical content sensitivity, we used manual annotation for validation (882 sentences) and test sets (996 sentences), with automated methods for the training set (2953 sentences). The schema extracts semantically meaningful SPO triples from patient utterances, while agent responses were excluded as their content is predefined in the system's knowledge base. The annotation schema was developed to be robust enough to handle special cases, including passive voice, compound Subjects and Predicates, subordinate clauses, and implied elements. In addition to that, special rules were applied to question-answer sequences to avoid redundant labeling. For example, if an answer was merely affirmative, only the question was annotated; if the answer contained substantial information, only the answer was annotated; if both contributed relevant content, both were labeled. The full annotation schema can be found in the accompanying repository [1].

Manual annotation was conducted using the INCEpTION tool (Klie et al., 2018) with three linguistic experts independently annotating a subset of the data. Fleiss' Kappa of 0.657 indicated substantial agreement (Landis, 1977), with disagreements resolved through majority voting and expert review.

For the training set, annotation was performed automatically using structured prompting with GPT-4o. We reused the annotation prompt from the manual phase and applied a dynamic chunking strategy to handle question-answer pairs appropriately. The model was run at a low `temperature` (0.2) to ensure consistency and `max_tokens` of 4096 (default) to avoid truncation in longer dialogues. This approach allowed us to scale the annotation process efficiently while maintaining alignment with the manually labeled schema. The resulting corpus includes 4831 annotated sentences in total.

5.2. Methods

We evaluated three complementary approaches spanning symbolic baselines to state-of-the-art neural models, examining trade-offs between interpretable rules, general-purpose LLM prompting, and domain-adapted training. *Rule-based syntactic parsing* uses a dependency parser to identify predicates and attach Subjects/Objects, offering interpretability but sensitivity to conversational variation. *GPT-4o* applies few-shot demonstrations at `temperature` 0.2 without fine-tuning to test generalization. *Fine-tuned BERT* adapts `bert-base-uncased` (Devlin et al., 2019) with a linear token classification head and Optuna-tuned hyperparameters, achieving validation macro-F1 of 0.7563 at learning rate 2.209×10^{-5} and batch size 1.

	Token-level (macro)			Triple matching (F1)				
System	**P**	**R**	**F1**	**Full**	**Partial**	**Subj**	**Pred**	**Obj**
Rule-based	0.62	0.50	0.55	0.00	0.29	0.29	0.20	0.00
GPT-4o	0.72	0.65	0.68	0.35	0.64	0.58	0.43	0.43
BERT	0.73	0.63	0.67	0.26	0.50	0.47	0.39	0.32

Table 6: Token-level SPO classification and triple matching results.

5.3. Evaluation Metrics

We report macro-averaged precision, recall, and F1 at the token level, plus triple matching metrics: *full* (all components match), *partial* (two components match), and *per-component matches* (one component match), accounting for informative partial triples.

6. Discussion

Our evaluation reveals complementary strengths when comparing LLM-based and human assessments. Both consistently rated fluency and clarity as strong, reflecting grammatically correct and understandable dialogues. However, differences emerged for naturalness and structure. GPT-4o assigned higher naturalness scores (4.45 vs. 3.95), while human raters flagged repetitive phrasing, rigid tone, and name overuse as reducing realism. Conversely, humans valued structure more (4.54 vs. 4.14), appreciating stepwise guidance and escalation. Although inter-rater agreement coefficients were low, this largely reflects ceiling effects: most ratings clustered between 4 and 5, limiting variability while still showing broad convergence that the dialogues were judged as good or excellent.

Triple extraction experiments provide further insights. As shown in Table 6, the rule-based approach while interpretable, struggled with conversational variation and produced very low *full match* scores (0.00). Both neural methods substantially outperformed this baseline. GPT-4o achieved the best token-level (F1 = 0.68) and triple-level results, with strong *partial match* performance (0.64), capturing meaningful fragments even when full triples were incomplete. Fine-tuned BERT performed slightly below GPT-4o overall (token-level F1 = 0.67; *partial match* = 0.50) but showed robustness to domain-specific patterns, demonstrating that domain adaptation benefits annotation style capture. A common challenge across models was low *full match* accuracy (GPT-4o = 0.35; BERT = 0.26), underscoring the difficulty of aligning all three triple

components consistently in conversational data. When examining single-component matches, GPT-4o achieved the strongest results across Subject (0.58), Predicate (0.43), and Object (0.43), while BERT followed closely, suggesting that both models can reliably identify individual components even when full triple alignment is incomplete.

These results demonstrate that our synthetic dataset is linguistically validated and effective for benchmarking triple extraction, while highlighting the need for hybrid strategies: LLM prompting offers broad generalization, while domain-adapted models capture contextual nuances more reliably.

7. Conclusion

This paper presented a methodology for generating and annotating a synthetic conversational dataset for Type 2 Diabetes management. Using demographically and medically grounded personas, we produced 256 multi-turn dialogues incorporating personalization, empathy, contextual awareness, and medically accurate advice. The dataset was enriched with Subject–Predicate–Object (SPO) annotations through manual expert labeling and automated LLM support, yielding 4831 annotated sentences.

Our evaluation combined LLM-based scoring and human expert judgments, confirming high linguistic quality while identifying naturalness as the main area for improvement. We benchmarked three triple extraction approaches—rule-based syntactic parsing, GPT-4o, and fine-tuned BERT with a linear token classification head—with GPT-4o achieving the strongest performance.

The dataset, annotation schema, and baseline results provide a new resource for developing and evaluating the knowledge extraction component of Hybrid Intelligence systems such as CHIP, and a reproducible benchmark for conversational triple extraction in patient-centered agents.

8. Limitations

Our findings highlight important limitations of the dataset itself. Patient turns tend to be well-formed and elaborate, lacking the spelling errors, informal phrasing, repairs, and fragmented utterances typical of real conversations. This may inflate fluency and clarity scores while masking difficulties in handling noisy input. Systems trained on this data may therefore need additional adaptation to handle the misalignments and ungrammatical constructions that characterize authentic patient communication.

Moreover, although CHIP is intended for patients in the Netherlands, the dataset was generated in English rather than Dutch due to resource availability. This language choice facilitates evaluation with widely available models but limits immediate applicability to Dutch-speaking populations, where additional adaptation and validation would be required.

Finally, the dataset has been evaluated intrinsically—through linguistic quality assessments and triple extraction benchmarking—but has not yet been tested as training data within a deployed dialogue system. A natural next step would be to evaluate whether models trained on this data improve the accuracy of knowledge graph updates when deployed within the full CHIP pipeline during real or Wizard-of-Oz patient–agent interactions. This falls outside the scope of this paper but remains an important direction for future validation.

9. Ethics Statement

All dialogues used in this study were synthetically generated using large language models (LLMs) and do not contain any real patient information. No personally identifiable data were used or collected at any stage. The human evaluation involved voluntary participation of domain experts who provided consent for their anonymized responses to be analyzed in aggregate form. The study complies with the ethical principles of the participating institutions and adheres to the data protection requirements of the General Data Protection Regulation (GDPR).

10. Bibliographical References

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. 2021. Dpard: rationale, design and initial results from the dutch national diabetes registry. *BMC Endocrine Disorders*, 21(1):122.

Centraal Bureau voor de Statistiek. 2024. Gezondheid en zorggebruik; persoonskenmerken. CBS StatLine. Accessed: 2024-06-02.

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

Anshul Chavda and Pushpak Bhattacharyya. Synthetic dialogue data generation: A comprehensive survey of methods, evaluation, and challenges.

Floris den Hengst, Shaad Alaka, and Bart A. Kamphorst. 2026. Collaborative hybrid intelligence

platform chip: A modular architecture for developing and testing personalized lifestyle support interactions. *SoftwareX*, 33:102536.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Bernd JW Dudzik, Jasper S van der Waa, Pei-Yu Chen, Roel Dobbe, Íngo MDR de Troya, Roos M Bakker, Maaike HT de Boer, Quirine TS Smit, Davide Dell'Anna, Emre Erdogan, et al. 2024. Hybrid intelligence supports application development for diabetes lifestyle management. *Journal of Artificial Intelligence Research*, 80:919–929.

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Samuel Holmes, Anne Moorhead, Raymond Bond, Huiru Zheng, Vivien Coates, and Michael McTear. 2019. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? In *Proceedings of the 31st European conference on cognitive ergonomics*, pages 207–214.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural instructions: Tuning language models with (almost) no human labor. *arXiv preprint arXiv:2212.09689*.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

JR Landis. 1977. The measurement of observer agreement for categorical data. *Biometrics*.

Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, et al. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

OpenAI. 2024. Gpt-4 technical report.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, pages 68772–68802. Curran Associates, Inc.

Tessa F Peerbolte, Rozanne JA van Diggelen, Pieter van den Haak, Kim Geurts, Luc JW Evers, Bastiaan R Bloem, Nienke M de Vries, and Sanne W van den Berg. 2025. Conversational agents supporting self-management in people with a chronic disease: Systematic review. *J Med Internet Res*, 27:e72309.

Krithika Ramesh, Nupoor Gandhi, Pulkit Madaan, Lisa Bauer, Charith Peris, and Anjalie Field. 2024. Evaluating differentially private synthetic data generation in high-stakes domains. *arXiv preprint arXiv:2410.08327*.

Selene Báez Santamarıa, Lea Krause, Lucia Donatelli, and Piek Vossen. 2023. The role of personal perspectives in open-domain dialogue.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2023. Data augmentation for conversational ai. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5220–5223.

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

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

11. Language Resource References

Steven Bedrick and A. Seza Doğruöz and Sergiu Nisioi. 2025. *A Typology of Synthetic Datasets for Dialogue Processing in Clinical Contexts*.

Trisha Das and Dina Albassam and Jimeng Sun. 2024. *Synthetic Patient-Physician Dialogue Generation from Clinical Notes Using LLM*.

Demszky, Dorottya and Movshovitz-Attias, Dana and Ko, Jeongwoo and Cowen, Alan and Nemade, Gaurav and Ravi, Sujith. 2020. *GoEmotions: A dataset of fine-grained emotions*.

Eisenberg, Joshua and Sheriff, Michael. 2020. *Automatic extraction of personal events from dialogue*.

Ahmad Rezaie Mianroodi and Amirali Rezaie and Niko Grisel Todorov and Cyril Rakovski and Frank Rudzicz. 2025. *MedSynth: Realistic, Synthetic Medical Dialogue-Note Pairs*.

Wang, Junda and Yao, Zonghai and Yang, Zhichao and Zhou, Huixue and Li, Rumeng and Wang, Xun and Xu, Yucheng and Yu, Hong. 2024. *NoteChat: A Dataset of Synthetic Patient-Physician Conversations Conditioned on Clinical Notes*. Association for Computational Linguistics.

Yu, Dian and Yu, Zhou. 2019. *Midas: A dialog act annotation scheme for open domain human machine spoken conversations*.