

Fahmni at AraHealthQA Track 1: Multi-Agent Retrieval-Augmented Generation and Multi-Label Classification for Arabic Mental Health Q&A

Caroline Sabty
German International University
caroline.sabty@giu.edu.eg

Mohamad Rasmy
MBZUAI
m.rasmy@mbzuai.ac.ae

Mohamed Eyad Badran
Gameball Company
mohamed.eyad2612@gmail.com

Nourhan Sakr
American University in Cairo
n.sakr@columbia.edu

Alia El Bolock
American University in Cairo
alia.elbolock@aucegypt.edu

Abstract

We present FAHMNI, a unified system for Arabic mental-health question answering developed for the AraHealthQA 2025 MentalQA Shared Task (Track 1). FAHMNI evaluates large language models (LLMs) on all subtasks: (1) multi-label classification of question types and (2) answer strategies, and (3) grounded answer generation. For Subtasks 1–2, we systematically compare Arabic-capable LLM families (Qwen3, SILMA) under zero-shot and few-shot prompting, few-shot learning with a frozen backbone, parameter-efficient fine-tuning (PEFT), and instruction tuning. To support Subtask 3, we implement a multi-agent, retrieval-augmented generation pipeline that routes queries between curated domain sources and controlled web search; an answer-style controller predicts the required strategy (Information, Direct Guidance, Emotional Support) and conditions the generator accordingly. Our best LLM configurations reach 0.507/0.404 (weighted-F1/Jaccard) on Subtask 1 with Qwen3+PEFT and 0.750/0.600 on Subtask 2 with SILMA+PEFT, while a strong fine-tuned MARBERT baseline remains competitive at 0.541/0.494 (Subtask 1) and 0.805/0.727 (Subtask 2). For Subtask 3, our multi-agent RAG system with SILMA attains an 0.652 BERTScore F1 and yields a 0.06 hallucination rate under our manual audit. These findings highlight both the viability and current limits of Arabic-capable LLMs for mental-health QA, and they motivate grounded, style-aware generation as a practical path for safe deployment.

1 Introduction

Despite the growing global awareness of mental health needs, Arabic remains severely underrepresented in mental health NLP resources. Existing work on Arabic mental health question answering (QA) is limited in both scale and task coverage, hindering the development of reliable digital support tools, e.g. triage, education, guided self-help, for

Arabic speakers. Complementary efforts on mental health text classification, such as cognitive distortions detection with data augmentation (Rasmy et al., 2024), highlight the importance of tailored augmentation for improving robustness in this sensitive domain. The AraHealthQA 2025 shared task (Track 1) (Alhuzali et al., 2025) addresses this gap by introducing Arabic *mental-health QA* across three subtasks: (1) multi-label classification of question types; (2) multi-label classification of answer strategies; and (3) answer generation (Alhuzali et al., 2024). To tackle all three subtasks, we develop FAHMNI, a unified system for Arabic mental-health QA. Our system leverages two modern Arabic-capable LLM families: **Qwen3** and **SILMA** (SILMA9BInstruct, 2024) / **Kashif** family (SILMA-AI, 2025), motivated by the strength of their predecessors in multilingual transfer for Arabic health retrieval and QA on the Massive Text Embedding Benchmark (MTEB) (Enevoldsen et al., 2025) and their competitive Arabic benchmarks including Arabic RAG-style QA on the Arabic Broad Leaderboard (Ouda, 2025).

Our approach for Subtasks 1 and 2 compares zero-shot prompting, few-shot prompting, and few-shot learning under both frozen-backbone, parameter-efficient fine-tuning (PEFT), and instruction tuning regimes. For Subtask 3, we design a multi-agent, retrieval-augmented answer generation system that dynamically routes queries, integrates curated domain resources, and invokes open-web retrieval when coverage is insufficient. We summarize our contributions as follows:

1. **Comprehensive evaluation** of state-of-the-art Arabic-capable LLMs on all three AraHealthQA subtasks, spanning prompting and fine-tuning strategies.
2. **A novel multi-agent, retrieval-augmented architecture** that moves beyond prior classification-only evaluations and offers

grounded answer generation.

3. **Reproducible resources** including code, configurations, and prompts to support future Arabic mental-health QA research.

Our system achieves competitive results across all three subtasks: For Subtask 1, we use Qwen3 under PEFT (weighted-F1 = 0.51; Jaccard = 0.4) and for Subtask 2, we choose SILMA under PEFT (weighted-F1 = 0.75; Jaccard = 0.6). Finally, for Subtask 3, the SILMA Kashif model reaches a BERTScore of 0.652. In practice, we encountered three recurring challenges: label overlap across clinically adjacent categories, dialectal and terminology variation, and limited data availability due to the small training split. Our code is available at <https://github.com/MHRasmy/AraHealthQA-2025-Track-1>.

2 Background

The shared task uses the MentalQA corpus of Arabic patient–doctor Q&A pairs annotated for seven *question types* (Diagnosis, Treatment, Anatomy & Physiology, Epidemiology, Healthy Lifestyle, Provider Choice, Other) and three *answer strategies* (Information, Direct Guidance, Emotional Support) (Alhuzali et al., 2024). The annotation study reports substantial reliability (Fleiss’ $\kappa = 0.61$ for question types; $\kappa = 0.96$ for answer strategies). Track 1 releases 500 Q&A posts with splits of 350 (train_dev) and 150 (test). Official metrics are weighted F1 and Jaccard for Subtasks 1–2 (multi-label classification), and BERTScore for Subtask 3 (grounded answer generation conditioned on classifications). For readers unfamiliar with MentalQA-style Q&A posts, we include illustrative Arabic examples in Appendix A.

Previous benchmarks (Alhuzali and Alasmari, 2025) compared classical SVM features, frozen PLM encoders, fine-tuned Arabic PLMs (e.g., AraBERT, CAMeLBERT, MARBERT), and GPT-3.5/4-based prompting. The fine-tuned MARBERT showed strongest classification performance, with few-shot prompting outperforming zero-shot. We adopt this model as a well-established baseline and extend the line of work by evaluating newer Arabic-capable LLMs (Qwen3, SILMA) under zero-shot, few-shot prompting, few-shot learning, fine-tuning, and instruction tuning regimes for Subtasks 1–2, and by operationalizing grounded answer generation for Subtask 3 via a multi-agent, retrieval-augmented design.

3 System Overview

In this work, we introduce FAHMNI, a single, modular architecture that couples classification and grounded generation, thereby addressing all three AraHealthQA Track 1 subtasks.

Subtasks 1–2 (multi-label classification). For Tasks 1 and 2, we evaluate five approaches with the Arabic-capable LLM families **Qwen3** and **SILMA**: *zero-shot prompting*, *few-shot prompting*, *few-shot learning*, *PEFT*, and *instruction tuning*. In zero-shot prompting, models receive only label definitions; few-shot prompting augments this with compact, label-balanced exemplars. To move beyond prompting without overfitting in a small-data setting, we train a shallow classification head over frozen LLM representations (“few-shot learning”). Finally, we perform instruction tuning in zero- and few-shot settings. This progression lets us quantify how much the task benefits from parametric specialization versus prompt conditioning under multi-label imbalance and clinically adjacent categories (e.g., Diagnosis vs. Treatment).

Subtask 3 (grounded answer generation via RAG). Given the sensitivity of mental-health counseling, responses should be *grounded*, *factual*, and *style-appropriate*. We, therefore, adopt retrieval-augmented generation (RAG) for Task 3, based on evidence that RAG improves faithfulness and reduces hallucinations on knowledge-intensive tasks (Lewis et al., 2020; Ayala and Bechard, 2024). Our pipeline (Fig. 1) is organized around a *decision agent*, which first inspects the query along with available candidate passages retrieved from the local knowledge base, then uses few-shot prompting (details in Appendix B) to select between a static, curated knowledge base and a dynamic web retrieval path.

(a) *Static domain-specific retrieval.* For well-scoped questions, the system consults a curated local knowledge base assembled from canonical references: DSM-5-TR (Association, 2022) for *Diagnosis*, OpenStax Anatomy & Physiology (Betts et al., 2024) for *Anatomy & Physiology*, CDC’s Principles of Epidemiology (Edition, 2006) for *Epidemiology*, and MedlinePlus articles for *Provider Choice* and general guidance. Retrieved passages are retrieved by similarity search (Qwen3-Embedding, 4B variant) and provided as grounding for the answer. Static retrieval yields high-precision responses but is limited by coverage gaps (e.g.,

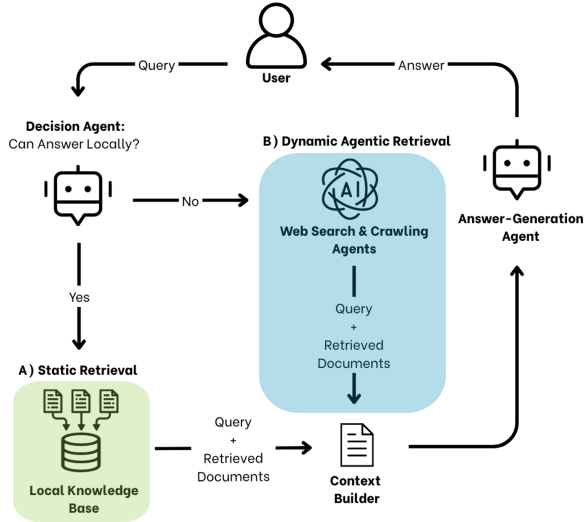


Figure 1: Multi-agent retrieval-augmented generation (RAG) pipeline for Subtask 3. A single LLM (SILMA or Qwen-3) serves *both* as the Decision/Generation agent: it first decides whether the query can be answered locally without retrieval; if not, it triggers retrieval, and later generates the final answer (details in Appendix A). Two retrieval paths are supported: (A) **Static Retrieval** from a curated local knowledge base for scoped domains, and (B) **Dynamic Agentic Retrieval** that launches web search and crawling agents to acquire evidence when curated coverage is insufficient. The retrieved documents are summarized by the agents and incorporated into the prompt, which is then provided to the LLM to produce a grounded response returned to the user.

Treatment and Healthy Lifestyle are too broad for a single canonical source).

(b) *Dynamic agentic retrieval*. For broader or open-ended queries, the decision agent triggers a web-based retrieval pipeline. Here, dedicated *Gemini-2.0-Flash* agents perform web search and crawling to acquire evidence from reliable sources (e.g., WHO, NIH, Mayo Clinic, CDC). Retrieved content is summarized by the agents, assembled into a context prompt, and then passed to the answering LLM (Qwen3 or SILMA), which generates the final grounded response.

4 Experimental Setup

Data splits. We follow the shared-task protocol: the training split contains 350 instances, which we partition into 300 for training and 50 for validation; the test set contains 150 instances. For few-shot classification, exemplars are chosen to cover *all* labels so the model observes at least one positive instance per class.

Hyperparameters. For fine-tuning in Tasks 1–2, we use a learning rate of 2×10^{-5} , batch size 8, and train for up to 10 epochs with early stopping on weighted F1 (validation split). These hyperparameters were chosen to match those in (Alhuzali and Alasmari, 2025) for consistency with the MARBERT baseline. We fix the random seed across all runs for reproducibility. For Task 3 generation, we set the temperature to 0 and disable sampling to obtain deterministic outputs for both model families.

Evaluation. Tasks 1–2 are evaluated with weighted F1 and the Jaccard index. Task 3 is evaluated with BERTScore (Zhang* et al., 2020).

Additional evaluation for Task 3 (RAG quality). Because mental health is a highly sensitive domain, we complemented standard metrics with domain-tailored ones to better capture answer quality and errors. Following Zhu et al. (Zhu et al., 2025), we report *Completeness* (coverage of extracted gold key points), *Hallucination* (contradictions), and *Irrelevance* (omissions). These metrics provide a granular view of factual reliability beyond BERTScore. Formal definitions and scoring details are given in Appendix C.

5 Results

5.1 Quantitative Performance

Table 1 reports official test-set results for Subtasks 1 (question-type classification) and 2 (answer-strategy classification) across baseline fine-tuning, few-shot prompting, parameter-efficient fine-tuning (PEFT), and instruction tuning.

For **Subtask 1**, the baseline fine-tuned model attains the strongest weighted F1 (**0.541**) and Jaccard (**0.494**). PEFT models follow (Qwen: F1 0.507; SILMA: F1 0.497), while few-shot prompting (Qwen) trails (F1 0.440). The instruction-tuned few-shot Qwen variant reaches F1 0.533 but a lower Jaccard 0.412, suggesting more partial label overlap than exact set matches.

For **Subtask 2**, the baseline fine-tuned model again leads (F1 **0.805**; Jaccard **0.727**). Among non-baseline settings, PEFT (SILMA) is strongest (F1 0.753; Jaccard 0.670), followed by instruction-tuned few-shot Qwen (F1 0.738; Jaccard 0.651). Empty predictions are rare and appear mainly in PEFT settings.

Task / Method	F1	Jac.	Empty
<i>Subtask 1: Question Type Classification</i>			
Baseline FT	0.541	0.494	0
Few-shot (Qwen)	0.440	0.453	0
PEFT (Qwen)	0.507	0.434	7
PEFT (SILMA)	0.497	0.422	7
Instr. Tuning Few-shot (Qwen)	0.533	0.412	0
<i>Subtask 2: Answer Strategy Classification</i>			
Baseline FT	0.805	0.727	0
Few-shot (Qwen)	0.622	0.572	0
PEFT (Qwen)	0.701	0.607	2
PEFT (SILMA)	0.753	0.670	1
Instr. Tuning Zero-shot (Qwen)	0.646	0.589	0
Instr. Tuning Few-shot (Qwen)	0.738	0.651	0

Table 1: Official test-set results for Subtasks 1 and 2. Best per subtask in bold.

5.2 Error Analysis

Table 2 shows the distribution of exact, partial, and wrong predictions. We expand here on why models make mistakes.

Subtask 1 (question types). The baseline FT has the highest partial-match rate (60.67%), which explains its strong F1 and Jaccard scores: it often identifies part of the correct set of question types, but misses others. Few-shot (Qwen) gives the highest exact rate (24.67%) but also the highest wrong rate (29.33%), meaning it sometimes predicts all labels correctly but more often misclassifies completely. PEFT variants stay competitive on partial matches but achieve fewer exact hits.

Looking at the labels, we see frequent misses on *Healthy lifestyle*, *Epidemiology*, and *Treatment*, while *Diagnosis* and *Treatment* are often added incorrectly. This indicates that the models sometimes confuse overlapping categories: for example, lifestyle-related questions are mistaken as treatment-related, and prognosis/etiology questions (epidemiology) are mistaken as diagnostic ones. The instruction-tuned few-shot Qwen reflects this tendency clearly: it achieves the highest partial rate (79.33%) but only 7.33% exact, as it often adds extra labels such as *Diagnosis* or *Treatment* while missing *Healthy lifestyle*. This increases recall but reduces exact agreement.

Subtask 2 (answer strategies). Here, the baseline FT achieves the best balance with the highest exact rate (48.67%) and the lowest wrong rate (3.33%). PEFT (Qwen) produces the most partial predictions (56.67%), often identifying one correct strategy but missing another. Across systems, the most common source of errors comes from *Information* and *Direct Guidance*: answers that mix fac-

tual knowledge with advice are difficult for models to consistently label, causing under-prediction or over-prediction of these two categories. Instruction-tuned few-shot Qwen improves over zero-shot by converting some wrong cases into partial matches, showing that in-context examples help the model separate advice from information.

Empty predictions. Empty outputs occur when all predicted scores fall below the decision threshold of 0.5. They are rare but appear mainly in PEFT runs (S1: 7 for Qwen, 7 for SILMA; S2: 2 for Qwen, 1 for SILMA). In these cases, the model is overly conservative, assigning low confidence to all categories and outputting no label.

Takeaways. Across both subtasks, the main challenges are (i) partial matches caused by overlapping categories, such as *Diagnosis* vs. *Treatment* or *Information* vs. *Direct Guidance*, and (ii) threshold-related errors that lead to either empty predictions or the addition of extra labels. These issues explain why the baseline FT remains the strongest overall: it provides more balanced predictions with higher exact matches, while instruction tuning (Subtask 1) trades exactness for broader coverage.

Task / Method	Exact %	Partial %	Wrong %
<i>Subtask 1</i>			
Baseline FT	22.67	60.67	16.67
Few-shot (Qwen)	24.67	46.00	29.33
PEFT (Qwen)	16.67	58.00	25.33
PEFT (SILMA)	16.67	57.33	26.00
Instr. Tuning Few-shot (Qwen)	7.33	79.33	13.33
<i>Subtask 2</i>			
Baseline FT	48.67	48.00	3.33
Few-shot (Qwen)	39.33	37.33	23.33
PEFT (Qwen)	33.33	56.67	10.00
PEFT (SILMA)	40.67	52.67	6.67
Instr. Tuning Zero-shot (Qwen)	36.67	45.33	18.00
Instr. Tuning Few-shot (Qwen)	38.67	53.33	8.00

Table 2: Error distribution for Subtasks 1 and 2. Best per column and subtask in bold.

Model	SILMA	Qwen
BERTScore ↑	0.652	0.645
Completeness ↑	0.567	0.6
Hallucination ↓	0.06	0.04
Irrelevance ↓	0.373	0.36

Table 3: Task 3 (RAG answer generation) results on the MentalQA test set. We report BERTScore F1, Completeness, Hallucination, and Irrelevance.

5.3 Task 3: Answer Generation (RAG)

Table 3 summarizes results for SILMA and Qwen-3. Both models perform similarly overall. SILMA attains a slightly higher BERTScore (0.652 vs. 0.645), while Qwen-3 achieves higher Completeness (0.600 vs. 0.567) and lower Hallucination (0.04 vs. 0.06) and Irrelevance (0.36 vs. 0.373). Qualitative best and worst examples for each model are provided in Appendix D.

The uniformly low Hallucination rates (≤ 0.06) indicate that generated answers rarely contain content that *contradicts* the gold key points, suggesting that the RAG pipeline effectively constrains factual errors. At the same time, completeness around 0.57–0.60 shows that only about three-fifths of the gold key information is covered, leaving a substantial fraction of gold content unaddressed (Irrelevance 0.36–0.373). This explains the moderate BERTScore values (≈ 0.65): limited key-point overlap and the inclusion of additional retrieved details (which are non-contradictory but not present in the references) dilute semantic alignment with the gold answers, lowering BERTScore despite the low Hallucination.

6 Conclusion

We presented FAHMNI, a unified system for Arabic mental-health question answering that combines multi-label classification (question types and answer strategies) with a retrieval-augmented, multi-agent generator. On Subtasks 1–2, classic Arabic PLMs remain a strong baseline: fine-tuned MARBERT delivers the best weighted F1 and Jaccard overall, while Arabic-capable LLMs (Qwen3, SILMA) with PEFT and instruction tuning are competitive under tighter compute and data budgets. On Subtask 3, both SILMA and Qwen3 yield similarly strong grounded generation with uniformly *low hallucination* rates (≤ 0.06), indicating faithful adherence to evidence. At the same time, mid-range BERTScore and ~ 0.6 completeness reveal recall gaps: answers are generally factual but do not fully cover gold key points, and extra retrieved details can dilute reference overlap.

Acknowledgments

This research was supported by Emirates NBD, Sharjah Electricity, Water & Gas Authority (SEWA), and the Technology Innovation Institute (TII), who served as the golden sponsors of the 5th Forum for Women in Research (QUWA): Together

Innovating to Shape the Future at the University of Sharjah.

References

- Hassan Alhuzali and Ashwag Alasmari. 2025. *Pre-trained language models for mental health: An empirical study on arabic qa classification*. *Healthcare*, 13(9).
- Hassan Alhuzali, Ashwag Alasmari, and Hamad Al-saleh. 2024. *Mentalqa: An annotated arabic corpus for questions and answers of mental healthcare*. *IEEE Access*, 12:101155–101165.
- Hassan Alhuzali, Farah Shamout, Muhammad Abdul-Mageed, Chaimae Abouzahir, Mouath Abu-Daoud, Ashwag Alasmari, Walid Al-Eisawi, Renad Al-Monef, Ali Alqahtani, Lama Ayash, Nizar Habash, and Leen Kharouf. 2025. *Arahealthqa 2025 shared task description paper*. In *Proceedings of ArabicNLP 2025*.
- American Psychiatric Association. 2022. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5-TR*. American Psychiatric Association Publishing.
- Orlando Ayala and Patrice Bechard. 2024. *Reducing hallucination in structured outputs via retrieval-augmented generation*. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 228–238, Mexico City, Mexico. Association for Computational Linguistics.
- J Gordon Betts, Kelly A Young, James A Wise, Eddie Johnson, Brandon Poe, Dean H Kruse, Oksana Korol, Jody E Johnson, Mark Womble, and Peter DeSaix. 2024. *Anatomy and physiology 2e*.
- Third Edition. 2006. *Principles of epidemiology*.
- Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrom, Roman Solomatin, and 67 others. 2025. *Mmteb: Massive multilingual text embedding benchmark*. *arXiv preprint arXiv:2502.13595*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. *Retrieval-augmented generation for knowledge-intensive nlp tasks*. *Advances in neural information processing systems*, 33:9459–9474.
- Karim Ouda. 2025. *ABBL: An advanced benchmark and leaderboard for comprehensive evaluation of arabic language models*. Accessed: 2025-07-19.

Mohamad Rasmy, Caroline Sabty, Nourhan Sakr, and Alia El Bolock. 2024. Enhanced cognitive distortions detection and classification through data augmentation techniques. In *Pacific Rim International Conference on Artificial Intelligence*, pages 134–145, Singapore. Springer Nature Singapore.

SILMA-AI. 2025. Silma kashif 2b instruct v1.0. <https://huggingface.co/silma-ai/SILMA-Kashif-2B-Instruct-v1.0>.

SILMA9BInstruct. 2024. Silma 9b instruct v1.0. <https://huggingface.co/silma-ai/SILMA-9B-Instruct-v1.0>.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. *Bertscore: Evaluating text generation with bert*. In *International Conference on Learning Representations*.

Kunlun Zhu, Yifan Luo, Dingling Xu, Yukun Yan, Zhenghao Liu, Shi Yu, Ruobing Wang, Shuo Wang, Yishan Li, Nan Zhang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. *Rageval: Scenario specific rag evaluation dataset generation framework*. *Preprint*, arXiv:2408.01262.

A Illustrative Q&A Examples (Arabic)

Example 1

Question type: Treatment

س: أعاني من القلق، ووصف لي الطبيب لوسترال .ه ملغم مرة يوميًا. لدي سؤال: هل قد تسبب هذه الجرعة زيادة في الوزن؟ وهل قد تؤدي إلى ضعف في الاداء الجنسي؟ (أنا مقبل على الزواج)

Answer strategy: Information

ج: قد يزيد لوسترال الشهية لدى بعض المرضى، بينما قد لا يؤثر في آخرين. وقد يقلل الرغبة الجنسية لدى بعض الرجال والنساء، لكن هذا لا يحدث بالضرورة في جميع الحالات.

Example 2

Question type: Diagnosis

س: أهلاً يا طبيب، أنا أعاني منذ نحو أسبوعين من عدة أعراض: حزن، وضيق

في التنفس، وأفكار عن الموت وإيذاء النفس، وعدم رغبة في الحياة. كما أشعر بالتعب أحيانًا.

Answer strategy: Direct Guidance

ج: راجعي طبيب أمراض نفسية، وسيساعدك في تجاوز الأزمة.

B Task 3: Few-Shot Prompt for Local Answerability

In our RAG pipeline, the answer-generation LLM (SILMA or Qwen-3) first acts as a *decision agent* that inspects the user query together with the candidate passages retrieved from the local knowledge base. Using a few-shot prompt (below), it outputs a single token: Yes if the local passages contain sufficient, explicit information to answer the query faithfully, and No otherwise (e.g., missing, partial, or ambiguous evidence). If the output is Yes, Task 3 proceeds with the *static* path, using the curated local knowledge-base documents; if No, it triggers the *dynamic agentic* retrieval path, as depicted in Fig. 1. We supply two illustrative few-shot exemplars to cover both outcomes. **(Yes)** The first exemplar uses context scraped by the *dynamic web retrieval* (web-scraping) agent; an author manually verified that the passages contain sufficient information to answer the training query faithfully. **(No)** The second exemplar uses context retrieved from the *local knowledge base*; an author verified that these passages are relevant but insufficient to answer the training query.

Few-shot prompt for deciding whether the query can be answered from local context only (Yes/No).

Role: Mental Health Question-Answering Assistant
Task: Determine whether the system can answer the user's mental-health question using ONLY the provided context passages.
Instructions:
- Analyze the context and determine whether it contains the specific

information required to answer the user's mental-health question.

- Provide a clear, concise decision indicating whether the system can answer the question based solely on the context.
- Your response must be exactly one word: either Yes or No.

Output Format:

- Answer: Yes/No

Study the examples and then respond to the last question.

Examples:

Input:

Context: {SUFFICIENT_CONTEXT: passages that contain the answer for the question below}

User Question: {Selected query from the training data}

Expected Output:

Answer: Yes

Input:

Context: {INSUFFICIENT_CONTEXT: passages that are relevant but do NOT contain the specific fact/criterion/instruction required to answer the question below}

User Question: {Selected query from the training data}

Expected Output:

Answer: No

Input:

Context: {Local Context}

User Question: {query}

C RAG Metrics and Evaluation Details

We provide formal definitions and implementation details for the RAG-specific metrics used in Task 3, following Zhu et al. (Zhu et al., 2025).

Key-point references. For each gold answer, we extract a set of concise *key points* with a vanilla LLM—here, Gemini-2.0-flash. These serve as reference units against which a system answer is judged. Let $K = \{k_1, \dots, k_m\}$ denote the key points for one item, and let A denote a system-generated answer.

1. **Completeness.** Measures how well the generated answer covers the ground-truth key points. Let $K = \{k_1, \dots, k_m\}$ be the set of

key points and A the generated answer:

$$\text{Comp}(A, K) = \frac{1}{|K|} \sum_{i=1}^{|K|} \mathbb{1}[A \text{ covers } k_i],$$

where $\mathbb{1}[A \text{ covers } k_i] = 1$ if A semantically includes or paraphrases the content of k_i ; otherwise 0.

2. **Hallucination.** Identifies contradictions between the generated answer and the key points:

$$\text{Hallu}(A, K) = \frac{1}{|K|} \sum_{i=1}^{|K|} \mathbb{1}[A \text{ contradicts } k_i],$$

where $\mathbb{1}[A \text{ contradicts } k_i] = 1$ if A asserts content that conflicts with k_i ; otherwise 0.

3. **Irrelevance.** Captures the proportion of key points that are neither covered nor contradicted:

$$\text{Irr}(A, K) = 1 - \text{Comp}(A, K) - \text{Hallu}(A, K),$$

i.e., key points that the answer omits or does not address.

Operationalization. We prompt the same vanilla LLM in a few-shot setting to (i) extract key points from the gold answer and (ii) judge coverage/contradiction for each k_i given A , with temperature = 0 for determinism.

D Qualitative Examples of Generated Answers

We present text-form qualitative examples for each model. In Subsection D.1, **SILMA**'s highest-scoring output (BERTScore F1=0.78) is annotated *Complete*, whereas its lowest (F1=0.54) is *Wrong* because the generated answer introduces hallucinated content that *contradicts* the reference. In Subsection D.2, **Qwen-3**'s highest-scoring output (F1=0.74) is also *Complete*, while its lowest (F1=0.58) is *Irrelevant*: the response is non-contradictory but fails to cover the reference's key points, largely due to a brief reply that offers only a single piece of advice for an open-ended query where multiple reasonable clinician responses are possible. These observations align with the aggregate metrics: Irrelevance remains non-trivial (0.373 for SILMA; 0.36 for Qwen-3), whereas Hallucination is consistently low (≤ 0.06). Thus, the systems typically produce factual, non-contradictory

responses; remaining errors are dominated by *coverage* rather than factuality, suggesting gains from retrieval/coverage enhancement, keypoint-aware planning, and length/structure control in generation.

D.1 SILMA Generated Responses (best & worst by BERTScore)

Best. BERTScore F1: 0.78 RAG Eval: Complete

Question

س: ماهي الأدوية التي تسبب تبول ليلي للكبار بدون وصفة طبية

Gold answer

ج: اذا كان المقصود الادوية النفسية ، فهناك العديد من الادوية منها المهدئات عموما و ادوية مجموعة البنزوديازيبين

Generated answer (SILMA)

ج: هناك العديد من الأدوية التي تسبب تبول ليلي للكبار بدون وصفة طبية، مثل المهدئات التي تحتوي على البنزوديازيبين. يجب مراجعة الطبيب قبل استخدام أي أدوية إذا كان المقصود الأدوية النفسية، فهناك العديد من الادوية منها المهدئات عموما و ادوية مجموعة البنزوديازيبين. يجب مراجعة الطبيب قبل استخدام أي أدوية.

Worst. BERTScore F1: 0.54 RAG Eval: Wrong (hallucination/contradiction)

Question:

س: اعاني من ادمان جاباتين ثلر منذ مده سنه من غير وصفه طبيه واريد الاقلاع عنه هل اقطعه فجأة ام بالتدريج وكيف اعيد ثقتي بنفسى بعد تركه ع

Gold answer

ج: يقطع فوراً إرادة قوية وتحمل لمدة أسبوعاً متتابعة نفسية وعلاج نفسي مكثف

Generated answer (SILMA)

ج: نعم، من المهم عدم قطع

Gabapentin فجأة. يجب عليك التحدث إلى طبيبك للحصول على المشورة اللازمة. قد تشعر بعدم الأمان، لكن التخلي عن *Gabapentin* بشكل تدريجي يمكنه تحسين حالتك النفسية. اختياراً، يجب عليك التحدث إلى طبيبك للحصول على المساعدة اللازمة. نعم، من المهم عدم قطع *Gabapentin* فجأة. يجب عليك التحدث إلى طبيبك للحصول على المشورة اللازمة. قد تشعر بعدم الأمان، لكن التخلي عن *Gabapentin* بشكل تدريجي يمكنه تحسين حالتك النفسية. اختياراً، يجب عليك التحدث إلى طبيبك للحصول على المساعدة اللازمة.

D.2 Qwen-3 Generated Responses (best & worst by BERTScore)

Best. BERTScore F1: 0.74 RAG Eval: Complete

Question

س: كنت اتناول ادوية نفسية وكنت اعاني من كثرة التفكير والقلق.. هل مع الوقت يمكن ان تتحسن الذاكرة وتعود لقوتها ام لا

Gold answer

ج: بالتأكيد تتحسن ذاكرتك، لكن عندما يخف القلق الشديد لأن كثرة التفكير والقلق يؤثر على الذاكرة ويضعفها.

Generated answer (Qwen-3)

ج: نعم، مع الوقت يمكن أن تتحسن ذاكرتك وتعود إلى قوتها، خاصة إذا تراجعت عن الأدوية النفسية تحت إشراف طبي وتقلصت أعراض القلق والتفكير الزائد. فكثرة القلق والتوتر تؤثر سلباً على وظائف الدماغ، بما في ذلك الذاكرة.

عندما يقل القلق وتحسن حالتك النفسية، يبدأ الدماغ في التعافي تدريجياً، مما يُعيد توازنه ويحسن الذاكرة. كما أن ممارسة الرياضة، والنوم الجيد، والتغذية

التوازن، والتأمل تُسهم في تعزيز الذاكرة وتحسين التركيز.

إذا كنت تتناولين أدوية نفسية، فمن الأفضل استشارة الطبيب النفسي قبل التوقف المفاجئ، لضمان خروج آمن من العلاج.

Worst. BERTScore F1: 0.58 RAG Eval: Irrelevant (coverage gap)

Question:

س: عندما أقوم بفعل شيء لا أقوم به وأصاب بالاحباط

Gold answer

ج: فقدان الشغف انتشر الآن بفضل وسائل التواصل الاجتماعي اقل موبايلك اسبوعاً قابل ناس حقيقيين وصحة طيبة والعلاج النفسي مفيد

Generated answer (Qwen-3)

ج: من الطبيعي أن تشعر بالاحباط عندما تخطط لفعل شيء ما ولا تتمكن من تنفيذه. الاحباط جزء من الحياة، لكن يمكن التعامل معه بطرق صحية. إليك بعض النصائح:

1 . **حدد السبب** : اسأل نفسك لماذا لم تتمكن من فعل ما خططت له. هل كان بسبب نقص الدافع، الوقت، أو التحديات الخارجية؟ فهم السبب يساعدك على التخطيط بشكل أفضل.

2 . **قسّم المهام** : إذا كانت المهمة كبيرة أو معقدة، جرب تقسيمها إلى خطوات صغيرة. هذا يسهل عليك البدء ويزيد من فرص النجاح.

3 . **لا تهمل الانجازات الصغيرة** : حتى لو لم تكمل المهمة بالكامل، اعترف بالخطوات التي قمت بها. كل شيء يُعد تقدماً.

4 . **كن لطيفاً مع نفسك** : لا تنتقد نفسك بقسوة. نحن جميعاً نواجه أياماً لا

تحقق فيها ما نخطط له. المهم هو أن تتعلم وتعيد المحاولة.

5 . **خطط مجدداً** : استخدم التجربة كفرصة لتحسين خططك. قد يكون من المفيد تقييم الوقت أو الطريقة التي تختارها.

إذا شعرت بالاحباط بشكل متكرر أو أثر على حماسك للحياة، قد يكون من المفيد التحدث مع مستشار نفسي. تذكر دائماً: كل خطوة صغيرة تُعد انتصاراً.