MARS: An Automatic Evaluation Framework for Cross-lingual RAG

Wesley Joosten*
University of Twente

Supervisors:
Dr. Shenghui Wang†
Dr. Faiza Bukhsh†
Rowan Terinathe MSc‡

5th February 2025

Abstract

Recently, there have been significant developments in the area of evaluating RAG (Retrieval-Augmented Generation) systems. Unfortunately, this research is limited mainly to English or monolingual systems. For multilingual RAG systems, evaluation is often limited to overall performance metrics such as accuracy, while multilingual RAG comes with additional unique challenges that are currently underexplored. We introduce MARS (Multilingual (Automatic) Assessment of RAG Systems), building on the developments in monolingual evaluation, especially the ARES (Automatic RAG Evaluation System) framework, for the granular evaluation of multilingual RAG systems. MARS can effectively evaluate existing metrics from the RAG triad in multilingual scenarios, as well as Language Consistency, a newly introduced metric to measure a unique challenge in multilingual RAG.

*Work done during internship at Info Support B.V.
†University of Twente
‡Info Support B.V.

1 Introduction

1.1 Motivation

Since the introduction of ChatGPT at the end of 2022, the usage of LLMs (Large Language Models) in daily life has rapidly increased (Duarte 2024; *Gartner Poll Finds 55% of Organizations are in Piloting or Production Mode with Generative AI* 2023; Uspenskyi 2024). RAG (Retrieval-Augmented Generation) (Lewis et al. 2020b) is a technology that enhances LLM systems by integrating external knowledge bases. RAG helps deal with the knowledge cutoff date and allows organisations to integrate LLMs with their knowledge bases.

Organisations worldwide increasingly use conversational agents powered by RAG to interface with their internal documentation (Singhal 2023; Smith 2024). Outside of the major world languages, organisations are often multilingual (O'Rourke and Brennan 2023). They might have documentation in both their countries' majority language and English or employ foreigners who do not speak the majority language. There are also many countries where more than one language is used in daily life (Tucker 1999). Such conversational agents should be multilingual in order to integrate into multilingual organisations flu-

ently. Unfortunately, these systems still do not perform as well in multilingual settings as in monolingual high-resource language settings (Asai et al. 2021b). RAG systems also generally struggle with faithfulness, not adequately using provided sources and extrapolating or hallucinating answers (Wu et al. 2024).

To measure the shortcomings of RAG systems, recent work develops frameworks to automatically evaluate RAG systems on several different metrics (Es et al. 2024; Ru et al. 2024; Saad-Falcon et al. 2024). Such automatic evaluation frameworks facilitate the iterative development and comparison of RAG systems. However, these frameworks are designed for monolingual settings and do not support multilingual evaluation. Multilingual systems require a different evaluation approach since both monolingual cases in multiple languages and cross-lingual cases must be evaluated. Furthermore, unique challenges in multilingual settings, such as cross-lingual retrieval, i.e. retrieving passages in another language as the questions, are not considered in existing methods. Hence, there is a lack of systems to adequately evaluate multilingual RAG systems, which halts the development and improvement of these systems. There is a strong need for thorough evaluation methods for multilingual RAG systems in order to compare methods and asses iterative improvements.

We intend to address this need by developing MARS (Multilingual (Automatic) Assessment of RAG Systems). MARS is based on ARES (Automatic RAG Evaluation System) (Saad-Falcon et al. 2024), which employs fine-tuned LLM judges, i.e. LLMs with binary classifier heads that "judge" whether a sample is positive or negative, for various metrics, requiring only a small amount of human annotations. It automatically evaluates different aspects of RAG systems, allowing users to compare different systems easily. ARES is developed for monolingual settings; it employs monolingual LLMs, and there is no support for cross-lingual evaluation. However, the system's modularity and low data requirements make it a suitable starting point for MARS. We identify all monolingual parts of ARES and adapt these to cope with multilingual scenarios. Additionally, we add a language consistency metric to measure a common shortcoming of cross-lingual RAG systems.

1.2 Challenges

The development of multilingual evaluation systems is halted by a lack of relevant data in languages other than English and more limited availability of multilingual language models. Furthermore, there is very little previous research into evaluating multilingual RAG systems. Hence, the strategies and scenarios for evaluation must be determined and developed mainly from scratch.

1.3 Research aims

In the development of MARS, we address the following research questions:

RQ1. How can the ARES framework be adapted to evaluate multilingual RAG systems and measure their unique challenges, and how effective is the adapted framework at this task?

RQ2. What role do dataset properties play in the effective evaluation of multilingual RAG systems?

RQ3. What insights into the limitations of current multilingual RAG systems can be gained using the adapted ARES framework?

1.4 Contributions

Our main contribution is the development of MARS, which we make available publicly[1]. MARS addresses the shortcomings of ARES to form a multilingual evaluation framework. It enables rapid iteration and improvement of multilingual RAG systems by allowing easy evaluation of new systems during development. To our knowledge, it is the first system to evaluate specific aspects of multilingual RAG systems rather than just their overall performance.

We comprehensively evaluate the MARS framework and discuss its strong and weak points. We give directions for the applicability and limitations

[1] `https://github.com/WJ44/MARS`

of the system. We give directions for the further development and validation of the system.

Additionally, we provide an extensive discussion of existing multilingual QA (Question Answering) datasets and their properties, summarising the role of these properties in evaluating cross-lingual RAG systems.

Lastly, we apply MARS to an existing cross-lingual RAG system to provide baseline scores. Using the output of MARS, we perform a qualitative analysis of the system's performance and gain insights into its limitations.

1.5 Outline

We begin by discussing related work on evaluating RAG systems in section 2. Next, section 3 describes the architecture of MARS and the changes that were made compared to ARES. The datasets used for our experiments are discussed in section 4. We describe our experiments in section 5 and discuss their results. Finally, we describe our conclusions and the limitations of our work in section 6, followed by possible directions for future work. We provide a list of acronyms used at the end.

2 Related work

2.1 Multilingual RAG systems

RAG systems are often used as Question Answering systems. Where multilingual QA is usually concerned with scenarios where passages and questions are in the same language, cross-lingual QA is concerned explicitly with retrieving evidence from knowledge bases in languages other than the original question. However, the terms are sometimes used interchangeably. Early solutions for cross-lingual QA, and many still, translate questions into English, use English QA systems to answer in English, and then translate back into the target language. This strategy is sensitive to errors in the machine translation and can only use English sources (Asai et al. 2021b).

Cross-lingual RAG methods have been developed, such as CORA (Asai et al. 2021b) and Sentri (Sorokin et al. 2022), where a single cross-lingual retriever and multilingual generator are used to answer questions in any language by retrieving evidence from any language. These methods outperform models that use translations. One notable problem these systems face is that they sometimes produce answers in the wrong language, often the language of retrieved evidence.

Currently, a common solution for RAG in practice is to use a tool like LangChain to "chain" off-the-shelf models. Popular LLMs such as GPT-4 are combined with available LLM-based embedding models. Since these popular models often have multilingual capabilities, they are also used for cross-lingual RAG. Such systems are primarily developed in industry, and there is very little research into their performance and shortcomings, which is further halted by a lack of thorough evaluation methods.

2.2 Evaluation of RAG

2.2.1 RAG triad

The RAG triad is a collection of metrics commonly used to evaluate RAG systems. They appear to be introduced by TruEra in their TruLens tool (Madzou 2024) but are also used by other systems such as RAGAs (Retrieval-Augmented Generation Assessment) and ARES (Es et al. 2024; Saad-Falcon et al. 2024). The RAG triad consists of Context Relevance (a metric to evaluate the retriever), Answer Relevance (a metric to evaluate the generator) and Answer Faithfulness (a metric to evaluate their interplay).

Context Relevance is a measure of the precision and specificity of the context retrieved by an RAG system (Gao et al. 2024). The context should be focussed, containing as little irrelevant information as possible (Es et al. 2024).

Answer Relevance measures whether generated answers pertain to the posed question and adequately address the core inquiry (Gao et al. 2024). Answers should exclusively contain information needed to answer the question (Es et al. 2024).

Answer Faithfulness, also called faithfulness or groundedness, measures whether generated answers remain faithful to the retrieved context (Gao et al. 2024). An answer is faithful if the claims made in

the answer can be inferred from the context (Es et al. 2024).

2.2.2 Automatic evaluation

Several automatic evaluation systems for RAG have been developed recently. RAGAs (Es et al. 2024) is a system that makes use of model-based evaluation, a relatively cheap strategy for testing the output of generative LLMs where an LLM is prompted to evaluate a system. RAGAs focuses on a setting where reference answers are unavailable, thus focusing on self-contained reference-free metrics: the RAG triad. These three quality aspects can be measured fully automatically by prompting an LLM using different prompting strategies. While RAGAs seems to do well, it relies on hand-crafted prompts with additional processing and prompts an LLM for every data point, making the system difficult to adapt and expensive in use.

Another evaluation method is RGB (Jiawei Chen et al. 2024). RGB evaluates RAG on four metrics: noise robustness, negative rejection, information integration and counterfactual robustness. RGB is a benchmark that contains data samples to asses the different metrics, making it difficult to adapt. Furthermore, the number of samples is limited, and a very small corpus is used, so scores on the benchmark might not reflect real-world performance.

AttrScore (Yue et al. 2023) is a framework designed for automatically evaluating attribution and identifying specific types of attribution errors. Two approaches are explored in AttrScore: prompting LLMs and fine-tuning LMs (Language Models) on simulated and repurposed data from related tasks. The prompting approach uses model-based evaluation and is similar to RAGAs. The primary challenge in fine-tuning LMs for evaluation is the lack of training data. The prompting approach is promising, but the fine-tuning approach is significantly more accurate. The system as is only supports English.

Concurrent with our research, RAGChecker (Ru et al. 2024) was released. RAGChecker distinguishes between two use cases: a user wanting to compare systems and pick the best and a developer wanting to improve a system interested in specific shortcomings. To address this, RAGChecker has both overall metrics of performance and granular metrics for both the retriever and generator, evaluating different aspects of the system. It extracts claims from long-form ground-truth answers and system responses and determines whether these claims are supported by retrieved passages using RefChecker (X. Hu et al. 2024). From this information, several different metrics are calculated. The overall performance metrics show promising alignment with human judgement. However, the system only supports monolingual scenarios.

2.2.3 ARES

Another recent solution is ARES (Saad-Falcon et al. 2024). ARES (Automatic RAG Evaluation System) is an automatic evaluation system that trains an LLM judge for each of the RAG triad. This strategy is similar to the fine-tuning approach in AttrScore. It promises substantially better evaluation precision and accuracy over other methods, such as RAGAs. ARES uses PPI (Prediction-Powered Inference) (Angelopoulos et al. 2023) to provide statistical guarantees and confidence intervals. It requires a set of passages from a target corpus, a small human reference validation set and few-shot examples of in-domain questions and answers. It can thus be applied to a new domain without requiring a large dataset of labelled examples. It works in three stages: generating a synthetic dataset using an LLM, fine-tuning a separate judge model for each metric, and evaluating the RAG system using PPI. While ARES is very promising for evaluating RAG systems, it is currently only applicable to monolingual RAG systems, as parts of the system are English and monolingual. It also does not measure challenges unique to cross-lingual RAG.

2.3 Evaluation of cross-lingual RAG

A thorough evaluation and comparison is lacking for all cross-lingual RAG methods. New RAG systems are commonly tested on a range of open-domain QA datasets, where accuracy is reported. Most papers only report accuracy on QA datasets, while some also report retrieval accuracy when golden passages are

known or BLEU scores (Papineni et al. 2002) for the generated answers. While this allows for comparing overall performance, it gives little insight into the system's capabilities and shortcomings.

Since RAG is an integrated system with multiple steps and many choices, more fine-grained evaluation is desired. Several benchmarks combine datasets from different tasks and domains (J. Hu et al. 2020; Liang et al. 2020). While this allows for a better comparison of systems, it still does not give much insight into how RAG systems might fail and what can be improved. While some papers include ablation tests, usually only accuracy is considered still. Metrics more suitable for RAG are required, which can indicate how and when the system fails and how it does well.

2.4 Multilingual benchmarks

XTREME (J. Hu et al. 2020) is a benchmark that combines different multilingual datasets for different tasks. It is, however, intended to test the applicability of machine learning models to different languages in monolingual settings, where all of the test data is in a single language. Furthermore, it is not specific to RAG.

A similar benchmark is XGLUE (Liang et al. 2020), which combines a multilingual pre-training corpus with existing multilingual evaluation datasets to test the capabilities of LLMs in different languages. Like XTREME, it is not intended for cross-lingual scenarios and is not specific to RAG.

2.5 Unresolved challenges

Cross-lingual RAG comes with new challenges in addition to the ones in monolingual RAG, such as language bias in cross-lingual retrieval and asymmetric knowledge availability in different languages (Asai et al. 2021a). To evaluate how good a multilingual RAG system is, metrics are needed that measure how well the system overcomes these challenges. How well the system performs in different languages on existing metrics such as the RAG triad needs to be measured, but also its cross-lingual capability, i.e. how well it can integrate knowledge available only in a language different from the one of the question.

To this end, MARS uses the existing metrics from the RAG triad in cross-lingual scenarios where passages and questions are not necessarily in the same language. MARS also addresses a common problem in cross-lingual RAG with a new metric: LC (Language Consistency). Cross-lingual RAG systems have a tendency to answer in a language other than the question (Li et al. 2023), which can be measured accurately using this metric.

3 MARS

Like ARES, MARS (Multilingual (Automatic) Assessment of RAG Systems) works in three stages: synthetic data generation, LLM judge training and evaluation. An overview of the system is shown in Figure 1. Synthetic data generation is done to create a domain-specific dataset to train the system in situations where only passages from a knowledge base are available. LLM judges are trained on this synthetic dataset to measure the performance of RAG systems on the specific knowledge base. Using these LLM judges, RAG systems can then be evaluated, aiding in the choice of system best suited to the user's specific needs.

We adapt these steps from ARES to facilitate ranking cross-lingual RAG systems. Furthermore, we introduce a new metric to measure a common shortcoming of cross-lingual RAG systems: Language Consistency. Lastly, we also make some changes to ARES unrelated to cross-lingual evaluation but intended to improve the system in general. These are mainly to differentiate Answer Faithfulness from Answer Relevance, as they are treated the same in ARES, and better align it with its definition.

In the following section, we attempt to answer the first part of RQ1. We begin by describing our Language Consistency metric, followed by the implementation of MARS, for each step, first describing the approach of ARES and then discussing any changes and additions made in MARS. We provide a global overview of all changes in Table 1. A short usage manual for the system is included in Appendix D.

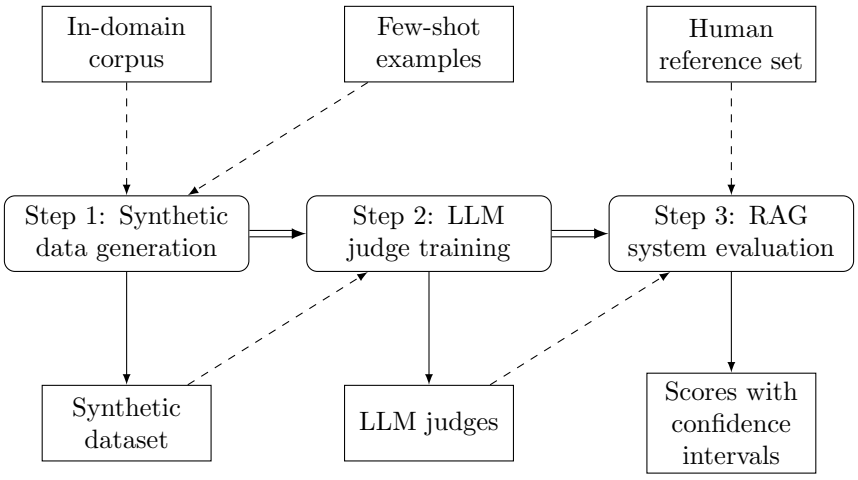

Figure 1: A diagram showing the overall system architecture of MARS.

3.1 Language Consistency

A common problem in cross-lingual RAG systems is Language Consistency. When asked a question in a particular language, but relevant knowledge is available only in passages in another language, a system is likely to wrongly generate an answer in this second language while we require an answer in the first (Li et al. 2023). A metric to measure the performance of RAG systems in this regard is added to the system. It is intended to estimate the ratio of questions answered in the correct language. To keep within the same architecture of the other metrics, Language Consistency follows the same three steps. More details on the implementation are given in the respective sections.

3.2 Step 1: Synthetic data generation

In ARES, the data used to train the LLM judges is synthetically generated based on the in-domain corpus also used by the RAG system to be evaluated as its knowledge base. Synthetic questions and corresponding answers are generated for each passage in the corpus. These samples are generated using few-shot examples. ARES uses FLAN-T5-XXL (Chung et al. 2024), an 11.3B parameter model, to generate the synthetic data, but another high-quality model can ultimately be used.

Following previous work (Saad-Falcon et al. 2024), low-quality questions can be filtered out by verifying whether a simple retriever can retrieve the original passage as the top result given the question. The question is removed if the passage is not retrieved as the top result. For Context Relevance, negatives are generated by sampling passages unrelated to a given synthetic question. For Answer Relevance and Answer Faithfulness, negatives are generated by sampling synthetically generated answers from other questions.

3.2.1 Changes unrelated to multilinguality

Differentiate Answer Faithfulness metric In ARES, negative training samples for Answer Faithfulness are constructed by sampling wrong answers. In this case, the LLM judge could learn to look only at whether the answer answers the question, just like Answer Relevance. However, in practice, a system might give answers relevant to the question but not supported by the retrieved passage.

Alternatively, the negative samples could be made by sampling random passages. However, the classifier could then learn only to consider whether the passage is relevant to the question, ignoring the answer, just like Context Relevance.

To force the classifier to consider whether the passage supports the answer, we use a mixture of two kinds of negative samples, where half of the negatives are constructed by sampling random passages and half by sampling random answers. This strategy forces the classifier to consider the relationship between the answer and the passage to get good performance in both scenarios while still allowing us to provide the answer to the LLM judge as additional potentially helpful information.

Change strategy for generating synthetic samples from passages Additionally, we make a few changes to the synthetic data generation to optimally use the corpus and ensure a similar amount of samples for each metric. We generate the same number of questions as there are passages, although there can be anywhere from 0 to 4 from each specific passage. An answer is generated for each question.

Each question is then used to create both a positive and negative sample for Context Relevance and Answer Relevance. For Answer Faithfulness, we randomly sample half of the questions for each strategy to generate negatives, leading to an intentional overlap in questions used for each strategy. The overlap is intended to further condition the classifier to the relationship between the answer and the passage.

Furthermore, while the ARES paper describes a distinction between weak and strong negative generation for the synthetic data, the strong negative generation is not present in the provided codebase. Consequently, in MARS, only the weak negative generation from ARES is used, which is the process described above. In strong negative generation, negative passages are sampled from the same document as the passage used to generate the question for Context Relevance, and negative answers are generated using few-shot prompts for Answer Relevance and Answer Faithfulness.

3.2.2 Adaptions for multilingual scenarios

To determine which adaptions should be made to the synthetic data generation step, we need to know what the synthetic data should look like in order to train LLM judges for evaluating cross-lingual RAG systems. In a bilingual setting, synthetic questions and answers are required where passages and questions are in the same language and where they are in a different language. Additionally, positive and negative samples are required for Language Consistency.

Language Consistency is primarily relevant in cases where a question and passage are in different languages, as otherwise, an RAG system is unlikely to generate an answer in the incorrect language. Hence, positive samples for Language Consistency are cases where a passage and question are in a different language, and the answer is in the question's language. Negative samples are cases where a passage and a question are in a different language, and the answer is in the passage's language. Since, in practice, the LLM judge will also be evaluating cases where questions and passages are in the same language, such samples are included as well so the classifier learns to handle them. Here, negative cases have answers in the wrong language, even though this is not the passage's language. We generate negatives by, for each generated question, also generating an answer in the wrong language using a few-shot prompt. Positive and negative Language Consistency samples can be positive or negative on any of the other metrics.

Change model for generating synthetic dataset We make several adaptions to achieve these desired changes to the synthetic data. First, a multilingual LLM is required to generate multilingual synthetic data, specifically, an instruction fine-tuned model that supports few-shot examples. In principle, any such model can be used if it supports the desired languages and is performant enough. The Aya 23 (Aryabumi et al. 2024) model (35B parameters) is used in this work since it is highly performant and its weights are publicly available. Aya 101 (Üstün et al. 2024) was also considered, as it is closer in architecture to the FLAN-T5-XXL (Chung et al. 2024) used in ARES, but its benchmark performance is significantly below Aya 23, as well as comparable English models. Additionally, we found Aya 101 to struggle with language consistency in generating synthetic data in explorative testing.

Add few-shot prompts for each passage-question language pair Secondly, changes are required in the generation of samples. These are mainly in the form of adaptions to the few-shot prompts used. In MARS, synthetic questions and answers are generated per passage-question language pair. A passage-question language pair is a combination of passages in one language and questions and answers in another or the same language. A bilingual setting gives four passage-question language pairs and, thus, four synthetic datasets. For each passage-question language pair, few-shot prompts are constructed by taking passages in the first language and questions and, where applicable, answers in the second language.

Add language to few-shot prompts The expected language is explicitly stated in the few-shot prompt to further condition the LLM to generate in the correct language, significantly decreasing synthetic data generation in the wrong language.

Add generation of synthetic answers in wrong language As mentioned previously, there is also a negative few-shot prompt where the LLM is tasked to generate answers in the wrong language.

We provide the few-shot prompts used in Appendix B.

Change retriever embedding model One last adaption to ARES was required. To judge the quality of synthetic questions, the system can check whether a simple retriever using an embedding model can retrieve the source passage used to generate the question as the first result. In a multilingual setting, this requires an embedding model that embeds samples from different languages into an aligned embedding space. BGE-M3 (Jianlv Chen et al. 2024) was chosen for this task. BGE-M3 was chosen after comparing the retrieval benchmark performance of popular options. It performs comparably across languages, allows for cross-lingual retrieval and even outperforms the text-embedding-ada-00 (Greene et al. 2022) used in ARES in English. The passages are all embedded using the model and stored in a FAISS index (Johnson et al. 2019). During synthetic question generation, the questions are embedded using the model and the nearest passage is retrieved from the index.

We provide some examples of synthetic positives and negatives in Appendix A.

3.3 Step 2: LLM judge training

The synthetic datasets are used to train a relatively lightweight LLM for each evaluation metric. In ARES DeBERTa-v3-Large (He et al. 2021) is used, a 304M parameter model, with an added classifier head. The judges are trained with a binary classification training objective. The human reference set is used as the validation set during training.

3.3.1 Changes unrelated to multilinguality

Differentiate Answer Faithfulness metric In ARES, no difference is made in training or inference between the Answer Relevance and Answer Faithfulness judges. Both see the complete passage-question-answer triple and use the same training data. Since Answer Relevance is about how well an answer answers the question, this should be independent of the passage. We change the training and inference procedures so the Answer Relevance judge only gets fed the question and answer. Answer Faithfulness, in turn, should be independent of the question since it is about whether the passage supports the answer, and the judge could be fed only the passage and the answer. This claim would hold if the answers were long-form. In reality, however, the answers are often short-form, e.g. only the word "yes", and the question is required to interpret the meaning of the answer and whether the passage supports it. The Answer Faithfulness judge, hence, still gets fed the complete triple. Alternatively, either MARS or the RAG system could include a processing step to transform short-form answers into long-form answers.

3.3.2 Adaptions for multilingual scenarios

The LLM judges must be adapted to handle cross-lingual scenarios. Since, in practice, the language of passages, questions or answers is unknown to MARS, the LLM judges must be able to judge samples regardless of their language. To this end, we train the LLM judges using the synthetic datasets from all passage-question language pairs together. This requires the LLM judges to handle multiple languages, and so, once again, requires a multilingual LLM.

Change model for LLM judges A multilingual version of the DeBERTa-V3 model used in ARES is available, mDeBERTa-V3 (He et al. 2021), which is shown to have good cross-lingual performance on new tasks. Unfortunately, no large version is available for this model, so we use the base version (86M parameters) instead. We stick to a model close in architecture and training regime as the one used in ARES since it is shown to perform well for the task considered. Other options, such as XLM-R (Conneau et al. 2020) or XLM-E (Chi et al. 2022), are available, but mDeBERTa-V3-base seems performant enough in preliminary testing; mDeBERTa-V3-base LLM judges reach similar validation set accuracy as the DeBERTa-V3-large judges in ARES. To train the LLM judges to handle the different language pairs, we combine the four synthetic datasets and train the judges on all of them, making the judges completely agnostic of the question, passage and answer language, which is necessary since in practice, the language is usually unknown.

Add language Consistency metric Other than adding an additional LLM judge for the Language Consistency metric, no further adaptions to the LLM judge training are required.

3.4 Step 3: RAG system evaluation

ARES uses PPI (Prediction-Powered Inference) (Angelopoulos et al. 2023) to combine the benefits of the LLM judges and the human annotations from the human reference set. To evaluate an RAG system, questions from an evaluation dataset are put to the RAG system up for evaluation. The retrieved passages and given answers are then saved with the questions as evaluation samples. The LLM judge for each metric then classifies each data sample as either positive or negative for that metric. They also classify each sample in the human reference set. PPI is then used to rectify the predictions by the LLM judge on the unlabelled samples and calculate a score with a confidence interval for each metric. We end up with four scores for the tested RAG system, one for Context Relevance, Answer Relevance, Answer Faithfulness and Language Consistency.

3.4.1 Changes unrelated to multilinguality

Differentiate Answer Faithfulness metric As described before, we change the system such that the Answer Relevance LLM judge only sees the question and answer, no longer the context.

3.4.2 Adaptions for multilingual scenarios

To evaluate cross-lingual RAG systems, hardly any adaptions to ARES are necessary other than requiring a human reference set with samples for each passage-question language pair.

4 Datasets

Two types of data needs can be distinguished: the data required to use the MARS framework in practice and the data required for the meta-evaluation of the framework's efficacy in an experimental setting. Note that no large labelled datasets are required to develop MARS or use it in practice; they are only required for the meta-evaluation of MARS itself. In the following section, we identify the data requirements for both scenarios and attempt to answer RQ2.

4.1 Data requirements for meta-evaluation

In order to validate MARS and the scores it gives, we need to evaluate MARS itself, which we will call meta-evaluation for clarity. We want to ensure MARS works well on each of the included metrics

Part	Change	Multi-lingual
Step 1	Differentiate Answer Faithfulness metric	No
Step 1	Change strategy for generating synthetic samples from passages	No
Step 1	Change model for generating synthetic dataset	Yes
Step 1	Add few-shot prompts for each passage-question language pair	Yes
Step 1	Add language to few-shot prompts	Yes
Step 1	Add generation of synthetic answers in wrong language	Yes
Step 1	Change retriever embedding model	Yes
Step 2	Differentiate Answer Faithfulness metric	No
Step 2	Change model for LLM judges	Yes
Step 2	Add Language Consistency metric	Yes
Step 3	Differentiate Answer Faithfulness metric	No

Table 1: Overview of changes made to ARES and whether they are done specifically to evaluate multilingual RAG systems.

across different scenarios. For a meta-evaluation of MARS, a ground-truth score is required for each metric to compare with MARS' output. This requires previously unseen labelled passage-question-answer triples, akin to the human reference set discussed later. Such samples, but usually only positives, are readily available in QA datasets normally used to train and evaluate QA systems. We use these labelled samples to construct mock RAG system output with controlled, known performance on each metric. To properly assess the performance of MARS, several relevant properties of QA datasets must be considered. These properties and their relevance to the experiments are described below.

4.1.1 Dataset properties

Labels Labels are required for each metric in MARS. However, in many cases, these labels can be inferred using other dataset properties. Context Relevance labels can be inferred from datasets that include gold passages for each question, Answer Relevance when they have a correct answer for each question, Answer Faithfulness when the answers are extracted from passages and Language Consistency when the language of passages, questions and answers is known. When labels are inferred, there are usually only positive samples. Negative samples can then be artificially constructed, although they are at risk of being unrepresentative of real-world cases.

Cross-linguality To best mimic a real-world bilingual scenario, a dataset is required that contains passages in two languages and questions and corresponding answers in those two languages. Crucially, samples should be present where the information used to answer a question is not in the same language as the question.

Parallel samples To allow a fair comparison between different scenarios, e.g., all passages in one language and questions in either another or the same language, it is best if the dataset contains parallel passages and questions, i.e., the same passages and questions available in multiple languages. Otherwise, the average difficulty of questions could vary between languages. Parallel questions additionally allow for the most control over experimental scenarios regarding the availability of information in either language.

Answer independance In many datasets, questions are tied to a specific passage, and the answer is a span within that passage. While this ensures that a question is answerable given the passage, it ties the answer to its particular phrasing (Longpre et al. 2021). This can make the samples artificially easy since a system does not need to handle paraphrasing. The importance of answer independence in evaluating MARS is non-trivial. While independent answers and questions best reflect real-world cases, datasets with independent questions and answers do not usually include labels for which questions are answerable using information from which passages or whether the question is answerable using the corpus at all. Hence, there is less control over the experimental scenarios.

Open retrieval Open retrieval is related to answer independence. Open-retrieval datasets assume

information can be found in a large corpus of knowledge, such as the whole of Wikipedia, instead of a relatively small set of predetermined passages (Asai et al. 2021a). Since MARS' intended use case is for settings with a specific knowledge base, open retrieval is somewhat out of scope.

Native-speaker questions Some datasets include questions as asked by native speakers, while others include either professional or machine translations. While native-speaker questions best reflect the natural distribution of questions from a specific language or culture, they disallow the possibility of parallel questions. Translations, however, can introduce translation artefacts and translationese (Clark et al. 2020; Longpre et al. 2021), lowering the dataset's quality.

Languages Datasets are available for an array of different languages. While we are not necessarily concerned with a specific language pair, in practice, it is crucial to know how well MARS performs in the users' desired languages.

Domain RAG can be applied to many different domains. Hence, it is interesting to know about the performance of a system in different domains. Knowing how well a RAG system performs in domains similar to the intended use case is especially useful.

Size The more extensive the dataset, the more robust the meta-evaluation will be. Furthermore, as mentioned later, there are some minimum size requirements for generating the synthetic dataset and constructing the human reference sets.

4.1.2 Importance of properties

As alluded to in the description of these properties, they do not all share the same level of importance. The evaluation should best reflect real-world scenarios while allowing for fair comparison and remaining practical. Cross-lingual samples and labels for each metric are a must for the meta-evaluation of MARS' capabilities. To best reflect real-world scenarios, questions asked by native speakers and answers independent of the context in an open-retrieval setting are desired. However, parallel samples are of more importance to allow for a fairer comparison between scenarios. Having labels for Context Relevance is somewhat incompatible with the notion of open retrieval since the dataset would contain samples only with a limited number of documents. These considerations lead us to a distinction between required and desired properties, where the desired properties do not share the same level of importance. An overview of the properties and their importance is given in Table 2. This table also shows whether the properties are unique to cross-lingual RAG.

Property	Required	Desired	Cross-lingual
Labels	Yes	N/A	No
Cross-linguality	Yes	N/A	Yes
Parallel samples	No	+++	Yes
Answer independence	No	++	No
Open retrieval	No	+	No
Native-speaker questions	No	++	Yes
Languages	Yes[2]	N/A	Yes
Domain	Yes[2]	N/A	No
Size	No	++	No

Table 2: an overview of the data properties considered for evaluating cross-lingual RAG systems using MARS. Here required means the system cannot function without it. The amount of crosses in the desired column indicates the relative importance.

4.1.3 Dataset choice

A comparison of existing datasets for these properties is shown in Table 3. Choosing the right datasets is challenging since we have to weigh our desired properties against each other, and few applicable multilingual datasets are available. Furthermore, we will have to contend with artificially creating negatives

[2]In practice, these are required to be specific to the use case, however, for our experiments, any domains and languages are usable.

since no dataset provides negatives for all of our metrics. Alternatively, we could construct our own dataset to suit our needs, but this requires human labelling hundreds to thousands of samples and is out of the scope of our research.

Keeping our considerations in mind, we choose to, in the first place, use MLQA (Lewis et al. 2020a) for the meta-evaluation of MARS, and XOR-AttriQA (Muller et al. 2023) for further evaluation of the Answer Faithfulness metric.

MLQA MLQA translates questions, guaranteed to be answerable by mined parallel pieces of text, into multiple languages. The passages are hence not translated. Answers are extracted spans from the parallel passages. This means we can use the dataset cross-lingually, using passages in one language to answer questions in another. Since the questions are parallel, we can compare different cross-lingual scenarios fairly. Furthermore, it has enough samples to construct a robust synthetic dataset and sufficiently large datasets for meta-evaluation. Unfortunately, answers are not independent from the text, potentially making them artificially easy. Questions are also not asked by native speakers since this disallows parallel samples. To ensure questions are answerable using passages in each language, the dataset uses a limited corpus of parallel passages and hence is not open-retrieval.

XOR-AttriQA XOR-AttriQA is a dataset constructed from CORA output on XOR-TyDi QA (Asai et al. 2021a). It uses the questions from XOR-TyDi QA, which are naturally elicited from native speakers of different languages and intended for open retrieval. It uses Wikipedia as the corpus for CORA, either in the same language as the questions or in English. It also translates any output and passages to English, making all samples parallel between English and one other language. This still allows for a fair comparison between different passage-question language pairs within a bilingual system but not across languages. Crucially, the CORA output is human labelled for Answer Faithfulness[3]

Challenges Ideally, the same dataset would be used for all metrics. However, the MLQA dataset contains no negatives for our metrics, so these have to be constructed artificially and might, thus, be unrealistic and artificially easy. The XOR-AttriQA dataset, however, while including real-world positives and negatives for Answer Faithfulness, is relatively small, allowing for a less robust evaluation.

4.2 Data requirements for evaluating RAG systems using MARS

Until now, we discussed the data required to validate MARS. We will now outline the data required to use MARS in practice. Three sets of data are required to use the system: a corpus of in-domain passages for generating the synthetic dataset, in-domain questions to pose to the RAG system and passage-question-answer triples for the human-reference set. Few-shot examples of passage-question-answer triples are also required to generate the synthetic data. The synthetic dataset itself is also a data need for the system as a whole.

4.2.1 Synthetic data generation

Since the intended use is to evaluate cross-lingual RAG systems, the system needs to be trained on cross-lingual data, meaning that passages are required in all the intended languages. In practice, this passage set would be the knowledge base that the RAG system uses. To generate cross-lingual synthetic data for training the LLM judges, the system requires this passage set along with few-shot examples for each language pair.

4.2.2 LLM judge training data

Data is needed to train the LLM judges. The synthetic data discussed above is used for this. For each passage, MARS generates positive and negative

[3]In truth, XOR-AttriQA does not include a metric called Answer Faithfulness but instead AIS (Attributable to Identified Sources), the definitions are almost identical, however.

samples for each metric. A few thousand samples are needed to train the LLM judges properly. Generating synthetic data specific to the knowledge base ensures that the system is perfectly adapted to the domain. The system can, however, also be applied to other domains without specific training since the judges are shown to generalize well to new domains (Saad-Falcon et al. 2024).

4.2.3 Evaluation of RAG systems

To evaluate an RAG system, MARS requires passage-question-answer triples as output by an RAG system. A set of in-domain questions is required, which can be fed into the RAG systems to be evaluated to collect these triples. This question set could, for instance, be collected from a system in use.

4.2.4 Human reference set

The system also requires a few hundred cross-lingual passage-question-answer triples labelled for each metric. These should reflect all possible passage-question language pairs. This human reference set should reflect the real distribution of questions about the knowledge base posed to the RAG system and the answers given by the system. In practice, these could be collected from the RAG system and then labelled. Saad-Falcon et al. 2024 find a minimum size for a proper evaluation to be around 150, while a more extensive human reference set improves the accuracy of the scores given by the system. This human reference set is also used as the validation set during LLM judge training.

5 Experiments

In this section, we first describe the setup shared between all of our experiments and then describe the experiments performed to validate MARS and answer the second part of RQ1 as well as RQ3. The source code for our experiments is available in the MARS GitHub repository[1].

5.1 Experimental setup

The experiments are limited to bilingual scenarios to limit the scope of our research. However, they could be extended to scenarios with more than two languages relatively easily.

The German-English parallel samples are taken from MLQA (Lewis et al. 2020a). From the test set and dev set separately, four evaluation datasets are created: English-English, English-German, German-English and German-German, where the first language denoted is the language of the passages and the second language is the language of the questions. These datasets are fully parallel, having the same context, questions and answers in each language. We also create a mixed dataset, randomly picking each sample from one of the four datasets. The mixed dataset best reflects most real-world use cases. We use the datasets resulting from the dev set of MLQA to construct the human reference sets. We use the datasets resulting from the test set for synthetic data generation and to construct mock RAG system output.

Synthetic questions and answers are generated for each of the four passage-question language pairs, resulting in four synthetic datasets. Four few-shot examples of queries and answers are taken out of the datasets. We sample 3000 passages from each of the datasets to form our corpus. For each dataset, we generate 3000 questions, and for each question, we generate an answer in both languages; these form the positives for Context Relevance, Answer Relevance and Answer Faithfulness. We sample 3000 negatives for Context Relevance, Answer Relevance and Answer Faithfulness separately. We use positive and negative samples from the other metrics with answers in the correct language as positives for Language Consistency and the wrong language as negatives, resulting in 18000 samples.

The four synthetic datasets are combined and used to train the LLM judges for each metric. In total, this means there are 48,000 samples for Context Relevance, Answer Relevance and Answer Faithfulness and 72,000 for Language Consistency, each with an equal amount of positives and negatives. The mixed human reference set is used as the validation set dur-

	Labels	Cross-lingual	Parallel samples	Answer inde-pendence	Open retrieval	Asked by native speakers	Languages	Domain	Size (train, dev, test)
XQA (Liu et al. 2019)	(CR), AR	No	No	Yes	Yes	Yes	en, zh, fr, pt, ru, de, ta, pl, uk	General knowledge	∼60k, ∼1k, ∼1k
MLQA (Lewis et al. 2020a)	CR, AR, LC	Yes	Yes	No	No	No	en, zh, hi, es, ar, de, vi	General knowledge	-, ∼500, ∼5k
XQuAD (Artetxe et al. 2020)	CR, AR, LC	No	Yes	No	No	No	en, zh, hi, es, ar, ru, de, tr, vi, th, el	General knowledge	-, -, ∼1k
TyDi QA (Clark et al. 2020)	CR, AR	No	No	No	No	Yes	en, ar, bn, ru, id, ja, te, sw, ko, th, fi	General knowledge	∼15k, ∼2k, ∼2k
XOR-TyDi QA (Asai et al. 2021a)	CR, AR, LC	Yes	No	No	Yes	Yes	(en), ar, bn, ru, ja, te, ko, fi	General knowledge	∼2k, ∼300, ∼300
XOR-AttriQA (Muller et al. 2023)	AF	Yes	Partly	No	Yes	Yes	(en), bn, ru, ja, te, fi	General knowledge	50, 100, ∼1k
EXAMS (Hardalov et al. 2020)	AR, LC	No	Partly	Yes	Yes	Mostly	es, ar, fr, pt, de, tr, vi, it, pl, hu, sr, bg, sq, hr, lt, mk	School exams	∼500, ∼200, ∼1k
MKQA (Longpre et al. 2021)	AR, LC	Yes	Yes	Yes	Yes	No	en, zh, es, ar, fr, pt, ru, de, ja, tr, vi, ko, it, th, pl, ms, nl, km, hu, sv, he, da, fi, no	General knowledge	-, -, 10k
CCQA (Huber et al. 2022)	CR, AR	No	No	Yes	Yes	Mostly	Many	Various	130M, -, -
Mintaka (Sen et al. 2022)	AR, LC	No	Yes	Yes	Yes	No	en, hi, es, ar, fr, pt, de, ja, it	General knowledge	-, -, 20k
xPQA (Shen et al. 2023)	CR, AR, (LC)	Yes	No	Yes	No	Yes	zh, hi, es, ar, fr, pt, de, ja, pa, ko, it, pl	Products	400, 100, 1k

Table 3: Comparison of datasets.
CR (Context Relevance), AR (Answer Relevance), AF (Answer Faithfulness), LC (Language Consistency).

ing training. We train for 10 epochs with a batch size of 32 and a learning rate of 5e-6. We employ early stopping when the accuracy on the validation set does not improve for three epochs. We train on an 80GB A100, taking a few hours.

We construct one human reference set for each passage-question language pair as well as the mixed case with a 50% accuracy on each of the metrics. Negatives are created in the same manner as for the mock RAG systems described below. We limit the human reference sets to 300 samples during training.

5.2 Mock RAG systems

To determine how well MARS can evaluate RAG systems, its output should be compared with ground-truth labels. Unfortunately, labelled datasets of cross-lingual RAG system outputs are not readily available for most metrics. Instead, we construct mock RAG system output. We use the test datasets to create positive and negative examples for each metric that mimic the output of potential RAG systems. We create these samples in a similar manner to the negatives for the synthetic dataset.

MLQA has questions with corresponding passages and ground-truth answers. We create negatives for Context Relevance by sampling passages from the same document that do not contain the answer for half of the questions and passages from other documents for the other half. For Answer Relevance, we create negatives by sampling answers to other questions. For Answer Faithfulness, we create negatives for half of the questions by sampling wrong passages the same way as for CR and by sampling answers from other questions for the other half. Negatives for LC are created by sampling answers from the parallel question in a second language. We construct datasets with different known performance ratios on each metric, from 50% to 70% with 5% increments. The size of these datasets is constrained to 500 samples per metric. We call these datasets mock RAG systems.

MARS is then used to evaluate these mock RAG systems, and the given score is compared with the known artificial performance of these "systems". To assess the performance of MARS, the Kendall rank correlation coefficient (Kendall's tau) (Kendall 1938) is used. Kendall's tau is a popular metric for comparing ranking systems. It is designed to measure the accuracy of pairwise rankings, comparing an experimental pairwise ranking with a perfect pairwise ranking. It is calculated as follows:

$$\tau = \frac{\#\text{concordant pairs} - \#\text{discordant pairs}}{\#\text{pairs}}$$

The accuracy of MARS in pairwise comparisons is important since, in practice, a system such as MARS would be used to see whether changes to an RAG system have a positive or negative impact or to compare two systems. Since the correct ranking of the mock RAG systems is known, we can calculate Kendall's tau between this perfect ranking and the ranking as determined by MARS' scores.

Samples from the datasets used as mock RAG systems are given in Appendix C.

5.2.1 Results of mock RAG system experiments

In Table 4, the results of MARS evaluation of the mock RAG systems are shown. Since, currently, no comparable cross-lingual evaluation framework exists to compare to, exploration is done on the impact of making ARES cross-lingual. MARS, as a system, when ranking in the English-English scenario, is almost identical to ARES, so this scenario can serve as a baseline.

We begin by looking at Kendall's tau. What can be learned from the results of this experiment is that MARS performs comparably when evaluating cross-lingual cases as it does monolingual cases. Furthermore, MARS shows promising results on the Language Consistency metric, albeit with somewhat varying performance. MARS has the most trouble ranking Answer Faithfulness. One should remember that the negative samples used in this experiment do not reflect real-world examples and are most likely artificially easy. They can, however, provide an indication of potential performance loss when evaluating cross-lingual systems. The experiment suggests that this impact is minimal.

Next to scores on a bilingual scenario in which the classifiers were trained, English-German, the table also shows ranking performance on an unseen bilingual scenario, English-Arabic. Arabic was chosen for no particular reason other than that it has a non-Latin script, which could pose an additional challenge. It is a language mDeBERTa-v3 was trained on, but we have not trained our LLM judges to evaluate our metrics. This means they have to rely on cross-lingual transfer for this specific task, something the model is shown to be good at for other tasks, such as NLI (Natural Language Inference). The results are promising, especially for Context Relevance and Language Consistency. It shows that MARS generalizes well to unseen languages. This generalization can potentially be improved further by training on more than one language pair. The performance loss is most significant for Answer Faithfulness, further suggesting that this is the hardest metric to measure.

Moving on to accuracy, we can see that the accuracy of the LLM judges is consistently high for Context Relevance and Answer Relevance and that this leads to reliable ranking performance when looking at Kendall's tau. The accuracy for Answer Faithfulness and Language Consistency is considerably lower. For Language Consistency, this still leads to mostly

reliable ranking performance, but for Answer Faithfulness, MARS struggles to rank the mock RAG systems correctly.

When looking at accuracy for our Arabic scenarios, we see only a slight drop in performance compared to our German mock RAG systems. We even see quite an increase in performance in Language Consistency. This makes some sense since Arabic is not related to English and additionally is in another script, while German is closely related to English. This probably makes it easier for our LLM judge to distinguish when answers are in the incorrect language. This is a promising result since it suggests that the LLM judge has indeed learned the task of judging whether answers are in the same language as the question.

When we look at the plots in Figure 2, we can see that for Context Relevance and Answer Relevance, the confidence intervals are very tight, and the score given by MARS, both before and after PPI is very close to the ground truth. For Answer Faithfulness, we see that the LLM judge greatly overestimates the Answer Faithfulness. PPI corrects this, but seems to "overshoot" for the mock RAG systems with higher ground-truth performance. For LC (Language Consistency), we see almost the opposite scenario. However, PPI can better deal with the underestimation of this LLM judge, with the ground-truth performance being within the 95% confidence interval in all cases. We only provide plots for the mixed case but see similar results in the other document-query language pairs.

Plots for Arabic are given in Appendix E (subsection E.1). We see similar patterns to those in the German scenarios. However, for the Arabic scenarios, the LLM judges overestimate Context Relevance, and PPI overcorrects this, actually pushing the score further away from the ground truth. We still see very strong results for AR in most cases. It is relevant to mention that the datasets used for the mock RAG systems are not completely parallel between the German and Arabic experiments since not enough questions in MLQA are three-way parallel with English, German and Arabic.

MARS seems to consistently underestimate the mock RAG systems at higher groud-truth performance. We hypothesize that this is due to the human reference set having a distribution of 50% correct and incorrect samples, and so the human reference set not aligning with the real-world distribution in the more performant scenarios. We experiment with a 70% human reference set as well to validate this hypothesis, the results of which are included in Appendix E (subsection E.2), and indeed in that case it instead overestimates in the less performant cases. It seems the 50% set works reasonably well for all ground truth performances, which means the human reference set can stay stable between experiments.

As can be seen in Table 4, only in two scenarios did PPI improve the ranking of mock RAG systems by MARS. However, it is clear from the plots that PPI helps considerably with getting an accurate score.

The results in the Arabic scenarios are promising since they point towards good transfer to unseen languages, contrary to the expectation posed in the ARES paper (Saad-Falcon et al. 2024). We have only trained on one language pair but can use the system on another. This is useful since data is scarce in many languages.

5.3 Baseline systems

We also try our system on a baseline cross-lingual RAG system. We are lucky since, for CORA, there are publicly available datasets with CORA system output on cross-lingual datasets. There is even a dataset with human labels for Answer Faithfulness for CORA output on XOR-TyDi QA: XOR-AttriQA. Unfortunately, no languages overlap between MLQA and XOR-AttriQA, and where XOR-AttriQA is open-retrieval, MLQA is not. This means that, without generating synthetic data for XOR-AttriQA and training new LLM judges, we apply our system to a different language and domain than it was trained for. Since XOR-AttriQA is an open-retrieval dataset, we do not have a specific corpus we could use to generate a synthetic dataset to train our LLM judges. We thus have to rely on the generalization abilities of our system and use the LLM judges as trained on MLQA.

For XOR-AttriQA, we use the val split per language as the human reference set. We perform much the same experiment as with the mock RAG systems,

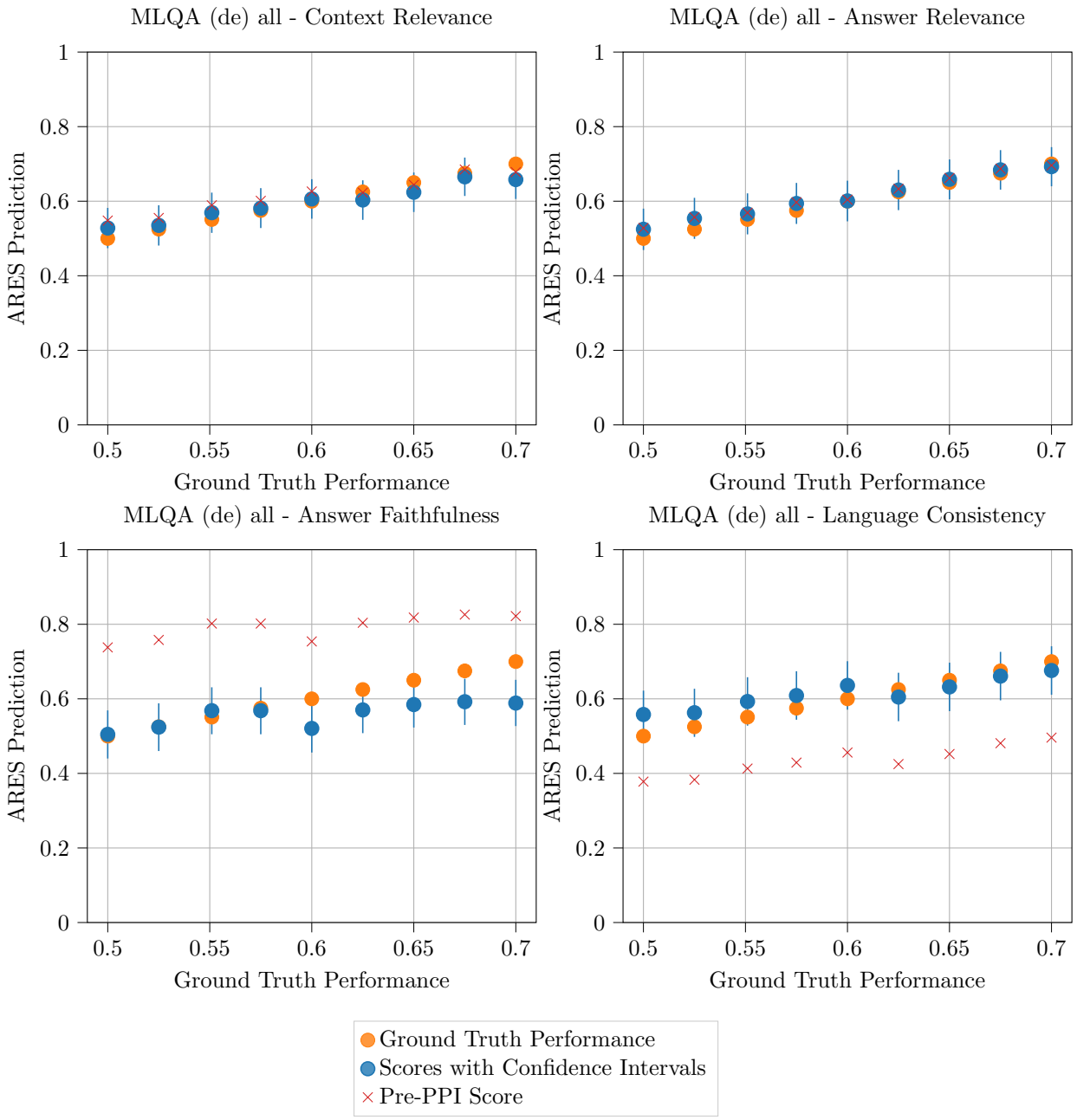

Figure 2: MARS output on mock RAG systems with mixed German-English samples.

		CR	AR	AF	LC
MLQA (de) all	Pre-PPI τ	0.89	1.00	0.76	0.83
	τ	0.89	1.00	0.76	0.83
	Acc.	92.6%	92.5%	77.6%	77.3%
MLQA (de) de-de	Pre-PPI τ	0.83	0.89	0.70	0.83
	τ	0.83	0.89	0.70	0.83
	Acc.	90.2%	91.6%	79.0%	78.1%
MLQA (de) de-en	Pre-PPI τ	0.83	0.94	0.50	0.83
	τ	0.83	0.94	0.50	0.83
	Acc.	90.4%	92.7%	77.6%	81.5%
MLQA (de) en-de	Pre-PPI τ	1.00	0.89	0.76	0.94
	τ	1.00	0.89	0.76	0.94
	Acc.	93.2%	91.6%	76.9%	78.3%
MLQA (de) en-en	Pre-PPI τ	0.94	0.94	0.83	0.76
	τ	0.94	0.94	0.83	0.76
	Acc.	93.7%	92.7%	77.6%	71.8%
MLQA (ar) all	Pre-PPI τ	0.94	0.94	0.37	0.94
	τ	0.94	0.94	0.39	0.94
	Acc.	91.3%	90.9%	76.4%	85.5%
MLQA (ar) ar-ar	Pre-PPI τ	1.00	0.89	0.93	0.94
	τ	1.00	0.89	0.93	0.94
	Acc.	90.6%	88.7%	77.8%	82.5%
MLQA (ar) ar-en	Pre-PPI τ	1.00	1.00	0.76	1.00
	τ	1.00	1.00	0.76	1.00
	Acc.	90.6%	93.1%	76.2%	94.3%
MLQA (ar) en-ar	Pre-PPI τ	1.00	0.89	0.78	0.94
	τ	1.00	0.89	0.78	0.94
	Acc.	89.3%	88.7%	73.0%	91.2%
MLQA (ar) en-en	Pre-PPI τ	1.00	1.00	0.93	0.89
	τ	1.00	1.00	0.94	0.89
	Acc.	94.2%	93.1%	77.9%	77.8%

Table 4: Results of the mock RAG system ranking. The table shows both Kendall's tau of the MARS ranking before and after PPI and the average accuracy of the LLM judges.

only now we have actual labelled RAG system output.

For our general CORA baseline scores, there is a publicly available dataset of CORA system output on XOR-TyDi QA. We feed this to MARS and receive a score for each metric in each language. The language mentioned is the language of questions. The corpus used consists of a mix of passages from all included languages. Due to a lack of labelled examples from XOR-TyDi QA, we use the MLQA German mixed human reference set.

5.3.1 Baseline system results

In Table 5, we show the results of MARS Answer Faithfulness on XOR-AttriQA. While we know from our previous experiments that MARS generalizes well to unseen languages, we see a significant drop in the accuracy of the LLM judge here. We expect this to be primarily because the negative samples it was trained on are artificially easy, often having answers or context completely unrelated to the question. Here, most answers answer the question and the retrieved context is relevant, but the answers are not grounded in the context and thus much more difficult to classify. However, we can recover much of the performance with our small human-labelled dataset of 150 samples, putting the Answer Faithfulness score very close to the ground-truth performance in almost all cases. A notable exception is Russian, where MARS vastly overestimates the performance of CORA, with the actual performance being outside of the 95% confidence intervals.

Our LLM judge's low accuracy leads us to an unfortunate conclusion. We cannot rely on the prediction of our LLM judge on specific samples, at least for Answer Faithfulness, which seems to be the most challenging metric. This means we cannot properly use the LLM judge output to gain insight into where cross-lingual RAG systems struggle in our qualitative evaluation. However, looking at the results, it seems that false negatives are rare, which makes sense if we consider the negatives in our training data artificially easy. This suggests that the samples labelled as negative by MARS can still be used for qualitative evaluation, although not to paint the whole picture.

In Table 6, we show baseline results of CORA on XOR-TyDi QA. If we are to believe MARS, CORA does very well on Answer Relevance but struggles with Answer Faithfulness. Furthermore, CORA scores well on Context Relevance in most languages, but results vary somewhat. The same holds for Language Consistency. CORA does worst in Japanese across the board and best in Bengali.

5.4 Qualitative evaluation

MARS can be used to identify examples where current multilingual RAG systems perform poorly. A subset of these examples will be analyzed to see whether patterns can be identified in the mistakes made by RAG systems.

We use our CORA baseline LLM judge output to

Language	Pair	Accuracy	Ground truth	Pre-PPI score	MARS Score
bn	all	30.6%	0.16	0.86	0.20±0.08
	bn-bn	27.9%	0.16	0.88	0.17±0.07
	bn-en	26.1%	0.16	0.90	0.20±0.08
	en-bn	32.0%	0.16	0.84	0.23±0.08
	en-en	37.5%	0.16	0.79	0.19±0.08
fi	all	29.6%	0.19	0.90	0.22±0.08
	en-en	31.3%	0.19	0.88	0.19±0.08
	en-fi	34.9%	0.19	0.83	0.18±0.09
	fi-en	28.8%	0.19	0.91	0.21±0.08
	fi-fi	24.7%	0.19	0.95	0.23±0.07
ja	all	30.9%	0.05	0.74	0.06±0.08
	en-en	36.5%	0.05	0.69	0.09±0.08
	en-ja	27.9%	0.05	0.78	0.03±0.07
	ja-en	29.1%	0.05	0.76	0.05±0.08
	ja-ja	26.0%	0.05	0.79	0.06±0.08
ru	all	26.8%	0.13	0.86	0.28±0.08
	en-en	27.1%	0.13	0.86	0.27±0.08
	en-ru	29.7%	0.13	0.83	0.27±0.09
	ru-en	27.4%	0.13	0.86	0.23±0.08
	ru-ru	25.7%	0.13	0.87	0.23±0.08
te	all	28.7%	0.10	0.81	0.13±0.08
	en-en	35.0%	0.10	0.75	0.14±0.08
	en-te	31.3%	0.10	0.79	0.12±0.08
	te-en	22.0%	0.10	0.88	0.11±0.07
	te-te	25.9%	0.10	0.84	0.13±0.08

Table 5: MARS results on XOR-AttriQA.

see if we can find any relevant patterns where CORA underperforms.

5.4.1 Results of qualitative evaluation

At first glance, we cannot discern any notable patterns in the negatively labelled samples for Context Relevance and Answer Relevance.

When we look at the negatives for Answer Faithfulness, a significant portion consists of dates that are indeed not present in the retrieved context. This suggests that CORA struggles with hallucinations around dates. Also, a significant portion of the negatives are yes/no answers. Unfortunately, it is difficult for us to validate whether these are true negatives without knowledge of the relevant languages.

We find interesting patterns when looking at the samples MARS marked as negative for Language Consistency. In Japanese, they are almost exclusively (Western Arabic) numbers, names (both in Latin script and non-Japanese names in Japanese script) and mostly the words "yes" and "no". The non-Japanese names in Japanese script are false negatives. The yes/no answers, however, are true negatives. CORA seems overly conditioned on yes/no questions in English, making it answer Japanese yes/no questions with "yes" or "no" instead of the Japanese equivalent. This same pattern holds across most languages.

Language	CR	AR	AF	LC
ar	0.90±0.03	0.98±0.03	0.69±0.05	0.83±0.05
bn	0.89±0.04	0.95±0.04	0.70±0.06	1.03±0.06
fi	0.92±0.04	0.95±0.04	0.73±0.05	0.83±0.06
ja	0.79±0.04	0.96±0.04	0.62±0.06	0.94±0.06
ko	0.83±0.05	0.95±0.04	0.68±0.06	1.03±0.06
ru	0.85±0.04	0.97±0.03	0.68±0.05	0.82±0.06
te	0.85±0.04	0.97±0.04	0.67±0.06	1.04±0.06

Table 6: MARS evaluation of CORA on XOR-TyDi QA.

In Arabic, we see some additional false negatives in the form of sentences. In Bengali, we see a few English answers that are not numbers or names but with no discernable pattern. In Finnish, we see a few answers in a third language, neither the language of the passage nor the question, which is interesting since it was unexpected. Korean and Russian hold to the pattern we see across languages with nothing further of note. In Telugu, apart from all other languages, we interestingly never see the word "yes".

We do not notice a difference in language consistency errors between cases where the system retrieved a document in the target language compared to a different language.

Unfortunately, it is difficult for us to do a more comprehensive qualitative analysis without knowledge of the languages in the dataset. We could have bulk-translated everything into English for analysis, but this was considered out of scope.

6 Conclusion

We adapt ARES by swapping out monolingual components for multilingual ones and expand the synthetic data generation step to generate multilingual

data. We add a Language Consistency metric to measure a common challenge in multilingual RAG systems. Additionally, we make some improvements to ARES unrelated to multilingualism. Together, these adaptions form MARS. We evaluate MARS in our experiments and validate our choices. We get promising results, generalizing the results from ARES to cross-lingual scenarios. While the scores given by MARS might not be entirely reliable by themselves, especially if the human reference set is not a good reflection of real-world sample distribution, MARS can rank systems accurately. With the creation and meta-evaluation of MARS, we answer RQ1.

Additionally, we explore the properties of different multilingual datasets and their role in evaluating multilingual RAG systems, primarily based on literature study, and provide an overview of available multilingual QA datasets, answering RQ2. We find that not all desired dataset properties can coexist and must be weighed against each other for each use case. Furthermore, we run into a lack of human-labelled data, leading to a reliance on artificial negatives. Nonetheless, our experiments show that our considerations for dataset choice can lead to valuable insights into the performance of our new evaluation system.

Using MARS system, we are also able to gain some insights into the limitations of current RAG systems, especially regarding LC (Language Consistency), concluding that they struggle with numbers, names and yes/no questions, especially across scripts. This answers RQ3.

By introducing MARS, the first system that can extensively evaluate multilingual RAG systems across a range of metrics, we allow researchers and practitioners to better understand otherwise obscure performance differences across different systems. Additionally, MARS can aid in the choices for either developing multilingual RAG systems or choosing existing systems, allowing the user to weigh the importance of different aspects of the system. It can point towards directions of potential improvements of RAG systems while highlighting in which areas a system already does well. Furthermore, it allows quantifying the language consistency of multilingual RAG systems, a recognized area of concern but previously underexplored.

By using MARS, practitioners can track the performance of their RAG systems across their development cycle, as well as changing use context. It fits nicely into the MLOps paradigm, allowing the scoring of new iterations of RAG systems as they are deployed. When the questions fed to MARS for scoring are updated periodically, MARS can aid in detecting concept drift.

6.1 Limitations

The most significant limitation in our research is our mock RAG system output to validate MARS. While using such mock RAG systems would not be a problem per se, our mock RAG systems are not representative of real-world systems. The negatives are created artificially instead of sourced from real RAG systems, likely making the samples artificially easy. While our results suggest that MARS can rank RAG systems well, its efficacy in real-world scenarios is yet to be proven, with the exception of Answer Faithfulness, for which we do have real-world data, albeit not much.

Furthermore, we use synthetic data to train our LLM judges. Again, this means the data does not necessarily represent real-world questions and answers. Especially the negatives, which are created similarly as for the mock RAG systems, are unrepresentative. The quality of the synthetic questions is unclear but could be judged using metrics like BLEU scores. Using synthetic training data is not necessarily an issue if it leads to well-trained classifiers, but as discussed in the previous paragraph, this is not easy to validate. The synthetic positive answers are also an issue since creating them relies on the ability of the LLM used to generate the synthetic data to give a correct answer given the question and the context, which is precisely the issue we are trying to measure. This potentially leads to incorrect positive samples. Additionally, due to how the mock RAG systems and synthetic data are created, they include answers that are numbers or names which are labelled as negative for Language Consistency. Since names and numbers are often the same across languages, these are mislabeled in both training and evaluation data. All in all, this means our synthetic data most likely has

incorrect positives and artificially easy negatives.

When generating the synthetic data, retrieval using an embedding model can be used to validate the quality of the question. When the retriever does not retrieve the passage used to generate the question as the top result, we discard the question. Unfortunately, during a very late stage of the research, we discovered a mistake in the codebase of ARES, which was also transferred to MARS, meaning this validation was not performed in our experiments. Additionally, just like with the answer generation, like in the section above, such retrieval is something we are trying to measure. Filtering out questions that do not retrieve the corresponding passage as the top result also means filtering out potentially good questions which are difficult to retrieve context for. This leads to our synthetic dataset only having relatively easy questions for retrieval.

In the implementation of the synthetic data generation as used in our experiments, the language of the passage is provided to the system. In practice, given a set of passages, the language is not always known. The system does not use the language directly, only to balance the training data across languages. While the system could easily be used without knowing the language of the documents, the impact of unbalanced training data is currently unknown.

We largely base our methodology and codebase on ARES. Unfortunately, during the development of MARS, we have found several errors in the ARES codebase, as well as discrepancies between the codebase and the methodology in the ARES paper. While we have carefully reviewed the parts of the ARES codebase we base our system on and are confident we have resolved any errors in logic, we cannot guarantee a proper comparison with ARES. While we would have liked to provide a performance comparison between MARS and ARES to show the impact of our changes on evaluation performance, we have been unable to get in contact with the authors to get more information and details on their experimental setup.

Lastly, we also inherit the system limitations from ARES, namely the need for expert labelling for the human reference set in specialized domains as well as the need for large GPUs for the generation of synthetic data and training of LLM judges.

6.2 Future work

In our experiments, we only compared MARS with real-world human judgement for Answer Faithfulness on a single dataset for a single RAG system. To more robustly validate the performance of MARS, it should be compared to human judgement for the other three metrics, preferably across various datasets and cross-lingual RAG systems. Such human judgements can be collected by taking samples from a real-world RAG system, for instance by applying it to one of the datasets considered in this work, and then human labelling each sample on each of the four metrics. This work's mock RAG system experiment can be repeated with real-world negatives instead of artificially created ones, providing a more robust validation.

While our qualitative evaluation is somewhat limited, we do believe that MARS can be used to identify where RAG systems struggle. To gather further insight, it would be interesting to make perturbations in test datasets used for such qualitative evaluations, such as limiting which information is available in which language. This way, differences in performance between different scenarios can be explored, such as when the system has to retrieve information cross-lingually as compared to monolingually.

An interesting direction for further improvement of MARS is reconsidering the synthetic data generation, especially the sampling of negatives. While using a synthetic dataset to train the LLM judges is one of the strengths of MARS since it removes the need for a large labelled dataset to use the system, the current process produces data which is not necessarily representative of real-world data. Further research can be put into how to generate representative synthetic negatives. A strategy for dealing with names and numbers for Language Consistency could also be helpful. A better way to validate the quality of synthetic questions can improve the quality of the synthetic dataset. Additionally, the quality of synthetic answers is currently not validated at all, which is also a good area for further improvement.

MARS as it is currently implemented is limited to

bilingual scenarios. It can, however, easily be expanded to handle cases where more than three languages are relevant. It would be interesting to explore the impact of including more language on both the performance of RAG systems, as well as MARS itself.

In our experiments, we only considered a single choice of LLM for the synthetic data generation and LLM judges, respectively. The choice of LLM could significantly impact the system's performance. Hence, it is a worthwhile direction for future work to explore different LLMs choices for both steps.

As mentioned in section 2, while conducting our research, RAGChecker was released. We believe RAGChecker to be a potentially more solid and extensive evaluation framework than ARES. It can be used both by developers and users, having separate metrics for both use cases, and does not rely on synthetic data. We believe a valuable research direction is to adapt RAGChecker to cross-lingual scenarios.

6.3 Recommendations

While the validation done on MARS in this work has some limitations, we believe MARS to be usable out-of-the-box on the comparison of two or more multilingual RAG systems. However, before MARS scores can be confidently interpreted on their own, the system should be more robustly validated on human-labelled data. Constructing a strong human reference set reflecting the real-world distribution of questions is important, as this significantly impacts the scores' accuracy.

Acronyms

AF Answer Faithfulness.
AR Answer Relevance.
ARES Automatic RAG Evaluation System.
CR Context Relevance.
LC Language Consistency.
LLM Large Language Model.
LM Language Model.
MARS Multilingual (Automatic) Assessment of RAG Systems.
NLI Natural Language Inference.
PPI Prediction-Powered Inference.
QA Question Answering.
RAG Retrieval-Augmented Generation.
RAGAs Retrieval-Augmented Generation Assessment.

References

Angelopoulos, Anastasios N., Stephen Bates, Clara Fannjiang, Michael I. Jordan and Tijana Zrnic (2023). 'Prediction-powered inference'. In: *Science* 382.6671 (10th Nov. 2023). Publisher: American Association for the Advancement of Science, pp. 669–674. DOI: 10.1126/science.adi6000. URL: https://www.science.org/doi/10.1126/science.adi6000.

Artetxe, Mikel, Sebastian Ruder and Dani Yogatama (2020). 'On the Cross-lingual Transferability of Monolingual Representations'. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter and Joel Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 4623–4637. DOI: 10.18653/v1/2020.acl-main.421. URL: https://aclanthology.org/2020.acl-main.421.

Aryabumi, Viraat, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin et al. (2024). *Aya 23: Open Weight Releases to Further Multilingual Progress*. 31st May 2024. DOI: 10.48550/arXiv.2405.15032. arXiv: 2405.15032[cs]. URL: http://arxiv.org/abs/2405.15032 (visited on 07/10/2024).

Asai, Akari, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi and Hannaneh Hajishirzi (2021a). 'XOR QA: Cross-lingual Open-Retrieval Question Answering'. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL-HLT 2021. Ed. by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy,

Steven Bethard et al. Online: Association for Computational Linguistics, June 2021, pp. 547–564. DOI: 10.18653/v1/2021.naacl-main.46. URL: https://aclanthology.org/2021.naacl-main.46.

Asai, Akari, Xinyan Yu, Jungo Kasai and Hanna Hajishirzi (2021b). 'One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval'. In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 7547–7560. URL: https://proceedings.neurips.cc/paper/2021/hash/3df07fdae1ab273a967aaa1d355b8bb6-Abstract.html.

Chen, Jianlv, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian and Zheng Liu (2024). *BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation*. 28th June 2024. DOI: 10.48550/arXiv.2402.03216. arXiv: 2402.03216[cs]. URL: http://arxiv.org/abs/2402.03216 (visited on 07/10/2024).

Chen, Jiawei, Hongyu Lin, Xianpei Han and Le Sun (2024). 'Benchmarking Large Language Models in Retrieval-Augmented Generation'. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 38.16 (24th Mar. 2024). Number: 16, pp. 17754–17762. ISSN: 2374-3468. DOI: 10.1609/aaai.v38i16.29728. URL: https://ojs.aaai.org/index.php/AAAI/article/view/29728.

Chi, Zewen, Shaohan Huang, Li Dong, Shuming Ma, Bo Zheng, Saksham Singhal et al. (2022). 'XLM-E: Cross-lingual Language Model Pre-training via ELECTRA'. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2022. Ed. by Smaranda Muresan, Preslav Nakov and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6170–6182. DOI: 10.18653/v1/2022.acl-long.427. URL: https://aclanthology.org/2022.acl-long.427.

Chung, Hyung Won, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus et al. (2024). 'Scaling Instruction-Finetuned Language Models'. In: *Journal of Machine Learning Research* 25.70 (2024), pp. 1–53. ISSN: 1533-7928. URL: http://jmlr.org/papers/v25/23-0870.html.

Clark, Jonathan H., Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev et al. (2020). 'TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages'. In: *Transactions of the Association for Computational Linguistics* 8 (2020). Ed. by Mark Johnson, Brian Roark and Ani Nenkova. Place: Cambridge, MA Publisher: MIT Press, pp. 454–470. DOI: 10.1162/tacl_a_00317. URL: https://aclanthology.org/2020.tacl-1.30.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán et al. (2020). 'Unsupervised Cross-lingual Representation Learning at Scale'. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter and Joel Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: https://aclanthology.org/2020.acl-main.747.

Duarte, Fabio (2024). *Number of ChatGPT Users (Jun 2024)*. Exploding Topics. 8th June 2024. URL: https://explodingtopics.com/blog/chatgpt-users (visited on 03/07/2024).

Es, Shahul, 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*. Ed. by Nikolaos Aletras and Orphee De Clercq. St. Julians, Malta: Association for Computational Linguistics, Mar. 2024, pp. 150–158. URL: https://aclanthology.org/2024.eacl-demo.16.

Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi et al. (2024). *Retrieval-Augmented Generation for Large Language Models: A Survey*. 4th Jan. 2024. DOI: 10.48550/arXiv.2312.10997. arXiv: 2312.10997[cs]. URL: http://arxiv.org/abs/2312.10997 (visited on 06/02/2024).

Gartner Poll Finds 55% of Organizations are in Piloting or Production Mode with Generative AI (2023). Gartner. 3rd Oct. 2023. URL: https://www.gartner.com/en/newsroom/press-releases/2023-10-03-gartner-poll-finds-55-percent-of-organizations-are-in-piloting-or-production-mode-with-generative-ai (visited on 03/07/2024).

Greene, Ryan, Ted Sanders, Lilian Weng and Arvind Neelakantan (2022). *New and improved embedding model*. 15th Dec. 2022. URL: https://openai.com/index/new-and-improved-embedding-model/ (visited on 12/12/2024).

Hardalov, Momchil, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev and Preslav Nakov (2020). 'EXAMS: A Multi-subject High School Examinations Dataset for Cross-lingual and Multilingual Question Answering'. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. EMNLP 2020. Ed. by Bonnie Webber, Trevor Cohn, Yulan He and Yang Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 5427–5444. DOI: 10.18653/v1/2020.emnlp-main.438. URL: https://aclanthology.org/2020.emnlp-main.438.

He, Pengcheng, Jianfeng Gao and Weizhu Chen (2021). *DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing*. arXiv.org. 18th Nov. 2021. URL: https://arxiv.org/abs/2111.09543v4 (visited on 07/10/2024).

Hu, Junjie, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat and Melvin Johnson (2020). 'XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation'. In: *Proceedings of the 37th International Conference on Machine Learning*. International Conference on Machine Learning. ISSN: 2640-3498. PMLR, 21st Nov. 2020, pp. 4411–4421. URL: https://proceedings.mlr.press/v119/hu20b.html.

Hu, Xiangkun, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu et al. (2024). *RefChecker: Reference-based Fine-grained Hallucination Checker and Benchmark for Large Language Models*. 23rd May 2024. DOI: 10.48550/arXiv.2405.14486. arXiv: 2405.14486. URL: http://arxiv.org/abs/2405.14486 (visited on 14/10/2024).

Huber, Patrick, Armen Aghajanyan, Barlas Oguz, Dmytro Okhonko, Scott Yih, Sonal Gupta et al. (2022). 'CCQA: A New Web-Scale Question Answering Dataset for Model Pre-Training'. In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Findings 2022. Ed. by Marine Carpuat, Marie-Catherine de Marneffe and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 2402–2420. DOI: 10.18653/v1/2022.findings-naacl.184. URL: https://aclanthology.org/2022.findings-naacl.184.

Johnson, Jeff, Matthijs Douze and Hervé Jégou (2019). 'Billion-Scale Similarity Search with GPUs'. In: *IEEE Transactions on Big Data* 7.3 (2019). Conference Name: IEEE Transactions on Big Data, pp. 535–547. ISSN: 2332-7790. DOI: 10.1109/TBDATA.2019.2921572. URL: https://ieeexplore.ieee.org/document/8733051.

Kendall, M. G. (1938). 'A New Measure of Rank Correlation'. In: *Biometrika* 30.1 (1938). Publisher: [Oxford University Press, Biometrika Trust], pp. 81–93. ISSN: 0006-3444. DOI: 10.2307/2332226. URL: https://www.jstor.org/stable/2332226.

Lewis, Patrick, Barlas Oguz, Ruty Rinott, Sebastian Riedel and Holger Schwenk (2020a). 'MLQA: Evaluating Cross-lingual Extractive Question Answering'. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter and Joel Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 7315–7330. DOI: 10.18653/v1/2020.acl-main.653. URL: https://aclanthology.org/2020.acl-main.653.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal et al. (2020b). 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks'. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associ-

ates, Inc., 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Li, Xiaoqian, Ercong Nie and Sheng Liang (2023). *From Classification to Generation: Insights into Crosslingual Retrieval Augmented ICL*. 2nd Dec. 2023. DOI: 10.48550/arXiv.2311.06595. arXiv: 2311.06595[cs]. URL: http://arxiv.org/abs/2311.06595 (visited on 14/02/2024).

Liang, Yaobo, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi et al. (2020). 'XGLUE: A New Benchmark Dataset for Cross-lingual Pre-training, Understanding and Generation'. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. EMNLP 2020. Ed. by Bonnie Webber, Trevor Cohn, Yulan He and Yang Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 6008–6018. DOI: 10.18653/v1/2020.emnlp-main.484. URL: https://aclanthology.org/2020.emnlp-main.484.

Liu, Jiahua, Yankai Lin, Zhiyuan Liu and Maosong Sun (2019). 'XQA: A Cross-lingual Open-domain Question Answering Dataset'. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL 2019. Ed. by Anna Korhonen, David Traum and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 2358–2368. DOI: 10.18653/v1/P19-1227. URL: https://aclanthology.org/P19-1227.

Longpre, Shayne, Yi Lu and Joachim Daiber (2021). 'MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering'. In: *Transactions of the Association for Computational Linguistics* 9 (2021). Ed. by Brian Roark and Ani Nenkova. Place: Cambridge, MA Publisher: MIT Press, pp. 1389–1406. DOI: 10.1162/tacl_a_00433. URL: https://aclanthology.org/2021.tacl-1.82.

Madzou, Lofred (2024). *What is the RAG Triad?* TruEra. URL: https://truera.com/ai-quality-education/generative-ai-rags/what-is-the-rag-triad/ (visited on 14/10/2024).

Muller, Benjamin, John Wieting, Jonathan Clark, Tom Kwiatkowski, Sebastian Ruder, Livio Soares et al. (2023). 'Evaluating and Modeling Attribution for Cross-Lingual Question Answering'. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2023. Ed. by Houda Bouamor, Juan Pino and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 144–157. DOI: 10.18653/v1/2023.emnlp-main.10. URL: https://aclanthology.org/2023.emnlp-main.10.

O'Rourke, Bernadette and Sara C. Brennan (2023). 'Multilingualism in the Workplace'. In: *Intercultural Issues in the Workplace: Leadership, Communication and Trust*. Ed. by Katerina Strani and Kerstin Pfeiffer. Cham: Springer International Publishing, 2023, pp. 179–191. ISBN: 978-3-031-42320-8. DOI: 10.1007/978-3-031-42320-8_12. URL: https://doi.org/10.1007/978-3-031-42320-8_12 (visited on 30/01/2025).

Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu (2002). 'BLEU: a method for automatic evaluation of machine translation'. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL '02. USA: Association for Computational Linguistics, 6th July 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://doi.org/10.3115/1073083.1073135.

Ru, Dongyu, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang et al. (2024). *RAGChecker: A Fine-grained Framework for Diagnosing Retrieval-Augmented Generation*. 16th Aug. 2024. DOI: 10.48550/arXiv.2408.08067. arXiv: 2408.08067[cs]. URL: http://arxiv.org/abs/2408.08067 (visited on 01/10/2024).

Saad-Falcon, Jon, Omar Khattab, Christopher Potts and Matei Zaharia (2024). *ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems*. 31st Mar. 2024. DOI: 10.48550/arXiv.2311.09476. arXiv: 2311.09476[cs]. URL: http://arxiv.org/abs/2311.09476 (visited on 06/05/2024).

Sen, Priyanka, Alham Fikri Aji and Amir Saffari (2022). 'Mintaka: A Complex, Natural, and Multi-

lingual Dataset for End-to-End Question Answering'. In: *Proceedings of the 29th International Conference on Computational Linguistics*. COLING 2022. Ed. by Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi et al. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 1604–1619. URL: https://aclanthology.org/2022.coling-1.138.

Shen, Xiaoyu, Akari Asai, Bill Byrne and Adria De Gispert (2023). 'xPQA: Cross-Lingual Product Question Answering in 12 Languages'. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*. ACL 2023. Ed. by Sunayana Sitaram, Beata Beigman Klebanov and Jason D Williams. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 103–115. DOI: 10.18653/v1/2023.acl-industry.12. URL: https://aclanthology.org/2023.acl-industry.12.

Singhal, Rahul (2023). *Council Post: The Power Of RAG: How Retrieval-Augmented Generation Enhances Generative AI*. Forbes. Section: Innovation. 30th Nov. 2023. URL: https://www.forbes.com/sites/forbestechcouncil/2023/11/30/the-power-of-rag-how-retrieval-augmented-generation-enhances-generative-ai/ (visited on 07/06/2024).

Smith, Matthew S. (2024). *AI chatbots spew out nonsense too often. But there's a solution: retrieval-augmented generation.* Business Insider. 16th May 2024. URL: https://www.businessinsider.com/retrieval-augmented-generation-making-ai-language-models-better-2024-5 (visited on 07/06/2024).

Sorokin, Nikita, Dmitry Abulkhanov, Irina Piontkovskaya and Valentin Malykh (2022). 'Ask Me Anything in Your Native Language'. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL-HLT 2022. Ed. by Marine Carpuat, Marie-Catherine de Marneffe and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 395–406. DOI: 10.18653/v1/2022.naacl-main.30. URL: https://aclanthology.org/2022.naacl-main.30.

Tucker, G. Richard (1999). *A Global Perspective on Bilingualism and Bilingual Education. ERIC Digest*. ERIC Number: ED435168. ERIC/CLL, 4646 40th Street, NW, Washington, DC 20016-1859, Aug. 1999. URL: https://eric.ed.gov/?id=ED435168 (visited on 18/11/2024).

Uspenskyi, Serhii (2024). *Large Language Model Statistics And Numbers (2024) - Springs*. 27th Feb. 2024. URL: https://springsapps.com/knowledge/large-language-model-statistics-and-numbers-2024,%20https://springsapps.ai/blog/large-language-model-statistics-and-numbers-2024 (visited on 03/07/2024).

Üstün, Ahmet, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude et al. (2024). *Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model*. 12th Feb. 2024. DOI: 10.48550/arXiv.2402.07827. arXiv: 2402.07827[cs]. URL: http://arxiv.org/abs/2402.07827 (visited on 02/05/2024).

Wu, Kevin, Eric Wu and James Zou (2024). *ClashEval: Quantifying the tug-of-war between an LLM's internal prior and external evidence*. 10th June 2024. DOI: 10.48550/arXiv.2404.10198. arXiv: 2404.10198. URL: http://arxiv.org/abs/2404.10198 (visited on 18/11/2024).

Yue, Xiang, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su and Huan Sun (2023). 'Automatic Evaluation of Attribution by Large Language Models'. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Findings 2023. Ed. by Houda Bouamor, Juan Pino and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 4615–4635. DOI: 10.18653/v1/2023.findings-emnlp.307. URL: https://aclanthology.org/2023.findings-emnlp.307.

A Synthetic data examples

In Table 7, we show some positive and negative examples from the synthetic dataset. The documents shown are passages from the in-domain corpus.

B Few-shot prompts

B.1 Synthetic questions

The few-shot prompt for generating synthetic questions is as follows, where the first part is repeated `n` times and taken from the few-shot examples:

```
Example n
Document ([passage_language]):  [passage]
Question ([question_language]):  [question]
Example n+1
Document ([passage_language]):  [passage]
Question ([question_language]):
```

B.2 Synthetic answers

The few-shot prompt for generating synthetic answers is as follows, where the first part is repeated `n` times and taken from the few-shot examples:

```
Example n
Document ([passage_language]):  [passage]
Question ([question_language]):  [question]
Answer ([answer_language]):  [answer]
Example n+1
Document ([passage_language]):  [passage]
Question ([question_language]):  [question]
Answer ([answer_language]):
```

This prompt is also used to generate negatives for Language Consistency; the examples are simply swapped with answers in another language, and the language labels are adjusted accordingly.

C Mock RAG examples

In Table 8, we show some positive and negative examples from the synthetic mock RAG systems generated from the MLQA dataset. As discussed, we create four different datasets, one for each passage-question language pair. The dataset from which the sample is taken is shown in the first column. The datasets are fully parallel, meaning each sample exists for each language pairing. We give samples from the different datasets for illustration.

Passage	Question	Answer	CR	AR	AF	LC
Ecuador annexed the Galápagos Islands on 12 February 1832, naming them the Archipelago of Ecuador. This new name added to several names that had been, and are still, used to refer to the archipelago. The first governor of Galápagos, General José de Villamil, brought a group of convicts to populate the island of Floreana, and in October 1832, some artisans and farmers joined them.	What was the name of the first governor of Galapagos?	General José de Villamil	Yes	Yes	Yes	Yes
Zur Gewährleistung hoher Qualitätsstandards wird ein vergleichsweise kompliziertes Verfahren angewandt. Zunächst wird von der Bookerpreisstiftung ein Beirat berufen, der einzig die Aufgabe hat, die jedes Jahr neu zu bestimmenden Juroren zu küren. In diesem Beirat sitzen obligatorisch: ein Vertreter der Schriftsteller, zwei Verleger, ein Literaturagent, ein Buchhändler, ein Bibliothekar sowie ein Moderator und Vorsitzender aus der Stiftung selbst. Die Juroren werden ausgewählt aus den Meinungsführern der Literaturkritiker, Schriftsteller, Literaturwissenschaftler und Persönlichkeiten des öffentlichen Lebens. Mehrfache Nominierungen als Jurymitglied sind über die Jahre eher die Ausnahme als der Regelfall geblieben.	What does Alice find in the rabbit's hole?	a room with many doors	No	NaN	No	Yes
Levantine Art was first discovered in Teruel in 1903. The Spanish prehistorian Juan Cabre was the first to study this art, defining it as a regional Palaeolithic art. Assessment as Palaeolithic was challenged for various reasons including the fact that no glacial fauna was depicted. Antonio Beltrán Martínez and others place the beginning of this art in the Epipaleolithic or Mesolithic, placing its heyday in the Neolithic period. Accepting a post-Paleolithic age for the art, Ripio devised a new chronological scheme in the 1960s, dividing the art into four stages:naturalistic, stylized static, stylized dynamic, and final phase of transition to the schematic.	who was the first to study this art?	Grenfell died	NaN	No	NaN	Yes
Wie in anderen europäischen Ländern kam es auch in Spanien nach dem Zweiten Weltkrieg, aus dem Franco das Land heraushalten konnte, zu einem langen wirtschaftlichen Nachkriegsboom. 1947 restaurierte Franco die Monarchie und ernannte Juan Carlos I. 1969 als Staatsoberhaupt zu seinem Nachfolger. Dieser leitete nach dem Tod des Diktators am 20. November 1975 einen Demokratisierungsprozess (span. Transición) ein. Durch die Verabschiedung einer Verfassung wurde Spanien 1978 zu einer parlamentarischen Monarchie. In der Endphase der Diktatur Francos und besonders während der Transition kam es zu massiven Terroraktionen der ETA und anderer linker wie auch rechter Terrorgruppen. Im Jahr 1981 erfolgte noch einmal ein Putschversuch („23-F“) von rechten Militärs und Teilen der paramilitärischen Guardia Civil gegen die demokratische Regierung.	How many fish species are there?	3 000	No	NaN	No	Yes
Am 26. Juli 1970 heiratete er Romina Power, die Tochter des Schauspielers Tyrone Power. Aus der Ehe gingen vier Kinder hervor: Cristel, geboren 1985, Romina Jr., geboren 1987, Yari, geboren 1973 und die älteste Tochter Ylenia Carrisi, geboren 1970. Die beiden wurden auch beruflich ein Paar; 1969 nahm er mit ihr Cori di Acqua di mare und 1970 Storia di due innamorati auf.	Wie viele Kinder hat Al Bano mit seiner ersten Frau?	22.8% were non-families	NaN	No	No	No

Table 7: Samples of synthetic data.

Dataset	Passage	Question	Answer	CR	AR	AF	LC
en-en	WWF Tag Team Championship (4 times) – with Shawn Michaels (1), Dude Love (1), The Undertaker (1), and Triple H (1)	What was the name of the third wrestler that made up the WWF Tag Team?	The Undertaker	Yes	Yes	Yes	Yes
en-en	Analysis of DNA is consistent with the hypothesis that Sumatran tigers became isolated from other tiger populations after a rise in sea level that occurred at the Pleistocene to Holocene border about 12,000–6,000 years ago. In agreement with this evolutionary history, the Sumatran tiger is genetically isolated from all living mainland tigers, which form a distinct group closely related to each other.	In EPIC what makes the decision of the order of the instructions?	compiler	No	NaN	NaN	NaN
de-en	Lawrence Lessig behauptet, dass Copyright ein Hindernis für kulturelle Produktion, Wissensverteilung und für technologische Innovation sei und dass dieses Gesetz nur auf private Interessen – entgegengesetzt zu öffentlichem Gut – abziele. Im Jahre 1998 reiste er durchs Land und gab hunderte Reden an Universitäten und entfachte somit die Bewegung. Dies führte zur Gründung des ersten Ortsverbands von Students for Free Culture am Swarthmore College.	What was founded at Swarthmore College?	ystävänpäivä	NaN	No	NaN	NaN
de-de	Der goldene Helm über dem Wappenschild ist ein Symbol der Souveränität Manitobas innerhalb der Kanadischen Konföderation. Helmdecke und Helmwulst sind beide in rot und weiß, den nationalen Farben Kanadas. Helmkleinod ist ein Biber, der in der rechten Pfote eine Kuhschelle (Anemone patens) hält, die offizielle Blume der Provinz. Auf seinem Rücken trägt er die Edwardskrone.	Was ist das offizielle Tier von Kanada?	Im September 2008	NaN	NaN	No	NaN
en-de	From the middle of the 19th century onwards, trade, industry and tourism gained momentum. Nevertheless, until the middle of the 20th century, agriculture dominated the canton. Today a great number of small and middle-sized businesses dominate the economy. The largest employer is the airplane constructor Pilatus. The small and middle-sized businesses work in a wide range of areas. Many specialize in machine construction, medical equipment, international trade, optics and electronics.	Wann begannen diese Wirtschaftssek zu wachsen?	19th century	NaN	NaN	NaN	No

Table 8: Samples of mock RAG systems.

D MARS usage manual

MARS is an evaluation tool that lets you score your multilingual RAG system on four different metrics: context relevance, answer relevance, answer faithfulness and language consistency. MARS works by using your own knowledge base to make sure its scores reflect your specific use case and only needs little labelled examples, reducing the need for you to create extensive test datasets. While the initial setup requires a GPU with considerable VRAM, using the system in practice can be done easily on conventional GPUs with just a few GBs of VRAM. MARS works in three steps: synthetic data generation, LLM judge training and RAG system evaluation. In order to use MARS to evaluate RAG systems, you will have to perform these three steps.

D.1 Data requirements

Before you start, you should ensure you have access to the right data to use MARS. MARS requires an in-domain corpus of passages, this would be the knowledge base you use for your RAG system. Secondly, MARS needs a few examples of questions that might be asked to your RAG system, as well as answers the RAG system would give. Additionally, it requires a labelled set of examples for each of the metrics from MARS you want to use. Ideally, this set would reflect the real-world distribution of questions asked to the RAG system and answers given by the RAG system, labelled as true or false for each metric. This set should contain at least around 150 examples, with a few hundred leading to better performance. You could construct this set by collecting data from your RAG system in practice and then human-labelling them. Lastly, to actually score your RAG system, MARS requires a set of (unlabelled) responses from your RAG system (including the question and context retrieved). Ideally, these are real-world questions, but benchmark questions can also be used.

D.2 Synthetic data generation

When you are sure you have the data you need, you first need to generate a synthetic dataset based on your own corpus. This requires access to a machine with considerable VRAM, as it includes running an LLM locally; we suggest using a VM with an 80GB A100 GPU, as this was used when developing MARS, so it is guaranteed to work. MARS requires a dataset of a few thousand questions, which should take a few hours to generate. A code example for the synthetic generation is shown below:

```
from mars import MARS

synth_config = {
    "document_filepaths": ["multilingual_data/mlqa_(de)_test_en_de.tsv"],
    "few_shot_prompt_filename": "multilingual_data/mlqa_(de)_test_few_shot_en_de.tsv",
    "synthetic_queries_filenames": ["multilingual_data/synthetic_queries_mlqa_(de)_test_en-de.tsv"],
    "documents_sampled": 3000,
    "model_choice": "CohereForAI/aya-23-35B",
    "document_language": "English",
    "query_language": "German",
}

mars = MARS(synthetic_query_generator=synth_config)
mars.generate_synthetic_data()
```

Currently, synthetic datasets have to be generated for each language pair separately.

D.3 LLM judge training

Now that you have your synthetic dataset, you can train your own LLM judges for each metric you want to use. Again, this requires a bit of compute and should take a few hours on an A100 GPU. A code example is shown below:

```
from mars import MARS

classifier_config = {
    "training_dataset": [
        "multilingual_data/synthetic_queries_mlqa_test_en-en.tsv",
        "multilingual_data/synthetic_queries_mlqa_test_en-de.tsv",
        "multilingual_data/synthetic_queries_mlqa_test_de-en.tsv",
        "multilingual_data/synthetic_queries_mlqa_test_de-de.tsv",
    ],
    "training_dataset_path": "multilingual_data/synthetic_queries_mlqa_test.tsv",
    "validation_set": ["multilingual_data/mlqa_(de)_dev_ratio_0.5_all.tsv"],
    "label_column": [
        "Context_Relevance_Label",
        "Answer_Relevance_Label",
        "Answer_Faithfulness_Label",
        "Language_Consistency_Label",
    ],
    "model_choice": "microsoft/mdeberta-v3-base",
    "num_epochs": 10,
    "patience_value": 3,
    "learning_rate": 5e-6,
    "assigned_batch_size": 32,
    "gradient_accumulation_multiplier": 32,
}

mars = MARS(classifier_model=classifier_config)
mars.train_classifier()
```

D.4 RAG system evaluation

With your LLM judges ready to go, you can now score your RAG system. This requires the sets of labelled and unlabelled responses. A code example is shown below:

```
from mars import MARS

ppi_config = {
    "evaluation_datasets": [
        "multilingual_data/mlqa_(de)_test_ratio_0.7_all.tsv",
    ],
    "checkpoints": [
        "checkpoints/microsoft-mdeberta-v3-base/Context_Relevance_Label_mlqa_dev_ratio_0.5_2024-09-30.pt",
        "checkpoints/microsoft-mdeberta-v3-base/Answer_Relevance_Label_mlqa_dev_ratio_0.5_2024-10-01.pt",
        "checkpoints/microsoft-mdeberta-v3-base/Answer_Faithfulness_Label_mlqa_dev_ratio_0.5_2024-10-02.pt",
        "checkpoints/microsoft-mdeberta-v3-base/Language_Consistency_Label_mlqa_dev_ratio_0.5_2024-10-02.pt",
    ],
    "labels": [
        "Context_Relevance_Label",
        "Answer_Relevance_Label",
```

```
        "Answer_Faithfulness_Label",
        "Language_Consistency_Label",
    ],
    "gold_label_paths": ["multilingual_data/mlqa_(de)_dev_ratio_0.5_all.tsv"],
    "model_choice": "microsoft/mdeberta-v3-base",
    "assigned_batch_size": 8,
    "prediction_filepaths": [
        "mlqa_(de)_test_ratio_0.7_all_output.tsv",
    ],
}

mars = MARS(ppi=ppi_config)
results = mars.evaluate_RAG()
print(results)
```

MARS will now label each sample, both labelled and unlabelled, and combine this information to compute a score for your RAG system. While these scores can be used as is to get an idea about the performance of your RAG system, they are best used for comparison between different RAG systems.

E Additional results

E.1 Arabic mock RAG graphs

Plots with results from the mock RAG experiments in an English-Arabic bilingual scenario are shown in Figure 3.

E.2 Mock RAG results with 70% correct human reference set

In Figure 4, plots for MARS scoring on the mock RAG systems is shown when a human reference set with 70% accuracy is used for PPI. We can clearly see that PPI now moves the scores higher than when the 50% set was used. It seems that in cases where the LLM judge is less accurate, it is important that the human reference set reflects the real-world distribution of samples. As can be seen in Table 9, the ranking performance remains unaffected. So, while a representative human reference set can aid in giving more accurate scores, it is less important for comparing different RAG systems.

		CR	AR	AF	LC
MLQA (de) all	τ 50%	0.89	1.00	0.76	0.83
	τ 70%	0.89	1.00	0.76	0.83
MLQA (de) de-de	τ 50%	0.83	0.89	0.70	0.83
	τ 70%	0.83	0.89	0.70	0.83
MLQA (de) de-en	τ 50%	0.83	0.94	0.50	0.83
	τ 70%	0.83	0.94	0.50	0.83
MLQA (de) en-de	τ 50%	1.00	0.89	0.76	0.94
	τ 70%	1.00	0.89	0.76	0.94
MLQA (de) en-en	τ 50%	0.94	0.94	0.83	0.76
	τ 70%	0.94	0.94	0.83	0.76

Table 9: Results of the mock RAG system ranking using a human reference set with 70% accuaracy. Results for 50% human reference set shown for comparison.

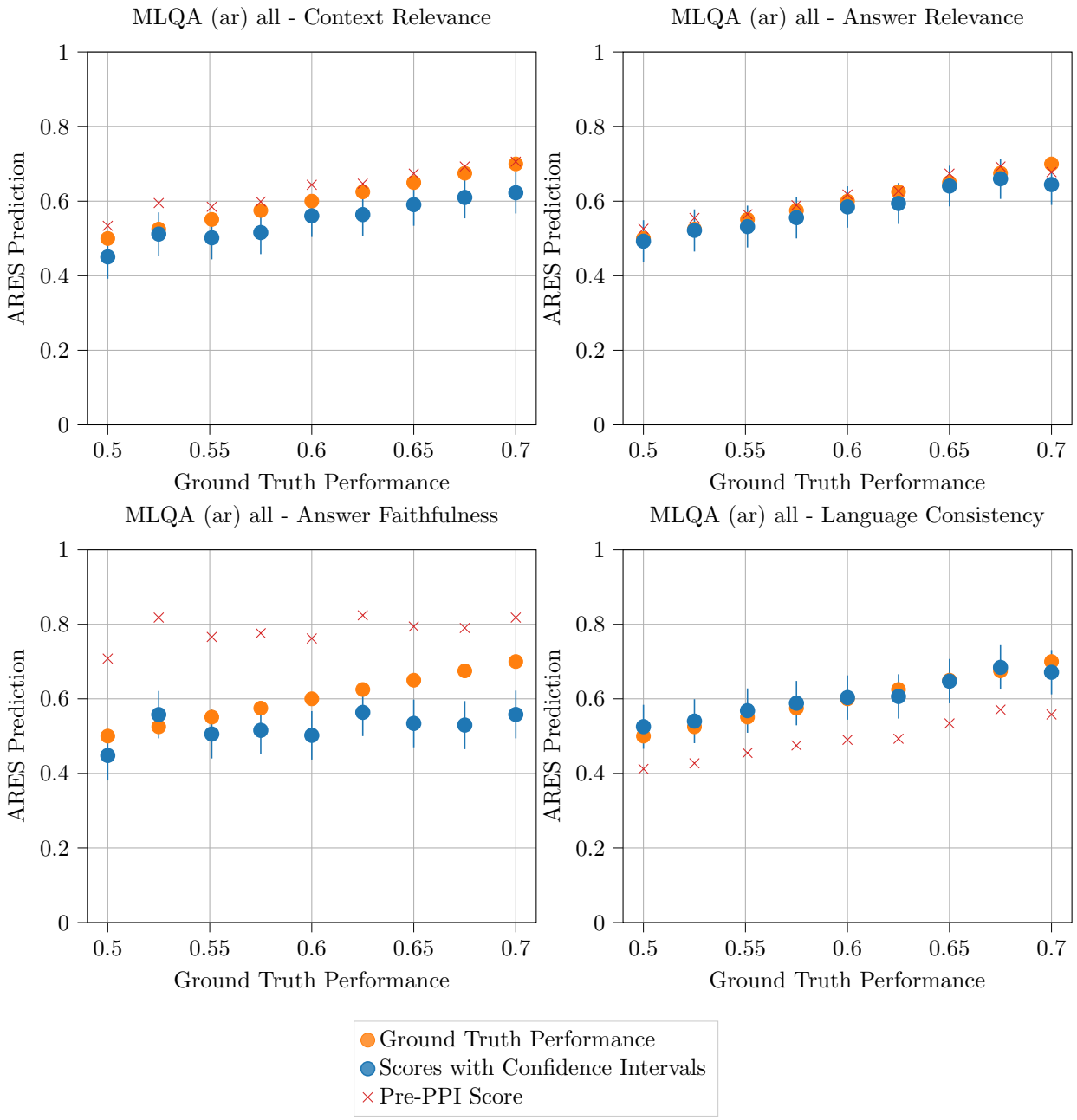

Figure 3: MARS output on mock RAG systems with mixed Arabic-English samples.

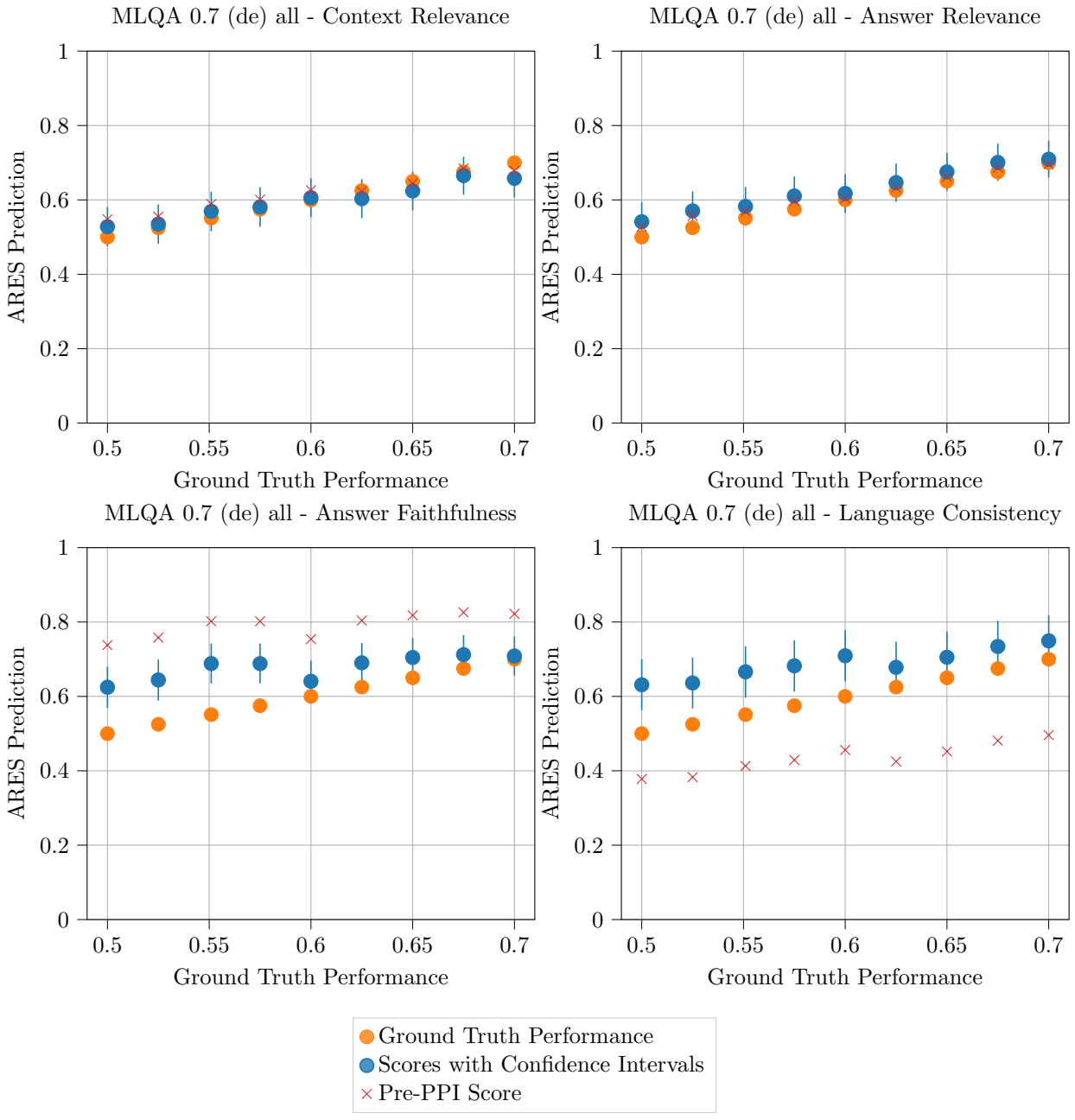

Figure 4: MARS output on mock RAG systems with mixed German-English samples using a human reference set with 70% accuaracy.