Archives available at journals.mriindia.com

International Journal on Advanced Electrical and Computer Engineering

ISSN: 2349-9338

Volume 15 Issue 01s, 2026

Evaluation of Artificial Intelligence in Answering Dermatological Medical Questions

[1]Abdullah ALSarrajie , [2]Akram ALsubari

[1 2] Department of Computer Science, Faculty of Applied Sciences, Ibb University, Ibb, Yemen

Email: [1] abdullah.alsarrajie@ibbuniv.edu.ye, [2] akram.alsubari@ibbuniv.com

Peer Review Information

Submission: 05 Dec 2025

Revision: 25 Dec 2025

Acceptance: 10 Jan 2026

Keywords

Arabic NLP, Dermatology Question Answering, Medical Question Answering, Transformer Models, Fine-tuning.

Abstract

This research presents a systematic evaluation of the adaptation of a large Arabic linguistic model (based on AraGPT2) to answer questions in the field of dermatology. The study aims to bridge the gap in specialized linguistic resources by fine-tuning the model to a newly created and purified dataset, collected using a hybrid methodology combining web scraping and filtered data enrichment. This dataset consists of 40,132 specialized question-answer pairs. The performance of the finely tuned model was quantitatively assessed using BERTScore, BLEU, Levenshtein distance, and two types of initial human evaluation. The quantitative results showed strong semantic performance, with the model achieving a BERTScore (F1) of 64.49%, confirming its ability to effectively understand medical context and meaning. In contrast, the verbal matching measures reflected a tendency toward free generation, scoring BLEU at 10.00% and Levenshtein distance at 28.13%. These results demonstrate that the model favors the free generation of new formulations over the verbatim retrieval of reference texts. Furthermore, the qualitative results of the proposed model showed a competitive overall performance of 4.01, achieving a high score of 4.55 in the criteria of linguistic clarity and readability for non-expert audiences. These results confirm that the model's primary contribution lies in its ability to enhance human comprehension (understanding) of complex medical data. The study also underscores the urgent need for subsequent clinical validation by field experts to ensure complete clinical accuracy and reliability.

Introduction

Contemporary approaches in artificial intelligence (AI), particularly with the emergence of large language models (LLMs) and transformer models [1], are witnessing unprecedented radical developments in their ability to perform natural language processing (NLP) and revolutionize complex cognitive applications. These models, which have become the benchmark methodology for understanding and generation tasks, promise to profoundly transform many sectors, most notably healthcare [2], where they are used to support clinical decisions and simplify access to medical knowledge. Leading models, such as ChatGPT-4, have demonstrated their ability to support clinical knowledge in subspecialties, successfully passing dermatology specialty certification exams in various linguistic contexts (English, Polish, and Korean) [3][4], establishing AI as a powerful tool for cognitive and clinical support. Dermatology is one of the specialties that greatly benefits from these technologies, as AI systems contribute to providing reliable and objective automated diagnostic tools, especially in the face of challenges such as the shortage of

dermatologists and the need to alleviate the burden of self-diagnosis [5].

In the context of the Arabic language, characterized by its morphological complexity and the challenges associated with the diversity between Standard Arabic (MSA) and colloquial Arabic (AD) [6][7], specially trained models such as AraGPT2 and AraBERT [8] have emerged. In medical question answering systems, developments in Arabic have been reviewed and challenges related to resource scarcity have been identified [9], with some resources developed for general disease detection [10], and language models tailored to specific domains such as infectious diseases [11]. In the field of dermatology specifically, global efforts have often focused on multimodal models that process both text and images [12][13], or on multilingual models that use translation from English [14]. Despite the proven success of large language models in assessing medical knowledge, there remains a clear shortcoming in the focus on generative question answering (GQA) systems specialized in Arabic dermatology. The existing literature lacks an in-depth evaluation of the effectiveness of fine-tuned Arabic transformer models specifically for this purpose. The majority of Arabic work remains either general in its medical field [10] or limited to extraction and classification tasks [9], leaving a clear knowledge and technical gap in providing Arabic speakers with dermatological information.

This paper aims to bridge this gap through a systematic evaluation of artificial intelligence (AI) performance in answering dermatology questions in Arabic. However, this research goes beyond simply measuring technical accuracy to include a qualitative assessment of the impact on the end consumer. The quality of the answer to the non-specialist reader's question (clarity and readability) places this study at the heart of the intersection of AI with human cognition and the ability to understand and control information. To achieve this, the study focuses on improving the AraGPT2 database model. This research primarily seeks to answer the question: How accurate and linguistically clear are the outputs of the modified Arabic-translated models in answering specialized dermatology questions?

Key contributions of the paper:

-*Construction and development of a new dataset:* A purified and specialized dataset for dermatology question answering (40,132 valid samples) was created and adapted to the answer generation environment.

-*Demonstration of strong semantic performance:* The modified AraGPT2-base model demonstrated strong semantic performance, achieving a BERTScore (F1) of 64.49%, confirming its ability to accurately understand medical context and meaning.

-*Confirmation of generative ability*: Verbal metrics revealed the model's tendency toward free generation of novel expressions rather than literal matching. The BLEU metric recorded a low performance (0.10%), while the Levenshtein distance (28.13%), confirming the generative nature of the model's output.

-*Achieving competitive superiority in clarit:* In initial human evaluation, the proposed model demonstrated competitive superiority in the linguistic clarity of medical content and readability for the general public, emerging as the best performer on this key qualitative criterion. The model's performance will be evaluated quantitatively via semantic and verbal metrics, in addition to a qualitative human evaluation focusing on clarity and general readability.

The rest of this paper is organized as follows: Section 2 reviews previous work, the technological background, and highlights research limitations in the specialized Arabic medical field. Section 3 describes the methodology used to build the dataset and the fine-tuning process of the AraGPT2-base model. Section 4 presents the results and discussion. Section 5 discusses conclusions. Section 6 discusses limitations and future directions.

Related Works

The field of natural language processing (NLP) has witnessed a pivotal shift with the dominance of transformer models, which have become the benchmark methodology for language comprehension and generation tasks [1]. In the Arabic context, characterized by morphological complexity and diversity between Modern Standard Arabic (MSA) and colloquial dialects (AD) [6], pre-trained models specific to the Arabic language have emerged, such as AraBERT and AraGPT2 [8]. These models have demonstrated their technical ability to accommodate Arabic linguistic complexities and their success in essential applications including text generation and classification [6]. Comprehensive reviews of Arabic Question Answering Systems (Arabic QAS) demonstrate the systematic evolution towards adopting deep learning models and transformers, confirming their position as a standard technological background for current research [1]. These efforts are complemented by recent efforts, such as the Jais-13B project [15] and the Noor project

[16], which focused on the large-scale pre-training phase and overall performance targets for bilingual models. In contrast to this technological progress, Arabic medical question answering systems face structural challenges compared to their global counterparts, most notably the scarcity of labeled and structured datasets in the Arabic medical field [9]. Although there have been some efforts to develop general resources for question answering (such as a dataset for detecting general diseases [10]) or to customize models for specific domains (such as InfectA-Chat for infectious diseases [11]), the literature reveals specific shortcomings in specialization and methodology:

- *Scope of specialization:* Works tend to focus on the general medical field rather than subspecialties such as dermatology.
- *Answering style:* Most available work is limited to classification tasks (such as that used in automated image-based diagnosis [5]) or answer extraction tasks [9], rather than the open generative QA style, which requires deep contextual understanding and comprehensive formulation.

Evaluations of large language models (LLMs) in dermatology have shown remarkable efficiency (e.g., ChatGPT-4 in certification exams [3][4]), but these evaluations have been almost exclusively in foreign languages (English, Polish, and Korean). Furthermore, the global approaches to answering dermatological questions have mostly been multimodal [12][13]or translation-based multilingual [14]. Some research has also focused on the use of parameter-efficient fine-tuning techniques (PEFT) to adapt open-source universal models to specialized Arabic tasks [17].

This critical review confirms the absence of in-depth research focused on specialized Arabic text generation in the field of dermatology. Accordingly, the fundamental research gap is the absence of a generative dermatological question answering system in Arabic based on systematically evaluated and tailored data. This paper makes a direct contribution that aims to bridge this gap by fine-tuning the AraGPT2-base model and evaluating it quantitatively and qualitatively.

Moethodology

To ensure the reproducibility of the experiment and support the results with strong evidence, a rigorous and structured methodology was adopted in this research. This methodology details the steps involved in constructing a specialized dataset, initial data processing, data partitioning and model structure, precise model tuning settings, and quantitative and qualitative model evaluation. **FIGURE 1** illustrates the main methodological stages followed in this study.

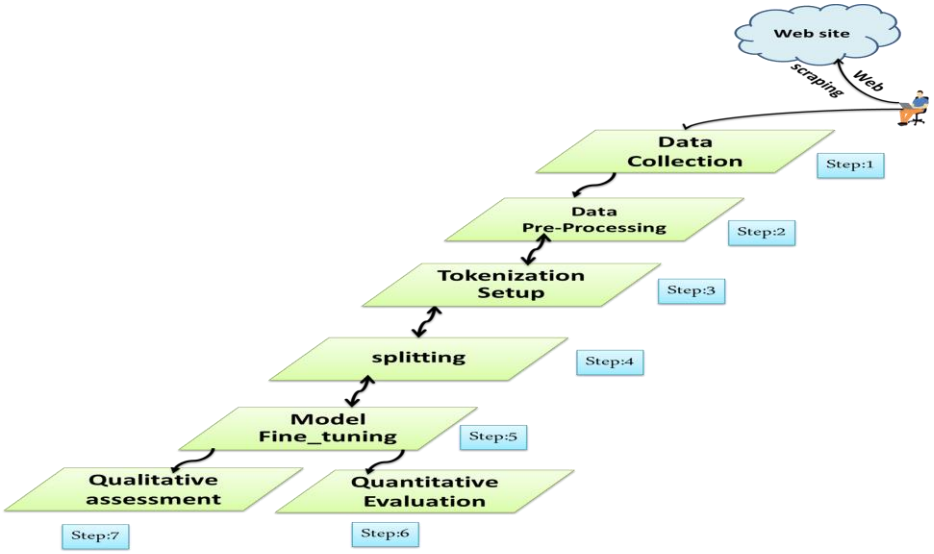

Figure 1. Steps of the methodology

1. Data Collection

Data collection is a fundamental step in any research in the field of Natural Language Processing (NLP), especially in specialized domains. To ensure the construction of a robust and focused dataset on dermatology, a hybrid methodology was adopted to collect data through sequential stages:

Web site. Data were collected from various sources, including websites specializing in dermatology. The collection process relied on three main platforms: Delta-Medlab[18], Raha

Health[19], and WebTeb[20]. This methodology resulted in the collection of a total of 20,000 specialized articles.

The distribution mechanism consisted of extracting 5,000 articles from the Delta-Medlab source, while the largest number of articles, totaling 9,000, was collected from the Raha Health website. WebTeb, in turn, contributed 6,000 articles to strengthen the database. This systematic sampling ensures a broad range of data inputs, enhancing the comprehensiveness and effectiveness of the analysis and models used in the study.

Web Scraping. is a methodology that aims to transform unstructured data from web pages into structured data that can be stored and analyzed in a central database [21].

This process was implemented using the Python programming language by developing a dedicated Web Scraping code, relying on the following basic libraries:

- *Beautiful Soup:* This library is a basic Python tool for parsing HTML and XML files. The library works in conjunction with an external parser to convert a complex HTML document into a structured parse tree of Python objects, facilitating navigation, searching, and modifying document elements [22].

- *Requests:* This library, the most important of these, is a tool for sending a standard, structured message (HTTP request) from the client program (scraper) to the hosting web server to retrieve content (usually the web page containing the required article or data). The request includes basic elements such as the request method (such as GET)[23].

Web scraping is one of the initial steps in data collection, where the content of several websites was extracted. The extraction process was defined to include the article title and content, while adhering to several controls to ensure the quality of the extracted data and its suitability for a question-answer task, where the title represents the question and the content represents the answer.

Among these controls were the automatic exclusion of articles with titles longer than the content, those in which the title or content is simply a date or link, those in which the title or content is empty, and those with a high similarity between the title and content.

In addition to several other controls adhered to during the web scraping process, approximately 20,000 articles were collected, with the quality of the extracted data rated between 90% good and 10% poor. TABLE 1 shows a summary of the characteristics of the extracted data.

Table 1. Summary of characteristics of extracted data

Characteristic	Details
Data Source	Web scraping from different websites.
Number of Scraped Data	41,810 articles.
Data Quality	90% Good / 10% Bad
Tools and Libraries	Requests library and other software libraries.
Data file	JSON.

Unstructured Data Transformation Methodology (QA-Pair Generation). To transform articles extracted from the web into a question-answer format suitable for fine-tuning (Fine-Tuning) for the Generative QA model, we adopted the following methodology:

- *Question Extraction (Query):* The article title was considered the best representation of the question or query guiding the reader. Additional filtering was performed to exclude titles that were too long or ambiguous and did not form a clear question or query.

- *Reference Answer Extraction:* To ensure that the answer was focused and appropriate for QA tasks, the first three paragraphs of the article's content were used as the reference answer. This methodological decision was made based on the belief that most extracted medical articles follow a standard web structure, where the main ideas and summary of the topic are presented at the beginning. This methodological decision was made based on a preliminary analysis of the average length of typical generative answers in the medical context and aims to provide a comprehensive and non-verbose reference for training the model for efficiency and brevity. In addition, This selection reduces the risk of including secondary and non-essential information in the reference answer, while ensuring coverage of key clinical points.

- *Automated Cleaning:* Cleaning and standardization techniques (as mentioned in Preprocessing Section 3.2) were applied to the extracted reference answer texts to remove redundant spaces, internal links, and any non-textual elements. This procedure ensures that

the Ground Truth reference answers used in training are as high-quality and accurate as possible within their linguistic context.

Dataset Augmentation. Due to the lack of diversity in the dataset extracted from the previous sites, a supplementary dataset was downloaded from the Mendeley Data Repository (site quote). This dataset, known as the AHQAD(Arabic Healthcare Question and Answer Dataset), contains large-scale healthcare data in Arabic[24]. Data samples that were not included in the dermatology dataset were excluded. After an initial screening of 41,760 datasets, random samples were selected, bringing the total number of dataset records to 21,810.

Limitations and Clinical Context. Methodologically, while the data collection and filtering process was rigorous and sourced, we acknowledge that the resulting dataset has not undergone formal verification by certified clinical dermatologists. Therefore, the current performance evaluation focuses primarily on the model's linguistic and semantic competence in generating coherent and clear responses, rather than ensuring the complete clinical accuracy of the generated content. This distinction is crucial in the context of AI in healthcare. Addressing this limitation through subsequent expert verification is a key priority in the "Future Work" section.

Ethical Considerations and Data Governance. This study adheres to all ethical and academic standards related to the responsible use of publicly available datasets.

- *Licensing and Attribution:* The supplementary dataset, the AHQAD(Arabic Healthcare Question and Answer Dataset), was obtained from the Mendeley Data Repository. This data is available to the public under a Creative Commons Attribution 4.0 International License (CC BY 4.0). The primary requirement of this license, attribution, was strictly adhered to, with the dataset and original author clearly documented in the bibliography to ensure academic and legal use of the resource[24].

- *Data Privacy and Abstraction:* The dataset used (AHQAD and web-based data) was anonymized of any personal or sensitive information, ensuring that individuals' privacy was not violated.

- *Scope of Use:* The developed language model was clearly intended as an experimental research tool. It should not be considered a substitute for clinical advice or diagnosis provided by qualified healthcare professionals.

-*Scope of Use and Transparency:* The developed language model is clearly intended as an experimental research tool, with its current validation focused on linguistic quality. It must be explicitly understood by users and readers that the model's outputs are not a substitute for clinical advice or diagnosis provided by qualified healthcare professionals.

Data Integration and Consolidation. After the data extraction process was successfully completed, the articles extracted from each site were compiled separately into JSON files. All these files were then merged into a unified JSON file. Later, this file was merged with the filtered dataset file from the Mendeley Data Repository platform, to ensure a unified data structure and facilitate subsequent processing.

2. Preprocessing

To ensure the quality of the dataset used in the study and improve the efficiency of the models used, a series of systematic preprocessing steps were applied to the data file:

- *Handling Missing Data:* All rows containing missing values (NaN) in the main columns, namely Question and Answer, were excluded and deleted.

- *Data Type Standardization and Spacing Cleansing:* The Question and Answer columns were explicitly converted to text data, with extra spaces removed from the beginnings and ends of the text.

- *Linguistic Content Check:* Rows were filtered to retain only those containing at least one letter from the Arabic or English alphabet. This procedure aims to exclude blank entries or entries containing only incomprehensible symbols.

- *Excluding Non-Textual Answers:* Additional filtering was applied to exclude any rows in which the answer consisted of only numbers or garbled punctuation symbols, ensuring that the answer was primarily textual.

After completing the cleaning and processing steps, the processed data was saved to a new .csv file to serve as the basis for all subsequent experimental and training phases. The final data size after applying this processing was 40,132 valid samples, while 1,678 irrelevant records were excluded.

3. Tokenization Setup

The encoding phase is a crucial step for converting processed text data into numerical vectors (numerical tensors) that the AraGPT2-base model can process and learn from. This phase was implemented within the QA Data set class in the PyTorch environment and followed two main steps:

Configure and customize the encoder . The GPT2TokenizerFast encoder, which is based on the Byte-Pair Encoding (BPE) algorithm, is pre-loaded from the aubmindlab/aragpt2-base

repository. Customization was necessary to adapt the model to the Conditional Language Modeling task:
-*Adding special token*:Three special tokens were inserted via tokenizer.add_special_tokens to define the input structure, as detailed in **TABLE 1.**

TABLE 1. Special symbols added to the AraGPT2 encoder for fine-tuning.

Convention	Methodological purpose
Start of sequence (BOS)	startoftext
End of sequence (EOS)	endoftext
Padding (PAD)	pad

Immediately after addition, the token embedding matrix in the GPT2LMHeadModel model was updated using the function model.resize_token_embeddings(len(tokenizer)) to synchronize the model vocabulary with the new tokens.
Data transformation and vector creation. The process of converting data into digital vectors is performed within the custom QADataset class, and follows the following sequential stages:
- *Instruction Template Construction.*The question and answer are combined into a unified text sequence that follows the structure of the instruction form, using special characters to specify context: <|startoftext|> Instruction: [Question] Answer: [Answer] <|endoftext|>
This template ensures that the model learns the relationship between the input (instruction) and the output (answer).
*Tokenization and standardization.*The encoder is applied to the merged sequence using the tokenizer() function specifying a set of critical parameters to standardize the input dimensions:
-*Trim (truncation=True):* Sets the length of the sequence to the defined maximum.
- *max_length and padding (max_length=128, padding="max_length"):* Standardizes the length of all sequences at 128 tokens by padding with <|pad|> to enable parallel training.

Creating input matrices and labels. Three main vectors (tensors) are extracted for each sample:
-*input_ids*: The digital vector representing subtokens.
-attention_mask: The binary vector (0 or 1) that instructs the attention mechanism to ignore filler codes.
-*labels:* The input_ids vectors themselves are used as labels, directing the model to perform the task of predicting the next token along the assembled sequence.
This process ensures that each data sample is ready for parallel processing by the model.

4. Dataset Splitting
To ensure a comprehensive and objective evaluation of the model's performance and mitigate the overfitting problem (which occurs when a model learns patterns specific to the training data but is unable to generalize them to new data), the processed dataset, which contained a total of 40,132 valid samples, was divided into three main sets according to standard machine learning practices.
This division represents the methodological basis for measuring the model's true performance and generalization ability. The largest set, representing 70% of the dataset (28,092 samples), was designated the training set. The validation set, representing 20% of the dataset (8,026 samples), was used to guide the training process and protect the model from overfitting. Finally, the test set, which comprised 10% of the dataset and totaled 4,014 samples, was kept completely neutral to provide an unbiased assessment of the model's ability to generalize to new data it had never seen. This systematic distribution ensures that the actual performance of the model is measured with high accuracy.

5. Fine-Tuning Model
. The fine-tuning phase represents a pivotal step in adapting pre-trained large language models (PLLMs) to specialized and domain-specific tasks. Although the AraGPT2-base model has a broad knowledge base gained from pre-training on a vast amount of Arabic text, it lacks the contextual depth and terminological precision needed to generate reliable answers in a specialized medical field such as dermatology. Accordingly, a transfer learning methodology was applied to fine-tune the AraGPT2-base model, where the model's weights are incrementally updated using the small, focused research dataset. This modification aims to transform the model from a general text generator into an expert system specialized in generative QA tasks in dermatology, while maintaining the general language capabilities it previously acquired. The following subsections discuss the architecture of the AraGPT2-base model and details the settings used in the tuning phase.
Model and Architecture. The AraGPT2-base model is built on the groundbreaking Transformer architecture, a neural network architecture originally proposed to completely eliminate the complex recurrence and convolutional neural network mechanisms prevalent in data sequence models.
The Transformer instead relies entirely on self-attention mechanisms to extract global

correlations between inputs and outputs. This reliance on self-attention enables significantly higher parallelization during training compared to recurrent models. 3 AraGPT2 is a large-scale Arabic language model pre-trained on over 77 GB of diverse Arabic text, including the Arabic Wikipedia, the OSCAR corpus, and other journalistic and literary sources[25]As shown in TABLE 2, the architectural specifications of the AraGPT2 model, several scalable models have been released.

Table 2. Architectural specifications of AraGPT2 model versions.

The model	Optimizer	Context Size	Embedding Size	Num of Heads	Num of Layers
AraGPT2-base	LAMB	1024	768	12	12
AraGPT2-medium	LAMB	1024	1024	16	24
AraGPT2-large	Adafactor	1024	1280	20	36
AraGPT2-mega	Adafactor	1024	1536	25	48

FIGURE 2. illustrates the structure of the AraGPT2-base model.

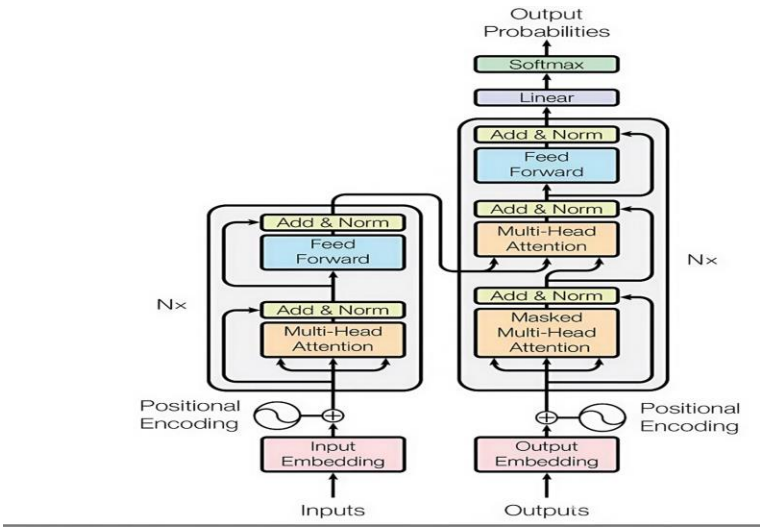

Figure 2. Transformer Architecture ARAGPT2[26].

Mathematical Basis of the Attention Mechanism. The basic architecture of the AraGPT2-base model is based on a Scaled Dot-Product Self-Attention (SDPSA) mechanism. This mechanism enables the model to determine weights that reflect the importance of each word relative to other words within a sequence, allowing it to accurately capture distant contextual associations without the need for iteration. It is implemented by calculating the dot product between the query vectors (Q) and the keys (K), then applying a Softmax function to obtain the weights, which are subsequently used to sum the value vectors (V). The model's power is enhanced by multi-head attention, which applies the process in parallel to capture different patterns of linguistic associations, while positional encoding is added to incorporate information about the temporal order of words.

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)$$

This equation explains how attention weights are assigned between elements of an input sequence, allowing the model to focus computational resources on the most important contextual connections[26].

FIGURE 3. illustrates the graphical mechanism of the self-attention process in the AraGPT2-base model.

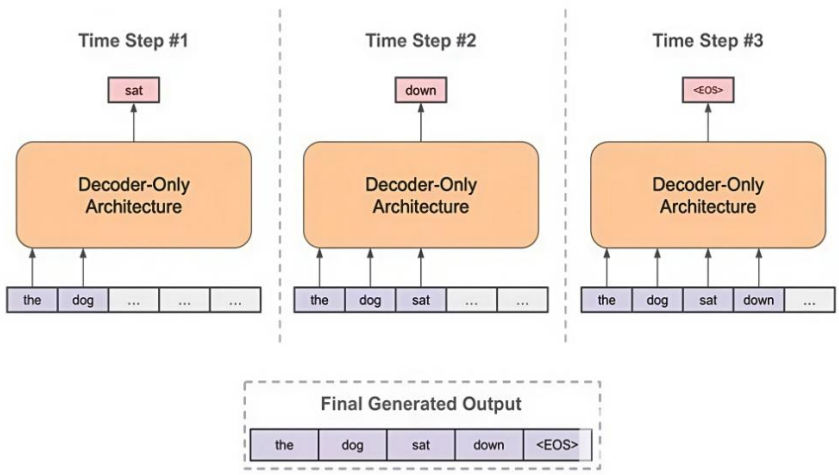

Figure 3. How to work ARAGPT2[26].

Fine-Tuning Configuration. The process of adapting the pre-trained Arabic language model, AraGPT2-base, to a specialized question-answering system in the medical field (dermatology) represents a systematic application of transfer learning principles. The critical parameters used in the fine-tuning process, shown in Table 3, reflect a practical application of advanced training strategies for large transformer models.

These values were chosen to meet three balanced goals: first, to overcome GPU memory limitations; second, to maintain a stable and efficient update rate for the model weights across the gradient; and third, to ensure that the model maintains its generalization ability by avoiding overfitting.

To achieve this, the batch size was set to 2 to meet available memory constraints, effectively offset by increasing the gradient accumulation steps to 64. This systematic arrangement results in an effective batch size of 128, a suitable number to ensure stable and reliable updates to the model parameters throughout the optimization process. Setting the maximum sequence length to 128 symbols is a logical choice for the Q&A task, as it balances the need for context with computational efficiency. The use of early stopping with a patient duration of 2 epochs is a precaution to regularize the model and protect it from overfitting to specialized data, especially since training is performed incrementally on chunks of 8,000 samples. Together, these settings ensure a structured and optimized training in terms of computational efficiency and final model performance.**TABLE 3** shows the parameter settings used in the fine-tuning phase.

Table 3. Fine-tuning parameter settings.

Parameter	Value
Number of Epochs	3 per Chunk
Batch Size	2
Gradient Accumulation Steps	64
Max Length	128
Learning Rate	Default (≈5e-5)
Early Stopping	Patience = 2
Chunk Size	8,000 samples
Device	GPU – CUDA

6. Quantitative Evaluation

To ensure a comprehensive and objective evaluation of the effectiveness of the fine-tuned model in the generative dermatology question-answering task, a set of specialized metrics was adopted to evaluate the quality of generated texts. These metrics provide a quantitative assessment of various aspects of the model's performance:

- BERTScore (F1): This metric was chosen to measure the semantic similarity between the generated answer and the reference answer. BERTScore relies on transformer models to understand context and meaning, making it more suitable for evaluating the quality of generative answers. Its focus on content and meaning makes it more suitable for evaluating generative answers than traditional semantic matching metrics[27].

- *BLEU (Bilingual Evaluation Understudy):* This metric is used to measure lexical similarity between texts based on n-grams. This measure is an indicator of the degree to which the grammatical and phonetic similarity of the generated responses matches the reference texts. A low score in open generation is used as evidence of a model's preference for generating new formulations rather than literal retrieval. This is an expected result and not necessarily an indicator of incorrectness[28].
- *Levenshtein Distance:* This measure measures the letter-by-letter similarity of text strings by calculating the minimum number of modifications (insertions, deletions, and substitutions) required to convert the generated text to the reference text. 5 Its score is used to confirm the limited literal matching, assess the extent to which the syntactic structure of the output differs from the reference text, and confirm the limited verbal matching[28].

7. Qualitative Assessment

To enhance the reliability of the quantitative results and provide an assessment of the quality of the qualitative outputs, a preliminary comparative study was conducted of the proposed model (Model B) against three of the world's leading large language models (LLMs): ChatGPT (A), DeepSeek (C), and Gemini (D). These models were selected to cover a wide range of currently available LLMs and to provide a solid basis for comparing the performance of the proposed model with other models selected for comparison.

Evaluation Objective and Methodology. The methodology was based on a specialized questionnaire distributed via a published link, in which nine (9) members of the general public participated as evaluators. This evaluation aimed to assess the quality of the generated responses and compare the proposed model with selected models in terms of linguistic clarity, grammatical consistency, and readability of the generated medical content from the perspective of a non-expert user. It is important to clarify that this initial human evaluation was specifically designed to assess linguistic clarity and readability for the general public, which is directly related to the human element of AI-patient interaction. This evaluation differs in scope from the evaluation of clinical accuracy, which should be conducted by qualified medical professionals. A five-point Likert scale was used to assess four key criteria, including the linguistic clarity of the medical content. To ensure methodological transparency, **TABLE 4.** shows the four criteria used by the human evaluators and how they were measured.

Table 4. Human qualitative assessment criteria.

Criterion	Brief description of the standard	Likert scale
Linguistic quality	Correctness of grammar and morphology, and consistency of formulation.	1 = Very weak, 5 = Excellent.
Linking the answer to the question	The extent to which the generated text directly answers the question posed.	1 = Very weak, 5 = Excellent.
Answer length	Balance between brevity and comprehensiveness.	1 = Very weak, 5 = Excellent.
Linguistic clarity of medical content	How easy it is to understand medical content from a non-specialist perspective.	1 = Very weak, 5 = Excellent.

To illustrate the nature of the inputs evaluated and to establish the basis upon which the general evaluators' assessment was built, **TABLE 5** provides an example of the questions used in the questionnaire used in the qualitative assessment.

Table 5. Sample qualitative assessment questions.

Question No.	Question
Q1	هل *عرق السوس مضر للوجه فقد سمعت خلطة تحتوي على عرق السوس توضع على الوجه وأرغب تجريبها ولكن بعد التأكد أنها تسبب الضرر للوجه؟*
Q2	اين ياتي فيروس الهربس وما طرق انتقاله بين الأشخاص عندما يظهر حول الفم وهل عدوى

	خطيرة؟
Q3	ماذا يحدث بعد جلسة الديرما وكم يبقى أثرها أثر مؤقت يستمر فترة طويلة بعد الجلسة؟
Q4	كيف تصنع بخاخ اكليل الجبل للشعر وما فوائد زيت اكليل الجبل للشعر يفيد تحسين نمو الشعر؟
Q5	متى يبدأ مفعول أوميغا للبشره علما أنني أتناول مكملات الأوميغا منذ أسبوع تقريبا ولم ألاحظ الفرق بعد؟

TABLE 6. shows a sample of the questions used in the survey. These questions were designed to address challenges that require generating open-ended responses, covering various aspects of diseases and natural remedies, in particular.

8. Evaluation Framework

While the quantitative evaluation (Levenshtein distance and BERTScore and BLEU) was conducted exclusively on our fine-tuned AraGPT2 model, the comparison with global models (ChatGPT, DeepSeek, and Gemini) was performed using a human-centered qualitative assessment. This methodology was chosen because these global models are closed-source (black-box), which limits the ability to access their internal probability distributions or ensure consistent tokenization required for certain automated metrics. Therefore, we prioritized a user-centric qualitative benchmarking to evaluate linguistic clarity and medical relevance in a real-world context, as this provides more meaningful insights into the model's practical utility for non-specialist Arabic speakers.

Results And Discussion

This section is designed to present the quantitative and qualitative results achieved and then discuss the performance implications of the modified AraGPT2-base model in the context of answering dermatological medical questions.

Model Training Results

The model training process on the question and answer dataset demonstrated positive and effective convergence behavior. This performance is attributed to the model's internal architecture and the hyperparameters used. Typically used text generation models (such as Transformer models) rely on self-attention mechanisms, which enable them to efficiently capture complex and long-range relationships between question and answer segments. This architecture contributed to a sharp and rapid decrease in the loss value in the initial training phases.

The results show that the model maintained good generalization ability, as the difference between the training loss and the validation loss remained small, demonstrating that it overcomes the high bias problem without falling into severe overfitting.

To understand the subtle changes in model performance, the training loss and validation loss values were monitored for each training epoch within each dataset. **TABLE 6** presents these values in detail.

Table 6. Training and validation losses.

Chunk	Epoch	Training Loss
(0–8000)	1	8.9464
	2	7.4158
	3	5.4399
(8000–16000)	1	2.4996
	2	1.7216
	3	1.6504
(16000–24000)	1	1.6973
	2	1.5867
	3	1.5441
(24000–28092)	1	1.6868
	2	1.6180
	3	1.5908

The performance is shown in the **TABLE 6** as follows:

- *Rapid Convergence:* In the first batch, the training loss decreased dramatically. The reduction in the training loss from 8.9464 to 5.4399 in the first batch indicates a sharp and effective reduction, demonstrating the efficiency of the optimizer algorithm used, which allowed the model to quickly adjust its weights to meet the requirements of the training set.

- Loss Gap: In most epochs, the validation loss remained slightly higher than the training loss (the actual loss in the third epoch of the final batch was 1.5908 versus 2.0275). This

difference indicates a slight and acceptable overfitting and reflects the model's reliable generalization ability. However, it cautions the need to use regularization techniques to control this discrepancy.

- *Loss Behavior Across Batches:*

It is noted that the validation loss reached its lowest level in the third epoch of the final batch, at 2.0275, indicating that this point represents the model's best performance on unseen data.

The high value of the training loss in the second epoch of the final batch (1.618) indicates a temporary spike. This spike may have resulted from a relatively difficult training batch load, or it may have been the result of an adjustment in the learning rate scheduling that forced the model to explore a new point in the parameter space. However, the model quickly recovered its performance in the subsequent epoch (1.5908).

- *Sustained Performance:* Across batches, the final losses stabilized at relatively low levels (below 2.1) after initial convergence, confirming that the model had learned an effective representation of the data, and that successive batch training was necessary to ensure exposure to the entire dataset.

As also shown in **FIGURE 4**, the graph of Table 7 shows the dynamic behavior of the training loss and validation loss values across the 12 training epochs, divided into chunks, providing a clear graphical view of the convergence process.

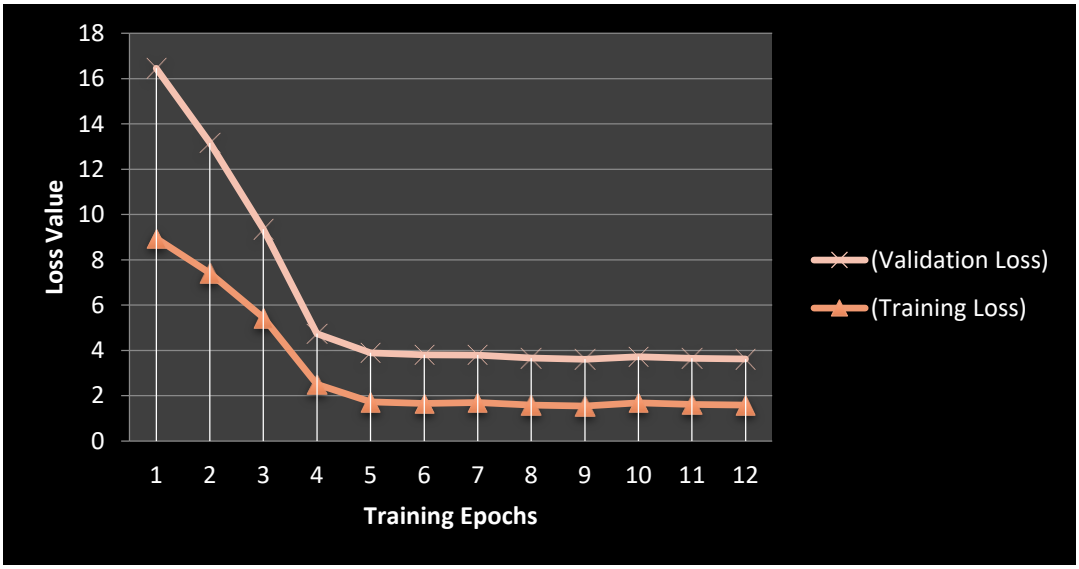

Figure 4. Training and validation loss curve.

Evaluation Results (Quantitative evaluation)

To provide an accurate and detailed numerical assessment of the model's performance, final numerical values for the main evaluation metrics were compiled. This serves as a basis for assessing the model's success in achieving the desired goals in terms of the semantic and syntactic quality of the output. To assess the actual performance of the proposed model, a variety of quantitative metrics were used, covering both semantic and syntactic aspects of the quality of the textual output. These metrics are essential for providing an objective and comprehensive assessment beyond mere verbal accuracy. These metrics include BERTSCORE (F1) for measuring semantic similarity, BLEU for estimating syntactic and syntactic matching accuracy, and Levenshtein distance. **TABLE 7** presents these final results, enabling quantitative analysis and direct comparison of the effectiveness of the different aspects of the model.

Table 7. Results of quantitative evaluation of model performance.

Scale	value
BERTSCORE (F1)	64.49%
BLEU	10.00%
Levenshtein distance	28.13%

TABLE 7. presents the values for the three quantitative evaluation metrics used. The table shows that the BERTSCORE (F1) measure dominates, achieving a score of 64.49%, confirming the model's strong performance in semantic similarity. In contrast, the BLEU value is lowest at 10.00%, highlighting the challenge the model faces in lexical exactness. The Levenshtein measure, on the other hand, achieves a score of 28.13%. The variance in the values in the table confirms that the model's main strength lies in understanding meaning rather than accurately mimicking the reference syntax.

The overall results for these measures, as shown in **FIGURE 5**, illustrate the discrepancy in model performance between deep meaning comprehension and syntactic accuracy.

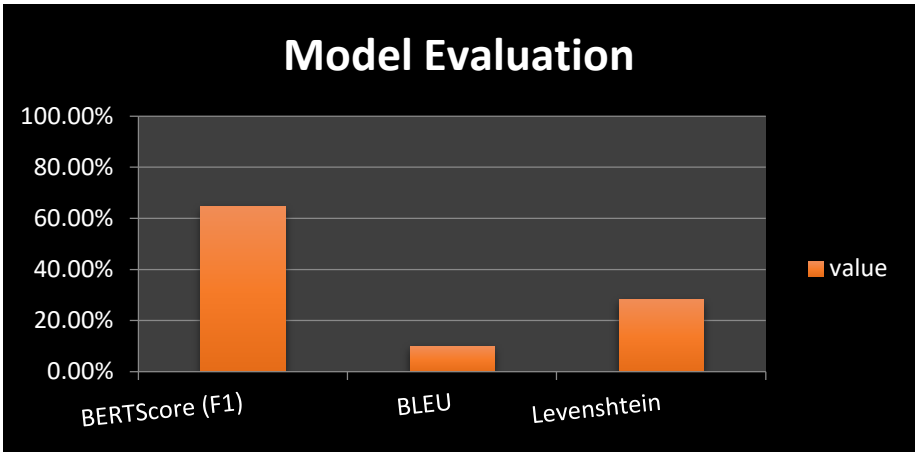

Figure 5. Comparison of model output quality assessment metrics.

FIGURE 5. provides a clear graphical representation of the achieved evaluation results. It is evident that the BERTSCORE (F1) measure achieved the highest value at 64.49%. This high score indicates the model's success in effectively capturing linguistic meanings and contexts, consistent with the use of transformer-based model architectures. In contrast, the BLEU measure showed the lowest performance at only 10.00%, while the Levenshtein measure reached 28.13%.

A decrease was recorded in the literal matching measures (BLEU 10.00%, Levenshtein distance 28.13%). This decrease does not reflect a weakness in semantic understanding (where the BERTScore F1 is excellent), but rather confirms the required capacity for free generation and creative paraphrasing. This characteristic is what allowed the model to generate responses with higher linguistic clarity than if it had relied on the literal retrieval of complex original reference texts.

The contrast between BERTSCORE and the BLEU and Levenshtein distance metrics reflects that the model generates texts with high semantic quality but lacks precise matching at the level of word sequences or lexical fidelity compared to the reference texts.

due to the restricted access to the internal weights and architectures of commercial black-box models, such as ChatGPT and Gemini and deepseek. This lack of transparency makes it technically unfeasible to calculate identical metrics like BERTScore or Levenshtein distance across all models using a uniform methodology. Therefore, a qualitative assessment and human evaluation were employed to ensure a fair and rigorous comparison within the context of practical, real-world utility.

Benchmarking (Initial Qualitative Evaluation)

Qualitative evaluation based on human judgment is a crucial step in assessing the actual performance of large language models (LLMs) in the context of specialized generative applications. Since the task of answering dermatological medical questions requires more than just verbal accuracy, an initial comparative study was conducted to measure the quality of the output of the fine-tuned model (the proposed model - B) against three leading global models (ChatGPT - A, DeepSeek - C, and Gemini - D). This evaluation aimed to measure the perceptual satisfaction and procedural effectiveness of the responses from a non-expert user's perspective, based on the four basic criteria.

Results of Statistical Analysis for Human Performance Evaluation. The systematic evaluation of human qualitative performance relied on the separation and analysis of basic statistical criteria. Arithmetic means (means) were used to estimate the overall efficiency of each model, and standard deviations (SDs) were separated to assess the consistency and statistical reliability of the performance of competing models.

Analyzing Mean Scores. **TABLE 8** presents the arithmetic means recorded for the performance of the four models across the five qualitative criteria, which reflect the overall performance of each model from the perspective of the non-specialist audience (the evaluators). **FIGURE 6** also shows a graphical representation of these means for visual comparison.

Table 8. Mean scores for the four models.

The model	Linguistic quality	Linking the answer to the question	Answer length	Linguistic clarity of medical content	Overall average
ChatGPT	4.17	4.38	4.14	4.20	4.22
Proposed model	3.58	3.68	3.86	4.55	4.01
DeepSeek	4.42	4.19	4.00	4.19	4.20
Gemini	3.75	3.67	4.25	4.45	4.03

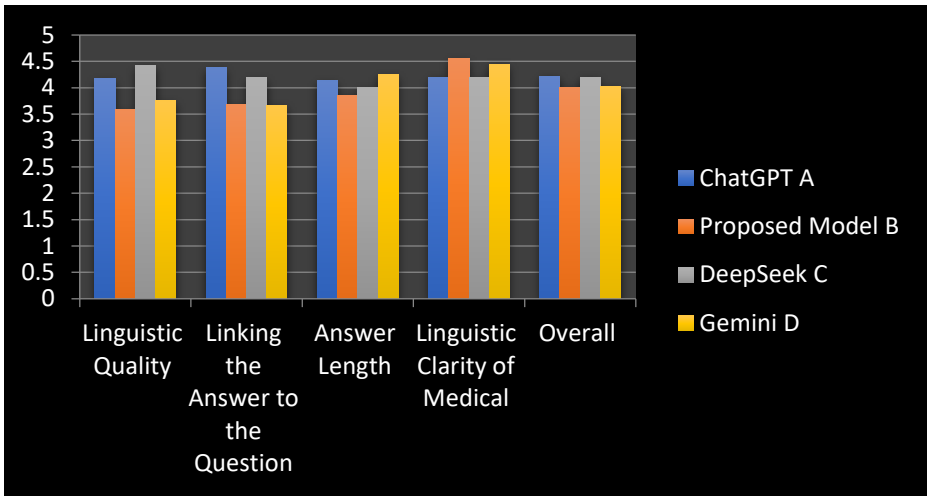

Figure 6. Comparison of mean scores for the four models.

By analyzing the data in Table 8 and **FIGURE 6**, the following results were observed:

- *Linguistic Quality:* DeepSeek model topped this criterion with an average score of 4.42.
- *Question-Answer Relevance:* ChatGPT model outperformed, recording the highest average score of 4.38, indicating its ability to accurately understand the context of the query.
- *Linguistic Clarity of Medical Content:* The proposed model demonstrated a significant qualitative superiority in this sensitive criterion, recording the highest average score of 4.55 among all models. This superiority effectively explains the fine-tuning process of adapting the language to meet the needs of the general audience and simplifying medical terminology.
- *Overall Performance:* ChatGPT model came in first with an overall average score of 4.22, followed by DeepSeek with an average score of 4.20. The proposed model achieved competitive performance with an overall average of 4.01.

Standard Deviation Scores Analysis. **TABLE 9** shows the standard deviation (SD) values associated with each mean, which are considered a key indicator of the degree of variance and statistical inconsistency between the evaluators' scores. A low value indicates a high level of agreement and reliability in the assessment. **FIGURE 7** also graphically represents these values.

Table 9. Standard deviations (SDs) for the four models.

The model	Linguistic quality(SD)	Linking the answer to the question(SD)	Answer length(SD)	Linguistic clarity of medical content(SD)	Overall average(SD)
ChatGPT	0.81	0.94	1.01	0.86	0.77
Proposed model	1.25	1.15	1.08	1.45	1.23
DeepSeek	0.76	0.76	1.13	0.76	0.78
Gemini	1.25	1.17	1.06	1.06	1.13

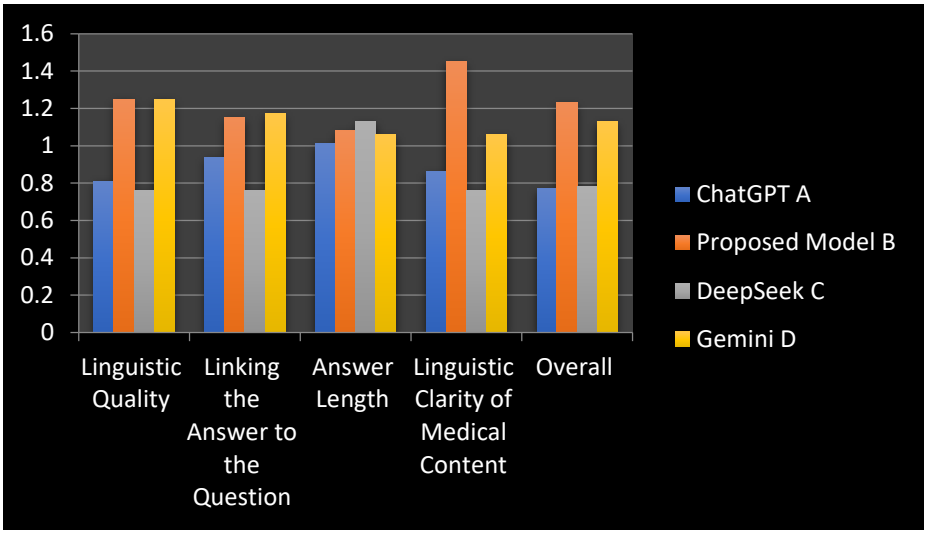

Figure 7. Standard deviation (SD) variance of the four models.

By analyzing the data in **TABLE 9** and **FIGURE 7**, the following results were observed:

- *Overall consistency:* DeepSeek model is the most consistent and reliable, recording the lowest standard deviation of 0.76 across three main criteria (linguistic quality, answer relevance, and medical clarity), confirming a high level of statistical agreement between raters regarding the quality of its output.

- *Variability in the proposed model:* Proposed model recorded the highest standard deviation for the linguistic clarity of medical content criterion, with a sharp value of 1.45. This value is the highest among all models and criteria, indicating a sharp decline in statistical agreement between raters regarding Model B's performance on this criterion, despite achieving the highest average.

The discrepancy between the mean and SD of the proposed model represents a pivotal point of analysis: despite its superiority in linguistic clarity (mean: 4.55), the high variance (SD: 1.45) is explained by the fact that this clarity was not comprehensive and consistent across all assessment cases.

The proposed model recorded the highest standard deviation (1.45) on the linguistic clarity criterion. This sharp variation in the residents' agreement suggests that the definition of "clarity" in medical content may differ significantly among non-specialist individuals, and does not necessarily indicate inconsistency in the model's outputs. We suggest that this variation could be significantly reduced by better calibrating the residents or using residents with a medical background, which is something we plan to address in our future research.

That is, the proposed model generated very simple and clear answers in some cases (which raised the mean), but failed miserably to address complexity or adhere to comprehensiveness in others (which raised the standard deviation). (This discrepancy is primarily due to the model's radical tendency toward free generation, which led to oversimplification or a failure to incorporate all the clinical details necessary for comprehensiveness in complex cases, and these answers were strongly rejected by the assessors.)

This discrepancy highlights the challenge of balancing linguistic specialization (represented by high clarity) with comprehensive consistency (represented by low standard deviation). Larger, more highly trained models such as DeepSeek and ChatGPT achieved higher statistical consistency (lower SD) at the expense of qualitative superiority in linguistic clarity, suggesting that the proposed model's fine-tuning came at the expense of general measures of consistency and comprehensiveness.

To provide a holistic view of the performance across all tested systems, **TABLE10** consolidates the quantitative metrics of the proposed model with the qualitative human-centered evaluation for both the proposed and global models (ChatGPT, DeepSeek, and Gemini). It is important to note that quantitative metrics such as BERTScore, BLEU, and Levenshtein distance were calculated exclusively for the fine-tuned AraGPT2 model, as the global models are closed-source (black-box), preventing identical automated benchmarking. This consolidated presentation allows for a direct comparison between technical semantic alignment and practical user-centric utility.

TABLE 10. Consolidated Performance Summary of the Proposed Model vs. Global LLMs.

Model	BERTScore (F1)	BLEU	Levenshtein Distance	Linguistic Quality	Answer-Question Linking	Answer Length	Medical Clarity	Overall Average
Proposed Model	64.49%	10.00%	28.13%	3.58	3.68	3.86	**4.55**	04.01
ChatGPT	N/A	N/A	N/A	4.17	4.38	4.14	4.20	4.22
DeepSeek	N/A	N/A	N/A	4.42	4.19	4.00	4.19	4.20
Gemini	N/A	N/A	N/A	3.75	3.67	4.25	4.45	04.03

The consolidated results in **TABLE 10** reveal several key insights. First, the proposed model achieved the highest score in Medical Clarity (4.55), outperforming larger global models. This suggests that domain-specific fine-tuning on a specialized Arabic dermatology dataset significantly enhances the model's ability to provide clear and accessible medical information to non-expert users.

Regarding the quantitative performance of the proposed model, the BERTScore (64.49%) indicates strong semantic understanding, while the BLEU (10.00%) and Levenshtein Distance (28.13%) reflect a lower lexical overlap. This divergence is not an indication of poor performance but rather confirms the model's tendency toward free generation—producing original, contextually relevant responses rather than verbatim retrieval from the training data. While global models like ChatGPT showed a higher overall qualitative average, our model's superiority in specialized medical clarity demonstrates its particular value for targeted healthcare communication in the Arabic language.

Qualitative Case Study. To examine the nature of variation in performance, particularly the weakness in the contextual relevance of the proposed model, **TABLE 11** presents a content analysis of the answers generated to the first question (Q1), as stated in the methodology (**TABLE 6**).

Table 11. COMPETING models' responses to the first question.

The model	السؤال: هل عرق السوس مضر للوجه فقد سمعت خلطة تحتوي على عرق السوس توضع على الوجه وأرغب تجريبها ولكن بعد التأكد أنها تسبب الضرر للوجه؟
ChatGPT	عرق السوس قد يكون له فوائد للبشرة مثل التفتيح وتقليل الالتهابات، لكن استخدامه المباشر أو بتركيز عالٍ قد يسبب تهيجاً أو تحسساً. يفضل استشارة طبيب الجلدية قبل استخدام أي خلطات منزلية.
Proposed model	علاج مناسب للبشرة الدهنية
DeepSeek	عرق السوس مفيد جدًا للوجه عند استخدامه بشكل صحيح. مستخلصه له خصائص مضادة للالتهابات ومبيضة للبشرة (يقلل التصبغات) ومهدئة. بشكل عام هو غير ضار، لكن يجب اختباره على منطقة صغيرة من الجلد أولاً. يفضل استخدام منتجات جاهزة ذات تركيز معروف بدلاً من الخلطات المنزلية غير الدقيقة.
Gemini	عرق السوس آمن للاستخدام الموضعي على الوجه لمعظم الأشخاص، ويُعرف بخصائصه المفتحة للبشرة والمضادة للالتهابات. ومع ذلك، قد تحدث بعض الآثار الجانبية النادرة مثل تهيج البشرة أو الحساسية. لذا، يفضل دائمًا إجراء اختبار حساسية على منطقة صغيرة من الجلد قبل استخدامه على كامل الوجه.

TABLE 11. shows that the proposed model failed to understand the context in this case, generating only an unrelated response ("A suitable treatment for oily skin"). The low contextual relevance explains the low scores on the relevance and response length criteria, and confirms that success in achieving linguistic clarity was not always accompanied by contextual accuracy.

Overall, the qualitative evaluation shows that the proposed model succeeded in achieving the specialization goal (linguistic clarity) but failed to match the overall improvements of the competing models (DeepSeek and Gemini) on the measures of overall proficiency and confidence. This discrepancy reinforces the conclusion that focus should be placed on enhancing the comprehensiveness of responses

and improving linguistic procedures to raise its overall performance and ensure the clinical integrity of responses.

Conclusion

This paper summarizes a comprehensive effort to evaluate and customize Arabic large language models (LLMs) for the task of answering specialized questions in the field of dermatology. The research succeeded in achieving its main goal by fine-tuning the AraGPT2-base model on a focused dataset.

Quantitative evaluation results confirmed the strong semantic performance of the proposed model, with a BERTScore F1 of 64.49%, demonstrating the model's ability to effectively understand medical context and meaning. In contrast, the verbal matching metrics (BLEU: 10.00%, Levinstein ratio: 28.13%) reflected the model's intrinsic tendency toward free generation of novel answers rather than literal matching to reference data.

As for qualitative performance, benchmarking against global models showed that the proposed model achieved a competitive overall average performance score of 4.01 in the initial human evaluation for a non-expert audience. More importantly, the model's superiority in linguistic specialization was confirmed, with the model achieving the highest score (4.55) in the linguistic clarity and readability criteria for the general audience. However, statistical analysis revealed high variance (SD = 1.45) in the human evaluation, indicating that the strength of linguistic clarity led to inconsistency in overall performance.

This study confirms the high feasibility of using customized Arabic models as a support tool for delivering understandable medical information to the general public, with a future focus on enhancing their clinical accuracy and the comprehensiveness of their responses to reduce variability and ensure content integrity.

Limitations And Future Directions

Despite the promising results, the current study encountered several structural and methodological limitations that will shape our future research agenda:

Main Limitations.

-Data Source and Clinical Validation: The dataset was generated from public web sources; this imposes a fundamental limitation on the immediate clinical validity of the reference responses. Currently, the study lacks direct verification by board-certified dermatologists to ensure medical safety.

-Initial Qualitative Assessment and Statistical Scope: Human evaluation was conducted by the general public to measure linguistic clarity. However, this assessment relied primarily on descriptive statistics and qualitative descriptions. The study did not yet incorporate rigorous statistical significance tests (e.g., p-values) to measure the reliability of agreement between raters.

-Model Size and Resource Constraints: Fine-tuning was performed using the AraGPT2-base model (135 million parameters). This size may not be sufficient to capture high cognitive and medical complexity compared to larger LLMs.

-Methodological Challenge (Clarity vs. Consistency): The high variability (SD = 1.45) in human evaluation highlights the difficulty of balancing free-generation clarity with the necessary consistency and safety of biomedical information.

Future Directions. Based on these limitations, future research recommendations prioritize establishing the clinical reliability of the model and expanding its capabilities:

-Rigorous Clinical Validation: Formal clinical validation is the immediate priority. We plan to involve a dedicated team of dermatologists to conduct a rigorous evaluation of the accuracy and safety of the generated responses.

-Enhanced Consistency Through RAG and PEFT: To mitigate hallucinations and the clarity-versus-consistency trade-off, we recommend integrating **(**RAG**)**. Furthermore, exploring **(**PEFT**)** techniques could optimize the model's performance while maintaining computational efficiency.

-Advanced Statistical Analysis: Future work will include inferential statistical analysis, such as significance testing between evaluators' scores, to provide a more robust scientific foundation for the results.

-Exploring Model Scaling: Fine-tuning will be tested using larger architectures (such as larger versions of AraGPT2 or Arabic-Llama constructs) to verify the impact of scaling on medical reasoning.

-Ethical and Normative Development: Developing a clinically validated test dataset and implementing methodologies to mitigate linguistic bias are essential steps toward ensuring the accountability of the model output.

References

[1] Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y.: TNT-Transformer in Transformer. Adv. Neural Inf. Process. Syst. 19, 15908–15919 (2021).

[2] Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., Wang, Y.: Artificial intelligence in healthcare: Past, present and future. Stroke Vasc. Neurol. 2, 230–243 (2017). https://doi.org/10.1136/svn-2017-000101.

[3] Lewandowski, M., Łukowicz, P., Świetlik, D., Barańska-Rybak, W.: ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the Specialty Certificate Examination in Dermatology. Clin. Exp. Dermatol. 49, 686–691 (2024). https://doi.org/10.1093/ced/llad255.

[4] Joh, H.C., Kim, M.H., Ko, J.Y., Kim, J.S., Jue, M.S.: Evaluating the Performance of ChatGPT in Dermatology Specialty Certificate Examination-style Questions: A Comparative Analysis between English and Korean Language Settings. Indian J. Dermatol. 69, 338–341 (2024). https://doi.org/10.4103/ijd.ijd_1050_23.

[5] Yazar, S., Corresponding, /, Goceri, E., Prof, A.: An Application for Automated Diagnosis of FacialDermatological Diseases. İzmir Kâtip Çelebi Üniversitesi Sağlık Bilim. Fakültesi Derg. 6, 91–99 (2021).

[6] Guellil, I., Saâdane, H., Azouaou, F., Gueni, B., Nouvel, D.: Arabic natural language processing: An overview. J. King Saud Univ. - Comput. Inf. Sci. 33, 497–507 (2021). https://doi.org/10.1016/j.jksuci.2019.02.006.

[7] Gummadi, V. P. K. (2020). API design and implementation: RAML and OpenAPI specification. Journal of Electrical Systems, 16(4).

[8] Alajmi, A., Altabaa, H., Abed, S., Ahmad, I.: Arabic Question Generation Using Transformers. ACM Trans. Asian Low-Resource Lang. Inf. Process. 24, 1–21 (2025). https://doi.org/10.1145/3701559.

[9] Khedimi, S., Bouziane, A., Bouchiha, D.: Advancements and challenges in Arabic question answering systems: a comprehensive survey. Brazilian J. Technol. 7, e75604 (2024). https://doi.org/10.38152/bjtv7n4-028.

[10] SAOUDI, Y., GAMMOUDI, M.M., YEFERNY, aoufik: FAAQA-QAD: A Frequently Asked Arabic Question Answering Dataset for Diseases Detection System. Commun. Int. Proc. (2023). https://doi.org/10.5171/2023.4115023.

[11] Selcuk, Y., Kim, E., Ahn, I.: InfectA-Chat, an Arabic Large Language Model for Infectious Diseases: Comparative Analysis. JMIR Med. Informatics. 13, (2025). https://doi.org/10.2196/63881.

[12] Yim, W.W., Fu, Y., Sun, Z., Abacha, A. Ben, Yetisgen, M., Xia, F.: DermaVQA: A Multilingual Visual Question Answering Dataset for Dermatology. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) . 15005 LNCS, 209–219 (2024). https://doi.org/10.1007/978-3-031-72086-4_20.

[13] Saeed, N.: MediFact at MEDIQA-M3G 2024: Medical Question Answering in Dermatology with Multimodal Learning. Clin. 2024 - 6th Work. Clin. Nat. Lang. Process. Proc. Work. 339–345 (2024). https://doi.org/10.18653/v1/2024.clinicalnlp-1.31.

[14] Bauer, M., Seibold, C.M., Kleesiek, J., Dada, A.: IKIM at MEDIQA-M3G 2024: Multilingual Visual Question-Answering for Dermatology through VLM Fine-tuning and LLM Translations. Clin. 2024 - 6th Work. Clin. Nat. Lang. Process. Proc. Work. 439–447 (2024). https://doi.org/10.18653/v1/2024.clinicalnlp-1.44.

[15] Al Wazrah, A., Altamimi, A., Aljasim, H., Alshammari, W., Al-Matham, R., Elnashar, O., Amin, M., AlOsaimy, A.: Evaluation of Large Language Models on Arabic Punctuation Prediction. Proc. - Int. Conf. Comput. Linguist. COLING. 144–154 (2025).

[16] Lakim, I., Almazrouei, E., Alhaol, I.A., Debbah, M., Launay, J.: A Holistic Assessment of the Carbon Footprint of Noor, a Very Large Arabic Language Model. 2022 Challenges Perspect. Creat. Large Lang. Model. Proc. Work. 84–94 (2022). https://doi.org/10.18653/v1/2022.bigscience-1.8.

[17] Fasha, M., Hammo, B., Sowan, B., Barham, H., Al-Nsour, E.: Parameter Efficient Fine Tuning Llama 3.1 for Answering Arabic Legal Questions: A Case Study on Jordanian Laws. 2025 1st Int. Conf. Comput. Intell. Approaches Appl. ICCIAA 2025 - Proc. (2025). https://doi.org/10.1109/ICCIAA65327.2025.11013375.

[18] Delta Medical Laboratories, https://delta-medlab.com/, last accessed 2025/10/29.

[19] Raha Center for Comprehensive Medical Care Services in Riyadh, https://rahahealth.com.sa/, last accessed 2025/10/29.

[20] WebTeb - Information I Trust, https://www.webteb.com/, last accessed 2025/10/29.

[21] Khder, M.A.: Web scraping or web crawling: State of art, techniques, approaches and

application. Int. J. Adv. Soft Comput. its Appl. 13, 144–168 (2021). https://doi.org/10.15849/ijasca.211128.11.

[22] Richardson, L.: Beautiful Soup Documentation Release 4.4.0. Media.Readthedocs.Org. 1–72 (2019).

[23] Farhadi, V., Mehmeti, F., He, T., Porta, T. La, Khamfroush, H., Wang, S., Chan, K.S.: Service Placement and Request Scheduling for Data-intensive Applications in Edge Clouds. Proc. - IEEE INFOCOM. 2019-April, 1279–1287 (2019). https://doi.org/10.1109/INFOCOM.2019.8737368.

[24] Gawbah, H.: AHD: Arabic Healthcare Dataset. 6, (2024). https://doi.org/10.17632/MGJ29NDGRK.6.

[25] Models – Hugging Face, https://huggingface.co/models, last accessed 2025/10/29.

[26] Mohiuddin, K., Welke, P., Alam, M.A., Martin, M., Alam, M.M., Lehmann, J., Vahdati, S.: Retention Is All You Need. Int. Conf. Inf. Knowl. Manag. Proc. 4752–4758 (2023). https://doi.org/10.1145/3583780.3615497.

[27] Al Hasan Rony, M.R., Kovriguina, L., Chaudhuri, D., Usbeck, R., Lehmann, J.: RoMe: A Robust Metric for Evaluating Natural Language Generation. Proc. Annu. Meet. Assoc. Comput. Linguist. 1, 5645–5657 (2022). https://doi.org/10.18653/v1/2022.acl-long.387.

Medium, https://medium.com/, last accessed 2025/10/29.