mucAI at AraHealthQA 2025: Explain–Retrieve–Verify (ERV) Workflow for Multi-Label Arabic Health QA Classification

Ahmed Abdou
Independent Researcher. Munich, Germany
ahmedabdou1789@gmail.com

Abstract

We present a simple, training-light pipeline for multi-label categorization of Arabic mental-health questions in the AraHealthQA 2025 MentalQA Track 1 (question and answer classification). Our method, Explain–Retrieve–Verify (ERV), couples a chain-of-thought LLM classifier with example-based retrieval and a verifier that arbitrates disagreements. The LLM first proposes candidate labels and rationales from a compact taxonomy prompt. A similarity agent then surfaces top-k nearest questions via multilingual sentence-transformer embeddings to induce case-based priors. A verification agent reconciles both signals to produce a final label set with a calibrated confidence, followed by a lightweight post-processor for code parsing and confidence clamping. ERV requires no fine-tuning or external data and runs efficiently at inference time. In shared-task evaluation, our system achieved 0.61 weighted F1-score for question classification and 0.73 for answer classification. A hybrid approach combining ERV with MARBERT further improves answer classification to 0.80 weighted F1-score.

1 Introduction

The growing burden of mental health conditions worldwide has created unprecedented challenges for healthcare systems. While mental health disorders affect diverse populations globally, access to adequate care remains severely limited, particularly in regions where cultural stigma and resource constraints compound the problem. These barriers underscore the need for scalable and supportive technologies that can assist practitioners in reaching underserved populations (Zolezzi et al., 2018). This gap has motivated the development of computational tools to support mental health professionals, particularly text mining and natural language processing (NLP) systems that can assist in diagnosis and triage rather than replace human expertise (Swaminathan et al., 2023).

While substantial progress has been made in English and other high-resource languages (Ghosh et al., 2020; Atapattu et al., 2022; Chaturvedi et al., 2023), Arabic remains under-studied in the mental health domain despite being spoken by more than 400 million people (Alhuzali et al., 2024; Guellil et al., 2021). This under-representation is critical, as Arabic presents unique challenges for NLP, including morphological richness, dialectal variation, and limited annotated resources.

Recent advances in Pre-Trained Language Models (PLMs) have revolutionized text classification and understanding in biomedical and clinical domains. More recently, Large Language Models (LLMs) (Brown et al., 2020) have introduced new possibilities by not only achieving strong predictive performance but also producing verbalized rationales for their decisions (Abu Daoud et al., 2025; Xie et al., 2025; Yang et al., 2023). This interpretability is particularly valuable in mental health applications, where transparency and human oversight are essential.

In this work, we present the Explain–Retrieve–Verify (ERV) workflow, a LLM-based workflow for Arabic mental health question and answer classification, developed for the AraHealthQA 2025 shared task. ERV operates in three steps: (i) the *Explain step*, where an LLM generates candidate labels with rationales; (ii) the *Retrieve step*, where semantically similar training examples are surfaced to provide case-based evidence; and (iii) the *Verify step*, where LLM reconciles both sources to produce final labels with calibrated confidence. On the official test set, ERV achieves a weighted F1-score of 0.61 and Jaccard score of 0.53 for question classification, outperforming fine-tuned baselines, and when combined with MARBERT (Abdul-Mageed et al., 2020) reaches 0.80 weighted F1 and 0.72 Jaccard for answer classification, the best overall performance.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 226–232

2 Task Definition

The AraHealthQA 2025 shared task (Alhuzali et al., 2025) focuses on Arabic health question-answering with two primary tracks. Track 1 (MentalQA) addresses classification of mental health questions and answers, while Track 2 handles general health topics. We participate in Track 1 sub-track 1 and 2: question classification and answer classification. Both sub-tasks are multi-label classification tasks where instances can belong to multiple categories simultaneously. The competition uses the MentalQA dataset (Alhuzali et al., 2024) with a train/dev/test split of 300/50/150 samples. Evaluation employs weighted F1-score and Jaccard index as primary metrics.

3 Data

The MentalQA dataset contains 500 Arabic question-answer pairs collected from Altibbi.com, focusing on mental health interactions posted between 2020-2021. The dataset encompasses interactions between patients seeking mental health guidance and professional doctors providing responses. Questions are classified into seven types: (A) Diagnosis, (B) Treatment, (C) Anatomy & Physiology, (D) Epidemiology, (E) Healthy Lifestyle, (F) Provider Choice, and (Z) Other. For detailed category definitions and examples, see (Alhuzali et al., 2024). Doctor responses are classified into three communication strategies: (1) Information, (2) Direct Guidance, and (3) Emotional Support. Complete strategy descriptions are available in the original dataset paper (Alhuzali et al., 2024).

4 Method

We present Explain–Retrieve–Verify (ERV) workflow, a training-free multi-agent pipeline that combines explicit reasoning, similarity-based retrieval, and consensus verification for Arabic medical text classification. The system operates through three sequential agents that provide complementary perspectives on multi-label classification decisions.

Explain The Explain step sends either the question or the answer to an LLM with chain-of-thought prompting with Arabic medical contexts. The agent outputs predicted labels, explanations, and confidence scores about the LLM own answer.

Retrieve The Retrieve step identifies semantically similar training examples through embedding-based similarity search. We pre-encode all training texts and cache these embeddings. For each input, we retrieve the k-nearest neighbors based on cosine similarity, and analyze their label patterns. The step passes the retrieved examples to an LLM asking to perform pattern analysis to suggest appropriate categories based on the given similar training cases

4.1 Verify

The Verify step reconciles the outputs from the Explain and Retrieve steps. We prompt LLM with three inputs: (i) the label predictions, rationales, and confidence score from the Explain step, (ii) the retrieved examples with their gold labels and the suggested categories from the Retrieve step, and (iii) the full task taxonomy. The verifier is instructed to compare the two sources of evidence, identify agreements and conflicts, and produce a final multi-label decision. It outputs the final label set, a calibrated confidence score, and a short reconciliation note explaining how disagreements were resolved.

5 Experimental Setup

For fine-tuned baselines, we trained MARBERT (Abdul-Mageed et al., 2020) and AraBERT-v02 (Antoun et al., 2020) using standard hyperparameters: learning rate 2×10^{-5}, batch size 16, 5 epochs with the best checkpoint selected by the highest weighted F1 on the validation set, and weight decay 0.01. For the ERV pipeline, we used GPT-4 as the underlying language model and employed a multilingual Sentence-BERT model (Reimers and Gurevych, 2019)[1] to compute semantic embeddings in the *Retrieve* step. The entire ERV system was implemented using the DSPy framework[2], which provides modular components for prompt engineering and multi-step reasoning workflows. All experiments were conducted on a single NVIDIA A100 GPU on Google Colab. The code used for the experiments is available on GitHub[3].

6 Results

We evaluate our ERV pipeline against fine-tuned Arabic language models on both question and answer classification sub-tasks using weighted F1-

[1] paraphrase-multilingual-mpnet-base-v2, https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

[2] https://dspy.ai/

[3] https://github.com/AhmedAbdel-Aal/mucAI-at-AraHealthQA-2025

Method	Weighted F1	Jaccard
MARBERT	0.56	0.51
AraBERT	0.56	0.51
ERV	**0.61**	**0.53**

Table 1: Question Classification Results on the test set.

Method	Weighted F1	Jaccard
MARBERT	0.76	**0.73**
AraBERT	0.74	0.68
ERV	0.73	0.61
MARBERT + ERV	**0.80**	0.72

Table 2: Answer Classification Results on the test set.

score and instance-based Jaccard index as primary metrics.

Table 1 shows the performance comparison for question classification on the test set. ERV achieves the best performance with a weighted F1-score of 0.61 and instance-based Jaccard of 0.53, outperforming both fine-tuned baselines. The improvement demonstrates the effectiveness of the multi-agent collaboration approach over single-model fine-tuning.

For answer classification (Table 2), fine-tuned models show stronger performance, with MARBERT achieving 0.76 weighted F1-score. ERV performs competitively at 0.73 weighted F1-score but falls behind the fine-tuned baselines on this subtask. Observing that MARBERT struggled to predict label 3 (Emotional Support) during development, we designed a hybrid approach combining MARBERT's expertise with ERV's pattern recognition capabilities. In this hybrid system, MARBERT serves as the primary classifier while ERV specifically handles the detection of emotional support responses. This combination achieves the best overall performance with 0.80 weighted F1-score and 0.72 instance-based Jaccard.

7 Discussion

7.1 Per Label Analysis

Comparing ERV directly with MARBERT reveals fundamentally different approaches to handling class imbalance and provides insights into why each method excels in different scenarios. Tables 3 and 4 present the detailed per-class performance comparison.

The comparison reveals fundamentally different precision-recall strategies between the two approaches. ERV consistently achieves higher recall across most question categories, particularly for dominant classes (A: 0.94 vs 0.79, B: 0.86 vs 0.74). For answers, ERV demonstrates an extreme high-recall strategy on Strategy 2 (Direct Guidance: 0.99 vs 0.86), while MARBERT shows higher recall on Strategy 1(0.89 vs 0.69).

The most striking difference lies in minority class handling. ERV shows remarkable ability to detect minority classes that MARBERT completely misses. For questions, ERV achieves non-zero performance on categories C (F1: 0.20) and F (F1: 0.15), while MARBERT scores zero on these categories. For answers, ERV detects emotional support responses (Strategy 3) with substantial performance (F1: 0.44, precision: 0.37, recall: 0.56), while MARBERT achieves zero performance. This represents a critical 44-point F1 advantage for detecting emotional support in mental health contexts. We hypothesize that this capability stems from ERV's similarity-based approach, which can identify minority class instances by matching them to semantically similar training examples. The hybrid approach validates this complementary strength, achieving the best overall performance (weighted F1: 0.80) by combining MARBERT's precision on majority classes with ERV's minority class detection capabilities.

7.2 Interpretability

A key advantage of large language models is their ability to verbalize reasoning, providing a level of transparency not available in neural fine-tuned models. Our ERV pipeline makes this explicit by combining rationales from the classification agent, evidence from retrieved examples, and the reconciliation notes from the verifier. This interpretability is particularly valuable in mental-health contexts, where system outputs should not only predict labels but also justify decisions in a way that is accessible to human reviewers. To illustrate, we show in Figure 1 a full example of the ERV workflow for question classification, and in Figure 2 an example of the ERV workflow for answer classification.

8 Limitations and Future Work

The current evaluation is constrained by testing single representatives of each modeling paradigm (MARBERT/AraBERT for fine-tuning, multilingual-mpnet-base for Encoding, and GPT-4 for ERV), which limits the generalizability of

Category	ERV			MARBERT			Support
	P	R	F1	P	R	F1	
A (Diagnosis)	0.67	0.94	0.78	0.68	0.79	0.73	84
B (Treatment)	0.63	0.86	0.73	0.72	0.74	0.73	85
C (Anatomy)	0.20	0.20	0.20	0.00	0.00	0.00	10
D (Epidemiology)	0.35	0.21	0.26	0.47	0.21	0.29	34
E (Lifestyle)	0.38	0.61	0.47	0.41	0.29	0.34	38
F (Provider Choice)	0.14	0.17	0.15	0.00	0.00	0.00	6
Z (Other)	0.00	0.00	0.00	0.00	0.00	0.00	3
Weighted Avg	0.54	0.71	0.61	0.57	0.57	0.56	260

Table 3: Per-class performance comparison for question classification

Strategy	ERV			MARBERT			Hybrid	Support
	P	R	F1	P	R	F1	F1	
1 (Information)	0.85	0.69	0.76	0.84	0.89	0.87	0.87	112
2 (Direct Guidance)	0.60	0.99	0.75	0.74	0.86	0.80	0.80	86
3 (Emotional Support)	0.37	0.56	0.44	0.00	0.00	0.00	0.44	18
Weighted Avg	0.71	0.80	0.73	0.73	0.81	0.77	0.80	216

Table 4: Per-class performance comparison for answer classification with hybrid results

our comparative findings. Future work should expand the experimental scope to include diverse Arabic language models of varying sizes and capabilities to establish more robust conclusions about the trade-offs between approaches.

While ERV demonstrates promising results, several limitations suggest important directions for future research. ERV requires multiple LLM calls per instance (three sequential steps plus similarity computation), resulting in significantly higher computational costs compared to single forward passes in fine-tuned models.

Our hypothesis that ERV's similarity-based retrieval drives its minority class detection capability suggests an intriguing research direction: interpolating fine-tuned models like MARBERT with k-nearest neighbors (kNN) at inference time. This approach could potentially provide MARBERT with non-zero performance on minority classes by incorporating retrieval-based evidence while maintaining its strong performance on majority categories.

For PLMs, performance is bounded by the small size of the used dataset. Future work can explore curating a larger corpus with balanced label coverage, and low-risk augmentation (back-translation, controlled paraphrasing).

The interpretable nature of ERV's three-step workflow creates opportunities for human-in-the-loop systems where medical professionals can review and refine step-by-step reasoning. Additionally, more sophisticated verification mechanisms could incorporate uncertainty quantification, confidence-aware voting, and learned arbitration strategies beyond the current simple consensus approach.

9 Conclusion

We present ERV (Explain–Retrieve–Verify), a three-step workflow for Arabic mental health question and answer classification. Our approach combines three sequential steps: the Explain step provides initial predictions through chain-of-thought reasoning, the Retrieve step identifies similar examples for evidence-based analysis, and the Verify step reconciles both signals to produce final classifications with calibrated confidence. Our experiments show that the ERV workflow improves over fine-tuned language models on the question classification task and provides complementary strengths for answer classification, especially in detecting emotional support strategies. Our work contributes to Arabic mental health NLP by demonstrating that collaborative three-step reasoning workflows can compete with fine-tuned models while offering better interpretability.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert &

marbert: Deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

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.

Hassan Alhuzali, Ashwag Alasmari, and Hamad Alsaleh. 2024. Mentalqa: An annotated arabic corpus for questions and answers of mental healthcare. *IEEE Access*.

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*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Thushari Atapattu, Mahen Herath, Charitha Elvitigala, Piyanjali de Zoysa, Kasun Gunawardana, Menasha Thilakaratne, Kasun De Zoysa, and Katrina Falkner. 2022. Emoment: An emotion annotated mental health corpus from two south asian countries. *arXiv preprint arXiv:2208.08486*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jaya Chaturvedi, Sumithra Velupillai, Robert Stewart, and Angus Roberts. 2023. Identifying mentions of pain in mental health records text: a natural language processing approach. *arXiv preprint arXiv:2304.01240*.

Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Cease, a corpus of emotion annotated suicide notes in english. In *Proceedings of the twelfth language resources and evaluation conference*, pages 1618–1626.

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2021. Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*, 33(5):497–507.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Akshay Swaminathan, Iván López, Rafael Antonio Garcia Mar, Tyler Heist, Tom McClintock, Kaitlin Caoili, Madeline Grace, Matthew Rubashkin, Michael N Boggs, Jonathan H Chen, and 1 others. 2023. Natural language processing system for rapid detection and intervention of mental health crisis chat messages. *NPJ digital medicine*, 6(1):213.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, and 1 others. 2025. Medical foundation large language models for comprehensive text analysis and beyond. *npj Digital Medicine*, 8(1):141.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. *arXiv preprint arXiv:2304.03347*.

Monica Zolezzi, Maha Alamri, Shahd Shaar, and Daniel Rainkie. 2018. Stigma associated with mental illness and its treatment in the arab culture: a systematic review. *International Journal of Social Psychiatry*, 64(6):597–609.

A Appendix A

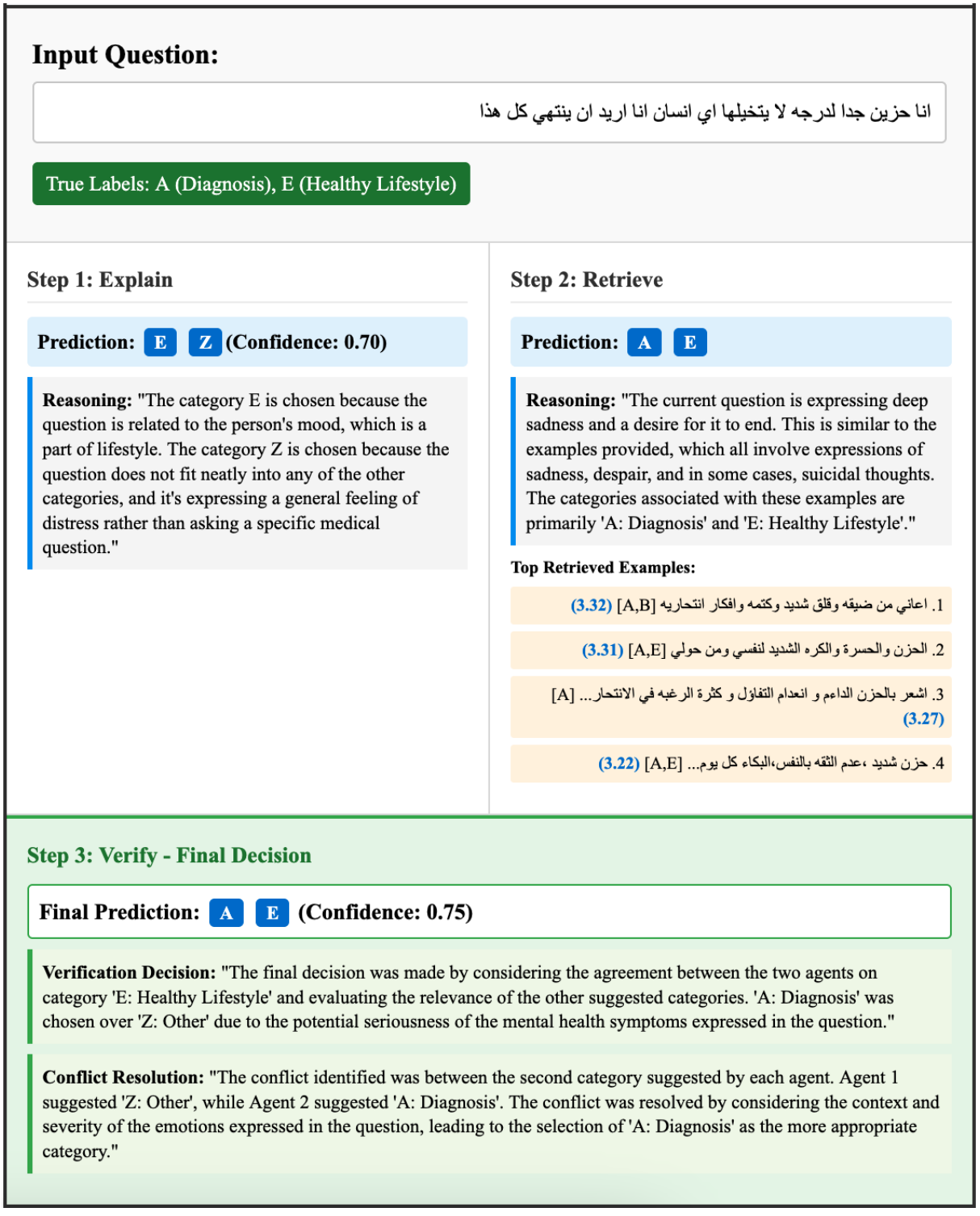

Figure 1: ERV pipeline workflow showing the three-step process for Arabic mental health question classification. The Explain step provides initial predictions through chain-of-thought reasoning, the Retrieve step identifies similar examples for pattern analysis, and the Verify step reconciles predictions to produce the final classification of A (Diagnosis) and E (Healthy Lifestyle).

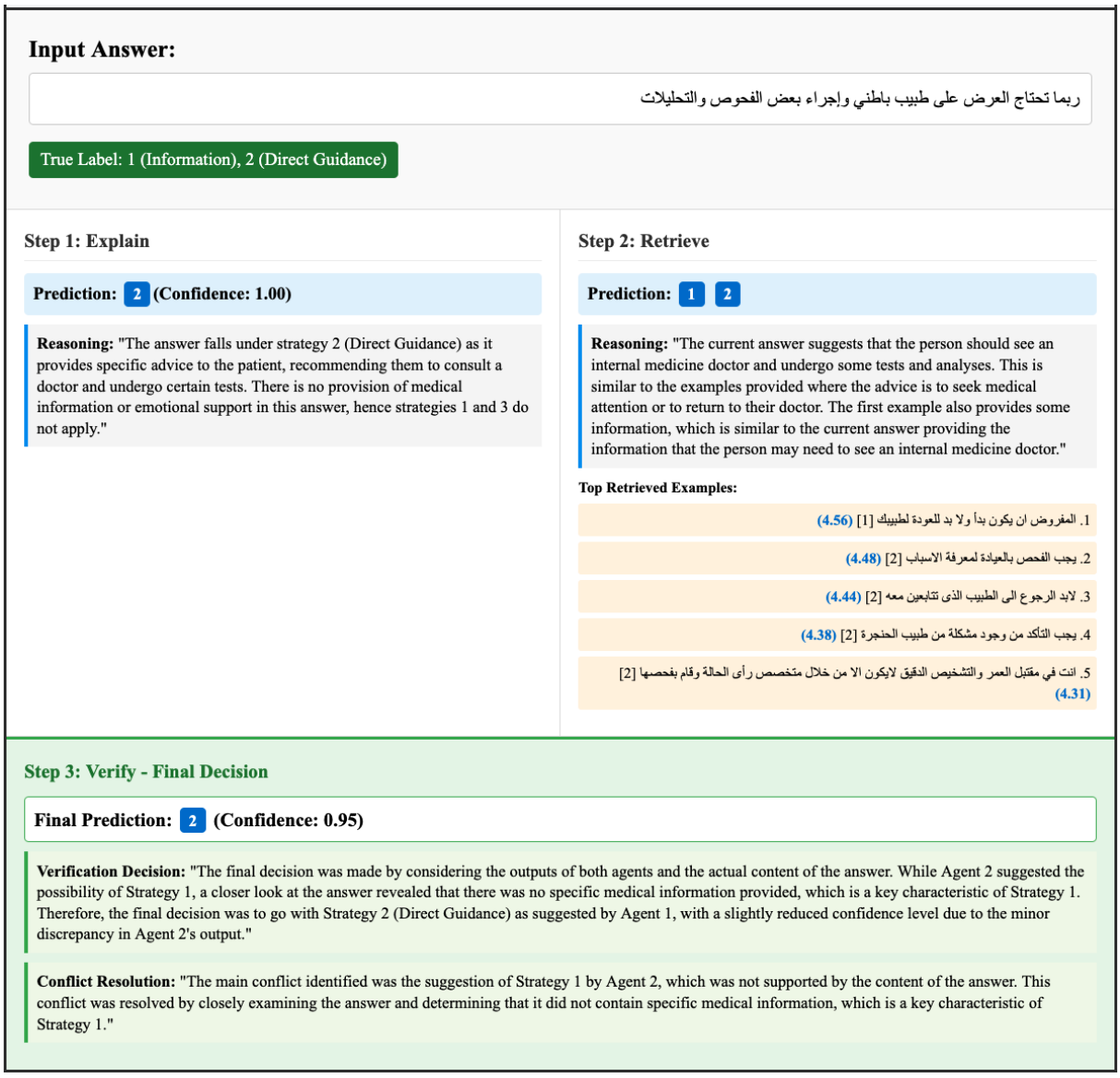

Figure 2: ERV workflow for answer strategy classification. The Explain step identifies direct guidance, the Retrieve step finds mixed evidence from similar examples, and the Verify step resolves the conflict by analyzing the absence of specific medical information, correctly classifying the answer as strategy 2 (Direct Guidance) while missing the 1 (Information) Strategy.