IRSC: A Zero-shot Evaluation Benchmark for Information Retrieval through Semantic Comprehension in Retrieval-Augmented Generation Scenarios

Hai $Lin^{1,2*}$, Shaoxiong $Zhan^{2*}$, Junyou Su^{3*} , Haitao $Zheng^{1,2}$, Hui $Wang^{1\dagger}$

¹PengCheng Laboratory
²Shenzhen International Graduate School, Tsinghua University
³Southern University of Science and Technology

Abstract

In Retrieval-Augmented Generation (RAG) tasks using Large Language Models (LLMs), the quality of retrieved information is critical to the final output. This paper introduces the IRSC benchmark for evaluating the performance of embedding models multilingual RAG tasks. The benchmark encompasses five retrieval tasks: retrieval, title retrieval, part-of-paragraph retrieval, keyword retrieval, and summary retrieval. Our research addresses the current lack of comprehensive testing and effective comparison methods for embedding models in RAG scenarios. We introduced new metrics: the Similarity of Semantic Comprehension Index (SSCI) and the Retrieval Capability Contest Index (RCCI), and evaluated models such as Snowflake-Arctic, BGE, GTE, and M3E. Our contributions include: 1) the IRSC benchmark, 2) the SSCI and RCCI metrics, and 3) insights into the cross-lingual limitations of embedding models. The IRSC benchmark aims to enhance the understanding and development of accurate retrieval systems in RAG tasks. All code and datasets are available at: https://github.com/Jasaxion/IRSC Benchmark

1 Introduction

The rapid advancements in large language models (LLMs) have demonstrated significant potential in natural language understanding and generation. However, these models still face challenges like factual hallucination, knowledge updating, and lack of domain-specific expertise (Chen et al., 2024b). To address these issues, incorporating external knowledge through Retrieval-Augmented Generation (RAG) has emerged as a promising approach (Chen et al., 2024b; Zhang et al., 2023).

RAG enhances LLMs by integrating retrieved information from external sources, which helps

mitigate hallucinations and provide more accurate, up-to-date responses (Chen et al., 2024b). Despite these advantages, existing benchmarks for evaluating RAG models are limited in scope and do not fully address the diverse needs of various retrieval tasks (Chen et al., 2024b; Zhang et al., 2023). Most benchmarks focus primarily on tasks such as semantic textual similarity (STS), clustering, and reranking, but fail to provide RAG comprehensive evaluations across different retrieval scenarios.

The IRSC Benchmark introduced in this study aims to fill this gap by evaluating Embedding models across five distinct retrieval tasks: query-based retrieval, title-based retrieval, part-of-paragraph retrieval, keyword-based retrieval and summarybased retrieval. This benchmark is designed to reflect realistic application scenarios of RAG, considering different types of queries and languages (English, Chinese, and Mixed-Language datasets) (Zhang et al., 2023). We evaluate models such as Snowflake-Arctic-Embed-S (Merrick et al., 2024), BGE-M3 (Chen et al., 2024a), and (Wang Yuxin, 2023) across different tasks and languages, providing insights into their strengths and weaknesses in real-world RAG applications. Additionally, this benchmark includes innovative evaluation metrics to capture model performance differences across tasks and languages.

And due to the differences in vector dimensions and values across various models, directly computing cosine similarity between vectors (Steck et al., 2024) is not feasible for comparing the semantic similarity between different models (Zhou et al., 2022). To address this, we propose the Similarity of Semantic Comprehension Index (SSCI) in this paper. SSCI measures the similarity of semantic understanding between the model's output and the ground truth.

Our contributions are as follows: 1. We propose a comprehensive IRSC Benchmark to evaluate the performance of embedding models in RAG

^{*}These authors contributed equally to this work.

[†]Corresponding author.

retrieval tasks and languages. 2. We introduce the SSCI and the Retrieval Capability Contest Index (RCCI) as innovative metrics to evaluate and compare models' semantic understanding and retrieval capabilities, respectively. 3. We conducted experiments on the retrieval effect of the model across languages and found the differences in the semantic understanding alignment of the model in different languages.

2 Related Work

The field of Retrieval-Augmented Generation (RAG) has gained significant attention, especially in addressing the limitations of Large Language models (LLMs) in providing accurate and contextually relevant information. This section reviews notable works in this domain and situates our contribution within the existing research.

Benchmarking in RAG

Chen et al. developed the Retrieval-Augmented Generation Benchmark (RGB) to evaluate LLMs on four abilities: noise robustness, negative rejection, information integration, and counterfactual robustness. Their findings highlight the need for nuanced evaluation metrics to improve RAG capabilities, as LLMs showed weaknesses in negative rejection, information integration, and handling false information(Chen et al., 2024b). However, RGB primarily focuses on robustness aspects and does not provide comprehensive coverage of different retrieval tasks, which is crucial for real-world RAG applications.

Multilingual Retrieval Datasets

The MIRACL dataset, introduced by Zhang et al., supports multilingual information retrieval with 700,000 human-annotated query-passage pairs across 18 languages. It aims to advance retrieval models that handle linguistic diversity and resource variability (Zhang et al., 2023). While MIRACL provides valuable multilingual data, it is mainly focused on query-passage retrieval and does not address other important retrieval tasks like keyword or title retrieval.

Evaluations of RAG Systems

Ogundepo et al. provided a comprehensive survey of current evaluation methods for RAG systems, emphasizing the importance of various retrieval tasks and metrics like nDCG, MRR, and MAP (Yu et al., 2024). Their work discusses challenges and future directions for robust RAG benchmarks. Despite their comprehensive survey, there is

a lack of practical benchmarks that integrate these varied metrics across different retrieval tasks.

Benchmark for Evaluation of Information Retrieval Models (BEIR)

Thakur et al. introduced an evaluation benchmark for retrieval models called BEIR, which includes a diverse set of information retrieval tasks across different domains and data types(Thakur et al., 2021). BEIR offers a collection of heterogeneous tasks and provides a unified and convenient framework for evaluating retrieval models based on natural language processing. However, BEIR only focuses on retrieval tasks between queries and paragraphs, and does not address the more complex retrieval tasks involving large-scale RAG (Retrieval-Augmented Generation) scenarios.

Massive Text Embedding Benchmark (MTEB)

Muennighoff et al. introduced MTEB, a benchmark evaluating text embedding models across tasks such as bitext mining, classification, clustering, reranking, retrieval, and semantic textual similarity(Muennighoff et al., 2023). Their findings highlight the need for specialized models tailored to specific retrieval scenarios. However, MTEB does not focus specifically on the integration of retrieved information for generation tasks, which is a critical component of RAG systems.

Multilingual Question Answering

The MKQA dataset, presented by Longpre et al., evaluates multilingual open-domain question answering systems with parallel questions in multiple languages(Longpre et al., 2021). It facilitates comparative analysis of retrieval and QA performance across different linguistic contexts. While useful for question answering, MKQA does not encompass the broader spectrum of retrieval tasks that are essential for RAG evaluations.

Current Work on RAG Model Evaluation

Our work extends these studies by proposing a novel Benchmark that evaluates retrieval performance across five tasks: query, keyword, title, summary, and part of paragraph retrieval. Unlike previous benchmarks, our dataset includes multilingual and cross-lingual components, addressing the need for robust evaluation in diverse linguistic environments. Our benchmark aims to fill gaps in existing methods by providing a comprehensive assessment of model performance in RAG tasks, focusing on cross-lingual retrieval and integrating various retrieval tasks into a unified framework.

3 The IRSC Benchmark

3.1 Desiderata

The IRSC benchmark is designed to evaluate the effectiveness of embedding models specifically within the context of Retrieval-Augmented Generation (RAG) tasks. Unlike traditional benchmarks that focus broadly on sentence or paragraph length, IRSC hones in on the unique needs of RAG applications, which require supplementing knowledge to queries. This benchmark emphasizes five key data types to cover most RAG tasks:

- 1. Focus on RAG-Specific Retrieval Tasks: Unlike traditional benchmarks, IRSC focuses on expanding a query or brief information into a detailed response.
- 2. Emphasis on Cross-lingual Capabilities: IRSC evaluates models in multiple languages, particularly English and Chinese, to handle Mixed-Language queries and adapt to cross-lingual environments.
- 3. Comprehensive Evaluation Metrics: Standard retrieval metrics (nDCG@10, MRR@10, MAP@10, precision@3, and recall@10) are used alongside new metrics like SSCI and RCCI for deeper insights into semantic comprehension and retrieval capabilities.
- 4. **Real-World Applicability:** IRSC focuses on real-world RAG tasks like retrieving detailed knowledge based on a query or summary, ensuring practical relevance and cross-lingual applicability.

Through these considerations, IRSC aims to set a new standard for evaluating embedding models in the context of RAG tasks, providing a more nuanced and applicable assessment framework.

3.2 Tasks and Evaluation

Figure 1 provides an overview of tasks and datasets available in IRSC. The benchmark consists of the following five task types:

1. **Query -> Paragraph:** Evaluates the model's ability to retrieve relevant paragraphs based on a given query. Datasets: MsMARCO(Bajaj et al., 2018), XQuAD(Artetxe et al., 2020), Xtreme(Hu et al., 2020), and MLQA(Lewis et al., 2020).

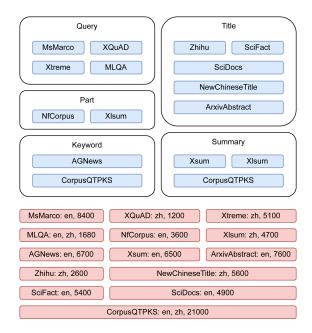


Figure 1: The IRSC Benchmark is structured around five primary task types, each designed to evaluate different aspects of a model's retrieval capabilities. The red labels indicate the languages and quantities of each dataset.

- 2. **Title -> Paragraph:** Tests the model's capability to find relevant paragraphs given a title. Datasets: zhihu*, New-Title-Chinese[†], Arxiv-Abstract(Clement et al., 2019), Sci-Docs(Muennighoff et al., 2023), and Sci-Fact(Muennighoff et al., 2023).
- 3. Part of Paragraph -> Paragraph: Evaluates sensitivity to text fragments, testing if the model can retrieve the full paragraph from a fragment. Datasets: nfcorpus(Muennighoff et al., 2023) (English) and xlsum(Joulin et al., 2017) (Chinese).
- Keyword -> Paragraph: Measures the model's ability to retrieve paragraphs based on keywords. Datasets: AG News(Zhang et al., 2015) and CorpusQTPKS.
- 5. **Summary -> Paragraph:** Evaluates performance in retrieving relevant paragraphs based on a summary. Datasets: XSum(Narayan et al., 2018) (English), xlsum(Joulin et al., 2017) (Chinese), and CorpusQTPKS.

Each task uses 5000 query-content pairs as evaluation data. The remaining samples form a unified

^{*}https://huggingface.co/datasets/suolyer/zhihu

 $^{^\}dagger$ https://huggingface.co/datasets/madao33/new-title-chinese

database for retrieval across all tasks. During scoring, queries are used to search the unified database, and retrieval performance is evaluated based on the precision of retrieved indices against the ground truth. This standardized approach ensures robust and fair evaluation of the model's retrieval capabilities across various tasks.

3.3 Evaluation Metrics

The primary evaluation metrics for IRSC include nDCG@10, MRR@10, MAP@10, precision@3, and recall@10. These metrics are used to evaluate model performance in information retrieval tasks:

Recall@10: Evaluates the fraction of relevant documents retrieved among the top 10 documents.

MRR@10 (Mean Reciprocal Rank): Evaluates the rank position of the first relevant document.

nDCG@10 (Normalized Discounted Cumulative Gain): Measures the ranking quality of the retrieved documents, taking into account the position of relevant documents in the ranking.

In addition to these standard metrics, we introduce new metrics to assess different aspects of model performance:

Similarity of Semantic Comprehension Index (SSCI)

The Similarity of Semantic Comprehension Index, averaged over multiple queries. It measures the difference in semantic understanding between the two models' outputs across all queries. A higher value indicates a greater disparity in the models' understanding of the given questions.

We define the average SSCI (\overline{SSCI}) as:

$$\overline{\text{SSCI}} = \frac{1}{Q} \sum_{q=1}^{Q} \frac{|m1_q - m2_q|}{n}$$

Retrieval Capability Contest Index (RCCI)

The Retrieval Capability Contest Index, averaged over multiple queries. It evaluates the differences in retrieval capabilities between the two models across all queries. A positive average score indicates that model 1 performs better overall, while a negative average score indicates that model 2 performs better overall. The magnitude of the score indicates the extent of the difference in performance.

We define the average RCCI (\overline{RCCI}) as:

$$\overline{\text{RCCI}} = \frac{1}{Q} \sum_{q=1}^{Q} \frac{m1_q - m2_q}{n}$$

Parameters

- n: The length of each query result vector minus one, representing the maximum index of the retrieval results.
- R1, R2: These are matrices representing the results retrieved by the two different models over Q queries. Each matrix has dimensions Q × (n + 1). Each element is a binary value (0 or 1), where 1 indicates the position of the correct answer for each query.

For query q, the vectors $R1_q$ and $R2_q$ are defined as follows:

$$R1_q = [r_{11q}, r_{12q}, r_{13q}, \dots, r_{1nq}, r_{1(n+1)q}]$$

$$R2_q = [r_{21q}, r_{22q}, r_{23q}, \dots, r_{2nq}, r_{2(n+1)q}]$$

• $m1_q$, $m2_q$: These represent the positions of the correct answer in $R1_q$ and $R2_q$ respectively for query q. If there is no 1 in the vector, it is assigned a value of -1.

For query q:

$$m1_q = \begin{cases} n - \operatorname{index}(R1_q, 1) & \text{if } 1 \in R1_q \\ -1 & \text{otherwise} \end{cases}$$

$$m2_q = \begin{cases} n - \operatorname{index}(R2_q, 1) & \text{if } 1 \in R2_q \\ -1 & \text{otherwise} \end{cases}$$

where index $(R_q, 1)$ denotes the index position of the element equal to 1 in the vector R_q .

By using these metrics, IRSC aims to provide a comprehensive evaluation framework for assessing the performance of embedding models across diverse retrieval tasks.

3.4 Model Descriptions

We evaluated 13 models using the IRSC Benchmark. Notably, MiniLM-L6-v2 models do not support Chinese.

- S-Arctic Series(Merrick et al., 2024): Includes S-Arctic-S, S-Arctic-M, and S-Arctic-L, designed for semantic embeddings in text retrieval tasks.
- **BGE Series**(Chen et al., 2024a): Includes BGE-M3 (multilingual) and BGE-Large (optimized for Chinese).

- **GTE Series**(Li et al., 2023): Comprises GTE-Small, GTE-Base, and GTE-Large, focusing on general text embeddings.
- M3E Series(Wang Yuxin, 2023): Includes M3E-Small, M3E-Base, and M3E-Large, designed for efficient multilingual text embeddings.
- MiniLM Series: MiniLM-L12(Reimers and Gurevych, 2019): A multilingual version of the MiniLM series, tailored for paraphrase identification and multilingual retrieval. MiniLM-L6*: A compact, efficient model focused on English for various NLP tasks.

4 Results

4.1 Experimental Setup

The experiments are conducted across three different language requirements: English, Chinese, and Mixed-Language (English + Chinese). For each language requirement, corresponding benchmark datasets are utilized to perform IRSC scoring experiments.

English: The evaluation involves English-specific benchmark datasets to test the retrieval performance of each model.

Chinese: The evaluation uses Chinese-specific benchmark datasets, ensuring the models' capabilities are tested in the Chinese language context.

Mixed-Language (English + Chinese): This mixed evaluation assesses the models' performance across both English and Chinese datasets, providing a comprehensive understanding of their crosslingual retrieval capabilities.

By employing this diversified language setup, we aim to provide a thorough and robust evaluation of each model's performance in retrieving relevant paragraphs based on various query types within the IRSC benchmark.

4.2 Experimental Results and Analysis

4.2.1 Benchmark Analysis

Based on the results in Table 1, we observe that the BGE-M3 model consistently outperforms other models across all metrics and categories, indicating its robustness and effectiveness in both Chinese and English retrieval tasks. Specifically, BGE-M3

*https://huggingface.co/ sentence-transformers/MiniLM-L6-v2 achieves the highest recall at 10 (r@10), mean average precision at 10 (m@10), and normalized discounted cumulative gain at 10 (n@10) in the Keywords, Title, Query, Part, and Summary categories. For instance, in the Keywords category, BGE-M3 has an impressive r@10 of 0.8668, m@10 of 0.8205, and n@10 of 0.8320.

Conversely, the S-Arctic series (S-Arctic-S, S-Arctic-M, S-Arctic-L) shows relatively lower performance compared to other models. Notably, S-Arctic-S performs better than S-Arctic-M and S-Arctic-L across most categories, but it still lags significantly behind models like BGE-M3, GTE-Small, and M3E-Base. For example, in the Summary category, S-Arctic-S achieves an r@10 of 0.5334, whereas BGE-M3 achieves an r@10 of 0.9812.

Models like GTE-Base and M3E-Base also demonstrate strong performance, particularly in the Keywords and Summary categories. GTE-Base achieves an r@10 of 0.7940 in the Keywords category and M3E-Base achieves an r@10 of 0.9644 in the Summary category, showing their potential effectiveness in specific retrieval contexts.

We observe an interesting performance pattern between the S-Arctic-S and M3E-Small models. Notably, S-Arctic-S performs significantly better than M3E-Small in the Keywords category. S-Arctic-S achieves an r@10 of 0.6302, while M3E-Small scores 0.3374. However, in Summary categories, M3E-Small significantly outperforms S-Arctic-S. M3E-Small achieves an r@10 of 0.8052 compared to S-Arctic-S's 0.5334. This pattern indicates that while S-Arctic-S excels in Keywords retrieval, it falls behind in other tasks such as Summary, where M3E-Small demonstrates superior performance.

The diverse performance across different models highlights the importance of selecting the appropriate model based on the specific retrieval task and the language requirements. Table 1 presents the results for Mixed-Language task. The results for the Chinese and English tasks will publish in our Github repository.

4.2.2 Radar Chart Analysis

To more intuitively showcase the performance of different models, we created radar charts 2 where the values for each capability are derived from the average of r@10, m@10, and n@10. These charts provide a clearer view of the comprehensive performance of each model across various tasks.

Model	Query			Title			Part			Keywords			Summary		
	r@10	m@10	n@10	r@10	m@10	n@10	r@10	m@10	n@10	r@10	m@10	n@10	r@10	m@10	n@10
S-Arctic-S	0.3067	0.2714	0.2815	0.3566	0.3125	0.3232	0.4588	0.4369	0.4423	0.6302	0.5759	0.5892	0.5334	0.5231	0.5256
S-Arctic-M	0.1379	0.1125	0.1196	0.0198	0.0145	0.0158	0.2746	0.2514	0.2570	0.2856	0.2339	0.2464	0.4554	0.4151	0.4247
S-Arctic-L	0.2238	0.1909	0.2002	0.0248	0.0186	0.0200	0.3126	0.2903	0.2957	0.3804	0.3342	0.3455	0.4394	0.4104	0.4175
BGE-M3	0.6972	0.6321	0.6495	0.8640	0.8149	0.8270	0.7964	0.7625	0.7708	0.8668	0.8205	0.8320	0.9812	0.9709	0.9735
GTE-Small	0.5099	0.4643	0.4771	0.7360	0.7005	0.7093	0.6916	0.6527	0.6622	0.7876	0.7258	0.7409	0.8284	0.7890	0.7986
GTE-Base	0.5163	0.4684	0.4817	0.7366	0.7027	0.7110	0.6980	0.6617	0.6706	0.7940	0.7333	0.7482	0.8282	0.7893	0.7987
GTE-Large	0.5205	0.4746	0.4874	0.7372	0.7022	0.7107	0.6984	0.6571	0.6672	0.7936	0.7338	0.7485	0.8180	0.7780	0.7877
M3E-Small	0.2292	0.1874	0.1981	0.2850	0.2267	0.2407	0.4942	0.4522	0.4624	0.3374	0.2874	0.2993	0.8052	0.7572	0.7689
M3E-Base	0.5912	0.5239	0.5415	0.7840	0.7167	0.7332	0.7562	0.7146	0.7249	0.8368	0.7777	0.7923	0.9644	0.9441	0.9491
M3E-Large	0.3415	0.2798	0.2957	0.5052	0.4113	0.4340	0.5788	0.5302	0.5421	0.5606	0.4701	0.4919	0.8964	0.8527	0.8634
MiniLM-L6	0.4589	0.4180	0.4297	0.6168	0.5842	0.5923	0.6066	0.5720	0.5805	0.7042	0.6326	0.6500	0.5484	0.5190	0.5260
MiniLM-L12	0.4934	0.4161	0.4363	0.6174	0.5294	0.5508	0.5708	0.5259	0.5368	0.5784	0.4846	0.5073	0.8728	0.8245	0.8365

Table 1: IRSC Benchmark Results of S-Arctic Series, BGE Series, GTE Series, M3E Series, and MiniLM Series in All Languages for All Tasks. **Metrics:** r@10 - Recall at 10, m@10 - MRR(Mean Reciprocal Rank) at 10, n@10 - nDCG(Normalized Discounted Cumulative Gain) at 10

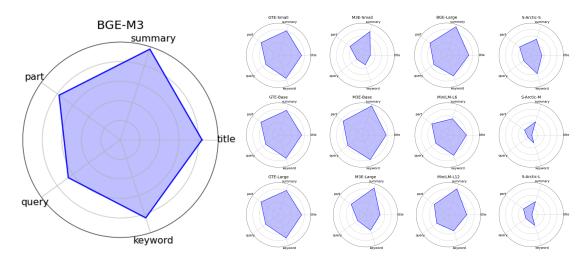


Figure 2: Comparative Performance Radar Charts of S-Arctic Series, BGE Series, GTE Series, M3E Series, and MiniLM Series Models Across IRSC Benchmark's Query, Title, Part, Keyword and Summary Tasks in Mixed-Language. **Metrics:** Average of Recall@10, MRR@10 and nDCG@10

From the charts, it is evident that BGE-M3 performs exceptionally well in all tasks (Query, Title, Part, Keywords, and Summary), demonstrating its comprehensive advantages across these five areas. The radar chart for BGE-M3 shows a balanced and extensive coverage, with particularly outstanding performance in the Summary tasks.

In contrast, the GTE series and M3E series models also show good performance. However, the S-Arctic series underperforms compared to the aforementioned models in all tasks, especially S-Arctic-M, which shows the lowest comprehensive performance across all tasks, indicating its lesser effectiveness in these tasks.

The radar charts clearly illustrate the comprehensive capabilities of each model across different tasks, with BGE-M3 standing out as the most optimal model in terms of performance.

4.2.3 Cross Language Analysis

In Table 2 , we also conducted experiments on the cross-lingual retrieval capabilities of different models using five IRSC tasks, with 1,000 randomly selected queries for each task. We obtained 5,000 data entries in both English and Chinese languages. Queries originally in English were translated into the target language (Chinese) and then searched within an entirely English database to obtain the Chinese to English (C2E) results in Table 2. The scores are the averages of $r@\,10$, $m@\,10$, and $n@\,10$.

From the results, several key observations can be made:

Performance Decline in Cross-Lingual Retrieval: Most models exhibit a decline in performance metrics when transitioning from monolingual (C2C or E2E) to cross-lingual (C2E or E2C) retrieval. This indicates a general challenge in maintaining semantic align-

Model	C2C C2E	E2E E2C
S-Arctic-S	0.0782 0.0068	0.5848 0.0462
S-Arctic-M	$0.1272 \mid 0.0014$	0.1441 0.0208
S-Arctic-L	$0.0882 \mid 0.0008$	$0.2008 \mid 0.0334$
BGE-M3	0.8630 0.6260	$0.8427 \mid 0.5964$
GTE-Small	0.4088 0.0569	0.8499 0.0620
GTE-Base	0.4048 0.0581	0.8613 0.0866
GTE-Large	0.4036 0.0651	0.8693 0.0888
M3E-Small	0.7486 0.0660	$0.1327 \mid 0.0190$
M3E-Base	$0.8026 \mid 0.3323$	0.7423 0.1578
M3E-Large	0.7648 0.2420	0.3659 0.0688
MiniLM-L6	0.1048 0.0209	$0.7942 \mid 0.0150$
MiniLM-L12	0.5586 0.3841	0.5872 0.4558

Table 2: IRSC Benchmark Results of the S-Arctic Series, BGE Series, GTE Series, M3E Series, and MiniLM Series in Cross Languages. **Metrics:** recall@10

ment across different languages.

- 2. **Superior Performance of BGE-M3**: The BGE-M3 model consistently demonstrates superior performance in both monolingual and cross-lingual retrieval tasks. Notably, its performance degradation from monolingual to cross-lingual retrieval is minimal. For instance, in C2C, it scores 0.8630, while in C2E, it scores 0.6260. Similarly, in E2E, it scores 0.8427, compared to 0.5964 in E2C.
- 3. **Significant Decline in M3E Series**: The M3E series models show a significant decrease in performance when moving to cross-lingual tasks. The most notable drop is observed in the M3E-Base model, which falls from 0.8026 in C2C to 0.3323 in C2E. This highlights a substantial challenge in the model's ability to align queries semantically across languages.
- 4. **Drastic Decline in GTE Series**: The GTE series models exhibit the most drastic decline in performance, especially in the E2C task. Scores around 0.85 in E2E drop below 0.1 in E2C, indicating a significant deficiency in the models' ability to handle cross-lingual semantic alignment from English to Chinese.
- 5. Mixed Performance in S-Arctic and MiniLM Series: The S-Arctic series models display varying levels of performance, with S-Arctic-S and S-Arctic-M performing poorly in C2E tasks. The MiniLM series also shows mixed results, with MiniLM-L12 performing

relatively well compared to MiniLM-L6 in cross-lingual tasks.

These findings underscore the need for further improvement in training vector models for cross-lingual query semantic alignment. Enhancing the models' ability to maintain semantic coherence across languages could lead to more effective and accurate cross-lingual retrieval systems. Future research should focus on developing techniques to bridge the semantic gap between languages, ensuring that models can perform consistently well in both monolingual and cross-lingual contexts.

4.3 SSCI & RCCI Analysis

4.3.1 SSCI

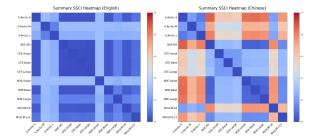


Figure 3: Comparative SSCI Heatmaps of the S-Arctic Series, BGE Series, GTE Series, M3E Series, and MiniLM Series in the IRSC Benchmark's Summary Subtask Across Chinese and English. Smaller values indicate more consistent model performance.

In Figure 3, we present detailed SSCI results for the Summary task across different languages and models. We observe that in the English language, most models display blue regions, indicating high consistency in semantic understanding among these models. In contrast, for the Chinese language, the SSCI values exhibit more red regions, suggesting lower consistency and greater divergence in semantic understanding among the models.

Furthermore, by examining the color distribution in Figure 3, we find that models within the same series generally exhibit better semantic understanding consistency, whereas models from different series are more likely to show divergence in understanding. From this analysis, we can draw several conclusions: there is a significant difference in semantic understanding consistency across languages, with models showing higher consistency in English compared to Chinese; models within the same series tend to have higher semantic understanding consistency, while different series of

models are more prone to divergences in understanding.

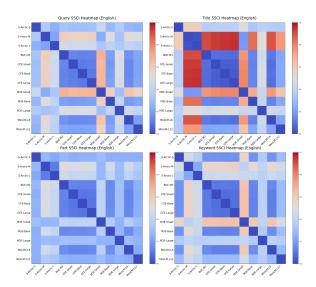


Figure 4: Comparative SSCI Heatmaps of the S-Arctic Series, BGE Series, GTE Series, M3E Series, and MiniLM Series in the IRSC Benchmark's Query, Title, Part and Keyword Subtasks in English. Smaller values indicate more consistent model performance.

Figure 4 illustrates the SSCI heatmaps for models on four tasks (excluding Summary) in the English language. From the figure 4, it is evident that different tasks exhibit varying degrees of divergence in model understanding. Unlike the Summary task, where most models show blue regions indicating high SSCI values, the IRSC tasks reveal different levels of red regions. This is particularly evident in the Title task, which shows extensive deep red regions, indicating significant divergence in semantic understanding among the models. The Title task imposes higher demands on the models' semantic understanding, highlighting the differences in model performance across different task types. While models show high consistency in the Summary task, possibly due to its clear objective and relatively smaller information processing requirements, the Title task requires complex context understanding and concise expression, leading to more pronounced divergences among the models.

From the analysis of Figure 4, we can derive that the Summary task shows high model consistency, whereas the IRSC tasks, especially the Title task, exhibit significant divergences, indicating that task complexity has a substantial impact on model consistency. Tasks requiring complex context understanding and concise expression (such as Title) reveal significant divergences in model perfor-

mance, exposing limitations in these areas. Future model training should focus on enhancing models' capabilities in complex context understanding and concise expression to address the challenges posed by complex tasks.

4.3.2 RCCI

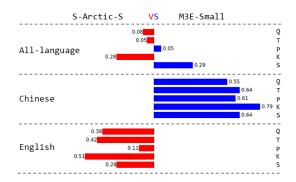


Figure 5: Comparison of RCCI Results Between S-Arctic-S and M3E-Small Across Mixed-Languages, Chinese, and English

The figure 5 presents a comparative analysis of the capabilities of two models, S-Arctic-S and M3E-Small, across multiple languages and evaluation metrics. The RCCI methodology has been employed to offer a detailed comparison between the models, which overcomes the limitations of simple average-based metrics such as r@10, m@10, and n@10 by highlighting finer differences in model performance.

Figure 5 illustrates the RCCI-based analysis clearly delineates the strengths and weaknesses of the two models. M3E-Small excels in the Chinese language, demonstrating robust performance across all metrics, which underscores its potential for applications requiring Chinese language proficiency. Conversely, S-Arctic-S exhibits a competitive edge in the English language metrics, particularly in the Q (Query), T (Title), and K (Keyword) metrics.

5 Conclusion

The IRSC Benchmark offers a comprehensive evaluation framework for embedding models in Retrieval-Augmented Generation (RAG) tasks. It includes five retrieval tasks: query-based, title-based, part-of-paragraph-based, keyword-based, and summary-based retrieval, in English and Chinese. Key contributions are the IRSC Benchmark, new metrics like the Similarity of Semantic Comprehension Index (SSCI) and Retrieval Capability Contest Index (RCCI), and a cross-lingual performance evaluation.

Experimental results show BGE-M3's superior performance across various metrics and tasks, highlighting its robust retrieval capabilities in monolingual and cross-lingual contexts. Diverse model performance emphasizes the importance of selecting models based on specific tasks and language needs. Cross-lingual retrieval challenges indicate a need for improved training in vector models for better semantic alignment. Visual tools like radar charts and heatmaps illustrate model strengths and weaknesses in different tasks and languages, showcasing the IRSC Benchmark's comprehensive evaluation.

Future research should focus on optimizing embedding models for complex tasks and improving cross-lingual semantic alignment.

6 Limitations

While the IRSC Benchmark provides a comprehensive evaluation framework for embedding models in Retrieval-Augmented Generation (RAG) tasks, there are several limitations that need to be addressed:

- 1. Language Scope: The benchmark primarily focuses on English and Chinese, which limits its applicability to other languages. While it provides insights into multilingual capabilities, extending the evaluation to a broader range of languages would offer a more holistic view of model performance in truly multilingual settings.
- Task Scope:: Although the benchmark covers five distinct retrieval tasks, real-world RAG applications might involve more complex and diverse scenarios. Expanding the range of tasks to include more specialized or domainspecific queries could provide a more comprehensive assessment.
- 3. **Model Variability**: The benchmark evaluates a selection of popular embedding models, but it does not encompass all existing models. New models and variations are continuously being developed, and the benchmark needs to be updated regularly to include these advancements.
- Interoperability with Other Systems: The benchmark does not assess how well these embedding models integrate with other systems and technologies used in RAG pipelines.

Evaluating interoperability and integration efficiency could provide a more practical measure of model utility.

Addressing these limitations in future research will help improve the robustness and applicability of the IRSC Benchmark, making it a more powerful tool for evaluating and developing embedding models in RAG tasks.

References

- Mikel Artetxe, 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*. Association for Computational Linguistics.
- Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.
- Jianly Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. arXiv preprint arXiv:2402.03216.
- Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024b. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.
- Colin B Clement, Matthew Bierbaum, Kevin O'Keeffe, and Alexander A Alemi. 2019. On the use of arxiv as a dataset. *On the use of the arXiv as a dataset*.
- Junjie Hu, 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 *International Conference on Machine Learning*, pages 4411–4421. PMLR.
- Armand Joulin, Edouard Grave, and Piotr Bojanowski Tomas Mikolov. 2017. Bag of tricks for efficient text classification. *EACL 2017*, page 427.
- Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards

- general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.
- Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.
- Luke Merrick, Danmei Xu, Gaurav Nuti, and Daniel Campos. 2024. Arctic-embed: Scalable, efficient, and accurate text embedding models. *Preprint*, arXiv:2405.05374.
- Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.
- Shashi Narayan, Shay Cohen, and Maria Lapata. 2018. Don't give me the details, just the summary! topicaware convolutional neural networks for extreme summarization. In 2018 Conference on Empirical Methods in Natural Language Processing, pages 1797–1807. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. 2024. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings* of the ACM on Web Conference 2024, pages 887–890.
- Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- He sicheng Wang Yuxin, Sun Qingxuan. 2023. M3e: Moka massive mixed embedding model.
- Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024. Evaluation of retrieval-augmented generation: A survey. *arXiv e-prints*, pages arXiv–2405.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.
- Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. Making a miracl: Multilingual information retrieval across a continuum of languages. arXiv preprint arXiv:2210.09984.

Kaitlyn Zhou, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky. 2022. Problems with cosine as a measure of embedding similarity for high frequency words. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 401–423.