# TRUMEDIQA: A Modular Trustworthy RAG Pipeline for Multilingual Medical Question Answering

**Meryem EL Fatimi**[1], **Ayoub Nainia**[2], **Jihad Zahir**[3]
[1] TICLab, International University of Rabat, Rabat, Morocco
[2] Institut de Systématique, Évolution, Biodiversité (ISYEB), Sorbonne Université, Paris, France
[3] UMMISCO, IRD, Computer Science Department, LISI, Cadi Ayyad University,
Faculty of Sciences, Marrakech, Morocco
meryem.elfatimi@uir.ac.ma, ayoub.nainia@sorbonne-universite.fr, j.zahir@uca.ac.ma

**Abstract**

Medical question answering systems must balance usefulness with safety, particularly in low-resource linguistic settings where robustness is limited and hallucinations can cause harm. We present TRUMEDIQA, a reproducible multilingual medical QA pipeline for Moroccan Darija, Arabic, French, and English, deployed on WhatsApp with text and voice interactions. TRUMEDIQA uses layered decision-making: (i) language identification, (ii) a pre-retrieval intent router that maps queries to one of 38 clinical FAQ categories to constrain retrieval, and (iii) post-retrieval LLM-based re-ranking that selects the best candidate answer or returns a null decision to trigger a safe fallback (abstention). Answers are retrieved from a curated FAQ knowledge base validated by medical professionals. We evaluate TRUMEDIQA with 21 participants submitting 290 questions across four languages. An expert annotator labels each interaction as relevant, acceptable, or irrelevant, and we also measure correct abstentions when no suitable answer exists in the knowledge base. An ablation study shows that routing and re-ranking improve the weighted relevance score from 0.25 to 0.94 and precision from 0.53 to 0.98 versus a naïve retrieval baseline, while increasing correct abstention on unanswerable queries from 4.38% to 69.77%.

**Keywords:** clinical question answering, multilingual NLP, retrieval-augmented generation

## 1. Introduction

Large Language Models (LLMs) have recently improved the fluency and apparent competence of medical conversational systems, creating opportunities to expand access to health information at scale. However, medical QA remains high-risk: models can produce plausible but incorrect statements, and this risk is amplified in low-resource linguistic settings where robustness is limited. In the Arabic context, recent benchmarks show that even Arabic-centric LLMs perform poorly on dialectal input compared to Modern Standard Arabic (MSA) (Mousi et al., 2025). Work targeting Moroccan Arabic (Darija) similarly reports that substantial adaptation is required to handle dialectal variation reliably (Qarah and Alsanoosy, 2025). These gaps can translate into unequal access to safe medical information and motivate clinical NLP systems that are both multilingual and cautious under uncertainty.

In this work, we study patient-facing medical question answering for adolescent reproductive health in a multilingual setting (Moroccan Darija, Arabic, French, and English). We introduce TRUMEDIQA, a reproducible WhatsApp-based QA system that combines retrieval with layered decision-making to reduce unsafe mismatches and to abstain when the knowledge base lacks a suitable answer. TRUMEDIQA retrieves answers from a curated FAQ knowledge base validated by medical professionals, and applies (i) language identification, (ii) a pre-retrieval intent router over 38 clinical FAQ categories to constrain retrieval, and (iii) post-retrieval LLM-based re-ranking that selects the best candidate answer or returns a null decision to trigger a safe fallback. This design aims to improve reliability without overclaiming coverage, which is critical in medical settings. TRUMEDIQA is intended for informational support rather than diagnosis or treatment advice. While individual components such as intent routing, constrained retrieval, and abstention are established patterns in RAG systems, their integration in a real-world multilingual deployment targeting a low-resource dialect (like Moroccan Darija) for a socio-culturally sensitive health domain represents a novel and underexplored setting that existing systems do not address.

This paper makes the following contributions:

- We present TRUMEDIQA, a layered multilingual medical QA system supporting Moroccan Darija, Arabic, French, and English, with both text and voice interactions on WhatsApp.
- We propose a modular decision pipeline combining intent-based query routing and post-retrieval LLM re-ranking with explicit abstention to reduce unsafe mismatches.
- We report a real-world evaluation (21 participants, 290 questions) with expert annotation, and an ablation study quantifying the impact of



*Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026*, pages 49–56
16 May 2026. 

routing and re-ranking on relevance, precision, and correct abstentions.

- We provide system details and a reproducible evaluation protocol (taxonomy, prompts, configuration), and describe our privacy/anonymization handling to facilitate replication in other low-resource health settings.

- We demonstrate that a safety-oriented layered RAG architecture can be effectively deployed in a low-resource dialectal setting, addressing a gap in multilingual clinical NLP where Moroccan Darija and similar dialects remain severely underserved.

The remainder of this paper is structured as follows: Section 2 reviews related work on medical QA and clinical NLP. Section 3 describes the trustworthiness dimensions that guide our design. Section 4 presents the TRUMEDIQA pipeline and knowledge base construction. Section **??** reports the experimental setup, and Sections 6 and 7 define the evaluation framework and metrics. Results are reported in Section 8, followed by discussion in Section 9 and conclusion in Section 10. Finally, we provide an ethics statement and limitations in Sections 11 and 12

## 2. Background and Related Work

### 2.1. Patient-facing medical QA and multilingual gaps

Medical question answering (MQA) aims to return reliable health information in response to natural-language queries. Prior to LLMs, many MQA systems relied on modular pipelines that combined question classification, retrieval over curated resources, and answer selection, often leveraging controlled terminologies such as UMLS and SNOMED CT for normalization and interpretability (Mutabazi et al., 2021; Zhu et al., 2018). While modern LLMs have improved fluency and coverage, reliability and safety remain central concerns in patient-facing settings.

A major limitation is linguistic imbalance: most benchmarks and resources are English-centric, which yields systematic performance gaps for other languages (Hamad et al., 2024; Baethge, 2008; Valkimadi et al., 2009). Multilingual medical QA benchmarks such as MedQA (Jin et al., 2020), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), and MultiMedQA (Singhal et al., 2025) show strong results on structured exam-style questions, yet performance degrades on open-ended queries and in linguistically diverse settings. Recent evaluations (e.g., MedExpQA) further report persistent disparities between English and non-English answers, even under translation (Alonso et al., 2024). These findings motivate multilingual systems that explicitly account for low-resource languages and dialectal variation.

### 2.2. Grounded LLM-based medical QA and retrieval pipelines

To mitigate hallucinations and improve factuality, a common strategy is to ground generation in external knowledge via retrieval-augmented generation (RAG) (Lewis et al., 2020). In medical QA, RAG typically retrieves from trusted sources (e.g., guidelines, curated FAQs, biomedical literature) and then generates or selects answers conditioned on retrieved evidence. Recent work explores retrieval and ranking improvements, including domain-specific retrievers and medically oriented RAG systems (Zhao et al., 2025; Gao et al., 2025). TRUMEDIQA follows this line but emphasizes a layered decision pipeline that combines intent-based routing and post-retrieval re-ranking to reduce mismatches in multilingual user queries.

### 2.3. Safety under uncertainty: calibration, answerability, and abstention

Even when grounded, LLMs can be overconfident and may produce misleading outputs (Tian et al., 2023; Nazi and Peng, 2024). Calibration and uncertainty estimation methods (e.g., temperature scaling, ensembles, and self-assessment prompts) aim to better align model confidence with correctness (Desai and Durrett, 2020; Balabanov and Linander, 2025; Xiong et al., 2024). In high-stakes medical contexts, this motivates *selective answering*: abstaining when evidence is insufficient or when a query is outside the supported scope. Our design operationalizes this principle through explicit abstention triggered by routing and re-ranking decisions, enabling the system to defer when the FAQ knowledge base does not contain an appropriate answer.

## 3. Trustworthiness by design

We define trustworthiness for multilingual medical QA as a set of system properties that can be operationalized and audited in the pipeline. In TRUMEDIQA, trustworthiness relies on four dimensions: (1) accuracy and safety, (2) explainability, (3) compliance with local socio-cultural and legal constraints, and (4) privacy. Table 1 summarizes how each dimension is instantiated in the system and how it is assessed in our evaluation.

The remainder of this paper focuses on the system components that instantiate these dimensions and on a real-world evaluation quantifying the trade-off between answer quality and responsible abstention.

## 4. Medical Knowledge Base Construction

TRUMEDIQA retrieves answers from a curated FAQ knowledge base covering adolescent reproductive health. The knowledge base is designed for patient-facing informational support and is intentionally scoped to avoid diagnosis, treatment recommendations, or urgent-care guidance; out-of-scope or emergency-like queries trigger abstention and a redirect message (Section 3).

**FAQ authoring and clinical validation.** We first compile a set of recurring user questions for the target domain, initially authored in Moroccan Darija. For each question, a medical professional authored and/or validated the corresponding answer to ensure local appropriateness and consistency with the intended scope. Answers may include optional media pointers (e.g., links or simple visual aids) when useful for comprehension.

**Multilingual collections.** To support multilingual access, we build separate FAQ collections for Moroccan Darija, Arabic, French, and English. Each collection contains language-specific question formulations paired with clinician-validated answers written for that language. This design avoids relying on cross-lingual retrieval as a default and enables language-aware routing and evaluation within each collection.

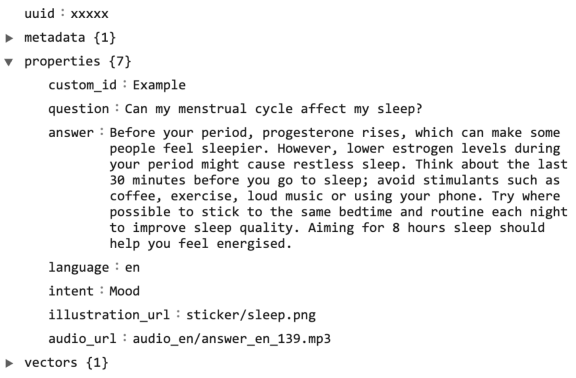


Figure 1: TRUMEDIQA Knowledge Base Entry.

**Intent taxonomy.** Each FAQ is manually assigned an intent label from a fixed taxonomy of 38 clinical FAQ categories. These labels provide an interpretable organization of the knowledge base and enable intent-based query routing that constrains retrieval to category-consistent candidates.

**Vector indexing and stored schema.** We store each language collection in Weaviate as a dedicated vector index. Each entry contains the question text, a clinician-validated answer, a language identifier, an intent category from our 38-category taxonomy, and optional media fields (e.g., `audio_url`, `sticker_url`; Figure 1). Within a given language, we embed question texts into dense vectors for similarity search and retrieve top-$k$ candidates at inference time, which are then

| Dimension | Operationalization in TRUMEDIQA | Evidence / assessment |
|---|---|---|
| Accuracy & safety | Clinician-validated FAQ knowledge base; intent-based routing to restrict retrieval space; post-retrieval re-ranking; explicit abstention when no suitable answer is found; safety-oriented fallback messages. | Expert annotation (relevant/acceptable/irrelevant); Precision and weighted Relevance Score; abstention analysis (Fallback Rate and correct abstentions). |
| Traceability (grounding) | Responses are linked to a retrieved FAQ entry; the system can surface the matched FAQ identifier and/or a supporting snippet; explicit abstention for uncovered queries. | Grounding coverage (% responses linked to an FAQ ID/snippet) and qualitative inspection; abstention correctness captured by correct-abstention metrics. |
| Local compliance | Domain scoping to adolescent reproductive health; standardized disclaimers and response templates; constrained answer set to avoid unavailable/inappropriate recommendations. | Qualitative analysis of compliance-related cases; failure modes reported in limitations. |
| Privacy | Anonymized logging; data minimization; restricted access to interaction logs; raw voice messages not retained beyond processing in our deployment. | Description of consent/anonymization and data handling; deployment constraints and limitations. |

Table 1: Trustworthiness dimensions and their operationalization in TRUMEDIQA.

passed to a post-retrieval re-ranking stage.

**Re-ranking module.** Retrieved top-$k$ candidates are passed to an LLM-based re-ranker for final selection. The re-ranker uses GPT-4o and receives a structured prompt containing the user question, the detected intent category, and the list of candidate question–answer pairs indexed from 0 to $k-1$. The system prompt instructs the model to act as an answer ranker rather than a generator: it must return the integer index of the most semantically relevant candidate, or `null` if no candidate adequately addresses the question. The model is constrained to respond with a single token only — an integer or the word `null`. A `null` response triggers the safe fallback mechanism described in Section 3. Temperature is set to 0 to ensure deterministic and reproducible ranking decisions.

**Voice responses.** To reduce latency for voice interactions, we pre-generate an audio rendition for each validated answer using a text-to-speech engine and store it as an `audio_url` field. When an answer is selected, TRUMEDIQA can return the text response and, when applicable, the corresponding audio response.

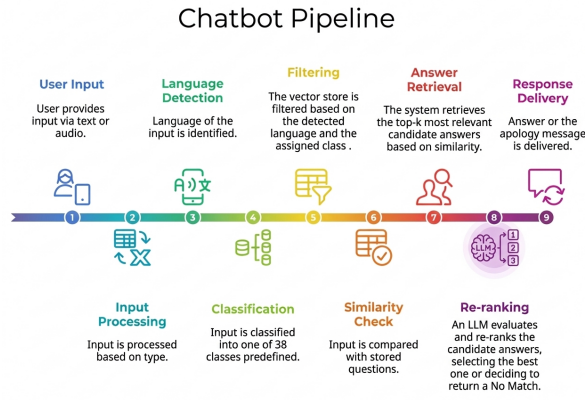


Figure 2: Overview of the TRUMEDIQA chatbot pipeline, from user input to response delivery. The pipeline combines language detection, intent-based classification, filtered retrieval, LLM re-ranking, and safe fallback decision process.

## 5. Experimental Setup

We evaluated TRUMEDIQA with 21 volunteers, who submitted 290 questions during a four-hour testing window via WhatsApp. Participants could ask questions in Arabic, Moroccan Darija, French, or English, and could freely alternate between languages. The language distribution reflects the target population: 220 questions were submitted in Arabic or Moroccan Darija — which participants used interchangeably and represent the primary language of our deployment context — 37 in English, 30 in French, and 3 involved code-switching across languages. This predominance of Arabic/Darija is consistent with our motivation: Moroccan users naturally favor Darija, a low-resource dialect, making it both the most representative and the most challenging language in our evaluation. Inputs were provided either as text messages or as voice messages (up to 60 seconds); voice messages were transcribed using Google Cloud Speech-to-Text before being passed to the pipeline. All interactions were anonymized for analysis and collected with participant consent; the system is intended for informational support rather than diagnosis or treatment advice.

We conducted an ablation study to quantify the contribution of two key components: intent routing (38-category taxonomy) and post-retrieval re-ranking. We compared three configurations (Table 3) on the same set of user questions. C1 is a naive retrieval baseline without routing or re-ranking. C2 adds intent routing but removes re-ranking, isolating the effect of routing. C3 corresponds to the complete system.

## 6. Evaluation Framework

All interactions were logged and anonymized to protect participant privacy. After the four-hour testing window, an expert evaluator manually annotated each question-answer (QA) pair. The evaluation focuses on whether the system took the *correct action* given the knowledge base coverage: answering when an entry exists and abstaining when it does not. Each interaction was assigned to one of three outcome classes:

- **Relevant:** The system took the correct action. This includes:
  - **Correct answer (answerable):** a direct,

| Config. | System variant | What it tests |
|---|---|---|
| C1 | Direct vector retrieval (no routing, no re-ranking). | Baseline retrieval quality without filtering. |
| C2 | Intent routing (38 categories) + retrieval; no re-ranking. | Contribution of routing to candidate quality. |
| C3 | Full system: routing + retrieval + LLM re-ranking (+ abstention). | End-to-end performance of TRUMEDIQA. |

Table 2: Ablation configurations for evaluating routing and re-ranking.

correct, and contextually appropriate answer when a suitable FAQ entry exists.

  - **Correct abstention (unanswerable):** abstention via a fallback message when no suitable answer exists in the knowledge base.

- **Acceptable:** The response was topically relevant and potentially useful but partially correct, incomplete, or insufficiently specific.
- **Irrelevant:** The response exhibited a clear mismatch between the question and the returned answer, or the system abstained despite the existence of a suitable knowledge base entry (i.e., incorrect abstention).

## 7. Evaluation Metrics

To assess TRUMEDIQA, we report four complementary metrics: (i) a weighted Relevance Score summarizing expert quality labels, (ii) precision over answered queries to quantify the rate of misleading answers, (iii) fallback rate to characterize abstention frequency, and (iv) abstention correctness to distinguish appropriate abstention from unnecessary fallback. Together, these metrics capture both answer quality and the system's ability to manage uncertainty.

**Relevance Score.** We compute a weighted relevance score (RS) from expert labels by assigning weights 2/1/0 to *Relevant*/*Acceptable*/*Irrelevant* and normalizing to $[0, 1]$:

$$RS = \frac{1 \cdot N_{\text{acceptable}} + 2 \cdot N_{\text{relevant}}}{2 \cdot N_{\text{total}}}. \quad (1)$$

**Precision (answered queries).** We define an *answered* query as one where the system returns an explicit answer rather than a fallback message. A false positive (FP) is an answered query labeled *Irrelevant* by the expert. A true positive (TP) is an answered query labeled *Relevant* or *Acceptable*. Precision is:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (2)$$

**Fallback Rate.** Fallback Rate (FR) measures how often the system abstains:

$$\text{FR} = \frac{N_{\text{fallback}}}{N_{\text{total}}}. \quad (3)$$

**Abstention correctness.** To evaluate whether abstentions are appropriate given knowledge base coverage, the expert additionally labels each query as *Answerable* or *Unanswerable* with respect to the FAQ knowledge base. Correct abstentions (true negatives, TN) are fallback cases labeled *Unanswerable*:

$$TN = \sum_{i=1}^{N} \mathbf{1}(\hat{a}_i = \text{Fallback}) \cdot \mathbf{1}(a_i = \text{Unanswerable}), \quad (4)$$

where $\hat{a}_i$ is the system action (answer vs fallback) and $a_i$ is the expert answerability label. Conversely, incorrect abstentions (false negatives, FN) are fallback cases labeled *Answerable*:

$$FN = \sum_{i=1}^{N} \mathbf{1}(\hat{a}_i = \text{Fallback}) \cdot \mathbf{1}(a_i = \text{Answerable}). \quad (5)$$

We report TN as a rate by normalizing by $N_{\text{unanswerable}}$ and FN as a rate by normalizing $N_{\text{answerable}}$ (FNR)..

| Configuration | RS | Prec. | TN (%) | FR (%) |
|---|---|---|---|---|
| C1: no routing, no re-ranking | 0.25 | 0.53 | 4.38 | 55 |
| C2: routing only (no re-ranking) | 0.68 | 0.74 | 26.47 | 12 |
| C3: full system | **0.94** | **0.98** | **69.77** | 15 |

Table 3: Ablation results. TN and FR are reported as percentages. TN(%) denotes correct abstentions among unanswerable queries, computed as $TN/N_{\text{unanswerable}}$. FR(%) denotes the overall fallback rate, computed as $(TN + FP)/N$.

## 8. Results

Table 3 reports the ablation results for routing and re-ranking. The naive baseline (C1) yields low overall quality (RS=0.25) and moderate precision (0.53), with a high fallback rate (55%), indicating frequent abstention and limited utility in this setting. Adding intent routing (C2) substantially improves RS (0.68) and precision (0.74) while reducing fallback to 12%, suggesting that constraining retrieval to category-consistent candidates reduces mismatches. The full system (C3) achieves the best performance (RS=0.94, precision=0.98) and markedly increases correct abstention (TN; Section 7), indicating improved alignment with knowledge base coverage limits. Figure 3 shows the distribution of expert labels across configurations, where the number of irrelevant responses drops sharply from C1 to C3.

### 8.1. Error Analysis

To better understand the residual failures of the full system (C3), we manually analyzed all cases labeled Irrelevant or Acceptable by the expert annotator. We identify three failure types.

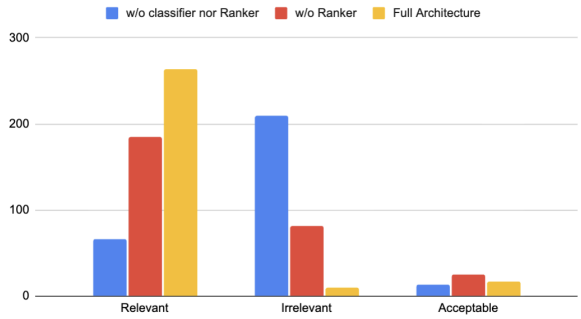


Figure 3: Expert evaluation of all configurations ( C1,C2 and C3)

**False abstention (14 cases, 52%).** The most frequent failure occurs when the ranker returns a null decision despite the existence of a suitable FAQ entry in the knowledge base. This typically arises when the retrieved candidates are semantically close but not close enough for the ranker to commit to a match, leading to an overly conservative abstention. This pattern was most prevalent in Darija queries, where the embedding model produces less reliable similarity scores due to the low-resource nature of the dialect.

**Wrong answer selected (8 cases, 30%).** In these cases the ranker returns an answer but selects the wrong FAQ entry even though the correct one exists in the knowledge base. This reflects retrieval ambiguity within a correctly identified intent category, where multiple candidates share surface similarity with the query.

**Incorrect answer when abstention was appropriate (5 cases, 19%).** In these cases the ranker returned an answer for a query that has no suitable match in the knowledge base, when a NO MATCH decision would have been correct. These cases involve questions that fall outside the scope of the knowledge base, where the ranker should have abstained but instead returned a partially related answer.

Overall, false abstentions account for the majority of C3 failures, suggesting that the ranker's conservatism — while beneficial for safety — occasionally withholds valid answers.

## 9. Discussion

The ablation results (Table 3) show a consistent improvement as routing and re-ranking are added (Figure 3). The weighted relevance score increases from $0.25$ in the naive baseline (C1) to $0.94$ in the full system (C3), and precision improves from $0.53$ to $0.98$. Together, these trends indicate that constraining retrieval to category-consistent candidates and re-ranking retrieved entries substantially reduces mismatches and misleading answers.

Beyond answer quality, the results also highlight the role of explicit abstention for safe behavior under limited knowledge base coverage. While the full system exhibits a modest fallback rate (15%), it achieves a substantially higher correct-abstention rate (TN; Section 7) than the baseline. This suggests that TRUMEDIQA more often withholds answers in cases where the knowledge base does not contain a suitable entry, rather than returning low-quality matches.

Overall, the comparison across configurations underscores the complementary value of intent routing and post-retrieval re-ranking in multilingual, low-resource medical QA. Routing improves candidate selection by narrowing the retrieval space, while re-ranking further improves contextual alignment and supports more reliable abstention decisions. These components jointly contribute to a better balance between usefulness and caution, which is critical in patient-facing medical settings.

## 10. Conclusion

We presented TRUMEDIQA, a reproducible multilingual medical QA pipeline for Moroccan Darija, Arabic, French, and English, deployed on WhatsApp with text and voice interactions. In a real-world evaluation (21 participants, 290 questions), an ablation study showed that combining intent routing with post-retrieval re-ranking substantially improves response quality compared to a naive retrieval baseline (RS: 0.25 to 0.94; precision: 0.53 to 0.98), while supporting responsible abstention when no suitable knowledge base entry exists.

Future work will focus on expanding coverage while preserving safety guarantees. One direction is to extend the curated FAQ knowledge base with additional clinician-reviewed content derived from medical documents, using human-in-the-loop validation for extracted passages. Another direction is to improve uncertainty handling by modeling answerability more explicitly and reducing unnecessary abstentions, as well as enabling controlled knowledge base updates with versioning and periodic clinical review.

## 11. Ethics Statement

TRUMEDIQA is a patient-facing informational QA system designed to provide general health education within a constrained scope (adolescent reproductive health). It is **not** intended for diagnosis, treatment recommendation, or emergency triage. To reduce the risk of harm, the system is designed to abstain when knowledge base coverage is insufficient and to present standardized fallback mes-

sages that encourage consulting qualified health professionals when appropriate.

The evaluation involved 21 volunteers interacting with the system via WhatsApp. All interactions were logged for research purposes and anonymized prior to analysis. We followed data minimization principles by restricting access to logs and avoiding the persistent storage of raw voice messages beyond processing. Because reproductive health topics can be sensitive, we emphasize that any deployment of similar systems should include clear user-facing disclosures, consent procedures, and appropriate safeguards for vulnerable users.

Finally, our work highlights equity considerations in clinical NLP: low-resource languages and dialects (e.g., Moroccan Darija) are under-served by current medical NLP systems, which can exacerbate unequal access to safe health information. Our goal is to contribute methods and evaluation practices that improve reliability for such settings.

## 12. Limitations

Our study has limitations. First, TRUMEDIQA relies on a curated FAQ knowledge base; therefore, system coverage is bounded by the scope and completeness of this resource. While abstention reduces unsafe outputs, unanswered questions may limit usefulness, and the system does not guarantee completeness even for in-scope topics.

Second, our evaluation is based on a single real-world testing session (21 participants, 290 questions) and a single expert annotator. Larger-scale studies, multiple annotators, and additional clinical expert review would strengthen reliability estimates and enable inter-annotator agreement analysis. Performance may also vary across languages and modalities (text vs. voice), and further stratified analysis is needed.

Third, the system uses LLM components for routing and re-ranking, which may exhibit instability across model versions and can introduce biases. In addition, parts of the stack may depend on proprietary services (e.g., messaging platform APIs, LLM endpoints, or TTS), which can affect reproducibility and long-term maintainability.

Finally, we evaluate informational QA rather than EHR-based clinical decision support. The system should not be used to replace professional medical care, and any real deployment would require careful governance, monitoring, and periodic clinical review of content and behavior.

## 13. Bibliographical References


Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial Intelligence in Medicine*, 155:102938.

Christopher Baethge. 2008. The languages of medicine. *Deutsches Arzteblatt International*, 105(3):37.

Oleksandr Balabanov and Hampus Linander. 2025. Uncertainty quantification in fine-tuned llms using lora ensembles.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers.

Yanjun Gao, Ruizhe Li, Emma Croxford, John Caskey, Brian W Patterson, Matthew Churpek, Timothy Miller, Dmitriy Dligach, and Majid Afshar. 2025. Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study. *JMIR AI*, 4:e58670.

Abdullah Ashraf Hamad, Jaber H Jaradat, Hamza K Alsalhi, and Ibraheem M Alkhawaldeh. 2024. Medical research production in native languages: A descriptive analysis of pubmed database. *Qatar Medical Journal*, 2024(1):21.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. Aradice: Benchmarks for dialectal and cultural capabilities in llms. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218.

Emmanuel Mutabazi, Jianjun Ni, Guangyi Tang, and Weidong Cao. 2021. A review on medical textual question answering systems based on deep learning approaches. *Applied Sciences*, 11(12).

Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering.

Faisal Qarah and Tawfeeq Alsanoosy. 2025. Evaluation of arabic large language models on moroccan dialect. *Engineering, Technology & Applied Science Research*, 15(3):22478–22485.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, Rezarta Islamaj, Aadit Kapoor, Xin Gao, and Zhiyong Lu. 2023. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1).

Politimi Valkimadi, Drosos Karageorgopoulos, Harissios Vliagoftis, and Matthew Falagas. 2009. Increasing dominance of english in publications archived by pubmed. *Scientometrics*, 81(1):219–223.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 2025. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, pages 4442–4457.

Xinhua Zhu, Xuechen Yang, and Hongchao Chen. 2018. A biomedical question answering system based on snomed-ct. In *Knowledge Science, Engineering and Management*, pages 16–28, Cham. Springer International Publishing.