Synthetic Arabic Medical Dialogues Using Advanced Multi-Agent LLM Techniques

Mariam ALMutairi, Lulwah AlKulaib, Melike Yildiz Aktas, Sara Alsalamah, Chang-Tien Lu

Virginia Tech
Falls Church, VA, USA
malmutairi, lalkulaib , melike, salsalamah, clu @vt.edu

Abstract

The increasing use of artificial intelligence in healthcare requires robust datasets for training and validation, particularly in the domain of medical conversations. However, the creation and accessibility of such datasets in Arabic face significant challenges, especially due to the sensitivity and privacy concerns that are associated with medical conversations. These conversations are rarely recorded or preserved, making the availability of comprehensive Arabic medical dialogue datasets scarce. This limitation slows down not only the development of effective natural language processing models but also restricts the opportunity for open comparison of algorithms and their outcomes. Recent advancements in large language models (LLMs) like ChatGPT, GPT-4, Gemini-pro, and Claude-3 show promising capabilities in generating synthetic data. To address this gap, we introduce a novel Multi-Agent LLM approach capable of generating synthetic Arabic medical dialogues from patient notes, regardless of the original language. This development presents a significant step towards overcoming the barriers in dataset availability, enhancing the potential for broader research and application in AI-driven medical dialogue systems.

1 Introduction

Healthcare is a crucial aspect of daily life, involving coordination among multiple specialized departments such as pharmacy, imaging, labs, and billing, alongside primary care provider notes. Each patient interaction results in a clinical note documenting the discussions, medical conditions, and future plans, essential for continuous care and communication with patients and the healthcare team (Husmann et al., 2022). Unlike regular meeting summaries, clinical notes are semi-structured, including concise, bullet-pointed phrases with medical terminology and references to external data from electronic medical records (Krishna et al., 2020).

AI integration in healthcare promises significant improvements in efficiency and patient outcomes through enhanced data processing. In the Arab world, AI can bridge gaps in healthcare delivery by automating medical dialogue generation from clinical notes, reducing the documentation burden on healthcare providers. This automation ensures comprehensive and real-time medical records, enhancing patient care quality.

However, developing AI technologies for Arabic medical dialogue generation faces challenges. Arabic's morphological richness and dialect diversity complicate NLP tool development. The lack of publicly available medical datasets in Arabic and privacy concerns further impede progress, requiring robust data handling protocols.

Many medical dialogue datasets available in English and translating them is not sufficient to capture the nuanced context-specific expressions and cultural subtleties of regional Arabic dialects. Google Translate and similar automated systems primarily focus on Modern Standard Arabic (MSA) rather than regional dialects, which can result in the loss of these important linguistic features. These dialects encapsulate unique social contexts that automated systems often overlook, thereby diminishing the authenticity and effectiveness of the communication. Moreover, translating dialects through such systems can strip away vital cultural aspects that are intrinsic to understanding and using the language properly.

In contrast, advanced language models can adeptly adapt to various dialects, including the Najdi dialect, to generate dialogues that are both accurate and contextually appropriate. Utilizing the Najdi dialect in synthetic dialogues ensures that the interactions closely mirror real-life conversations, thus providing more realistic and practical training scenarios for medical professionals. The use of dialect-specific expressions and idioms can resonate more deeply with native speakers, enhanc-

Proceedings of The Second Arabic Natural Language Processing Conference, pages 11–26
August 16, 2024 ©2024 Association for Computational Linguistics

ing the clarity and empathy of medical communication compared to MSA. This approach not only improves the precision of conveyed information but also fosters a more genuine connection in medical interactions.

Recent efforts focus on advanced multi-agent large language models (LLMs) to generate synthetic Arabic medical dialogues. These models, trained on diverse datasets, achieve high linguistic and medical accuracy. Synthetic data generation offers a solution to the scarcity of Arabic medical training data, providing feasible alternatives to real conversations and aiding in training robust AI models. This approach ensures AI systems remain adaptable and continue evolving with medical advancements and changing healthcare practices.

Our work contributes significantly to the field of Arabic Natural Language Processing (NLP) and AI-driven healthcare solutions through several innovative approaches:

- **Development of a Multi-Agent LLM System:** We introduce a novel multi-agent large language model system that is specifically designed to generate synthetic Arabic medical dialogues. This system uses state-of-the-art techniques to ensure that the generated dialogues are linguistically and medically accurate, making it highly suitable for real-world healthcare applications.

- **Creation of Synthetic Medical Dialogue Datasets:** Addressing the acute shortage of publicly available medical datasets in Arabic, this work utilizes advanced AI to generate synthetic datasets from clinical notes. These datasets are crucial for training and improving AI models tailored to the Arabic language and healthcare needs.

- **Rigorous Evaluation of Generation Tasks:** This study establishes a comprehensive evaluation framework for the Generation tasks, using both quantitative metrics such as ROUGE and BERTScore and qualitative assessments by domain experts. This dual approach ensures that the generated conversations adhere to technical accuracy and maintain practical usability and relevance in clinical settings.

The generated medical dialogue dataset is accessible via Hugging Face [1]

2 Related Work

The adoption of electronic health records (EHRs) in the Arab world has enhanced health information availability and interoperability (HASANAIN et al., 2015; Zarouni et al., 2022), but increased documentation workloads for clinicians. Many find EHR documentation more time-consuming than traditional methods, leading to delays and incomplete records. The quality of electronic notes is often questioned due to readability and completeness issues, and the use of copy-and-paste practices (Michalopoulos et al., 2022). Clinicians sometimes dictate notes during patient visits, reducing patient interaction and perceived empathy (Schaaf et al., 2021). Employing medical assistants or scribes can help but requires significant investment and faces high turnover (Yan et al., 2016).

Automatic summarization technologies have gained interest due to improvements in speech-to-text technologies, widespread EHR implementation, and AI advancements, particularly transformer models (Ando et al., 2022). Early applications used statistical machine translation, recurrent neural networks (RNNs), and advanced transformer-based models (Jiang et al., 2021; Pilault et al., 2020; Egonmwan and Chali, 2019). Recent efforts in Arabic text summarization include tailored transformer models like AraGPT2 (Antoun et al., 2020b), AraBERT (Antoun et al., 2020a), AraT5 (Elmadany et al., 2022), and Arabic Pegasus (Alsuhaibani). These models process sequential data and capture Arabic language complexities, excelling in both extractive and abstractive summarization tasks.

The field faces challenges due to the lack of publicly available medical conversation datasets essential for training and evaluating these systems (Varshney et al., 2023). This lack of data stems from the personal and sensitive nature of medical recordings. While private entities may develop their datasets, it limits open comparison of algorithms and outcomes. Advances by large language models like GPT-4, Claude, and Gemini show promise, but the lack of common datasets restricts comprehensive evaluation in Arabic NLP (Al Oudah et al., 2019).

Specialized datasets have significantly advanced Arabic NLP, enhancing language understanding and processing. Ali et al. (2023) introduced the

[1] The generated medical dialogue dataset is available on Hugging Face through this https://huggingface.co/datasets/Mars203020/arabic_medical_dialogue

largest Arabic corpus, sourced from over 500 GB of text, improving language modeling and fine-tuning advanced Arabic models like those based on GPT-3 architecture. This corpus improved performance in benchmark tests by 4.5% to 8.5% over the multilingual BERT model (Koubaa et al., 2024). The AGS dataset by Atef et al. (2023) for abstractive text summarization, with 142,000 article-summary pairs, supports high-quality text summarization technologies.

In medical Arabic NLP, Hammoud et al. (2021) developed a dataset for Arabic medical text classification with 2,000 documents across 10 disease categories, supporting machine learning models for Arabic medical data (Mounsef et al., 2022). Despite advancements, challenges persist in dataset availability and accessibility, highlighting the need for collaborative efforts to develop and share domain-specific datasets publicly. This is crucial for applications like generating datasets from clinical notes to test Arabic text summarization (Chouikhi and Alsuhaibani, 2022).

3 Experiment

3.1 Data Source

For our Arabic medical dialogue generation, we used the ACI-Benchmark (ACI-bench) dataset Yim et al. (2023), known as the Ambient Clinical Intelligence Benchmark corpus. This dataset benchmarks automatic visit note generation from doctor-patient conversations and includes three components: Virtual Assistant (VirtAssist), Virtual Scribe (VirtScribe), and Ambient Clinical Intelligence (ACI).The ACI-bench dataset was created and validated by domain experts, including medical doctors, physician assistants, medical scribes, and clinical informaticians. Clinical notes were generated automatically and then checked and rewritten by experts to ensure accuracy. Data cleaning involved removing unsupported sentences and correcting transcription errors. The ACI-bench dataset includes conversation transcripts and corresponding clinical notes, partitioned into four sections: subjective, objective_exam, objective_results, and assessment_and_plan. This structure mirrors typical clinical documentation, making it relevant for practical applications. The dataset is the largest publicly available for model-assisted clinical note generation, with 207 dialogue-note pairs.

We used the ACI-Benchmark as an English baseline to generate synthetic Arabic medical dialogues based on the clinical notes. This allowed us to compare our models' outputs against a well-established dataset, ensuring high-quality Arabic dialogues evaluated for medical accuracy, communication effectiveness, and adherence to clinical documentation standards.

Despite its strengths, the ACI-bench dataset has limitations. It was produced synthetically by a limited number of content creators, which may not fully capture the diversity of health topics, speech variations, and note formats in real-world settings. This highlights the need for more representative samples to improve AI-assisted note-generation systems.

In conclusion, the ACI-bench dataset is a significant contribution to AI-assisted clinical note generation, providing a robust benchmark for evaluating generative models in healthcare. Future research can build on this dataset to enhance the accuracy and efficiency of automatic clinical note-generation systems.

3.2 Dialogue Generation

To generate synthetic Arabic medical dialogues in the Saudi Najdi dialect, we implemented a Multi-Agent system using Claude-3-Opus-20240229 (Anthropic, 2024) and GPT-4 (OpenAI et al., 2024). This sophisticated approach aims to produce high-quality, culturally relevant medical conversations that can serve as valuable tools for training and educational purposes within the medical field. Each agent in the system plays a crucial role in ensuring the final dialogues are comprehensive, accurate, and engaging, adhering to both medical standards and cultural nuances. Sample data from the generated dataset demonstrate the effectiveness of our approach and are presented in Figure 2 in the appendix.

Generation Agent: The process begins with the Generation Agent, where an ACI-benchmark clinical note in English serves as the foundation for creating the Arabic medical dialogue. Using Claude-3-Opus-20240229, we prompt the model to generate conversation in Arabic, specifically tailored to the Saudi Najdi dialect. We chose Claude-3-Opus because it performed better with the Najdi dialect compared to GPT-4, based on our evaluation of a small sample set. The prompt (details in Appendix A.1) includes detailed instructions to ensure the conversation covers approximately 50 exchanges and around 3000 words. The generated dialogue starts with a warm greeting from the doctor and continues with inquiries into the patient's current condition and specific concerns. Key aspects of the clinical note are discussed, including the patient's chief complaint, history of present illness, current medications, past medical history, past surgical history, examination findings, results from any diagnostic tests, and the proposed treatment or follow-up plan. The conversation aims to be clear, culturally sensitive, and engaging, avoiding medical jargon and instead using layman's terms to explain medical terms and findings. Small conversational elements and culturally relevant expressions are integrated to enhance the natural flow and authenticity of the conversation.

Improvement Agent: Once the initial conversation is generated, it is passed to the Improvement Agent. In this stage, the previously generated conversation and the clinical note are further refined to enhance its depth and educational value. The prompt (details in appendix A.2) for this node emphasizes clarifying medical terms and diagnostic results comprehensively, integrating realistic conversational flow, and encouraging the patient to describe their medical history, symptoms, and current medications in accessible language. The conversation is restructured to explore each medical topic thoroughly, ensuring logical flow and consistency. Empathy and professionalism are demonstrated throughout, making the patient feel understood and supported. The dialogue is expanded to include discussions on lifestyle impacts, potential complications, and preventive measures, all while maintaining cultural sensitivity and minimizing repetition. This enhancement process ensures that the conversation is not only informative but also engaging and supportive, reflecting a high standard of medical professionalism.

Evaluation Agent: The final step involves the Evaluation Agent, where the enhanced dialogues are assessed for quality and accuracy. This evaluation is conducted using two advanced models: Claude-3-Opus-20240229 and GPT-4-Turbo-2024-04-09. Each dialogue is evaluated based on multiple criteria in Table 1, including medical accuracy and completeness, communication, and rapport, structure and flow, language and terminology, and patient engagement and education. For a chat to be accepted, it must achieve an average score of above 4.5 out of 5 from both models, which is 90%. This rigorous evaluation process ensures that only the highest quality dialogues, which meet stringent standards of medical accuracy, cultural relevance, and patient engagement, are approved for use. The high threshold of 4.5/5 is chosen to ensure exceptional quality, minimize errors, and maximize the educational value and practical applicability of the dialogues. The result is a collection of synthetic Arabic medical dialogues that are valuable resources for medical training and education, helping to improve communication skills, cultural competence, and overall patient care in the medical field. the details of the evaluation prompt is in appendix A.2.1.

Evaluation Criteria for Medical Conversation and Clinical Note Similarity
1. Medical Accuracy and Completeness
1.1. Symptoms and Complaints
1.2. Medical History
1.3. Diagnosis and Treatment
2. Communication and Rapport
2.1. Active Listening
2.2. Clarity and Explanations
2.3. Empathy and Respect
3. Structure and Flow
3.1. Logical Progression
3.2. Transitions and Coherence
3.3. Time Management
4. Language and Terminology
4.1. Appropriate Language
4.2. Medical Terminology
4.3. Cultural Sensitivity
5. Patient Engagement and Education
5.1. Patient Participation
5.2. Patient Education
5.3. Addressing Concerns

Table 1: Evaluation Criteria for Medical Conversation and Clinical Note Similarity

4 Automatic Evaluation

To evaluate the quality and accuracy of the generated Arabic medical dialogues, we employed two primary metrics: ROUGE (Lin, 2004) and BERTScore Zhang et al. (2020). Additionally, we compared the generated Arabic results with the ACI-benchmark dataset. The following sections detail the evaluation criteria and processes used.

4.1 ROUGE

The Multilingual ROUGE Score (MRouge), as developed by Hasan et al. (2021), assesses the similarity between the generated conversation and the reference text. For evaluating Arabic texts, the clinical note was translated using Google Translate, and a medical student subsequently reviewed the translation to ensure accuracy. MRouge facilitates the measurement of extractiveness metrics, thereby demonstrating how much information in the dialogue is derived from the clinical note.

4.2 BERTScore

BERTScore is used to evaluate the quality of the generated dialogues by comparing them to the reference text. We used the model `xlm-roberta-large` by Conneau et al. (2020) since it is multilingual and supports both Arabic and English. BERTScore measures the semantic similarity between the generated conversation and the reference, ensuring that the output maintains the intended meaning and context.

4.3 Factuality Evaluation

To ensure the factual accuracy of the generated dialogues, we utilized the GPT-4-Turbo and Claude-3-Opus models, aiming to derive information from the clinical note rather than allowing the model to hallucinate content. The evaluation criteria, as detailed in Table 1, are divided into five main categories.

The first category, Medical Accuracy and Completeness, assesses symptoms and complaints, medical history, and diagnosis and treatment, ensuring all medical information is accurately and thoroughly addressed. The second category, Communication and Rapport, focuses on active listening, clarity and explanations, and empathy and respect, which are crucial for establishing a strong doctor-patient relationship. The third category, Structure and Flow, evaluates logical progression, transitions and coherence, and time management, ensuring the dialogue flows naturally and efficiently covers all necessary topics.

The fourth category, Language and Terminology, assesses the use of appropriate language, medical terminology, and cultural sensitivity, ensuring the conversation is accessible to the patient and respects cultural nuances. Finally, Patient Engagement and Education evaluates patient participation, patient education, and addressing concerns, ensuring the patient is actively involved and understands their medical situation.

Each criterion is scored on a scale from 0 to 5, with detailed comments provided for comprehensive feedback. Scores are averaged to calculate the overall scores for the five main categories. The final evaluation score is determined by averaging these five overall scores, providing a comprehensive measure of the dialogue's quality and effectiveness. This structured approach helps maintain high standards in the generated medical dialogues, fostering accurate, empathetic, and clear communication between doctors and patients.

5 Human Evaluation

The expert evaluation involved five medical practitioners assessing whether machine-generated conversations matched real patient-physician encounters, with a focus on medical common sense, knowledge, and logic. Each evaluator was assigned 10 generated conversations and reviewed them against the corresponding clinical notes. They rated the dialogues using a Likert scale (1-5), as outlined in Table 2.

The criteria for evaluation included whether the conversation's symptoms, complaints, medical history, diagnosis, and treatment were consistent with the clinical notes. Additionally, they assessed the clarity and smoothness of the conversation's structure and flow. Evaluators also considered if the conversation provided comprehensive answers to the patient's queries or concerns and offered appropriate patient advice. They examined adherence to ethical guidelines and standards for health communication and evaluated whether the language and medical terminology were suitable for the patient. Effective communication and rapport, demonstrated through active listening, clear explanations, empathy, and respect, were also key factors.

In addition to the ratings, evaluators provided feedback on what they liked and disliked about the dialogues and suggested areas for improvement. This feedback was used to determine the overall quality and effectiveness of the machine-generated dialogues, ensuring they meet professional standards for patient-physician interactions.

6 Results

6.1 ROUGE Scores

The ROUGE scores, presented in Table 3, compare the performance of our multi-agent LLM in generating Arabic medical dialogues against the English dialogues from the ACI-Benchmark dataset. The ROUGE-1 score for Arabic-generated dialogues was 0.236, while for English dialogues, it was 0.341. ROUGE-2 scores were significantly lower

Evaluation Criteria
1. The symptoms, complaints, medical history, diagnosis, and treatment in the conversation are based on the clinical note.
2. The conversation follows a clear structure and flows smoothly.
3. The conversation provides comprehensive answers to the patient's queries or concerns and offers appropriate patient advice.
4. The conversation adheres to ethical guidelines and standards for health communication.
5. The language and medical terminology used are appropriate for the patient.
6. The conversation demonstrates effective communication and rapport by actively listening to the patient's concerns, providing clear and thorough explanations, and showing empathy and respect.

Table 2: Evaluation Criteria for Health Conversations

for both languages, with Arabic scoring 0.042 and English 0.134. The ROUGE-L score, which measures the longest common subsequence, was 0.122 for Arabic and 0.199 for English. ROUGE-Lsum, which focuses on the summary level, showed scores of 0.220 for Arabic and 0.319 for English.

Arabic		English	
Metric	Value	Metric	Value
ROUGE1	0.236	ROUGE1	0.341
ROUGE2	0.042	ROUGE2	0.134
ROUGEL	0.122	ROUGEL	0.199
ROUGELsum	0.220	ROUGELsum	0.319

Table 3: Comparison of ROUGE scores between Arabic and English Generated Dialogue Datasets

6.2 BERT Scores

The BERT scores for the generated dialogue datasets reveal significant insights into the similarity between different language pairs. Using the English-English pair, which compares the English notes with the English dialogues in the ACI dataset, as the baseline with a score of 0.819, the Arabic-Arabic pair achieved a higher similarity score of 0.834, indicating strong semantic consistency in the generated Arabic dialogues. The Arabic-English pair, which measures cross-lingual similarity, obtained a score of 0.809, demonstrating a slightly lower but comparable level of semantic alignment between the Arabic and English dialogues. These results, as presented in Table 4, underscore the effectiveness of the generation process in maintaining semantic consistency across both monolingual and cross-lingual comparisons.

6.3 Evaluation Using LLMs Results

Table 5 shows the evaluation results comparing Claude-3-Opus and GPT-4-Turbo. The evaluation criteria Table included Medical Accuracy and Completeness, Communication and Rapport, Structure and Flow, Language and Terminology, and Patient Engagement and Education. Claude-3-Opus achieved an overall score of 4.93, while GPT-4-Turbo scored a perfect 4.99.

BERT Scores	
Metric	Value
Arabic - Arabic	0.834
Arabic - English	0.809
English - English	0.819

Table 4: BERT-Scores for Arabic and English Generated Dialogue Datasets

6.4 Human Evaluation Results

The human evaluation of our system for generating Arabic clinical conversations from English clinical notes using a Multi-Agent Large Language Model (LLM) showed strong performance (see Figure 1). Most evaluators (36 strongly agree, 9 agree) confirmed that the conversations accurately reflected the clinical notes. The structure and flow were highly rated, with 40 strongly agreeing and 4 agreeing. In terms of comprehensiveness, 46 evaluators strongly agreed that the conversations provided thorough answers and appropriate advice. Adherence to ethical guidelines was confirmed by 42 strongly agreeing and 5 agreeing. Language and medical terminology were deemed appropriate by 35 strongly agreeing and 5 agreeing, though a few indicated room for improvement. Effective communication and rapport were also strong points, with 44 strongly agreeing and 2 agreeing. These results highlight the efficacy of our multi-agent LLM system while identifying areas for refinement to ensure greater consistency and quality

	Claude-3-Opus	GPT-4-Turbo
Medical Accuracy and Completeness	4.77	4.99
Communication and Rapport	5.00	4.99
Structure and Flow	4.91	5.00
Language and Terminology	5.00	5.00
Patient Engagement and Education	4.97	5.00
Overall Score	**4.93**	**4.99**

Table 5: Comparison of Claude-3-Opus and GPT-4-Turbo Evaluation Results

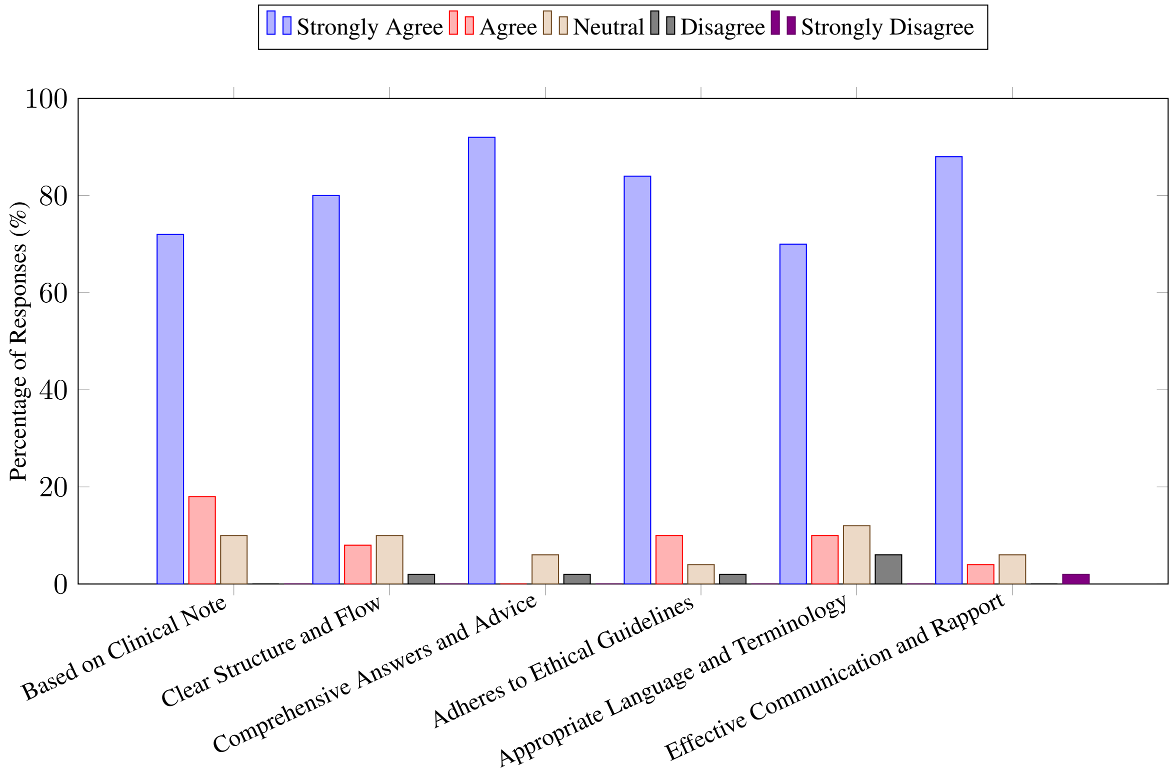

Figure 1: Human Evaluation Results of Arabic Clinical Conversations

7 Discussion

The results demonstrate the performance of our multi-agent LLM system in generating Arabic medical dialogues when compared to English dialogues from the ACI-Benchmark dataset.

7.1 ROUGE Scores Analysis

The lower ROUGE-2 scores in both languages indicate challenges in generating precise bi-gram matches, which might be attributed to the complex medical terminology and the structure of dialogues. The significant gap between ROUGE-1 and ROUGE-2 suggests that while the system can generate relevant words, it struggles with the accurate construction of phrases. The disparity between ROUGE-L and ROUGE-Lsum scores for both languages also suggests that while individual sentences may be relatively well-constructed, the overall summary coherence is lacking. Although ROUGE is the de facto metric for similar tasks, it has several limitations (Schluter, 2017). Achieving optimal extractive summarization with respect to ROUGE is NP-hard, making it difficult to find the perfect summary. Additionally, perfect 100% ROUGE scores are unattainable for higher-quality datasets because of averaging across multiple reference summaries, and there is no clear understanding of what constitutes a perfect score with ROUGE. Furthermore, humans cannot achieve perfect ROUGE scores due to diversity in content selection among different human-generated summaries. Therefore, ROUGE should be used with caution and in conjunction with other evaluation metrics to provide a more comprehensive assessment of generation quality.

7.2 BERT Scores Analysis

The BERT-Scores highlight the Multi-Agent LLM's ability to generate internally consistent dialogues within Arabic, with the Arabic-Arabic score being the highest among the comparisons. This indicates that the model is adept at maintaining contextual relevance and semantic coherence within the same language. The cross-language comparison score (Arabic-English) indicates that while the model can translate and adapt dialogues, there is a notable decrease in similarity, suggesting room for improvement in cross-linguistic translation and contextual understanding. This decrease in similarity may stem from differences in linguistic structures, idiomatic expressions, and cultural context that are not perfectly captured by the model.

7.3 Evaluation Using LLM Results Comparison

The evaluation results comparing Claude-3-Opus and GPT-4-Turbo further underline the strengths and areas for improvement of the multi-agent LLM system. Claude-3-Opus achieved high scores across all categories, with an overall score of 4.87, indicating a strong performance in generating medical dialogues. However, GPT-4-Turbo achieved a perfect score in all categories, highlighting its superior performance in generating accurate, engaging, and contextually appropriate medical dialogues.

Medical Accuracy and Completeness: Claude-3-Opus scored 4.71, while GPT-4-Turbo scored 4.99 in this category. This difference indicates that while both models perform well in ensuring medical accuracy, GPT-4-Turbo has a slight edge in the completeness and precision of the medical information provided. This could be attributed to GPT-4-Turbo's larger training dataset or more sophisticated algorithms for medical information retrieval and synthesis.

Communication and Rapport: In the Communication and Rapport category, Claude-3-Opus scored 4.94, compared to GPT-4-Turbo's 4.99. Both models show strong performance in this area, indicating their capability to generate dialogues that foster a positive patient-provider relationship. However, GPT-4-Turbo's slight advantage may be due to better handling of empathy, tone, and conversational flow, which are critical in medical dialogues.

Structure and Flow: Claude-3-Opus scored 4.86 in Structure and Flow, while GPT-4-Turbo achieved a perfect score of 5.00. This indicates that GPT-4-Turbo produces more logically structured and smoothly flowing dialogues. The multi-agent LLM system could benefit from enhancements in maintaining the narrative structure and logical progression of medical dialogues to match GPT-4-Turbo's performance.

Language and Terminology: Both models performed exceptionally well in the Language and Terminology category, with Claude-3-Opus scoring 4.94 and GPT-4-Turbo scoring 5.00. This suggests that both models are adept at using appropriate medical terminology and maintaining clarity and precision in language. GPT-4-Turbo's slight advantage could be due to more extensive training on medical texts and better contextual understanding.

Patient Engagement and Education: Claude-3-Opus scored 4.91 in Patient Engagement and Education, while GPT-4-Turbo scored 5.00. This category assesses how well the models can engage patients and provide educational information. GPT-4-Turbo's perfect score indicates a superior ability to generate dialogues that are not only informative but also engaging and reassuring for patients.

7.4 Human Evaluation Analysis

The human evaluation of our system for generating Arabic clinical conversations from English clinical notes using a Multi-Agent Large Language Model (LLM) revealed several strengths and areas for improvement. Evaluators confirmed that the conversations accurately reflected the clinical notes, with most strongly agreeing or agreeing on this aspect. The conversations were generally perceived as having a clear structure and flow, and they provided comprehensive answers and appropriate advice.

However, evaluators identified specific areas for improvement. For instance, there were instances of inaccuracies, such as a patient mentioning that their right knee hurt more, whereas the notes indicated bilateral pain. Another significant area for improvement is the simplification of medical terminology. Evaluators noted that some terms were difficult for patients to understand and needed simplification in three generations. This complexity in language likely stems from the model's translation process, which wasn't trained on the real-life usage of these terms. Despite these challenges, the system demonstrated effective communication and rapport, particularly in listening to and addressing patient concerns accurately. The brief and to-the-point nature of the conversations was appreciated, though some evaluators suggested that additional context, such as mentioning the potential need for antibiotics depending on test results, would enhance the usefulness of the conversations.

Overall, while the system shows promise in generating accurate and contextually appropriate Arabic medical dialogues, attention to simplifying medical terminology and ensuring consistent and comprehensive information throughout the conversation is essential. Addressing these issues will help enhance patient understanding and improve the overall effectiveness of the generated dialogues.

8 Conclusion

The study explores using a multi-agent LLM system to generate Arabic medical dialogues from English clinical notes. While promising in generating accurate and engaging dialogues, further refinement is needed. Comparing these dialogues with English ones from the ACI-Benchmark dataset and evaluation results from Claude-3-Opus and GPT-4-Turbo highlights areas for improvement, especially in translation and contextual consistency. A limitation is the synthetic nature of the ACI-Benchmark dataset, which may not reflect real-world clinical diversity. Future research should address these issues by developing a cross-lingual similarity measure for longer texts and medical terms in Arabic. We are making the dataset publicly available and expanding it with other clinical notes to ensure a more comprehensive and representative dataset. Improving these aspects will enhance patient-provider communication in Arabic-speaking regions.

References

Abrar Al Oudah, Kholoud Al Bassam, Heba Kurdi, and Shiroq Al-Megren. 2019. Wajeez: An extractive automatic arabic text summarisation system. In *Social Computing and Social Media. Design, Human Behavior and Analytics: 11th International Conference, SCSM 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26-31, 2019, Proceedings, Part I 21*, pages 3–14. Springer.

Abbas Raza Ali, Muhammad Ajmal Siddiqui, Rema Algunaibet, and Hasan Raza Ali. 2023. A large and diverse arabic corpus for language modeling. *Procedia Computer Science*, 225:12–21.

Mohammed Alsuhaibani. Fine-tuned pegasus: Exploring the performance of the transformer-based model on a diverse text summarization dataset.

Kenichiro Ando, Takashi Okumura, Mamoru Komachi, Hiromasa Horiguchi, and Yuji Matsumoto. 2022. Exploring optimal granularity for extractive summarization of unstructured health records: Analysis of the largest multi-institutional archive of health records in japan. *PLOS Digital Health*, 1(9):e0000099.

Anthropic. 2024. Claude 3 haiku: Our fastest model yet.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020a. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020b. Aragpt2: Pre-trained transformer for arabic language generation. *arXiv preprint arXiv:2012.15520*.

Abdelrahman Atef, Fahd Seddik, and Abdulrahman Elbedewy. 2023. Ags: Arabic gpt summarization corpus. In *2023 International Conference on Electrical, Communication and Computer Engineering (ICECCE)*, pages 1–8.

Hasna Chouikhi and Mohammed Alsuhaibani. 2022. Deep transformer language models for arabic text summarization: A comparison study. *Applied Sciences*, 12(23).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Elozino Egonmwan and Yllias Chali. 2019. Transformer-based model for single documents neural summarization. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 70–79.

AbdelRahim Elmadany, Muhammad Abdul-Mageed, et al. 2022. Arat5: Text-to-text transformers for arabic language generation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (Volume 1: Long papers)*, pages 628–647.

Jaafar Hammoud, Aleksandra Vatian, Natalia Dobrenko, Nikolai Vedernikov, Anatoly Shalyto, and Natalia Gusarova. 2021. New arabic medical dataset for diseases classification. In *Intelligent Data Engineering and Automated Learning–IDEAL 2021: 22nd International Conference, IDEAL 2021, Manchester, UK, November 25–27, 2021, Proceedings 22*, pages 196–203. Springer.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Rihab A HASANAIN, Kirsten Vallmuur, and Michele Clark. 2015. Electronic medical record systems in saudi arabia: knowledge and preferences of healthcare professionals. *Journal of Health Informatics in Developing Countries*, 9(1).

Severin Husmann, Hugo Yèche, Gunnar Rätsch, and Rita Kuznetsova. 2022. On the importance of clinical notes in multi-modal learning for ehr data. *arXiv preprint arXiv:2212.03044*.

Jiawen Jiang, Haiyang Zhang, Chenxu Dai, Qingjuan Zhao, Hao Feng, Zhanlin Ji, and Ivan Ganchev. 2021. Enhancements of attention-based bidirectional lstm for hybrid automatic text summarization. *IEEE Access*, 9:123660–123671.

A Koubaa, A Ammar, L Ghouti, O Necar, and S Sibaee. 2024. Arabiangpt: Native arabic gpt-based large language model.

Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. 2020. Generating soap notes from doctor-patient conversations using modular summarization techniques. *arXiv preprint arXiv:2005.01795*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. 2022. Medicalsum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749.

Jinane Mounsef, Maheen Hasib, and Ali Raza. 2022. Building an arabic dialectal diagnostic dataset for healthcare. *International Journal of Advanced Computer Science and Applications*, 13(7).

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, ukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, ukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong

Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Jonathan Pilault, Raymond Li, Sandeep Subramanian, and Christopher Pal. 2020. On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9308–9319.

Thomas Schaaf, Longxiang Zhang, Alireza Bayestehtashk, Mark Fuhs, Shahid Durrani, Susanne Burger, Monika Woszczyna, and Thomas Polzin. 2021. Are you dictating to me? detecting embedded dictations in doctor-patient conversations. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 587–593. IEEE.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

Deeksha Varshney, Aizan Zafar, Niranshu Kumar Behera, and Asif Ekbal. 2023. Knowledge grounded medical dialogue generation using augmented graphs. *Scientific Reports*, 13(1):3310.

Chen Yan, Susannah Rose, Michael B Rothberg, Mary Beth Mercer, Kenneth Goodman, and Anita D Misra-Hebert. 2016. Physician, scribe, and patient perspectives on clinical scribes in primary care. *Journal of general internal medicine*, 31:990–995.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Acibench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10(1):586.

Sumaya Al Zarouni, Noor Majed Al Mheiri, Kalthoom Al Blooshi, Yousif Al Serkal, Neema Preman, Sadaf Ahsan Naqvi, and Yasir Khan. 2022. Impact of an electronic medical record-based automated screening program for critical congenital heart disease: Emirates health services, united arab emirates. *BMC Medical Informatics and Decision Making*, 22(1):165.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A Appendix

A.1 Generation Prompt

Generate a 20-minute conversation in Arabic with a Saudi Najdi accent between a doctor and a patient based on the provided clinical note. The conversation should consist of approximately 50 exchanges and 3000 words. The clinical note includes details about the patient's chief complaint, history of present illness (HPI), current medications, past medical history, past surgical history, examination findings, results from any diagnostic tests, and the planned treatment or follow-up.

The conversation should begin with a warm greeting from the doctor, followed by an inquiry into the patient's current condition and specific concerns. The doctor should discuss the key aspects of the clinical note, explaining any medical terms or findings in a way that is understandable to the patient, avoiding medical jargon. This includes reviewing the results of physical examinations and any tests, discussing the significance of these findings, and outlining the proposed treatment plan. Ensure that all the information in the clinical note is included in the conversation without missing any details.

Instructions for the doctor in the conversation should include:

- Establishing a connection with the patient by asking about their overall well-being and specific symptoms in layman's terms.
- Encouraging the patient to express feelings, uncertainties, or the need for clarifications about the medical terms or procedures mentioned. Phrases like "أنا مو متأكد وش تقصد بهذا" (I'm not sure what you mean by that) or "ممكن توضح أكثر؟" (Could you explain that further?) should be included to enhance realism.
- Delving into the patient's medical history and any relevant details that could impact their current health issue, ensuring all responses are directly related to and consistent with the clinical note provided.
- Explaining the results of any examinations or tests and their implications for the patient's health in simple terms.
- Discussing the treatment options, including medications, physical therapy, or any procedures if applicable, and explaining why each is recommended using layman's terms.
- Providing clear instructions for home care, lifestyle adjustments, and medication management.
- Setting up a follow-up plan to monitor progress or further evaluate the condition.
- Ensuring that the patient has a chance to ask questions and clarifying any doubts they may have, responding in a manner that respects the patient's understanding and concerns.
- The conversation should be engaging, empathetic, and informative, reflecting a professional and caring interaction between the doctor and the patient.
- Ensure all the clinical information provided is included in the conversation.
- Integrate modal particles like " أمم" (hmm), "أيوه" (yes), "طيب"(okay), and "إن شاء الله" (God willing) to mimic natural conversational flow.
- Show empathy and understanding using phrases like "أنا متفهم قلقك" (I understand your concern), "احنا هنا عشان نساعدك" (We are here to help you), and "أنت قوي" (You're strong).
- If the clinical note mentions Labor Day, change it to Saudi National Day to reflect the local culture.

- Include culturally relevant greetings and expressions like "السلام عليكم" (Peace be upon you), "الحمد لله" (Praise be to God), and "ما شاء الله" (God has willed it).
- Incorporate references to Islamic beliefs and practices where appropriate, such as mentioning prayer or seeking guidance from Allah.
- Use Saudi Najdi dialect expressions and vocabulary to enhance the authenticity of the conversation.

clinical note: {clinical_note}
output should start with <conv> and end with </conv>
\n \n Assistant:

A.2 Improvement Prompt

Enhance the depth and educational value of the following clinical simulation conversation between a patient and physician, conducted in Arabic using the Najdi dialect. The conversation should model effective communication, medical professionalism, cultural sensitivity, and be accessible and engaging for the patient over a span of approximately 20 minutes (3000-4000 words). Guidelines for Conversation Enhancement:

- Clarify medical terms and diagnostic results comprehensively, providing detailed explanations and care instructions.
- Integrate modal particles like "أمم" (hmm), "أيوه" or "نعم" (yes), and "حسنًا" or "طيب" (okay) to mimic realistic conversational flow.
- Encourage the patient to describe their medical history, symptoms, and current medications in an accessible manner.
- Use transitional phrases to connect topics naturally and ensure the conversation flows logically.
- Structure the dialogue to explore each medical topic thoroughly, allowing both the physician and patient to discuss issues in-depth.
- Organize the conversation into sections based on the clinical scenario, each lasting a few minutes.
- Demonstrate empathy and professionalism consistently, making the patient feel understood and supported.
- Ensure the dialogue builds logically, with each response adding to the narrative or providing new insights.
- Address the patient in a culturally sensitive manner, respecting their background and experiences.
- Expand on typical topics to include discussion of lifestyle impacts, potential complications, and preventive measures.
- Minimize repetition; each exchange should introduce new concepts or elaborate on previous points to educate and inform.
- Use the detailed clinical notes as a foundational tool to align the conversation with the patient's medical history and symptoms.
- Avoid unexplained medical jargon or abbreviations. Strive for clarity to ensure the patient fully understands their condition and treatment plan.

Simulation Materials:

Conversation: {conversation} Clinical Note: {clinical_note}

Please rewrite the provided conversation, incorporating the guidelines above to enhance its educational value, realism, cultural relevance, and patient engagement. The output should be entirely in Arabic using the Najdi dialect. output should start with <conv> and end with </conv> \n\n Assistant:

A.2.1 Evaluation Prompt

Evaluation Criteria for Medical Conversation and Clinical Note Similarity

Medical Accuracy and Completeness

1.1 Symptoms and Complaints - Score (0-5): [Enter score]
- Comments: [Enter comments]
1.2 Medical History - Score (0-5): [Enter score]
- Comments: [Enter comments]
1.3 Diagnosis and Treatment - Score (0-5): [Enter score]
- Comments: [Enter comments]

Communication and Rapport

2.1 Active Listening- Score (0-5): [Enter score]
- Comments: [Enter comments]
2.2 Clarity and Explanations - Score (0-5): [Enter score]
- Comments: [Enter comments]
2.3 Empathy and Respect - Score (0-5): [Enter score]
- Comments: [Enter comments]

Structure and Flow

3.1 Logical Progression - Score (0-5): [Enter score]
- Comments: [Enter comments]
3.2 Transitions and Coherence - Score (0-5): [Enter score]
- Comments: [Enter comments]
3.3 Time Management - Score (0-5): [Enter score]
- Comments: [Enter comments]

Language and Terminology

4.1 Appropriate Language - Score (0-5): [Enter score]
- Comments: [Enter comments]
4.2 Medical Terminology - Score (0-5): [Enter score]
- Comments: [Enter comments]
4.3 Cultural Sensitivity - Score (0-5): [Enter score]
- Comments: [Enter comments]

Patient Engagement and Education

5.1 Patient Participation - Score (0-5): [Enter score]
- Comments: [Enter comments]
5.2 Patient Education - Score (0-5): [Enter score]
- Comments: [Enter comments]
5.3 Addressing Concerns - Score (0-5): [Enter score]
- Comments: [Enter comments]

Scoring System:
0: Not addressed at all
1-2: Partially addressed with significant gaps or inaccuracies
3-4: Adequately addressed with minor gaps or inaccuracies

5: Fully and accurately addressed

Overall Scores:

Medical_Accuracy_and_Completeness: [Average of 1.1, 1.2, 1.3]
Communication_and_Rapport: [Average of 2.1, 2.2, 2.3]
Structure_and_Flow: [Average of 3.1, 3.2, 3.3]
Language_and_Terminology: [Average of 4.1, 4.2, 4.3]
Patient_Engagement_and_Education: [Average of 5.1, 5.2, 5.3]

Final_Evaluation_Score: [Average of the five overall scores]

Medical Conversation: {conversation}
Clinical Note: {clinical_note}

A.3 Generated Conversation Example

Clinical Note	Generated Conversation
CHIEF COMPLAINT Right knee pain. MEDICAL HISTORY The patient has a history of diabetes. She has been doing pretty good with her diet. She states that she forgets to check her sugars quite a bit. REVIEW OF SYSTEMS Musculoskeletal: Reports right knee pain and swelling. PHYSICAL EXAM Respiratory - Auscultation of Lungs: Clear bilaterally. Cardiovascular No murmurs, gallops. Musculoskeletal - Examination of the right knee: Some swelling present. - Palpation: Some pain to palpation on the medial aspect of the right knee, and a little bit of pain on the lateral aspect of the right knee. - Range of Motion: Limited range of motion as well as pain on both flexion and extension of the knee. - Special Testing: McMurray's Test: Negative. ASSESSMENT AND PLAN 1. Right knee pain. - Medical Reasoning: I am concerned about a torn MCL due to pain on ambulation and trouble with weightbearing, as well as the pop she heard. - Patient Education and Counseling: We discussed treatment options today including bracing, anti-inflammatories, and icing. - Medical Treatment: I am going to put her in a straight leg brace and I will prescribe some Mobic. She can start taking that as a pain reliever and to try to get some of the swelling down. I want her to ice her knee once an hour for about 15 minutes. - Additional Testing: I am also going to send her out for an MRI. 2. Type 2 diabetes. - Medical Reasoning: The patient states that her type 2 diabetes are well-managed. - Medical Treatment: I am also going to get a refill on the metformin that she has been taking 500 mg. - Additional Testing: We are going to recheck her hemoglobin A1c. Patient Agreements: The patient understands and agrees with the recommended medical treatment plan.	د: السلام عليكم يا أم محمد، كيف حالك اليوم؟ م: وعليكم السلام يا دكتور، الحمدلله بخير بس الركبة اليمين تعورني كثير وتضايقني. د: أمم، أنا آسف إنك تعانين من هالألم. قولي لي أكثر عن الوجع في ركبتك. متى بدأ بالضبط وكيف كان الإحساس؟ م: والله يا دكتور بدأ من أسبوع تقريبا لما كنت أمشي. فجأة حسيت كأن في شي انقطع أو انفرقع بركبتي، وبعدها على طول صارت تعورني وتنتفخ. د: أيوه فهمت عليك. وكيف كان تأثير الألم على حركتك ومشيك بعد الحادثة؟ م: صراحة يا دكتور صار صعب علي كثير إني أمشي أو أتحرك بشكل طبيعي. خصوصا لما أحاول أثني ركبتي أو أفردها، أحس بألم شديد. د: طيب، خليني ألقي نظرة على ركبتك... أمم في انتفاخ واضح، وكمان ألم لما ألمس المنطقة من الجهة الداخلية والخارجية للركبة. وأشوف إن حركة الركبة محدودة مع وجود ألم. م: أيوه يا دكتور، هذا بالضبط اللي أحسه وأعاني منه. بس وش تتوقع يكون السبب؟ أنا خايفة يطلع شي خطير لا سمح الله. د: أتفهم تماما قلقك وخوفك يا أم محمد، بس إن شاء الله ما فيه داعي للهلع. من خلال الفحص السريري، أنا أشك بشكل أولي إنه يمكن يكون عندك تمزق بالرباط الداخلي للركبة، وذلك بسبب طبيعة الألم والانتفاخ وصعوبة تحميل الوزن على الرجل بعد سماع صوت الفرقعة. م: يا ساتر يا رب... الله يستر. طيب وش راح يكون العلاج يا دكتور؟ أنا محتاجة ركبتي سليمة عشان أقدر ألحق على عيالي وبيتي. د: لا تشيلين هم أبدا يا أم محمد، احنا هنا عشان نساعدك ونقف معك إن شاء الله. في الوقت الحالي، العلاج الأولي حيشمل وضع دعامة طبية على رجلك المصابة لتثبيت الركبة، وكمان حنعطيك أدوية مضادة للالتهاب عشان تخفف الألم والانتفاخ. حكتب لك دواء اسمه "موبيك" وهو فعال جدا في هالحالات. كمان لازم تحرصين تحطين ثلج على ركبتك كل ساعة لمدة 15 دقيقة تقريبا. م: إن شاء الله يا دكتور حسوي زي ما قلت. بس في شي ثاني ممكن نسويه؟ د: نعم طبعا، أنا كمان حأطلب لك تصوير رنين مغناطيسي للركبة عشان نتأكد تماما من التشخيص ونقدر نحدد مدى الإصابة بالرباط بشكل دقيق. النتيجة حتساعدنا نقرر إذا في حاجة لأي تدخل إضافي أو لأ. م: طيب الله يطمنك يا دكتور. تسلم والله، بس عندي سؤال ثاني بعد إذنك... أنا مريضة بالسكر وآخذ حبوب الميتفورمين. وش رأيك فيها؟ المفروض أكمل عليها؟ د: ما شاء الله عليك يا أم محمد، من زين إنك مهتمة بصحتك ومتابعة لعلاج السكر. إن شاء الله السكر عندك مسيطر عليه. عشان نتأكد أكثر ونطمن، أنا حطلب إعادة فحص الهيموجلوبين السكري. وبالنسبة للميتفورمين، حجدد لك الوصفة الطبية بنفس الجرعة الحالية 500 ملغم، واستمري عليها زي ما انتي. م: تمام يا دكتور، الله يعطيك العافية. بس للأسف ساعات كثير أنسى أو أتكاسل عن قياس السكر بانتظام، الله يستر بس. د: أنا متفهم يا أم محمد إن الالتزام بمراقبة السكر بشكل يومي ممكن يكون صعب أو يتطلب مجهود، بس صدقيني هو شي مهم جدا لصحتك على المدى الطويل. حاولي قد ما تقدرين تخلينها عادة يومية منتظمة، وتسجلين القراءات في دفتر عشان نقدر أنا وإنتي نتابع وضعك الصحي بشكل أفضل. وإذا احتجتي أي مساعدة أو عندك أي استفسار، أرجوك لا تترددين أبدا في إنك تتواصلين معي. م: إن شاء الله يا دكتور، ما تقصر والله. ربي يجزاك كل خير على اهتمامك وحرصك. الله يسعدك. بس بغيت أسألك متى المفروض أرجع أشوفك مرة ثانية؟ د: حنحدد موعد المتابعة القادم بعد ما تستلمين نتيجة الرنين المغناطيسي إن شاء الله. حنشوف سوا التطورات في حالة ركبتك، ونقرر إذا في حاجة لأي خطوات علاجية إضافية. المهم إنك تلتزمين بكل النصائح والتعليمات اللي اتفقنا عليها اليوم. وثقي تماما بالله، أنا متأكد إنك قوية وقادرة تتجاوزين هالمحنة وتتحسنين بإذن الله. م: أمين يا رب العالمين، من كل قلبي. ما أعرف كيف أشكرك يا دكتور، فرجت همي وطمنتني والله. في أمان الله وحفظه. د: في أمان الكريم يا أم محمد. ربي يسلمك ويقومك بالسلامة يارب. مع السلامة.

Figure 2: Sample of Clinical Note and Generated Conversation.