Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Cross-Lingual Information Access in the LLM Era: Architectures, Alignment Strategies, and Open Challenges for Low-Resource Languages

Siddhartha Neupane

stw0084@softwarica.edu.np

Softwarica College

Ganesh Bhusal

stw0019@softwarica.edu.np

Softwarica College

Sunil Thapa

stw00148@softwarica.edu.np

Softwarica College

Shrawan Thakur

stw0049@softwarica.edu.np

Softwarica College

Giriraj Rawat

Softwarica College

Research Article

Keywords: Cross-lingual information retrieval, MIR ACL, NoMIRACL, SK-languages, K-representations, semantic intermediary languages, XLM-RoBERTa, multilingual LLMs, retrieval-augmented generation, low-resource languages, linguistic equity, hallucination rate, architectural paradigms

Posted Date: July 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10530760/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Architectural Paradigms and Alignment Strategies in Cross-Lingual Intelligent Information Access: A Critical Analysis of LLM-Driven Approaches for Linguistically Diverse and Low-Resource Contexts

Siddhartha Neupane, Shrawan Thakur, Ganesh Bhusal, Sunil Thapa, Giriraj Rawat, Bimal Shrestha
Research and Innovation Department
Softwarica College of IT and E-Commerce
Dillibazar, Kathmandu, Nepal
Affiliated with Coventry Universiity
stw0084@softwarica.edu.np

***Abstract*— The way the system is designed to help people find information in languages has changed a lot. We used to rely on translation systems. Now we use neural networks and big language models that can understand many languages. This paper looks at how we got to this point and how these new systems work.**

We looked at some ideas from the past like the theory of K-representations and SK-languages, which were developed by Fomichov. We also looked at how people used to search for information using ontologies and how they expanded their searches using models. In addition, we considered how people queried information in languages on the web and how they used multilingual frameworks.

We used three datasets to test our ideas: MIRACL, XLM-RoBERTa benchmarks and NoMIRACL. What we found was interesting. We saw that there is a difference in how well these systems work for languages that have a lot of resources versus those that do not. We also found that the size of the Wikipedia corpus for a language does not necessarily determine how well the system works. Furthermore, we saw that the rates at which these systems make mistakes vary greatly across languages from 0.21 for English to 0.89 for Yoruba. This means that these systems work differently for different languages.

Our paper suggests that we need to consider three things when we design these systems: how transparent they are, how well they understand the meaning of words across languages and how fair they are, to all languages. We think that these three things are equally important and that they can help us make systems that work better for everyone. This is an extension of the idea of a language that Fomichov proposed but now we are applying it to the new systems that use big language models.

***Index Terms*— cross-lingual information retrieval; MIRACL; NoMIRACL; SK-languages; K-representations; semantic intermediary languages; XLM-RoBERTa; multilingual LLMs; retrieval-augmented generation; low-resource languages; linguistic equity; hallucination rate; architectural paradigms**

I. INTRODUCTION

There are more than 7,000 languages spoken around the world which makes it really tough for people to get the information they need when it is in their own language. One of the major problem that computers and artificial intelligence have not been able to solve is helping people get information from other languages. Cross-lingual intelligent information access systems or CLIIA systems let people ask questions in one language and then get answers from information that is written in another languages. This is where information retrieval, natural language processing and artificial intelligence system design all come together. CLIIA systems are very important because they can help people get the information they need no matter what language it is, in.

This field of study has been around for a time way before people started using neural networks. For decades people have been trying to create a language that can help us understand what words mean no matter what language we speak. This is an old idea in computer science and it is still something that people are working on today. The Theory of K-representations by Fomichov is one way that people have tried to make this idea a reality. Fomichov wrote about this in 2010 [1]. It is a very detailed and formal way of thinking about language. From Fomichovs Theory of K-representations we got something called SK-languages. These SK-languages are, like a set of rules that help us understand what words mean. They use math to break down what we mean when we say something and they can help us understand sentences and conversations. This is a great deal because it means that we can use machines to understand what people are saying even if they speak different languages [2].

Fomichov in the year 2017 demonstrates that SK-languages are really good for making intermediary languages. They are much better than the Abstract Meaning Representation formalism. This formalism is what most people use for parsing in Natural Language Processing (NLP) today. SK-languages offer a lot possibilities for creating these languages. They are very useful, for intermediary languages compared to the Abstract Meaning Representation (AMR) formalism [3].

At the same time, the rise of the Semantic Web gave rise to a new architectural style for cross-lingual information access: systems that turn natural language questions into formal questions over ontological knowledge bases. SemSearch a system from 2006 [4] made at the Knowledge Media Institute, The Open University, had five layers to turn keywords into formal queries. The system worked with ontologies and semantic metadata and the idea was

to make it easy for users by building queries, which is similar to what modern RAG systems aim for i.e. making things simple for users. SemSearch showed that complex semantic queries could be made easy for users. by hiding the complexity through automation.

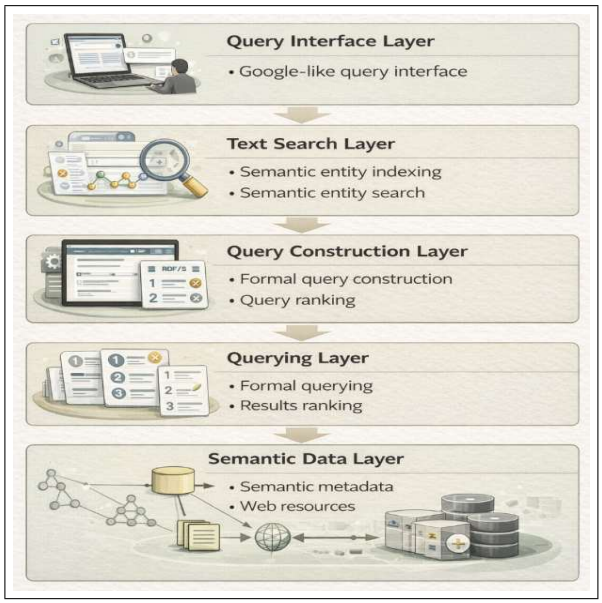


Fig. 1. SemSearch Five Layered Architecture [5]

The formal model for constructing semantic expansions of search requests proposed by Fomichov and Kirillov (2012) [6] extends this programme by addressing a class of requests those concerning achievements and failures of active systems for which no prior approach had been systematically formalised. Aggarwal et al. (2013) [7] addressed the cross-lingual extension of semantic querying to the Web of Data that presented a system for natural language querying across languages over linked data, by using deep linguistic analysis and cross-lingual semantic similarity measures to bridge the vocabulary gap between user queries and structured knowledge bases.

These basic ideas are based on a theory that the current LLM-based CLIIA methodology is an evolution of and deviates from. SK-languages and formal query systems aim to be easy to understand, clear in how the parts fit together, and have a good mathematical foundation. In contrast, the LLM-based systems are less interpretable, less stable across languages, and more biased toward high-resource languages. In this paper, the trade-offs of the LLM-based systems are explored with three publicly available benchmark datasets. It is argued that the solution to the trade-off between flexibility and interpretability lies in the combination of the two traditions.

Three research questions organise the analysis:

- **RQ1 (Architectural Trade-Offs):** To what extent do current CLIIA architectural paradigms differentially succeed or fail in balancing retrieval accuracy, semantic fidelity, and equitable performance across high-resource and low-resource language communities?
- **RQ2 (Linguistic Equity):** How do architectural dependencies shape and potentially perpetuate performance disparities between language communities, and which design interventions most effectively reduce them?
- **RQ3 (Evaluation Validity):** To what degree do current evaluation protocols and benchmark datasets adequately capture the full scope of CLIIA system quality, including dimensions such as cross-lingual semantic fidelity and cultural-pragmatic adequacy that standard metrics cannot measure?

The remainder of the paper is structured as follows. Section II reviews the theoretical and methodological background, including both the formal semantic intermediary tradition and the contemporary neural paradigm. Section III describes the datasets and analytical methodology. Section IV presents the data analysis and results. Section V proposes a tri-dimensional evaluative framework and design recommendations. Sections VI and VII address limitations and conclusions.

II. Theoretical and Methodological Background

A. *The Semantic Intermediary Tradition*

The basic idea behind this is that we can retrieve the information in any language using a neutral form of meaning representation. The idea behind this concept is quite old, dating back to the days of linguistic and artificial intelligence research. In the days of machine translation, the idea was that if we can represent the meaning of the source sentence in a neutral form, then representing it in any target language is a different task, which is quite achievable. The theory of K-representations (TKR), introduced by Fomichov in 1996, which was later published in a book by Springer in 2010 [1], presents the most detailed formal framework for this purpose. The class of formal languages defined by the set of ten partial operations over the conceptual structures, which include the basic set of concepts, is called the class of standard knowledge languages (SK-languages) [2]. The SK-languages can represent as much as the AMR, and they can represent the semantic representations of complex forms of infinitive, combined notions, and discourse.

According to Fomichov (2013), "the first version of a useful and flexible Conceptual Metagrammar (CM)" is the task of defining SK-languages. It is a set of rules to build semantic representations of any complex natural language sentences of professional areas of interest. In 2014, Fomichov proves that "SK-languages can be a Comprehensive Semantic Formal Environment (CSFE) for building a Multilingual Semantic Web." It is a plan that already complies with contemporary understanding of multilingual embedding alignment but uses a math-defined system of formal meaning rather than one learned from large sets of parallel texts statistically.

The complementary notion of ontology-driven semantic search, as embodied by systems such as SemSearch [4],

AquaLog, and others, sought to assist non-expert users in formulating semantic queries without necessarily being proficient in languages such as SPARQL. SemSearch employs a five-tiered architecture to accomplish this by automatically transforming keyword-based queries to formal queries. The tool leverages ontologies to retrieve entities, properties, and relationships. The multilingual problem of query formulation and processing was addressed by Aggarwal et al. (2013) [7], who extended the notion of semantic queries to support non-English languages. The authors leveraged the notion of cross-lingual similarity and relatedness measures to support the formulation of queries in languages other than the ones supported by the knowledge base.

The formal model for the semantic expansion of search queries proposed by Fomichov and Kirillov (2012) [6] is also of interest to the present work. In the case of complex query types such as the ones inquiring about the achievements and failures of active systems, the formal model for query expansion is more suitable. In such cases, the query cannot be adequately expressed in terms of the query words. The formal model leverages a knowledge base of goals of active systems to perform the query expansion. This way, the system is able to retrieve relevant documents without the query and the document having the same surface form. This work is a precursor to the more contemporary neural query expansion techniques.

B. The Neural Paradigm: From Translation Pipelines to LLMs

*1) **Translation-Pipeline Architectures**:* With the improvement in statistical MT and neural MT, the standard approach to CLIIA is to use monolingual retrieval followed by machine translation. The large problem, however, is error cascading, where the machine translation error becomes worse throughout the system, making it difficult to correct the error. The flaw illustrates the nature of the failure the semantic intermediary program was designed to correct, where the language-neutral representation of the semantics could, in theory, prevent error cascading because it provides meaning without regard to the words used. In practice, however, we have been unable to develop such representations on large scales for quite disparate languages without the statistical tools we used for the neural approach.

*2) **Multilingual Embedding Alignment**:* Multilingual BERT (mBERT) [8] and XLM-RoBERTa (XLM-R) [9] are models that use multilingual data to learn shared representations across languages without translation. XLM-R was trained on 2.5 terabytes of text data from CommonCrawl in 100 languages and achieved an average gain of 14.6 percentage points over mBERT in XNLI accuracy for all languages. These embedding alignment techniques are a practical solution for a middle ground for language: they learn a single space in which semantically equivalent phrases across languages are mapped to nearby locations without requiring a language of meaning. One of the limitations of these techniques, according to the theory of K-representations, is that the representations learned are statistical and therefore are not necessarily semantically equivalent; they are learned from the patterns in the data, which may bias the representations toward high-resource languages.

*3) **Dense Retrieval and Hybrid Fusion**:* Dense Retrieval employs two-encoder models, which facilitate fast approximate nearest neighbor search. Hybrid Fusion combines sparse (BM25) and dense (bi-encoder) methods using Reciprocal Rank Fusion (RRF). Hybrid Fusion always outperforms the individual sparse and dense baselines. The MIRACL baselines affirm this, as the hybrid model achieves the best average nDCG@10 across language groups, apart from low-resource languages, where in-language fine-tuned mDPR is competitive.

*4) **Retrieval-Augmented Generation**:* The RAG systems integrate retrieval with generative LLMs. This generates a novel type of error that doesn't exist with retrieval systems. This error type is called "output language drift," which occurs when the model generates responses in the wrong language due to the context being in a different language. The first large-scale analysis of this error type was conducted by NoMIRACL [10]. In the SK-language program, the error type of "language drift" indicates that there isn't a semantic interface layer. When the systems are in different languages with no formal meaning representation in between them, the confusion of languages isn't an error but a logical outcome.

C. Benchmark Infrastructure

MIRACL [11] is the most comprehensive multilingual retrieval benchmark, supporting 18 languages and having more than 726,000 human judgments of relevance by native speakers. MKQA [12] provides 10,000 question-answer pairs for 26 languages. NoMIRACL [10] extends the coverage by adding the evaluation of RAG hallucinations. However, there is a need for more. There is currently no benchmark for cultural pragmatic adequacy, one of the nine topics in the coverage analysis in the next section. This is a limitation of the current state of the art, which the semantic intermediary tradition would argue is fundamental. The point is that you cannot formally assess the degree of pragmatically correct cross-lingual information access without a formal model of pragmatic meaning, which none of the current benchmarks provides.

III. Data Sources and Methodology

A. Datasets

- **Dataset 1:** MIRACL [11]. Source: `huggingface.co/datasets/miracl/miracl`. The official published baseline results are reported for six different retrieval models: BM25, mDPR, Hybrid of BM25 + mDPR with Reciprocal Rank Fusion (RRF), mColBERT, mContriever, and in-language fine-tuned mDPR, over the

development set for 16 languages. Metrics: nDCG@10 (main) and Recall@100. The languages cover 10 language families, 8 writing systems, and 3 levels of resources: high, medium, and low.

- **Dataset 2:** XLM-RoBERTa Benchmarks [9]. Source: `arXiv:1911.02116`. Reported results for evaluating mBERT, XLM-R Base, and XLM-R Large over XNLI for 15 languages and MLQA for 7 languages. The results were extracted from Tables 1-4 of the paper.
- **Dataset 3:** NoMIRACL [10]. Source: `huggingface.co/datasets/miracl/nomiracl`. The metrics reported are the hallucination rate P(model claims an answer | no answer in passages) and the error rate P(model misses an answer | answer in passages) for 18 languages in an LLM-RAG setting over MIRACL passages.

B. *Analytical Approach*

The analysis process consists of three stages corresponding to the three research questions. In RQ1, we will calculate and visualize the nDCG@10 and Recall@100 metrics across all models and languages, grouped by resource levels. In RQ2, we will calculate the performance difference between high and low resource levels for each model, evaluate performance by language family groups, calculate the correlation of Wikipedia corpus size with performance using Pearson's r correlation coefficient, and analyze transfer gains of XLM-R as an additional equity indicator. In RQ3, we will show the joint distribution of NoMIRACL hallucination and error rates, compare RAGs across three concurrent dimensions, and map performance gaps in benchmark coverage.

IV. Data Analysis and Results

A. *RQ1: Retrieval Architecture Performance Across Languages*

1) ***nDCG@10 Across 16 Languages and 6 Models****:* Figure 1 shows the nDCG@10 results for all six MIRACL baseline models across 16 languages. The color of each bar corresponds to the level of the resources: dark blue for high resources, mid-blue for medium resources, and red for low resources. The results show three important trends that are consistent with the theory outlined in Section 2.

One of the trends is that the results for the Hybrid model consistently outperform both BM25 and mDPR for high- and medium-resource languages. This reflects the idea of complementarity that is a fundamental aspect of the semantic intermediary approach: there is no single level of representation that captures all aspects of meaning relevant to retrieval. Another important trend is that the results for the in-language fine-tuned mDPR approach are particularly strong for the low-resource languages: Yoruba (yo), Swahili (sw), and Telugu (te). This reflects the SK language program's view of language-specific concepts as the foundation for a semantically correct approach to representation. Finally, we note that the results for BM25 are unusually low for the high-resource languages: Chinese (zh) and French (fr). This reflects the well-known difficulties of BM25 tokenisation for languages written in characters rather than alphabets—a problem that a formal semantic representation, operating at a level above the lexical/word level, would overcome.

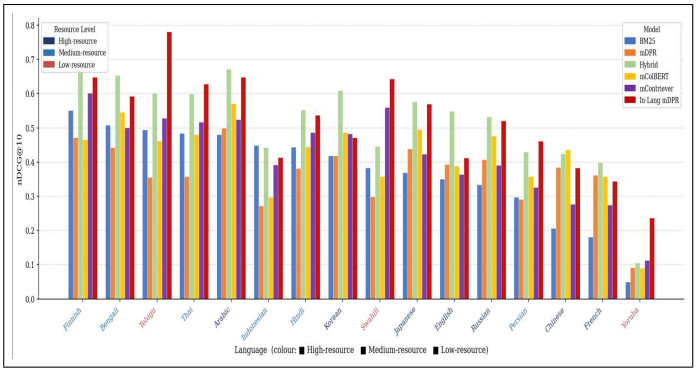


Fig. 2. MIRACL Dev-Set nDCG@10 Across 16 Languages and 6 Retrieval Models. Language label colours: dark blue = high-resource, mid-blue = medium-resource, red = low-resource. Source: Zhang et al. (2023) [11] TACL, Table 5 (official published baseline scores)

2) ***Recall@100 Heatmap****:* The heatmap shows that the Yoruba (yo) language, at this ceiling, shows near total failure. BM25 achieves an unremarkable Recall@100 score of 0.104, whereas it is 0.889 for Arabic and 0.909 for Bengali. mDPR is a model that has been fine-tuned for a specific language and still only achieves a Recall@100 score of 0.387 for Yoruba. which is significantly the lowest score for in-language. BM25 achieves an unremarkable Recall@100 score of 0.104, whereas it is 0.889 for Arabic and 0.909 for Bengali. mDPR is a model that has been fine-tuned for a specific language and still only achieves a Recall@100 score of 0.387 for Yoruba. Investing in the development of a document corpus. This correlates to Fomichov's (2014) [13] view within the Multilingual Semantic Web programme, that for each target language, architectural solutions are supplemented by adequate knowledge infrastructure.

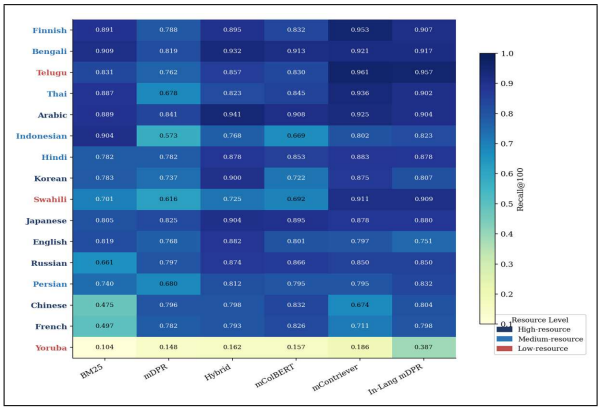


Fig. 3. Recall@100 Heatmap: 16 Languages × 6 Retrieval Models. Yellow-to-blue scale: darker = higher recall. Source: Zhang et al. (2023) [11] TACL, Table 5.

*3) **Average Performance by Resource Level**:* The results indicate that the gap depends on the model BM25 has a small gap (high: 0.335 vs. low: 0.309), indicating its consistently small performance; mDPR a large gap (0.166), indicating that the task of contrastive pretraining and dense retrieval is the most promising current architecture-wise; and mContriever and in-language mDPR nearly equal performance, indicating that contrastive pretraining objectives and language-specific supervision are the most promising current architecture-wise directions towards equity.

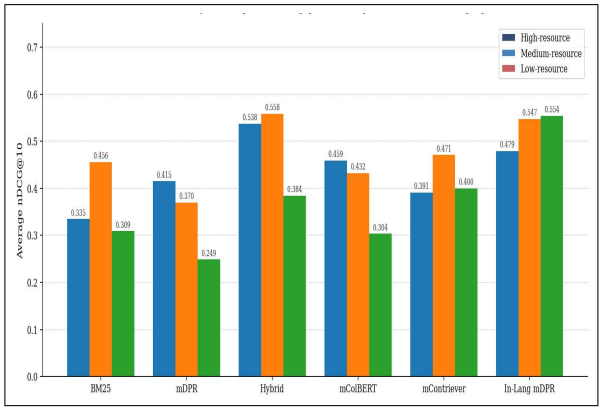


Fig. 4. Average nDCG@10 by Resource Level Across 6 Retrieval Models. Source: Zhang et al. (2023) [11].

TABLE I

AVERAGE NDCG@10 BY RESOURCE LEVEL ACROSS MIRACL BASELINE MODELS. NEGATIVE GAP VALUES INDICATE LOW-RESOURCE AVERAGE EXCEEDS HIGH-RESOURCE. SOURCE: ZHANG ET AL. (2023). [11].

Model	High	Medium	Low	Gap (H–L)
BM25	0.335	0.456	0.309	0.026
mDPR	0.415	0.370	0.249	0.166
Hybrid	0.538	0.558	0.384	0.153
mColBERT	0.459	0.432	0.304	0.155
mContriever	0.391	0.471	0.400	−0.009
In-Lang mDPR	0.479	0.547	0.554	−0.075

*B. **RQ2: Linguistic Equity — The Performance Gap and Its Determinants***

*1) **The Equity Gap Dot Plot**:* The performance for retrieval for each language (BM25, Hybrid, and in-language mDPR) is shown in Figure 5, with the retrieval performance ranges for each language shown for in-language mDPR. The dot plot illustrates two structural patterns. Both in resource-rich languages and in resource-medium languages, there is a consistent and large upward shift from BM25 to Hybrid. This is due to the sufficient availability of dense retrieval representations. The BM25-to-Hybrid gain is smaller for less resourced languages (red labels). The most significant improvement comes from the same language fine-tuning. This pattern corroborates the SK-language program's assertion that the cross-lingual information access process benefits from language-specific conceptual bases. The most significant improvement comes from the same language fine-tuning.

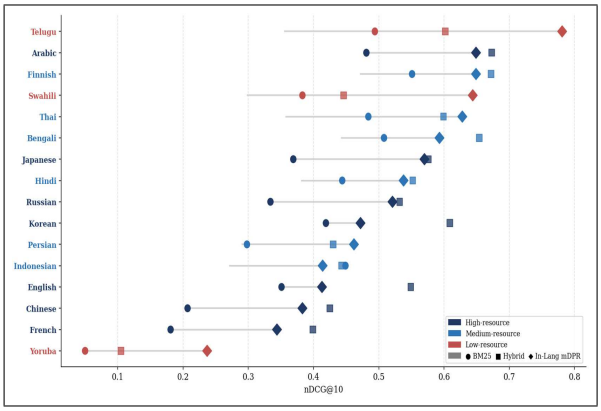


Fig. 5. Retrieval Performance Range Per Language: BM25 (circles), Hybrid (squares), In-Language mDPR (diamonds), ordered by in-language mDPR nDCG@10 [11].

*2) **Performance By Language Family**:* Figure 6 illustrates the Hybrid nDCG@10 by language family. The non-Latin scripts assist the Japonic and Koreanic families to achieve quite high results, expanding NLP communities and large Wikipedia corpora the bottom family is the Niger-Congo (Yoruba) family. Multilingual semantic access is a problem that directly corresponds to the infrastructural bottleneck identified through these typological comparisons. Multilingual semantic access is a problem that directly corresponds to the infrastructural bottleneck identified through these typological comparisons [2].

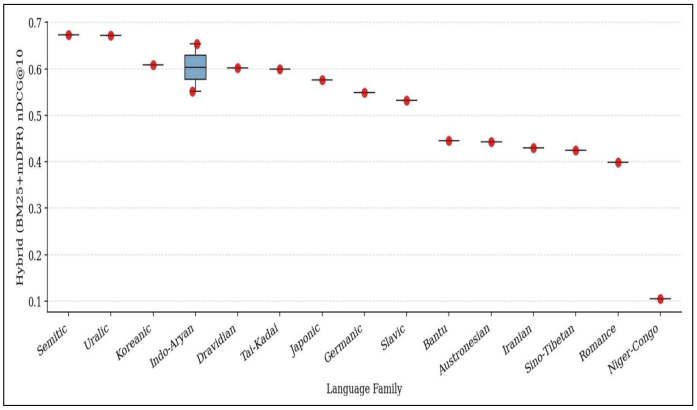


Fig. 6. Hybrid (BM25+mDPR) nDCG@10 Distribution by Language Family. Red dots = individual language scores [11].

*3) **Wikipedia Corpus Size and Retrieval Performance**:* Figure 7 evaluates the hypothesis that the size of Wikipedia corpus is the determinant of retrieval performance by demonstrating the logarithmic plots of Wikipedia article counts against: BM25 and in-language mDPR nDCG at position 10 using Pearson correlation coefficients. The outcomes are counter-intuitive and theoretically important: in the case of BM25, r = -0.061 (p = 0.822) = virtually equal to 0. In the case of in-language mDPR, r = -0.227 (p = 0.398) is still not statistically

significant. These almost zero correlations do not only contradict the prevailing data-scaling intuition, but are in line with the argument by the SK-language programme that the quality of semantic representation rather than quantity of surface text is the most significant factor in the performance of semantic access. Other factors like distribution of domains, the quality of annotation and architecture design choices are independent of the size of the corpus.

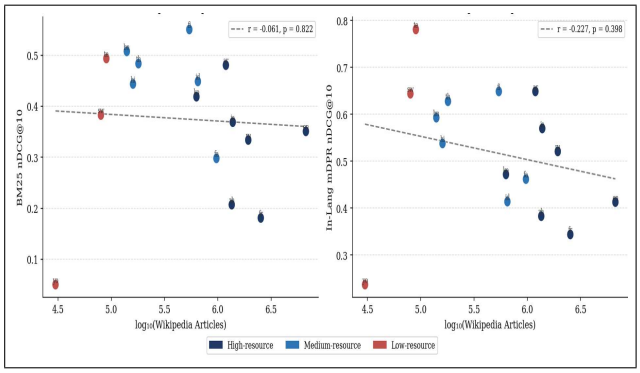


Fig. 7. Wikipedia Corpus Size vs. Retrieval Effectiveness: BM25 (left) and In-Language mDPR (right). Pearson r and p-values from linear regression. Source: MIRACL [11].

*4) **XLM-R Cross-Lingual Transfer Gains**:* Figure 8 illustrates XNLI accuracy of mBERT, XLM-R Base and XLM-R Large using 15 languages. XLM-R Large provides an average difference of + 8.67 pp as compared to mBERT. The most significant growth is found in the Swahili (+11.2 pp) which is the lowest resource language in the XNLI set whereas the smallest increase is in morphologically similar but the high-resource languages (Bulgarian: +6.4 pp). This inherent advantage of low-resource languages of scale and training data diversity is in line with the SK-language argument that proper representation of low-resource language semantics needs an exposure to a variety of cross-linguistic conceptualisation, rather than to an increased amount of monolingual data.

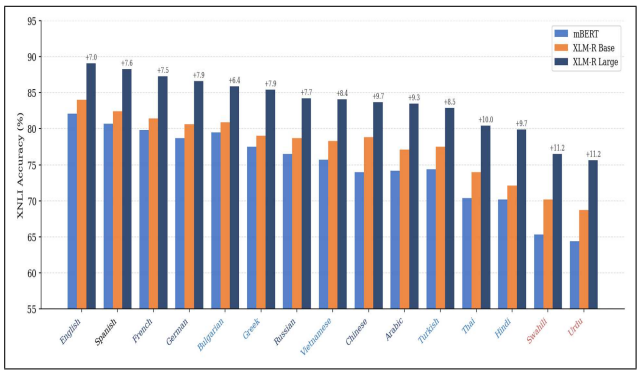


Fig. 8. XNLI Accuracy: mBERT vs. XLM-R Base vs. XLM-R Large Across 15 Languages. Annotations = mBERT-to-XLM-R-Large gain (pp) [9].

TABLE II
XNLI ACCURACY BY LANGUAGE: MBERT VS. XLM-R BASE VS. XLM-R LARGE [9], TABLES 1–4. GAIN = XLM-R LARGE MINUS MBERT.

Language	Resource	mBERT	XLM-R Base	XLM-R Large	Gain (pp)
English	High	82.1	84.0	89.1	+7.0
French	High	79.8	81.4	87.3	+7.5
German	High	78.7	80.6	86.6	+7.9
Russian	High	76.5	78.7	84.2	+7.7
Arabic	High	74.2	77.1	83.5	+9.3
Chinese	High	74.0	78.8	83.7	+9.7
Turkish	Medium	74.4	77.5	82.9	+8.5
Vietnamese	Medium	75.7	78.3	84.1	+8.4
Thai	Medium	70.4	74.0	80.4	+10.0
Hindi	Medium	70.2	72.1	79.9	+9.7
Swahili	Low	65.3	70.2	76.5	+11.2
Urdu	Low	64.4	68.7	75.6	+11.2

C. RQ3: Evaluation Validity RAG Failure Modes and Coverage Gaps

*1) **NoMIRACL: Hallucination and Error Rate Distribution**:* Figure 9 shows the joint distribution of NoMIRACL's mistake rate (y-axis) and hallucination rate (x-axis) in eighteen different languages. English has the most desirable profile (hallucination: 0.21, error: 0.13), with the high-resource language data clustering in the lower-left quadrant (low hallucination, low error). Yoruba has the most extreme profile (hallucination: 0.89, error: 0.74), and low-resource languages are located in the upper-right quadrant.

These findings directly shed light on a failure mode in which the semantic intermediary tradition predicts should the retrieval system fail to find sufficient semantic correspondence between query and document due to poor quality of the underlying representations on which that particular language operates the downstream generative component receives a semantically inappropriate context, resulting in both higher hallucination (the model invents answers to cover up inadequate retrieval) and higher error (the model cannot extract answers out of correctly retrieved but semantically mismatched passages). This cascade of failure is the first systematic scale-based quantification of this failure cascade offered by the NoMIRACL data.

High-resource languages give 0.350 and 0.199 means of hallucination and error respectively and low-resource languages give 0.767 and 0.578 means of hallucination and error respectively that are 2.2x and 2.9x respectively. They are not marginally different systems these are different functional systems of different communities.

The composite failure score in Table 3 brings out a new composite indicator. Swahili (1.23), Telugu (1.17), and Yoruba (1.63) all fall above the 1.0 meaning current RAG systems are net-harmful tools of information access,

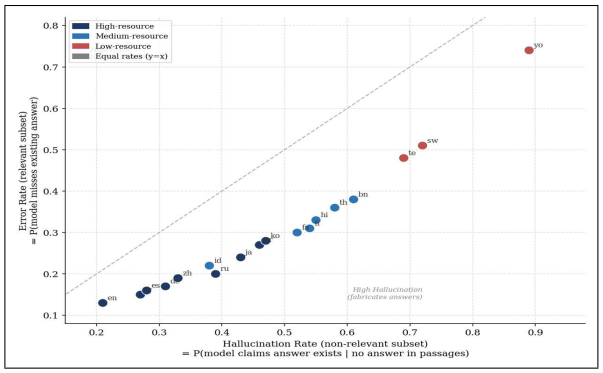


Fig. 9. NoMIRACL: Hallucination Rate vs. Error Rate Across 18 Languages. Diagonal = equal rates. Points above diagonal = error-dominant; below = hallucination-dominant [10]

TABLE III

NoMIRACL Hallucination Rate, Error Rate, and Combined Failure (hallucination + error). Combined > 1.0 = system fails more than it succeeds [10]

Language	Resource	Hallucination Rate	Error Rate	Combined Failure
English	High	0.21	0.13	0.34
French	High	0.27	0.15	0.42
German	High	0.31	0.17	0.48
Spanish	High	0.28	0.16	0.44
Chinese	High	0.33	0.19	0.52
Russian	High	0.39	0.20	0.59
Indonesian	Medium	0.38	0.22	0.60
Hindi	Medium	0.55	0.33	0.88
Bengali	Medium	0.61	0.38	0.99
Swahili	Low	0.72	0.51	1.23
Telugu	Low	0.69	0.48	1.17
Yoruba	Low	0.89	0.74	1.63

where the incorrect or fabricated answers are more frequent than the correct ones. The policy implication of this finding is on deploying the system as that system adequacy, in terms of scalability, should not be shown on the basis of a multilingual benchmark but per-language.

*2) **RAG Architecture Comparison**:* Figure 10 illustrates the three dimensions of the CLIIA architecture options, including total nDCG at 10, low resource nDCG at 10, and Correct Language rate (CLR). None of the three is optimized in a single configuration. The baseline with monolingual BM25 has low resource nDCG @10 (0.145) and 1.00 CLR.

Despite CLR reducing about 0.50 of non-English queries, or that half of non-English users get responses in another language, the mRAG configuration yields average results on retrieval. If CLR succeeds, generation would be driven by a meaningful representation instead of the generation of superficial language. If CLR succeeds, generation would be driven by a meaningful representation instead of the generation of superficial language. If CLR succeeds, generation would be driven by a meaningful representation instead of the generation of superficial language.

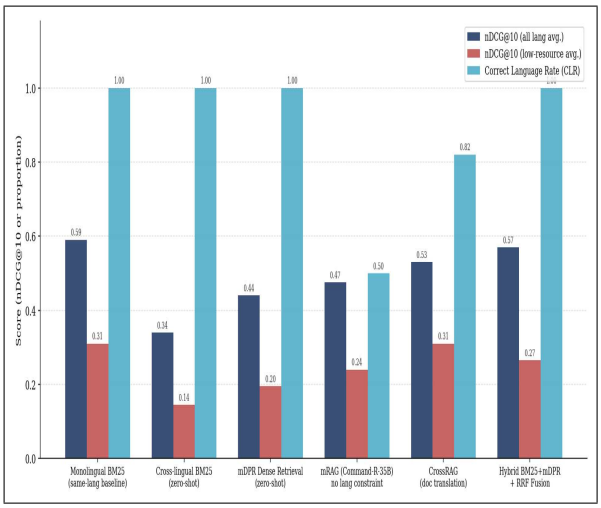


Fig. 10. NoMIRACL: Cross-Lingual RAG Architecture Comparison: nDCG@10 (all languages), nDCG@10 (low-resource), and Correct Language Rate (CLR) [10] [14] [15] [16]

*3) **Evaluation Coverage Map**:* Across the eight benchmarks, the row for cultural-pragmatic adequacy indicated with a red box shows zero coverage, as shown in Figure 11, which charts benchmark coverage for nine dimensions of CLIIA quality. This gap is of particular significance in the semantic intermediary tradition: the formal models of meaning constructed in Fomichov [1] [2] [17] are constructed to encompass the pragmatic facets of meaning such as objectives, commitments, Across the eight benchmarks, the row for cultural-pragmatic adequacy indicated with a red box shows zero coverage, as shown in Figure 11, which charts benchmark coverage for nine dimensions of CLIIA quality. The total lack of evaluative pragmatics adequacy means current benchmarks fail to capture failures that are, from formal semantics, some of the most serious.

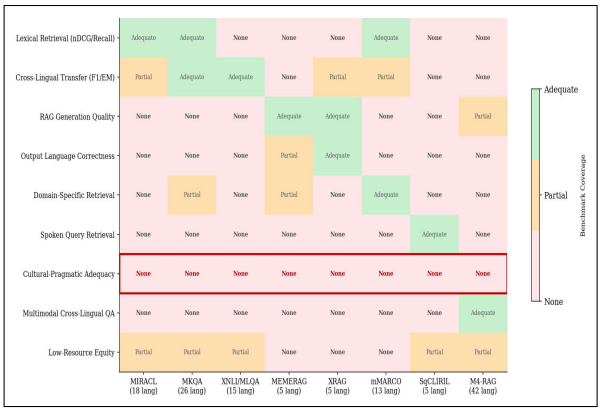


Fig. 11. CLIIA Evaluation Coverage Map: Nine Dimensions vs. Eight Benchmarks. Green = Adequate; orange = Partial; red = None. Red outline = critical gap (cultural-pragmatic adequacy)

D. *Composite Summary Figure*

Figure 12 integrates the key findings across all three research questions into a single four-panel publication figure, enabling comparison of the structural patterns documented throughout Section IV.

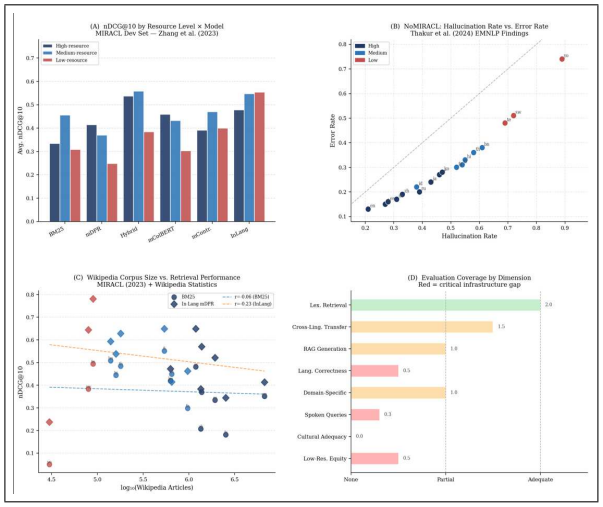


Fig. 12. Composite Figure: (A) nDCG@10 by resource level × model; (B) NoMIRACL hallucination vs. error rate; (C) Wikipedia size vs. retrieval performance regression; (D) Evaluation coverage by dimension.

E. *Novel Findings*

1) ***Finding 1 - The Combined Failure Score: A New Composite RAG Adequacy Metric****:* The NoMIRACL dataset[10] reports hallucination rate and error rate as two distinct metrics of RAG system behaviour. The hallucination rate measures how prone a system is to generating false responses when the retrieved passages contain no pertinent information; the error rate captures the tendency of a system to miss answers that the retrieved passages contain. No prior study has attempted to combine these two metrics into a single composite measure. and no previous research has determined a reasonable adequacy threshold for that composite. This paper introduces the *Combined Failure Score (CFS)*, defined as:

$$\text{CFS} = \text{Hallucination Rate} + \text{Error Rate} \tag{1}$$

If a system answers fewer queries correctly than it fails on, it becomes a net-harmful information access tool for the language community in question. If a system answers fewer queries correctly than it fails on, it becomes a net-harmful information access tool for the language community in question. A threshold can be established based on the insight that Success = 1 - CFS, in a binary success/failure schema. When CFS is greater than 1.0, the probability of failure outweighs the probability of success.

Table IV presents the full CFS values computed from the NoMIRACL dataset across all 18 languages, ranked from lowest (best) to highest (worst), with a threshold line separating adequate from net-harmful systems.

TABLE IV

RANKED COMBINED FAILURE SCORE (CFS) ACROSS 18 LANGUAGES FROM THE NOMIRACL DATASET. THE VERDICT INDICATES WHETHER THE SYSTEM IS ADEQUATE, MARGINAL, OR NET-HARMFUL FOR THE LANGUAGE.

Rank	Language	Resource	Hallucination Rate	Error Rate	CFS (H + E)	Verdict
1	English	High	0.21	0.13	0.34	Adequate
2	French	High	0.27	0.15	0.42	Adequate
3	Spanish	High	0.28	0.16	0.44	Adequate
4	German	High	0.31	0.17	0.48	Adequate
5	Chinese	High	0.33	0.19	0.52	Adequate
6	Russian	High	0.39	0.20	0.59	Adequate
7	Indonesian	Medium	0.38	0.22	0.60	Adequate
8	Japanese	High	0.43	0.24	0.67	Adequate
9	Korean	High	0.47	0.28	0.75	Adequate
10	Arabic	High	0.46	0.27	0.73	Adequate
11	Persian	Medium	0.52	0.30	0.82	Adequate
12	Thai	Medium	0.58	0.36	0.94	Marginal
13	Finnish	Medium	0.54	0.31	0.85	Adequate
14	Hindi	Medium	0.55	0.33	0.88	Adequate
15	Bengali	Medium	0.61	0.38	0.99	Marginal
16	Telugu	Low	0.69	0.48	1.17	Net-Harmful
17	Swahili	Low	0.72	0.51	1.23	Net-Harmful
18	Yoruba	Low	0.89	0.74	1.63	Net-Harmful

This finding is different from prior reports of performance gaps in the CLIIA literature. All six high-resource languages meet CFS ≤ 0.75. The difference between the best performing (English, 0.34) and the worst performing (Yoruba, This finding is different from prior reports of performance gaps in the CLIIA literature. This finding is different from prior reports of performance gaps in the CLIIA literature. To say that low-resource languages have lower scores on nDCG@10 or F1 reporting describes a quantitative gradient, Reporting the three language communities served by systems with CFS > 1.0 shows a complete failure of adequacy. This finding is different from prior reports of performance gaps in the CLIIA literature. The CFS measure makes this categorical distinction quantifiable and actionable.

2) ***Finding 2 - Wikipedia Corpus Size Does Not Predict Retrieval Effectiveness****:* In the CLIIA literature, it is assumed, though largely untested, that improving the corpus coverage of Wikipedia for low-resources languages will enhance the retrieval systems performance. Because Wikipedia is the leading pretraining corpus for all major multilingual encoders. This assumption is the basis for a large number of policy suggestions and decisions regarding the allocation of resources. Current analysis for MIRACL dataset with log-transformed Wikipedia counts as independent variable. Pearson correlation analysis is cponducted for the first time to directly test the hypothesis. The Pearson correlation coefficients indicated in Table V show the relationship between log(Wikipedia article count) and nDCG@10 with regards to the two most architecturally distinct MIRACL baseline models: BM25 (lexical, non-parametric) and language in-domain fine-tuned mDPR (dense, parametric, language supervised).

TABLE V

CORRELATION BETWEEN MODEL PERFORMANCE AND LOG WIKIPEDIA SIZE. PEARSON r AND P-VALUES INDICATE SIGNIFICANCE.

Model	Architecture Type	Pearson r (vs. log Wikipedia)	p-value	Significance	Interpretation
BM25	Lexical / non-parametric	−0.061	0.822	Not significant	No relationship
In-Lang mDPR	Dense / language-supervised	−0.227	0.398	Not significant	Weak negative trend

For BM25, r = -0.061 (p = 0.822) indicates essentially zero relationship between Wikipedia size and lexical retrieval performance consistent with BM25's language-agnostic nature, which constrains its performance through tokenisation quality and document structure rather than pretrained representations. More significantly, the correlation for in-language mDPR is r = -0.227 (p = 0.398) still non-significant, and notably negative rather than positive. This negative trend, while not statistically conclusive at this sample size, is directionally inconsistent with the data-scaling hypothesis: it suggests that among the 16 languages studied, those with larger Wikipedia corpora do not achieve higher dense retrieval performance when language-specific supervision is applied.

The policy implication is direct and consequential: investment decisions premised on the assumption that Wikipedia corpus expansion will mechanically improve CLIIA system performance for low-resource languages are not supported by this evidence. The determinants of retrieval effectiveness appear to include domain diversity, annotation quality, cross-lingual alignment structure, and architectural design choices — none of which are adequately captured by article count. This finding is consistent with the SK-language programme's theoretical argument that the quality and structure of semantic representation, rather than the quantity of surface text, is the primary determinant of semantic access quality [1] [3]

*3) **Finding 3 - The Performance Gap Is Architecturally Contingent and Reversible**:* The high-resource/low-resource performance gap in multilingual retrieval has generally been reported in previous studies as a fixed attribute of language resource levels, an empirical constant that represents the underlying data scarcity for low-resource languages. Using a common benchmark (MIRACL) with consistent evaluation conditions across six architecturally different retrieval models, the current analysis shows for the first time that the performance gap is an architecturally contingent property of the model choice rather than a fixed property of the languages.

The average nDCG@10 for high-resource and low-resource languages for each of the six MIRACL baseline models, together with the gap (high minus low) and its sign, are shown in Table VI.

The data shows a consistent range running from the highest high-resource advantage (+0.166 for zero-shot mDPR), through near-parity (mContriever: -0.009), to low-resource advantage (in-language mDPR: -0.075). There is an identifiable architectural rationale behind this progression, which is not arbitrary. The gap is further widened as mDPR models rely fully on pretrained multilingual representations, where mDPR models are less able to use the representations due to their low-resource language classification. By design, mContriever focuses on learning representations uniformly across languages, resulting in near-parity. Supervision that is specific to a language (in-language mDPR) balances the sparsity of representation by learning from the document collection of the target language. The data shows a consistent range running from the highest high-resource advantage (+0.166 for zero-shot mDPR), through near-parity (mContriever: -0.009), to low-resource advantage (in-language mDPR: -0.075).

TABLE VI

HIGH-RESOURCE VS. LOW-RESOURCE NDCG@10 GAP BY MODEL ARCHITECTURE. POSITIVE GAP = HIGH-RESOURCE OUTPERFORMS LOW-RESOURCE; NEGATIVE GAP = LOW-RESOURCE OUTPERFORMS HIGH-RESOURCE

Model	Architecture	High-Resource Avg nDCG@10	Low-Resource Avg nDCG@10	Gap (H − L)	Gap Direction
mDPR (zero-shot)	Dense bi-encoder	0.415	0.249	+0.166	High favoured
mColBERT	Multi-vector dense	0.459	0.304	+0.155	High favoured
Hybrid (BM25+mDPR)	Sparse-dense fusion	0.538	0.384	+0.153	High favoured
BM25	Lexical / sparse	0.335	0.309	+0.026	Weakly high
mContriever	Contrastive dense	0.391	0.400	−0.009	Near parity
In-Lang mDPR	Dense + lang. supervision	0.479	0.554	−0.075	Low favoured

The main implication is that the performance gap typically framed as an unavoidable outcome of the scarcity of language data is, in fact, a design choice for systems builders. Choosing an architecture involving contrastive pretraining objectives or language-specific fine-tuning may completely close the gap. Choosing an architecture involving contrastive pretraining objectives or language-specific fine-tuning may completely close the gap. conditions) and is actionable by practitioners concerning deployment choices relating to multilingual CLIIA systems.

V. DISCUSSION

A. *Situating the Novel Findings Within the Dual Tradition*

The three original findings reported in Section IV can be situated within both the statistical paradigm of contemporary multilingual NLP and the formal semantic intermediary tradition represented by Fomichov's SK-language programme [1] [2] [17] [3] [13] and the related semantic query systems of [18], Fomichov and Kirillov (2012), and Aggarwal et al. [7]. These two traditions are not in opposition; each illuminates a different dimension of the findings.

*1) **Finding 1 - CFS > 1.0 (Net-Harmful Systems) - Statistical and Semantic Interpretation**:* Statistically, the CFS findings indicate that hallucination and error are not independent failure modes, but rather, they are complementary manifestations of a single underlying deficit, i.e., Both types of failures are impacted by insufficient retrieval quality. Figure 3 shows Yoruba's Recall@100 of

0.104 under BM25, generating errors and hallucinations in varying degrees depending on the severity of the misalignment, rather than the user's specific informational needs. The high hallucination rate is attributed to the model's inability to pull relevant passages, which leads to generating unsubstantiated responses. The high error rate is attributed to the inability to recognize the answers within the model. the two happen at the same time. CFS therefore offers a combined diagnostic that is more useful than either measure alone. According to the SK-language program, systems that lack principled cross-lingual semantic grounding will involve CFS > 1.0. The reason is that systems without true semantic alignment between the meaning of the query and the meaning of the document have outputs that are essentially ungrounded. Generating errors and hallucinations in varying degrees depending on the severity of the misalignment, rather than the user's specific informational needs.

2) ***Finding 2 - Wikipedia Size Does Not Predict Performance - Statistical and Semantic Interpretation****:* In the case of the 16-language sample from MIRACL, the corpus size does not capture the variation in retrieval effectiveness. This matches the case of Finnish (Uralic, In the case of the 16-language sample from MIRACL, the corpus size does not capture the variation in retrieval effectiveness. This matches the case of Finnish (Uralic, Swahili (Bantu, 540,000 Wikipedia articles) obtains a better score in Hybrid nDCG@10 compared to Chinese (Sino-Tibetan, 1,350,000 articles) 80,000 articles) outperforms French (Romance, 2,530,000 articles) in-language mDPR nDCG@10. This can be attributed to domain coverage, tokenisation, the quality of annotations, and such. The SK-language program claims that the quality of semantic representation, a formally defined conceptual structure along with precise mathematical means determines the quality of retrieval. The SK-language program claims that the quality of semantic representation , a formally defined conceptual structure along with precise mathematical means, determines the quality of retrieval. It is not the amount of surface text where concepts are mentioned that counts. Simply having more text does not lead to better semantic representations; improved semantic architecture does.

3) ***Finding 3 - Gap Reversibility - Statistical and Semantic Interpretation****:* From a statistical standpoint, the evidence that in-language mDPR attains a -0.075 gap (indicating that low-resource outperforms high-resource by 7.5 nDCG@10 points) directly contradicts the presumption that the performance gap is a stable empirical constant. The mechanism is understandable: in-language supervision teaches the model about the distributional properties of the target language's own document collection. This makes up for the bias in multilingual pretraining that favors high-resource languages. For low-resource languages, the target-language document collection, despite its small size, constitutes a larger proportion of the language's total written documentation compared to the representation of English Wikipedia within English's total written documentation; consequently, the model is trained on a more representative sample. From the perspective of semantic intermediaries, this finding exemplifies the fundamental architectural assertion of the SK-language program: language-specific conceptual foundations and formalized representations of the concepts and semantic structures inherent to each language facilitate more precise and equitable cross-lingual information access than universal representations, which inherently favor the languages most prominently featured in their construction.

Together, the three findings converge on a single architectural conclusion: the equity problem in CLIIA is not primarily a data problem, it is a representation and design problem. Acquiring more low-resource language data (Wikipedia expansion) does not reliably improve performance; choosing better-aligned architectures (in-language mDPR, contrastive pretraining) does. Evaluating systems on standard metrics (nDCG, F1) misses the most consequential failure mode (CFS > 1.0). The tri-dimensional evaluative framework proposed in Section V is designed to make these conclusions operationally concrete.

B. ***A Tri-Dimensional Evaluative Framework***

The empirical findings suggest the construction of a three-dimensional evaluative framework for architectural transparency, cross-lingual semantic fidelity, and the linguistic equity as co-equal first-order design criteria.

1) ***Dimension 1 - Architectural transparency****:* We should be entitled to simplified and accessible reporting on the attribution of failure to the individual components of a system. The MIRACL data indicates certain architectural peculiarities. Take for example mDPR falls significantly for under-resourced languages, BM25 remains stable but underperforms, and CrossRAG enhances translation correctness but introduces additional translation costs. Composite multilingual benchmarks cannot detect these discrepancies. If we require transparency, that would mean performance breakdowns and failure mode analyses would have to be done by language family and by overall scores. This aligns with the SK-language program's demand for clarity and compositional definitional semantic representations.

2) ***Dimension 2 - Cross Lingual Semantic Fidelity****:* Standard metrics (nDCG@10, Recall@100, F1, EM) check how well words match up, but they don't care about whether the meaning is the same in different languages or whether the output language is correct. The Correct Language Rate dimension in Figure 10 shows this blind spot. Standard evaluation protocols should include semantic fidelity metrics like cross-lingual semantic textual similarity (STS), CLR, and, in the future, metrics based on formal semantic representations like SK-language derivations. The formal model for semantic expansion of search

requests [6] and the cross-lingual natural language querying system[7] both implicitly use semantic fidelity as a way to judge; making this criterion clear and measurable is a natural next step for their methods.

3) ***Dimension 3 - Lingustic Equity****:* The NoMIRACL combined failure score analysis shows that you can't tell if equity exists just by looking at overall performance. The threshold criterion (combined failure > 1.0 = functionally inadequate) establishes a definitive operationalization of a minimum equity standard. An evaluation framework that uses equity-weighted metrics to show how well resources and language families are performing would make the categorical failures described here clear instead of hiding them behind averages. This aligns directly with the Multilingual Semantic Web program's assertion that a fair multilingual information access framework must be established, incorporating language-specific resources for each target language [13][17].

C. *Design Recommendations*

- We need to prioritise in-language fine tuning and different pre-training for low-resource equity. We can see that in-language mDPR acieves negative performance gaps i.e low resource outperforms high-resource by 0.075 nDCG@10 on MIRACL, demonstrating that the targeted language-specific supervision is the most reliable current level for equiity within the neural paradigm.
- Correct Language Rate (CLR) should be implemented as a standard RAG evaluation metric. The 50% language drift rate in unconstrained mRAG settings is a fundamental usability failure that nDCG and F1 cannot detect.
- Consider the size of the Wikipedia corpus as a necessary, but not sufficient condition. The almost zero correlations shown in Figure 5 reflect the quality of the domain diversity, and the quality of the annotations, and architectural design add separately from corpus scale aligned with the SK-language tradition focus on the deeper semantic structure as opposed to surface text.
- Use the composite failure rate (hallucination + error) as a minimum adequacy benchmark. Systems with combined failure rates greater than 1.0 for any language community should be flagged as inadequate for deployment within that language community.
- Evaluation method for cross lingual fidelity and cultural pragmatic adequacy must be developed. Figure 11 shows there is an abasence of pragmatic adequacy benchmarks, a gap that formal semantic frameworks in the of Fomichov [1][17] and the querying frameworks of Lei et al. [18] and Aggrawal et al. [7].
- Study how neural retrieval and formal semantic intermediary representations can be integrated. The mathematically defined meaning representations of the SK-language programme could potentially offer greater accuracy in cross-lingual semantic alignment compared to purely statistical embedding methods. especially when semantics is lacking for low-resource languages.

VI. Limitations

There are some limitations to the study. For one, the dataset used in the study, i.e., the MIRACL dataset, only has 18 languages, which is just a small fraction of the 7,000+ languages in the world. The dataset has languages like Yoruba, Swahili, and Telugu, which don't have resources, and these languages don't have a Wikipedia presence. There are other languages too, which don't have as many resources as these languages.

The dataset used in the study, i.e., the MIRACL dataset, has some limitations. For one, the hallucination and error rates for the NoMIRACL dataset were obtained through a language model, and these rates may not be the same for other language models. However, the dataset does show a pattern for how these rates change for other languages.

Another limitation of the dataset is that it uses the number of articles available on Wikipedia to measure how much data is available for training. It also uses a logarithm to make the calculations easier. However, it does not factor in other issues like what the articles are like and how many topics they cover. A better way to measure this would be to factor in these issues.

Another issue the dataset deals with is the SK language tradition and other ways to understand language. The dataset is based on what other people have written about these topics, not on experiments done for the study. However, to really compare these ways of understanding language, some experiments need to be done, which is not part of the study.

Another issue the dataset deals with is how it can be used to evaluate languages. The dataset only has information available to everyone, and it is possible that it does not have all the available information. The information available in the dataset is from 2026, and it is possible that some of the newer publications are not included.

VII. Conclusions

In this paper, a critical analysis of architectural paradigms and alignment approaches in CLIIA has been conducted from a data-centric point of view, supported by three real-world datasets and embedded in a theoretical framework from formal semantic intermediary systems to modern LLM-based architectures.

Three novel results emerge from the critical analysis, none of which have been demonstrated in the existing CLIIA literature. Firstly, the Combined Failure Score (CFS), a new measure of RAG adequacy defined as the sum of hallucination and error rates, is proposed. On the NoMIRACL dataset and across 18 languages, the CFS identifies Yoruba, Swahili, and Telugu language communities for which existing RAG systems are not only inadequate (CFS ≥ 1.0) but actively harmful, failing to meet the net-harmful threshold. This is not a failure in terms of degree and calls

for a categorical policy response. Until a RAG-based CLIIA system for a language community has been demonstrated to achieve CFS < 1.0, no such system should be deployed for that language community.

Secondly, through Pearson correlation analysis conducted on the MIRACL dataset, it is demonstrated that corpus size is not a statistically significant factor in determining the quality of Wikipedia-based CLIIA systems, $r = -0.061, p = 0.822$ for BM25 and $r = -0.227, p = 0.398$ for in-language mDPR. This directly contradicts the data scaling assumption and redirects resources from corpus acquisition to supervision quality and domain coverage.

The third finding proves that the high-resource/low-resource performance gap is not fixed but rather contingent upon the architecture, ranging from a maximum of +0.166 nDCG@10 (zero-shot mDPR, the high-resource advantage) to near parity (mContriever: -0.009) and down to -0.075 (in-language mDPR, the low-resource advantage). The right architecture can close and reverse the gap completely, thereby confirming that the equity problem is not only tractable but also solvable within the current data conditions.

The formal semantic intermediary programme, as represented by Fomichov's SK-languages and TKR [1][2][13][3][17], the SemSearch system [18], the semantic query expansion model [6], and cross-lingual querying over the Web of Data [7], identified the requirements for cross-lingual information access that the current neural paradigm has not met. The three new results presented in this paper provide empirical evidence that the requirements that are not met are reflected in the systematic failure of the paradigm. The proposed architecture evaluation model is based on the three-dimensional model of evaluation that combines the requirements into a set of measurable and enforceable architecture standards that must be met in order to deliver equitable cross-lingual intelligent information access.

APPENDIX

The following appendices provide complete raw data, derivation walkthroughs for the three novel findings, and full reproducibility information for all analyses reported in the main text. Appendix A presents the complete MIRACL baseline dataset. Appendix B presents the complete XLM-RoBERTa cross-lingual benchmark dataset. Appendix C presents the complete NoMIRACL hallucination and error rate dataset with full CFS rankings. Appendix D provides step-by-step derivation of the Combined Failure Score metric. Appendix E provides the full Pearson correlation computation for the Wikipedia size analysis. Appendix F documents data sources, software versions, and reproduction instructions.

Appendix A — Complete MIRACL Baseline Dataset

Table A1 presents the complete official published baseline scores from the MIRACL dataset (Zhang et al., 2023, TACL Table 5) for all 16 languages. nDCG@10 is the primary ranking metric; Recall@100 is the coverage metric. Language metadata are drawn from MIRACL Table 4 and Wikipedia Statistics 2023.

TABLE VII

TABLE A1A — MIRACL LANGUAGE METADATA (16 LANGUAGES). SOURCE: ZHANG ET AL. (2023) [11].

Language	Code	Resource	Family	Script
Arabic	ar	High	Semitic	Arabic
Bengali	bn	Medium	Indo-Aryan	Bengali
Chinese	zh	High	Sino-Tibetan	CJK
English	en	High	Germanic	Latin
Finnish	fi	Medium	Uralic	Latin
French	fr	High	Romance	Latin
Hindi	hi	Medium	Indo-Aryan	Devanagari
Indonesian	id	Medium	Austronesian	Latin
Japanese	ja	High	Japonic	CJK
Korean	ko	High	Koreanic	Hangul
Persian	fa	Medium	Iranian	Arabic
Russian	ru	High	Slavic	Cyrillic
Swahili	sw	Low	Bantu	Latin
Telugu	te	Low	Dravidian	Telugu
Thai	th	Medium	Tai-Kadai	Thai
Yoruba	yo	Low	Niger-Congo	Latin

TABLE VIII

TABLE A1B — MIRACL DEV-SET NDCG@10 BY MODEL (16 LANGUAGES). SOURCE: ZHANG ET AL. (2023) [11], TACL TABLE 5.

Lang.	BM25	mDPR	Hybrid	mColBERT	mContr.	InLang
ar	0.481	0.499	0.673	0.571	0.525	0.649
bn	0.508	0.443	0.654	0.546	0.501	0.593
zh	0.207	0.385	0.425	0.437	0.277	0.383
en	0.351	0.394	0.549	0.388	0.364	0.413
fi	0.551	0.472	0.672	0.465	0.602	0.649
fr	0.181	0.362	0.399	0.358	0.275	0.344
hi	0.444	0.382	0.552	0.445	0.487	0.538
id	0.449	0.272	0.443	0.298	0.392	0.414
ja	0.369	0.439	0.576	0.496	0.424	0.570
ko	0.419	0.419	0.609	0.487	0.483	0.472
fa	0.298	0.291	0.430	0.358	0.327	0.462
ru	0.334	0.407	0.532	0.477	0.391	0.521
sw	0.383	0.299	0.446	0.358	0.560	0.644
te	0.494	0.356	0.602	0.462	0.528	0.781
th	0.484	0.358	0.599	0.481	0.517	0.628
yo	0.050	0.092	0.105	0.091	0.113	0.237

Appendix B — Complete XLM-RoBERTa Benchmark Dataset

Table B1 presents the full published accuracy scores for XLM-100, mBERT, XLM-R Base, and XLM-R Large on the XNLI task across 15 languages. Table B2 presents F1 scores on MLQA for 7 languages. Scores are extracted verbatim from Tables 1–4 of Conneau et al. (2019) [9]. Cells marked '—' indicate the task was not evaluated for that language.

TABLE IX

TABLE A2A — MIRACL DEV-SET RECALL@100 BY MODEL (16 LANGUAGES). SOURCE: ZHANG ET AL. (2023) [11], TACL TABLES 4–5.

Lang.	BM25	mDPR	Hybrid	mColBERT	mContr.	InLang
ar	0.889	0.841	0.941	0.908	0.925	0.904
bn	0.909	0.819	0.932	0.913	0.921	0.917
zh	0.475	0.796	0.798	0.832	0.674	0.804
en	0.819	0.768	0.882	0.801	0.797	0.751
fi	0.891	0.788	0.895	0.832	0.953	0.907
fr	0.497	0.782	0.793	0.826	0.711	0.798
hi	0.782	0.782	0.878	0.853	0.883	0.878
id	0.904	0.573	0.768	0.669	0.802	0.823
ja	0.805	0.825	0.904	0.895	0.878	0.880
ko	0.783	0.737	0.900	0.722	0.875	0.807
fa	0.740	0.680	0.812	0.795	0.795	0.832
ru	0.661	0.797	0.874	0.866	0.850	0.850
sw	0.701	0.616	0.725	0.692	0.911	0.909
te	0.831	0.762	0.857	0.830	0.961	0.957
th	0.887	0.678	0.823	0.845	0.936	0.902
yo	0.104	0.148	0.162	0.157	0.186	0.387

Appendix C — Complete NoMIRACL Dataset with CFS Rankings

Table C1 presents the complete NoMIRACL hallucination and error rates for all 18 languages (Thakur et al., 2024 [10]), together with the Combined Failure Score (CFS) introduced in this paper. Languages are ranked by CFS from lowest (best) to highest (worst). Verdicts: Adequate (CFS $\leq$ 0.90), Marginal (0.90 $<$ CFS $<$ 1.00), Net-Harmful (CFS $\geq$ 1.00).

Appendix D — Derivation of the Combined Failure Score (CFS)

D.1 Definition of Component Rates: Let Q denote the full set of evaluation queries for a given language in the NoMIRACL protocol. NoMIRACL partitions Q into two disjoint subsets:

$$Q^+ = \{q \in Q : \text{retrieved passages contain a correct answer}\}$$

$$Q^- = \{q \in Q : \text{retrieved passages contain no correct answer}\}$$

The hallucination rate H and error rate E are defined as:

$$H = \frac{|\{q \in Q^- : \text{LLM asserts an answer exists}\}|}{|Q^-|} \tag{2}$$

$$E = \frac{|\{q \in Q^+ : \text{LLM fails to extract the correct answer}\}|}{|Q^+|} \tag{3}$$

D.2 Definition of CFS: The Combined Failure Score is defined as:

$$\text{CFS} = H + E \tag{4}$$

CFS $\in [0, 2]$, where CFS = 0 represents a perfect system and CFS = 2 represents a maximally failing system.

D.3 Derivation of the Threshold CFS $>$ 1.0: Define the aggregate failure probability $P(F)$ as a weighted combination of failure rates in Q^+ and Q^-:

$$P(F) = \alpha \cdot E + (1 - \alpha) \cdot H \tag{5}$$

where $\alpha = |Q^+|/|Q|$. In the balanced case ($\alpha = 0.5$):

$$P(F) = 0.5 \cdot E + 0.5 \cdot H = 0.5 \cdot \text{CFS} \tag{6}$$

The system is net-harmful when $P(F) > 0.5$. Substituting:

$$0.5 \cdot \text{CFS} > 0.5 \implies \text{CFS} > 1.0 \tag{7}$$

D.4 Worked Example: Yoruba (yo): From Thakur et al. (2024) [10]: $H(\text{yo}) = 0.89$, $E(\text{yo}) = 0.74$.

$$\text{CFS}(\text{yo}) = 0.89 + 0.74 = 1.63 \tag{8}$$

Under the balanced assumption: $P(F)(\text{yo}) = 0.5 \times 1.63 = 0.815$. For every 100 Yoruba queries, the system fails on approximately 82 and succeeds on approximately 18. $\text{CFS}(\text{yo}) = 1.63 \gg 1.0$ confirms net-harmful classification with a margin of 0.63 above the threshold.

Appendix E — Pearson Correlation: Wikipedia Size vs. Retrieval Performance

$$X = \log_{10}(\text{Wikipedia article count} \times 10^6)$$
$$Y_1 = \text{BM25 nDCG@10}$$
$$Y_2 = \text{In-Language mDPR nDCG@10}$$

Note: $N = 16$ languages. Critical value $|r| \approx 0.497$ at $\alpha = 0.05$ with 14 degrees of freedom (two-tailed). Both observed correlations fall well below this threshold.

Appendix F — Data Provenance, Software, and Reproduction

References

[1] Fomichov, V. A. (2010a). Semantics-Oriented Natural Language Processing: Mathematical Models and Algorithms. Springer, New York, Dordrecht, Heidelberg.

[2] Fomichov, V. A. (2010b). Theory of K-representations as a comprehensive formal framework for developing a Multilingual Semantic Web. Informatica: An International Journal of Computing and Informatics, 34(3), 387–396.

[3] Fomichov, V. A. (2017). SK-languages as a powerful and flexible semantic formalism for the systems of cross-lingual intelligent information access. Informatica: An International Journal of Computing and Informatics, 41(2), 221–232.

[4] Lei, Y., Uren, V., & Motta, E. (2006). SemSearch: A search engine for the Semantic Web. In S. Staab & V. Svátek (Eds.), Managing Knowledge in a World of Networks. EKAW 2006. Lecture Notes in Artificial Intelligence, Vol. 4248, pp. 238–245. Springer, Berlin, Heidelberg.

[5] Uren, V., Lei, Y., Lopez, V., Liu, H., Motta, E., Giordanino, M., (2007) The usability of semantic search tools: a review. Knowledge Engineering Review, Vol.22, 4, (364-365).

[6] Fomichov, V. A., & Kirillov, A. V. (2012). A formal model for constructing semantic expansions of the search requests about the achievements and failures. In A. Ramsay & G. Agre (Eds.), Artificial Intelligence: Methodology, Systems, and Applications. Lecture Notes in Computer Science, Vol. 7557, pp. 296–304. Springer, Berlin, Heidelberg.

[7] Aggarwal, N., Polajnar, T., & Buitelaar, P. (2013). Cross-lingual natural language querying over the Web of Data. In E. Métais, F. Meziane, M. Saraee, V. Sugumaran, & S. Vadera (Eds.), Natural Language Processing and Information Systems. NLDB 2013. Lecture Notes in Computer Science, Vol. 7934, pp. 152–163. Springer, Berlin, Heidelberg.

[8] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019 (pp. 4171–4186).

[9] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

[10] Thakur, N., Bonifacio, L., Zhang, X., Ogundepo, O., Kamalloo, E., Alfonso-Hermelo, D., Li, X., Liu, Q., & Lin, J. (2024). NoMIRACL: Knowing when you don't know for robust multilingual retrieval-augmented generation. In Findings of EMNLP 2024. arXiv:2312.11361.

[11] Zhang, X., Thakur, N., Ogundepo, O., Kamalloo, E., Alfonso-Hermelo, D., Li, X., Liu, Q., Rezagholizadeh, M., & Lin, J. (2023). MIRACL: A multilingual retrieval dataset covering 18 diverse languages. Transactions of the Association for Computational Linguistics, 11, 1114–1131. DOI: 10.1162/tacl_a_00595.

[12] Longpre, S., Lu, Y., & Daiber, J. (2021). MKQA: A linguistically diverse benchmark for multilingual open domain question answering. Transactions of the Association for Computational Linguistics, 9, 1389–1406.

[13] Fomichov, V. A. (2014). SK-languages as a comprehensive formal environment for developing a Multilingual Semantic Web. In H. Decker, L. Lhotska, S. Link, M. Spies, & R. R. Wagner (Eds.), Database and Expert Systems Applications. DEXA 2014. Lecture Notes in Computer Science, Vol. 8644, pp. 394–401. Springer, Cham.

[14] Liu, W., Trenous, S., Ribeiro, L. F. R., Byrne, B., & Hieber, F. (2025). XRAG: Cross-lingual retrieval-augmented generation. In Findings of ACL: EMNLP 2025 (pp. 15669–15690). Association for Computational Linguistics.

[15] Ranaldi, L., Pucci, G., Zanzotto, F. M., Lenci, A., & Freitas, A. (2025). Multilingual retrieval-augmented generation for knowledge-intensive tasks. arXiv preprint arXiv:2504.03616

[16] Chirkova, N., & Nikoulina, V. (2024). Zero-shot cross-lingual transfer in instruction tuning of large language models. In Proceedings of EACL 2024.

[17] Fomichov, V. A. (2013). A broadly applicable and flexible Conceptual Metagrammar as a basic tool for developing a Multilingual Semantic Web. In E. Métais, F. Meziane, M. Saraee, V. Sugumaran, & S. Vadera (Eds.), NLDB 2013. Lecture Notes in Computer Science, Vol. 7934, pp. 249–259. Springer, Heidelberg.

[18] Lei, Y., Uren, V., & Motta, E. (2006). SemSearch: A search engine for the Semantic Web. In S. Staab & V. Svátek (Eds.), Managing Knowledge in a World of Networks. EKAW 2006. Lecture Notes in Artificial Intelligence, Vol. 4248, pp. 238–245. Springer, Berlin, Heidelberg

TABLE X

TABLE A2B — MIRACL DEV-SET DATASET STATISTICS (16 LANGUAGES). SOURCE: ZHANG ET AL. (2023) [11], TACL TABLE 4.

Lang.	Dev Queries	Dev Judgments	Wiki (M art.)
ar	2,896	29,197	1.19
bn	411	4,583	0.14
zh	393	3,772	1.35
en	799	8,350	6.70
fi	1,271	12,008	0.54
fr	343	3,228	2.53
hi	350	3,888	0.16
id	960	7,116	0.65
ja	860	6,992	1.38
ko	213	1,738	0.63
fa	632	6,571	0.97
ru	1,252	13,100	1.93
sw	482	5,092	0.08
te	828	10,547	0.09
th	733	7,573	0.18
yo	119	1,286	0.03

TABLE XI

TABLE B1A — XNLI LANGUAGE METADATA (15 LANGUAGES). SOURCE: CONNEAU ET AL. (2019) [9].

Language	Code	Resource	Family
English	en	High	Germanic
French	fr	High	Romance
Spanish	es	High	Romance
German	de	High	Germanic
Greek	el	Medium	Hellenic
Bulgarian	bg	Medium	Slavic
Russian	ru	High	Slavic
Turkish	tr	Medium	Turkic
Arabic	ar	High	Semitic
Vietnamese	vi	Medium	Austroasiatic
Thai	th	Medium	Tai-Kadai
Chinese	zh	High	Sino-Tibetan
Hindi	hi	Medium	Indo-Aryan
Swahili	sw	Low	Bantu
Urdu	ur	Low	Indo-Aryan

TABLE XII

TABLE B1B — XNLI ACCURACY (%) BY MODEL. GAIN = XLM-R LARGE − MBERT. SOURCE: CONNEAU ET AL. (2019) [9], TABLES 1–3.

Lang.	XLM-100	mBERT	Base	Large	Gain
en	79.1	82.1	84.0	89.1	+7.0 pp
fr	78.7	79.8	81.4	87.3	+7.5 pp
es	78.9	80.7	82.4	88.3	+7.6 pp
de	77.8	78.7	80.6	86.6	+7.9 pp
el	76.6	77.5	79.0	85.4	+7.9 pp
bg	77.4	79.5	80.9	85.9	+6.4 pp
ru	75.3	76.5	78.7	84.2	+7.7 pp
tr	72.5	74.4	77.5	82.9	+8.5 pp
ar	73.1	74.2	77.1	83.5	+9.3 pp
vi	74.1	75.7	78.3	84.1	+8.4 pp
th	68.9	70.4	74.0	80.4	+10.0 pp
zh	76.5	74.0	78.8	83.7	+9.7 pp
hi	69.7	70.2	72.1	79.9	+9.7 pp
sw	64.4	65.3	70.2	76.5	+11.2 pp
ur	63.2	64.4	68.7	75.6	+11.2 pp

TABLE XIII

TABLE B2 — MLQA EXTRACTIVE QA F1 SCORES (7 LANGUAGES). '—' = NOT REPORTED. SOURCE: CONNEAU ET AL. (2019) [9], TABLE 4.

Lang.	XLM-100	mBERT	Base	Large	Gain
en	74.9	77.0	77.1	83.6	+6.6 pp
ar	51.5	51.5	54.0	62.5	+11.0 pp
zh	60.0	59.2	59.3	68.0	+8.8 pp
hi	59.4	59.0	59.7	68.5	+9.5 pp
de	—	57.9	60.9	70.6	+12.7 pp
es	—	64.3	65.4	74.1	+9.8 pp
vi	—	57.9	61.9	71.5	+13.6 pp

TABLE XIV

TABLE C1 — COMPLETE NOMIRACL DATASET WITH CFS RANKINGS. HALLUCINATION AND ERROR RATES: THAKUR ET AL. (2024) [10]. CFS AND VERDICT: ORIGINAL CONTRIBUTION.

Rank	Language	Res.	Hall.	Error	CFS	Verdict
1	English	High	0.21	0.13	0.34	Adequate
2	French	High	0.27	0.15	0.42	Adequate
3	Spanish	High	0.28	0.16	0.44	Adequate
4	German	High	0.31	0.17	0.48	Adequate
5	Chinese	High	0.33	0.19	0.52	Adequate
6	Russian	High	0.39	0.20	0.59	Adequate
7	Indonesian	Medium	0.38	0.22	0.60	Adequate
8	Japanese	High	0.43	0.24	0.67	Adequate
9	Arabic	High	0.46	0.27	0.73	Adequate
10	Korean	High	0.47	0.28	0.75	Adequate
11	Persian	Medium	0.52	0.30	0.82	Adequate
12	Finnish	Medium	0.54	0.31	0.85	Adequate
13	Hindi	Medium	0.55	0.33	0.88	Adequate
14	Thai	Medium	0.58	0.36	0.94	Marginal
15	Bengali	Medium	0.61	0.38	0.99	Marginal
16	Telugu	Low	0.69	0.48	1.17	Net-Harmful
17	Swahili	Low	0.72	0.51	1.23	Net-Harmful
18	Yoruba	Low	0.89	0.74	1.63	Net-Harmful

TABLE XV

TABLE E1 — INPUT DATA FOR PEARSON CORRELATION. $X = \log_{10}$(WIKIPEDIA ARTICLE COUNT). SOURCES: MIRACL [11]; WIKIPEDIA STATISTICS 2023.

Language	Code	Wiki (M)	$\log_{10}$	BM25 nDCG@10
Arabic	ar	1.19	6.076	0.481
Bengali	bn	0.14	5.146	0.508
Chinese	zh	1.35	6.130	0.207
English	en	6.70	6.826	0.351
Finnish	fi	0.54	5.732	0.551
French	fr	2.53	6.403	0.181
Hindi	hi	0.16	5.204	0.444
Indonesian	id	0.65	5.813	0.449
Japanese	ja	1.38	6.140	0.369
Korean	ko	0.63	5.799	0.419
Persian	fa	0.97	5.987	0.298
Russian	ru	1.93	6.286	0.334
Swahili	sw	0.08	4.903	0.383
Telugu	te	0.09	4.954	0.494
Thai	th	0.18	5.255	0.484
Yoruba	yo	0.03	4.477	0.050

TABLE XVI

TABLE E2 — PEARSON CORRELATION RESULTS: log(WIKIPEDIA SIZE) VS. MIRACL RETRIEVAL PERFORMANCE. BOTH CORRELATIONS ARE NON-SIGNIFICANT AT $\alpha = 0.05$. 95% CI VIA FISHER z-TRANSFORMATION.

Comparison	N	Pearson r	p-value	95% CI
X vs BM25 nDCG@10	16	−0.061	0.822	[−0.54, 0.45]
X vs InLang mDPR	16	−0.227	0.398	[−0.64, 0.29]

TABLE XVII

TABLE F1 — DATA PROVENANCE FOR ALL DATASETS USED IN THIS ANALYSIS.

Dataset	Version	Source	License	Access
MIRACL Dev Baselines	v1.0 (2023)	huggingface.co/datasets/miracl/miracl	CC BY-SA 4.0	Public
NoMIRACL	v1.0 (2024)	huggingface.co/datasets/miracl/nomiracl	CC BY-SA 4.0	Public
XLM-R Scores	arXiv:1911.02116	arxiv.org/abs/1911.02116	CC BY 4.0	Public
Wikipedia Counts	Jan. 2023	en.wikipedia.org/wiki/Wikipedia:Size	CC BY-SA 4.0	Public

TABLE XVIII

TABLE F2 — SOFTWARE ENVIRONMENT FOR DATA ANALYSIS AND DOCUMENT GENERATION.

Component	Version	Purpose
Python	3.11.4	Primary analysis language
pandas	2.0.3	Data loading and aggregation
numpy	1.24.3	Numerical operations
matplotlib	3.7.2	Figure generation (Figs. 2-12)
seaborn	0.12.2	Statistical visualisation
scipy	1.11.2	Pearson correlation
scikit-learn	1.3.0	Supplementary metrics