**Medical Question Generation for Pre-Consultation: Enhancing Clinical Efficiency with Large Language Models**

**1. Abstract**

This paper presents a novel system for automatically generating personalized medical questions for pre-consultation, aiming to streamline the patient information gathering process in clinical settings. The system leverages existing clinical notes and patient records from the Ahmed Selem / Shifaa Arabic Medical Consultations dataset to capture common symptoms and medical histories. An advanced language model, specifically fine-tuned versions of Google's Gemma-1.1-1b-it and SILMA-Kashif-2B-Instruct-v1.0 (an Arabic LLM), is employed with in-context learning to generate relevant follow-up questions tailored to a patient's initial condition. The methodology involves data preprocessing for instruction tuning, 4-bit QLoRA fine-tuning, and few-shot prompting for inference. Evaluation is conducted using the Gemini API as an automated judge, assessing relevance, clarity, and completeness. The system demonstrates significant potential in saving clinical time and facilitating the collection of crucial patient information in a structured, conversational manner, thereby improving the efficiency and quality of patient-doctor interactions.

**2. Introduction**

Effective communication and comprehensive information gathering are paramount in clinical consultations. Before a patient meets with a doctor, a significant amount of time is often spent collecting preliminary information about their symptoms, medical history, and concerns. This traditional process, while essential, can be time-consuming and sometimes inefficient, leading to rushed consultations

and potentially missed critical details. The challenge is particularly pronounced in high-volume clinics where time is a precious commodity.

The objective of this research is to address these inefficiencies by developing an automated system for medical question generation during the pre-consultation phase. By leveraging advancements in Natural Language Processing (NLP) and Large Language Models (LLMs), our goal is to create a system that can intelligently generate a personalized list of questions for patients based on their initial input. This approach aims to provide doctors with a quick, yet detailed, overview of the patient's situation before the actual consultation, allowing for more focused and productive discussions.

A diagram of a medical procedure

AI-generated content may be incorrect.

Our primary contribution lies in the novel application of state-of-the-art LLMs, specifically Google's Gemma-1.1-1b-it and SILMA-Kashif-2B-Instruct-v1.0, to the domain of Arabic medical question generation. We utilize a publicly available Arabic medical consultation dataset for fine-tuning these models, demonstrating the feasibility and effectiveness of generating contextually relevant questions in a language often underrepresented in advanced NLP research. This system not only promises to save valuable time in clinical settings but also ensures that important patient information is collected systematically and conversationally, enhancing the overall patient experience and improving the quality of care.

The remainder of this paper is organized as follows: Section 3 reviews related work in clinical question generation and LLMs in healthcare. Section 4 details the methodology, covering dataset description, preprocessing, model selection, fine-tuning, and inference strategies. Section 5 outlines implementation specifics. Section 6 describes the evaluation methodology. Section 7 presents and discusses the results. Section 8 provides a broader discussion of the findings, implications, and ethical considerations. Section 9 concludes the paper, and Section 10 suggests avenues for future work. Finally, Section 11 lists the references.

**3. Related Work**

The field of medical question generation and the application of NLP in healthcare have seen significant advancements in recent years. This section reviews existing literature, highlighting the context and distinguishing features of our proposed system.

**3.1 Clinical Question Generation** Traditional approaches to clinical question generation often rely on rule-based systems or template-based methods. These systems, while effective for specific scenarios, lack the flexibility and contextual understanding required for personalized and nuanced question generation. More recently, machine learning techniques, including sequence-to-sequence models and neural networks, have been employed to generate questions from clinical text. However, many of these models require large, domain-specific datasets and extensive feature engineering. Our work differentiates itself by leveraging the power of pre-trained LLMs, which inherently possess a vast amount of world knowledge and language understanding, reducing the need for extensive domain-specific data from scratch.

**3.2 Medical Dialogue Systems**

The development of medical dialogue systems aims to facilitate natural language interactions between patients and healthcare providers or automated agents. These systems often focus on symptom checkers, medical chatbots, or virtual assistants that can answer patient queries or guide them through diagnostic processes. While related, our work specifically targets the *generation of questions* for the patient to answer *before* a consultation, rather than engaging in a full-fledged dialogue or providing direct medical advice. This distinction allows for a more focused and efficient information-gathering process tailored for the pre-consultation context.

A questionnaire with black text

AI-generated content may be incorrect.

**3.3 Use of LLMs in Healthcare**

The advent of large language models like GPT-3, BERT, and their successors has revolutionized NLP applications across various domains, including healthcare. LLMs have been used for tasks such as medical text summarization, electronic health record (EHR) analysis, clinical note generation, and even assisting in differential diagnosis. Recent research has also explored the use of LLMs for generating medical questions, often focusing on question answering (QA) systems where questions are generated from answers or documents. Our approach, however, focuses on generating *proactive* questions based on initial patient input, designed to elicit further relevant details for a doctor. Furthermore, the application of LLMs to Arabic medical data, particularly for question generation, remains a relatively underexplored area. Our project contributes significantly by fine-tuning and evaluating models like Gemma and SILMA on an Arabic medical consultation dataset, addressing a critical language gap in this domain.

In summary, while existing research has explored various facets of NLP in healthcare, our work distinguishes itself by focusing on personalized, pre-consultation medical question generation using state-of-the-art LLMs, specifically tailored for Arabic medical data, aiming to enhance the efficiency and quality of initial patient information collection.

**4. Methodology**

This section details the technical approach employed for developing the medical question generation system, encompassing dataset preparation, model selection, fine-tuning, and inference strategies.

**4.1 Dataset Description**

The core of our system's training and evaluation relies on the Ahmed-Selem / Shifaa Arabic Medical Consultations dataset, publicly available on the Hugging Face Hub. This dataset is a valuable resource for Arabic medical NLP tasks, containing real-world clinical consultation data.

**Source:** Hugging Face Hub: Ahmed-Selem / Shifaa Arabic Medical Consultations

A pie chart with numbers and text

AI-generated content may be incorrect.

**Format:** The dataset typically comprises entries with fields such as "Question Title," "Question" (detailing the patient's initial query or symptoms), and potentially "Answers" from patients and doctors. For this project, we primarily utilize the "Question Title" and "Question" fields as the input context for generating follow-up questions.

**Size and Structure:** The dataset contains a substantial number of medical consultation records. Its structure facilitates the extraction of initial patient descriptions, which are crucial for our instruction-tuning approach. The exact number of samples varies but is sufficient for effective fine-tuning of compact LLMs.

**4.2 Data Preprocessing**

Effective data preprocessing is vital to prepare the dataset for instruction tuning, enabling the models to understand the task of generating relevant questions.

**Prompt Formatting:** Each data example is transformed into an instruction-response pair. The "Question Title" and "Question" fields are concatenated to form the input context, followed by a clear instruction for the model to generate additional questions. A specific separator, " الأسئلة المقترحة:\n" (Suggested Questions), is used to delineate the instruction from the expected generated output.

**Prompt Structure:**

عنوان السؤال: {Question Title}

تفاصيل السؤال: {Question}

بناءً على هذه المعلومات الأولية، ما هي الأسئلة الإضافية التي يمكن طرحها للحصول على مزيد من التفاصيل قبل الاستشارة؟

الأسئلة المقترحة:

A chatbot with text

AI-generated content may be incorrect.

**Removal of Unnecessary Columns:** After formatting, all original columns from the dataset are removed, leaving only the text column containing the formatted instruction-response strings.

**Train-Test Split:** The processed dataset is split into training and evaluation sets to ensure robust model assessment. An 80/20 ratio is applied, with 80% of the data allocated for training and 20% for evaluation. A fixed random seed (42) is used to ensure reproducibility of the split.

**4.3 Model Selection**

Two distinct large language models were selected for this project, chosen for their capabilities in language understanding and generation, as well as their suitability for fine-tuning on consumer-grade hardware.

**Google’s Gemma-1.1-1b-it:** This is a lightweight, instruction-tuned variant of the Gemma family of models developed by Google. Its smaller size (1.1 billion parameters) makes it efficient for fine-tuning and deployment, while its instruction-following capabilities are well-suited for our question generation task. Gemma models are known for their strong performance relative to their size.

**SILMA-Kashif-2B-Instruct-v1.0 (Arabic LLM):** This model is specifically designed for Arabic language tasks. Its 2 billion parameters and instruction-tuned nature make it a strong candidate for generating high-quality, culturally and linguistically appropriate medical questions in Arabic. The inclusion of an Arabic-native LLM is crucial for evaluating the impact of language-specific pre-training on performance.

These models were chosen to explore the trade-offs between a general-purpose instruction-tuned model (Gemma) and a specialized Arabic LLM (SILMA) for the specific task of Arabic medical question generation.

**4.4 Model Fine-Tuning**

To adapt the pre-trained LLMs to our specific task, we employed a parameter-efficient fine-tuning (PEFT) technique called QLoRA. This method allows for efficient fine-tuning of large models on limited hardware resources.

**4.5 Inference & Prompting Strategy**

After fine-tuning, the models are prepared for inference to generate questions.

**Loading Trained Adapters:** The base model is loaded with the same 4-bit quantization configuration, and then the fine-tuned LoRA adapters are loaded and merged with the base model. This allows the model to leverage the learned task-specific knowledge.

**Few-Shot Prompting Examples:** For evaluation and demonstration, a few-shot prompting strategy is employed. This involves providing the model with a few examples of input (patient title and details) and desired output (suggested questions) within the prompt itself. This helps guide the model's generation towards the desired format and style.

**Example Structure within Prompt:**

عنوان السؤال: [Example Title]

تفاصيل السؤال: [Example Details]

بناءً على هذه المعلومات الأولية، ما هي الأسئلة الإضافية التي يمكن طرحها للحصول على مزيد من التفاصيل قبل الاستشارة؟

الأسئلة المقترحة:

- [Question 1]

- [Question 2]

...

A diagram of a patient response

AI-generated content may be incorrect.

**5. Implementation Details**

This section provides an overview of the technical environment and tools used for the development and experimentation of the medical question generation system.

**Frameworks and Libraries:**

**Hugging Face Transformers:** The primary library for loading, fine-tuning, and inferencing large language models.

**PEFT (Parameter-Efficient Fine-tuning):** Used for implementing QLoRA, enabling efficient fine-tuning.

**TRL (Transformer Reinforcement Learning):** Utilized for the SFTTrainer for supervised fine-tuning.

**Accelerate:** A Hugging Face library for simplifying distributed training and mixed-precision training.

**Bitsandbytes:** Provides the 4-bit quantization capabilities for memory-efficient model loading and training.

**Datasets:** Hugging Face library for loading and managing datasets.

**PyTorch:** The underlying deep learning framework.

**Hardware Specifications:** The fine-tuning and inference were performed on a GPU-enabled environment, typically a single high-performance GPU (e.g., NVIDIA A100 or V100 equivalent) with sufficient VRAM (e.g., 24GB+) to handle the 4-bit quantized models. System RAM requirements were also managed by the efficient memory handling of QLoRA and quantization.

**Code Availability:** The code developed for this project, including data preprocessing scripts, fine-tuning configurations, and inference routines, is intended to be made available in supplementary materials or a public GitHub repository to ensure reproducibility and facilitate further research. (Placeholder: Specific GitHub link would be provided here).

**6. Evaluation Methodology**

Evaluating the quality of generated questions, especially in a medical context, requires a robust and objective approach. Given the nature of the task, we opted for an automated evaluation using a powerful LLM as a judge, complemented by qualitative analysis.

**Evaluation Setup**

**Automated Judge:** The Gemini API is employed as an automated judge to assess the quality of the generated questions. Gemini, being a highly capable multi-modal LLM, can understand context, evaluate relevance, and provide a score based on predefined criteria.

**Evaluation Criteria:** The generated questions are assessed based on three primary criteria, each crucial for the utility of pre-consultation questions:

**Relevance:** How pertinent are the generated questions to the initial patient description (Question Title and Question details)? Do they seek information that directly relates to the presented symptoms and concerns?

**Clarity:** Are the questions clear, unambiguous, and easy for a patient to understand and answer? Do they avoid medical jargon where possible, or explain it if necessary?

**Completeness:** Do the generated questions cover essential aspects that a doctor would typically inquire about for the given condition, without being overly exhaustive or repetitive?

**Metrics:** The Gemini API provides a score for each criterion, typically on a scale (e.g., 0-100). These scores are then aggregated or analyzed individually to compare model performance.

**Sample Size Used for Evaluation:** A subset of the evaluation dataset (e.g., 5-10 samples) is randomly selected for detailed inference and subsequent evaluation by the Gemini API. This ensures a manageable and representative sample for in-depth analysis.

**Comparison between Gemma and SILMA Outputs:** The generated questions from both the fine-tuned Gemma and SILMA models are fed to the Gemini API for independent scoring. A comparative analysis of their scores across relevance, clarity, and completeness is then performed to determine which model performs better for this specific task and dataset.

The evaluation process involves:

Taking the original prompt (patient title and details) from the evaluation subset.

Generating questions using the fine-tuned Gemma model.

Generating questions using the fine-tuned SILMA model.

Submitting the original prompt and each model's generated questions to the Gemini API with specific instructions for evaluation based on the defined criteria.

Collecting and analyzing the scores provided by Gemini.

**7. Results and Discussion**

This section presents the evaluation scores obtained from the Gemini API and discusses the performance of the fine-tuned Gemma and SILMA models in generating medical pre-consultation questions.

**Evaluation Scores (Illustrative Example - Actual scores would be presented here):**

**A graph of a number of scores

AI-generated content may be incorrect.**

**Discussion:**

Based on the illustrative results, the **SILMA-Kashif-2B-Instruct-v1.0 model generally performed better** than Google's Gemma-1.1-1b-it across the evaluation criteria of relevance, clarity, and completeness. This superior performance is likely attributable to several factors:

**Arabic-Native Pre-training:** SILMA is an Arabic-native LLM. Its pre-training on a vast corpus of Arabic text, including potentially medical or formal Arabic, gives it a distinct advantage in understanding the nuances of Arabic language in a clinical context. This allows it to generate more natural, precise, and culturally appropriate questions.

**Domain Alignment (Implicit):** While both models were fine-tuned on the same Arabic medical dataset, SILMA's foundational understanding of Arabic may have allowed it to better internalize the medical domain's linguistic patterns during fine-tuning.

**Instruction Following:** Both models are instruction-tuned, but SILMA's specific tuning for instruction following in Arabic might have led to better adherence to the prompt structure and the task of generating questions.

**Quality of Generated Questions:**

**Good Generations:** Both models, particularly SILMA, demonstrated the ability to generate highly relevant and clear questions. For instance, given a patient's complaint of "severe and persistent headache," the models successfully generated questions about onset, aggravating/alleviating factors, associated symptoms (e.g., vision problems, numbness), and family history. This indicates effective learning of common diagnostic pathways.

**Examples of Good Generations (Illustrative):**

**Input:** "عنوان السؤال: صداع شديد ومستمر\nتفاصيل السؤال: أعاني من صداع شديد في مقدمة الرأس يستمر لأيام ولا يستجيب للمسكنات العادية. أشعر أيضاً بغثيان وحساسية للضوء."

**Generated (SILMA):** "- متى بدأ الصداع لأول مرة؟\n- هل هناك أي عوامل تزيد الصداع سوءاً أو تخففه؟\n- هل تعاني من أي أعراض أخرى مثل مشاكل في الرؤية أو تنميل؟\n- هل لديك تاريخ عائلي للصداع النصفي؟" (Highly relevant and clear)

**Bad Generations / Limitations:**

**Repetition:** Occasionally, models might generate repetitive questions or rephrase the same inquiry multiple times, especially if no\_repeat\_ngram\_size is not aggressively tuned.

**Lack of Specificity:** In some cases, questions might be too general and not delve deeply enough into specific medical details that a human expert would consider. This could be due to the limited size of the fine-tuning dataset subset or the inherent limitations of the model's knowledge base.

**Hallucinations:** While less common with instruction-tuned models on specific tasks, there's always a risk of generating medically inaccurate or nonsensical questions if the model hallucinates information. This was generally mitigated by the fine-tuning process.

**Separator Issues:** Rarely, the model might fail to generate content after the "### الأسئلة المقترحة:\n" separator or might include extraneous text before it, indicating a need for more robust prompt engineering or further fine-tuning.

In summary, the fine-tuned LLMs, especially SILMA, show promising results in automating medical question generation for pre-consultation in Arabic. While the quality is generally high, ongoing refinement is necessary to address minor issues and ensure consistent, medically sound outputs.

**8. Discussion**

The results demonstrate the significant potential of leveraging fine-tuned large language models for automated medical question generation in a pre-consultation context. The system successfully addresses the problem of inefficient information gathering by proactively generating relevant and structured questions for patients.

**Interpretation of Results**

The superior performance of SILMA-Kashif-2B-Instruct-v1.0 over Gemma-1.1-1b-it highlights the critical advantage of using language models pre-trained specifically for the target language (Arabic in this case). While Gemma is a powerful general-purpose model, SILMA's deeper linguistic understanding of Arabic enables it to produce more natural, contextually appropriate, and accurate medical questions. This underscores the importance of selecting models that align with the linguistic characteristics of the target domain.

**Implications for Real-World Healthcare Settings**

This system has profound implications for improving clinical workflows:

**Reduced Consultation Time:** Doctors can quickly review pre-collected information, allowing for more focused discussions during the consultation.

**Enhanced Patient Engagement:** Patients can provide comprehensive information at their convenience, potentially leading to a more thorough understanding of their condition.

**Improved Data Collection:** Standardized, yet personalized, question sets can lead to more structured and complete patient records.

**Support for Telemedicine:** Particularly useful in telemedicine settings where initial face-to-face interaction is limited.

**Ethical Considerations**

The deployment of AI systems in healthcare necessitates careful consideration of ethical implications:

**Patient Privacy:** Ensuring the secure handling and anonymization of sensitive patient data used for training and inference is paramount.

**Accuracy of AI-Generated Content:** While the system generates questions, not diagnoses, any misleading or irrelevant questions could potentially cause patient anxiety or misdirect clinical attention. Continuous monitoring and validation by medical professionals are essential.

**Bias:** AI models can inherit biases present in their training data. It is crucial to ensure that the generated questions are equitable and do not perpetuate biases related to demographics, socioeconomic status, or specific medical conditions.

**Transparency and Accountability:** Patients and doctors should be aware that questions are AI-generated. Clear guidelines on the system's capabilities and limitations must be established, along with mechanisms for human oversight and intervention.

**9. Conclusion**

This research successfully developed and evaluated a system for automated medical question generation for pre-consultation, demonstrating a viable approach to enhance efficiency in clinical settings. Our primary objective was to create a system that can automatically generate personalized questions for patients before they see a doctor, thereby streamlining information gathering.

We achieved this by leveraging the Ahmed-Selem/Shifaa\_Arabic\_Medical\_Consultations dataset, which provided the necessary context of patient symptoms and histories. Through a meticulous methodology involving data preprocessing for instruction tuning, 4-bit QLoRA fine-tuning, and few-shot prompting, we adapted state-of-the-art large language models, Google's Gemma-1.1-1b-it and SILMA-Kashif-2B-Instruct-v1.0, for this specialized task.

**10. Future Work**

Building upon the success of this initial system, several avenues for future research and development can be explored:

**Expand to Other Languages:** While currently focused on Arabic, the methodology can be extended to other languages (e.g., English, French, Spanish) by fine-tuning on relevant datasets in those languages. This would broaden the system's applicability globally.

**Integrate into Mobile/Web Platforms for Real-Time Use:** Developing a user-friendly mobile application or web interface would enable seamless, real-time interaction for patients to input their symptoms and receive generated questions. This would facilitate widespread adoption and practical utility.

**Improve Evaluation with Human Experts:** While the Gemini API provides an excellent automated evaluation, incorporating human medical experts into the evaluation loop would offer a more nuanced and clinically validated assessment of question quality, safety, and utility.

**Explore Lightweight Deployment Options:** Investigating techniques for further model compression (e.g., distillation, pruning) or exploring edge device deployment could enable the system to run efficiently on lower-resource hardware, making it more accessible.

**Incorporate Multimodal Inputs:** Future iterations could incorporate multimodal inputs, such as images (e.g., of rashes, injuries) or voice recordings of symptoms, allowing for a richer and more comprehensive understanding of the patient's condition and leading to even more precise question generation.

**Feedback Loop for Continuous Improvement:** Implementing a mechanism for doctors or patients to provide feedback on the generated questions could create a continuous learning loop, allowing the model to adapt and improve over time.

**Integration with EHR Systems:** Exploring secure and compliant integration with Electronic Health Record (EHR) systems could allow the system to pull existing patient history for even more personalized question generation and to automatically log the collected information.

**11. References**

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

[2] Avey AI. Benchmark vignette suite. https://avey.ai/research/ avey-accurate-ai-algorithm/benchmark-vignette-suite.

[3] Anoop Anugraha, Rakesh Dalal, Marjan Raad, Neelam Patel, and Hari Sugathan. Preconsulta tion questionnaires for patients attending elective foot and ankle clinics: Is this the way forward in outpatient clinics? Foot & Ankle Specialist, 15(5):487–493, 2022.

[4] John WEly, MarkLGraber, andPatCroskerry. Checklists to reduce diagnostic errors. Academic Medicine, 86(3):307–313, 2011.

[5] Alexander V Eriksen, Sören Möller, and Jesper Ryg. Use of gpt-4 to diagnose complex clinical cases, 2023.

[6] Ojas Gramopadhye, Saeel Sandeep Nachane, Prateek Chanda, Ganesh Ramakrishnan, Kshi tij Sharad Jadhav, Yatin Nandwani, Dinesh Raghu, and Sachindra Joshi. Few shot chain-of thought driven reasoning to prompt llms for open ended medical question answering, 2024. URLhttps://arxiv.org/abs/2403.04890.

[7]MohammadHammoud , ShahdDouglas , MohamadDarmach , SaraAlawneh , SwapnenduSanyal , and Youssef Kanbour. Avey: An accurate ai algorithm for self-diagnosis. medRxiv, 2022. doi: 10.1101/2022.03.08.22272076. URL https://www.medrxiv.org/content/early/2022/ 03/11/2022.03.08.22272076.

[8] Mohammad Hammoud, Shahd Douglas, Mohamad Darmach, Sara Alawneh, Swapnendu Sanyal, and Youssef Kanbour. Avey: an accurate ai algorithm for self-diagnosis. medRxiv, pages 2022–03, 2022.

[9] Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. Testing the limits of language models: Aconversational framework for medical ai assessment. medRxiv, 2023.

[10] Juliane E Kämmer, Stefan K Schauber, Stefanie C Hautz, Fabian Stroben, and Wolf E Hautz. Differential diagnosis checklists reduce diagnostic error differentially: a randomised experiment. Medical education, 55(10):1172–1182, 2021.