SHEFAA: A Fine-Tuned Answering System for Arabic Medical Questions

Yasser Alharbi, Eyad Alatifi, Nawaf Alandijany, Abdulaziz AbuTaleb, Abdullah Alharbi
Department of Computer Science and Artificial Intelligence,
Umm Al-Qura University, Saudi Arabia
yasser.luq@outlook.com eyad.alatifi@gmail.com nwandijany@gmail.com
abdulaziz.h.abutaleb@hotmail.com abdullahomalharbi@gmail.com

Abstract

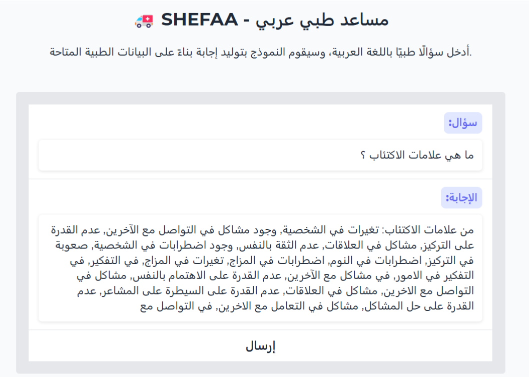
SHEFAA is an Arabic medical question-answering system fine-tuned on the **AHQAD dataset**. Our approach leverages **Jais-family-256m** and **QLoRA** to efficiently adapt a large language model using limited computational resources. We apply extensive data cleaning and normalization to address Arabic's morphological complexity and spelling inconsistencies. The system is evaluated on heldout test data using BLEU, ROUGE, and BERTScore, demonstrating competitive performance in generating coherent, medically focused answers. By tackling an underrepresented language and specialized domain, SHEFAA highlights the potential of parameter-efficient fine-tuning strategies for broader, low-resource, and domain-specific NLP applications.

Keywords: Arabic, NLP, Medical, Fine-Tuning, QLoRA, Jais

1 Introduction

Arabic medical question answering (Q&A) is both a linguistically and contextually challenging task in Natural Language Processing (NLP). Unlike English-based Q&A systems, Arabic Q&A must handle morphological variations, diacritic ambiguity, and different dialects. The proliferation of medical information online has heightened the need for automated systems that can provide accurate, concise, and context-aware answers to Arabic speakers seeking reliable healthcare advice. In this project, we present **SHEFAA**, an Arabic healthcare Q&A system that fine-tunes a transformer-based large language model on the AHQAD dataset (a comprehensive collection of Arabic medical questions and answers sourced from Kaggle).

By leveraging **Jais-family-256m** as the foundational model and applying **QLoRA** for parameter-efficient fine-tuning, we address common challenges in Arabic text processing—such as letter normalization and diacritic handling—while ensuring our model remains both memory-efficient and performant. We chose this task for our term project because it not only meets a growing demand for reliable Arabic medical content but also provides a valuable opportunity to apply and refine NLP techniques in a real-world setting.



SHEFAA - مساعد طبي عربي

أدخل سؤالًا طبيًا باللغة العربية، وسيقوم النموذج بتوليد إجابة بناءً على البيانات الطبية المتاحة.

ما هي علامات الاكتئاب ؟

من علامات الاكتئاب: تغيرات في الشخصية, وجود مشاكل في التواصل مع الآخرين, عدم القدرة على التركيز, مشاكل في العلاقات, عدم الثقة بالنفس, وجود اضطرابات في الشخصية, صعوبة في التركيز, اضطرابات في النوم, اضطرابات في المزاج, تغيرات في المزاج, في التفكير, في التفكير في الامور, في مشاكل مع الآخرين, عدم القدرة على الاهتمام بالنفس, مشاكل في التواصل مع الاخرين, مشاكل في العلاقات, عدم القدرة على السيطرة على المشاعر, عدم القدرة على حل المشاكل, مشاكل في التعامل مع الاخرين, في التواصل مع

Figure 1: Illustrative example of SHEFAA's response to a medical question.

2 Literature Review

In the field of medical question answering (QA), prior research has examined how large language models can be adapted or fine-tuned to produce concise and accurate responses. We highlight three key papers that inform the design and methodology of our own system.

2.1 MedLM: Exploring Language Models for Medical Question Answering Systems

This paper compares general-purpose large language models (e.g., GPT-2, GPT-3.5) with domain-specific ones (e.g., T5) for closed-book medical QA tasks. The authors utilize two data sources—MedQuAD and Icliniq—to conduct fine-tuning experiments with distilled language models and prompt-based approaches. They introduce static prompting (the same set of Q&A examples) versus dynamic prompting (retrieving top-k similar examples) and find that dynamic retrieval yields more consistent gains in BLEU and ROUGE scores. Although GPT models often produce factual, expansive answers, smaller distilled variants are easier to control yet more prone to hallucinations.

2.2 Answer Generation for Retrieval-based Question Answering Systems

Shifting away from strict answer-sentence selection, this work proposes a generative approach called GenQA. Rather than selecting a single top sentence, the model integrates multiple high-ranking sentences from a RoBERTa-based ranker and uses a seq2seq architecture (T5 or BART) to generate a concise response. By training on short-answer datasets like MSNLG and testing on WikiQA and ASNQ, the authors demonstrate up to 32% improvements in human-judged correctness over traditional baselines.

2.3 Fine-Tuning LLMs for Reliable Medical Question-Answering Services

Focusing on fine-tuning large models such as LLaMA-2 and Mistral with domain reliability in mind, this paper introduces a technique employing low-rank adapters, rank stabilization, and noise injection for robust feature learning at higher ranks. It also proposes a retrieval-augmented method combining a "Question Rewrite" node and selective retrieval tokens to ensure only relevant external content is incorporated, yielding a 13% improvement in BLEU-4 and a 15% improvement in ROUGE-L scores compared to baseline methods for medical QA tasks.

Taken together, these three papers emphasize the importance of domain adaptation (whether through specialized data or rank-stabilized adapters) and effective retrieval/generation strategies. For Arabic medical QA, we similarly rely on domain-specific data cleaning and normalization, as well as parameter-efficient fine-tuning (QLoRA). By considering these insights, our project builds a system that targets both high-accuracy medical answers and efficient deployment in low-resource settings.

3 Data

Our experiments use the **AHQAD** Arabic Healthcare Q&A dataset[1]. Below, we describe our collection and processing pipeline.

3.1 Data Collection and Cleaning

We begin with the raw CSV file (`AHQAD.csv`) from the Kaggle repository [6], which initially has **808,472** of Q&A instances. We then apply the following steps:

- **Removing Unnecessary Columns:** The *Unnamed: 0* column is dropped.

Question	Category	Answer (Truncated)
ما هي علامات ضغط الدم المرتفع؟	الضغط والقلب	تشمل الأعراض الصداع الشديد...
كيف أتخلص من حب الشباب بسرعة؟	الأمراض الجلدية	يمكن استخدام الكريمات الموضعية...
ما طرق تسهيل الولادة الطبيعية؟	الحمل والولادة	تمارين التنفس والمشي تساعد في...

Table 1: Examples of Q&A pairs in our cleaned dataset. Each row shows a sample question, its medical category, and a truncated answer.

Split	Number of Q&A Pairs	Percentage
Train	572,149	80%
Validation	71,519	10%
Test	71,519	10%

Table 2: Approximate distribution of Q&A pairs after cleaning and stratified sampling **(full dataset, 90 categories)**.

- **Handling Missing Values:** Rows with null *Question* or *Answer* fields are removed.
- **Dropping Duplicates:** We remove any repeated questions or answers to ensure uniqueness.
- **Text Normalization:** We trim leading/trailing whitespace, remove punctuation and special characters, replace newline characters with spaces, and normalize Arabic letters (e.g., converting أ/إ to ا).

After cleaning, we are left with a high-quality set of **715,187** question-answer pairs in Modern Standard Arabic.

3.2 Data Splits and Sampling

Due to limitations in resources, we sample approximately 10% of the remaining data in a stratified manner, ensuring balanced representation across different categories (e.g., الحمل والولادة, أمراض القلب, الأمراض الجلدية, etc.). We then split this 10% subset into:

- **Training Set (80%)**
- **Validation Set (10%)**
- **Test Set (10%)**

For illustration, Table 1 shows sample rows from the cleaned data. Tables 2 and 3 summarize the split sizes for the full dataset and the 10% sample, respectively.

Each instance is {*Question*, *Category*, *Answer*}, with categories derived from the original AHQAD metadata. By sampling 10% and performing an 80-10-10 split, we preserve a representative range of healthcare topics while reducing computational overhead during fine-tuning.

[1] Kaggle Link

Split	Number of Q&A Pairs	Percentage
Train	57,212	80%
Validation	7151	10%
Test	7151	10%

Table 3: Approximate distribution using 10 percent of Q&A pairs after cleaning and stratified sampling **(85 categories)**.

3.3 Data Characteristics

We observe moderate class imbalance across categories (e.g., more questions about pregnancy than dermatology), but the stratified sampling helps mitigate extreme skew. Many answers are short paragraphs providing medical advice or factual information. The wide topical coverage (e.g., nutrition, chronic diseases, pediatric care) ensures a broad challenge for the language model.

4 Evaluation Metrics

We adopt three primary metrics to evaluate the quality and accuracy of generated answers in Arabic medical Q&A: **BLEU**, **ROUGE**, and **BERTScore**. These metrics have been employed in prior work on question answering and text generation.

4.1 BLEU Score

Definition. BLEU (Bilingual Evaluation Understudy)[7] is a precision-focused metric that calculates n-gram overlap between the generated text and the reference text. Intuitively, BLEU measures how many words and phrases in the generated text appear in the reference (to some maximum order n). The standard BLEU formula includes a brevity penalty (BP) and geometric average of n-gram precisions:

$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{1}$$

where p_n is the precision for n-grams, w_n is the weight (often $1/N$), and BP is the brevity penalty to avoid overly short outputs.

4.2 ROUGE

Definition. ROUGE (Recall-Oriented Understudy for Gisting Evaluation)[7] focuses on the overlap of n-grams, sequences, and/or word pairs between the generated text and the reference text. Unlike BLEU, ROUGE is more recall-oriented, measuring how much of the reference text is captured by the generated text.

Formula. The ROUGE-N score for an n-gram is calculated as:

$$\text{ROUGE-N} = \frac{\sum_{s \in \text{RefSumm}} \sum_{n\text{-gram} \in s} \text{Count}_{\text{match}}(n\text{-gram})}{\sum_{s \in \text{RefSumm}} \sum_{n\text{-gram} \in s} \text{Count}(n\text{-gram})} \tag{2}$$

where:

- $\text{Count}_{\text{match}}(n\text{-gram})$ is the number of overlapping n-grams between the candidate and reference summaries.
- $\text{Count}(n\text{-gram})$ is the total number of n-grams in the reference summary.

4.3 BERTScore

BERTScore[8] leverages deep contextualized embeddings to assess the semantic similarity between generated and reference texts. Instead of direct token matching, BERTScore computes a soft alignment between tokens in the reference and hypothesis using embedding similarity. For Arabic, one can leverage multilingual or Arabic-specific pretrained transformers (e.g., AraBERT).

The precision, recall, and F1-score for BERTScore are computed as follows:

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \tag{3}$$

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \tag{4}$$

$$F_{\text{BERT}} = 2\frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \tag{5}$$

Explanation: - **R_{BERT} (Recall)**: Measures how well each token in the reference text (x) is captured by the most similar token in the generated text ($\hat{x}$), using cosine similarity in the embedding space. - **P_{BERT} (Precision)**: Measures how well each token in the generated text ($\hat{x}$) aligns with the most similar token in the reference text (x). - **F_{BERT} (F1-score)**: The harmonic mean of precision and recall, providing a balanced measure of alignment between reference and hypothesis.

This approach allows BERTScore to capture **semantic similarity** beyond exact token matching, making it highly effective for evaluating text generation in Arabic using transformer-based models like **AraBERT**.

In summary, BLEU and ROUGE provide complementary perspectives on precision and recall at the token/phrase level, while BERTScore captures deeper semantic alignment. By reporting all three, we obtain a more comprehensive view of our model's performance on Arabic healthcare Q&A.

5 Simple Baseline: TF-IDF Retrieval

A straightforward way to characterize our data and establish a benchmark is through a TF-IDF retrieval baseline. Below, we summarize the key steps and results, without diving into the implementation code:

Baseline Approach. We vectorize all training questions using TF-IDF and, for each new question, retrieve the single most similar training question based on cosine similarity. The corresponding answer from the training set is then used as the "predicted" answer. Due to the large size of the validation set (71,519 entries), we evaluate on a sample of 1,000 questions for efficiency.

Results and Example. On average, the TF-IDF baseline achieves modest performance, with BLEU-1 around 0.118 and BLEU-4 as low as 0.036. ROUGE-L remains minimal at about 0.005 in F1. Meanwhile, BERTScore F1 hovers around 70%, indicating some lexical overlap with correct answers, but limited deeper semantic alignment. TF-IDF can capture relevant medical keywords, but in more complex or ambiguous queries, the baseline often yields irrelevant matches.

Test question:	**Predicted answer:**	**True answer:**
سؤال: السلام عليكم انا عمري ١٥ دائما تيجيني تشنجات في اصابع رجلي وفي اصابع يدي ومني عارف من ايش وهي ايش | التصنيف: الأمراض العصبية	الإجابة: عليک فحص مقدار الكلسيم فى الدمب	الإجابة: اجراء فحص الكلسيم

Table 4: Example of TF-IDF answer in the cleaned dataset.

Metric	Score (Avg)
BLEU-1	0.118
BLEU-2	0.081
BLEU-4	0.036
ROUGE-1 (F1)	0.005
ROUGE-2 (F1)	0.003
ROUGE-L (F1)	0.005

Table 5: BLEU and ROUGE scores for the simple baseline model.

Limitations. Because TF-IDF relies solely on surface-level term frequency and does not incorporate contextual embeddings, it may fail when questions do not share exact keywords with the training data. Additionally, it cannot generate novel or more nuanced answers, limiting its utility for complex, multi-faceted medical queries.

Metric	Score (Avg)
BERTScore Precision %	71.08
BERTScore Recall %	70.86
BERTScore F1 %	70.78

Table 6: BERTScore results for the simple baseline model.

Conclusion of Baseline. While simplistic, the TF-IDF retrieval baseline demonstrates how direct keyword matching can achieve partial success on a fraction of test queries. Its modest performance underscores the need for more advanced models (e.g., QLoRA fine-tuning of a large language model) that capture deeper linguistic and semantic relationships—particularly essential in medical Q&A tasks.

This analysis highlights the importance of incorporating contextual and semantic understanding, which simple baselines like TF-IDF retrieval lack.

6 Experimental Design

In this section, we describe our implementation of a published baseline, detail the extensions we explored for Arabic medical Q&A, and analyze system errors. All results are reported on a subset of 1,000 validation or test examples unless otherwise noted, following the evaluation metrics (BLEU, ROUGE, BERTScore). All code and experimental details are available in our GitHub repository .[2]

6.1 Motivation and Approach

To adapt a large language model for Arabic medical question-answering while maintaining computational efficiency, we employ a parameter-efficient fine-tuning strategy using **QLoRA**. This approach allows us to optimize the model while significantly reducing GPU memory requirements.

We fine-tune the **Jais-family-256m** model by applying **4-bit quantization** using `bitsandbytes`, which enables low-memory adaptation without compromising performance. Additionally, we incorporate **low-rank adaptation (LoRA)** to fine-tune only a subset of model parameters, maintaining the efficiency of the base model while improving domain-specific understanding. The training configuration remains as follows:

- **Learning Rate:** 1×10^{-4}
- **Batch Size:** 8

[2] https://github.com/yasser-alharbi/SHEFAA

- **Training Epochs:** 1

To train the model effectively with limited computational resources, we utilize **10% of the AHQAD** Arabic healthcare Q&A dataset. The dataset is further sampled to **1,000 queries** each for validation and testing, ensuring a representative yet manageable evaluation set. This setup allows for efficient fine-tuning while maintaining strong performance in medical question-answering.

Prompt Construction. Each training example comprises *Question*, *Category*, and *Answer*. We build prompts using the following template:

```
سؤال: {Question}\
التصنيف: {Category}\
الإجابة: {Answer}\
```

6.2 Error Analysis

To better understand model errors, we manually examined incorrect predictions. We categorized these errors into four main types:

1. **Hallucinated Responses (40%)**: The model sometimes generated plausible but factually incorrect medical advice.
2. **Incomplete Answers (30%)**: Some generated responses lacked key medical details present in the ground truth.
3. **Ambiguity (20%)**: Certain questions had multiple valid interpretations, leading to mismatched references.
4. **Spelling/Diacritics Issues (10%)**: Though rare, Arabic diacritic errors occasionally appeared.

> **Example Hallucination: Question**: ما هو أفضل علاج لمرض السكري؟
> **Gold Answer**: يعتمد على نوع السكري، يشمل الأنسولين أو أدوية تنظيم السكر.
> **Generated**: شرب الماء الدافئ مع الليمون يومياً يمكن أن يشفي من السكري.

6.3 Results

The generated answer presents misleading medical claims, highlighting the need for factual verification strategies.

Metric	Test, After Cleaning	Valid, After Cleaning
BLEU-1	0.037	0.033
BLEU-2	0.015	0.014
BLEU-4	0.006	0.006
ROUGE-1 (F1)	0.001	0.002
ROUGE-2 (F1)	0.000	0.001
ROUGE-L (F1)	0.001	0.002

Table 7: Performance comparison (BLEU, ROUGE) for the QLoRA baseline (10% data, 1 epoch) *after cleaning* on a 1,000-sample subset. The left column shows test-set results; the right column shows validation-set results.

Metric	Test, After Cleaning	Valid, After Cleaning
BERTScore P	61.40%	60.75%
BERTScore R	61.59%	60.66%
BERTScore F1	61.33%	60.57%

Table 8: Performance comparison (BERTScore) for the QLoRA baseline (10% data, 1 epoch) *after cleaning* on a 1,000-sample subset. The left column shows test-set results; the right column shows validation-set results.

6.4 Extension: Fine-Tuning on Uncleaned Data

In order to further assess the impact of our data preprocessing pipeline, we conducted an additional experiment in which the model is fine-tuned using the uncleaned version of the AHQAD dataset. This extension directly employs the raw data—without the normalization, duplicate removal, and punctuation cleaning steps described in Section 3—to determine how much the lack of preprocessing affects performance.

6.4.1 Experimental Setup

We adopted the same parameter-efficient fine-tuning framework (QLoRA on the Jais-family-256m model) as detailed in Section 6, with one key difference: the raw AHQAD data is used directly, without any cleaning or normalization. All other aspects (including prompt construction, training configuration and evaluation metrics) mirror those used in our primary experiments.

6.4.2 Results

Table 9 presents the performance metrics for the fine-tuning experiment on the uncleaned dataset, evaluated on a 1,000-sample subset of the validation set.

6.4.3 Analysis

Table 9 shows that fine-tuning on uncleaned data leads to a significant drop in performance. In particular, BLEU scores are very low (0.020 on test and 0.017 on validation), indicating that the generated responses share

Metric	Test (Before)	Valid (Before)
BLEU-1	0.020	0.017
BLEU-2	0.010	0.008
BLEU-4	0.004	0.003
ROUGE-1 (F1)	0.000	0.000
ROUGE-2 (F1)	0.000	0.000
ROUGE-L (F1)	0.000	0.000
BERTScore P	60.93%	60.42%
BERTScore R	65.99%	65.78%
BERTScore F1	63.18%	62.83%

Table 9: Performance metrics for fine-tuning on uncleaned (raw) data (10%).

few common n-grams with the reference answers. Similarly, ROUGE scores are zeros, suggesting minimal overlap in key phrases. The BERTScore results further confirm these findings, with F1 scores of 63.18% on test and 62.83% on validation. This drop in semantic similarity reveals that the model struggles to capture the underlying meaning when trained on noisy, uncleaned data. Overall, these results emphasize the importance of effective data cleaning in improving the quality and reliability of fine-tuning for Arabic medical question answering.

7 Final Results

Metric	Baseline (10%) Test, After Cleaning	Baseline (10%) Valid, After Cleaning	Baseline (10%) Test, Before Cleaning	Baseline (10%) Valid, Before Cleaning	TF-IDF Baseline (100%) Test
BLEU Scores					
BLEU-1	0.037	0.033	0.020	0.017	0.118
BLEU-2	0.015	0.014	0.010	0.008	0.081
BLEU-4	0.006	0.006	0.004	0.003	0.036
ROUGE Scores (F1)					
ROUGE-1 (F1)	0.001	0.002	0.000	0.000	0.005
ROUGE-2 (F1)	0.000	0.001	0.000	0.000	0.003
ROUGE-L (F1)	0.001	0.002	0.000	0.000	0.005
BERTScores (%)					
BERTScore P	61.40%	60.75%	60.93%	60.42%	71.08%
BERTScore R	61.59%	60.66%	65.99%	65.78%	70.86%
BERTScore F1	61.33%	60.57%	63.18%	62.83%	70.78%

Table 10: Performance comparison (BLEU, ROUGE, BERTScore) for the QLoRA baseline (10% data, 1 epoch) before/after cleaning vs. a TF-IDF baseline using 100% data. All metrics are computed on a 1,000-sample subset of either test or validation sets.

8 Limitations

While **SHEFAA** demonstrates promising results for Arabic medical question-answering, several limitations remain:

- **Domain Complexity:** Medical knowledge is vast, and our model is fine-tuned on a single dataset (AHQAD). Certain rare or highly specialized conditions may be insufficiently represented, limiting the model's coverage and potentially leading to incomplete advice.

- **Potential for Misinformation:** Despite efforts to refine domain-specific knowledge, the model can still produce inaccurate or misleading answers. This risk is heightened given the inherent complexity of healthcare advice, where incorrect guidance can have serious consequences.

- **Limited Training Data Usage:** Due to constrained computational resources and GPU memory, we restricted our experiments to only **10%** of the available AHQAD dataset and **1 epoch**. While parameter-efficient methods like QLoRA mitigate the memory requirements, this reduced sample size inevitably results in sparser coverage of many medical subtopics. Consequently, the model may underperform or hallucinate when encountering less frequent healthcare domains during inference.

- **Ethical and Privacy Concerns:** Medical Q&A inherently involves sensitive personal information. Our model does not currently incorporate mechanisms for data anonymization or compliance with regulatory frameworks, highlighting the need for careful deployment in real-world medical settings.

9 Conclusions

In this paper, we introduced SHEFAA, an Arabic healthcare question-answering system that leverages a transformer-based large language model fine-tuned via QLoRA on the AHQAD dataset. Our work addressed key challenges in processing Arabic text—such as letter normalization, diacritic handling, and spelling inconsistencies—by incorporating a robust data cleaning and preprocessing pipeline. Through a series of experiments, we demonstrated that effective preprocessing is critical: fine-tuning on cleaned data yielded significantly better performance across evaluation metrics (BLEU, ROUGE, and BERTScore) compared to training on uncleaned, raw data.

Additionally, our system outperformed a simple TF-IDF retrieval baseline, particularly in capturing semantic relationships and generating more contextually relevant responses. Despite these improvements, our error analysis revealed ongoing challenges, including hallucinated responses, incomplete answers, ambiguity in multi-interpretable questions, and occasional diacritic errors.

Overall, the findings underscore the importance of domain-specific data preparation and parameter-efficient fine-tuning techniques for developing reliable and memory-efficient medical Q&A systems in low-resource languages like Arabic. Future work will focus on enhancing factual verification mechanisms and addressing the identified error types to further improve the quality and reliability of the generated medical advice.

References

[1] Niraj Yagnik et al. MedLM: Exploring Language Models for Medical Question Answering Systems. *2024*.

[2] Chao-Chun Hsu et al. Answer Generation for Retrieval-based Question Answering Systems. In *ACL-IJCNLP*, 2021.

[3] Ali Anaissi et al. Fine-Tuning LLMs for Reliable Medical Question-Answering Services. *2024*.

[4] Alasmari, A., Alhumoud, S., & Alshammari, W. (2024). AraMed: Arabic Medical Question Answering using Pretrained Transformer Language Models. *OSACT 2024 Workshop*, 50–56. © 2024 ELRA Language Resource Association: CC BY-NC 4.0.

[5] Moussa, A., Khattab, H., and Khattab, A. (2022). *ArMed: An Arabic Medical Reading Comprehension Dataset*. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, Gyeongju, Republic of Korea, pages 1235–1247.

[6] AHQAD Kaggle. Arabic Healthcare Q&A Dataset (AHQAD). Accessed 2024.

[7] GeeksforGeeks, "Understanding BLEU and ROUGE Score for NLP Evaluation," Available at: GeeksforGeeks, Accessed: February 2, 2025.

[8] A. Sojasingarayar, "BERTScore Explained in 5 Minutes," Medium, January 15, 2024. Accessed: February 2, 2025.

[9] Tran, M.-N., Nguyen, P.-V., Nguyen, L., & Dinh, D. (2024). ViMedAQA: A Vietnamese Medical Abstractive Question-Answering Dataset and Findings of Large Language Model. *ACL 2024 Student Research Workshop*, 270–278.

[10] Hugging Face. (2023). *Transformers – Question Answering*. `https://huggingface.co/docs/transformers/tasks/question_answering`. Accessed on June 11, 2023.

[11] Ben Abacha, A., & Demner-Fushman, D. (2019). *A Question-Entailment Approach to Question Answering*. BMC Bioinformatics, 20(1), 511:1–511:23.

[12] Guo, Q., Cao, S., & Yi, Z. (2022). *A Medical Question Answering System Using Large Language Models and Knowledge Graphs*. International Journal of Intelligent Systems, 37(11), 8548–8564.

[13] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). *BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining*. Bioinformatics, 36(4), 1234–1240.

[14] Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). *BLEU: A Method for Automatic Evaluation of Machine Translation*. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

[15] Lin, C.-Y. (2004). *ROUGE: A Package for Automatic Evaluation of Summaries*. In *Text Summarization Branches Out*, pages 74–81.

[16] Liévin, V., Hother, C. E., & Winther, O. (2022). *Can Large Language Models Reason About Medical Questions?* `https://arxiv.org/abs/2207.08143`.

[17] Sallam, M. (2023). *The Utility of ChatGPT as an Example of Large Language Models in Healthcare Education, Research and Practice: Systematic Review on the Future Perspectives and Potential Limitations*. medRxiv, 2023-02.

[18] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ...Natarajan, V. (2022). *Large Language Models Encode Clinical Knowledge*. `https://arxiv.org/abs/2212.13138`.

[19] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. `https://arxiv.org/abs/2106.09685`.

[20] Han, Z., Gao, C., Liu, J., Zhang, J., & Zhang, S. Q. (2024). *Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey*. `https://arxiv.org/abs/2403.14608`.

[21] Antoun, W., Baly, R., and Hajj, H. (2020). *AraBERT: Transformer-based Model for Arabic Language Understanding*. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT)*, co-located with LREC, Marseille, France, pages 9–15.

[22] Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). *QLoRA: Efficient Finetuning of Quantized LLMs*. `arXiv:2305.14314`.

[23] Khalifa, S., Alali, A., and Shaalan, K. (2021). *AraELECTRA: Pre-Training Transformers for Arabic Language Understanding Using ELECTRA*. In *Proceedings of the 1st Workshop on Arabic Language Technologies and Applications (ArLTA)*, pages 10–19.

[24] Pasha, A., Al-Badrashiny, M., Al-Rahee, S. M., Diab, M., El Kholy, A., Habash, N., Pooleery, M., Rambow, O., and Roth, R. (2014). *MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic*. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland, pages 1094–1101.