Supervised Fine-Tuning Approach for Medical Question-Answering using Qwen Instruct

AMJAD JUMAAH FRHAN[1,2], MOHAMMED A.S. AL-HITAWI [3 *], HIBA A, ABU-ALSAAD[4 *]
[1]University of AL Mashreq, Department of Cybersecurity Engineering Technology,
10015, Baghdad,
IRAQ

[2]Department of Education and Islamic Studies, 10011, Sunni Endowment Diwan
[3]Artificial Intelligence Department, College of Information Technology, University of Fallujah,
31001, Fallujah,
IRAQ

[4] Director of Computer Engineering Dept. Mustansiriyah University, 10011,
Baghdad,
IRAQ

*Corresponding author
https://orcid.org/0009-0009-7905-0978

Abstract: Large Language Model (LLM) became promising tool for supporting medical applications based on natural language processing. Except that models trained for general purposes suffer from the limitation of accuracy and reliability when it deals with sensitive medical question answering (MQA). The model is evaluated using Exact Match and F1 score. The goal of this study is to develop LLM to answer the MQA tasks. Through applying SFT with Parmerter Efficient Fine Tuning (PEFT) specifical LoRA were applied. The results showed significant qualitative improvement and Notable accuracy in the answers. And organizing medical content, by reducing cognitive mistakes capered to the base model. The experiments refer to guided personalization for LLM represent essential step towards building medical systems.

Key-Words: Parameter Efficient Fine Tuning (PEFT), Supervised Fine Tuning (SFT), Medical QA, Qwen2.5, LoRA, Generative Models, Natural Language Processing.

Received: June 29, 2025. Revised: October 23, 2025. Accepted: November 21, 2025. Published: April 15, 2026.

1 Introduction

Large language Models (LLM) perceptions fast development last years. It is widely used by language understanding, such as translation, summarization, code generation, questions answering. In medical fields the need arises to smart systems which are capable of providing accurate information and safe for users, whether they are patients or healthcare workers. However, the use for general model in medical area faces great challenges. The circumscribed specialization is probably the most notable of each of them, and there is a probability to produce inaccurate information or misleading. From here it appears the needs for customizing these models for specific medical tasks using Supervised Finetuning (SFT) approach. Which allows the models for behavior guidance towards narrative knowledge field and highly sensitive such as medical questions answering. All experiment available on: https://github.com/Mohammed20201991/MQA-SFT

The goal of this study is to present a framework to downstream Qwen2.5-7B-Instruct model [1] to answer medical question in English language, with documentation for the main pipeline starting from data collection, preprocessing, training, evaluation, testing and deployment. The source code is publicly available on MQA-SFT [2]. Additional to answer the question whether fine-tuning with LoRA can

improve medical QA over baseline models? The fig.1 shows some visualized tokens.

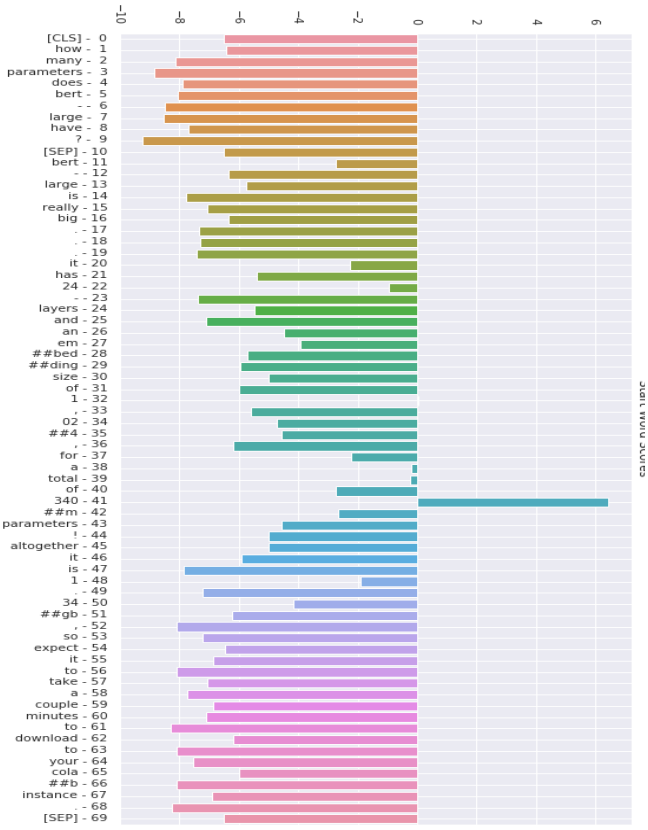

Fig.1 show score for some visualized tokens.

This investigation checks if it improves accuracy with limited data as LoRA is widely used, but its impact on medical QA is still unclear.

The main contributions of this study are:

1- Medical Question Answering App based LoRA.
2- Evaluate the model performance using real data.
3- Comparing fine-tuning vs tradition approach.
4- Enhanced the model outcome quality in terms of accuracy and structure.

2 Related Work

Previous studies focused on using LLM in medication such as BioBERT [3], ClinicalBERT [4], which depends on pre-training weights sharing trained on medical data. As addressed in modern research applying technologies such SFT and Instruction tuning to enhance the performance of the language models in specific task.

This study suggests a model that combines both natural language processing and knowledge representation to enhance the accuracy and reliability for medical question answering in the neuroscience field. with the simulation for the used pattern by the specialists (doctors in this field) to reduce errors. and showed that it outperforming the traditional approaches [19].

This study introduced the BERT model, which is dependent on the bidirectional representation for the text orientation through the pre-training on unlabeled data. Which allows deep language context understanding. Those models can be adapted easily for many tasks in a downstream approach, unlike auto-regressive ones like GPT, which follow one direction. Bert showed very accurate performance on text classification, next word and sentence prediction and achieved surprisingly advanced performance (consider state-of-the-art in the last few years), specifically of natural language understanding and prediction [22][21].

3 Methodology

3.1 Task Definition

The mission is to train language model to answer medical questions including Symptoms, Diagnostic methods, Tests, General treatments, Preventive health. With confirmation that the models don't introduce a definitive diagnosis and does not replace medical consultation. In Fig.2. LoRA adaptations are placed into transformer attention layers for cost-effective fine-tuning, although the Qwen2.5-7B-Instruct network stays frozen. When being implemented as a secure, non-diagnostic medical QA assistance, the final clinical-specialized model is examined using exact match, F1 score, and objective analysis.

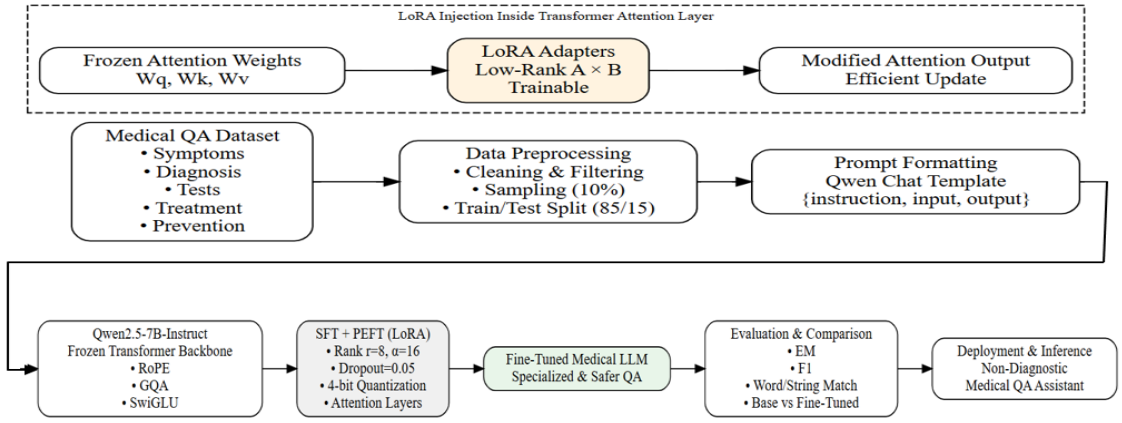

Fig.2. The main proposed methodology from training to deployment.

3.2 Dataset

We collected this data [6] due it is scientific accuracy, clarity of wording, integrity of medical content, relevance of data to supervised learning. And it is structure suitable for SFT format that has features and labels. High-quality data formatted as input-output pairs suitable for SFT as shown in Fig.1 it includes

{"instruction": "...", "input": "...", "output": "..."}

The dataset was loaded and filtered to keep question and answer columns. After sampling 10%, it was converted into a Hugging Face dataset and split (85% training, 15% testing). Text formatting followed the Qwen chat structure. Fig.3 shows sample for medical data we use [15]. Intended only as a non-diagnostic assistant and with no sensitive patient information

it uses public data.

	instruction	input	output
0	Answer the medical question in a safe, general...	What symptoms might indicate dehydration?	Feeling very thirsty, dry mouth, dark urine, d...
1	Answer the medical question in a safe, general...	What are typical symptoms of seasonal flu?	Fever, body aches, fatigue, cough, and sore th...
2	Answer the medical question in a safe, general...	Context: The person has had mild symptoms for ...	Common symptoms include a runny or stuffy nose...
3	Answer the medical question in a safe, general...	What is the role of insulin (general)?	Insulin is a hormone that helps cells take in ...
4	Answer the medical question in a safe, general...	Note: What does 'chronic' mean in health conte...	Chronic describes a condition that lasts a lon...
5	Answer the medical question in a safe, general...	Please explain: What is the function of the he...	The heart pumps blood throughout the body, del...
6	Answer the medical question in a safe, general...	Please explain: What are symptoms of anemia in...	Tiredness, paleness, shortness of breath with ...
7	Answer the medical question in a safe, general...	Note: What are common symptoms of a cold?	Common symptoms include a runny or stuffy nose...

Fig.3. Show sample of the used data.

In addition to the figure above we could see in Table 1 some basics statistical information to understand the data distribution.

Table 1. (samples [500]), Question Length (q_len), and Answer Length (a_len).

Statistical	q_len	a_len
Count	500	500
Mean	59.38	122.07
Std Dev	29.61	43.68
Min	21	71
25%	41	86
50%	53	114
75%	65	150
Max	171	238

LLMs are aware about the sequence length in the next Fig.4, shows the questions length in addition to answers length.

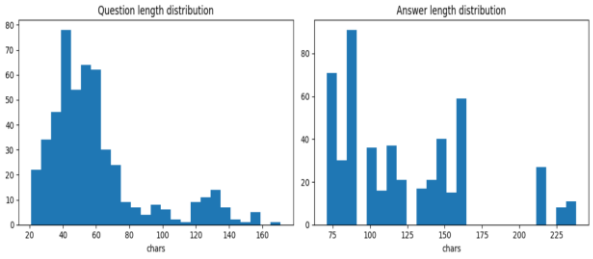

Fig.4. Question length vs. Answer Length.

How is the prompt look like?

<|im_start|> user: What causes Glaucoma? <|im_end|>

<|im_start|> assistant: <think> </think>

"Nearly 2.7 million people have glaucoma, a leading cause of blindness in the United States. Although anyone can get glaucoma, some people are at higher risk …. "

<|im_end|>

3.3 Base Model Selection

We choose Qwen2.5-7B-Instruct model due of what it possesses. Balance between volume and performance, huge ability to understand instructions, compatible with PEFT techniques and can be trained on middle resources (RAM and GPU) availability [10].

3.4 Model Architecture

Qwen model is a dense, decoder-only transformer language model with approximately 3 billion parameters, implemented for efficient general-purpose NLP tasks, see Table 2. It utilizes the advantages of a 36-layer modern transformer framework, Rotatable Positional Embeddings (RoPE) for extended context managing (up to ~32K tokens), Grouped-Query Attention (GQA) to decrease KV-cache memory, SwiGLU triggering, and RMSNorm normalization. The framework is capable of used for estimation, fine-tuning, and distribution on low GPU resources in particular when quantized, and can be made available with open weights with the Qwen Research License. Frequent ecosystems like Hugging Face Transformers accommodate it [7].

Table 2. demonstrate the main hyper-parameters for Queen architected [17].

Aspect	Details
Model size	~3.1B parameters
Architecture	Decoder-only dense Transformer
Layers	36
Attention	Grouped-Query Attention (48 heads, 8 KV heads)
Context length	Up to ~32K tokens
Activation / Norm	SwiGLU / RMSNorm
Resources	*~6–8 GB VRAM (FP16), ~2–4 GB (quantized)*
License	Qwen Research License

3.5 Fine-Tuning Setup

We Setup Environment by using a framework like Hugging Face transformers and PEFT (Parameter-Efficient Fine-Tuning) libraries. LoRA (Low-Rank Adaptation) it is highly recommended for efficient training and fine-tuning on the mentioned data with the hyper-parameters we choose as shown in Table 3.

Table 3. The hyper-parameters used during this study.

Hyper-parameter	Value
Learning Rate	2e-4
Batch Size	4
Epochs	10
Optimizer	AdamW
Max Sequence Length	256
Max Steps	400
dropout	0.05
alpha	16
Rank (r)	8
quantization	4-bit
target modules	attention layers

4 Results and Evaluation

The results showed good enhancement for accurate medical terms, the answers become more organized and sequenced, Reduce hallucinations, safer and transparency in language for clients.

4.1 Evaluation method

Based on the nature of medical task, we denote qualitative evaluation through some evaluation metrics as listed below and through comparing generated answer from fine-tuned vs base model see eq. (1) and eq. (2) as auxiliary analysis metrics.

$$Word\ match\ score = \frac{number\ of\ matching\ words}{number\ of\ groundtruth\ answer\ words} \quad (1)$$

$$String\ match\ score = \frac{length\ of\ matching\ string}{length\ of\ groundtruth\ answer\ string} \quad (2)$$

Exact Match (EM) measures whether the predicted answer exactly matches the ground-truth answer after normalization (e.g., lowercasing, removing punctuation, articles, extra spaces) as shown in eq. (3).

$$EM = \begin{cases} 1, & if\ \hat{A} = A \\ 0, & otherwise \end{cases} \quad (3)$$

Where:

- $\hat{A}$ = predicted answer
- A = ground-truth answer

For a dataset with N questions, we use summation in eq. (4):

$$EM_{dataset} = \sum_{i=1}^{N} EM_i \quad (4)$$

F1 Score (Token-Level):

$$Precesion = \frac{Number\ of\ common\ tokens}{Number\ of\ tokens\ in\ \hat{A}} \quad (5)$$

$$Recall = \frac{Number\ of\ common\ tokens}{Number\ of\ tokens\ in\ A} \quad (6)$$

If there are no overlapping tokens, then: F1 = 0

$$F1 = \frac{2 * Precesion * Recall}{Precesion + Recall} \quad (6)$$

Quantitative Evaluation

Table 4. Quantitative evaluation presents the comparison on the test set results between the base and the fine-tuned model.

Model	Exact Match (EM)	F1 Score
Qwen2.5-7B (Base)	0.42	0.58
Fine-Tuned (LoRA)	0.61	0.74

for domain adaptation demonstrating the effectiveness of LoRA-based supervised fine-tuning, The results show a good enhancement in both given metrics.

4.2 Comparative examples

After training we do qualitative evaluation by loading the fine-tuned model and generate 5 example outputs that clearly demonstrate the specialization achieved by the SFT process. And then compared these to the same 5 prompts run on the base model. Table 5 shows the significant enhancement using the proposed approach. with fewer hallucinations, the fine-tuned model performed better than the base model.

Table 5. Base Model vs Fine-Tuned Model Comparison

Prompt	Base Model Output (Summary)	Fine-Tuned Model Output (Summary)
1. What is the difference between glaucoma and cataracts?	Gives a general eye-disease explanation; does not clearly contrast both conditions.	Provides a structured comparison: glaucoma = optic nerve damage due to pressure; cataracts = clouding of lens; symptoms, causes, and diagnostics clearly separated.
2. How is acute appendicitis diagnosed?	Basic definition; misses physical exam signs and imaging steps.	Includes clinical symptoms, RLQ tenderness, lab findings, ultrasound/CT imaging, and when surgery is indicated.
3. What are the complications of uncontrolled diabetes?	Mentions only a few complications like neuropathy.	Lists all major complications: retinopathy, nephropathy, neuropathy, cardiovascular risks, infections, DKA, poor wound healing.
4. Explain the treatment protocol for bacterial pneumonia.	Gives vague treatment suggestions; lacks structured protocol.	Provides step-by-step management: empiric antibiotics, severity scoring (CURB-65), supportive care, hospitalization criteria, and follow-up.
5. What symptoms indicate a possible heart attack?	Gives a short, generic description.	Lists medically accepted MI symptoms: chest pressure, radiating pain, sweating, dyspnea, nausea, emergency steps (call emergency services).

5 Conclusion

SFT plus LoRA significantly boosted knowledge of the medical questions-answering field. When compared to the original model, the amended model generated more reliable, structured, and scientifically precise responses. Intensity and scope being limited to database content have some limitations. As possible, the performance in the future could be enhanced by combining different data type and update the weights for knowable database continuously. In addition to test the model performance on real data to ensure it capability for generalization in different environments.

This research has several drawbacks: due to limited resources it used a smaller dataset, which may affect generalization. The expert medical review lacked evaluation in this study.

6 Future Work

Extend the dataset to contains more accurate desaies type, make qualitve evaluation with medical workers participation who are specialized in, integrating automated medical verification mechanisms, and compare the performance with others medical models. Choose of suitable open-source base LLM (e.g., a smaller model from the Llama, Mistral, or Gemma families, or a model suitable for consumer GPUs) that can be fine-tuned. Using different machine learning approach such unsupervised, Sem-supervised, reinforcement learning and so on [18].

Bert score metrics can be utilized as trusted new introduced metric [22] as showed in eq. (7,8) and (9).

$$P = \left(\frac{1}{|x|}\right) * \sum_{\{x_i \in x\} \backslash max_{\{y_j \in y\}\cos}} cos(x_i, y_j) \quad 7$$

$$R = \left(\frac{1}{|y|}\right) * \sum_{\{y_j \in y\} \backslash max_{\{x_i \in x\}} \cos} cos(y_j, x_i) \quad 8$$

$$F1 = 2 * \frac{(P * R)}{(P + R)} \quad 9$$

Two stages approach could be utilized the increase model performance by generating more data using generative models and fine-tune with human (real data) [23].

Acknowledgement:
The authors would like to thank the university of Fallujah Mustansiriyah University, and the University of AL Mashreq for them support.

References:

[1] Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., .& Qiu, Z. (2025). Qwen3 technical report. arXiv preprint arXiv:2505.09388.

[2] Muzhe Guo, Muhao Guo, Edward T. Dougherty, and Fang Jin. 2023. MSQ-BioBERT: Ambiguity Resolution to Enhance BioBERT Medical Question-Answering. In Proceedings of the ACM Web Conference 2023 (WWW '23). Association for Computing Machinery, New York, NY, USA, 4020–4028. https://doi.org/10.1145/3543507.3583878

[3] Huang, K., Altosaar, J., & Ranganath, R. (2019). Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342.

[4] N. A. Mohammed et al., "Recognizing Phishing in Emails by Using Natural Language Processing & Machine Learning Techniques," 3rd International Conference on Cyber Resilience (ICCR), Dubai, UAE, 2025, pp. 1–7, doi: 10.1109/ICCR67387.2025.11292212.

[5] Ben Abacha, A., Demner-Fushman, D. A question-entailment approach to question answering. BMC Bioinformatics 20, 511 (2019). https://doi.org/10.1186/s12859-019-3119-4.

[6] Qwen Team. (2024). Qwen2.5: A Party of Foundation Models. Retrieved from https://qwenlm.github.io/blog/qwen2.5/.

[7] Qwen Team, "Qwen2.5-7B-Instruct Model Card," Alibaba Cloud, 2024. [Online]. Available: https://huggingface.co/Qwen/Qwen2.5-7B-Instruct.

[8] Y. Bai et al., "Qwen Language Models: Enhancing Multilingual, Multi-Domain Large Language Models," Alibaba Cloud Research, 2023.

[9] E. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," arXiv preprint arXiv:2106.09685, 2021.

[10] H. Liu et al., "PEFT: Parameter-Efficient Fine-Tuning Library," Hugging Face, 2023. [Online]. Available: https://github.com/huggingface/peft.

[11] T. Dettmers et al., "BitsAndBytes: 8-bit and 4-bit Quantization for Large Language Models," Hugging Face, 2022. [Online]. Available: https://github.com/TimDettmers/bitsandbytes.

[12] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," Proceedings of EMNLP, 2020.

[13] P. Rajpurkar et al., "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in Proc. EMNLP, 2016.

[14] A. Ben Abacha and D. Demner-Fushman, "MedQuAD: Medical Question Answering Dataset," National Library of Medicine, 2019.

[15] National Institutes of Health, "NIHSeniorHealth Medical Encyclopedia Articles," U.S. Department of Health and Human Services.

[16] Hugging Face, "Accelerate: Distributed and Efficient Training," 2022. [Online]. Available: https://github.com/huggingface/accelerate.

[17] ApXML. (n.d.). Qwen2.5-3B: Specifications and GPU VRAM Requirements. Retrieved from https://apxml.com/models/qwen2-5-3b.

[18] A. J. Frhan, "Hybrid intelligence learning and signature-based framework for zero-day malware intrusion detection," International Journal of Computers, vol. 2025, no. 10, pp. 284–293,2025, https://www.iaras.org/iaras/filedownloads/ijc/2025/006-0030(2025).pdf.

[19] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

[20] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

[21] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

[22] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

[23] Al-Hitawi, M. A., & Gyöngyössy, N. M. (2026). Enhancing Transformer-Based Language Models for Hungarian Handwritten Text Recognition. *F1000Research*, *15*, 181. https://f1000research.com/articles/15-18.