From RAG to Agentic: Validating Islamic-Medicine Responses with LLM Agents

Mohammad Amaan Sayeed [*1] **Mohammed Talha Alam** [*1] **Raza Imam** [*1] **Shahab Saquib Sohail** [2]
Amir Hussain [3]

Abstract

Centuries-old Islamic medical texts like Avicenna's Canon of Medicine and the Prophetic Tibb-e-Nabawi, encode a wealth of preventive care, nutrition, and holistic therapies, yet remain inaccessible to many and underutilized in modern AI systems. Existing language-model benchmarks focus narrowly on factual recall or user preference, leaving a gap in validating culturally grounded medical guidance at scale. We propose a unified evaluation pipeline, Tibbe-AG, that aligns 30 carefully curated Prophetic-medicine questions with human-verified remedies and compares three LLMs (LLaMA-3, Mistral-7B, Qwen2-7B) under three configurations: direct generation, retrieval-augmented generation, and a *scientific* self-critique filter. Each answer is then assessed by a secondary LLM serving as an agentic judge, yielding a single 3C3H quality score. Retrieval improves factual accuracy by 13%, while the agentic prompt adds another 10% improvement through deeper mechanistic insight and safety considerations. Our results demonstrate that blending classical Islamic texts with retrieval and self-evaluation enables reliable, culturally sensitive medical question-answering.

1. Introduction

Islamic literatures have contributed significantly to medical science. Classic texts written in the 11$^{\text{th}}$ century, such as The Canon of Medicine by Avicenna (Avicenna, 2005) and Prophetic Tibb-e-Nabawi (Saeed & Grapes, 2015; Junaid & Ali, 2019; Mohammad, 1983), provide detailed guidelines for preventive care, balanced lifestyle, and holistic health. The Canon of Medicine established systematic observation, diagnosis, and treatment protocols that influenced both Eastern and Western practices, while Tibb-e-Nabawi preserves prophetic wisdom by emphasizing hygiene, nutrition, and natural remedies. Historically confined to manuscripts, these sources continue to shape modern healthcare by promoting ethical and patient-centered care.

Research Gap: LLM-based medical agents, grounded in Tibb-e-Nabawi and The Canon of Medicine, can serve as interactive tools for educating students and scholars of Unani medicine. By integrating these sources into modern language models, it is possible to achieve enhanced contextual understanding that combines clinical data with centuries of ethical and holistic medical wisdom. Such agents can also generate diagnostic recommendations and therapeutic guidance that are both scientifically robust and culturally sensitive, addressing limitations observed in current Unani practices (Mehdi et al., 2022). Ultimately, if scaled effectively, these agents could deliver accessible, personalized, and ethically grounded healthcare solutions that benefit diverse populations from regions like Indian-subcontinent where a huge portion of people rely on Unani medicine (Pernau, 2012), while honoring a rich cultural heritage.

Related Work: Computational medicine integrated with age-old medical insights can enhance diagnostic precision and therapeutic outcomes (Esteva et al., 2017).For example, combining traditional knowledge with modern diagnostics improves treatment accuracy (Zhang et al., 2020), and deep learning can extract and operationalize data from traditional remedies (Zhang et al., 2023). Furthermore, (Chammas et al., 2022; Alrehali et al., 2020) integrated Arabic manuscripts points and learning models toward semantic extraction of Islamic texts, bridging the gap between historical wisdom and contemporary computational methods.

Contribution: Overall, we introduce Tibbe-AG, an agentic RAG framework that blends dense retrieval from classical Islamic medical texts with an explicit self-critique prompt to the same base LLM, yielding scientifically validated, culturally grounded healthcare guidance. We curate a focused benchmark of 30 Prophetic-medicine QA pairs drawn from *Tibb-e-Nabawi* (Shamsi, 2016) and (Mufti A.H.Elias, 2012), and used 3C3H (Hugging Face, 2024) with secondary LLMs as judge. Experiments across three base LLMs demonstrate that Tibbe-AG significantly outperforms both direct inference and standard RAG baselines.

[*]Equal contribution [1]Mohamed bin Zayed University of Artificial Intelligence, UAE [2]VIT Bhopal University, India [3]Edinburgh Napier University, UK. Correspondence to: Mohammed Talha Alam <mohammed.alam@mbzuai.ac.ae>.

Published at the 4th Muslims in Machine Learning (MusIML) Workshop (ICML-25), Vancouver, Canada, July 2025

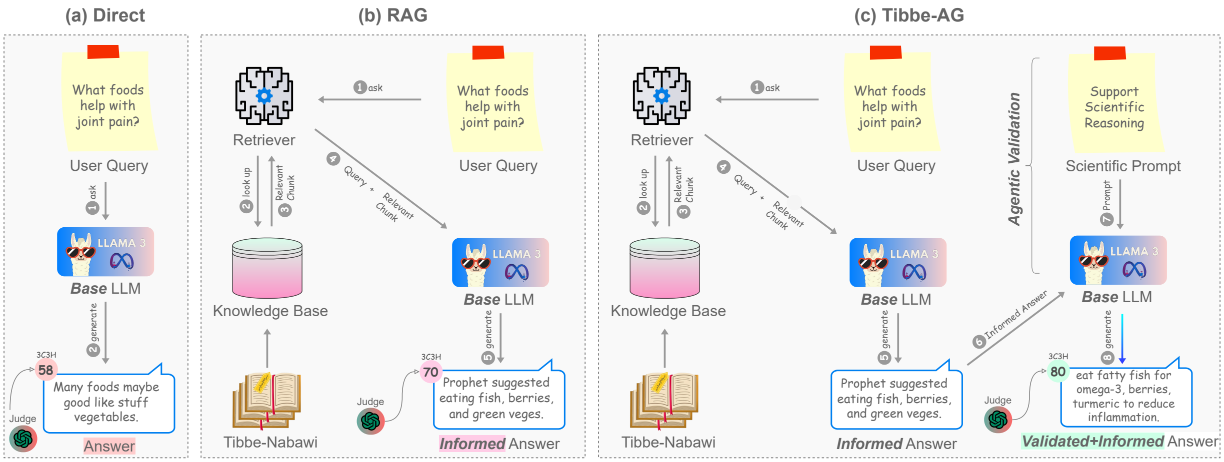

Figure 1. **Overview of TibbeAG:** (a) In the Direct inference, the base LLM generates an answer from the user query alone, often producing unvalidated or hallucinated content. (b) The RAG setting augments the prompt with top-k passages from the Tibb-e-Nabawi corpus to ground the response, yet still lacks an explicit mechanism to verify factual consistency or safety. (c) TibbeAG combines dense retrieval with an additional self-critique prompt to the same base LLM, yielding a final answer that is scientifically validated.

2. Methodology

We propose a question–answering pipeline, Tibbe-AG, grounded in classical Islamic medical knowledge such as *Tibb-e-Nabawi*. Given a health-related question q (for example, `What foods help with joint pain?`), our framework first retrieves relevant passages from a curated knowledge base and then employs a two-stage LLM process to generate and validate a final, high-quality answer.

Retrieval: To ensure that generated answers remain firmly rooted in authentic Prophetic-medicine teachings, we employ a dense retriever i.e., ChromaDB, that converts both query and corpus passages into a shared embedding space via a Transformer-based encoder. Formally, let

$$R(q) = \{d_1, d_2, \dots, d_k\}, \quad d_i \in K, \tag{1}$$

denote the top-k passages selected by cosine-similarity ranking. Each d_i is drawn from a preprocessed corpus K of classical texts, annotated with metadata (e.g., citation, context, severity). By thresholding similarity scores and filtering out redundant or low-information fragments, this step guarantees that the downstream LLM sees only the most germane and credible evidence, minimizing off-topic or factual drift.

Initial Answer (A_0): The first *Base* model, LLM_0, ingests the concatenated query q and retrieved context $R(q)$ to output answer A_0 as:

$$A_0 = \mathrm{LLM}_0(q,\ R(q)). \tag{2}$$

Under the hood, LLM_0 is prompted with a structured template that interleaves user question, document excerpts, and explicit ‘extract-and-summarize’ instructions. Internally, it attends over each passage’s provenance tokens to align outputs with source assertions. The resulting draft A_0 synthesizes actionable recommendations (e.g., `consume 1-2 teaspoons of raw honey daily`) while preserving direct textual traces, thus facilitating traceability and citation of classical sources.

Refinement (A_f): To guard against hallucinations and to enrich mechanistic and safety rationale, we append an explicit validation prompt q_{val} to the base LLM’s input. The final answer is then:

$$\begin{aligned} A_f &= \mathrm{LLM}_0(q,\ R(q),\ A_0,\ q_{\text{val}}) \\ &= \mathrm{LLM}_0(q,\ R(q),\ \mathrm{LLM}_0(q,\ R(q)),\ q_{\text{val}}). \end{aligned} \tag{3}$$

This agentic step performs three sub-tasks using validation prompt as it directs the model to (i) fact-check each A_0 segment against $R(q)$, (ii) inject mechanistic context (e.g. ginger’s effect on COX-2 pathways), and (iii) filter or flag unsafe recommendations (e.g. drug–herb interactions). By prompting the same LLM to re-evaluate its draft against retrieved evidence and explicitly apply mechanistic and safety checks, the additional validation step curbs hallucinations and ensures more reliable, evidence-anchored, and scientifically coherent response.

Evaluation (3C3H): To assess answer quality over a test set of n samples, we compute the aggregated 3C3H score as

$$\text{3C3H} = \frac{1}{6n}\sum_{i=1}^{n} c_{1i}\left(1+c_{2i}+\frac{c_{3i}-1}{4}+\frac{h_{1i}-1}{4}+\frac{h_{2i}-1}{4}+\frac{h_{3i}-1}{4}\right), \tag{4}$$

where c_{1i} and c_{2i} are the Correctness and Completeness scores for sample i, c_{3i} is Conciseness, h_{1i}, h_{2i}, and h_{3i} are Helpfulness, Harmlessness, and Honesty, each component lies in $[0, 1]$ (El Filali et al., 2024). Each criterion is evaluated by the judge model (e.g., GPT4.5 or Gemini) itself, based on the consistency checks for each answer generated by base LLM.

3. Dataset and Setup

3.1. Dataset

Validating Islamic-medicine responses via LLMs demands a specialized benchmark that captures the unique terminology and safety considerations of Prophetic treatments. To this end, we focused on a carefully curated question-answering dataset of *30 Prophetic-medicine questions*, to assess our experimental setup, drawn from two classical sources:

1 *Cures from the Qur'aan and Rasulullaah* as presented in *Madrasah in Just 5 Minutes* (Mufti A.H.Elias, 2012),

2 *Tibb-e-Nabawi* (Medical Guidance & Teachings of Prophet Muhammed) (Shamsi, 2016).

Specifically, our curation process proceeded in three steps:

A **Section Extraction:** From the first source, we extracted the entire section titled *"(9) Cures from the Qur'aan and Rasulullaah"* using a PyMuPDF-based script, isolating all remedy descriptions and their original citations. From the second source, we parsed the Prophetic-medicine chapters to compile an analogous list of remedies, ensuring comparable structure (remedy statement + reference).

B **Question Generation and Selection:** We converted each remedy description into a question of the form "`What Prophetic remedy is recommended for <ailment>?`", yielding an initial pool of approximately 120 candidate questions. We then *manually filtered* this pool to 30 questions, balancing across five broad categories, *nutritional therapies, herbal remedies, ritual supplications, hygiene practices,* and *wound treatments*, to ensure coverage of the full spectrum of Islamic-medicine teachings.

C **Representativeness & Feasibility:** Our 30 questions draw equally from Qur'ānic/Rasulullaah-based cures and the broader Tibb-e-Nabawi corpus, thereby spanning both spiritual and herbal dimensions, yet remain tractable for exhaustive evaluation (3 settings × 3 models × 3 judges). Each prompt carries its exact source (e.g. Sūrah, ḥadīth collection, or Tibb-e-Nabawi chapter and verse), enabling precise lookup and citation during retrieval and self-critique.

3.2. Evaluation Setup

To empirically validate the superiority of our agentic Tibbe-AG framework, we compare *three inference settings* on the same test questions q from our dataset as:

E1 In the **Direct** setting[1], the base model receives only the question q and produces an answer $A_f = \mathrm{LLM}_0(q)$ without any external grounding.

[1]We use '*Direct*' throughout to denote direct inference, where model receives only the user query and directly outputs an answer.

Table 1. Response quality across Direct, RAG, and Tibbe-AG on four key criteria. Only Tibbe-AG meets all, demonstrating both evidence grounding and safety.

Response Generation →	**Direct**	**RAG**	**Tibbe-AG**
Cites authentic sources	✗	✓	✓
Provides actionable specifics	✗	✓	✓
Includes scientific validation	✗	✗	✓
Includes clinical safety cues	✗	✗	✓

E2 The **RAG** setting augments the prompt with top-k passages retrieved from the Tibb-e-Nabawi corpus, yielding $A_f = \mathrm{LLM}_0(q,\ R(q))$.

E3 Finally, our **Tibbe-AG** (Agentic) pipeline applies the same retrieval step as E2 and but additionally prompts A_0 with q_{val} to refine the draft answer as:
$A_f = \mathrm{LLM}_0(q,\ R(q),\ A_0,\ q_{\mathrm{val}})$. Detailed in Sec. 2.

This controlled comparison among three settings is essential: by isolating the effects of plain inference, retrieval grounding, and agentic self-critique, we can precisely quantify each component's contribution under the 3C3H metric. Only through this three-way ablation can we demonstrate that retrieval alone improves factual grounding, and that the subsequent judge step yields statistically significant gains in Completeness, Harmlessness, and overall answer reliability.

3.3. Base and Judge Models

For the base-model component of each pipeline, we conduct experiments using three state-of-the-art 7B-parameter LLMs: **Mistral-7B** (Jiang et al., 2023),**Qwen-7B** (Bai et al., 2023), and **LLaMA-3-7B** (Grattafiori et al., 2024). In our primary evaluation, the judge model is instantiated as GPT4.5 (OpenAI, 2024), which critiques and refines every answer generated by the Direct, RAG, and Tibbe-AG settings. To verify that our findings are not idiosyncratic to a single judge, we also perform ablation studies replacing **GPT-4.5** with 3 alternative evaluators, **O4-mini-hi**, **Claude-4** (Anthropic, 2024) and **Gemini** (Google, 2025), while keeping all other components fixed. This multi-judge protocol demonstrates that Tibbe-AG's improvements in 3C3H scores remain consistent across different evaluation models, reinforcing the robustness of our agentic framework.

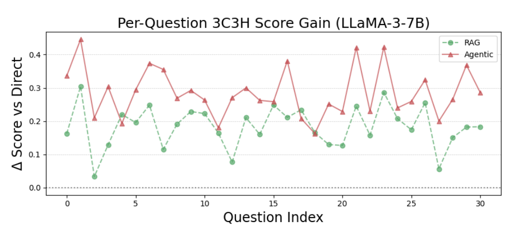

Figure 2. Comparison of *per-sample 3C3H score* gain among: Direct, RAG, and Agentic. Each point represents a unique query, highlighting performance improvements with *Tibbe-AG* (Agentic).

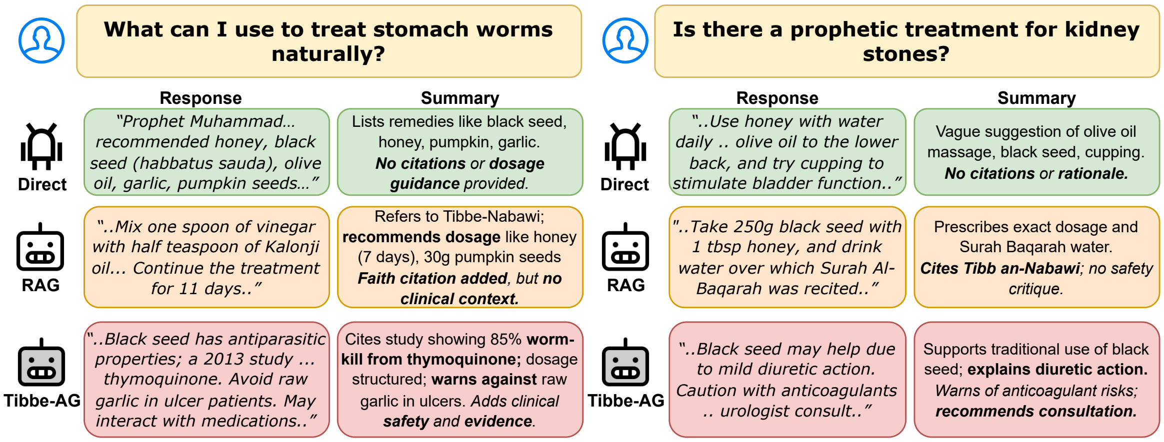

Figure 3. **Qualitative comparison of representative response excerpts across three inference settings.** Direct responses list remedies without justification; RAG adds faith-based specificity and references; Tibbe-AG integrates scientific critique, safety reasoning, and actionable guidance.

Table 2. Average 3C3H scores across inference settings. Tibbe-AG demonstrates consistent improvements in both faith-based and scientific dimensions. **Mean** is the average of four Judge LLMs.

Base Model ↓	Method ↓	GPT4.5	O4-mini	Claude-4	Gemini	Mean
Qwen	Direct	0.44	0.48	0.43	0.44	0.45
	RAG	0.62	0.62	0.75	0.76	0.69
	Tibbe-AG	**0.73**	**0.74**	**0.85**	**0.86**	**0.80**
Mistral	Direct	0.48	0.53	0.45	0.46	0.48
	RAG	0.65	0.66	0.76	0.77	0.71
	Tibbe-AG	**0.76**	**0.77**	**0.86**	**0.87**	**0.82**
LLaMA-3	Direct	0.49	0.58	0.47	0.48	0.50
	RAG	0.67	0.70	0.78	0.79	0.73
	Tibbe-AG	**0.77**	**0.79**	**0.88**	**0.89**	**0.83**

4. Results and Discussion

Quantitative Results: As shown in Table 2, Tibbe-AG consistently achieved higher average 3C3H scores across all tested base models (LLaMA-3, Mistral-7B, Qwen2-7B) when compared to both Direct inference and standard RAG approaches. For instance, with LLaMA-3 as the base model and GPT4.5 as the judge, Tibbe-AG scored 0.77, surpassing Direct (0.49) and RAG (0.67) by significant margins. This trend of improvement is robust across different judge models, as the mean results are consistently higher with our approach, indicating that the benefits of Tibbe-AG are not tied to a specific evaluator, reinforcing generalizability.

Qualitative Results: Direct responses often lack grounding and specifics, while RAG, though an improvement, typically falls short on scientific validation and safety considerations. Figure 3 provides examples of these distinctions. For instance, when asked about treating stomach worms, the Direct response is generic. The RAG response offers more specific remedies from Tibb-e-Nabawi but lacks clinical context or safety warnings. Tibbe-AG, however, not only suggests remedies like black seed but also cites a relevant study on thymoquinone's efficacy, provides dosage considerations, and crucially, warns about contraindications (e.g., raw garlic in ulcer patients) and potential medication interactions. Similarly, for kidney stones, Tibbe-AG explains the potential mechanism (diuretic action of black seed), warns about anticoagulant risks, and advises consulting a urologist, demonstrating a more comprehensive and responsible approach.

Discussion: The enhanced performance of Tibbe-AG can be attributed to its two-stage process. First, the retrieval step grounds the LLM in the relevant classical Islamic medical texts, which helps keep the information accurate and true to the source. This directly combats the tendency for LLMs to "hallucinate" when generating answers directly. Then comes the crucial agentic self-critique. By prompting the LLM to review its initial answer against the retrieved evidence and to actively think about mechanistic details and safety, Tibbe-AG polishes the response into something more coherent, scientifically plausible, and safe. This design directly addresses a key shortcoming in many existing systems, which often don't adequately validate medical guidance that's grounded in specific cultural contexts.

5. Conclusion

Our work successfully introduces Tibbe-AG, a novel framework that significantly enhances the generation of reliable and culturally sensitive Islamic medical guidance using LLMs. By effectively blending classical texts with retrieval and agentic self-critique, Tibbe-AG paves the way for AI systems that make the rich heritage of Islamic medicine more accessible and applicable in modern contexts. While our initial 30-question dataset provided a strong foundation, future efforts will focus on expanding this dataset, further refining Tibbe-AG's agentic capabilities, and conducting user studies. These steps will be crucial in evolving Tibbe-AG into an even more robust tool, contributing to a more inclusive and culturally competent approach in healthcare.

References

Alrehali, B., Alsaedi, N., Alahmadi, H., and Abid, N. Historical arabic manuscripts text recognition using convolutional neural network. In *2020 6th conference on data science and machine learning applications (CDMA)*, pp. 37–42. IEEE, 2020.

Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf. Accessed: May 20, 2025.

Avicenna. *The Canon of Medicine*. Kazi Publications, 2005.

Bai, J., Bai, S., Chu, Y., Cui, Z., et al. Qwen technical report, 2023. URL https://arxiv.org/abs/2309.16609.

Chammas, M., Makhoul, A., Demerjian, J., and Dannaoui, E. A deep learning based system for writer identification in handwritten arabic historical manuscripts. *Multimedia Tools and Applications*, 81(21):30769–30784, 2022.

El Filali, A., Sengupta, N., Abouelseoud, A., Nakov, P., Fourrier, C., AI, I., and MBZUAI. Rethinking LLM evaluation with 3c3h: Aragen benchmark and leaderboard, December 2024. URL https://huggingface.co/blog/leaderboard-3c3h-aragen. Accessed: 2025-05-30.

Esteva, A., Kuprel, B., Novoa, R., Ko, J., Swetter, S., Blau, H., and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542 (7639):115–118, 2017.

Google. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

Grattafiori, A., Dubey, A., Jauhri, A., et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Hugging Face. 3c3h: Arabic reasoning and faith-aligned evaluation leaderboard, 2024. URL https://huggingface.co/blog/leaderboard-3c3h-aragen. Accessed: May 14, 2025.

Jiang, A. Q., Sablayrolles, A., Mensch, A., et al. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Junaid, M. and Ali, S. Analysis of historical islamic medical manuscripts. In *Proceedings of the International Conference on Islamic Sciences*, pp. 145–150, 2019.

Mehdi, S., Sultana, A., Heyat, M. B. B., Chola, C., Akhtar, F., Gutema, H. K., Al-qadasi, D. M., and Baig, A. A. A review of amenorrhea toward unani to modern system with emerging technology: current advancements, research gap, and future direction. *Computational Intelligence in Healthcare Applications*, pp. 121–135, 2022.

Mohammad, A. Tibb-e-nabawi: Medical teachings in prophetic tradition. *Islamic Journal of Medical History*, 1:34–42, 1983.

Mufti A.H.Elias. *5-Minute Madrasa in English*. 2012. URL https://www.islamicstudiesresources.com/uploads/1/9/8/1/19819855/5-minute-madrasa-in-english.pdf. Accessed: 2025-05-14.

OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Pernau, M. The indian body and unani medicine: Body history as entangled history. In *Images of the Body in India*, pp. 97–108. Routledge India, 2012.

Saeed, A. and Grapes, M. Prophetic guidance on health: An analysis of tibb-e-nabawi. *Journal of Islamic Medical Studies*, 4(2):98–105, 2015.

Shamsi, D. M. S. *Tibb-e-Nabawi: Medical Guidance & Teachings of Prophet Muḥammad*. 2016. URL https://ia600507.us.archive.org/29/items/TibENabi1/Tib-e-nabi-%20(1).pdf. Accessed: May 14, 2025.

Zhang, H., Ni, W., Li, J., Zhang, J., et al. Artificial intelligence–based traditional chinese medicine assistive diagnostic system: validation study. *JMIR medical informatics*, 8(6):e17608, 2020.

Zhang, S., Wang, W., Pi, X., He, Z., and Liu, H. Advances in the application of traditional chinese medicine using artificial intelligence: a review. *The American journal of Chinese medicine*, 51(05):1067–1083, 2023.