Generating High Quality Synthetic Data for Dutch Medical Conversations

Cecilia Kuan, Aditya Kamlesh Parikh, Henk van den Heuvel
Center for Language Studies, Radboud University
Nijmegen, Netherlands
{cecilia.kuan, aditya.parikh, henk.vandenheuvel}@ru.nl

Abstract

Medical conversations offer insights into clinical communication often absent from Electronic Health Records. However, developing reliable clinical Natural Language Processing (NLP) models is hampered by the scarcity of domain-specific datasets, as clinical data are typically inaccessible due to privacy and ethical constraints. To address these challenges, we present a pipeline for generating synthetic Dutch medical dialogues using a Dutch fine-tuned Large Language Model, with real medical conversations serving as linguistic and structural reference. The generated dialogues were evaluated through quantitative metrics and qualitative review by native speakers and medical practitioners. Quantitative analysis revealed strong lexical variety and overly regular turn-taking, suggesting scripted rather than natural conversation flow. Qualitative review produced slightly below-average scores, with raters noting issues in domain specificity and natural expression. The limited correlation between quantitative and qualitative results highlights that numerical metrics alone cannot fully capture linguistic quality. Our findings demonstrate that generating synthetic Dutch medical dialogues is feasible but requires domain knowledge and carefully structured prompting to balance naturalness and structure in conversation. This work provides a foundation for expanding Dutch clinical NLP resources through ethically generated synthetic data.

Keywords: Synthetic Medical Dialogues, clinical NLP, Prompt Engineering

1. Introduction

Recent developments in Natural Language Processing (NLP) have greatly advanced text analysis, especially in the medical domain. Specifically, analyzing physician-patient conversations through clinical NLP can enrich research datasets and provide data-driven insights into patient-initiated concerns, which are often absent from Electronic Health Records (EHRs) (Weiner et al., 2024; Lalleye et al., 2020; Alshaikhdeeb et al., 2025). Such conversational data have also been shown to capture interactional detail and patient narratives valuable for prediction and research (Pyne et al., 2023), while potentially reducing clinician workload (Lukac et al., 2025; Zhang et al., 2024).

The performance of NLP models strongly depends on the size, quality, and domain alignment of their training data. However, access to healthcare datasets is restricted by privacy regulations such as the General Data Protection Regulation[1] (GDPR) (Marino et al., 2025; Gassenn et al., 2025), and anonymization processes can still risk potential re-identification (El Emam et al., 2009).

Our broader objective is to convert unstructured Dutch medical data, both audio and text, into structured resources that enable research, analysis, and interoperability. As shown in Figure 1, it involves multiple processing steps, including transcription, anonymization, medical entity recognition, and ontology mapping.

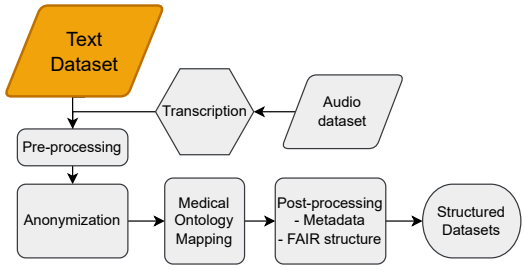

Figure 1: Broader Project Workflow - this study focuses on the Text Dataset component (highlighted)

Currently, our experiments rely on a small subset of audio–transcription pairs for controlled testing, which limits large-scale fine-tuning or comprehensive evaluation. This constraint reflects a broader challenge faced by the clinical NLP community: the scarcity of publicly available sensitive datasets (Hiebel et al., 2023).

Figure 2 illustrates the workflow of this study, which is motivated by the broader objective and focuses specifically on generating synthetic Dutch medical text dialogues as a resource to support and evaluate the pipeline. To mitigate data scarcity, generating high-quality synthetic data offers a practical alternative by enabling model fine-tuning, benchmarking, and system validation without the privacy constraints of real clinical data. Synthetic data can therefore serve as a privacy-compliant substitute

[1] https://gdpr-info.eu/

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10750–10763
11–16 May 2026.

for real datasets, supporting broader research collaboration and performance benchmarking in Dutch clinical NLP (Ive et al., 2020).

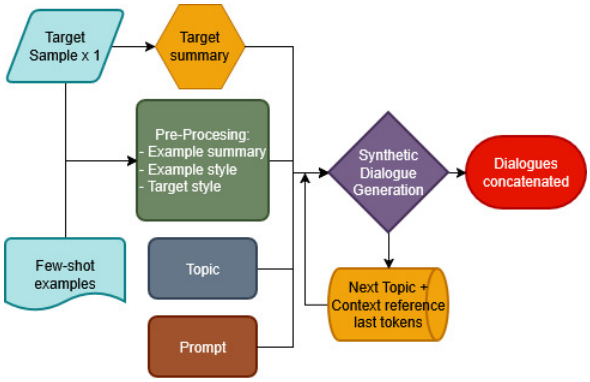

Figure 2: Synthetic Dialogue Text Generation Workflow

This paper presents a pipeline for generating synthetic Dutch medical text dialogues using a Large Language Model (LLM) fine-tuned with a Dutch dataset[2] by leveraging real clinical conversations as a linguistic and structural reference with two-shot samples for guidance. Our key research question is:

> To what extent can a Dutch LLM produce synthetic medical dialogues that match real-world data for supporting clinical NLP pipeline development?

To the best of our knowledge, no prior work has explored synthetic Dutch data generation using LLMs beyond clinical reports and EHR. Existing work on other languages and domains is reviewed in Section 2. The proposed pipeline and evaluation provide a novel framework for generating and assessing synthetic data that can advance Dutch clinical research and NLP development. By combining quantitative and qualitative evaluation, this work also explores the limits of automatic metrics in measuring linguistic quality, emphasizing the importance of human review in assessing complex linguistic patterns in natural conversations.

2. Related Work

One major challenge in clinical NLP is data scarcity and privacy concerns surrounding patient information. Synthetic medical data generation provides a valuable alternative under these constraints. This approach has been explored for synthetic medical dialogues and clinical notes in several studies (Das et al., 2024; Mianroodi et al., 2025). In addition, models developed using synthetic data have shown performance comparable to or exceeding models developed by real-world data (Mianroodi et al., 2025; Ive et al., 2020).

Previous work on synthetic text generation include synthetic medical data generation in English (Meoni et al., 2024; Melamud and Shivade, 2019; Ive et al., 2020; Naguib et al., 2024), medical dialogue generation in English and other languages except Dutch (Hiebel et al., 2023; Wang et al., 2023; Lozoya et al., 2024; ALMutairi et al., 2024; Mousavi et al., 2021), synthetic EHR generation in Dutch (Libbi et al., 2021), and automatic medical reporting (summary) in Dutch (Van Zandvoort et al., 2023). This review confirms that synthetic Dutch medical dialogue generation remains unexplored.

3. Methodology

In this section, we describe the experimental setup of synthetic Dutch medical dialogue generation in detail. Supporting materials are provided in the Appendices.

3.1. Data

In this work, we use a real-life dataset containing transcriptions of patient-doctor conversations in the nephrology domain from Nivel Institute[3]'s archive collection as target samples for the LLM to adapt linguistic features for text generation. The nephrology focus reflects the scope of the broader project, which targets this clinical specialty. The manual transcriptions were created as part of the HoMed (Homo Medicinalis) project[4]. Due to data protection constraints, only the textual transcriptions of the original audio recordings are accessible for research use.

The selected dataset originally contained nine transcriptions of nephrology consultations. Two files were used as few-shot examples, and the remaining seven were used as source material for synthetic dialogue generation. Several of these seven files were chunked into 1,000-token segments to accommodate the LLM's context window size, resulting in a total of nine usable text files. Each synthetic dialogue was modeled on the structure and linguistic style of its corresponding real dialogue, maintaining a one-to-one mapping between the reference and generated dialogues. File sizes range from 1,000 to 4,000 words and 170 to 650 speaker turns. The mean word count across target sample files is 2,708, and the mean number of turns is 396.

[2]Generated corpus available at: https://doi.org/10.34973/mvpm-9987

[3]https://nictiz.nl/

[4]https://homed.ruhosting.nl/publications

3.2. Model

Meta's Llama-3-8B-Instruct (Dubey et al., 2024)[5] was initially chosen for its open-source accessibility and suitability for secure, privacy-preserving experiments. The smaller model was used due to computing constraints. While early outputs showed some inconsistency and repetition, these observations motivated further adaptation toward domain-specific fine-tuning for improved dialogue quality.

To explore alternatives, we selected Llama-3-ChocoLlama-8B-Instruct (Meeus et al., 2024)[6] (ChocoLlama in this paper), an instruction-tuned model fine-tuned on Dutch-translated instruction datasets. This model also offers open-source accessibility and local deployment, addressing privacy constraints. According to testing results reported by Meeus et al. (2024), it is the best performing model in the ChocoLlama family. The base model, Llama-3-ChocoLlama-8B-Base, was adapted from Meta's Llama-3-8B and fine-tuned on an extensive corpus of native Dutch text totaling approximately 32 billion tokens. As an instruction-tuned model, ChocoLlama aims to improve instruction following and the coherence of Dutch outputs, helping to address the limited Dutch coverage in general multilingual LLMs. The model is released under the CC BY-NC 4.0 license, which permits non-commercial use only.

3.3. Experiment Setup

Figure 2 shows the workflow for synthetic Dutch medical dialogue generation, which involves preparing examples and controlled information to guide the LLM in generating dialogues that are contextually relevant and linguistically coherent. Major steps are as follows:

- Target sample summary generation and preprocessing - target sample summaries were generated as a reference for linguistic style and overall dialogue structure. Summaries reduced token consumption within the model's context window limit of approximately 8,000 tokens, leaving space for the prompt, few-shot examples, style specifications, and sliding window context. Bullet-point summaries also provided more structured guidance than raw dialogue text. Summaries were reviewed manually to ensure quality before using them for dialogue generation.

- Few-shot learning example preprocessing - two dialogue files were used for few-shot learning. For each file, an initial segment (approximately 400 tokens) was summarized as input, while a later, non-overlapping segment (approximately 1,200 tokens) served as output, with a 100-token gap to avoid overlap between segments. Input-output pairs were designed to demonstrate stylistic features - turn structure, tone, sentence length - rather than content coherence.

- Other controlled information - medical specialty domain (nephrology in this study) and four topics (*symptomen* / symptoms, *medicatiegebruik* / medical use, *leefstijl* / lifestyle, *laboratoriumuitslagen* / laboratory results) are provided for dialogue generation

- Prompt and above information are collected and provided to LLM for text generation. Prompt information is provided in Section 3.4.

- Dialogue generation - One dialogue will be generated with each topic along with the other information. In order to maintain contextual continuity, the last 150 words of the last generated dialogue will be passed onto the LLM as part of the instruction for next dialogue generation.

- Generated Dialogue concatenation - dialogues generated using given topics will be concatenated as the final dialogue.

3.4. Prompt Engineering

To identify an effective prompt, we first referenced the dialogue-generation prompt from Mianroodi et al. (2025) to draft an initial version. Prompt engineering typically requires several iterations and adjustments to reach the desired output quality (Zhou et al., 2022). Following the approach of Tang et al. (2023), we instructed the LLM to generate four variations and conducted a small-scale test to select the most suitable version. The selected version was then refined manually until it produced the intended responses.

The final prompt defined the LLM's role, speaker roles, the subject of the dialogue (medical domain, e.g., nephrology or oncology), a predefined list of conversational topics, limits on turn length, approximate total number of turns and words, a single sentence per turn (to mimic the target sample style), and integration of medical terminology. The prompt was designed to encourage professional yet natural dialogue suitable for clinical contexts. All prompts were written in Dutch. The complete prompt, which strongly influences dialogue quality and fluency, is provided in the Appendix 10.1.

[5] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[6] https://huggingface.co/ChocoLlama/Llama-3-ChocoLlama-8B-instruct

3.5. Evaluation Methods

To assess the quality of synthetically generated dialogues, both quantitative and qualitative evaluations were conducted. Quantitative analysis assessed dialogue-level characteristics of fluency and realism, including turn alternation, greeting and closing phrases repetition, role consistency, Average Sentence Length (ASL), average Sentences Per Turn (SPT), topic coverage, and lexical diversity. Qualitative analysis focused on the practicality and clinical usability of the generated dialogues through human evaluation.

Quantitative Analysis

Turn alternation rate and the repetition of greeting or closing phrases were used to identify potential structural errors and assess dialogue organization and fluency–important for understanding downstream usability. Alternation rate was calculated as the proportion of speak switches within each dialogue. A perfect rate of 1.0 indicates strict alternation, which is rare in real conversations where short interjections (e.g.,"yes", "no") and overlaps commonly occur (Clark and Tree, 2002; Shriberg, 2001).

Role consistency evaluated how well speaker roles were represented through their lexical choices, since physician-patient communication show systematic vocabulary differences between the two roles. Prior studies have found that physicians typically use more technical and domain-specific terminology, whereas patients more often describe symptoms, experiences, or emotional state (Whittaker, 2024; Schillinger et al., 2021; Pires and Cavaco, 2014). This distinct lexical difference formed the basis for evaluating role consistency through keyword matching.

Keyword matching was performed between the generated dialogues and predefined role-specific lexicons. Relevant vocabularies were extracted from the Dutch medical ontology, Nictiz and SNOMED International (2025), using procedure terms for physicians and clinical finding or symptom terms for patients. The most frequent items in nephrology conversation transcripts (unseen during text generation) were identified from these vocabularies, and the top 300 items were selected for each role. As no prior work has evaluated role consistency in synthetic Dutch medical dialogues, no direct benchmarks are available. Role consistency scores were interpreted relative to a heuristic range of 0.05–0.35. Keyword lists for doctors and patients are provided in the Appendix 10.2.

The ASL of the target sample was provided to the LLM as a reference, and prompts instructed one sentence per turn. ASL and SPT were measured to assess compliance with these structural instructions. Sentence length was defined as the number of words per sentence, ASL as the mean of all sentence lengths, and SPT as the mean number of sentences per turn.

Keyword-based evaluation, a common metric in Natural Language Generation, was adapted to assess topic generation since topics were explicitly defined in the prompt (Sun et al., 2023). Representative keywords were compiled for each topic, and their occurrences in the generated dialogues served as an evaluation metric. For medical topics such as *symptomen* (symptoms), *medicatiegebruik* (medication use), *laboratoriumuitslagen* (laboratory results), keywords were extracted from clinically validated Dutch SNOMED CT ontology by Nictiz and SNOMED International (2025)[7]. For the non-medical topic "*leefstijl*" (lifestyle), keywords were derived using semantic relations from the Open Dutch WordNet (Postma et al., 2016) by CLTL, Vrije Universiteit Amsterdam (2016). The compiled keyword lists are provided in the Appendix 10.2.

Lexical diversity measures linguistic richness, where excessive word repetition indicates reduced variety in natural conversations. The Type-Token Ratio (TTR), defined as the ratio of unique words to total words in a text (Rosillo-Rodes et al., 2025), was used as a baseline measure for lexical diversity. However, it is known to decline as repetition increases naturally in longer text (Bestgen, 2024). To address this limitation, the Mean Segmental Type–Token Ratio (MSTTR) was also computed, estimating average TTR across fixed-size windows within a text (Räsänen and Kocharov, 2025). A window length of 50 words was used to quantify lexical diversity in the generated medical dialogues. Given the mean dialogue length of 2,708 words, lower overall TTR values were expected, whereas specialized medical terminology may lead to relatively higher MSTTR scores.

Qualitative Analysis

Human evaluation assessed contextual richness, natural language use, lexical appropriateness, and clinical relevance–aspects that are difficult to quantify through automated metrics (Tam et al., 2024). Such evaluations are particularly valuable when no ground-truth or reference data are available for validation. A scoring rubric was used to standardize ratings across evaluators. The rubric, adapted from Fraile Navarro et al. (2025) and customized to evaluate synthetic medical dialogues in this study, included five categories: coherence, consistency, fluency, relevancy, and clinical use. Each was rated on a five-point scale, where 0 indicates no adherence and 5 indicates full adherence to the category criteria. The complete rubric is provided in the Ap-

[7]NL release 20250930, https://nictiz.nl/wat-we-doen/activiteiten/terminologie/snomed/, last visited: October 2025

pendix 10.3.

Five native Dutch-speaking reviewers participated in the evaluation, four of whom were medical practitioners. The non-medical reviewer did not score the clinical use category. For each aspect, mean and standard deviation scores were computed, and Inter-Rater Reliability (IRR) was assessed using Krippendorff's α(Krippendorff, 2011), which accommodates multiple raters and data types. Interpretation followed established guidelines (Krippendorff, 2018; Marzi et al., 2024), where $\alpha < 0$ indicates poor agreement, 0.00–0.20 slight, and 0.21–0.40 fair agreement.

Finally, correlations between quantitative and qualitative scores were examined using Spearman's rank correlation coefficient (ρ). Positive ρ values indicate alignment between automatic and human scores, whereas negative values denote inverse relationships.

4. Results and Discussion

4.1. Quantitative Results

We evaluated the generated dialogues using structural, lexical, and topic-based metrics. Figure 3 summarizes the distribution of word and turn counts across all generated dialogues.

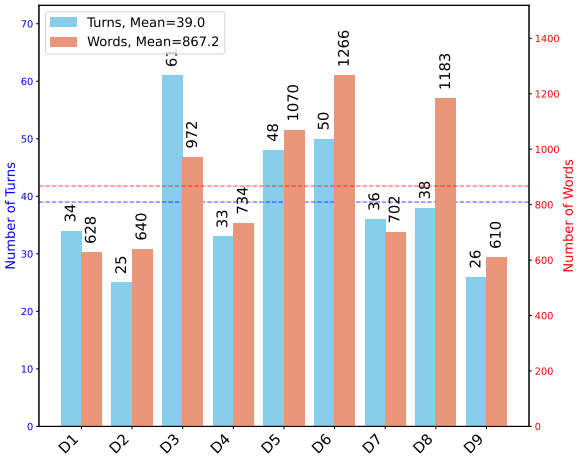

Figure 3: Synthetic Dialogue Statistics - Number of Words and Turns. D1–D9 denote individual-generated dialogues.

The summary of results is presented in Table 1.

Alternation Rate. The mean alternation rate across the nine generated dialogues is 0.973. The near-perfect score suggests that the generated dialogues may be overly structured and show scripted turn-taking behavior rather than a real-life conversational flow.

Greeting/Closing Detection. A total of 21 greeting occurrences and nine closing occurrences were detected across the nine generated dialogues. This high count, which exceeds the number of generated dialogues, suggests greetings were overused.

Metric	Mean	SD
Alternation rate	0.973	0.021
Role consistency	0.012	0.007
ASL	16.18	4.03
Average SPT	2.14	0.37
Topic coverage	0.889	0.132
TTR	0.377	0.04
MSTTR	0.834	0.01

Metric	Total Qty	
Greeting Detection	21	-
Closing Detection	9	-

Table 1: Quantitative Evaluation Summary

Role Consistency. Keyword matching between each role-specific lexicon and the generated dialogues resulted in a mean overlap score of 0.012. Figure 4 shows a box plot of role consistency scores across the nine dialogues. The clustering of scores within the blue box suggests relatively similar word choice for both doctor and patient across dialogues, with one outlier indicating that one dialogue contained comparatively more role-specific vocabulary. All scores fall below the heuristic baseline reference (gray shaded area).

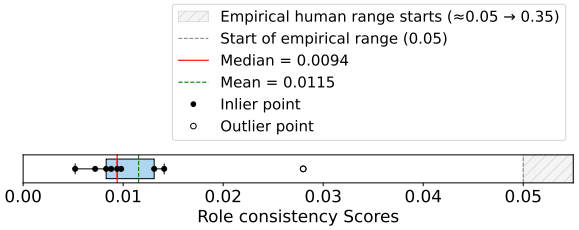

Figure 4: Role Consistency - Average per Dialogue

Figure 5 shows the proportion of lexicon match per role within each dialogue. Two dialogues contained no keyword matches for the patient role, while one dialogue showed a higher proportion of patient-specific words.

ASL and Average SPT. The ASL of each dialogue ranged between 15 to 20 words, with a mean of 16.18 words, notably higher than the target sample's ASL of seven words. This suggests that the model did not fully adhere to the instruction to produce shorter sentences. The mean SPT was 2.14, with six out of nine dialogues exceeding 1.5, suggesting deviations from the one-sentence-per-turn instruction. Figure 6 illustrates the ASL and SPT scores across generated dialogues.

Topic Coverage. Keyword matching achieved a mean topic coverage score of 0.889, with reasonable adherence to the prompt instructions. However, topic distribution varied considerably across

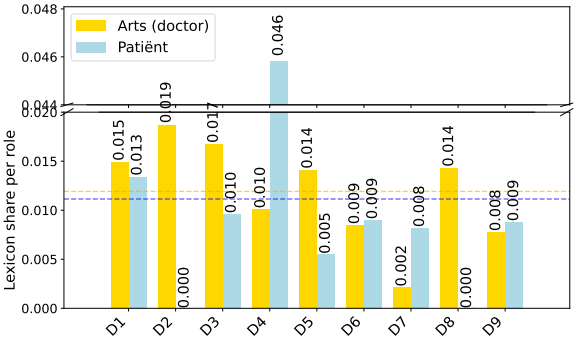

Figure 5: Role Consistency - Roles

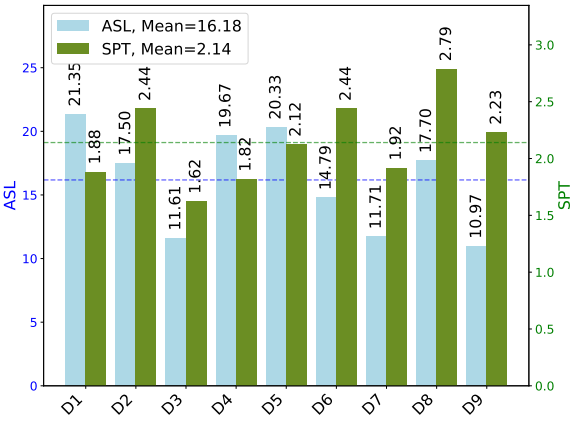

Figure 6: Scores: ASL and SPT

dialogues, see Figure 7. Discussions on laboratory results were missing in four dialogues, and with minimal coverage in three others, lifestyle and medication use dominated in many dialogues.

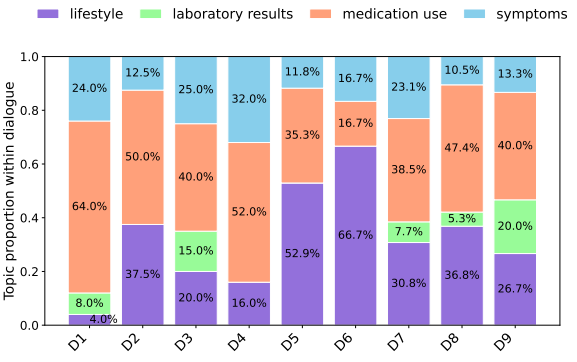

Figure 7: Topic Coverage - Proportion per Dialogue

Lexical diversity Figure 8 presents the TTR and MSTTR[8] for all generated dialogues. The scores were largely consistent across dialogues, with a mean TTR = 0.38 and MSTTR = 0.83. The relatively low TTR is expected, given the average dialogue length, as lexical repetition increases with text size, whereas the high MSTTR reflects local lexical variation introduced by specialized medical terminology.

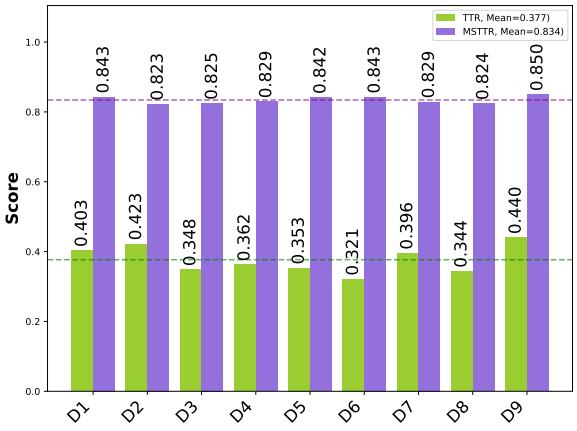

Figure 8: Scores: TTR & MSTTR

4.2. Qualitative Results

Human raters evaluated the dialogues across five qualitative categories. Figure 9 shows the overall score distribution from all raters, with each box representing a category. Scores generally range between two and three, with some exceptions. Coherence and Consistency both exhibit median scores of two, indicating that most ratings cluster around this value. The means for these categories are likely skewed by outliers, represented by red dots in the figure. In contrast, Relevance scores are more tightly concentrated around three, showing that this category received mostly average ratings. Clinical Use scores tend to be low, falling mostly between one and two, with an outlier reflecting a higher score that is not representative of the majority. Fluency received the most varied ratings, with 50% of data spread between two and four.

The heatmap in Figure 10 illustrates the mean score per category by each rater, reflecting potential biases, where C1–C4 denote native Dutch medical practitioners and N1 denotes a native Dutch speaker. Specifically, Rater C4 consistently scored higher for fluency and relevance compared to other raters.

To further understand rater C4's impact, Figure 11 compares IRR measured by Krippendorff's α with and without this rater. Aside from the Fluency category, inclusion of rater C4 had little effect on overall reliability. However, α values remain low

[8] Covington and McFall (2010) proposed the Moving-Average TTR (MATTR), which employs overlapping sliding windows and provides slightly higher precision than MSTTR. In this study, the difference between MSTTR and MATTR was minimal (0.003), and only MSTTR is reported

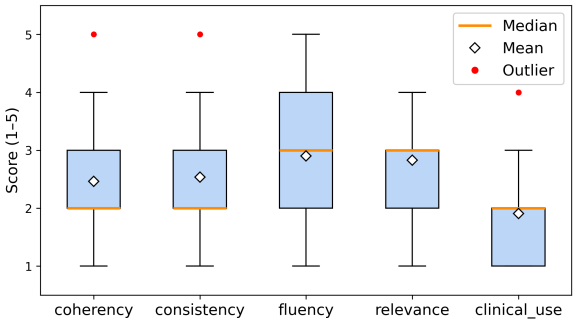

Figure 9: Human Evaluation - Scores per Category

across all categories (consistently below 0.12, with 60% of scores below zero), indicating substantial disagreement among raters.

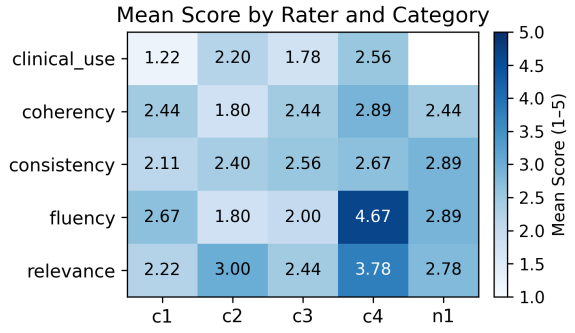

Figure 10: Human Evaluation - per Rater in each Category

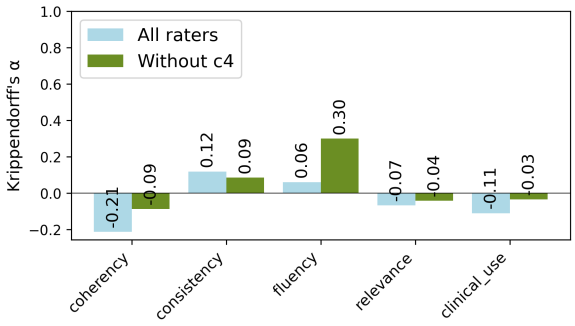

Figure 11: Human Evaluation - Inter Rater Reliability

4.3. Comparative Analysis and Discussion

This section integrates quantitative and qualitative findings to interpret how structural, lexical, and pragmatic aspects jointly influence dialogue quality. Overall, the results reveal that while the model followed the expected conversational structure defined by the prompt, it failed to reflect domain-specific language use and natural variations.

The near-perfect alternation rate suggests highly regular speaker switching, reflecting scripted rather than spontaneous interaction. Closer adherence to the one-sentence-per-turn (SPT) rule would likely lower this value and yield a more natural conversational flow. Frequent greetings, in contrast to the limited number of closings, likely stem from the generation pipeline where dialogues were concatenated from topic-based segments. These effects indicate an over-structured dialogue organization that may be potentially mitigated through prompt refinement or post-processing to smooth transitions. Longer ASL and higher SPT show that the model did not fully adhere to the prompt, representing a separate limitation from the structural overuse seen in alternation rate and greeting use.

Zero keyword matches for some dialogues, low role-specific lexicon overlap, and uneven topic coverage suggest that the model struggles to fully capture context-specific vocabulary. In contrast, uneven topic coverage mirrors real clinical dialogue patterns, where topic shifts do not always follow fixed patterns (Robinson et al., 2016; Ten Have, 2002).

Measure	D9	D3	D6
Length	610	972	1266
MSTTR	0.85	0.825	0.843
TTR	0.44	0.384	0.321
ASL	10.97	11.61	14.79
Turns	26	61	50
SPT	2.44	1.62	2.44

Table 2: Comparison of Selected Dialogues by Key Metrics

Table 2 compares three dialogues of varying lengths and their respective scores for MSTTR and ASL. It reveals that dialogue length alone does not determine lexical richness; shorter text achieved comparable MSTTR values to longer ones. It also implies that measured lexical diversity (measured by MSTTR) likely reflects the complexity of medical terminology rather than the role-specific vocabulary use.

It is important to note that lexical validation through context-specific lexicon lists and MSTTR captures only the presence of relevant terms, not their semantic correctness or contextual appropriateness. These measures may also reflect limitations in the lexicon sets or the inherent complexity of Dutch clinical language. Further qualitative analysis is therefore needed to assess contextual accuracy in synthetic dialogues.

Human ratings averaged 2.53, contrasting with high quantitative scores such as alternation rate and lexical diversity measured by MSTTR. Low IRR

(Figure 11) likely reflects rubric ambiguity and the subjective nature of evaluation, yet also signals the challenge of capturing conversational naturalness through quantitative measures.

Correlation analysis (Figure 12) confirmed weak alignment between automatic and human assessments: fluency and clinical use correlated moderately with MSTTR and role consistency, whereas relevance showed a negative relation (ρ = –0.31). Given the limited number of dialogues, these correlations should be interpreted with caution. These discrepancies highlight that numeric metrics capture pattern regularity but not semantic or pragmatic naturalness.

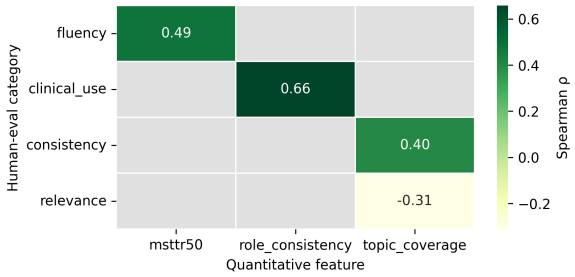

Figure 12: Spearman Correlation (ρ) - Qualitative vs Quantitative Scores

These quantitative-qualitative discrepancies echo rater comments, noting issues such as unclear domain focus ("not always clear about the subject being nephrology"), unnatural word choice resembling translated English, and inconsistencies in typical expressions ("errors in typical Flemish expressions"). Additional remarks include multiple greetings, abrupt openings or endings, and unnatural speaker turn-taking patterns different from real clinical conversations. The perception of translation-like output noted by reviewers likely stems from using a model fine-tuned on translated rather than naturally spoken data. Overall, these findings highlight specific weaknesses in the current pipeline - structural over-regularity, lack of domain vocabulary, and limited fluency.

5. Conclusion and Future Work

We proposed a pipeline leveraging a Dutch dataset-fine-tuned LLM to generate synthetic medical dialogues. In answer to our research question, findings indicate that while a Dutch LLM can feasibly produce synthetic medical dialogues that support clinical NLP pipeline development, the generated data do not yet match real-world dialogue quality. Achieving higher quality requires attention to model selection and prompt design, which strongly influence the linguistic features and overall quality of the generated dialogues. Models fine-tuned on translated datasets may negatively affect fluency, whereas overly structured prompts can lead to unnatural or rigid dialogues. These results contribute to ongoing efforts in clinical NLP to mitigate data scarcity and privacy constraints by utilizing realistic, language-specific synthetic datasets.

The main limitations of our study include the lack of domain-specific fine-tuning data, limited computational resources, and constraints on qualitative review process. Future work will focus on improving dialogue generation quality through refined prompt engineering and enhanced human evaluation protocols, including keyword refinement and inter-rater calibration. In addition, downstream development will explore synthetic audio dialogue generation and medical ontology mapping to further improve the realism and clinical usability of the generated dialogues.

Direct comparison with prior work was not possible, as, at the time of submission, no existing studies had addressed synthetic Dutch medical dialogue generation using LLMs. Cross-language comparisons (e.g., with English synthetic dialogue systems) were not pursued, as differences in language and model characteristics would obscure meaningful interpretation.

6. Ethical Considerations and Limitations

Synthetic medical dialogues offer an ethically aware alternative in contexts where data scarcity and privacy concerns restrict the development of clinical NLP. Such corpora can be shared in accordance with FAIR principles (Findable, Accessible, Interoperable, Reusable)[9], promoting data sharing and reproducibility without exposing sensitive information.

However, the high similarity observed across generated dialogues suggests that the model may have been adapting to its own synthetic output through iterative testing. Other potential risks include the replication of bias embedded in the training data and potential inaccuracies in clinical content. All dialogues are nephrology-specific, which may limit generalizability to other medical specialties.

Additionally, the few-shot design for dialogue generation prioritized stylistic features over content coherence—input and output segments were drawn from the same file, but likely covered different topics. This may have contributed to high scores on structural metrics (alternation rate, lexical diversity) but lower qualitative ratings for coherence and fluency.

[9]https://www.go-fair.org/fair-principles/

7. Acknowledgements

This research was supported by the MediSpeech project funded by ITEA4 under contract number 22032.

We thank qualitative evaluators - Amir Chaman Baz, Lex Dingemans, Sandra van Dulmen, Edwin Geleijn, Henk van den Heuvel - for their comments on synthetic dialogues, which have led to many findings and further improvements.

8. Bibliographical References

Mariam ALMutairi, Lulwah AlKulaib, Melike Aktas, Sara Alsalamah, and Chang-Tien Lu. 2024. Synthetic arabic medical dialogues using advanced multi-agent llm techniques. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 11–26.

Basel Alshaikhdeeb, Ahmed Abdelmonem Hemedan, Soumyabrata Ghosh, Irina Balaur, and Venkata Satagopam. 2025. Generation of synthetic clinical text: A systematic review. *arXiv preprint arXiv:2507.18451*.

Yves Bestgen. 2024. Back to basics in measuring lexical diversity: Too simple to be true. *Applied Linguistics*, 45(5):926–932.

Herbert H Clark and Jean E Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Michael A Covington and Joe D McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100.

Trisha Das, Dina Albassam, and Jimeng Sun. 2024. Synthetic patient-physician dialogue generation from clinical notes using llm. *arXiv preprint arXiv:2408.06285*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Khaled El Emam, Fida K Dankar, Régis Vaillancourt, Tyson Roffey, and Mary Lysyk. 2009. Evaluating the risk of re-identification of patients from hospital prescription records. *The Canadian journal of hospital pharmacy*, 62(4):307.

David Fraile Navarro, Enrico Coiera, Thomas W Hambly, Zoe Triplett, Nahyan Asif, Anindya Susanto, Anamika Chowdhury, Amaya Azcoaga Lorenzo, Mark Dras, and Shlomo Berkovsky. 2025. Expert evaluation of large language models for clinical dialogue summarization. *Scientific reports*, 15(1):1195.

Aline E Gassenn, Luis GM Andrade, Douglas Teodoro, and Jose F Rodrigues-Jr. 2025. Medical dialogue audio transcription: Dataset and benchmarking of asr models. In *Dataset Showcase Workshop (DSW)*, pages 71–82. SBC.

Nicolas Hiebel, Olivier Ferret, Karen Fort, and Aurélie Névéol. 2023. Can synthetic text help clinical named entity recognition? a study of electronic health records in french. In *The 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023)*.

Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf N Cardinal, Angus Roberts, Robert Stewart, and Sumithra Velupillai. 2020. Generation and evaluation of artificial mental health records for natural language processing. *NPJ digital medicine*, 3(1):69.

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

Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*, 4th edition. SAGE Publications, Inc., Thousand Oaks, CA.

Fréjus AA Laleye, Gaël de Chalendar, Antonia Blanié, Antoine Brouquet, and Dan Behnamou. 2020. A french medical conversations corpus annotated for a virtual patient dialogue system. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 574–580.

Claudia Alessandra Libbi, Jan Trienes, Dolf Trieschnigg, and Christin Seifert. 2021. Generating synthetic training data for supervised de-identification of electronic health records. *Future Internet*, 13(5):136.

Daniel Lozoya, Alejandro Berazaluce, Juan Perches, Eloy Lúa, Mike Conway, and Simon D'Alfonso. 2024. Generating mental health transcripts with sape (spanish adaptive prompt engineering). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5096–5113.

Paul J Lukac, William Turner, Sitaram Vangala, Aaron T Chin, Joshua Khalili, Ya-Chen Tina Shih, Catherine Sarkisian, Eric M Cheng, and John N Mafi. 2025. A randomized-clinical trial of two ambient artificial intelligence scribes: Measuring

documentation efficiency and physician burnout. *medRxiv*, pages 2025–07.

Simeone Marino, Ruth Cassidy, Joseph Nanni, Yuxuan Wang, Yipeng Liu, Mingyi Tang, Yuan Yuan, Toby Chen, Anik Sinha, Balaji Pandian, et al. 2025. Medical data sharing and synthetic clinical data generation–maximizing biomedical resource utilization and minimizing participant re-identification risks. *NPJ Digital Medicine*, 8(1):526.

Giacomo Marzi, Marco Balzano, and Davide Marchiori. 2024. K-alpha calculator–krippendorff's alpha calculator: a user-friendly tool for computing krippendorff's alpha inter-rater reliability coefficient. *MethodsX*, 12:102545.

Matthieu Meeus, Anthony Rathé, François Remy, Pieter Delobelle, Jens-Joris Decorte, and Thomas Demeester. 2024. Chocollama: Lessons learned from teaching llamas dutch. *arXiv preprint arXiv:2412.07633*.

Oren Melamud and Chaitanya Shivade. 2019. Towards automatic generation of shareable synthetic clinical notes using neural language models. *arXiv preprint arXiv:1905.07002*.

Simon Meoni, Éric De la Clergerie, and Théo Ryffel. 2024. Generating synthetic documents with clinical keywords: A privacy-sensitive methodology. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING 2024*, pages 115–123.

Ahmad Rezaie Mianroodi, Amirali Rezaie, Niko Grisel Todorov, Cyril Rakovski, and Frank Rudzicz. 2025. Medsynth: Realistic, synthetic medical dialogue-note pairs. *arXiv preprint arXiv:2508.01401*.

Seyed Mahed Mousavi, Alessandra Cervone, Morena Danieli, and Giuseppe Riccardi. 2021. Would you like to tell me more? generating a corpus of psychotherapy dialogues. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 1–9.

Marco Naguib, Xavier Tannier, and Aurélie Névéol. 2024. Few-shot clinical entity recognition in english, french and spanish: masked language models outperform generative model prompting. *arXiv preprint arXiv:2402.12801*.

Carla M Pires and Afonso M Cavaco. 2014. Communication between health professionals and patients: review of studies using the rias (roter interaction analysis system) method. *Revista da Associação Médica Brasileira*, 60(2):156–172.

Marten Postma, Emiel Van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. 2016. Open dutch wordnet. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 302–310.

Yvette Pyne, Yik Ming Wong, Haishuo Fang, and Edwin Simpson. 2023. Analysis of 'one in a million'primary care consultation conversations using natural language processing. *BMJ Health & Care Informatics*, 30(1):e100659.

Okko Räsänen and Daniil Kocharov. 2025. A pipeline for stochastic and controlled generation of realistic language input for simulating infant language acquisition. *Behavior Research Methods*, 57(10):275.

Jeffrey D Robinson, Alexandra Tate, and John Heritage. 2016. Agenda-setting revisited: When and how do primary-care physicians solicit patients' additional concerns? *Patient Education and Counseling*, 99(5):718–723.

Pablo Rosillo-Rodes, Maxi San Miguel, and David Sánchez. 2025. Entropy and type-token ratio in gigaword corpora. *Physical Review Research*, 7(3):033054.

Dean Schillinger, Nicholas D Duran, Danielle S McNamara, Scott A Crossley, Renu Balyan, and Andrew J Karter. 2021. Precision communication: Physicians' linguistic adaptation to patients' health literacy. *Science advances*, 7(51):eabj2836.

Elizabeth Shriberg. 2001. To 'errrr'is human: ecology and acoustics of speech disfluencies. *Journal of the international phonetic association*, 31(1):153–169.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. 2023. Evaluating large language models on controlled generation tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168, Singapore. Association for Computational Linguistics.

Thomas Yu Chow Tam, Sonish Sivarajkumar, Sumit Kapoor, Alisa V Stolyar, Katelyn Polanska, Karleigh R McCarthy, Hunter Osterhoudt, Xizhi Wu, Shyam Visweswaran, Sunyang Fu, et al. 2024. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ digital medicine*, 7(1):258.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*.

Paul Ten Have. 2002. Sequential structures and categorical implications in doctor–patient interaction: ethnomethodology and history. *Structure and Emergence of Professionalized" Praxis*.

Daphne Van Zandvoort, Laura Wiersema, Tom Huibers, Sandra van Dulmen, and Sjaak Brinkkemper. 2023. Enhancing summarization performance through transformer-based prompt engineering in automated medical reporting. *arXiv preprint arXiv:2311.13274*.

Junda Wang, Zonghai Yao, Avijit Mitra, Samuel Osebe, Zhichao Yang, and Hong Yu. 2023. Umass_bionlp at mediqa-chat 2023: Can llms generate high-quality synthetic note-oriented doctor-patient conversations? *arXiv preprint arXiv:2306.16931*.

Michael Weiner, Mindy E Flanagan, Katie Ernst, Ann H Cottingham, Nicholas A Rattray, Zamal Franks, April W Savoy, Joy L Lee, and Richard M Frankel. 2024. Accuracy, thoroughness, and quality of outpatient primary care documentation in the us department of veterans affairs. *BMC Primary Care*, 25(1):262.

P Whittaker. 2024. Do physicians and patients speak different languages: quantitative evidence from linguistic analysis. *European Heart Journal*, 45(Supplement_1):ehae666–3566.

Longxiang Zhang, Caleb D Hart, Susanne Burger, and Thomas Schaaf. 2024. Annotate the way you think: An incremental note generation framework for the summarization of medical conversations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1173–1186.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. In *The eleventh international conference on learning representations*.

9. Language Resource References

CLTL, Vrije Universiteit Amsterdam. 2016. *Open Dutch WordNet (ODWN)*. Computational Lexicology and Terminology Lab (CLTL), VU Amsterdam. CLTL, VU Amsterdam. PID http://wordpress.let.vupr.nl/odwn/.

Nictiz and SNOMED International. 2025. *SNOMED CT Netherlands Edition*. Nictiz (Netherlands Release Center). Nictiz, Netherlands Edition. PID https://nictiz.nl/wat-we-doen/activiteiten/terminologie/snomed/. National release of SNOMED CT for the Netherlands.

10. Appendices

10.1. Prompt Used For Synthetic Dutch medical Dialogue Generation

All prompts are written in Dutch.

System Prompt, Dutch:

> Je bent een behulpzame medisch onderzoeksassistent die Nederlandstalige medische dialogen genereert. Gebruik alleen 'Patiënt:' en 'Arts:' als sprekerlabels. Gebruik alleen algemeen aanvaarde medische feiten en vermijd het verzinnen van informatie. Corrigeer indien nodig misverstanden in het gesprek. Wees informatief en nauwkeurig binnen het kader van medische kennis. Ga naadloos door zonder gesprekken opnieuw te beginnen of te groeten. Geef geen introductie, vraag geen toestemming, onderbreek niet en produceer alleen dialoog. Geen uitleg, geen voorbeelden, geen meta-commentaar, geen samenvattingen.

System Prompt, English translation:

> You are a helpful medical research assistant who generates Dutch medical dialogues. Use only 'Patient:' and 'Doctor' as speaker labels. Use only generally accepted medical facts and avoid fabricating information. Correct misunderstandings in the conversation as needed. Be informative and accurate within the framework of medical knowledge. Continue seamlessly without restarting conversations or greeting others. Do not introduce yourself, ask for permission, interrupt, and only generate dialogue. No explanations, no examples, no meta-commentaries, no summaries.

User Prompt, Dutch:

> Schrijf een natuurlijke, informele dialoog tussen een patiënt en een art over nefrologie. De dialoog moet minstens 140 gespreksuitingen en ongeveer 1,000 woorden bevatten. Behandel hoofdonderwerpen uit de lijst: symptomen, medicatiegebruik, leefstijl, laboratoriumuitslagen. Gebruik minstens 5 relevante medische vaktermen. Geen opsommingen; integreer lijsten in vraag-en-antwoord. Laat onderwerpen natuurlijk overlopen, gebruik verduidelijkingen en herhalingen voor misverstanden. Het gesprek moet vloeiend zijn zonder herhaalde begroetingen of onderbrekingen. Vervolg het gesprek zonder het opnieuw te starten of scenario's opnieuw te introduceren. Speciale kenmerken: veel korte beurten (1–3 woorden), zoals 'ja', 'nee', 'oké', 'hm', afgewisseld met langere, informatieve antwoorden. Let op: GEEN uitleg, GEEN voorbeelden, GEEN meta-commentaar, GEEN introductie. Alleen de dialoog.

User Prompt, English translation:

> Write a natural, informal dialogue between a patient and a physician about nephrology. The dialogue should contain at least 140 utterances and approximately 1,000 words. Cover main topics from the list: symptoms, medication use, lifestyle, laboratory results. Use at least 5 relevant medical terms. No lists; integrate lists into questions and answers. Allow topics to flow naturally, use Clarifications and repetitions for misunderstandings. The conversation should flow smoothly without repeated greetings or interruptions. Continue the conversation without restarting it or reintroducing scenarios. Special features: Many short turns (1–3 words), such as 'yes', 'no', 'okay', 'hm', interspersed with longer, informative answers. Note: NO explanations, NO examples, NO meta-commentaries, NO introductions. Just the dialogue.

10.2. Quantitative Evaluation - Keyword Lists

Keyword Lists for Role Consistency Evaluation
Role - Doctors, Dutch

> chemotherapie, dialyse, operatie, immunotherapie, therapie, meten, bestralen, vasthouden, transplantatie, injectie, oefenen, hemodialyse, vervangen, voorlichting, vruchtwaterpunctie, bloedtransfusie, euthanasie, roesje, voorschrijven, vervolgonderzoek, fysiotherapie, ondersteuning, screening, discussie, verwijderen, eerste hulp, punctie, PET, conservatieve therapie, beleid, vaccinatie, infusie, voetverzorging, assisteren, bloedtest, evaluatie, voedingsadvies, aanpassing, delegeren, palliatieve zorg, prenatale screening, revisie.

Role - Doctors, English translation

> chemotherapy, dialysis, surgery, immunotherapy, therapy, measuring, irradiating, holding, transplantation, injection, exercising, hemodialysis, replacing, information, amniocentesis, blood transfusion, euthanasia, sedation, prescribing, follow-up examination, physiotherapy, support, screening, discussion, removal, first aid, puncture, PET, conservative therapy, policy, vaccination, infusion, foot care, assisting, blood test, evaluation, nutritional advice, adjustment, delegation, palliative care, prenatal screening, revision.

Role - Patients, Dutch

> pijn, probleem, ziekte, zwanger, diarree, downsyndroom, hoesten, koorts, speelt, misselijkheid, bevalling, stress, kanker, wil niet, dood, astma, hoest, allergie, geen klachten, bronchitis, trekt, lachen, vermoeidheid, schrijft, drinkt, gevoelig, buikpijn, slaapt, hoofdpijn, slaat, huilen, vloeibaar.

Role - Patients, English translation

> pain, problem, illness, pregnant, diarrhea, Down syndrome, coughing, fever, playing, nausea, childbirth, stress, cancer, doesn't want to, death, asthma, cough, allergy, no complaints, bronchitis, pulls, laughs, fatigue, writes, drinks, sensitive, stomach ache, sleeps, headache, hits, cries, fluid.

Keyword Lists for Topic Coverage Evaluation
Topic - symptomen / symptoms, Dutch

> pijn, hoesten, koorts, misselijkheid, hoest, vermoeidheid, buikpijn, hoofdpijn, spierpijn, keelpijn, vermoeid, duizeligheid, maagpijn, neuropathische pijn, brandende pijn, stekende pijn, aangezichtspijn, abdominale spierpijn, acromioclaviculaire gewrichtspijn.

Topic - symptomen / symptoms, English translation

> pain, cough, fever, nausea, cough, fatigue, abdominal pain, headache, muscle pain, sore throat, tired, dizziness, stomach pain, neuropathic pain, burning pain, stabbing pain, facial pain, abdominal muscle pain, acromioclavicular joint pain.

Topic - medicatiegebruik / medication use, Dutch

> medicijnen, kuur, medicijn, medicatie, pil, slikken, paracetamol, recept, dosering, tablet, antibiotica, antibioticum, ibuprofen, toediening, antibiotica-inhalatietherapie, antibiotische therapie, behandelen met bètareceptorantagonist, behandelen met erytropoëtinereceptoragonist, desensitisatiekuur door injectie.

Topic - medicatiegebruik / medication use, English translation

> medicine, treatment, drug, medication, pill, swallow, paracetamol, prescription, dosage, tablet, antibiotic, antibiotic, ibuprofen, administration, antibiotic inhalation therapy, antibiotic therapy, treatment with beta-receptor antagonist, treatment with erythropoietin receptor agonist, desensitization treatment by injection.

Topic - laboratoriumuitslagen / laboratory results, Dutch

glucose, vitamine d, hb, creatinine, cholesterol, bloedtest, kalium, bilirubine, celonderzoek, cervixcytologisch onderzoek, chromosoomonderzoek, crp, d-dimeer, ferritine, ft4, genetisch onderzoek, glucosetolerantietest, hba1c, natrium.

Topic - laboratoriumuitslagen / laboratory results, English translation

Glucose, vitamin D, HB, creatinine, cholesterol, blood test, potassium, bilirubin, cell analysis, cervical cytology, chromosome analysis, CRP, D-dimer, ferritin, FT4, genetic testing, glucose tolerance test, HB1C, sodium.

Topic - leefstijl / lifestyle, Dutch

suiker, gewicht, slaap, roken, suikerziekte, voeding, stress, dieet, wandelen, inspanning, beweging, zout, rook, oefenen, alcohol, afvallen, sporten, spanning, drank, ontspannen, drankje, afval, fysiek, spannen, aankomen, sport, overgewicht, wijn, bier, voedsel, sportman, voeden, oefening, borrel, sterke drank, lichaamsgewicht, ontspanning, gewichtsverlies, koolhydraat, slaperig.

Topic - leefstijl / lifestyle, English translation

sugar, weight, sleep, smoking, diabetes, nutrition, stress, diet, walking, effort, exercise, salt, smoke, practice, alcohol, lose weight, exercise, stress, drink, relax, drink, waste, physical, tense, gain weight, sport, overweight, wine, beer, food, athlete, nourish, exercise, drink, spirits, body weight, relaxation, weight loss, carbohydrate, sleepy.

10.3. Qualitative Evaluation Rubric

Coherency

Points	*Definition and Examples*
1-2	Frequent order/timing mistakes (non-sequential turns, confusing switches, inconsistent tense). Dialogue feels unnatural and hard to follow. Ex - doctor and patient turns are frequently mixed up, inserted after goodbye:: "Arts: Goed, laten we nu eens kijken naar wat u al doet en wat we kunnen verbeteren."
3	Mostly sequential and moderately coherent, but some abrupt or unnatural/unclear transitions between topics (e.g. sudden topic change, unclear turn boundaries). Ex - "Patiënt: Ik probeer elke twee uur een half kopje te drinken. Arts: En wat betreft uw voeding?"
4-5	Logically flows, correct turn-taking, consistent tenses, naturally progressing conversation; mostly coherent with minor jumps between topics without clear transition.. Ex - "Patiënt: Ik probeer elke twee uur een half kopje te drinken. Arts: En wat betreft uw voeding?"

Consistency

Points	*Definition and Example*
1-2	Frequent factual errors, hallucinated medical info, missing or contradictory details with previous statement and context. Lacks accuracy in medical facts, treatment options, or role actions. Ex - contains hallucinated diseases/treatment - "Arts: Je zegt dat je allergisch bent en Diazepam vijf keer per week neemt."
3	Some inconsistencies, minor errors or missing pieces, e.g., partial topic coverage, minor fact mistakes, not linking blood pressure to kidney function after discussing both. Ex - "Patiënt: Ik probeer mijn waterinname te verhogen, maar soms vergeet ik." (General lifestyle is relevant, but lacks specific detail or contradicts earlier plan.)
4-5	Consistent and accurate; mostly or all facts align with scenario and medical knowledge, topics covered.

Fluency

Points	*Definition and Example*
1-2	Use of Dutch is awkward or ungrammatical, regardless of content. Poor quality sentences, long sentences, repetition in sentences. Translations-like errors. Ex - "Het rug pijn is niet goed voelen en ik zijn moe altijd is van pijn."
3	Readable with occasional awkwardness, odd phrasing, or grammar slips; generally understandable. Ex - "Ja, ik ben nemen de pillen en ik voel minder, maar niet blij altijd."
4-5	Fluent, natural, and idiomatic Dutch; well-structured and appropriate medical register. Ex - "Ik neem ibuprofen en probeer rust te nemen. Maar het helpt niet echt."

Relevance

Points	*Definition and Example*
1-2	Major topics missing (symptoms, medication use, lifestyle, and laboratory results.), many irrelevant details, or missing clinical actions. Ex - "Patiënt: Ik houd van voetbal en mijn huisdier is een kat." (Unrelated information to context and/or of topics being discussed.)
3	Some or all target topics touched but detail may be shallow, or topic coverage uneven; some off-topic content. Lacks full coverage. Ex - "Arts: Wat betreft beweging, misschien kunt u een beetje meer wandelen."
4-5	Every target topic is addressed with sufficient detail; all info relevant to consultation. Ex - "Arts: We kunnen naar fysiotherapie of yoga kijken. Ook uw labresultaten wijzen op een geleidelijk probleem." (Addresses symptoms, medication, lifestyle, and lab results directly.)

Clinical Use

Points	*Definition and Example*
1-2	Dialogue is unrealistic for actual clinical setting (e.g., unsafe advice, fundamental errors, implausible behaviors). Ex - unsafe or implausible advice is given ex - "Arts: Neem gewoon meer pijnstillers, zoveel als u wilt, en de rest is niet belangrijk."
3	Could work in clinic with edits; mostly safe and realistic, but has notable weaknesses. Ex - "Arts: Probeer rustig te blijven en misschien wat meer water drinken." (Safe but incomplete; lacks specific clinical recommendation or next steps.)
4-5	Realistic, safe, plausible for clinical scenario; could be useful for annotation/training. Ex - Arts: We zullen uw bloeddruk monitoren en indien nodig medicatie voorschrijven. Houd een dagboek bij van uw symptomen en neem contact op als ze verergeren."