!MSA at AraHealthQA 2025 Shared Task: Enhancing LLM Performance for Arabic Clinical Question Answering through Prompt Engineering and Ensemble Learning*

Mohamed Younes, Seif Ahmed, Mohamed Basem
Faculty of Computer Science, MSA University, Egypt
{mohamed.tarek61, seifeldein.ahmed, mohamed.basem1}@msa.edu.eg

Abstract

We present our systems for Track 2 (General Arabic Health QA, MedArabiQ) of the AraHealthQA-2025 shared task, where our methodology secured 2nd place in both Sub-Task 1 (multiple-choice question answering) and Sub-Task 2 (open-ended question answering) in Arabic clinical contexts. For Sub-Task 1, we leverage the Gemini 2.5 Flash model with few-shot prompting, dataset preprocessing, and an ensemble of three prompt configurations to improve classification accuracy on standard, biased, and fill-in-the-blank questions. For Sub-Task 2, we employ a unified prompt with the same model, incorporating role-playing as an Arabic medical expert, few-shot examples, and post-processing to generate concise responses across fill-in-the-blank, patient-doctor Q&A, GEC, and paraphrased variants.

1 Introduction

The MedArabiQ benchmark (Abu Daoud et al., 2025), part of the AraHealthQA-2025 shared task (Alhuzali et al., 2025), evaluates large language models (LLMs) on Arabic medical question answering, addressing the critical need for reliable AI-driven clinical tools in Arabic-speaking regions where digital healthcare resources are scarce. Track 2, General Arabic Health QA (MedArabiQ), tests models on general medical knowledge, from foundational topics like physiology to advanced areas like neurosurgery, across two subtasks. Sub-Task 1 (classification) involves selecting correct answers from predefined options for 300 development samples, split into standard multiple-choice questions, bias-injected questions (e.g., confirmation, cultural, or recency bias), and fill-in-the-blank with choices, evaluated by accuracy on a 100-question test set. Sub-Task 2 (generation) requires free-text responses for 400 development samples, covering fill-in-the-blank without choices, patient-doctor Q&A from the AraMed corpus (Alasmari et al., 2024), grammatically corrected Q&A, and LLM-paraphrased questions, assessed via BLEU, ROUGE, and BERTScore on a 100-question test set.

Arabic medical question answering poses unique challenges for current LLMs due to limited training data in Modern Standard Arabic (MSA) and dialectal variations, which often lead to poor generalization on clinical tasks. Additionally, culturally sensitive or biased questions require nuanced reasoning, while diverse question formats (e.g., fill-in-the-blank, open-ended consultations) demand robust adaptation to varying linguistic and contextual demands. Existing models often struggle with these complexities, as they are predominantly trained on English-centric or general-domain data, lacking domain-specific Arabic medical knowledge.

Our approach innovatively combines targeted prompt engineering and ensemble techniques with the Gemini 2.5 Flash model. We develop a unified methodology that addresses both classification and generation tasks in Arabic medical QA without requiring task-specific fine-tuning, leveraging carefully designed prompts and ensemble strategies to handle the complexities of Arabic medical language and diverse question formats.

2 Background

Track 2 (General Arabic Health QA, MedArabiQ) of the AraHealthQA-2025 shared task (Abu Daoud et al., 2025) evaluates large language models on Arabic medical question answering, addressing the need for reliable AI-driven clinical tools in Arabic-speaking regions. The task spans 12 medical domains: Biochemistry, Histology, Embryology, Microbiology, Neurosurgery, OBGYN, Oncology, Ophthalmology, Pediatrics, Pharmacol-

*https://github.com/AraHealthQA_2025

Proceedings of The Third Arabic Natural Language Processing Conference, pages 176–183

ogy, Physiology, and Pulmonology. We participated in both subtasks of Track 2, leveraging prompt engineering and ensemble techniques to achieve robust performance.

2.1 Task Details

Sub-Task 1 (classification) involves selecting the correct option from multiple-choice questions (MCQs) in Modern Standard Arabic (MSA). The dataset includes 300 development samples (100 each for standard MCQs, biased MCQs with biases like recency or status quo, and fill-in-the-blank with choices) and 100 test samples. Input is an MSA question with 4–5 options, and output is the correct option's text. Representative examples are summarized in Table A.1.

Sub-Task 2 (generation) requires free-text responses to prompts in MSA or dialectal Arabic, with 400 development samples (100 each for fill-in-the-blank without choices, patient-doctor Q&A, grammatical error correction (GEC), and LLM-modified Q&A) and 100 test samples, sourced from Arabic medical school exams, notes, and the AraMed corpus (Alasmari et al., 2024). Representative examples are summarized in Table A.2 .

2.2 Related Work

Arabic NLP faces challenges due to limited resources and dialectal variations (Abdul-Mageed et al., 2021). Prior work on Arabic medical QA (Alasmari et al., 2024) provides datasets like AraMed but lacks focus on handling biases or diverse question types. Prompt engineering techniques, such as Chain-of-Thought (CoT) prompting (Wei et al., 2022), improve reasoning in English-centric tasks but are underexplored in Arabic medical contexts. Recent work has explored prompt engineering for Arabic NLP tasks, such as stance detection, demonstrating the effectiveness of tailored prompts for LLMs in handling Arabic text (Al Hariri and Abu Farha, 2024). Similarly, few-shot learning with transformer models (Devlin et al., 2019) has advanced general NLP, but its application to Arabic clinical scenarios remains limited.

Medical question answering often relies on retrieval-augmented approaches (Lewis et al., 2020), which integrate external knowledge bases for open-domain tasks. However, such methods are less effective for Arabic medical QA due to the scarcity of structured medical knowledge in Arabic and the complexity of handling biases like recency or status quo. Our unified prompt for Sub-Task 2, addressing diverse question types without fine-tuning, and ensemble voting for Sub-Task 1, tackling biases, offer novel solutions tailored to the resource-scarce and culturally nuanced Arabic medical domain.

3 System Overview

We describe the methods we used for each subtask, the design choices that made them work well in Arabic medical settings, and how to reproduce them step-by-step.

3.1 Sub-Task 1: Classification (MCQ)

Model and settings. All systems use the same model (Gemini 2.5 Flash) for consistent outputs.

Systems (different approaches).

- Arabic Few-Shot (AFS): Arabic instruction prompt + 6 examples from different medical areas; output limited to a single Arabic letter from {أ، ب، ج، د، هـ}.
- English Translation + Answer (ETA): translate the Arabic question to English using a specific translation prompt, then answer with the same letter format.
- Refinement + Answer (RFA): rewrite the Arabic question for clarity (adds 15–25 word explanations for each option without changing meaning), then answer with the same letter format. Examples of the data refinement process are shown in Table A.3.
- Arabic Zero-Shot (AZS): Arabic instruction prompt without examples (baseline, not used in the final combination).

Ensemble (majority voting). We ensembled AFS, ETA, and RFA by simple vote counting over the answer choices $\mathcal{C}$={أ، ب، ج، د، هـ}. Given prediction functions f_i and input x:

$$\hat{y} = \arg\max_{c \in \mathcal{C}} \sum_{i=1}^{3} \mathbf{1}[f_i(x)=c]. \quad (1)$$

Ties are broken by a fixed priority RFA > AFS > ETA. This combination strategy provides reliable predictions across different question types. Ensemble methods have been shown to improve question answering performance by combining multiple classifiers, leading to more robust predictions (Chu-Carroll et al., 2003).

Output cleaning and standardization. We map any predicted character to the standard set {أ، ب، ج، د، هـ} (e.g., fix Arabic punctuation/spacing and Latin "A/B/C/D/E" if ever produced). We also remove extra tokens to ensure single-letter output format.

Challenges and solutions.

- Arabic variety and formatting: Examples cover multiple medical areas and different answer lengths; strict output rules and cleaning avoid problems.
- Prompt and dataset biases: Using three different approaches (native Arabic, English translation, refined Arabic) reduces single-prompt bias through voting.

3.2 Sub-Task 2: Generation

Model and settings. Same model. A single unified Arabic instruction + few-shot prompt handles: fill-in-the-blank (no choices), patient–doctor Q&A, grammar error correction (GEC), and LLM-rewritten Q&A.

Unified prompting and formatting. The prompt requires:

- Fill-in-the-blank: return only the missing word(s); if multiple blanks, separate answers with a comma and a space.
- Patient–doctor Q&A: brief, helpful advice; clearly recommend in-person care when needed.
- Avoid extra introductions or conclusions; keep Arabic medical terms unchanged.

This setup provides consistent performance across different generation tasks.

Output cleaning steps. For fill-in-the-blank tasks, we split answers by commas and clean up spacing. For consultations, we keep medical terms and maintain a proper clinical tone. All outputs go through Arabic text cleaning to handle different dialects. Additionally, we remove any markdown formatting (e.g., **bold**, *italic*, bullet points) that the model may produce to ensure clean, plain-text responses suitable for medical contexts, as well as not affecting the BERTScore.

Example selection. Examples cover multiple medical areas (drug studies, anatomy, clinical cases) and include both formal and dialect Arabic. Each example shows the desired output format and medical reasoning level.

Challenges and solutions.

- Different formats: One prompt with high-quality examples and clear output rules ensures consistency across types without fine-tuning.
- Arabic language complexity: Carefully chosen examples and consistent decoding reduce errors and inconsistencies.
- Safety/clinical tone: The prompt guides toward brief, careful advice and marks cases needing doctor follow-up.

Reproducibility notes. Use the exact prompt templates provided in Appendix B; keep the examples unchanged; do minimal, consistent output cleaning as specified above. All runs use Gemini 2.5 Flash with the decoding settings specified in Table A.4.

4 Experimental Setup

4.1 Data and Splits

We follow the official AraHealthQA-2025 Track 2 (MedArabiQ) setup and evaluated directly via the organizers' API on the official test sets (ST1: 100 items, ST2: 100 items). The provided development sets (ST1: 300 items; ST2: 400 items) were used only to guide prompt design, select few-shot examples, and perform sanity checks. No fine-tuning or external training data was used.

4.2 Preprocessing

We applied only input-side, minimal steps to ensure consistent prompts and data cleanliness:

- Standardize Arabic punctuation and whitespace in the input text while preserving medical terminology and numbers.
- Normalize option labels and bullet symbols in MCQ questions to a consistent form before prompting.

4.3 Post-processing

We applied lightweight output-side normalization for evaluation stability:

- MCQ: map any predicted symbol to the canonical set {أ، ب، ج، د، هـ} and strip extra tokens.

- Generation: remove markdown (bold/italic/bullets), standardize commas and spaces, and keep a concise clinical tone.

Removing markdown formatting from generated text is essential, as structured formatting can introduce noise that affects evaluation metrics like BERTScore by altering token representations (Tang et al., 2024).

4.4 Prompting Configurations

For Sub-Task 1, we use three complementary prompts: Arabic Few-Shot (AFS), English Translation + Answer (ETA), and Refinement + Answer (RFA). Predictions are combined via simple majority vote with a fixed tie-breaker (RFA $>$ AFS $>$ ETA). For Sub-Task 2, a single unified Arabic instruction with few-shot examples handles fill-in-the-blank, patient–doctor Q&A, GEC, and paraphrased inputs.

4.5 Evaluation Metrics

- Sub-Task 1 (MCQ): Accuracy
- Sub-Task 2 (Generation): BERTScore

5 Results

We present our official results from the AraHealthQA-2025 shared task evaluation , analyzing performance across both subtasks and examining the effectiveness of our ensemble approach.

5.1 Sub-Task 1: Classification Results

Our ensemble approach achieved 76% accuracy on the official test set, securing 2nd place in the classification task. Table 1 presents detailed performance breakdown for each individual approach and the final ensemble.

Individual system performance. The Refinement + Answer (RFA) approach performed best among individual systems at 74% accuracy, demonstrating the effectiveness of question clarification and option explanation in Arabic medical contexts. The Arabic Few-Shot (AFS) approach achieved 71% accuracy, showing strong baseline performance with domain-specific examples. The English Translation + Answer (ETA) approach scored 69% accuracy, indicating some information loss during translation despite maintaining medical terminology.

Ensemble effectiveness. The 3-system ensemble (RFA + AFS + ETA) improved performance by 2 percentage points over the best individual system, reaching 76% accuracy. This demonstrates successful bias reduction through diverse prompt strategies, with the RFA approach providing clarity, AFS maintaining Arabic medical context, and ETA offering cross-lingual reasoning perspectives.

5.2 Sub-Task 2: Generation Results

Our unified prompting approach achieved 86.953% BERTScore on the official test set, securing 2nd place in the generation task. The approach used a single Arabic instruction prompt with few-shot examples, casting the model as an Arabic medical expert to handle diverse question formats including fill-in-the-blank, patient-doctor consultations, grammatical error correction, and paraphrased questions. This unified strategy proved effective across all question types without requiring task-specific fine-tuning, demonstrating the power of well-designed prompting for Arabic medical contexts. Table 2 summarizes the performance.

5.3 Ablation Studies

Ensemble composition. Removing individual systems from the 3-way ensemble showed: RFA removal (-3% accuracy), AFS removal (-2% accuracy), ETA removal (-1% accuracy), confirming the value hierarchy and ensemble complementarity.

Post-processing impact. Arabic text normalization and markdown removal improved Sub-Task 2 BERTScore by approximately 2-3%, demonstrating the importance of output standardization for evaluation metrics.

6 Conclusion

We presented a compact, prompt-engineering-based pipeline for Arabic clinical QA that performs robustly across diverse formats without fine-tuning. A small ensemble improves Sub-Task 1 classification, while a unified instruction guides Sub-Task 2 generation. Future extensions include retrieval augmentation with vetted Arabic medical sources, broader model diversity, and human-in-the-loop validation to mitigate ambiguity and domain gaps.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi.

Table 1: Sub-Task 1 (Classification) official results on test set. All experiments used Gemini 2.5 Flash.

Approach	Description	Accuracy (%)	Ranking
English Translation (ETA)	Translate to English then answer	69.0	–
Arabic Few-Shot (AFS)	Arabic instruction + 6 medical examples	71.0	–
Refinement + Answer (RFA)	Question clarification + option explanation	74.0	–
Ensemble (Final)	**3-way majority vote (RFA + AFS + ETA)**	**76.0**	**2nd**

Note: Individual systems not submitted separately; ensemble represents official submission.

Table 2: Sub-Task 2 (Generation) official results on test set using unified prompting approach.

Approach	Description	BERTScore (%)
Unified Arabic Prompting	Single prompt with Arabic medical expert role-playing, few-shot examples, handles all question formats	**86.953**
Final Ranking	**Official AraHealthQA-2025 shared task**	**2nd place**

BERTScore combines BLEU, ROUGE, and semantic similarity metrics.

2021. Arbert & marbert: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4510–4521.

Mouath Abu Daoud, Chaimae Abouzahir, Leen Kharouf, Walid Al-Eisawi, Nizar Habash, and Farah E Shamout. 2025. Medarabiq: Benchmarking large language models on Arabic medical tasks. *arXiv e-prints*, pages arXiv–2505.

Youssef Al Hariri and Ibrahim Abu Farha. 2024. SMASH at StanceEval 2024: Prompt engineering LLMs for Arabic stance detection. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 800–806, Bangkok, Thailand. Association for Computational Linguistics.

Ashwag Alasmari, Sarah Alhumoud, and Waad Alshammari. 2024. Aramed: Arabic medical question answering using pretrained transformer language models. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT)*, pages 50–56.

Hassan Alhuzali, Farah Shamout, Muhammad Abdul-Mageed, Chaimae Abouzahir, Mouath Abu-Daoud, Ashwag Alasmari, Walid Al-Eisawi, Renad Al-Monef, Ali Alqahtani, Lama Ayash, and 1 others. 2025. Arahealthqa 2025 shared task description paper. *arXiv preprint arXiv:2508.20047*.

Jennifer Chu-Carroll, Krzysztof Czuba, John Prager, and Abraham Ittycheriah. 2003. In question answering, two heads are better than one. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 24–31.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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 (NAACL-HLT)*, pages 4171–4186.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wentau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 9459–9474.

Xiangru Tang, Yiming Zong, Jason Phang, Yilun Zhao, Wangchunshu Zhou, Arman Cohan, and Mark Gerstein. 2024. Struc-Bench: Are large language models good at generating complex structured tabular data? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 12–34, Mexico City, Mexico. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V

Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

A Tables

This appendix contains tables referenced in the main text.

A.1 Classification Examples

Table 3: Examples on classification problem (Sub-Task 1).

Type	Inputs	Outputs
Multiple Choice Questions	كل ما يلي صحيح عن السفلس ماعدا: أ. يتميز الطور الأول بسلبية الاختبارات المصلية؛ ب. يتميز الطور الأول بقرحة صلبة مؤلمة على الأعضاء التناسلية؛ ج. يتميز الطور الثاني باندفاعات على الجلد والأغشية المخاطية؛ د. 25% من الأجنة تموت بعد الولادة من أم مصابة؛ هـ. يعاني الطفل المصاب بالزهري الخلقي من أسنان هوتشنسن.	ب
Fill-in-the-blank with choices	الحمرة هي عدوى جلدية تسببها _____، وهي تصيب عادةً _____ الوجه. أ. المكورات العنقودية الذهبية، البشرة؛ ب. العقديات B و C، الأدمة الشبكية؛ ج. العقديات A و G، الأدمة الحليمية؛ د. العصيات سلبية الجرام، الأنسجة تحت الجلد.	ج

A.2 Generation Examples

Table 4: Examples on generation problem (Sub-Task 2).

Type	Input	Output
Fill-in-the-blank	الحمرة هي عدوى جلدية تسببها _____، وهي تصيب عادةً _____ الوجه.	العقديات A و G، الأدمة الحليمية
Patient-Doctor Q&A	انا امرأة عمري 24 سنة، اشعر بألم ف بطني شديد الألم اشعر بعصر ف بطن و مغص و غثيان و فقدان شهيه...	يرجى عمل تحليل البراز وموافتنا بالنتيجة لتحديد العلاج...
Grammatical Error Correction (GEC)	انا امرأة عمري 24 سنة، اشعر بألم ف بطني شديد الألم اشعر بعصر ف بطن و نعرات ف البطن...	يرجى عمل تحليل البراز وموافتنا بالنتيجة لتحديد العلاج...
LLM Paraphrasing	انا امرأة عمري 24 سنة، لدي ألم في بطني مصحوب بمغص وغثيان وفقدان للشهية...	يرجى عمل تحليل البراز وموافتنا بالنتيجة لتحديد العلاج...

A.3 Data Refinement Examples

Table 5: Data refinement examples showing improvements in question clarity and formatting.

Version	Issue	Question Text
Original	Unclear formatting	في التهاب المشيمية تصادف الأشكال الالتهابية التالية: (الخاطئة) أ. التهاب مشيمية نتحي ب. التهاب مشيمية منتشر ج. التهاب مشيمية أمامية د. التهاب مشيمية مركزي ه. التهاب مشيمية زاوي
	Ambiguous phrasing	كل ما يخص النيجيرية الدجاجية صحيح ما عدا: أ. تعيش هذه المتحولة بشكل حر في الماء والتربة ب. تسبب حالات التهاب سحايا ودماغ بدئي العائل الناقل الذباب ج. يعيش هذا الطفيلي في المياه المعدنية الساخنة ناقصة الأكسجة د. تعتبر النيجلرية التجاجية أسرع وأشد إمراضية من الشوكيية
Refined	Clear formatting	في التهاب المشيمية تصادف الأشكال الالتهابية التالية، ما عدا: أ. التهاب مشيمية نتحي (التهاب موضعي محدود في منطقة معينة) ب. التهاب مشيمية منتشر (التهاب يشمل مناطق واسعة) ج. التهاب مشيمية أمامي (التهاب في الجزء الأمامي من المشيمية) د. التهاب مشيمية مركزي (التهاب في المنطقة المركزية) ه. التهاب مشيمية زاوي (مصطلح غير دقيق طبياً)
	Enhanced clarity	كل ما يلي صحيح عن النيجلرية الدجاجية ما عدا: أ. تعيش بشكل حر في الماء والتربة (كائن حي مجهري حر المعيشة) ب. تسبب التهاب السحايا والدماغ الأولي (عدوى خطيرة في الجهاز العصبي) ج. العائل الناقل هو الذباب (معلومة خاطئة - لا ينتقل عبر الذباب) د. تعيش في المياه المعدنية الساخنة قليلة الأكسجة (بيئة خاصة للنمو) ه. أسرع وأشد إمراضية من الشوكيية (خصائص مرضية مميزة)
Fill-in-blank	Missing context	نستخدم أغشية مصنوعة من _____ في تقنيات التبقيع. أ. النايلون أو السيللوز ب. site acyl Amino ج. site Peptide د. الاسيتونتريل مع مادة TEAA
	Clear context	في تقنيات التبقيع المخبرية، نستخدم أغشية مصنوعة من _____: أ. النايلون أو السيللوز (مواد ماصة للبروتينات) ب. acyl Amino site (موقع ربط الأحماض الأمينية) ج. site Peptide (موقع تكوين البتيدات) د. الأسيتونيتريل مع TEAA (مذيبات كروماتوغرافية)

A.4 Hyperparameters

Table 6: Decoding hyperparameters used for all experiments with Gemini 2.5 Flash.

Parameter	Value
Temperature (τ)	0.1
Top-p	0.8
Top-k	40

B Prompt Templates

This appendix contains the complete prompt templates used in our experiments for reproducibility.

Table 7: Complete prompt templates used in Sub-Task 1 and Sub-Task 2.

Prompt Type	Template
Arabic Few-Shot (AFS)	أنت مساعد طبي خبير وموثوق للغاية. مهمتك هي الإجابة بدقة لا متناهية على الأسئلة الطبية المقدمة باللغة العربية، مع الالتزام التام بتنسيق الإجابة المطلوب. نوع الأسئلة التي ستتلقاها: أسئلة الاختيار من متعدد: تتضمن سؤالاً وخيارات إجابة مرقمة بأحرف عربية (أ، ب، ج، د،). أسئلة إكمال الفراغ: تتضمن جملة أو فقرة بها فراغ واحد أو أكثر، وتُتبع بخيارات إجابة مرقمة. Few-shot examples: المثال 1 (علم الأدوية): السؤال: هـ. لا يجوز مشاركة الكازولين مع البكتين. الإجابة الصحيحة: هـ [... 5 more examples] التعليمات الأساسية للإجابة: 1. الفهم الشامل: اقرأ السؤال وجميع الخيارات المتاحة بعناية فائقة. 2. استخدام المعرفة: استعن بمعرفتك العميقة والموثوقة في المجالات الطبية. 3. تحديد الإجابة الصحيحة: اختر الخيار الأنسب. 4. صيغة الإجابة المطلوبة (صارمة): يجب أن تكون إجابتك حرفاً عربياً واحداً فقط.
Translation (ETA)	You are a medical translation expert. Translate the following Arabic medical question into English following these exact requirements: 1. Maintain the medical accuracy and terminology 2. Format the question properly with options A, B, C, D, E 3. Use "**except**" formatting when the question asks for the wrong/false option 4. Keep the medical context and meaning intact 5. Use proper English medical terminology
Refinement (RFA)	أنت خبير في الطب وتحرير النصوص الطبية. مهمتك هي تحسين وضوح وسلاسة الأسئلة الطبية التالية باللغة العربية مع الحفاظ على: 1. المعنى الطبي الدقيق 2. تنسيق الخيارات (أ، ب، ج، د، ه) 3. الفراغات للأسئلة من نوع "املأ الفراغ" 4. الأرقام والرموز العلمية 5. المصطلحات الطبية باللغة الإنجليزية كما هي. مطلوب إضافي: أضف شرحاً مختصراً (15-25 كلمة) لكل خيار من الخيارات لتوضيح المعنى الطبي.
Generation (Sub-Task 2)	أنت طبيب خبير ومستشار صحي موثوق، ومتخصص في تقديم إجابات طبية دقيقة ومحترفة باللغة العربية. مهمتك هي الإجابة على استفسارات طبية متنوعة، تتراوح بين إكمال الفراغات والرد على استشارات المرضى. التعليمات الأساسية: 1. التحليل الدقيق: اقرأ السؤال أو الاستشارة بعناية فائقة لفهم السياق الطبي المطلوب. 2. استحضار المعرفة: استخدم معرفتك المتعمقة في الطب والعلوم السريرية. 3. صيغة الإجابة المطلوبة: لأسئلة إكمال الفراغ: أجب فقط بالكلمة أو الكلمات المطلوبة. للاستشارات الطبية المفتوحة: قدم إجابة مباشرة ومفيدة. 4. الالتزام بالمصطلحات: استخدم المصطلحات الطبية الصحيحة باللغة العربية. 5. التجنب: لا تكتب أي تفسير أو شرح إضافي. 6. التخصيص: انتبه للمعلومات التي تخص المريض.