"*Knowing When You Don't Know*": A Multilingual Relevance Assessment Dataset for Robust Retrieval-Augmented Generation

Nandan Thakur[1], **Luiz Bonifacio**[1,3], **Xinyu Zhang**[1], **Odunayo Ogundepo**[1], **Ehsan Kamalloo**[1], **David Alfonso-Hermelo**[2], **Xiaoguang Li**[2], **Qun Liu**[2], **Boxing Chen**[2], **Mehdi Rezagholizadeh**[2], **Jimmy Lin**[1]

[1] David R. Cheriton School of Computer Science, University of Waterloo, Canada
[2] Huawei Noah's Ark Lab [3] FEEC-Unicamp, Brazil

Abstract

Retrieval-Augmented Generation (RAG) grounds Large Language Model (LLM) output by leveraging external knowledge sources to reduce factual hallucinations. However, prior work lacks a comprehensive evaluation of different language families, making it challenging to evaluate LLM robustness against errors in external retrieved knowledge. To overcome this, we establish **NoMIRACL**, a human-annotated dataset for evaluating LLM robustness in RAG across 18 typologically diverse languages. NoMIRACL includes both a non-relevant and a relevant subset. Queries in the non-relevant subset contain passages judged as non-relevant, whereas queries in the relevant subset include at least a single judged relevant passage. We measure relevance assessment using: (i) *hallucination rate*, measuring model tendency to hallucinate, when the answer is not present in passages in the non-relevant subset, and (ii) *error rate*, measuring model inaccuracy to recognize relevant passages in the relevant subset. In our work, we observe that most models struggle to balance the two capacities. Models such as LLAMA-2 and Orca-2 achieve over 88% hallucination rate on the non-relevant subset. Mistral and LLAMA-3 hallucinate less but can achieve up to a 74.9% error rate on the relevant subset. Overall, GPT-4 is observed to provide the best tradeoff on both subsets, highlighting future work necessary to improve LLM robustness. NoMIRACL dataset and evaluation code are available at: https://github.com/project-miracl/nomiracl.

1 Introduction

Retrieval-Augmented Generation (RAG) (Guu et al., 2020; Lewis et al., 2020; Izacard and Grave, 2021; Borgeaud et al., 2022) is a promising way to incorporate external knowledge via a first-stage retrieval system. RAG instills information from reliable knowledge corpora (provided as external passages) to generate accurate and faithful responses (Shuster et al., 2021; Gao et al., 2023b; Liu et al., 2024). Ever since the advent of Large Language Models (LLMs), such as GPT-3 (Brown et al., 2020) or LLAMA-2 (Touvron et al., 2023), they are the de-facto choice for answer generation in RAG, due to their unprecedented progress in text generation and understanding (Brown et al., 2020; Li et al., 2024; Chang et al., 2024; Guo et al., 2023). RAG grounds the LLM-generated answer, thereby avoiding previously observed factual hallucination (Maynez et al., 2020; Raunak et al., 2021) and outdated knowledge (Cao et al., 2021; He et al., 2023) in LLMs.

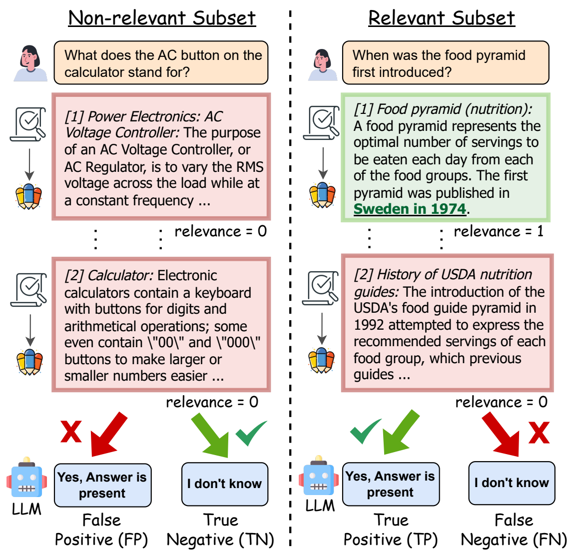

Figure 1: LLM robustness evaluation as a binary tree in NoMIRACL. When dealing with queries in the non-relevant subset, the LLM is expected to disregard all noisy passages and refrain from answering (TN). Conversely, for queries in the relevant subset, the LLM should recognize the relevant passage and provide a valid answer (TP).

A challenging issue in RAG is to provide robust and reliable LLM-generated answers. The answer generation stage is dependent on the first-

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 12508–12526
November 12-16, 2024

stage information retrieval system. The retrieval system poses challenges in accurately retrieving relevant information when evaluated either on zero-shot domains (Thakur et al., 2021) or low-resource languages (Zhang et al., 2023). The incorrect or non-relevant information contained in retrieved passages can frequently mislead the LLM to hallucinate (Adlakha et al., 2024; Chen et al., 2024; Shi et al., 2023; Yoran et al., 2024; Yu et al., 2023). Prior work lacks a comprehensive evaluation of LLM reasoning capabilities in multiple languages. As a result, it remains unclear to which extent LLMs hallucinate across both high- or low-resource languages.

To facilitate research in this direction, we present **NoMIRACL**, a large multilingual human-annotated dataset containing over 56,000 (including both non-relevant and relevant samples) to evaluate LLM robustness against errors in first-stage external information, i.e., retrieved passages, across 18 typologically diverse languages. To construct the dataset, we hired 31 native speakers as human annotators (Zhang et al., 2023). NoMIRACL contains two subsets, non-relevant and relevant. The non-relevant subset contains all queries with no known answers, i.e., all top-k retrieved passages manually judged as non-relevant. Conversely, the relevant subset contains queries with known answers, i.e., at least one of the top-k passages is manually judged as relevant.

To better understand the LLM robustness in NoMIRACL, we conduct experiments with several existing powerful and multilingual-focused LLMs (*e.g.*, GPT-4, Mistral, LLAMA-3). We conduct our experiments using the top-k oracle passages retrieved using a hybrid retrieval system from a language-specific Wikipedia corpus (Zhang et al., 2023). We use a zero-shot "vanilla" prompt template for prompting all LLMs. Our key findings are: First, LLMs such as LLAMA-2, Aya-101, and Orca-2 observe a surprisingly high 88% hallucination rate on the non-relevant subset. Second, the Mistral and LLAMA-3 series of models hallucinate less but perform worse on the relevant subset. Overall, GPT-4 is found to provide the optimal performance tradeoff across both subsets.

To understand our experimental findings better, we conduct an empirical analysis on NoMIRACL (en) to analyze the blind spots in a subset of LLMs. We observe LLAMA-2-7B and 13B interestingly repeat the query and prompt instructions on average by at least 25%. Mistral-7B always provides a rationale in their output generation by over 88%. In addition, we conduct different prompting techniques and observe that supervised fine-tuning LLMs on the NoMIRACL development set can be tricky.

Pred. / Subset	Non-Relevant	Relevant
LLM: Yes, answer is present	FP ✗	TP ✓
LLM: I don't know	TN ✓	FN ✗

Hallucination rate = FP / (FP + TN) ¦ Error rate = FN / (FN + TP)

Figure 2: Confusion matrix for robustness evaluation with NoMIRACL. More details are provided in (§2.1); (Subset) denotes the ground-truth in NoMIRACL; (Pred.) denotes the LLM output prediction.

To summarize, our contributions are: (i) We introduce NoMIRACL, a novel multilingual dataset to evaluate LLM hallucinations against first-stage retrieval errors in RAG. (ii) We evaluate several powerful multilingual LLMs and observe challenges in LLM robustness by often hallucinating an answer within non-answerable passages in the non-relevant subset and the inability to recognize relevant passages in the relevant subset. (iii) We conduct thorough manual inspections on LLM's generation results, and find several hallucination patterns for each genre of the LLM; We hope NoMIRACL can serve as a valuable dataset towards a much-needed LLM robustness evaluation.

2 Background and Problem Identification

A challenging issue in RAG is to provide robust and reliable LLM-generated output against a first-stage information retrieval system. Overreliance of the LLM on the content of the retrieved passages (i.e., tendency to extract information from passages) can be limiting when passages are noisy or non-relevant (Yu et al., 2023; Yoran et al., 2024; Shi et al., 2023).

RAG Background. Retrieval-augmented generation (RAG) (Lewis et al., 2020; Guu et al., 2020) involves a two-stage inference pipeline. In the first stage, given the retrieval system and the user query, the retrieval system provides the subset of top-k passages retrieved from an external data corpus C. For the next stage, the user query with the top-k retrieved passages is provided to the LLM, which generates an output summarizing the answer for the query and citing the relevant passages.

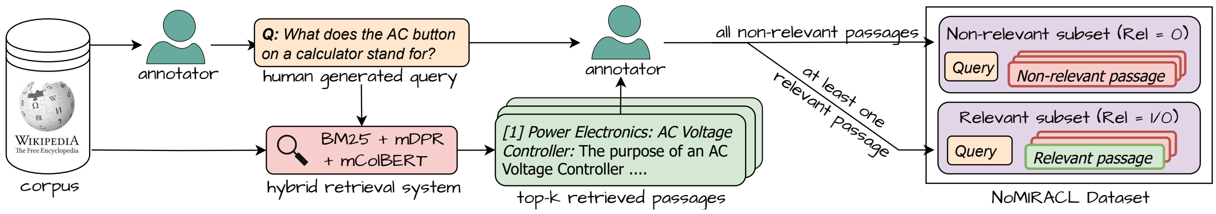

Figure 3: An overview of the data construction procedure (for English) involved in NoMIRACL.

2.1 Robustness Evaluation

We conduct our evaluation strategy as a contingency table (as shown in Figure 2) to robustly evaluate the LLM behavior in both answerable and non-answerable scenarios using a binary classification task, by comparing LLM predictions against the ground truth provided by human annotators.

Definitions. NoMIRACL contains two subsets, we denote them as either non-relevant (F) or relevant (T). The non-relevant subset contains queries with no-known answers, i.e., all top-k passages are non-relevant, while the relevant subset contains queries with known answers, i.e., at least one of the top-k passages is relevant. Similarly, we denote the LLM prediction as either positive (P) indicating the model finds the passage relevant to answer the query and similarly negative (N) denotes the model does not find any passage relevant (i.e., containing the answer) for the query.

Confusion Matrix. In our confusion matrix (cf. Figure 2), for our non-relevant subset, True Negative (TN) denotes when the model correctly predicts queries with no-known answers using the non-relevant retrieved passages, whereas False Positive (FP) denotes when the model prediction is incorrect on the non-relevant subset. Similarly, True Positive (TP) denotes when LLM correctly predicts queries with known answers within the retrieved passages, whereas False Negative (FN) denotes when the model prediction is incorrect (i.e., the model finds no answer) on the relevant subset.

Evaluation Metrics. Following prior works (Adlakha et al., 2024; Chen et al., 2024), we assess LLM robustness as a binary classification task using two metrics: (i) *hallucination rate* and (ii) *error rate*. First, we compute the hallucination rate (in %) $= \mathrm{FP}/(\mathrm{FP}+\mathrm{TN})$ which measures the model's tendency to hallucinate an answer, when no answer is available in all of the passages in the non-relevant subset. Next, we measure the error rate (in %) $= \mathrm{FN}/(\mathrm{FN}+\mathrm{TP})$ which measures the model's ability to identify the answer present within the passages in the relevant subset.

3 NoMIRACL Dataset

As the goal of NoMIRACL is to understand to which extent LLMs tend to hallucinate across different languages, our dataset contains 18 diverse languages with a myriad of both correct or answerable queries (relevant) subset and hallucinated or unanswerable queries (non-relevant). We describe our dataset construction procedure in (§3.1), fold creation in (§3.2) and languages covered and dataset usage in (§3.3). An overview of our data construction procedure is shown in Figure 3.

3.1 Data Construction Procedure

NoMIRACL is constructed using the same procedure utilized to develop MIRACL (Zhang et al., 2023). The data construction occurs in two stages, following (Zhang et al., 2023; Clark et al., 2020). In the first stage, the annotator (a native language speaker) writes a well-formed query for each individual prompt text. Each prompt is a short text snippet containing the first 100 words from a language-specific Wikipedia corpus. Next, for each human-generated query, top-k passages are retrieved from the corpus using a hybrid multilingual retrieval system (more details in §4.1). In the second stage, annotators assess the binary relevance judgment of the top-k query–passage pairs, either relevant (relevance = 1) or non-relevant (relevance = 0). For additional details in data construction, such as quality control, we would like to refer the reader to Appendix C.

Non-relevant Subset. Annotators generate queries based on certain dataset guidelines, however, occasionally the human-generated queries cannot be answered with the external corpus, which leads to the scenario where none of the top-k passages are relevant, i.e., none contains the answer. These queries with unknown answers may occur due to the following reasons: (i) queries can be either generic or specific for information to be present in Wikipedia, for e.g. "*What does the AC button on a calculator stand for?*" retrieves the Wikipedia

Split / ISO	ar	bn	de	en	es	fa	fi	fr	hi	id	ja	ko	ru	sw	te	th	yo	zh	Total
***Non-relevant Subset**: Queries with all human-judged non-relevant passages*																			
Development	228	495	171	289	245	760	98	1,016	1,016	474	211	1,577	268	508	480	323	1,678	1,085	10,922
Test	291	630	218	367	311	968	125	1,294	1294	603	269	2,006	342	646	610	412	2,136	1,381	13,903
***Relevant Subset**: Queries with at least one human-judged relevant passage*																			
Development	2,896	411	305	799	648	632	1,271	343	350	960	860	213	1,252	482	828	733	119	393	13,495
Test	1,405	1,130	712	1,790	1,515	1,476	801	711	819	611	1,141	1,417	718	465	793	650	663	920	17,737

Table 1: Dataset Statistics for NoMIRACL. The dataset contains two subsets for all 18 languages: (i) Non-relevant subset, where queries contain all human-judged non-relevant passages. (ii) Relevant subset, where queries contain at least one relevant human-judged passage. Both subsets are split into disjoint development and test splits.

page on Calculator,[1] but it does not contain information about the AC button; (ii) spelling mistakes in query generation. We construct the non-relevant subset with queries with all top-k passages judged as non-relevant, i.e., with a relevance of 0.

Relevant Subset. Queries with known answers, i.e. at least one of top-k retrieved passages marked by the annotator as relevant to provide sufficient information to answer the query. We construct the NoMIRACL relevant subset with queries with at least one relevant passage, i.e., all query-passage pairs have been judged as either relevant with a relevance of 1 or non-relevant with relevance of 0.

3.2 Fold Creation

In NoMIRACL, we split the non-relevant and relevant subsets to form disjoint development and test splits. Detailed statistics can be found in Table 1.

Development Split. For queries present in the relevant subset, we reuse the queries from the MIRACL development split (Zhang et al., 2023) for all 18 languages. For queries in the non-relevant subset, we randomly sample a disjoint set containing 44% of the queries from the whole non-relevant subset.

Test Split. For queries present in the relevant subset, we reuse the queries from the MIRACL test-B split (Zhang et al., 2023) for all 18 languages.[2] For queries in the non-relevant subset, we utilize the other disjoint set, containing 56% of the queries from the whole non-relevant subset.

3.3 Languages Covered and Dataset Usage

NoMIRACL covers 18 diverse typological languages (Zhang et al., 2023).[3] The languages along with their ISO codes are: Arabic (ar), Bengali (bn), German (de), English (en), Spanish (es), Persian (fa), Finnish (fi), French (fr), Hindi (hi), Indonesian (id), Japanese (ja), Korean (ko), Russian (ru), Swahili (sw), Thai (th), Yoruba (yo), Chinese (zh).

From Table 1, we observe an uneven amount of queries present in each language. To avoid this non-uniformity and budget constraints (see subsection 4.3), in our experiments, we limit the maximum number of 250 queries for each language and subset (if available) in NoMIRACL.

4 Experimental Setup

4.1 Evaluation Setup and Metrics

In NoMIRACL, we assess LLM relevance as either hallucination or error, using an input query, a vanilla prompting technique, and top-k (oracle) retrieved and relevance judged passages.

Retrieved Passages. For each query in NoMIRACL, a maximum of $k = 10$ passages are retrieved and judged by our annotators. We follow the hybrid retrieval setup in Zhang et al. (2023), which includes three different multilingual retriever models: (i) BM25 (Robertson and Zaragoza, 2009), a lexical retriever, previously shown to be robust across domains and languages (Thakur et al., 2021; Zhang et al., 2022). We use the BM25 implementation available in Anserini (Yang et al., 2018) with default parameters ($k_1 = 0.9$ and $b = 0.4$) and the corresponding language-specific analyzer. (ii) mDPR (Karpukhin et al., 2020), a dense retriever, using mBERT (Devlin et al., 2018) as the backbone and fine-tuned on MS MARCO with the Tevatron toolkit (Gao et al., 2023a). (iii) mColBERT (Khattab and Zaharia, 2020), a multi-vector retriever, fine-tuned following Khattab and Zaharia (2020) using mBERT as backbone and fine-tuned on MS MARCO. The top-k passages are ranked using an ensemble fusion by normalizing and averaging each model score within the range of $[0, 1]$.

Evaluation Objective. In our work, we evaluate LLM relevance as a response string y in a binary classification setup. Following prior evaluation

[1] https://en.wikipedia.org/wiki/Calculator

[2] We left out the MIRACL test-A split, as it contains queries for only 10 out of the 18 languages available.

[3] NoMIRACL covers 10 families (from Niger-Congo to Indo-European) and 11 scripts (from Latin to Devanagari) covering diversity from the perspective of linguistic characteristics.

strategies in (Adlakha et al., 2024; Yu et al., 2023), we use the input query q_i, a vanilla prompt template Q, and a set of top-k annotated passages P_k. We prompt the LLM to evaluate if q_i can be answered using any passage in P_k. The LLM generates an answer output containing either y = "`I don't know`" as negative (N) or "`Yes, answer is present`" as positive (P). The output is tagged "`Invalid`" if it does not fall in either one of the above. Recall from (§2.1), we calculate the *hallucination rate* (in %) = $\text{FP}/(\text{FP}+\text{TN})$, which measures the error in rejecting information from non-relevant passages and the *error rate* (in %) = $\text{FN}/(\text{FN}+\text{TP})$, which measures the error in identifying relevant passages amongst noisy ones.

4.2 Evaluation Models

We evaluate eleven state-of-the-art LLMs with a strong focus of multilingual instruction capabilities, including both open and closed-sourced. All model checkpoints can be found in Table 5.

(1) OpenAI: We include three closed-book LLM variants: GPT-3.5-turbo, GPT-4, and GPT-4o (OpenAI, 2023) using the Azure OpenAI service. **(2) Mistral:** We include two variants: (i) `Mistral-7B-Instruct-v0.3`, the latest 7B instruction-tuned parameter model (Jiang et al., 2023) and (ii) `Mixtral-8x7B-v0.1`, a sparse Mixture-of-Expert (MoE) model (Jiang et al., 2024). **(3) Orca-2:** In the Orca-2 series (Mitra et al., 2023), we include both `Orca-2-7B` and `Orca-2-13B`. **(4) Aya**: `Aya-101` (Üstün et al., 2024) is a recently introduced multilingual LLM containing 13B parameters and trained with 101 languages and `Aya-23-35B` finetuned across 23 languages (Aryabumi et al., 2024). **(5) LLAMA-2:** In the LLAMA-2 series (Touvron et al., 2023), we include three chat variants: `Llama-2-7b-chat-hf`, `Llama-2-13b-chat-hf`, and `Llama-2-70b-chat-hf` instruction tuned chat models. **(6) LLAMA-3:** Following up on the LLAMA-2 series, we include both the instruction tuned models (Dubey et al., 2024): `Meta-Llama-3-8B-Instruct` and `Meta-Llama-3-70B-Instruct`.

4.3 Experimental Settings

We execute the generation of GPT-4o, GPT-4, and GPT-3.5-turbo, using the OpenAI service (API version `2023-05-15`) deployed on Microsoft Azure[4]

```
I will give you a question and several
contexts containing information about the
question. Read the contexts carefully. If
any of the contexts answers the question,
respond as either "Yes, answer is present"
or "I don't know":

QUESTION: {query}

CONTEXTS:
[1] {Passage title}: {Passage text}
[2] {Passage title}: {Passage text}
...
[10] {Passage title}: {Passage text}

OUTPUT:
```

Figure 4: Vanilla zero-shot prompt template used in our experiments for LLM hallucination evaluation for all 18 languages in NoMIRACL. The instruction is provided in English, similar to Ahuja et al. (2023).

and LLAMA-3 series using the AnyScale API service. We maintain a maximum input sequence length of 4096 tokens for a fair evaluation amongst all models. We set a low-temperature score = 0.1 for a deterministic output, and a top-p sampling ratio = 0.95. We output a maximum of 50 tokens.

Vanilla Prompting. The choice of prompt significantly influences the performance and LLMs have been shown brittle to prompting variations, training examples, or long context setups (Liu et al., 2024). In our work, we evaluate all baselines using a *zero-shot monolingual listwise* prompting strategy. We construct a vanilla prompt template using all top-k (oracle) passages (available in NoMIRACL) as a list of contexts along with the input query, both in the same language. We evaluate all LLMs zero-shot, as we cannot fit few-shot exemplars due to insufficient context length (maximum of 4096 sequence length). Our template provides a short description in English describing the task (Ahuja et al., 2023). Our vanilla prompt template used in our experiments is shown in Figure 4.

Reducing Costs. Running GPT-4 is expensive and LLAMA-3-70B is rather slow at inference. For long contexts and low-resource languages, the costs can even multiply. To limit, we did not exceed our prompt above 4096 tokens. To effectively fit all $k = 10$ passages within the vanilla zero-shot prompt, we truncate each passage to use the first 375 tokens. Next, due to budget constraints, we keep our evaluation to a maximum of 250 randomly sampled queries for all languages in both NoMIR-

[4]We compared Azure API with OpenAI API across four languages in NoMIRACL and observed no noticeable difference between different GPT-4 API version providers.

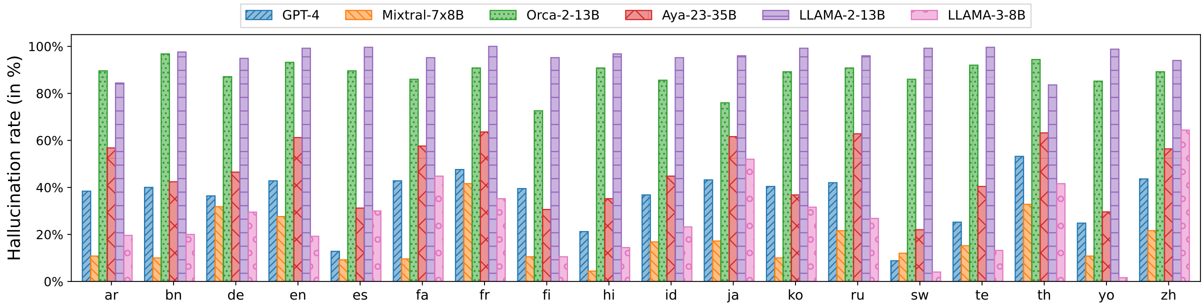

Figure 5: Hallucination rate (in %) = FP/(FP + TN) on the non-relevant subset (F) in NoMIRACL test split. The non-relevant subset contains queries with no known answers, i.e., all top-k (where $k = 10$) passages are judged by a human annotator as non-relevant. A majority of LLMs (except Mistral) hallucinate on the non-relevant subset. Lower the hallucination rate is better. The best model in each category is plotted (see Figure 8 for all models).

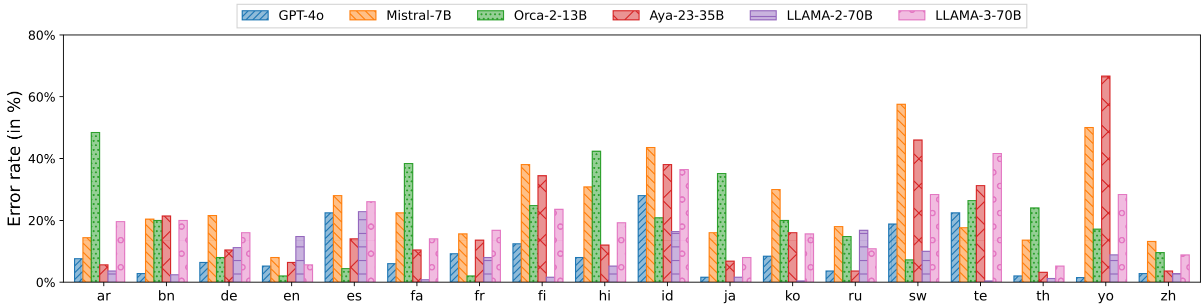

Figure 6: Error rate (in %) = FN/(FN + TP) on the relevant subset (T) in NoMIRACL test split. The relevant subset contains queries with known answers, i.e., at least one of the top-k (where $k = 10$) passages are judged by a human annotator as relevant. On average, a majority of LLMs (except Mistral and Aya-101) have a lower error rate by accurately identifying the answer. Lower the error rate is better. The best model in each category is plotted (see Figure 9 for all models).

ACL relevant and non-relevant split. We end up providing $\approx$ 20K API calls producing an expense of $1,474 (in USD) including miscellaneous costs.

5 Experimental Results

We discuss our LLM robustness evaluation results using the hallucination rate on the non-relevant subset in (§5.1) and using the error rate on the relevant subset in (§5.2), and compare overall both the relevant and non-relevant capacities in (§5.3).

5.1 NoMIRACL Non-relevant Subset

Figure 5 shows hallucination rates on the NoMIRACL non-relevant subset for a maximum of 250 queries evaluated (each language) on all 18 languages for best LLM in each category (for all model results, please refer to Figure 8). Our findings indicate that all LLMs (except Mistral) hallucinate that an answer is present across all languages, thereby indicating their poor ability to abstain from answering. On average, the lowest hallucination rate of 17.4% is observed by Mixtral-7x8B, followed by LLAMA-3-8B-Instruct with 26.8%. GPT-4 achieves a 35.5% hallucination rate, which highlights the challenge of identifying non-relevant passages. LLAMA-2, Orca-2, and Aya-101 perform much worse on average across all languages, by achieving a hallucination rate of more than 80%. We hypothesize that LLMs perform poorly to identify non-relevant passages as they are highly similar to the query, but do not contain the exact answer.

Overall, the lowest hallucination rates are observed in Swahili and Yoruba. We hypothesize that queries in low-resource languages (smaller Wikipedia corpus) contain retrieved information, likely to be non-relevant (easier negative), thereby making it easier for the LLM to judge as "I don't know". GPT-3.5 (cf. Figure 8) is observed with the highest deviation across languages with a hallucination rate as low as 25.2% on Swahili (sw) to 95.2% on Bengali (bn). Overall, all LLMs are found to perform poorly on NoMIRACL non-relevant subset, indicating our dataset is very challenging in robustness evaluation for LLMs.

5.2 NoMIRACL Relevant Subset

Figure 6 shows error rates on NoMIRACL relevant subset for a maximum of 250 queries (each language) on 18 languages for best LLM in each category Our findings indicate that all LLMs (except Mistral) identify the answer present within the relevant passage. On average, Aya-101 achieves the highest error rate of 62.5%. The lowest error rates are observed by LLAMA-2-70B and GPT-4o which are lower than 10%. Overall, Aya-101, Mixtral-7x8B, and LLAMA-3-8B perform worse on average by observing more than a 40% error rate. Overall, LLMs (except Mistral and Aya-101) perform well and do not suffer from errors in identifying answers in the NoMIRACL relevant subset.

5.3 NoMIRACL Overall Comparison

A robust LLM should be able to identify the answer captured within retrieved passages in the relevant subset and abstain from answering when none of the retrieved passages contain the answer in the non-relevant subset. To measure performance across both dimensions in NoMIRACL, we plot the average model accuracy across both the non-relevant (x-axis) and relevant subset (y-axis) for all tested models in Figure 7. Overall, LLMs positioned in the top-right corner provide an optimal performance on both subsets. A majority of LLMs (such as LLAMA-2, and Orca-2) in the top-left corner perform well on the relevant subset, however, hallucinate and struggle to perform well on the non-relevant subset, indicating their inability to accurately judge non-relevant passages.

On the other hand, Mistral and LLAMA-3 suffer less from hallucination on the non-relevant subset but observe a higher error rate (over 40%) on the relevant subset, indicating they are not confident in identifying passages containing the answer. Aya-101 is unable to perform well in either of the subsets. GPT-4 provides a good tradeoff balancing both a low hallucination and error rate on NoMIRACL relevant and non-relevant subsets, however is expensive to compute at scale for inference.

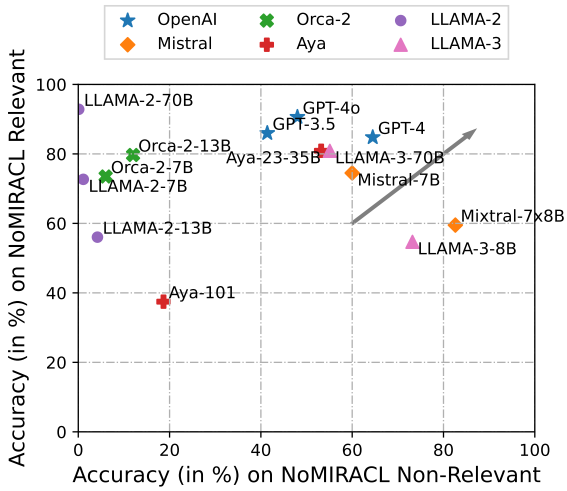

Figure 7: Plot measuring average model accuracy across all languages on relevant (y-axis) and non-relevant subset (x-axis) in NoMIRACL. Performance towards the top-right corner (denoted by the arrow) is better.

Non-relevant Subset. As shown in Table 2 (above), we observe a uniform distribution of the hallucination pattern of failed samples for a majority of LLMs by answering “Yes, answer is present” with or without additional explanation. LLMs such as LLAMA-2-7B and LLAMA-2-13B suffer from hallucinations by often repeating the question or instruction in their generation output. Mistral-7B interestingly always provides a rationale or explanation in their model response, whereas Aya-101, uses implicit memory heavily to directly provide an answer instead of grounding the answer from within the retrieved passages. Lastly, models such as Orca-2-7B tend to change the output generation style and often use synonyms such as “No, answer is not present” instead of “I don’t know”.

Relevant Subset. As shown in Table 2 (below), similar to the non-relevant subset, we observe a uniform distribution of accurate LLM responses. Interestingly, GPT-4 and Orca-2-13B overall only provide a single output classification token, whereas models such as GPT-3.5 or Mixtral-8x7B provide an additional rationale or explanation. Similar to the non-relevant subset, LLAMA-2-7B and 13B models repeat the instruction in their output and Aya-101 sometimes uses implicit memory.

6 Empirical Analysis

In this section, we conduct an empirical analysis of LLM outputs, with both the non-relevant subset (containing hallucinations) and the relevant subset (containing errors). We conduct our ablation study on the English (en) subset in NoMIRACL. We categorize each LLM output pattern as either positive or accurate (highlighted in green), unable to understand instruction (highlighted in orange), or either a hallucination or error (highlighted in red).

7 Further Studies

Prompt Optimization. Prompting is crucial in handling the robustness evaluation of multilingual-focused LLMs. Techniques such as Chain-of-Thought (CoT) (Wei et al., 2022) or algorithmi-

	OpenAI		Mistral		Orca-2		Aya	LLAMA-2		
	GPT-4	GPT-3.5	8x7B	7B	13B	7B	Aya-101	70B	13B	7B
Empirical results on the non-relevant subset: *Queries with all human-judged non-relevant passages*										
(i) Perfectly answers "I don't know"	56.8%	54.8%	67.2%	1.6%	3.2%	9.2%	15.6%	0.0%	0.0%	0.8%
(ii) "I don't know" with explanation	0.0%	0.0%	3.6%	88.4%	0.0%	2.0%	0.0%	0.0%	0.0%	0.0%
(iii) Uses a synonym of "I don't know"	0.4%	0.0%	1.6%	1.2%	6.0%	7.6%	0.0%	0.4%	0.0%	2.4%
(iv) Refuses to answer	0.4%	0.8%	0.0%	0.8%	3.2%	0.0%	6.4%	0.0%	0.0%	0.0%
(v) Repeats question or instruction	0.0%	0.0%	0.0%	0.0%	0.0%	3.2%	0.0%	2.0%	64.8%	30.8%
(vi) Conversation	0.0%	0.0%	0.4%	0.0%	0.0%	0.4%	0.0%	0.0%	0.0%	0.4%
(vii) Answers "Yes" w. or w.o. explanation	42.4%	44.4%	27.2%	8.0%	87.6%	68.4%	67.2%	97.6%	34.8%	65.6%
(viii) Uses implicit memory to answer	0.0%	0.0%	0.0%	0.0%	0.0%	8.8%	10.8%	0.0%	0.0%	0.0%
Empirical results on the relevant subset: *Queries with at least one human-judged relevant passage*										
(i) Perfectly answers "Yes, answer is present"	94.8%	45.2%	22.8%	5.2%	97.2%	71.2%	51.6%	1.2%	3.6%	7.2%
(ii) "Yes, answer is present" with explanation	0.0%	46.0%	56.0%	31.6%	0.8%	11.2%	0.0%	98.8%	46.8%	63.2%
(iii) Refuses to answer	0.4%	0.8%	0.0%	0.0%	0.0%	0.0%	1.2%	0.0%	0.0%	0.0%
(iv) Repeats question or instruction	0.0%	0.0%	0.4%	0.0%	0.0%	0.4%	0.0%	0.0%	48.8%	24.8%
(v) Conversation	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.6%
(vi) Answers "No" w. or w.o. explanation	4.8%	8.0%	20.8%	60.8%	0.8%	2.8%	4.0%	0.0%	0.0%	1.2%
(vii) Uses implicit memory to answer	0.0%	0.0%	0.0%	2.4%	1.2%	14.4%	42.0%	0.0%	0.0%	0.4%

Table 2: Empirical results on the complete NoMIRACL English (en) non-relevant (above) and relevant (below) subsets. The analysis is bracketed into three categories, where green category denotes an accurate response, orange denotes limitations in understanding the instruction and red denotes model hallucination or error respectively.

cally optimizing prompts using DSPy (Khattab et al., 2023) highlight the necessity of prompt optimization. Although optimizing for the prompt is certainly challenging and expensive to evaluate all LLMs across 18 languages relevant and non-relevant subsets, we experiment with three listwise variations techniques inspired by Thomas et al. (2024). The prompt template changes are listed in Figure 10: (i) **role**, we highlight the role of LLM as an evaluator within the prompt at the beginning, (ii) **repeat**, we repeat the task instructions at the end of the prompt to remind the LLM, and (iii) **explanation**, we ask the LLM model to provide a step-by-step explanation and then answer and require 400 output tokens to fit both the LLM reasoning and the answer.

We evaluate Mistral-7B with three prompt variations independently on NoMIRACL. The complete results are listed in Table 3. On average, both role and repeat techniques help reduce the error rate in the NoMIRACL relevant subset by 6.3% and 15.2% but overall increase the hallucination rate by 8.7% and 15.9% respectively. On the other hand, prompting with explanation decreases the hallucination rate by 9.7% but increases the error rate by 8.3%. These results show that prompting is user dependent, the user will be required to choose their technique depending on whether they wish to be better on the non-relevant subset by reducing the hallucination rate or the relevant subset by reducing the error rate.

Fine-tuning on NoMIRACL. In this section, following prior works such as Chain-of-Verification (CoVe; Dhuliawala et al., 2024) or Chain-of-Noting (CoN; Yu et al., 2023), we investigate the following research question: *Does fine-tuning on the NoMIRACL development set help increase robustness?*

We experiment with two open-sourced LLMs: Mistral-7B and LLAMA-3 (8B). We Supervised Fine-Tune (SFT) LoRA adapters (Hu et al., 2022) on the development set of NoMIRACL for all 18 languages (randomly sampled 90% train, 10% development) using 4-A6000 GPUs each containing 48GB RAM with PEFT.[5] Our hyperparameter settings are listed in Table 6. We were unable to fine-tune larger models (greater than 8B parameters) due to computational budget restrictions.

As shown in Table 4, we observe LLAMA-3 (8B) to be quite unstable after SFT. Fine-tuning helps to reduce the error rate of LLAMA-3 (8B) (an improvement of 10.6%) but can hurt its performance on the hallucination rate (drop up to 17.9%). For a few languages mentioned in Table 7 such as Arabic (ar) the LLM always outputs "Yes, answer is present", whereas for Bengali (bn) heavily relies on "I don't know". On the other hand, SFT deteriorates Mistral-7B on both relevant and non-relevant datasets. Overall, we demonstrate SFT is tricky and careful experimentation is required to achieve the best out of fine-tuning on the NoMIRACL development subset for a binary classification task output ("Yes, answer is present" or "I don't know").

8 Related Work

Retrieval-Augmented Generation. The knowledge stored in a large language model (LLM) is commonly outdated (He et al., 2023), and prone

[5]https://github.com/huggingface/alignment-handbook

	ar	bn	de	en	es	fa	fr	fi	hi	id	ja	ko	ru	sw	te	th	yo	zh	Avg.
Hallucination Rates (in %) on NoMIRACL test split (non-relevant subset)																			
Original	40.0	63.2	38.2	42.8	17.2	52.4	47.6	16.1	39.6	30.8	44.4	28.8	41.6	14.8	74.0	58.8	23.6	45.2	40.0
(+ Role)	35.6	60.0	53.5	62.4	40.4	38.0	71.2	29.8	38.4	49.2	59.2	41.2	55.6	21.2	72.0	54.4	32.4	62.8	48.7
(+ Repeat)	47.2	72.4	56.7	50.8	35.2	69.2	70.0	50.0	49.6	48.8	65.2	48.4	56.0	35.2	78.4	72.0	43.6	58.4	55.9
(+ Explanation)	28.0	33.6	30.4	34.8	16.8	39.6	42.4	31.5	18.0	27.2	32.8	26.8	31.6	22.8	36.4	34.0	27.2	31.6	30.3
Error Rates (in %) on NoMIRACL test split (relevant subset)																			
Original	14.4	20.4	21.6	8.0	28.0	22.4	15.6	38.0	30.8	43.6	16.0	30.0	18.0	57.6	17.6	13.6	50.0	13.2	25.5
(+ Role)	15.2	23.6	10.0	3.2	14.4	23.2	7.2	29.2	25.2	32.8	9.6	21.6	10.8	47.2	17.6	17.2	33.8	4.4	19.2
(+ Repeat)	8.4	12.4	9.2	5.2	12.4	4.8	5.6	6.0	16.4	16.8	1.6	12.0	6.8	22.8	14.0	6.0	21.6	3.6	10.3
(+ Explanation)	28.8	39.6	26.0	16.8	35.6	35.6	20.0	32.0	46.8	46.8	24.4	33.6	24.8	50.4	48.8	31.2	48.0	18.8	33.8

Table 3: Hallucination and error rates on the NoMIRACL test split (non-relevant and relevant subsets) with three types of prompting techniques on Mistral-7B (v0.3). The changes in the prompt template are listed in Figure 10.

Model	**w/o SFT**	**w/ SFT**
Non-Relevant Subset: Hallucination Rates (in %)		
Meta-Llama-3-8B-Instruct	**26.8**	44.7 (– 17.9)
Mistral-Instruct-7B-v0.3	**40.0**	44.3 (– 4.3)
Relevant Subset: Error Rates (in %)		
Meta-Llama-3-8B-Instruct	45.3	**34.7** (+ 10.6)
Mistral-Instruct-7B-v0.3	**25.5**	46.1 (– 20.6)

Table 4: Supervised fine-tuning on the NoMIRACL development split with Llama-3 (8B) and Mistral-7B (v0.3) LLMs.

to hallucinations by generating factually incorrect output (Maynez et al., 2020; Raunak et al., 2021). By grounding on external knowledge, a retrieval-augmented LLM can generate better and more trustworthy output (Guu et al., 2020; Lewis et al., 2020; Izacard and Grave, 2021; Borgeaud et al., 2022). Retrieval-augmented generation has achieved remarkable results in various tasks such as open-domain question answering (ODQA) (Lewis et al., 2020; Izacard and Grave, 2021; Trivedi et al., 2023), argument extraction (Du and Ji, 2022) and code generation (Zhou et al., 2023). Real-world products such as Bing Search and LangChain have incorporated RAG applications.

LLM Evaluation. Prior work explores adding perturbation in passages and shows that LLM performance can be influenced when exposed to different tasks, such as question answering (QA) (Jia and Liang, 2017; Petroni et al., 2020; Creswell et al., 2023), logical reasoning (Misra et al., 2023) or arithmetic reasoning (Shi et al., 2023; Kumar et al., 2021). In examining controllability and robustness, Li et al. (2023) observes that LLMs disregard contextual information, showing that LLM output can be influenced by non-relevant context. Adlakha et al. (2024) observes complementary results from our work, where they observe LLM can be rather faithful when provided non-relevant passages in QA datasets such as NQ (Kwiatkowski et al., 2019). Knowing that prompting LLMs with non-relevant data can result in misguided responses, Yu et al. (2023) recently introduced a new prompting technique, Chain-of-Noting (CON) and Yoran et al. (2024) fine-tuned the LLM explicitly, both aimed to improve LLM robustness in RAG when non-relevant information is provided.

Related Datasets. Datasets focused on addressing unanswerable queries such as SQuAD 2.0 (Rajpurkar et al., 2018) were created adversarially to look similar to datasets with answerable queries. Similarly, Conversational QA datasets such as CoQA (Reddy et al., 2019) and QuAC (Choi et al., 2018) also contain unanswerable queries. A concurrent work proposes RGB, a RAG benchmark to evaluate LLM robustness in English and Chinese (Chen et al., 2024).

9 Conclusion

We introduce NoMIRACL, a multilingual human-labeled dataset for relevance assessment of LLM robustness as a binary relevance identification task in 18 languages. Our multilingual dataset is human-annotated and constructed with 31 native speakers. We provide two subsets in NoMIRACL, the non-relevant subset, where queries contain all judged non-relevant passages, and the relevant subset, where queries contain at least one relevant judged passage to measure the hallucination on the non-relevant and error on the relevant subset. Our experimental results indicate that existing LLMs are not robust, as we observe challenges in LLM robustness in either hallucination or error. GPT-4 achieves the best model and performance tradeoff across both subsets. NoMIRACL can facilitate research in understanding to which extent LLMs tend to hallucinate, ultimately paving the way for building more effective and robust multilingual-focused LLMs in the future.

10 Limitations

NoMIRACL is not perfect and like other datasets have limitations. We describe our limitations below and keep it as future work to improve our dataset.

1. Human Errors in Dataset Construction. Our dataset has been fully constructed using humans, thereby it may contain human errors. We conducted additional quality checks on a subset of the NoMIRACL dataset to validate its question quality and relevance judgment as explained in Appendix C.

2. Evaluation Setup. In our work, we evaluate whether a passage is relevant or non-relevant for a given query, instead of evaluating actual answer spans. Reliable and accurate answers for a given query require domain experts as annotators. Annotators can potentially highlight short extractive spans of answers within relevant passages, however, non-extractive queries can either contain multiple answers or a long-form answer, making it difficult to highlight a relevant answer span. Therefore, for NoMIRACL, we focus on evaluating top-k passages as information contexts, which are judged for their relevancy by a data annotator.

3. Limited to Wikipedia. NoMIRACL is currently developed using language-specific Wikipedia as the corpora. Wikipedia may not be the ideal choice for real-world applications across languages. For example, the English BEIR benchmark (Thakur et al., 2021) includes diversity within its domains (all English) and contains more real-world domains such as Medical, etc. However, we keep it as future work to extend NoMIRACL to diverse domains for the following reasons: (i) *scarcity of corpora across languages*: for low-resource languages such as Bengali or Yoruba, finding a suitable large enough text corpora is difficult with limited choices. (ii) *no uniformity across domains*: certain European languages have more legal domain corpora available, whereas news articles for African languages. This will introduce non-uniformity in information across languages. (iii) *limited budget*: constructing NoMIRACL was expensive involving several annotators involved for about 4–6 months. Extending to more domains would require additional budgets and human effort to be able to implement.

Acknowledgements

We would like to thank our annotators, without whom NoMIRACL could not have been built. We would also like to thank Akintunde Oladipo for providing the necessary Microsoft Azure credits for evaluating OpenAI models. This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada, a gift from Huawei, and Cloud TPU support from Google's TPU Research Cloud (TRC).

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating correctness and faithfulness of instruction-following models for question answering. *Trans. Assoc. Comput. Linguistics*, 12:681–699.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Uttama 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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4232–4267. Association for Computational Linguistics.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan N. Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *CoRR*, abs/2405.15032.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33:*

Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6491–6506. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3):39:1–39:45.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17754–17762. AAAI Press.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Jonathan H. Clark, Jennimaria Palomaki, Vitaly Nikolaev, Eunsol Choi, Dan Garrette, Michael Collins, and Tom Kwiatkowski. 2020. Tydi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Trans. Assoc. Comput. Linguistics*, 8:454–470.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2023. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 3563–3578. Association for Computational Linguistics.

Xinya Du and Heng Ji. 2022. Retrieval-augmented generative question answering for event argument extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4649–4666. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023a. Tevatron: An efficient and flexible toolkit for neural retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 3120–3124. ACM.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6465–6488. Association for Computational Linguistics.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is ChatGPT to human experts? comparison corpus, evaluation, and detection. *CoRR*, abs/2301.07597.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Hangfeng He, Hongming Zhang, and Dan Roth. 2023. Rethinking with retrieval: Faithful large language model inference. *CoRR*, abs/2301.00303.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Gautier Izacard and Edouard Grave. 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: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 874–880. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *CoRR*, abs/2401.04088.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. DSPy: Compiling declarative language model calls into self-improving pipelines. *CoRR*, abs/2310.03714.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 39–48. ACM.

Vivek Kumar, Rishabh Maheshwary, and Vikram Pudi. 2021. Adversarial examples for evaluating math word problem solvers. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2705–2712. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix X. Yu, and Sanjiv Kumar. 2023. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1774–1793. Association for Computational Linguistics.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. Pre-trained language models for text generation: A survey. *ACM Comput. Surv.*, 56(9).

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguistics*, 12:157–173.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July*

5-10, 2020, pages 1906–1919. Association for Computational Linguistics.

Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. COMPS: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2928–2949, Dubrovnik, Croatia. Association for Computational Linguistics.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andrés Codas, Clarisse Simões, Sahaj Agrawal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching small language models how to reason. *CoRR*, abs/2311.11045.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Fabio Petroni, Patrick S. H. Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. In *Conference on Automated Knowledge Base Construction, AKBC 2020, Virtual, June 22-24, 2020*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1172–1183. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3784–3803. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large language models can accurately predict searcher preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 1930–1940. ACM.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10014–10037. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model.

In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15894–15939. Association for Computational Linguistics.

Ellen M. Voorhees. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 315–323. ACM.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Peilin Yang, Hui Fang, and Jimmy Lin. 2018. Anserini: Reproducible ranking baselines using Lucene. *ACM J. Data Inf. Qual.*, 10(4):16:1–16:20.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. 2023. Chain-of-note: Enhancing robustness in retrieval-augmented language models. *CoRR*, abs/2311.09210.

Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. 2022. Towards best practices for training multilingual dense retrieval models. *CoRR*, abs/2204.02363.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Shuyan Zhou, Uri Alon, Frank F. Xu, Zhengbao Jiang, and Graham Neubig. 2023. DocPrompting: Generating code by retrieving the docs. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

A Appendix

The following supplementary sections in the appendix are arranged as follows:

- Appendix B provides information on the NoMIRACL dataset release.
- Appendix C provides additional construction details in NoMIRACL, including corpora preparation and annotator hiring details.
- Appendix D describes steps we took for quality control during the dataset construction.
- Appendix E provides model checkpoints and additional experimental results.

B Details on NoMIRACL Dataset Release

Licensing. The NoMIRACL dataset is based on language-specific Wikipedia. We follow the same license as Wikipedia for NoMIRACL: Creative Commons Attribution-ShareAlike 4.0 Unported License (CC BY-SA 4.0).[6] Overall, the license allows both researchers and industry alike to access the dataset, and allow them to copy and redistribute the dataset for future work.

Examples. A randomly sampled example for each of the non-relevant and relevant subsets of the NoMIRACL dataset for English (`en`) has been provided in Table 8 and Table 9 respectively.

C Additional Data Construction Details

Corpora Preparation. For each NoMIRACL language, we follow the same passage corpora provided in MIRACL (Zhang et al., 2023). Out of the 18 languages, 11 of the existing languages common in Mr. TyDi (Zhang et al., 2022) use the raw Wikipedia dump from early 2019 and the rest of the languages used in MIRACL use a release from March 2022. In MIRACL, all Wikipedia articles are parsed using WikiExtractor[7] and segmented into passages based on natural discourse units using two consecutive newlines in the wiki markup as the delimiter.

Annotator Hiring Details. An important feature of NoMIRACL is that our dataset was *not* constructed via crowd-sourced workers similar to MIRACL (Zhang et al., 2023). We overall hired 31 annotators (both part-time and full-time) across all languages in NoMIRACL. Each annotator was interviewed and evaluated to be a native speaker of their language, based on a carefully constructed onboarding and training process. Overall our hiring process and NoMIRACL data construction in total took around 6 months. We offered annotators the hourly rate of \$18.50 per hour (converted into USD). For reference, the local minimum wage is \$11.50 USD/hr.

[6] `https://creativecommons.org/licenses/by-sa/4.0/`
[7] `https://github.com/attardi/wikiextractor`

D Quality Control

To ensure high data quality, we conduct a manual assessment executed by human reviewers (hired part-time) on a random subset of NoMIRACL annotations, following MIRACL (Zhang et al., 2023). We conducted our quality control in two phases.

Phase I. In this phase, reviewers were given both the prompts and the generated queries and filled up a checklist to determine whether the quality of the queries met our requirements. Criteria include the examination of the query itself (e.g., spelling, syntax, and fluency, etc.) and whether the query could be answered directly by the prompt, which we wanted to avoid to generate more informative queries, following (Clark et al., 2020; Zhang et al., 2023). To evaluate this, we measured the lexical overlap between the queries and their corresponding prompts. We found the overlaps primarily occur in entities or stopwords and thus concluded that the generated queries are reasonably different from the given prompts.

Phase II. In this phase, reviewers were provided the same guidance as annotators performing the relevance assessment. They were asked to label a randomly sampled subset of the query–passage pairs from our annotated batch. The degree of agreement on the overlapping pairs is used to quantify the quality of the relevance labels. Overall, we observe on average agreements of over 80% on query–passage relevance, which is consistent with the IR literature dating back many decades (Voorhees, 1998).

E Checkpoints and Additional Results

All multilingual-focused LLM checkpoints used in our experiments for both closed and open-sourced can be found in Table 5. Hyperparameter choices during NoMIRACL supervised fine-tuning LLMs are listed in Table 6 and experimental results in Table 7. LLM evaluation results for both the non-relevant and relevant subsets for all models can be found in Figure 8 and Figure 9 respectively. Figure 10 shows template changes for prompt optimization ablation experiments, including (i) role, (ii) repeat, and (iii) explanation prompts.

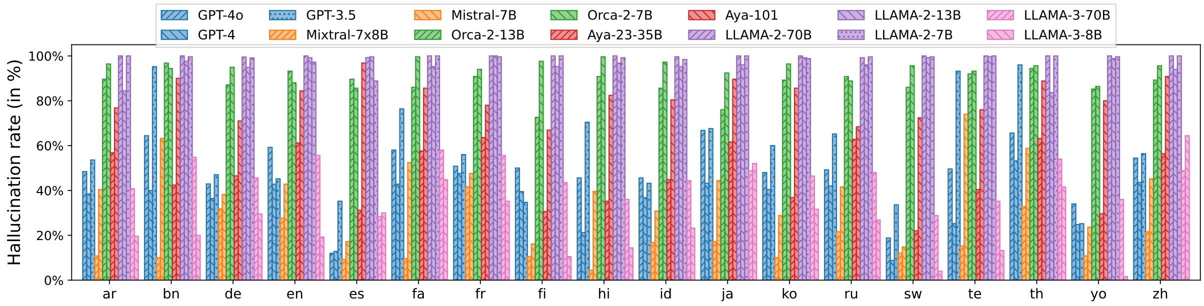

Figure 8: Hallucination rate (in %) = $\text{FP}/(\text{FP} + \text{TN})$ on the non-relevant subset in NoMIRACL test split. The non-relevant subset contains queries with no-known answers, i.e., all top-k (where $k = 10$) passages are judged by a human annotator as non-relevant. On average, most LLMs (except Mistral) hallucinate on the non-relevant subset. Lower the hallucination rate is better.

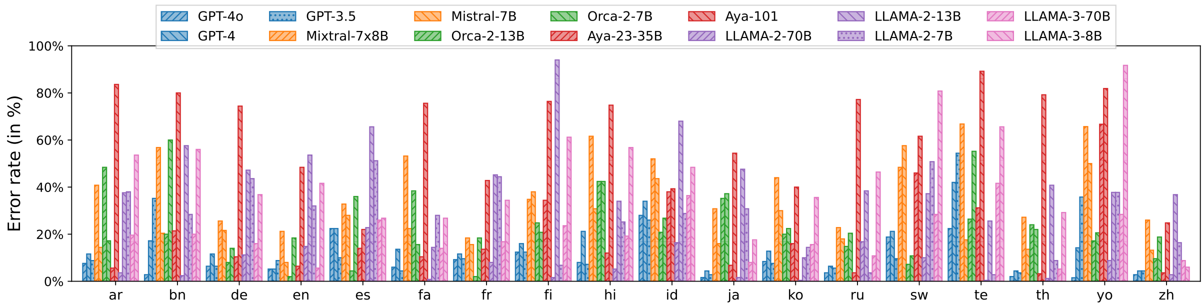

Figure 9: Error rate (in %) = $\text{FN}/(\text{FN} + \text{TP})$ on the relevant subset in NoMIRACL test split. The relevant subset contains queries with known answers, i.e., at least one of the top-k (where $k = 10$) passages judged by a human annotator is relevant. On average, most LLMs (except Mistral and Aya-101) have a lower error rate, i.e., can accurately identify the relevant answer. Lower the error rate is better.

Model	**Model Checkpoints (Link)**
	OpenAI baseline models
GPT-4o	learn.microsoft.com/en-us/azure/ai-services/openai/
GPT-4	learn.microsoft.com/en-us/azure/ai-services/openai/
GPT-3.5	learn.microsoft.com/en-us/azure/ai-services/openai/
	Mistral baseline models
Mixtral-8x7B	huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1
Mistral-7B (v0.3)	huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
	Orca-2 baseline models
Orca-2-13B	huggingface.co/microsoft/Orca-2-13b
Orca-2-7B	huggingface.co/microsoft/Orca-2-7b
	Aya baseline models
Aya-101	huggingface.co/CohereForAI/aya-101
Aya-23-35B	huggingface.co/CohereForAI/aya-23-35B
	LLAMA-2 baseline models
LLAMA-2-70B	huggingface.co/meta-llama/Llama-2-70b-chat-hf
LLAMA-2-13B	huggingface.co/meta-llama/Llama-2-13b-chat-hf
LLAMA-2-7B	huggingface.co/meta-llama/Llama-2-7b-chat-hf
	LLAMA-3 baseline models
LLAMA-3 (70B)	huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct
LLAMA-3 (8B)	huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

Table 5: All models and checkpoint links used for NoMIRACL evaluation.

Hyperparameter	**Value**
`use_peft`	`true`
`torch_dtype`	`bfloat16`
`lora_r`	`64`
`lora_alpha`	`16`
`lora_dropout`	`0.05`
`lora_target_modules`	`{q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj}`
`learning_rate`	`3.0e-06`
`lr_scheduler_type`	`cosine`
`max_seq_length`	`4096`

Table 6: Hyperparameter settings chosen during LoRA supervised fine-tuning (SFT) Mistral-7B (v0.3) and LLAMA-3 (8B) instruct models on the NoMIRACL development split.

```
You are an evaluator checking whether the question contains the answer within the provided
contexts or not. I will give you a question and several contexts containing information about the
question. Read the contexts carefully. If any of the contexts answers the question, respond as
either “Yes, answer is present” or “I don’t know”:

QUESTION: {query}

CONTEXTS:
[1] {Passage title}: {Passage text}
[2] {Passage title}: {Passage text}
...
[10] {Passage title}: {Passage text}

Please remember to read all the contexts carefully.  If any of the contexts answers the
question: {query}, respond as either “Yes, answer is present” or “I don’t know”.

OUTPUT:
```

```
Read the query and the contexts carefully and provide a step-by-step explanation for your answer.
If any of the contexts answers the question, respond as either “Yes, answer is present” or “I
don’t know”. You must strictly follow the output format with ## Reasoning: ... ## Answer: “Yes,
answer is present” OR “I don’t know”.

QUESTION: {query}

CONTEXTS:
[1] {Passage title}: {Passage text}
[2] {Passage title}: {Passage text}
...
[10] {Passage title}: {Passage text}

OUTPUT:
```

Figure 10: All prompt ablations used in our experiments for LLM hallucination evaluation for all 18 languages in NoMIRACL on both the relevant and non-relevant subsets. Role prompt appends the role of the LLM at the beginning of the prompt (highlighted in blue). Repeat prompt highlights the task by repeating instructions at the end of the prompt (highlighted in red). Explanation prompt asks the model to provide a reasoning path and finally answer the question (highlighted in violet).

	ar	bn	de	en	es	fa	fr	fi	hi	id	ja	ko	ru	sw	te	th	yo	zh	Avg.
Hallucination Rates (in %) on NoMIRACL test split (non-relevant subset)																			
Llama-3 (8B)	19.6	20.0	29.5	19.2	30.0	44.8	35.2	10.5	14.4	23.2	52.0	31.6	26.8	4.0	13.2	41.6	1.6	64.4	26.8
Llama-3 (8B) (w/ SFT)	83.0	12.5	36.9	44.0	46.8	65.6	36.9	70.0	61.5	53.1	41.7	9.7	85.4	42.2	0.0	82.7	12.0	20.0	44.7
Mistral-7B (v0.3)	40.4	63.2	38.2	42.8	17.2	52.4	47.6	16.1	39.6	30.8	44.4	28.8	41.6	14.8	74.0	58.8	23.6	45.2	40.0
Mistral-7B (v0.3) (w/ SFT)	46.8	33.2	34.1	73.6	33.6	52.4	48.0	42.7	26.0	59.2	52.0	47.6	62.8	35.6	31.2	40.0	45.2	33.2	44.3
Error Rates (in %) on NoMIRACL test split (relevant subset)																			
Llama-3 (8B)	53.6	56.0	36.8	41.6	26.8	26.8	34.4	61.2	56.8	48.4	17.6	35.6	46.4	80.8	65.6	29.2	91.7	6.0	45.3
Llama-3 (8B) (w/ SFT)	6.0	70.8	32.4	31.2	24.0	14.6	21.2	8.8	14.9	40.6	24.0	77.2	8.3	32.5	87.5	4.2	75.5	50.4	34.7
Mistral-7B (v0.3)	14.4	20.4	21.6	8.0	28.0	22.4	15.6	38.0	30.8	43.6	16.0	30.0	18.0	57.6	17.6	13.6	50.0	13.2	25.5
Mistral-7B (v0.3) (w/ SFT)	46.4	60.8	46.0	12.4	46.8	36.4	35.6	42.0	69.6	37.2	37.6	42.0	25.2	69.6	66.4	56.8	52.0	46.8	46.1

Table 7: Complete SFT results using the NoMIRACL development dataset for LLAMA-3 (8B) and Mistral-7B (v0.3) LLMs across all languages in NoMIRACL. Lower the hallucination and error rates (%) is better.

Query	Judged Passages	Relevance
What is the name of the winner of the Abel Prize of 2022?	[1] **Abel Prize**: The Abel Prize () is a Norwegian prize awarded annually by the King of Norway to one or more outstanding mathematicians. It is named after Norwegian mathematician Niels Henrik Abel (1802–1829) and directly modeled after the Nobel Prizes. It comes with a monetary award of 6 million Norwegian kroner (NOK) (€635,000 or $740,000).	0
	[2] **MIT Department of Mathematics**: The current faculty of around 50 members includes Wolf Prize winner Michael Artin, Abel Prize winner Isadore Singer, and numerical analyst Gilbert Strang.	0
	[3] **Abel Prize**: After interest in the concept of the prize had risen in 2001, a working group was formed to develop a proposal, which was presented to the Prime Minister of Norway in May. In August 2001, the Norwegian government announced that the prize would be awarded beginning in 2002, the two-hundredth anniversary of Abel's birth. Atle Selberg received an honorary Abel Prize in 2002, but the first actual Abel Prize was awarded in 2003.	0
	[4] **Abel Prize**: The prize was first proposed in 1899, to be part of the celebration of the 100th anniversary of Niels Henrik Abel's birth in 1902. Shortly before his death in 1899, the Norwegian mathematician Sophus Lie proposed establishing an Abel Prize when he learned that Alfred Nobel's plans for annual prizes would not include a prize in mathematics. King Oscar II was willing to finance a mathematics prize in 1902, and the mathematicians Ludwig Sylow and Carl Størmer drew up statutes and rules for the proposed prize. However, Lie's influence waned after his death, and the dissolution of the union between Sweden and Norway in 1905 ended the first attempt to create an Abel Prize.	0
	[5] **Eötvös Loránd University**: Eötvös Loránd University (, ELTE) is a Hungarian public research university based in Budapest. Founded in 1635, ELTE is one of the largest and most prestigious public higher education institutions in Hungary. The 28,000 students at ELTE are organized into eight faculties, and into research institutes located throughout Budapest and on the scenic banks of the Danube. ELTE is affiliated with 5 Nobel laureates, as well as winners of the Wolf Prize, Fulkerson Prize and Abel Prize, the latest of which was Abel Prize winner Endre Szemerédi in 2012.	0
	[6] **Abel Prize**: Anyone may submit a nomination for the Abel Prize, however, self-nominations are not permitted. The nominee must be alive; however, if the awardee dies after being declared as the winner, the prize will be awarded posthumously.	0
	[7] **Abel Prize**: The Norwegian Academy of Science and Letters declares the winner of the Abel Prize each March after recommendation by the Abel Committee, which consists of five leading mathematicians. Both Norwegians and non-Norwegians may serve on the Committee. They are elected by the Norwegian Academy of Science and Letters and nominated by the International Mathematical Union and the European Mathematical Society. The committee is of 2018 chaired by Norwegian mathematician Hans Munthe-Kaas (University of Bergen), and was before that, headed by Professor John Rognes.	0
	[8] **Hans Munthe-Kaas**: Munthe-Kaas received Exxon Mobil Award for best PhD at NTNU, 1989, and the Carl-Erik Frõberg Prize in Numerical Mathematics 1996 for the paper Ïie–Butcher theory for Runge–Kutta Methods.̈ Munthe-Kaas is elected member of the Norwegian Academy of Science and Letters, the Royal Norwegian Society of Sciences and Letters and the Norwegian Academy of Technological Sciences. Munthe-Kaas is the chair of the international Abel prize committee (2018-2022), he is President of the Scientific Council of Centre International de Mathématiques Pures et Appliquées (CIMPA) (2017–present) and he is Editor-in-Chief of Journal Foundations of Computational Mathematics (2017–present). Munthe-Kaas was secretary of Foundations of Computational Mathematics (2005–2011) and member of the Board of the Abel Prize in Mathematics (2010–2018).	0
	[9] **Science and technology in Hungary**: Among Hungary's numerous research universities, the Eötvös Loránd University, founded in 1635, is one of the largest and the most prestigious public higher education institutions in Hungary. The 28,000 students at ELTE are organized into eight faculties, and into research institutes located throughout Budapest. ELTE is affiliated with 5 Nobel laureates, as well as winners of the Wolf Prize, Fulkerson Prize and Abel Prize, the latest of which was Abel Prize winner Endre Szemerédi in 2012.	0
	[10] **Abel Prize**: The Abel Prize's history dates back to 1899, when its establishment was proposed by the Norwegian mathematician Sophus Lie when he learned that Alfred Nobel's plans for annual prizes would not include a prize in mathematics. In 1902 King Oscar II of Sweden and Norway indicated his willingness to finance a mathematics prize to complement the Nobel Prizes, but the establishment of the prize was prevented by the dissolution of the union between Norway and Sweden in 1905. It took almost a century before the prize was finally established by the Government of Norway in 2001, and it was specifically intended ẗo give the mathematicians their own equivalent of a Nobel Prize.T̈he laureates are selected by the Abel Committee, the members of which are appointed by the Norwegian Academy of Science and Letters.	0

Table 8: Randomly sampled example of a query on "*What is the name of the winner of the Abel Prize of 2022?*" and top-10 judged passages in English (en) from the non-relevant subset (test split) in NoMIRACL. Titles of each passage are marked in **bold**. The relevance judgment has been annotated manually by a native speaker.

Query	Judged Passages	Relevance
In which country Praia dos Pescadores is?	[1] **Praia dos Pescadores (Albufeira)**: Praia dos Pescadores or the "Fishermans Beach" is a blue flag beach on the Atlantic south coast of the Algarve, in the district of Bairro dos Pescadores (Neighborhood of the Fisherman), Albufeira which is within the Municipality of Albufeira, Portugal. The beach is one of the two beaches which front the town of "Albufeira" with "Praia do Túnel" at the western end and "Praia dos Pescadores"" lying to the eastern end of the towns seafront. The town and its beaches are located west by road of the regions capital of Faro. In the days before Albufeira had a harbour and mariner the "Praia dos Pescadores" was where all the local fishermen operated from and the beach scene would have been very different to the site you see today. Then the beach would have been full of brightly painted fishing boats pulled up on this beach when not at sea and much of the tourist activities took place on the "Praia do Túnel". Today the "Praia dos Pescadores" is now used for tourism and is a very busy beach especially in the summer season.	1
	[2] **Praia do Túnel (Peneco)**: Praia do Túnel is a beach on the Atlantic south coast of the Algarve, in the town of Albufeira which is within the Municipality of Albufeira, Portugal. The beach is also known as "Praia do Peneco" and is one of the two beaches which front the town of "Albufeira" with "Praia do Túnel" at the western end and "Praia dos Pescadores" lying to the eastern end of the towns seafront. The town and its beaches are located west by road of the regions capital of Faro. The beach gets its name from a 20 meter long tunnel next to the tourist office in the middle of Albufeira which cuts through the cliffs linking the towns square to the beach. At the western end of the beach there is a promenade which ends at the cave known as the Xorino Grotto. According to 13th-century legend, the cave was used as shelter by the Moors after the Christian conquest of Albufeira. As well as the tunnel there are several other points of access to the beach including a lift, ramps and steps [...]	0
	[3] **Praia dos Pescadores (Albufeira)**: The beach is in length and is wide at low tide. The beach is divided by a protruding cliff from Praia do Túnel at the western end of the seafront. To the beaches eastern boundary is the Praia do Inatel and it is divided from that beach by a concrete pier which covers the outflow of the Ribeira de Albufeira (Albufeira River). There are also cliffs at the eastern end and to the back of the beach there is an amphitheatre of white houses of the district of Bairro dos Pescadores. The beach can also be accessed by an outdoor foot escalator from the Pau da Bandeira bluff located south of Bairro dos Pescadores down to the beach and Albufeira old town.	0
	[4] **Praia dos Pescadores (Albufeira)**: Praia dos Pescadores is an easily accessed beach with its large hard surface square at beach level. There are two car park's near-by, one of which, is at beach level, a short distance along the Avenida 25 de Abril within the old town. The second car park is at the top of the cliffs at Bairro dos Pescadores and is accessed via the outdoor escalator. To the back of the western end of the beach there a variety of restaurants many of which specialise in the local fish and seafood. The beach has several licensed concessions with opportunities to hire parasols and sun loungers. There are also many organised beach and water sport concessions from volleyball to boat trips and Parasailing. The beach also has toilet and shower facilities. During the summer months the beach is patrolled by lifeguards. In recent years the beach has been the focal point for the new year celebrations in the town. A temporary concert stage is erected on the Largo 25 de Abril and concerts are held to celebrate the new year. In the past the celebration has seen international bands appearing such as British reggae/pop band UB40 in 2009. The celebrations cumulate with a firework display held just of the beach on boats and pontoons just of the shoreline.	0
	[5] **Praia do Penedo**: Praia do Penedo is a beach within the Municipality of Aljezur, in the Algarve, Portugal. The beach is on the western Seaboard in the north west of the Algarve. The beach is south west of the village of Aljezur, and is north west, by road, from the regions capital of Faro. The beach of Praia do Penedo is inside the Vicentine Coast Natural Park, an area of outstanding natural beauty.	0
	[6] **Praia do Norte**: Praia do Norte is a civil parish of the municipality of Horta, located along the northern coast between Cedros and Capelo, on the Portuguese island of Faial, in the archipelago of the Azores. The population in 2011 was 250, in an area of . It is the least populous parish on the island, reached along the Ëstrada RegionalË.R.1-1ª regional roadway from the urban centre of Horta.	0
	[7] **Praia das Conchas, São Tomé and Príncipe**: Praia das Conchas is a settlement in the western part of the Lobata District on São Tomé Island in São Tomé and Príncipe. Its population is 174 (2012 census). Established as a plantation (r̈oça)̈, Praia das Conchas lies 2 km from the coast, 3 km west of Guadalupe. There is a smaller seaside settlement also called P̈raia das Conchas,̈ 3.5 km to the north.	0
	[8] **Praia das Gatas**: Praia das Gatas (Portuguese meaning b̈each of the cats)̈ is a sandy beach in the northeastern part of the island of Boa Vista in Cape Verde. The nearest village is Fundo das Figueiras, 5km to the southwest. It forms a part of Northern Nature Park (P̈arque Natural do Norte)̈. The small island Ilhéu dos Pássaros lies off the coast at the Praia das Gatas.	0
	[9] **Praia (Santa Cruz da Graciosa)**: Praia (officially São Mateus da Praia) is a Portuguese civil parish in the municipality of Santa Cruz da Graciosa, on the island of Graciosa, in the Azores. It still retains its former name locally, owing to the parish having once been the historical municipality of Praia. The population in 2011 was 836, in an area of 12.82 km².	0
	[10] **Praia Harbor**: Praia Harbor () is the port of the city of Praia in the southern part of the island of Santiago, Cape Verde. It is situated in a natural bay of the Atlantic Ocean. Since the latest modernization in 2014, it has 2 long quays, 3 shorter quays, a quay for fishing boats with fish processing installations, 2 container parks, 2 roll-on/roll-off ramps, and a passenger terminal. The total length of the quays is 863 m, and the maximum depth is 13.5 m. The port of Praia played an important role in the colonization of Africa and South America by the Portuguese. With 817,845 metric tonnes of cargo and 85,518 passengers handled (2017), it is the second busiest port of Cape Verde, after Porto Grande (Mindelo).	0

Table 9: Randomly sampled example of a query on "*In which country Praia dos Pescadores is?*" and top-10 judged passages in English (en) from the relevant subset (test split) in NoMIRACL. Titles of each passage are marked in **bold**. The relevance judgment has been annotated manually by a native speaker.