HEALTH-PARIKSHA: Assessing RAG Models for Health Chatbots in Real-World Multilingual Settings

Varun Gumma[♠] **Anandhita Raghunath**[*◇]
Mohit Jain[†♠] **Sunayana Sitaram**[†♠]
[♠]Microsoft Corporation [◇]University of Washington
varun230999@gmail.com, sunayana.sitaram@microsoft.com

Abstract

Assessing the capabilities and limitations of large language models (LLMs) has garnered significant interest, yet the evaluation of multiple models in real-world scenarios remains rare. Multilingual evaluation often relies on translated benchmarks, which typically do not capture linguistic and cultural nuances present in the source language. This study provides an extensive assessment of 24 LLMs on real world data collected from Indian patients interacting with a medical chatbot in Indian English and 4 other Indic languages. We employ a uniform Retrieval Augmented Generation framework to generate responses, which are evaluated using both automated techniques and human evaluators on four specific metrics relevant to our application. We find that models vary significantly in their performance and that instruction tuned Indic models do not always perform well on Indic language queries. Further, we empirically show that factual correctness is generally lower for responses to Indic queries compared to English queries. Finally, our qualitative work shows that code-mixed and culturally relevant queries in our dataset pose challenges to evaluated models.

1 Introduction

Large Language Models (LLMs) have demonstrated impressive proficiency across various domains. Nonetheless, their full spectrum of capabilities and limitations remains unclear, resulting in unpredictable performance on certain tasks. Additionally, there is now a wide selection of LLMs available. Therefore, evaluation has become crucial for comprehending the internal mechanisms of LLMs and for comparing them against each other.

Despite the importance of evaluation, significant challenges still persist. Many widely-used benchmarks for assessing LLMs are contaminated (Ahuja et al., 2024; Oren et al., 2024; Xu et al., 2024), meaning that they often appear in LLM training data. Some of these benchmarks were originally created for conventional Natural Language Processing tasks and may not fully represent current practical applications of LLMs (Conneau et al., 2018; Pan et al., 2017). Recently, there has been growing interest in assessing LLMs within multilingual and multicultural contexts (Ahuja et al., 2023, 2024; Faisal et al., 2024; Watts et al., 2024; Chiu et al., 2024). Traditionally, these benchmarks were developed by translating English versions into various languages. However, due to the loss of linguistic and cultural context during translation, new benchmarks specific to different languages and cultures are now being created. However, such benchmarks are few in number, and several of the older ones are contaminated in training data (Ahuja et al., 2024; Oren et al., 2024). Thus, there is a need for new benchmarks that can test the abilities of models in real-world multilingual settings.

LLMs are employed in various fields, including critical areas like healthcare. Jin et al. (2024) translate an English healthcare dataset into Spanish, Chinese, and Hindi, and demonstrate that performance declines in these languages compared to English. This highlights the necessity of examining LLMs more thoroughly in multilingual contexts for these important uses.

In this study, we conduct the first comprehensive assessment of multilingual models within a real-world healthcare context. We evaluate responses from 24 multilingual and Indic models using 750 questions posed by users of a health chatbot in five languages (Indian English and four Indic languages). All the models being evaluated function within the same RAG framework, and their outputs are compared to doctor-verified ground truth responses. We evaluate LLM responses on four metrics curated for our application, including factual correctness, semantic similarity, coherence,

[*]Work done during an internship at Microsoft
[†]Equal Advising

and conciseness and present leaderboards for each metric, as well as an overall leaderboard. We use human evaluation and automated methods (LLMs-as-a-judge) to compute these metrics by comparing LLM responses with ground-truth reference responses or assessing the responses in a reference-free manner.

Our results suggest that models vary significantly in their performance, with some smaller models outperforming larger ones. Factual Correctness is generally lower for non-English queries compared to English queries. We observe that instruction-tuned Indic models do not always perform well on Indic language queries. Our dataset contains several instances of code-mixed and culturally-relevant queries, which models sometimes struggle to answer. The contributions of our work are as follows:

- We evaluate 24 models (proprietary as well as open weights) in a healthcare setting using queries provided by patients using a medical chatbot. This guarantees that our dataset is not contaminated in the training data of any of the models we evaluate.

- We curate a dataset of queries from multilingual users that spans multiple languages. The queries feature language typical of multilingual communities, such as code-switching, which is rarely found in translated datasets, making ours a more realistic dataset for model evaluation.

- We evaluate several models in an identical RAG setting, making it possible to compare models in a fair manner. The RAG setting is a popular configuration that numerous models are being deployed in for real-world applications.

- We establish relevant metrics for our application and determine an overall combined metric by consulting domain experts - doctors working on the medical chatbot project.

- We perform assessments (with and without ground truth references) using LLM-as-a-judge and conduct human evaluations on a subset of the models and data to confirm the validity of the LLM assessment.

2 Related Works

Healthcare Chatbots in India Within the Indian context, the literature has documented great diversity in health seeking and health communication behaviors based on gender (Das et al., 2018), varying educational status, poor functional literacy, cultural context (Islary, 2018), stigmas (Wang et al.) etc. This diversity in behavior may translate to people's use of medical chatbots, which are increasingly reaching hundreds of Indian patients at the margins of the healthcare system (Mishra et al., 2023). These bots solicit personal health information directly from patients in their native Indic languages or in Indic English. For example, (Ramjee et al., 2024) find that their CataractBot deployed in Bangalore, India yields patient questions on topics such as surgery, preoperative preparation, diet, exercise, discharge, medication, pain management, etc. Mishra et al. (2023) find that Indian people share "*deeply personal questions and concerns about sexual and reproductive health*" with their chatbot SnehAI. Yadav et al. (2019) find that queries to chatbots are "*embedded deeply into a communities myths and existing belief systems*" while (Xiao et al., 2023) note that patients have difficulties finding health information at an appropriate level for them to comprehend. Therefore, LLMs powering medical chatbots in India and other Low and Middle Income Countries are challenged to respond lucidly to medical questions that are asked in ways that may be hyperlocal to patient context. Few works have documented how LLMs react to this linguistic diversity in the medical domain. Our work begins to bridge this gap.

Multilingual and RAG evaluation Several previous studies have conducted in-depth evaluation of Multilingual capabilities of LLMs by evaluating across standard tasks (Srivastava et al., 2022; Liang et al., 2023; Ahuja et al., 2023, 2024; Asai et al., 2024; Lai et al., 2023; Robinson et al., 2023), with a common finding that current LLMs only have a limited multilingual capacity. Other works (Watts et al., 2024; Leong et al., 2023) include evaluating LLMs on creative and generative tasks. Salemi and Zamani (2024) state that evaluating RAG models require a joint evaluating of the retrival and generated output. Recent works such as Chen et al. (2024); Chirkova et al. (2024) benchmark LLMs as RAG models in bilingual and multilingual setups. Lastly, several tools and benchmarks have also been built for automatic evaluation of RAG, even in medical domains (Es et al., 2024; Tang and Yang, 2024; Xiong et al., 2024a,b), and we refer the readers to Yu et al. (2024) for such a comprehensive list and survey.

LLM-based Evaluators With the advent of large-scale instruction following capabilities in LLMs, automatic evaluations with the help of these models is being preferred (Kim et al., 2024a,b; Liu et al., 2024; Shen et al., 2023; Kocmi and Federmann, 2023). However, it has been shown that it is optimal to assess these evaluations in tandem with human annotations as LLMs can provide inflated scores (Hada et al., 2024b,a; Watts et al., 2024). Other works (Zheng et al., 2023; Watts et al., 2024) have employed GPT-4 alongside human evaluators to leaderboards to assess other LLMs. Ning et al. (2024) proposed an innovative approach using LLMs for peer review, where models evaluate each other's outputs. However, a recent study by Doddapaneni et al. (2024) highlighted the limitations of LLM-based evaluators, revealing their inability to reliably detect subtle drops in input quality during evaluations, raising concerns about their precision and dependability for fine-grained assessments. In this work, we use LLM-based evaluators both with and without ground-truth references and also use human evaluation to validate LLM-based evaluation.

3 Methodology

In this study, we leveraged a dataset collected from a deployed medical chatbot. Here, we provide an overview of the question dataset, the knowledge base employed for answering those questions, the process for generating responses, and the evaluation framework.

3.1 Data

The real-world test data was collected by our collaborators as part of an ongoing research effort that designed and deployed a medical chatbot, hereafter referred to as HEALTHBOT, to patients scheduled for cataract surgery at a large hospital in urban India. An Ethics approval was obtained from our institution prior to conducting this work, and once enrolled in the study and consent was obtained, both the patient and their accompanying family member or attendant were instructed on how to use HEALTHBOT on WhatsApp. Through this instructional phase, they were informed that questions could be asked by voice or by text, in one of 5 languages - English, Hindi, Kannada, Tamil, Telugu. The workflow of chatting with HEALTHBOT was as follows: Patients sent questions through the WhatsApp interface to HEALTHBOT. Their questions were transcribed automatically (using a speech recognition system) and translated (using an off-the-shelf translator) into English if needed, after which GPT-4 was used to to produce an initial response by performing RAG on the documents in the knowledge base (KB, see below). This initial response was passed to doctors who reviewed, validated, and if needed, edited the answer. The doctor approved answer is henceforth referred to as the ground truth (GT) response associated with the patient query.

Our evaluation dataset was curated from this data by including all questions sent to HEALTHBOT along with their associated GT response. Exclusion criteria removed exact duplicate questions, those with personally identifying information, and those not relevant to health. Additionally, for this work, we only consider questions to which the GPT-4 answer was directly approved by the expert as the "*correct and complete answer*" without additional editing on the doctors' part. The final dataset contained 749 question and GT answer pairs that were sent in to HEALTHBOT between December 2023 to June 2024. In the pool, 666 questions were in English, 19 in Hindi, 27 in Tamil, 14 in Telugu, and 23 in Kannada. Note that, queries written in the script of a specific language were classified as belonging to that language. For code-mixed and Romanized queries, we determined whether they were English or non-English based on the matrix language of the query.

The evaluation dataset consists of queries that (1) have misspelled English words, (2) are code-mixed, (3) represent non-native English, (4) are relevant to the patient's cultural context and (5) are specific to the patient's condition. We provide some examples of each of these categories.

Examples of misspelled queries include questions such as "*How long should saving not be done after surgery?*" where the patient intended to ask about shaving, and "*Sarjere is don mam?*" which the attendant used to inquire about the patient's discharge status. Instances of code mixing can be seen in phrases like "*Agar operation ke baad pain ho raha hai, to kya karna hai?*" meaning "*If there is pain after the surgery, what should I do?*" in Hindi-English (Hinglish). Other examples include "*Can I eat before the kanna operation?*" where "*kanna*" means eye in Tamil, and "*kanna operation*" is a well understood, common way of referring to cataract surgery, and "*In how many days can a patient take Karwat?*" where "*Karwat*" means turning over in sleep in Hindi.

Indian English was used in a majority of the English queries, making the phrasing of questions different from what they would be with native English speech. Examples are as follows - "*Because I have diabetes sugar problem I am worried much*", "*Why to eat light meal only? What comes under light meal?*" and "*Is the patient should be in dark room after surgery?*" Taking a shower was commonly referred to as "*taking a bath*", and eye glasses were commonly referred to as "*goggles*", "*spex*" or "*spectacles*".

Culturally-relevant questions were also many in number, for example questions about specific foods were asked like "*Can he take chapati, Puri etc on the day of surgery?*" and "*Can I eat non veg after surgery?*" ("*non-veg*" is a term used in Indian English to denote eating meat). Questions about yoga were asked, like "*How long after the surgery should the Valsalva maneuver be avoided?*" and "*Are there any specific yoga poses I can do?*". The notion of a patient's native place or village was brought up in queries such as "*If a person gets operated here and then goes to his native place and if some problem occurs what shall he do ?*" or "*Can she travel by car with AC for 100 kms ?*".

3.2 Knowledge Base

The documents populating the knowledge base (KB) were initially curated by doctors at the hospital where HEALTHBOT was deployed. This consisted of 12 PDF documents that were converted into text files and manually error checked. The documents included Standard Operating Provedure manuals, standard treatment guidelines, consent forms, frequently-asked-question documents, insurance information, etc. Following this initial curation, doctors that were with HEALTHBOT were able to select question-answer pairs to be added to KB after the bot was deployed. In this manner, the knowledge available to GPT-4 in the KB grew over time. Therefore, every question that was asked by patients was associated with a different version of the KB being used for answer generation. This detail was incorporated into our evaluation in order to compare the verified ground truth data with the generated response in an accurate manner. All KB documents were chunked to a maximum length of 1000 tokens, and embedded in VectorDB[1] using the TEXT-EMBEDDING-ADA-002[2]. Subsequently, for each query, the top 3 most relevant chunks are extracted, and the models are queried with this data.

3.3 Models

We chose 24 models including proprietary multilingual models, as well as Open-weights multilingual and Indic language models for our evaluation. A full list of models can be found in Table 1.

3.4 Response Generation

We use the standard Retrieval-Augmented-Generation (RAG) strategy to elicit responses from all the models. Each model is asked to respond the given query by extracting the appropriate pieces of text from the knowledge-base chunks. During prompting, we segregate the chunks into RAWCHUNKS and KBUPDATECHUNKS symbolizing the data from the standard sources, and the KB updates. Then model is explicitly instructed to prioritize the information from the most latest sources, i.e. the KBUPDATECHUNKS (if they are available). The exact prompt using for generation is provided in Appendix X. Note that each model gets the same RAWCHUNKS and KBUPDATECHUNKS, which are also the same that are given to the GPT-4 model in the HEALTHBOT, based on which the GT responses are verified.

3.5 Response Evaluation

We used both human and automated evaluation to evaluate the performance of models in the setup described above. GPT-4o[3] was employed as an LLM evaluator. We prompted the model separately to judge each metric, as Hada et al. (2024b,a) show that individual calls reduce interaction and influence among and their evaluations.

3.5.1 LLM Evaluation

In consultation with domain experts working on the HEALTHBOT, we curated metrics that are relevant for our application. We limit ourselves to 3 classes (Good - 2, Medium - 1, Bad - 0) for each metric, as a larger number of classes could hurt interpretability and lower LLM-evaluator performance. The prompt used for each of our metrics are available in Appendix A.2, and a general overview is provided below.

[1] https://www.trychroma.com

[2] https://platform.openai.com/docs/guides/embeddings/embedding-models

[3] https://openai.com/index/hello-gpt-4o/

Models	Languages Tested	Availability
GPT-4	All	Proprietary
GPT-4o	All	Proprietary
microsoft/Phi-3.5-MoE-instruct	All	Open-weights
CohereForAI/c4ai-command-r-plus-08-2024	All	Open-weights
Qwen/Qwen2.5-72B-Instruct	All	Open-weights
CohereForAI/aya-23-35B	All	Open-weights
mistralai/Mistral-Large-Instruct-2407	All	Open-weights
google/gemma-2-27b-it	All	Open-weights
meta-llama/Meta-Llama-3.1-70B-Instruct	All	Open-weights
GenVRadmin/llama38bGenZ_Vikas-Merged	All	Indic
GenVRadmin/AryaBhatta-GemmaOrca-Merged	All	Indic
GenVRadmin/AryaBhatta-GemmaUltra-Merged	All	Indic
GenVRadmin/AryaBhatta-GemmaGenZ-Vikas-Merged	All	Indic
Telugu-LLM-Labs/Indic-gemma-7b-finetuned-sft-Navarasa-2.0	All	Indic
ai4bharat/Airavata	En, Hi	Indic
Cognitive-Lab/LLama3-Gaja-Hindi-8B-v0.1	En, Hi	Indic
BhabhaAI/Gajendra-v0.1	En, Hi	Indic
manishiitg/open-aditi-hi-v4	En, Hi	Indic
abhinand/tamil-llama-7b-instruct-v0.2	En, Ta	Indic
abhinand/telugu-llama-7b-instruct-v0.1	En, Te	Indic
Telugu-LLM-Labs/Telugu-Llama2-7B-v0-Instruct	En, Te	Indic
Tensoic/Kan-Llama-7B-SFT-v0.5	En, Ka	Indic
Cognitive-Lab/Ambari-7B-Instruct-v0.2	En, Ka	Indic
GenVRadmin/Llamavaad	En, Hi	Indic

Table 1: List of models tested. “En” for English, “Hi” for Hindi, “Ka” for Kannada, “Ta” for Tamil, “Te” for Telugu, and “All" refers to all the aforementioned languages. All Indic models are open-weights.

- FACTUAL CORRECTNESS (FC): As Doddapaneni et al. (2024) had shown that LLM-based evaluators fail to identify subtle factual inaccuracies, we curate a separate metric to double-check facts like dates, numbers, procedure and medicine names.
- SEMANTIC SIMILARITY (SS): Similarly, we formulate another metric to specifically analyse if both the prediction and the ground-truth response convey the same information semantically, especially when the they are in different languages.
- COHERENCE (COH): This metric evaluates if the model was able to stitch together appropriate pieces of information from the three data chunks provided to yield a coherent response.
- CONCISENESS (CON): Since the knowledge base chunks extracted and provided to the model can be quite large, with important facts embedded at different positions, we build this metric to assess the ability of the model to extract and compress all these bits of information relevant to the query into a crisp response.

Among the metrics presented above, FACTUAL CORRECTNESS and SEMANTIC SIMILARITY use the GT response verified by doctors as a reference, while COHERENCE and CONCISENESS are reference-free metrics. In order to arrive at a combined score for each model, we asked two doctors who collaborate on the HEALTHBOT to assign weights to the first four metrics according to their importance and used an average of the percentages for each metric as the final coefficient to compute the AGGREGATE (AGG). Both doctors gave the maximum weight to FACTUAL CORRECTNESS followed by SEMANTIC SIMILARITY while COHERENCE and CONCISENESS were given lower and equal weightage.

3.5.2 Human Evaluation

Following previous works (Hada et al., 2024b,a; Watts et al., 2024), we augment the LLM evaluation with human evaluation and draw correlations between the LLM evaluator and human evaluation for a subset of the models (PHI-3.5-MOE-INSTRUCT, MISTRAL-LARGE-INSTRUCT-2407, GPT-4O, META-LLAMA-3.1-70B-INSTRUCT, INDIC-GEMMA-7B-FINETUNED-SFT-NAVARASA-2.0). These models were selected based on results from early automated evaluations, covering a range of scores and representing models of interest.

The human annotators were employed by

KARYA, a data annotation company and were all native speakers of Indian languages that we evaluated. We selected a sample of 100 queries from English, and all the queries from Indic languages for annotation, yielding a total of 183 queries. Each instance was annotated by one annotator for SEMANTIC SIMILARITY between the model's response and the GT response provided by the doctor. The annotations began with a briefing about the task and each of them was given a sample test task, and were provided some guidance based on their difficulties and mistakes. Finally, the annotators were asked to evaluate the model response based on the metric[4], query, and ground-truth response on a scale of 0 to 2, similar to the LLM-evaluator.

4 Results

In this section, we present the outcomes of both the LLM and human evaluations. We begin by examining the average scores across all our metrics including the combined metric for English queries, followed by results for queries in other languages. Next, we examine the ranking of models based on scores given by human annotators and compare these rankings based on scores provided by the LLM evaluator. Lastly, we conduct a qualitative analysis of the outcomes and describe noteworthy findings.

4.1 LLM evaluator results

We see from Table 2 that for English, the best performing models is the QWEN2.5-72B-INSTRUCT model across all metrics. Note that it is expected that GPT-4 performs well, as the ground truth responses are based on responses generated by GPT-4. The PHI-3.5-MOE-INSTRUCT model also performs well on all metrics, followed by MISTRAL-LARGE-INSTRUCT-2407 and OPEN-ADITI-HI-V4, which is the only Indic model that performs near the top even for English queries. Surprisingly, the META-LLAMA-3.1-70B-INSTRUCT model performs worse than expected on this task, frequently regurgitating the entire prompt that was provided. In general, all models get higher scores on conciseness and many models do well on coherence.

For the non-English queries, which are far fewer in number compared to English (Tables 3, 5, 6, 4 in Appendix A.1), we find that models such as AYA-23-35B perform near the top for Hindi along with proprietary and large open weights models such as QWEN2.5-72B-INSTRUCT and MISTRAL-LARGE-INSTRUCT-2407, outperforming many of the fine-tuned Indic LLMs. The GEMMA-2-27B-IT model also outperforms many Indic models in the Indic setting, compared to its performance in English. This shows that some instruction-tuned Indic LLMs may not perform well in the RAG setting. We also find that compared to English, models get lower values on FC on Indic queries, which is concerning as it is rated as the most important metric by doctors.

4.2 Comparison of human and LLM evaluators

We perform human evaluation on five models on the SEMANTIC SIMILARITY (SS) task and compare human and LLM evaluation by inspecting the ranking of the models in Appendix A.3. We find that for all languages except Telugu, we get identical rankings of all models. Additionally, we also measure the Percentage Agreement (PA) between the human and LLM-evaluator, details of which can be found in the Appendix A.1 and find it to be consistently higher than 0.7 on average across all languages and models. This shows the reliability of our LLM-based evaluation for SEMANTIC SIMILARITY which uses the GT response as a reference.

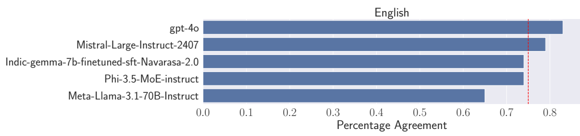

Figure 1: Percentage Agreement between human and LLM-evaluators for English. The red line indicates the average PA across models.

4.3 Qualitative Analysis

One of the authors of the paper performed a qualitative analysis of responses from the evaluated LLMs on 100 selected patient questions. The questions were chosen to cover a range of medical topics and languages. Thematic analysis involved (1) initial familiarization with the queries and associated LLM responses, (2) theme identification where 5 themes were generated and (3) thematic coding where the generated themes were applied to the 100 question-answer pairs. We briefly summarize these results below.

The five generated themes across queries were

[4]The formulation and wording of the metric was slightly simplified to the annotators to better understand it.

Model	AGG	COH	CON	FC	SS
QWEN2.5-72B-INSTRUCT	1.46	1.86	1.96	1.62	1.43
GPT-4	1.40	1.71	1.95	1.56	1.36
PHI-3.5-MOE-INSTRUCT	1.29	1.65	1.93	1.43	1.22
MISTRAL-LARGE-INSTRUCT-2407	1.29	1.60	1.95	1.42	1.24
OPEN-ADITI-HI-V4	1.27	1.69	1.85	1.37	1.22
LLAMAVAAD	1.16	1.34	0.97	1.36	1.20
ARYABHATTA-GEMMAGENZ-VIKAS-MERGED	1.12	1.48	1.65	1.22	1.07
KAN-LLAMA-7B-SFT-V0.5	1.01	1.39	1.64	1.07	0.97
GEMMA-2-27B-IT	1.00	1.28	1.88	1.07	0.91
ARYABHATTA-GEMMAORCA-MERGED	0.97	1.32	1.62	1.03	0.92
LLAMA3-GAJA-HINDI-8B-V0.1	0.91	0.63	1.65	1.09	0.98
GPT-4O	0.91	1.08	1.78	0.98	0.87
AYA-23-35B	0.91	1.09	1.65	1.00	0.83
GAJENDRA-V0.1	0.88	1.21	1.38	0.93	0.85
C4AI-COMMAND-R-PLUS-08-2024	0.82	1.15	1.48	0.85	0.74
TAMIL-LLAMA-7B-INSTRUCT-V0.2	0.81	1.13	1.50	0.83	0.75
AIRAVATA	0.80	1.03	1.38	0.85	0.78
AMBARI-7B-INSTRUCT-V0.2	0.73	0.86	1.11	0.76	0.82
META-LLAMA-3.1-70B-INSTRUCT	0.65	0.55	1.12	0.77	0.67
TELUGU-LLAMA2-7B-V0-INSTRUCT	0.51	0.60	1.12	0.53	0.53
LLAMA38BGENZ_VIKAS-MERGED	0.51	0.52	1.09	0.55	0.53
INDIC-GEMMA-7B-FINETUNED-SFT-NAVARASA-2.0	0.35	0.32	0.53	0.40	0.39
ARYABHATTA-GEMMAULTRA-MERGED	0.32	0.38	1.19	0.31	0.27
TELUGU-LLAMA-7B-INSTRUCT-V0.1	0.04	0.00	0.58	0.03	0.00

Table 2: Metric-wise scores for English. The Proprietary , Open-Weights and Indic models are highlighted appropriately. All Indic models are open-weights.

(1) misspelling of English words, (2) code-mixing, (3) non-native English, (4) relevance to cultural context and (5) specificity to the patient's condition.

For queries that involve misspellings (such as "*saving*" and "*sarjere*" mentioned in Section 3.1), many evaluated LLM were not able to come up with an appropriate response. For the query with the word "*saving*", responses varied from "*The patient should not be saved for more than 15 days after the surgery*" to "*Saving should not be done after surgery*" to "*You should not strain to pass motion for 15 days after the surgery. If you are constipated, it is recommended to consult the doctor*". All of these responses deviate from the GPT-4 generated GT, which said "*You can have a shave after the cataract surgery. However, you should avoid having a head bath or shampoo for 15 days post-surgery.*"

In cases of code mixing and Indian English, LLMs were more robust in their responses than to misspellings. The term "*Kanna operation*" was well understood by most models, and Hinglish queries were responded to in a manner that mirrored the GT answer. However, less common terms were not understood when mixed with English. The query with the word "*Karwat*" as mentioned in 3.1 received responses ranging from "*you can start cooking after 1 week*" to "*I'm sorry, but I cannot provide an answer to your question. The information you are seeking is not relevant to cataract surgery or the guidelines for post-operative care*" to "*be careful when children get near you*". Most of the evaluated LLMs understood the use of "*sugar*" in reference to diabetes, as well as sentences following different syntax than would be common in native English.

The responses for culturally-relevant questions varied greatly between evaluated LLMs. For example, to the question on appropriateness of chapati and puri on the day of surgery, some LLMs approved, saying "*Yes, he can take chapati, Puri etc on the day of cataract surgery*" while others were against this, saying "*You should have a light meal before the surgery. Avoid heavy or oily foods like chapati and Puri on the day of your cataract* *surgery. It's best to stick to easily digestible foods. If you have any specific dietary concerns, please discuss them with your healthcare team*". Questions relating to returning to a "*native place*" were met with refusals by around half of the evaluated LLMs.

Questions that were specific to the patient's condition were also responded to in a diverse manner by the evaluated LLMs. For example, the query "*Can aztolet20 (atorvastatin and clopidogrel) tablet be taken post surgery*" had the GT response "*I do not know the answer to your question. If this needs to be answered by a doctor, please schedule a consultation*" as there was no mention of this medication in the KB. However, some LLMs approved it's use, responding "*Yes, you can take the aztolet20 (atorvastatin and clopidogrel) tablet post cataract surgery. However, it is important to follow your doctor's prescription and instructions*" while others responded with mentions of medication that was referred to in the KB, "*If you are referring to the IMOL tablet, it is usually taken when you experience pain. However, for other medications, please share the name so I can provide a more accurate answer. Always remember to follow your doctor's prescription*." Around half refused to answer the question, mirroring the GT.

5 Discussion

In this study, we evaluated 24 models on healthcare-related queries in the RAG setting. Our findings revealed many insights which we share below:

Difference in model scores We find that the models that we evaluate vary widely in their scores. This indicates that not all models are suitable for use in the healthcare setting, and we find that some models perform worse than expected. For example, GPT-4O and META-LLAMA-3.1-70B-INSTRUCT perform worse than smaller models on this task.

English vs. Multilingual Queries Although the number of non-English queries is small, we find that some Indic models perform better on English queries than non-English queries. We also observe that the Factual Correctness score is lower for non-English queries than English queries on average, indicating that models find it difficult to answer non-English queries accurately. This may be due to the cultural and linguistic nuances present in our queries.

Multilingual vs. Indic models We evaluate several models that are specifically fine-tuned on Indic languages and on Indic data and observe that they do not always perform well on non-English queries. This could be because several instruction tuned models are tuned on synthetic instruction data which is usually a translation of English instruction data. A notable exception is the AYA-23-35B model, that contains manually created instruction tuning data for different languages and performs well for Hindi. Additionally, several multilingual instruction tuning datasets have short instructions, which may not be suitable for complex RAG settings, which typically have longer prompts and large chunks of data.

Human vs. LLM-based evaluation We conduct human evaluation on a subset of models and data points and observe strong alignment with the LLM evaluator overall, especially regarding the final ranking of the models. However, for certain models like MISTRAL-LARGE-INSTRUCT-2407 (for Telugu) and META-LLAMA-3.1-70B-INSTRUCT (for other languages), the agreement is low. It is important to note that we use LLM-evaluators both with and without references, and assess human agreement for SEMANTIC SIMILARITY which uses ground truth references. This suggests that LLM-evaluators should be used cautiously in a multilingual context, and we plan to broaden human evaluation to include more metrics in future work.

Evaluation in controlled settings with uncontaminated datasets We evaluate 24 models in an identical setting, leading to a fair comparison between models. Our dataset is curated based on questions from users of an application and is not contaminated in the training dataset of any of the models we evaluate, lending credibility to the results and insights we gather.

Locally-grounded, non-translated datasets Our dataset includes various instances of code-switching, Indian English colloquialisms, and culturally specific questions which cannot be obtained by translating datasets, particularly with automated translations. While models were able to handle code-switching to a certain extent, responses varied greatly to culturally-relevant questions. This underscores the importance of collecting datasets from target populations while building models or systems for real-world use.

6 Limitations

Our work is subject to several limitations.

- Because our dataset is derived from actual users of a healthcare bot, we couldn't regulate the ratio of English to non-English queries. Consequently, the volume of non-English queries in our dataset is significantly lower than that of English queries, meaning the results on non-English queries should not be considered definitive. Similarly, since the HEALTHBOT is available only in four Indian languages, we also could not evaluate on languages beyond these. The scope of our HEALTHBOT setting is currently confined to queries from patients at one hospital in India, resulting in less varied data. We intend to expand this study as HEALTHBOT extends its reach to other parts of the country.

- While we evaluated numerous models in this work, some were excluded from this study for various reasons, such as ease of access. We aim to incorporate more models in future research.

- Research has indicated that LLM-based evaluators tend to prefer their own responses. In our evaluations, we use GPT-4O, and there may be a bias leading to higher scores for the GPT-4O model and other models within the GPT family. Although not investigated in prior research, it is also conceivable that models fine-tuned with synthetic data generated by GPT-4O might receive elevated scores. We urge readers to keep these in mind while interpreting the scores. In future work, we plan to use multiple LLM-evaluators to obtain more robust results.

- Finally, our human evaluation was limited to a subset of models and data, and a single metric due to time and budget constraints. In future work, we plan to incorporate more human evaluation, as well as qualitative analysis of the results.

7 Ethical Considerations

We use the framework by Bender and Friedman (2018) to discuss the ethical considerations for our work.

Institutional Review All aspects of this research were reviewed and approved by the Institutional Review Board of our organization and also approved by KARYA.

Data Our study is conducted in collaboration with KARYA, that pays workers several times the minimum wage in India and provides them with dignified digital work. Workers were paid 15 INR per datapoint for this study. Each datapoint took approximately 4 minutes to evaluate.

Annotator Demographics All annotators were native speakers of the languages that they were evaluating. Other annotator demographics were not collected for this study.

Annotation Guidelines KARYA provided annotation guidelines and training to all workers.

Compute/AI Resources All our experiments were conducted on 4 $\times$ A100 80Gb PCIE GPUs. The API calls to the GPT models were done through the Azure OpenAI service. We also acknowledge the usage of ChatGPT and GitHub CoPilot for building our codebase, and for refining the writing of the paper.

8 Acknowledgements

We thank Aditya Yadavalli, Vivek Seshadri, the Operations team and Annotators from KARYA for the streamlined annotation process. We also extend our graditude to Bhuvan Sachdeva for helping us with the HEALTHBOT deployment, data collection and organization process.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume*

1: Long Papers), pages 2598–2637, Mexico City, Mexico. Association for Computational Linguistics.

Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2024. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1771–1800, Mexico City, Mexico. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17754–17762.

Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. 2024. Retrieval-augmented generation in multilingual settings. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 177–188, Bangkok, Thailand. Association for Computational Linguistics.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, et al. 2024. Culturalbench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms. *arXiv preprint arXiv:2410.02677*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Moumita Das, Federica Angeli, Anja J. S. M. Krumeich, and Onno C. P. van Schayck. 2018. The gendered experience with respect to health-seeking behaviour in an urban slum of kolkata, india - international journal for equity in health.

Sumanth Doddapaneni, Mohammed Safi Ur Rahman Khan, Sshubam Verma, and Mitesh M. Khapra. 2024. Finding blind spots in evaluator llms with interpretable checklists. *arXiv preprint arXiv: 2406.13439*.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. Dialectbench: A nlp benchmark for dialects, varieties, and closely-related languages. *arXiv preprint arXiv:2403.11009*.

Rishav Hada, Varun Gumma, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024a. METAL: Towards multilingual meta-evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2280–2298, Mexico City, Mexico. Association for Computational Linguistics.

Rishav Hada, Varun Gumma, Adrian Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2024b. Are large language model-based evaluators the solution to scaling up multilingual evaluation? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1051–1070, St. Julian's, Malta. Association for Computational Linguistics.

Jacob Islary. 2018. Health and health seeking behaviour among tribal communities in india: A socio-cultural perspective.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. In *Proceedings of the ACM on Web Conference 2024*, pages 2627–2638.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024a. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024b. Prometheus 2: An open source language model specialized in evaluating other language models. *Preprint*, arXiv:2405.01535.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models. *Preprint*, arXiv:2309.06085.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Yang Liu, Meng Xu, Shuo Wang, Liner Yang, Haoyu Wang, Zhenghao Liu, Cunliang Kong, Yun Chen, Yang Liu, Maosong Sun, and Erhong Yang. 2024. Omgeval: An open multilingual generative evaluation benchmark for large language models. *Preprint*, arXiv:2402.13524.

Ritwik Mishra, Rajiv Ratan Shah, Pushpendra Singh, Jasmeet Kaur, and Simranjeet Singh. 2023. [link].

Kun-Peng Ning, Shuo Yang, Yu-Yang Liu, Jia-Yu Yao, Zhen-Hui Liu, Yu Wang, Ming Pang, and Li Yuan. 2024. Pico: Peer review in llms based on the consistency optimization. *arXiv preprint arXiv: 2402.01830*.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. 2024. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations*.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1946–1958.

Pragnya Ramjee, Bhuvan Sachdeva, Satvik Golechha, Shreyas Kulkarni, Geeta Fulari, Kaushik Murali, and Mohit Jain. 2024. Cataractbot: An llm-powered expert-in-the-loop chatbot for cataract patients.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Alireza Salemi and Hamed Zamani. 2024. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2395–2400, New York, NY, USA. Association for Computing Machinery.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv: 2206.04615*.

Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *Preprint*, arXiv:2401.15391.

Hua Wang, Sneha Gupta, Arvind Singhal, Poonam Muttreja, Sanghamitra Singh, Poorva Sharma, and Alice Piterova. An artificial intelligence chatbot for young people's sexual and reproductive health in india (snehai): Instrumental case study.

Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. 2024. Pariksha : A large-scale investigation of human-llm evaluator agreement on multilingual and multi-cultural data. *arXiv preprint arXiv: 2406.15053*.

Ziang Xiao, Q. Vera Liao, Michelle Zhou, Tyrone Grandison, and Yunyao Li. 2023. Powering an ai chatbot with expert sourcing to support credible health information access.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024a. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2024b. Improving retrieval-augmented generation in medicine with iterative follow-up questions. *arXiv preprint arXiv:2408.00727*.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. 2024. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv: 2406.04244*.

Deepika Yadav, Prerna Malik, Kirti Dabas, Pushpendra Singh, Delhi Deepika YadavIndraprastha Institute of Information Technology, Delhi Prerna MalikIndraprastha Institute of Information Technology, Delhi Kirti DabasIndraprastha Institute of Information Technology, and Delhi Pushpendra SinghIndraprastha Institute of Information Technology. 2019. Feedpal: Understanding opportunities for chatbots in breastfeeding education of women in india.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024. Evaluation of retrieval-augmented generation: A survey. *arXiv preprint arXiv: 2405.07437*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Appendix

A.1 LLM-evaluator scores for non-English languages

A.2 Prompts

A.3 Comparison of human and LLM-evaluator ranking

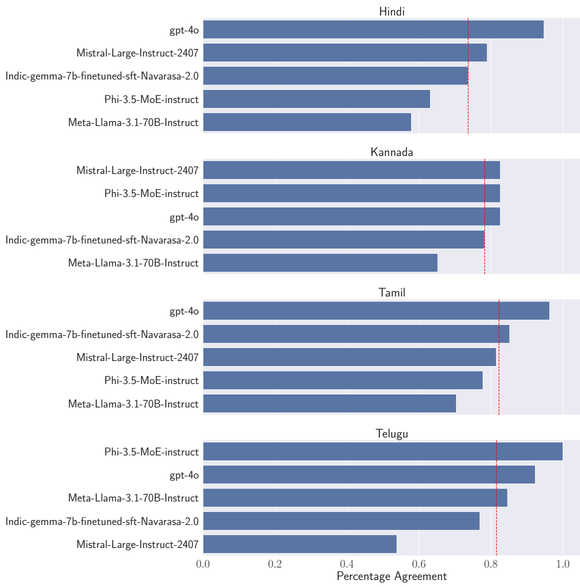

Figure 2: Percentage agreement between human and LLM-evaluators for Indic languages

```
- You are an cataract chatbot whose primary goal is to help patients
undergoing or undergone a cataract surgery.
- If the query can be truthfully and factually answered using the
knowledge base only, answer it concisely in a polite and professional
way. If not, then just say: ‘I do not know the answer to your
question. If this needs to be answered by a doctor, please schedule
a consultation.‘
- Incase of a conflict between raw knowledge base and new knowledge
base, always prefer the new knowledge base, and the latest source
in the new knowledge base. Note that, either the raw knowledge base
or the new knowledge base can be empty.
- The provided query is in {query_lang}, and you must always respond
in {response_lang}.
- Do not generate any other opening or closing statements or remarks.
```

Figure 3: System Prompt for Generation

```
- You are a helpful, unbiased **evaluator** that judges the quality
of the response generated by the model given a query, relevant
knowledge base chunks, ground-truth reference, and a metric to
evaluate the response. Note that, not all the information will be
provided to you in every case, and you must evaluate the response
based only on the information provided to you.
- The metric will be always provided to you in a **JSON** format,
and you have to use that metric to evaluate the response. You
**MUST NOT** change or digress from the metric provided to you. -
In each case, you **MUST ALWAYS** prioritize the knowledge from
the new/updated knowledge base over the raw knowledge base.
- **IF** a reference ground truth is provided, you **MUST** take
it as the most optimal response and evaluate the response based on
the metric provided to you.
- In all cases, the knowledge base will serve as the **ONLY**
knowledge source for you to generate the response, and you **MUST
NEVER** use any of your internal knowledge to evaluate the response
for factuality and information retrieval.

- Your output **MUST** be a **JSON** dictionary with the following
keys:

    - Score: The score of the response based on the metric provided
      to you. The score should be an integer value from 0 to 2,
      as mentioned in the metric.

    - Justification: A brief justification (in English) of
      the score you have assigned the response. Your
      justification **MUST** always reference the relevant pieces
      from the answer, query, and knowledge base chunks for
      interpretability.
```

Figure 4: System prompt for evaluation

Model	AGG	COH	CON	FC	SS
GPT-4	1.21	1.74	1.79	1.26	1.16
QWEN2.5-72B-INSTRUCT	1.20	1.89	1.95	1.21	1.11
MISTRAL-LARGE-INSTRUCT-2407	1.18	1.53	1.79	1.26	1.16
GEMMA-2-27B-IT	0.93	1.11	1.89	1.05	0.79
AYA-23-35B	0.92	0.95	1.79	1.05	0.84
ARYABHATTA-GEMMAGENZ-VIKAS-MERGED	0.89	1.11	1.32	1.00	0.84
PHI-3.5-MOE-INSTRUCT	0.81	1.11	1.74	0.79	0.79
GPT-4O	0.76	0.74	1.79	0.84	0.74
ARYABHATTA-GEMMAORCA-MERGED	0.64	1.00	1.21	0.58	0.68
AIRAVATA	0.63	0.84	1.26	0.68	0.53
LLAMA3-GAJA-HINDI-8B-V0.1	0.60	0.79	1.26	0.63	0.53
OPEN-ADITI-HI-V4	0.56	0.89	1.00	0.47	0.63
LLAMAVAAD	0.55	0.47	0.21	0.68	0.63
C4AI-COMMAND-R-PLUS-08-2024	0.52	0.95	1.47	0.47	0.37
META-LLAMA-3.1-70B-INSTRUCT	0.48	0.47	1.16	0.53	0.47
GAJENDRA-V0.1	0.38	0.47	0.68	0.37	0.42
LLAMA38BGENZ_VIKAS-MERGED	0.32	0.21	1.00	0.32	0.37
ARYABHATTA-GEMMAULTRA-MERGED	0.31	0.37	1.00	0.32	0.26
INDIC-GEMMA-7B-FINETUNED-SFT-NAVARASA-2.0	0.24	0.11	0.53	0.26	0.32

Table 3: Metric-wise scores for Hindi

Model	AGG	COH	CON	FC	SS
QWEN2.5-72B-INSTRUCT	1.29	1.87	1.96	1.35	1.22
GPT-4	1.18	1.78	1.96	1.30	0.91
MISTRAL-LARGE-INSTRUCT-2407	1.09	1.39	1.96	1.22	0.96
GEMMA-2-27B-IT	0.92	1.30	1.91	1.04	0.65
GPT-4O	0.88	0.96	2.00	1.00	0.74
ARYABHATTA-GEMMAORCA-MERGED	0.51	0.57	1.13	0.52	0.52
META-LLAMA-3.1-70B-INSTRUCT	0.48	0.43	0.78	0.57	0.48
KAN-LLAMA-7B-SFT-V0.5	0.47	0.52	1.04	0.48	0.48
LLAMA38BGENZ_VIKAS-MERGED	0.47	0.52	1.00	0.43	0.57
INDIC-GEMMA-7B-FINETUNED-SFT-NAVARASA-2.0	0.24	0.35	0.39	0.26	0.22
PHI-3.5-MOE-INSTRUCT	0.20	0.26	1.22	0.17	0.09
ARYABHATTA-GEMMAULTRA-MERGED	0.13	0.17	0.70	0.09	0.13
AMBARI-7B-INSTRUCT-V0.2	0.05	0.04	0.13	0.04	0.09

Table 4: Metric-wise scores for Kannada

Model	AGG	COH	CON	FC	SS
QWEN2.5-72B-INSTRUCT	1.29	1.87	1.96	1.35	1.22
GPT-4	1.18	1.78	1.96	1.30	0.91
MISTRAL-LARGE-INSTRUCT-2407	1.09	1.39	1.96	1.22	0.96
GEMMA-2-27B-IT	0.92	1.30	1.91	1.04	0.65
GPT-4O	0.88	0.96	2.00	1.00	0.74
ARYABHATTA-GEMMAORCA-MERGED	0.51	0.57	1.13	0.52	0.52
META-LLAMA-3.1-70B-INSTRUCT	0.48	0.43	0.78	0.57	0.48
KAN-LLAMA-7B-SFT-V0.5	0.47	0.52	1.04	0.48	0.48
LLAMA38BGENZ_VIKAS-MERGED	0.47	0.52	1.00	0.43	0.57
INDIC-GEMMA-7B-FINETUNED-SFT-NAVARASA-2.0	0.24	0.35	0.39	0.26	0.22
PHI-3.5-MOE-INSTRUCT	0.20	0.26	1.22	0.17	0.09
ARYABHATTA-GEMMAULTRA-MERGED	0.13	0.17	0.70	0.09	0.13
AMBARI-7B-INSTRUCT-V0.2	0.05	0.04	0.13	0.04	0.09

Table 5: Metric-wise scores for Tamil

Model	AGG	COH	CON	FC	SS
GPT-4	1.14	1.64	2.00	1.29	0.86
QWEN2.5-72B-INSTRUCT	1.11	1.57	1.71	1.29	0.86
MISTRAL-LARGE-INSTRUCT-2407	1.03	1.36	2.00	1.14	0.86
GEMMA-2-27B-IT	0.91	1.21	2.00	1.00	0.71
META-LLAMA-3.1-70B-INSTRUCT	0.61	0.43	1.00	0.79	0.57
GPT-4O	0.54	0.57	1.86	0.57	0.43
PHI-3.5-MOE-INSTRUCT	0.44	0.57	1.86	0.43	0.29
LLAMA38BGENZ_VIKAS-MERGED	0.33	0.14	1.50	0.36	0.29
ARYABHATTA-GEMMAORCA-MERGED	0.29	0.29	0.93	0.29	0.29
ARYABHATTA-GEMMAULTRA-MERGED	0.26	0.29	1.71	0.21	0.14
INDIC-GEMMA-7B-FINETUNED-SFT-NAVARASA-2.0	0.19	0.29	0.57	0.21	0.07
TELUGU-LLAMA-7B-INSTRUCT-V0.1	0.09	0.00	1.71	0.00	0.00
TELUGU-LLAMA2-7B-V0-INSTRUCT	0.00	0.00	0.00	0.00	0.00

Table 6: Metric-wise scores for Telugu

```
"name": "Coherence",
"description": "Coherence assesses the logical flow of the response, ensuring that one idea leads smoothly to the next. A coherent response should
present information in a structured manner, making it easy for the reader to follow the thought process without confusion.",

"scoring": {
    "0": {
        "(a)": "The response is highly disorganized and lacks a clear structure, making it difficult to follow.",
        "(b)": "Sentences or ideas appear out of order or are disconnected, resulting in a confusing or jarring reading experience.",
        "(c)": "The overall message is unclear due to poor organization."
    },
    "1": {
        "(a)": "The response has some structure but includes noticeable breaks in the logical flow.",
        "(b)": "Transitions between ideas may be abrupt, or there may be gaps in the reasoning, forcing the reader to make extra effort
        to follow along.",
        "(c)": "While the main point is evident, the flow is inconsistent."
    },
    "2": {
        "(a)": "The response is well-organized and flows logically from one idea to the next.",
        "(b)": "Each point builds naturally on the previous one, creating a clear and cohesive narrative.",
        "(c)": "The reader can easily follow the thought process without having to backtrack or piece together disjointed information."
    }
}
```

Figure 5: Metric description: COHERENCE

```
"name": "Conciseness",
"description": "This metric evaluates how effectively the response conveys its message without unnecessary repetition or extraneous details. A concise response is brief yet comprehensive, avoiding long-winded explanations and focusing on the core message. However, it must not sacrifice clarity or completeness in the pursuit of brevity.",

"scoring": {
    "0": {
        "(a)": "The response is overly verbose, including repeated information, irrelevant details, or excessive explanations.",
        "(b)": "It takes far longer than necessary to convey the intended message, making it inefficient and difficult to read."
    },
    "1": {
        "(a)": "The response is somewhat concise but includes some unnecessary information or redundant points.",
        "(b)": "While the main message is clear, the response could be made more efficient by removing repetition or streamlining explanations."
    },
    "2": {
        "(a)": "The response is highly concise, delivering all relevant information in a brief and efficient manner.",
        "(b)": "There is no repetition, and every sentence serves a clear purpose.",
        "(c)": "The message is conveyed succinctly, without sacrificing clarity or detail."
    }
}
```

Figure 6: Metric description: CONCISENESS

```
"name": "Factual Accuracy",
"description": "This metric assesses the factual correctness of the response, focusing on whether the information provided aligns with verified facts from the ground-truth answer and the available knowledge base. It evaluates both numerical and phrase-based facts, ensuring that key factual elements such as data points, dates, and specific terminology are accurate and verifiable. The evaluation emphasizes the accuracy of important details that are crucial for the validity of the response.",

"scoring": {
    "0": {
        "(a)": "The response contains one or more significant factual errors.",
        "(b)": "Key facts, numbers, or data points are incorrect, misleading, or fabricated, and the response does not align with the ground-truth or the knowledge base.",
        "(c)": "The factual inaccuracies could lead to misunderstandings or incorrect conclusions."
    },
    "1": {
        "(a)": "The response is partially accurate but contains minor factual inaccuracies or omissions.",
        "(b)": "While the majority of facts are correct, some important details may be misstated or missing.",
        "(c)": "The response captures the general truth but lacks precision or completeness in key factual areas."
    },
    "2": {
        "(a)": "The response is factually accurate, with all critical facts, figures, and details aligned with the ground-truth answer and knowledge base.",
        "(b)": "There are no factual errors, and the information is presented with precision and correctness, making the response highly reliable."
    }
}
```

Figure 7: Metric description: FACTUAL CORRECTNESS

```
"name": "Semantic Similarity",
"description": "This metric assesses the core meaning and factual alignment between the prediction and ground-truth. It evaluates whether critical information such as factual knowledge, numbers, and key phrases match, prioritizing factual accuracy and the alignment of essential concepts over stylistic or surface-level similarities.",

"scoring": {
    "0": {
        "(a)": "he prediction does not align with the ground truth in terms of key facts, numbers, or critical phrases.",
        "(b)": "The core meaning of the prediction diverges entirely from the ground-truth.",
        "(c)": "The differences would lead to misunderstandings or incorrect conclusions about the core message."
    },
    "1": {
        "(a)": "The prediction contains some similarities to the ground truth, with some key facts, numbers, and phrases being correctly aligned.",
        "(b)": "However, the prediction is missing some information or contains some added information.",
        "(c)": "This causes the prediction to fail at encapsulating the entire core meaning present in the ground truth."
    },
    "2": {
        "(a)": "The prediction is semantically similar to the ground-truth, with key facts, numbers, and phrases correctly aligned.",
        "(b)": "Any differences are minor and do not significantly alter the core meaning or factual accuracy.",
        "(c)": "The essential message of the prediction matches that of the ground-truth."
    },
}
```

Figure 8: Metric description: SEMANTIC SIMILARITY

Language	Human Ranking	LLM Ranking
English	PHI-3.5-MOE-INSTRUCT (1.30), MISTRAL-LARGE-INSTRUCT-2407 (1.28) GPT-4O (0.90), META-LLAMA-3.1-70B-INSTRUCT (0.88), INDIC-GEMMA-7B-FINETUNED-SFT-NAVARASA-2.0 (0.62)	PHI-3.5-MOE-INSTRUCT (1.22), MISTRAL-LARGE-INSTRUCT-2407 (1.14), GPT-4O (0.87), META-LLAMA-3.1-70B-INSTRUCT (0.61), INDIC-GEMMA-7B-FINETUNED-SFT-NAVARASA-2.0 (0.41)
Hindi	MISTRAL-LARGE-INSTRUCT-2407 (1.21), PHI-3.5-MOE-INSTRUCT (0.95), GPT-4O (0.68), META-LLAMA-3.1-70B-INSTRUCT (0.58), INDIC-GEMMA-7B-FINETUNED-SFT-NAVARASA-2.0 (0.53)	MISTRAL-LARGE-INSTRUCT-2407 (1.16), PHI-3.5-MOE-INSTRUCT (0.79), GPT-4O (0.74), META-LLAMA-3.1-70B-INSTRUCT (0.47), INDIC-GEMMA-7B-FINETUNED-SFT-NAVARASA-2.0 (0.32)
Kannada	MISTRAL-LARGE-INSTRUCT-2407 (0.96), GPT-4O (0.91), META-LLAMA-3.1-70B-INSTRUCT (0.74), INDIC-GEMMA-7B-FINETUNED-SFT-NAVARASA-2.0 (0.35), PHI-3.5-MOE-INSTRUCT (0.17)	MISTRAL-LARGE-INSTRUCT-2407 (0.96), GPT-4O (0.74), META-LLAMA-3.1-70B-INSTRUCT (0.48), INDIC-GEMMA-7B-FINETUNED-SFT-NAVARASA-2.0 (0.22), PHI-3.5-MOE-INSTRUCT (0.09)
Tamil	MISTRAL-LARGE-INSTRUCT-2407 (1.37), GPT-4O (1.07), PHI-3.5-MOE-INSTRUCT (1.04), META-LLAMA-3.1-70B-INSTRUCT (0.48), INDIC-GEMMA-7B-FINETUNED-SFT-NAVARASA-2.0 (0.19)	MISTRAL-LARGE-INSTRUCT-2407 (1.26), GPT-4O (1.04), PHI-3.5-MOE-INSTRUCT (0.96), META-LLAMA-3.1-70B-INSTRUCT (0.48), INDIC-GEMMA-7B-FINETUNED-SFT-NAVARASA-2.0 (0.19)
Telugu	MISTRAL-LARGE-INSTRUCT-2407 (1.31), META-LLAMA-3.1-70B-INSTRUCT (0.62), INDIC-GEMMA-7B-FINETUNED-SFT-NAVARASA-2.0 (0.38), GPT-4O (0.38), PHI-3.5-MOE-INSTRUCT (0.15)	MISTRAL-LARGE-INSTRUCT-2407 (0.77), META-LLAMA-3.1-70B-INSTRUCT (0.46), GPT-4O (0.31), PHI-3.5-MOE-INSTRUCT (0.15), INDIC-GEMMA-7B-FINETUNED-SFT-NAVARASA-2.0 (0.08)

Table 7: Human and LLM ranking according to the direct assessment. The value in the bracket denotes the average score of the metric SEMANTIC SIMILARITY which was used for the evaluation.