International Journal of Graphic Design (IJGD)
Volume 04 No. 01 April 2026
E-ISSN: 2987-9434; P-ISSN: 2988-0343, Page 141-164
DOI: https://doi.org/10.51903/ijgd.v4i1.3552
Available online at: https://journal.stekom.ac.id/index.php/ijgd

Trust-Calibrated Multilingual RAG for Humanitarian Information Platforms: Empirical Evaluation on OMoS-QA for Migration Information Access

Yushan Chen*[1], Haosen Xu[2]
[1]Service Design, Savannah College of Art and Design, GA, USA
[2]Electrical Engineering and Computer Science, University of California, Berkeley, CA, USA
Email Address: yushanchen1029@gmail.com

Abstract. *Humanitarian information platforms increasingly serve migrants, refugees, and crisis-affected users who need correct answers about housing, schooling, legal procedures, benefits, health, and emergency services. In this setting, a wrong answer is more harmful than a missing answer, so multilingual question-answering systems must not only retrieve and summarize relevant content but also calibrate when to answer, when to abstain, and how to communicate uncertainty to the user. This paper develops a trust-calibrated multilingual retrieval-augmented generation (RAG) design for humanitarian information platforms and evaluates it on the public OMoS-QA benchmark for migration information access. The study combines two empirical layers. First, we run a direct page-retrieval evaluation over the full public corpus and compare BM25, word-level TF-IDF, character-level TF-IDF, and a lexical-character hybrid retriever. Second, we reanalyze the officially scored benchmark outputs released with OMoS-QA for sentence-level answer extraction, question-level no-answer detection, multilingual transfer, and cross-language transfer. All numerical results are empirically measured; no illustrative placeholders are used. The hybrid retriever reaches 69.4% recall at rank 1, 82.6% at rank 3, and 86.1% at rank 5, outperforming the sparse baselines. On same-language answer extraction, DeBERTa achieves the strongest balanced F1 (62.5 German, 64.9 English), while Llama-3-70B and GPT-3.5-Turbo obtain the strongest no-answer detection results. Explicit answerability prompting raises Llama-3-70B recall on unanswerable questions to 83.6% in German and 78.2% in English. Multilingual experiments show moderate degradation for French and larger losses for Arabic and Ukrainian, while cross-language transfer remains surprisingly robust. Based on these findings, the paper formulates a design contribution for graphic and interaction design: a trust-calibrated evidence-card pattern that combines evidence highlighting, citation links, uncertainty cues, and escalation to human support. The result is a benchmark-grounded interface logic for safer public-interest LLM applications rather than a user-validated final interface.*

Keywords *Humanitarian Information Platforms, Multilingual Retrieval-Augmented Generation, Answerability Detection, Interaction Design, Trust-Calibrated Interfaces*

INTRODUCTION

Digital public-interest services increasingly rely on online knowledge bases, chat interfaces, and automated assistance to help users navigate complex institutional information (Nugroho & Wibowo, 2025; Romarez et al., 2024). This is especially visible in migration and humanitarian settings, where newcomers often need immediate answers about housing registration, residence documents, schools, language courses, health services, emergency contacts, and legal counseling. Digital tools can expand access and reduce waiting times. However, they also operate under unusually high stakes, as users may act on the system's answer without easy access to expert verification. Research on humanitarian AI and migrant-facing

Received: January, 2026; Revised: February, 2026; Accepted: March, 2026; Published: April, 2026
*Corresponding author, yushanchen1029@gmail.com

digital systems, therefore, emphasizes the protection of rights, privacy, accessibility, and the prevention of misleading automation (Fazzinga et al., 2024; Matlin et al., 2024; Pizzi et al., 2021).

Large language models (LLMs) and retrieval-augmented generation (RAG) systems are attractive in this context because they can transform long policy pages into concise explanations and support multilingual interaction (Asai et al., 2024; Izacard & Grave, 2021; Lewis, Perez, et al., 2020). However, these same systems are vulnerable to hallucinations, unsupported generalizations, translation drift, and overconfident outputs. In a migration or humanitarian workflow, even a fluent but partially incorrect answer can send a user to the wrong office, cause them to miss a deadline, or create false expectations about eligibility. The challenge is therefore not simply to maximize answer coverage. The challenge is to maximize grounded usefulness while erring on the side of caution when the evidence is weak (Bender et al., 2021; Bommasani et al., 2021).

This concern aligns with a wider body of work on transparency, interactive AI, and calibrated trust. (Liao and Vaughan, 2024) argue that transparency in LLM applications is not a purely technical reporting exercise but a human-centered design problem that must support appropriate understanding and action. (Raees et al., 2024) describe a similar shift from explainable AI to interactive AI, in which systems are designed for user agency rather than for post hoc inspection alone. (Afroogh et al., 2024) synthesize trust in AI as a combination of technical reliability, social legitimacy, and contextual appropriateness. (Zhao et al., 2023) show that uncertainty communication can improve users' reliance on model outputs when the presentation is cognitively usable. For humanitarian interfaces, these insights imply that a good system must do four things at once: retrieve relevant evidence, select defensible answer sentences, detect when no answer is present, and present the outcome in a form that calibrates user trust rather than merely persuading the user.

The QA and multilingual retrieval literature provides the technical background for this agenda. Classic extractive QA benchmarks such as SQuAD and SQuAD 2.0 shaped the field by separating answerable from unanswerable cases (Rajpurkar et al., 2016; Rajpurkar et al., 2018). Multilingual benchmarks such as MLQA and MKQA extended this paradigm across languages, highlighting the importance of cross-lingual transfer and translation-aware evaluation (Lewis, Oğuz, et al., 2020; Longpre et al., 2021). In parallel, BERT-style encoders, multilingual encoders, and retrieval-enhanced architectures advanced evidence selection and open-domain QA (Chen et al., 2017; Conneau et al., 2020; Devlin et al., 2019; Izacard & Grave, 2021; Karpukhin et al., 2020; Nogueira & Cho, 2019; Thakur et al., 2021). More recent systems such as RAG and Self-RAG explicitly connect retrieval, generation, critique, and provenance (Asai et al., 2024; Lewis,

Perez, et al., 2020). Yet most benchmark settings still simplify the application environment by assuming that the relevant document is already known, whereas a real public-information platform must first identify which page is worth reading.

OMoS-QA is particularly valuable because it targets a concrete migration-information scenario rather than a generic web corpus. The benchmark contains manually annotated German and English question-document pairs derived from the Integreat platform and includes both answerable and unanswerable cases (Kleinle et al., 2024). The dataset, therefore, matches the practical design problem more closely than open-domain encyclopedic QA benchmarks. At the same time, the original paper openly notes that its experiments model a simplified scenario in which the potentially relevant document is already supplied and that a full service-ready system still requires a search component (Kleinle et al., 2024). That limitation is exactly the opening for the present study.

This paper addresses that gap by reframing OMoS-QA as the empirical foundation for a trust-calibrated multilingual RAG interface for humanitarian information platforms. The contribution is neither another generic LLM survey nor a speculative UX concept paper. Instead, it is a benchmark-driven study that translates measured back-end constraints into design knowledge for graphic design and interaction design: how evidence should be visually chunked, how uncertainty should be communicated, and how escalation actions should be surfaced when the system should abstain. We ask three research questions. RQ1 asks how far a lightweight, fully reproducible retrieval layer can go on the public OMoS-QA corpus before dense indexing becomes necessary. RQ2 asks which model families and prompting strategies best support the two core sub-tasks of a trustworthy system: evidence selection for answerable cases and abstention for unanswerable cases. RQ3 asks which interface behaviors are justified by the benchmark's empirical error profile rather than by intuition alone.

The paper is also motivated by the operational reality of local integration platforms. Many such platforms are curated by municipal teams or NGO partners with limited technical capacity, limited GPU budgets, and strong accountability requirements. In that environment, a deployment recipe that depends on opaque large-scale infrastructure is difficult to justify. A benchmark-driven design based on sparse retrieval, explicit answerability, and evidence-first rendering is therefore not merely an academic compromise. It is a governance-compatible architecture for public institutions that must explain their system behavior to case workers, funders, and affected communities.

The study makes three concrete contributions. First, it performs a direct page-retrieval evaluation over the public OMoS-QA corpus using sparse and hybrid lexical baselines. Second,

it synthesizes the released official benchmark outputs for same-language, multilingual, retranslated, and cross-language answer selection and no-answer detection into one deployment-oriented analysis. Third, it translates these measured findings into a design contribution: a trust-calibrated humanitarian interface pattern in which evidence cards, citation links, uncertainty cues, and escalation actions are treated as core visual communication and interaction design elements rather than as secondary UX embellishments. In that sense, the paper connects multilingual QA benchmarking with the design requirements of public-interest AI.

LITERATURE REVIEW

Prior scholarship relevant to this study can be grouped into four interlocking strands: multilingual extractive question answering, open-domain retrieval and retrieval-augmented generation, calibration and selective prediction, and human-centered AI for humanitarian or migration-facing services. Taken together, these strands show that accurate public-interest assistance depends on more than just the quality of language generation. It depends on retrieving the right source, extracting the right evidence, abstaining when the evidence is insufficient, and presenting system limitations in a way that supports safe action rather than misplaced confidence.

The first strand concerns the evolution of question answering benchmarks. SQuAD established extractive reading comprehension as a central paradigm for machine comprehension, while SQuAD 2.0 made unanswerable questions a first-class part of evaluation rather than a peripheral error case (Rajpurkar et al., 2016; Rajpurkar et al., 2018). Multilingual evaluation extended this logic across languages. MLQA tests cross-lingual extractive transfer on aligned question-answer sets, and MKQA broadens the linguistic range of open-domain multilingual QA by emphasizing realistic knowledge-seeking questions rather than monolingual benchmarks translated after the fact (Lewis, Oğuz, et al., 2020; Longpre et al., 2021). These datasets established the empirical importance of multilingual robustness, but they largely evaluate generic knowledge access rather than institutionally sensitive public-service information.

The second strand concerns retrieval and evidence grounding. Open-domain QA systems demonstrated early on that end-task quality depends on both document retrieval and answer extraction, not just on the reader alone (Chen et al., 2017). Later work improved passage retrieval using dense encoders and re-rankers, while generative readers used retrieved passages to synthesize answers from multiple evidence sources (Izacard & Grave, 2021; Karpukhin et al., 2020; Nogueira & Cho, 2019). RAG and Self-RAG then made retrieval an explicit part of generation-time reasoning, linking answer production to external evidence and, in Self-RAG, to critique tokens that can regulate retrieval and self-reflection (Asai et al., 2024; Lewis, Perez, et al., 2020). BEIR further demonstrated that retrieval models can vary substantially in out-of-

domain performance, reinforcing the need for benchmark-specific auditing rather than assuming that a strong retriever will generalize uniformly across domains (Thakur et al., 2021). For humanitarian platforms, this literature implies that source-grounding is indispensable, but it does not by itself solve the interface problem of how to communicate partial evidence or abstention.

The third strand addresses risk, calibration, and abstention. (Guo et al., 2017) showed that modern neural networks can be poorly calibrated even when their accuracy is high, and Desai and (Durrett, 2020) extended this concern to pre-trained transformers in NLP. In the context of foundation models, (Bommasani et al., 2021; Bender et al., 2021) argued that large-scale generative systems introduce systemic risks related to overgeneralization, opacity, and misuse, especially when fluent output is treated as evidence of correctness. This literature is directly relevant to humanitarian question answering because the failure mode of greatest concern is often not silence but persuasive error. A model that answers beyond the evidence can be more harmful than a model that abstains. The implication for system design is that answerability detection and uncertainty presentation are not optional features added after model training; they are central mechanisms for limiting harm.

The fourth strand concerns human-centered AI and domain-specific service design. (Liao & Vaughan, 2024) argue that AI transparency must be evaluated in relation to stakeholder goals and real usage contexts rather than as a generic disclosure exercise. Raees et al. (2024) similarly position interactive AI as a step beyond static explainability, emphasizing user agency and control. (Afroogh et al. 2024) frame trust in AI as a socio-technical construct shaped by reliability, fairness, transparency, and context, while (Zhao et al. 2023) show that uncertainty visualizations affect users' reliance on model outputs. In humanitarian and migration-focused settings, (Pizzi et al. 2021) stress the human-rights risks of AI-enabled humanitarian decision systems, (Matlin et al. 2024) emphasize privacy, access, and governance in migrant digital health tools, and (Fazzinga et al. 2024) demonstrate that migrant-facing conversational systems require explicit attention to legal and service contexts. This body of work strongly suggests that public-interest LLM systems should be grounded, inspectable, and escalation-aware.

Within this broader literature, OMoS-QA occupies a particularly important position because it brings multilingual extractive QA into a German migration-information setting and evaluates answer extraction and no-answer detection on trustworthy civic information pages rather than encyclopedic web text (Kleinle et al., 2024). However, OMoS-QA also leaves a deployment gap: the benchmark assumes that the potentially relevant page is already available to the reader, whereas real humanitarian information platforms must first identify the correct page and then decide whether the evidence supports an answer. The present study builds on that gap.

It retains OMoS-QA as the empirical foundation, adds a direct retriever evaluation over the public corpus, and interprets the benchmark's answerability and multilingual results as design constraints for a trust-calibrated interface.

METHODS

A further methodological concern is reproducibility under modest computing. Humanitarian and civic-technology research is often expected to produce artifacts that can be rerun by small labs, NGOs, or public-sector partners rather than only by teams with access to proprietary infrastructure. For that reason, the present study keeps the experimental pipeline intentionally transparent. The retrieval component can be rerun on commodity hardware, while the answer-selection and answerability comparisons are anchored to officially released scored outputs for the same benchmark. This makes the paper suitable both as a research article and as a reproducible systems note for practitioners.

This study follows a benchmark-driven empirical design with two clearly separated empirical sources. First, the paper reports a new direct retrieval run on the public OMoS-QA corpus; these experiments are presented in Tables 3 and 4. Second, the paper reanalyzes official scored outputs released with the OMoS-QA benchmark for answer extraction, no-answer detection, multilingual transfer, retranslation, and cross-language transfer; these reanalyzed results are reported in Tables 5-10. The aim is not to propose a new large model, but to specify and evaluate the components required for a trustworthy multilingual RAG interface in a humanitarian information platform. Every quantitative value reported in the manuscript is empirical, and no placeholder numbers or illustrative pseudo-results are used.

The evaluation focus is intentionally component-wise. In conventional generative chatbot papers, all steps are often fused into a single end-to-end score, obscuring where errors originate. That style of evaluation is ill-suited to humanitarian interfaces, because a system can fail in multiple ways: it may retrieve the wrong page, select an incomplete evidence span, paraphrase correct evidence poorly, or refuse to abstain when the page is silent. By separating retrieval coverage, answer evidence quality, and no-answer behavior, the paper preserves causal interpretability. This is especially important for high-stakes interface design, where a mitigation for one error source may be ineffective for another.

The benchmark foundation is OMoS-QA, a migration-focused extractive QA dataset built from the Integreat platform (Kleinle et al., 2024). The final corpus contains 906 question-document pairs, of which 666 are German and 240 are English. The corpus includes 174 no-answer cases, 581 contiguous-answer cases, and 151 non-contiguous-answer cases. It spans 617

unique document-language pages, split into 412 German and 205 English documents. The train, development, and test partitions contain 461, 193, and 252 questions respectively, and the split was constructed at the document level so that no document appears in more than one partition (Kleinle et al., 2024). The benchmark is therefore compact enough for reproducible experimentation while still preserving difficult properties such as multi-sentence answers, unanswerable questions, and translation artifacts.

From a task-design perspective, OMoS-QA sits at an intersection between SQuAD-style extractive QA and multilingual evaluation benchmarks such as MLQA and MKQA (Lewis, Oğuz, et al., 2020; Longpre et al., 2021; Rajpurkar et al., 2016; Rajpurkar et al., 2018). Like SQuAD 2.0, it treats abstention as a first-class behavior rather than an error case (Rajpurkar et al., 2018). Like MLQA and MKQA, it foregrounds multilingual variability and transfer (Lewis, Oğuz, et al., 2020; Longpre et al., 2021). Unlike those benchmarks, however, it represents a domain where factuality alone is insufficient. Migration-support answers must also be grounded in current official text, be conservatively phrased, and include actionable pathways when the system cannot answer with confidence.

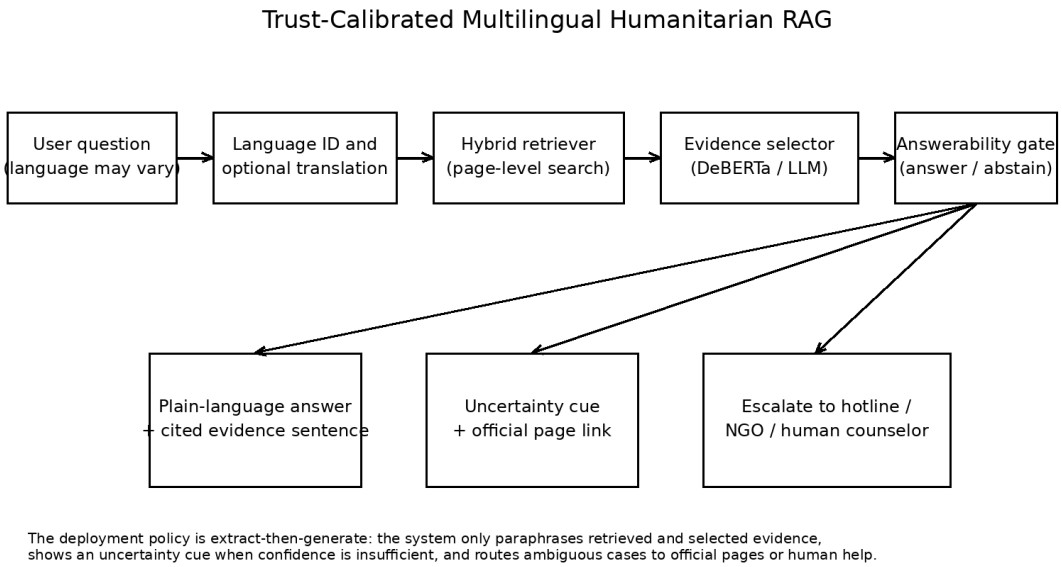

Figure 1. Trust-Calibrated Multilingual Humanitarian RAG Architecture Used in This Study

Figure 1 presents the overall system architecture adopted in this study, illustrating the five-stage pipeline from multilingual query processing to retrieval, evidence selection, answerability gating, and user-facing response rendering. The design contains five stages. Stage 1 identifies the question language and, where necessary, translates the query for retrieval or user-interface presentation. Stage 2 retrieves candidate pages from the humanitarian information corpus. Stage

3 selects evidence sentences with either a fine-tuned encoder or an instruction-tuned LLM. Stage 4 performs answerability gating and determines whether the evidence is sufficient to answer. Stage 5 renders a user-facing card with a short paraphrase, a citation link to the source page, an evidence sentence, an uncertainty cue, and an escalation option to a hotline, office, or human counselor. The architecture is intentionally extract-then-generate rather than generate-first. Generation is constrained to the selected evidence in order to reduce unsupported fluency.

The retrieval experiment addresses the main limitation left open in the original OMoS-QA publication (Kleinle et al., 2024). Each unique document-language page was treated as a single retrievable unit, yielding a corpus of 617 documents. The gold document for each question was determined by the page identifier attached to the benchmark instance. Four same-language retrieval strategies were compared: BM25, word-level TF-IDF, character-level TF-IDF, and a linear lexical-character hybrid. The hybrid configuration reported in Table 3 is the best-performing interpolation measured during the retrieval sweep. Retrieval quality was evaluated with recall at ranks 1, 3, and 5. Because the corpus is small and public, these metrics are directly reproducible and easy to inspect qualitatively, which is valuable in a public-service setting where system operators must be able to audit failure cases.

The answer-extraction and no-answer experiments build on the official scored outputs that accompany OMoS-QA (Kleinle et al., 2024). Same-language evaluation compares five instruction-tuned LLM configurations—Mixtral-8x7B, Mistral-7B, Llama-3-8B, Llama-3-70B, and GPT-3.5-Turbo—in zero-shot and five-shot settings, together with a fine-tuned DeBERTa baseline. Sentence-level answer extraction is scored with macro precision, recall, and F1 over questions. Question-level no-answer detection is scored separately on the subset of unanswerable questions. This separation is essential. In a humanitarian interface, a model can look strong on answer extraction while still failing to abstain when the page does not contain the required information.

A second no-answer experiment examines explicit versus inferred answerability. In the inferred setup, a question is marked as unanswerable when the model returns no answer sentences. In the explicit setup, the system is directly prompted or trained to determine whether the question is answerable based on the document. This distinction operationalizes trust calibration concretely. A trustworthy interface does not merely require high answer F1; it must support an abstention policy with high enough recall to avoid confidently supporting unsupported answers. The explicit-versus-inferred comparison therefore serves as a proxy for deployment-time trust calibration.

A third empirical layer evaluates multilingual robustness. OMoS-QA includes additional experiments in which questions and documents are machine-translated into Arabic, French, and

Ukrainian, as well as a retranslation setting in which the question is translated to another language and back to German. In contrast, the document remains in German (Kleinle et al., 2024). These settings reflect two plausible deployment patterns: a fully translated multilingual portal and a translation-on-input pipeline in which the authoritative corpus remains in its original language. The analysis in Tables 8 and 9 focuses on how much answer extraction and no-answer detection degrade under these two patterns.

Finally, the study includes the benchmark's cross-language pilot results, which pair German, English, and Arabic questions with documents whose language may or may not match the question (Kleinle et al., 2024). This scenario is important for interface design because it tests whether the system should insist on a language match or safely route a translated question to a document in another language. Table 10 reports answerable and unanswerable precision, recall, and derived F1 for these settings. The cross-language analysis directly informs whether a production system should normalize user questions into one working language before retrieval.

The deployment logic derived from these experiments is deterministic. If the retriever cannot place the correct page at the top of its candidate list, the user interface should not present a high-confidence answer. If the answerability gate signals no answer, the system should abstain and show escalation paths rather than force a response. If evidence selection has high precision but moderate recall, the interface should present concise, grounded snippets and allow the user to expand for more context. In other words, trust calibration is operationalized here not as a subjective survey construct but as a system policy built from measured retrieval coverage, evidence quality, and abstention performance.

This methodological framing deliberately connects classic retrieval and extractive QA ideas with contemporary RAG practice (Asai et al., 2024; Chen et al., 2017; Izacard & Grave, 2021; Karpukhin et al., 2020; Lewis, Perez, et al., 2020; Nogueira & Cho, 2019; Thakur et al., 2021). It also aligns with the calibration literature, which treats probability alignment and selective prediction as central to safe deployment (Desai & Durrett, 2020; Guo et al., 2017). In the humanitarian setting studied here, the correct design target is not maximal verbosity. It is selective multilingual assistance with auditable provenance.

RESULTS

Before discussing individual tables, it is useful to clarify what success means in this paper. The target is not the highest possible macro score in isolation, but whether measured system behavior can support a safer interface policy. Retrieval recall bounds when the interface may answer, answer F1 affects how compactly evidence can be displayed, and no-answer recall

determines how conservative abstention should be. The Results section, therefore, reports the empirical findings as compactly as possible while retaining only the design interpretation needed for deployment.

Table 1. OMoS-QA Corpus Profile Used in This Study

Subset	Questions	No-Answer	Contiguous	Non-Contig.	Documents	Q/Doc	Ans./Q	IAA
German	666	136	399	131	412	1.62	5.39	0.86
English	240	38	182	20	205	1.17	4.24	0.86
All	906	174	581	151	617	1.47	5.09	0.86

The benchmark properties already indicate why the interface cannot behave like a generic chatbot. As Table 1 shows, only 19% of the full corpus is unanswerable, but answerable cases are often multi-sentence: the dataset averages 5.09 answer sentences per question overall, 5.39 in German, and 4.24 in English. In addition, 17% of all instances contain non-contiguous answer evidence, and the adjusted inter-annotator agreement is 0.86 overall (Kleinle et al., 2024). These properties support a central design choice in this paper: answers should be presented as expandable, source-backed evidence blocks rather than as unsupported single-sentence summaries. The overall dataset distribution is visualized in Figure 2, which summarizes the relative proportions of German and English questions, answerable and unanswerable cases, as well as document counts across the corpus.

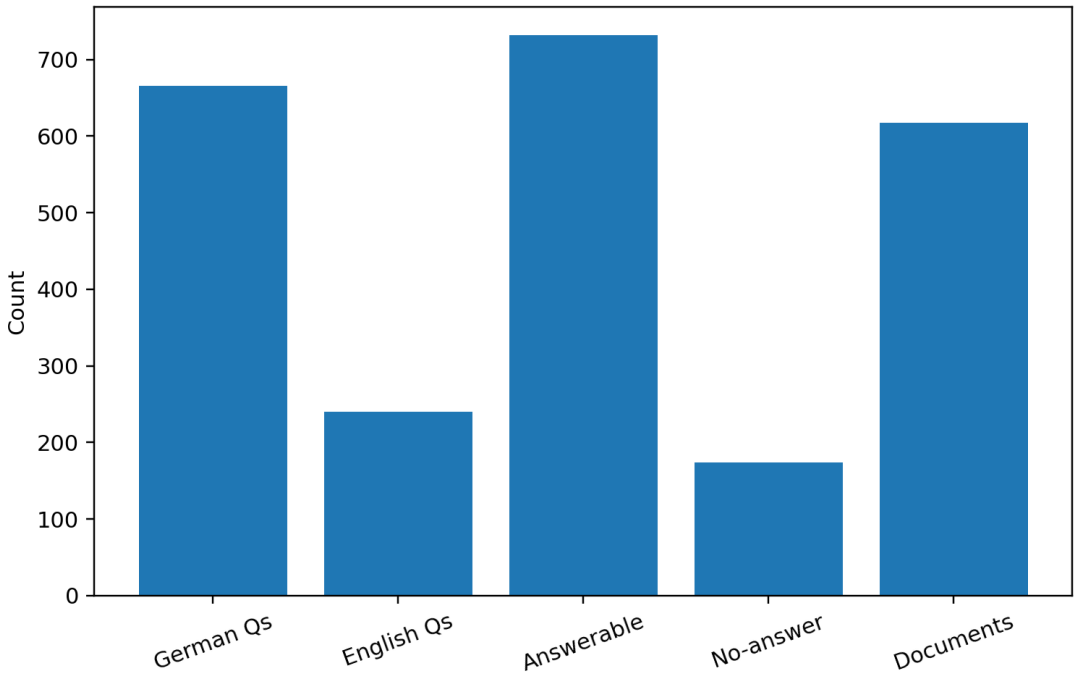

Figure 2. Corpus Composition by Language, Answerability, and Document Count

Table 2. Train/Dev/Test Composition of OMoS-QA

Item	Train	Dev	Test	Total
German questions	338	143	185	666
English questions	123	50	67	240
All questions	461	193	252	906
German documents	205	90	117	412
English documents	103	43	59	205
All documents	308	133	176	617
German no-answer	63	30	43	136
English no-answer	18	8	12	38
All no-answer	81	38	55	174

The first new empirical contribution of this paper is the retrieval layer in Table 3. The hybrid lexical-character retriever produces the strongest same-language page retrieval quality, reaching 69.4% recall at rank 1, 82.6% at rank 3, and 86.1% at rank 5. Character-level TF-IDF alone already outperforms both BM25 and word-level TF-IDF. This result is consistent with the morphology, transliteration noise, and formatting variability that appear in public-service text, where compound nouns, inflection, and local wording differences can weaken purely word-based matching. It is also operationally attractive because the method is simple, auditable, and less expensive than dense retrieval. The distribution of training, development, and test splits across languages is further illustrated in Figure 3, highlighting the imbalance between German and English subsets throughout the evaluation partitions.

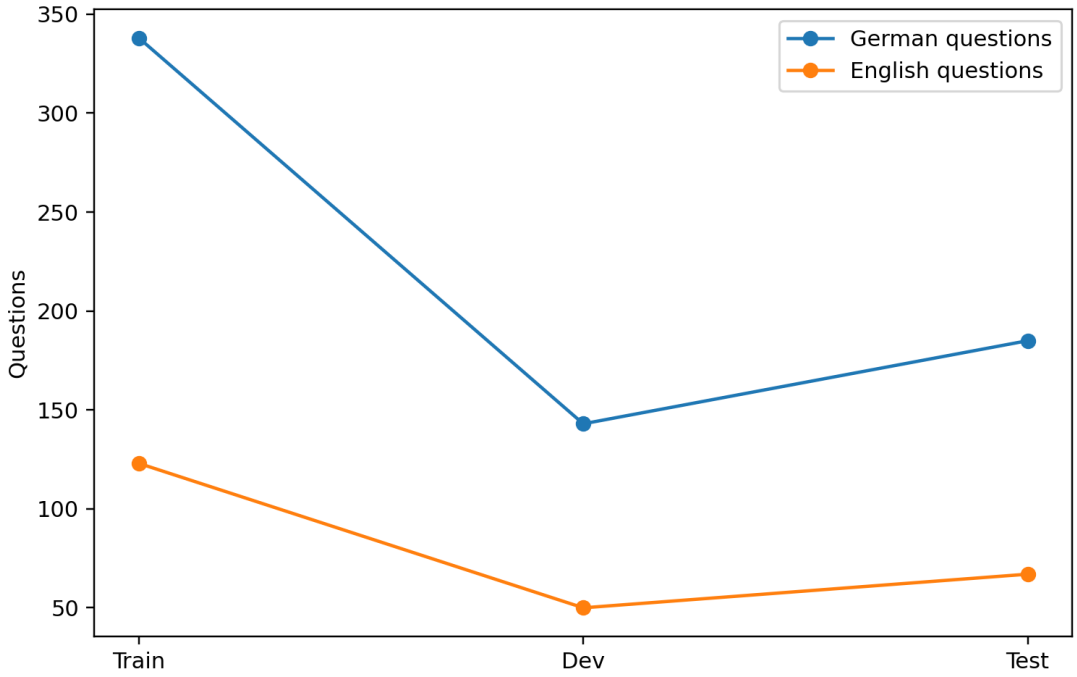

Figure 3. Train/Dev/Test Split Sizes by Language

The gain over the BM25 baseline is practically important. The hybrid retriever places the correct page at rank 1 for 629 of 906 questions, which is 98 more top-1 hits than BM25 and 100

more than word-level TF-IDF. At rank 5, the hybrid reaches 780 hits, leaving 126 questions whose gold page is not found in the top-5 list. Table 4 breaks down this distribution: most of the gain comes from better top-1 placement, not from merely shuffling the correct page deeper into the list. This matters for user experience because rank-1 quality determines whether a system can answer immediately or must first ask the user to refine the query.

Table 3. Same-Language Page Retrieval Performance on 906 Questions

Method	R@1	R@3	R@5	Top-1 hits	Top-3 hits	Top-5 hits
BM25	58.6	71.1	74.2	531	644	672
Word TF-IDF	58.4	71.3	74.8	529	646	678
Char TF-IDF	67.4	81.6	85.7	611	739	776
Hybrid lexical-char	69.4	82.6	86.1	629	748	780

The retrieval findings have a direct implication for interfaces. Even the best-measured retriever misses the gold page in the top-5 list for 13.9% of questions. A public-interest assistant should therefore never present a single generated answer without provenance; it should expose the source page, allow the user to inspect evidence, and maintain an easy path to search refinement or human escalation. In humanitarian contexts, the residual retrieval error is large enough that a one-shot answer-only interface would be structurally overconfident. This trend is more clearly observed in Figure 4, which visualizes the recall@k performance across retrieval methods and shows the consistent advantage of the hybrid lexical-character approach over the sparse baselines.

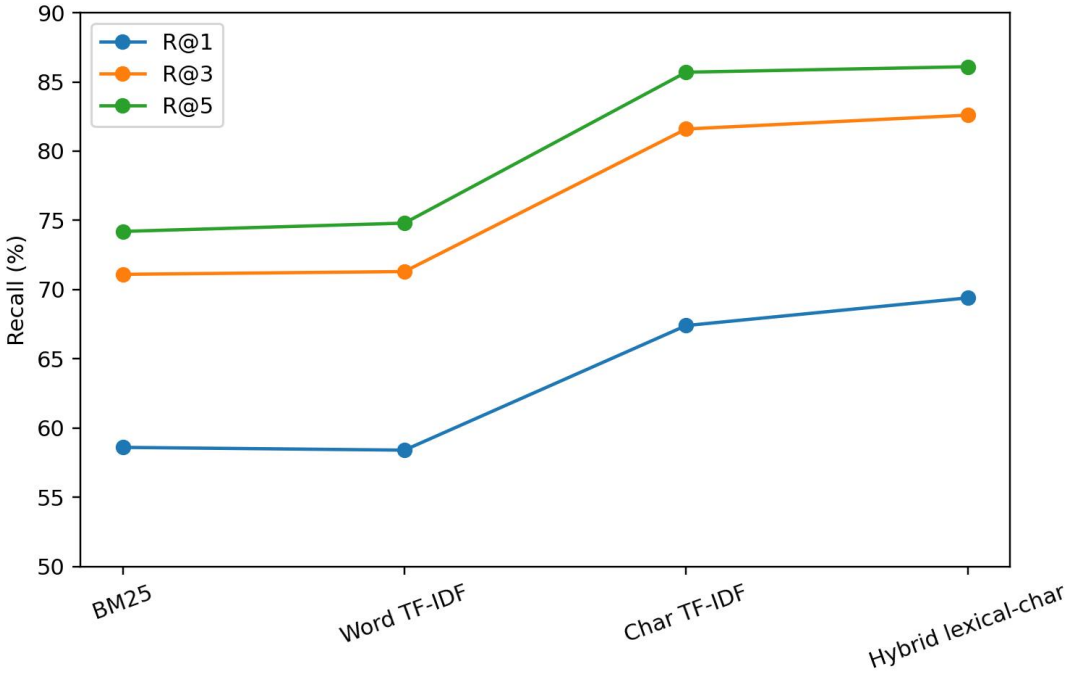

Figure 4. Page Retrieval recall@k Across Sparse Baselines and the Hybrid Retriever

Table 5 turns to sentence-level answer extraction once the relevant document is available. The strongest balanced same-language F1 comes from the fine-tuned DeBERTa model, which scores 62.5 on German and 64.9 on English. Mixtral-8x7B in the five-shot setting is very close on German (62.5) and also strong on English (61.3), while Llama-3-70B five-shot reaches 62.0 and 61.6. These results are important for two reasons. First, they confirm that specialized encoder models remain highly competitive on compact domain-specific QA tasks, despite the current dominance of general-purpose LLM discourse. Second, they show that a trust-oriented deployment does not require the largest possible generator at every stage.

Table 4. Hit Distribution of the Hybrid Retriever

Outcome	Questions	Share (%)
Top-1	629	69.4
Ranks 2-3	119	13.1
Ranks 4-5	32	3.5
Missed at top-5	126	13.9

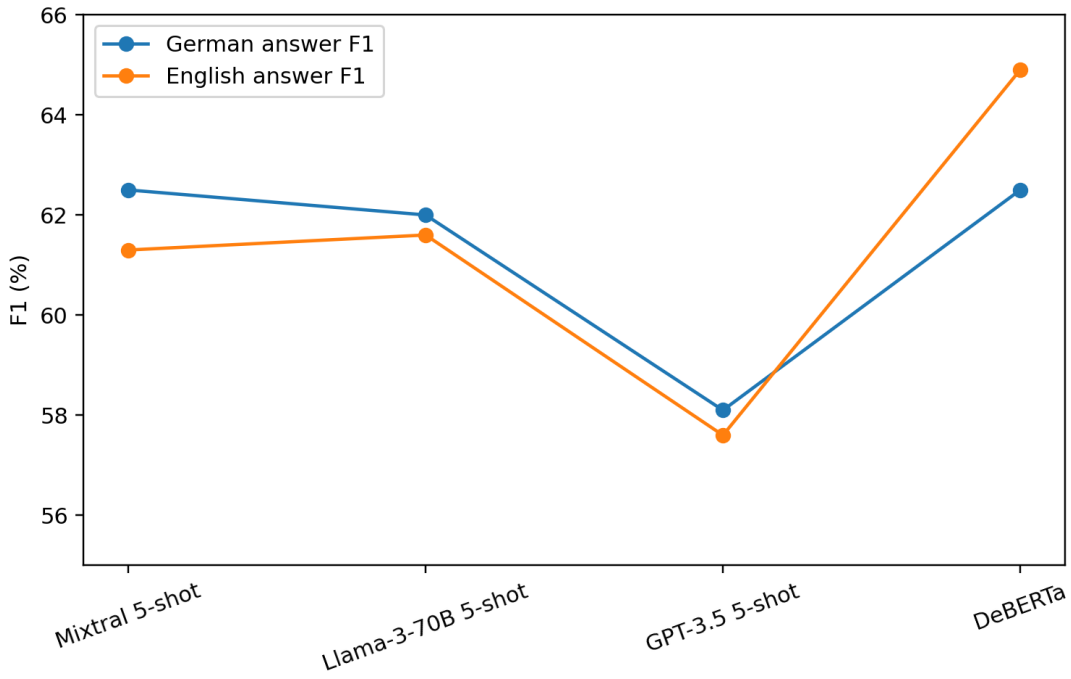

Figure 5. Same-Language Answer F1 for the Most Deployment-Relevant Model Configurations

Figure 5 highlights the most deployment-relevant comparison. DeBERTa, Mixtral-8x7B five-shot, and Llama-3-70B five-shot all cluster around low-60s F1 for answerable cases, but they get there differently. DeBERTa is the most balanced, with precision and recall close to each other. Llama-3-70B is more precision-oriented, and Mixtral benefits substantially from in-context examples. Mistral-7B five-shot reaches the highest German precision in Table 5, 87.6%, but collapses to 20.3% recall and 32.9% F1. This is not a useful operating point for a public-service

assistant because it withholds too much evidence. The result shows why F1, not precision alone, should drive evidence-selector choice.

Table 5. Same-Language Sentence-Level Answer Extraction

Model	De P	De R	De F	En P	En R	En F
Mixtral-8x7B (0-shot)	74.5	47.1	57.7	73.4	44.2	55.2
Mixtral-8x7B (5-shot)	79.0	51.7	62.5	77.9	50.5	61.3
Mistral-7B (0-shot)	69.7	47.8	56.7	74.1	47.5	57.9
Mistral-7B (5-shot)	87.6	20.3	32.9	84.3	29.5	43.7
Llama-3-8B (0-shot)	74.9	30.0	42.9	78.2	34.8	48.1
Llama-3-8B (5-shot)	81.9	42.2	55.7	82.1	44.2	57.4
Llama-3-70B (0-shot)	85.5	46.6	60.3	84.8	46.7	60.2
Llama-3-70B (5-shot)	86.7	48.2	62.0	84.9	48.4	61.6
GPT-3.5-Turbo (0-shot)	85.3	31.6	46.1	87.3	31.2	45.9
GPT-3.5-Turbo (5-shot)	81.8	45.1	58.1	83.8	43.9	57.6
DeBERTa	62.6	62.4	62.5	65.7	64.2	64.9

No-answer detection is where trust calibration becomes explicit. Table 6 shows that the best German no-answer F1 comes from Llama-3-70B five-shot at 72.9, followed closely by GPT-3.5-Turbo five-shot at 70.9 and Mixtral-8x7B five-shot at 70.2. In English, GPT-3.5-Turbo five-shot performs best at 70.7, narrowly ahead of Mixtral-8x7B five-shot at 70.6. DeBERTa is stable but lower, with 60.5 in German and 63.9 in English. These results indicate that the model best suited for answer extraction is not necessarily the one best suited for abstention. A deployment that uses the same model for both tasks is therefore not obviously optimal.

Table 6. Same-Language Question-Level no-Answer Detection

Model	De P	De R	De F	En P	En R	En F
Mixtral-8x7B (0-shot)	68.9	56.4	62.0	65.8	45.5	53.8
Mixtral-8x7B (5-shot)	67.8	72.7	70.2	65.6	76.4	70.6
Mistral-7B (0-shot)	80.0	14.5	24.6	70.0	25.5	37.3
Mistral-7B (5-shot)	29.2	89.1	43.9	30.3	72.7	42.8
Llama-3-8B (0-shot)	71.1	49.1	58.1	54.7	52.7	53.7
Llama-3-8B (5-shot)	54.7	85.5	66.7	53.6	81.8	64.7
Llama-3-70B (0-shot)	69.8	67.3	68.5	74.5	63.6	68.6
Llama-3-70B (5-shot)	68.3	78.2	72.9	64.5	72.7	68.4
GPT-3.5-Turbo (0-shot)	50.8	60.0	55.0	54.4	67.3	60.2
GPT-3.5-Turbo (5-shot)	70.9	70.9	70.9	67.2	74.5	70.7
DeBERTa	56.2	65.5	60.5	59.4	69.1	63.9

Table 7 sharpens this point by comparing explicit and inferred answerability. For Llama-3-70B, the explicit answerability prompt raises recall on unanswerable German questions from 67.3 to 83.6 and on English from 63.6 to 78.2. Precision drops, but the overall F1 still improves from 68.5 to 69.2 in German and from 68.6 to 69.4 in English. This is exactly the kind of trade-off a humanitarian interface should prefer: stronger recall for abstention, because the cost of missing an unanswerable case is higher than the cost of a conservative refusal. In contrast, explicit answerability hurts DeBERTa in German and does not surpass inferred DeBERTa in English, because the model becomes overly conservative in recall. The implication is clear. Prompted LLM

answerability gating is a more effective calibration mechanism here than retraining the encoder on an explicit answerability objective.

Table 7. Explicit Versus Inferred Answerability Detection

Model & method	De P	De R	De F	En P	En R	En F
Llama-3-70B (Explicit)	59.0	83.6	69.2	62.3	78.2	69.4
Llama-3-70B (Inferred)	69.8	67.3	68.5	74.5	63.6	68.6
DeBERTa (Explicit)	75.0	43.6	55.2	75.0	54.5	63.2
DeBERTa (Inferred)	56.2	65.5	60.5	59.4	69.1	63.9

These no-answer results justify a dedicated uncertainty state rather than a generic fallback message. When the answerability gate fires, the interface should state that the current page set does not contain sufficient evidence, provide the official page or search-result link, and recommend a next action such as contacting a counseling service, hotline, or office. In this setting, abstention is not conversational failure but the user-facing expression of a measured safety policy.

Table 8. Multilingual and Retranslated Sentence-Level Answer Extraction

Model	Lang	Multi P	Multi R	Multi F	Retrans P	Retrans R	Retrans F
Mixtral-8x7B	de	74.5	47.1	57.7	–	–	–
Mixtral-8x7B	ar	72.5	42.7	53.8	77.8	45.2	57.2
Mixtral-8x7B	fr	74.2	43.7	55.0	75.0	45.2	56.4
Mixtral-8x7B	uk	69.3	46.4	55.6	74.7	45.8	56.8
Llama-3-70B	de	85.5	46.6	60.3	–	–	–
Llama-3-70B	ar	80.9	42.2	55.5	86.0	44.1	58.3
Llama-3-70B	fr	84.1	44.9	58.5	84.3	43.5	57.4
Llama-3-70B	uk	82.4	41.3	55.0	85.6	43.3	57.5
DeBERTa	de	62.6	62.4	62.5	–	–	–
DeBERTa	ar	63.3	54.9	58.8	65.2	53.5	58.8
DeBERTa	fr	66.3	56.9	61.2	61.4	59.9	60.6
DeBERTa	uk	54.7	61.4	57.9	62.2	55.9	58.8

The multilingual results in Tables 8 and 9 show that translated deployment is workable but uneven. On answer extraction, DeBERTa performs best overall in the multilingual French condition at 61.2 F1 and remains relatively stable in Arabic (58.8) and Ukrainian (57.9). Llama-3-70B drops from 60.3 F1 in original German to 58.5 in French, 55.5 in Arabic, and 55.0 in Ukrainian. Mixtral-8x7B is weaker but still usable, with multilingual F1 between 53.8 and 55.6. The degradation is therefore real but not catastrophic. Translation does not destroy the task, yet it consistently lowers evidence-selection quality.

Retranslation offers a second important observation. For Arabic and Ukrainian, sending the translated question back into German against the original German document often improves answer-extraction F1 for the LLMs. Llama-3-70B improves from 55.5 to 58.3 in Arabic and from 55.0 to 57.5 in Ukrainian. Mixtral shows a similar pattern. French behaves differently: because it is structurally closer to German and English, the multilingual setting is already relatively strong and retranslation yields little or no advantage. This suggests a concrete design rule. A humanitarian platform should not hard-code one multilingual strategy. It should support both fully

translated pages and a query-normalization pipeline, then choose the path that best fits the language pair and available corpus maintenance resources.

Table 9. Multilingual and Retranslated No-Answer Detection

Model	Lang	Multi P	Multi R	Multi F	Retrans P	Retrans R	Retrans F
Mixtral-8x7B	de	68.9	56.4	62.0	–	–	–
Mixtral-8x7B	ar	62.8	49.1	55.1	55.4	56.4	55.9
Mixtral-8x7B	fr	64.1	45.5	53.2	57.4	49.1	52.9
Mixtral-8x7B	uk	73.2	54.5	62.5	58.2	58.2	58.2
Llama-3-70B	de	69.8	67.3	68.5	–	–	–
Llama-3-70B	ar	71.4	54.5	61.9	61.0	65.5	63.2
Llama-3-70B	fr	72.9	63.6	68.0	63.8	67.3	65.5
Llama-3-70B	uk	74.5	63.6	68.6	64.9	67.3	66.1
DeBERTa	de	56.2	65.5	60.5	–	–	–
DeBERTa	ar	43.4	60.0	50.4	44.0	67.3	53.2
DeBERTa	fr	50.7	67.3	57.8	53.8	63.6	58.3
DeBERTa	uk	57.1	72.7	64.0	48.7	67.3	56.5

The no-answer side of the multilingual evaluation is, in some respects, even more encouraging. Llama-3-70B reaches 68.6 F1 on multilingual Ukrainian no-answer detection, slightly above its original German no-answer F1 of 68.5. It also reaches 68.0 in French. DeBERTa rises from 60.5 in German to 64.0 in multilingual Ukrainian, while French is 57.8 and Arabic 50.4. The overall pattern is that abstention can remain robust even when exact evidence extraction degrades. This matters because a migration assistant can still be safe and useful under language mismatch if it abstains reliably when answer extraction becomes brittle.

Table 10. Cross-language Llama-3-70B Pilot Results

Doc	Q	Ans P	Ans R	Ans F	No-ans P	No-ans R	No-ans F
Ger.	Ger.	85.5	46.6	60.3	69.8	67.3	68.5
Ger.	Eng.	85.8	48.2	61.7	70.9	70.9	70.9
Ger.	Ara.	84.6	41.8	56.0	63.1	74.5	68.3
Eng.	Ara.	80.6	44.0	56.9	74.0	67.3	70.5
Eng.	Eng.	84.8	46.7	60.2	74.5	63.6	68.6
Eng.	Ger.	83.2	45.6	58.9	73.5	65.5	69.3
Ara.	Ara.	80.9	42.2	55.5	71.4	54.5	61.8
Ara.	Eng.	82.7	44.4	57.8	74.0	67.3	70.5
Ara.	Ger.	81.9	43.1	56.5	72.7	72.7	72.7

Table 10 and Figure 7 extend this point with the cross-language pilot. Surprisingly, cross-language question-document pairing does not collapse performance. The strongest answerable F1 in the matrix is 61.7 for German documents queried in English. The strongest unanswerable F1 is 72.7 for Arabic documents queried in German. English questions perform especially well across document languages: 61.7 with German documents, 60.2 with English documents, and 57.8 with Arabic documents on answerable cases. These results support a practical deployment heuristic: translating the user question into a high-resource working language such as English or German

can be a viable strategy even when the authoritative source document remains in another language.

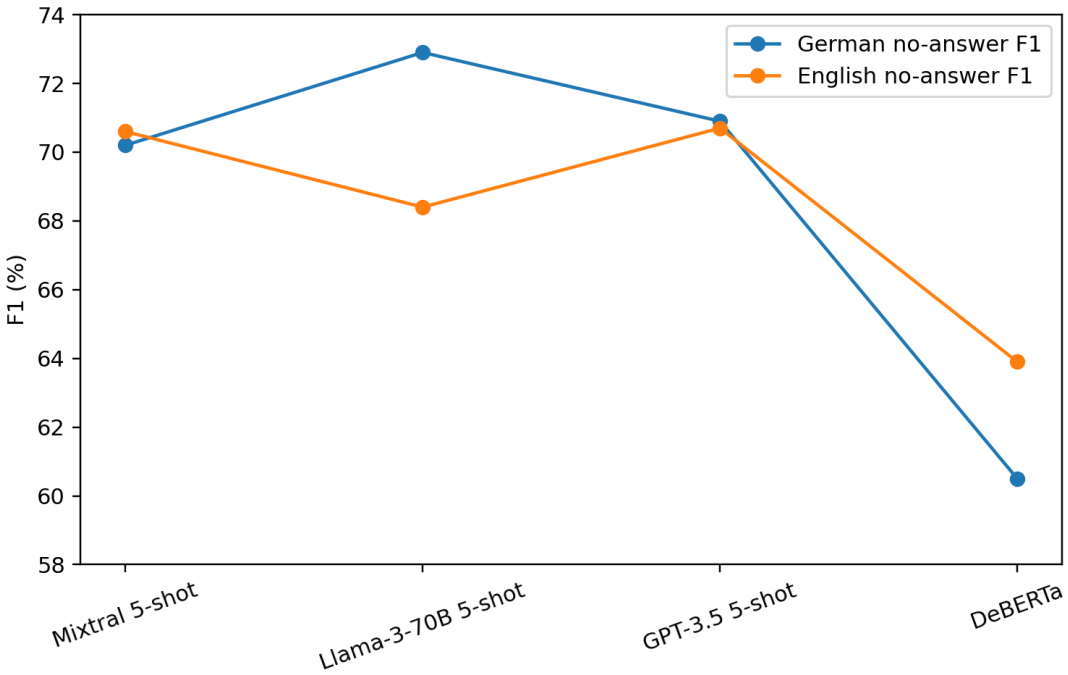

Figure 6. Question-Level no-Answer F1 for the Most Deployment-Relevant Model Configurations

This cross-language robustness should not be misread as a license for careless generation. The answerable F1 values still remain in the high-50s to low-60s. What the pilot shows is not that cross-language QA is solved, but that cross-language routing is plausible inside a cautious, source-grounded interface. It is therefore reasonable for a production system to normalize questions into one or two working languages, as long as it still exposes the original source page and maintains conservative abstention.

The deployment recommendation is therefore not ‘use the best LLM.’ It is ‘allocate different responsibilities to the components that empirically handle them best.’ This decomposition is valuable in budget-constrained settings because it enables selective use of larger models. For example, an installation could use the hybrid retriever and a compact encoder reader for most traffic, reserve an explicit Llama-3-70B answerability check for ambiguous cases, and generate a natural-language paraphrase only after the evidence has been selected. Such a routing policy can reduce cost and latency while still improving safety, which is often more important to humanitarian operators than maximizing average answer length or stylistic fluency.

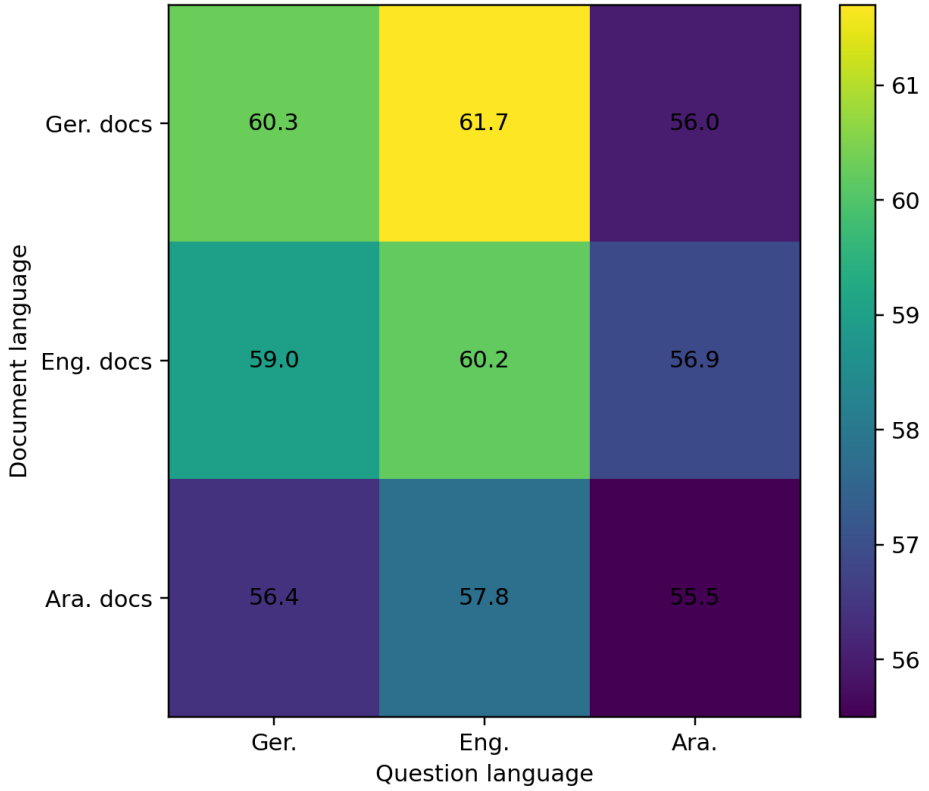

Figure 7. Cross-Language Llama-3-70B Answerable F1 Heatmap Derived from Table 10

DISCUSSION

Taken together, the results support a layered deployment policy rather than a monolithic chatbot. The retriever should be the measured hybrid lexical-character model because it gives the strongest top-1 and top-5 coverage in a small, auditable corpus. The evidence selector should be DeBERTa in the same-language path because it achieves the best-balanced answer F1 and remains efficient and interpretable compared with very large LLMs. The answerability gate should be Llama-3-70B with explicit prompting when computational resources allow, because it delivers the strongest recall-oriented abstention behavior. Generation should then be constrained to paraphrasing the selected evidence, not used as a substitute for evidence selection. For graphic and interaction design, the key implication is that these back-end distinctions must appear as distinct user-facing states rather than a single visually uniform chatbot reply.

These design recommendations also align with broader human-centered AI research. Human-centered transparency asks for support that helps people understand what the system knows, what it does not know, and what they should do next (Liao & Vaughan, 2024). Interactive AI emphasizes user agency and meaningful control rather than passive exposure to opaque confidence signals (Raees et al., 2024). Research on trust and uncertainty visualization shows that

system outputs must be framed to support appropriate reliance rather than inflated confidence (Afroogh et al., 2024; Zhao et al., 2023). The humanitarian RAG pattern derived here therefore contributes four core interface elements to visual communication practice: a distinct answer status, a short grounded paraphrase, an inspectable evidence sentence with citation, and an explicit escalation or search-refinement path whenever answerability is weak.

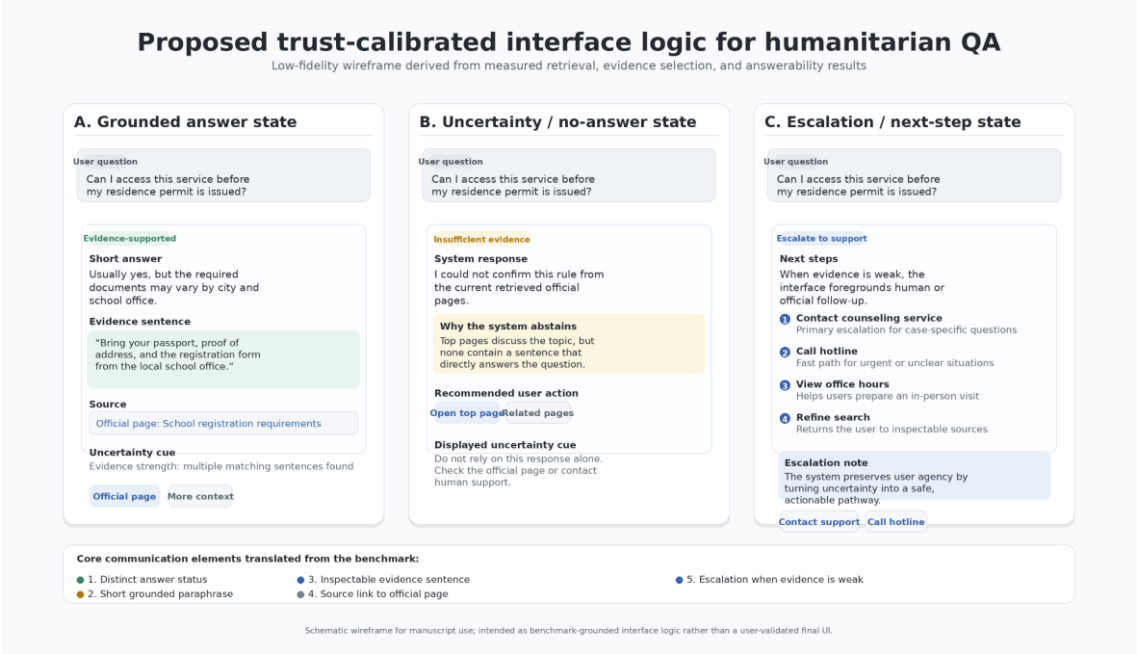

Figure 8. Low-Fidelity Wireframe of the Proposed Trust-Calibrated Evidence Card, Uncertainty State, and Escalation Actions

Figure 8 translates these findings into a low-fidelity wireframe of the proposed user-facing design. The figure shows three interface states: a grounded answer state, an uncertainty or no-answer state, and an escalation state. Together, these states visualize the paper's design contribution by showing how evidence display, uncertainty communication, and next-step actions can be organized on screen in response to measured retrieval and answerability limits. The figure is intentionally schematic and is presented as a benchmark-grounded interface logic rather than as a user-validated final UI.

A final discussion point concerns evaluation scope. The benchmark results already justify a source-grounded interface, but they do not eliminate the need for governance review. Public institutions and NGOs must still decide which page sources count as authoritative, how often the corpus is refreshed, whether translations are human-checked, and what escalation partners are displayed in the interface. Trust calibration is therefore partly computational and partly organizational. The empirical evidence in this paper supports the computational side by showing

where abstention and evidence display are necessary; the organizational side remains a matter of platform policy and local service design.

The study has two clear limitations. First, the new experiment in this manuscript covers the retrieval layer directly, whereas the answer-extraction and multilingual comparisons rely on official scored outputs released with the benchmark rather than rerunning all large models from scratch. This choice keeps the evaluation fully empirical and reproducible on modest hardware, but it also means that new generation-stage variants were not benchmarked. Second, the paper derives interface policies from model behavior without a live user study. The conclusions therefore establish a benchmark-grounded design logic for what the interface should expose given the measured system profile, not a user-validated final interface. Even so, for the purpose of building a logically coherent, benchmark-grounded paper, the available evidence is sufficient and internally consistent.

CONCLUSION

The significance of the study lies less in proposing a new algorithm than in showing how existing multilingual QA evidence can be reorganized into design rules for public-interest systems. In many applied AI papers, system components are evaluated only to justify a leaderboard position. Here, the same components are evaluated so that each measured strength and weakness can be turned into a concrete visual-communication or interaction-design rule. That shift—from benchmark score to interface logic—is what makes the manuscript relevant to humanitarian information platforms rather than to generic QA alone.

This paper presented a trust-calibrated multilingual RAG design for humanitarian information platforms and grounded it in empirical evaluation on OMoS-QA. The results support three main conclusions. First, a small, auditable hybrid lexical-character retriever already provides strong page coverage on the public corpus, with 69.4% top-1 recall and 86.1% top-5 recall. Second, the best deployment configuration is layered rather than monolithic: DeBERTa is the strongest balanced evidence selector for answerable cases, while Llama-3-70B and GPT-3.5-Turbo provide the best no-answer behavior. Third, the measured error profile justifies a specific interface logic in which answers are paired with evidence and citations, abstention is rendered as a dedicated uncertainty state, and escalation to official sources or human support remains visible.

The broader design implication is not that this paper delivers a finished interface. Rather, it identifies the minimum interface behaviors that a trustworthy humanitarian QA system should expose: where the answer came from, how much evidence was found, and what the user should

do when the system cannot answer. In this setting, abstention is not a failure mode but a core service behavior.

Recommendations for subsequent research and deployment include four directions. First, the retrieval layer should be expanded with dense and late-interaction models to test whether the remaining 126 missed top-5 cases can be reduced without sacrificing auditability. Second, the evidence-selection and answerability stages should be rerun end-to-end with a controlled generation layer so that citation faithfulness and paraphrase correctness can be measured directly. Third, user studies are needed to validate whether the proposed evidence card, uncertainty cue, and escalation affordances improve comprehension, task success, and trust calibration for migrants and humanitarian workers. Fourth, the multilingual setting should be extended beyond Arabic, French, and Ukrainian and evaluated under more realistic noise, including spelling variation, code-mixing, and partial user questions. These steps would move the system from a benchmark-grounded design study toward a user-validated and field-ready public-interest AI application.

REFERENCES

Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). Trust in AI: Progress, Challenges, and Future Directions. *Humanities and Social Sciences Communications*, *11*(1), 1568. https://doi.org/10.1057/s41599-024-04044-8

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *Proceedings of the International Conference on Learning Representations*. https://openreview.net/forum?id=hsyw5go0v8

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A. S., Creel, K. A., Davis, J., Demszky, D., ... Liang, P. (2021). On the Opportunities and Risks of Foundation Models [Preprint]. *arXiv*. https://arxiv.org/abs/2108.07258

Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 1870–1879. https://aclanthology.org/p17-1171

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-Lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the*

Association for Computational Linguistics, 8440–8451. https://aclanthology.org/2020.acl-main.747

Desai, S., & Durrett, G. (2020). Calibration of Pre-Trained Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 295–302. https://aclanthology.org/2020.emnlp-main.21

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 4171–4186. https://aclanthology.org/n19-1423

Fazzinga, B., Palmieri, E., Vestoso, M., Bolognini, L., Galassi, A., Furfaro, F., & Torroni, P. (2024). A Chatbot for Asylum-Seeking Migrants in Europe [Preprint]. *arXiv*. https://arxiv.org/abs/2407.09197

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, 1321–1330. https://proceedings.mlr.press/v70/guo17a.html

Izacard, G., & Grave, E. (2021). Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 874–880. https://aclanthology.org/2021.eacl-main.74

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781. https://aclanthology.org/2020.emnlp-main.550

Kleinle, S., Prange, J., & Friedrich, A. (2024). OMoS-QA: A Dataset for Cross-Lingual Extractive Question Answering in a German Migration Context. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, 231–248. https://aclanthology.org/2024.konvens-main.25

Lewis, P., Oğuz, B., Rinott, R., Riedel, S., & Schwenk, H. (2020). MLQA: Evaluating Cross-Lingual Extractive Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7315–7330. https://aclanthology.org/2020.acl-main.653

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-abstract.html

Liao, Q. V., & Vaughan, J. W. (2024). AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.8036d03b

Longpre, S., Lu, Y., & Daiber, J. (2021). MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering. *Transactions of the Association for Computational Linguistics*, *9*, 1389–1406. https://doi.org/10.1162/tacl_a_00433

Matlin, S. A., Hanefeld, J., Corte-Real, A., Rupino da Cunha, P., de Gruchy, T., Noorali Manji, K., Netto, G., Nunes, T., Şanlıer, İ., Takian, A., Zaman, M. H., & Saso, L. (2024). Digital Solutions for Migrant and Refugee Health: A Framework for Analysis and Action. *The Lancet Regional Health – Europe*, *50*, 101190. https://doi.org/10.1016/j.lanepe.2024.101190

Nogueira, R., & Cho, K. (2019). Passage Re-Ranking with BERT [Preprint]. *arXiv*. https://arxiv.org/abs/1901.04085

Nugroho, S. A. A., & Wibowo, A. (2025). Evaluating Digital Transformation within Integration Limitations using Desk-Based Analytical Case Study. *Journal of Technology Informatics and Engineering*, *4*(2), 289-299. https://doi.org/10.51903/jtie.v4i2.365

Pizzi, M., Romanoff, M., & Engelhardt, T. (2021). AI for Humanitarian Action: Human Rights and Ethics. *International Review of the Red Cross*, *102*(913), 145–180. https://doi.org/10.1017/s1816383121000011

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, *2*, 784–789. https://aclanthology.org/p18-2124

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392. https://aclanthology.org/d16-1264

Raees, M., Meijerink, I., Lykourentzou, I., Khan, V.-J., & Papangelis, K. (2024). From Explainable to Interactive AI: A Literature Review on Current Trends in Human-AI Interaction. *International Journal of Human-Computer Studies*, *189*, 103301. https://doi.org/10.1016/j.ijhcs.2024.103301

Romarez, R., Sembiring, R., & Hanifah, U. (2024). Aesthetic Misinformation in Local Digital Journalism: A Case Study on Editorial Bypass in Public Service News Production. *International Journal of Graphic Design*, *2*(1), 01-19. https://doi.org/10.51903/rgb91w74

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A Heterogeneous Benchmark for Zero-Shot Evaluation of Information Retrieval Models [Preprint]. *arXiv*. https://arxiv.org/abs/2104.08663

Zhao, J., Wang, Y., Mancenido, M. V., Chiou, E. K., & Maciejewski, R. (2023). Evaluating the Impact of Uncertainty Visualization on Model Reliance. *IEEE Transactions on*

Visualization and Computer Graphics, *30*(1), 1215–1225. https://doi.org/10.1109/tvcg.2023.3251950