Article

Not peer-reviewed version

A Comprehensive Evaluation of Embedding Models and LLMs for IR and QA across English and Italian

Ermelinda Oro *

Posted Date: 27 February 2025

doi: 10.20944/preprints202502.2143.v1

Keywords: multilingual embeddings; information retrieval; large language models; natural language processing; question answering; retrieval-augmented generation; language model evaluation, cross-lingual; English-Italian benchmarking

Article

A Comprehensive Evaluation of Embedding Models and LLMs for IR and QA Across English and Italian

Ermelinda Oro [1,2,*], Francesco Maria Granata [2] and Massimo Ruffolo [2]

[1] Institute for High Performance Computing and Networking, National Research Council, via P. Bucci 8/9C, Rende (CS), 87036, Italy
[2] Altilia srl, TechNest Innovative Companies Incubator of University of Calabria, Piazza Vermicelli, Rende (CS), 87036, Italy
* Correspondence: ermelinda.oro@cnr.it

Abstract: This study presents a comprehensive evaluation of state-of-the-art embedding techniques and large language models (LLMs) for enhancing Information Retrieval (IR) and Question Answering (QA) tasks across multiple languages, with a focus on English and Italian. Our work addresses a critical gap in the current literature by providing empirical evidence of model performance across linguistic boundaries. For IR tasks, we evaluate 12 embedding models across diverse datasets including SQuAD, DICE, SciFact, ArguAna, and NFCorpus. For QA tasks, we employ 4 LLMs (GPT4o, LLama-3.1 8B, Mistral-Nemo, and Gemma-2b) in a retrieval-augmented generation (RAG) pipeline, evaluating on SQuAD, CovidQA, and NarrativeQA datasets, including cross-lingual scenarios. Results demonstrate that multilingual models achieve competitive performance compared to language-specific ones, with embed-multilingual-v3.0 attaining top nDCG@10 scores of 0.90 and 0.86 for English and Italian respectively. In QA tasks, Mistral-Nemo excels in answer relevance (0.91-1.0) while maintaining strong groundedness (0.64-0.78). Our findings reveal that: (1) multilingual embedding models effectively bridge cross-lingual performance gaps, (2) model size does not consistently correlate with performance, and (3) QA systems exhibit a critical trade-off between answer relevance and factual groundedness. Our evaluation framework combines traditional metrics with novel LLM-based assessments, establishing new benchmarks for multilingual language technologies and providing actionable insights and practical guidelines for deploying IR and QA systems in real-world applications.

Keywords: multilingual embeddings; information retrieval; large language models; natural language processing; question answering; retrieval-augmented generation; language model evaluation, cross-lingual; English-Italian benchmarking

1. Introduction

The exponential growth of digital information has made efficient information access and retrieval a critical challenge. Two key technologies have emerged to address this need: Information Retrieval (IR) and Question Answering (QA). IR systems excel at searching through large data collections to find relevant content, while QA systems go a step further by extracting and formulating precise answers to specific queries. Together, these technologies form the backbone of modern information access systems, enabling users to navigate and extract meaning from vast amounts of digital content. These capabilities are essential for applications ranging from enterprise search to personal digital assistants, making IR and QA fundamental technologies in our data-driven world. The landscape of IR and QA has been transformed by recent breakthroughs in Natural Language Processing (NLP). Two technological advances have been particularly influential: Large Language Models (LLMs) and sophisticated embedding techniques. LLMs have revolutionized text understanding and generation capabilities, while embedding techniques have enabled more nuanced semantic search and retrieval. A key innovation emerging from these advances is Retrieval-Augmented Generation (RAG), which combines the strengths of both technologies. RAG systems enhance LLMs' capabilities by grounding their responses in retrieved relevant information, offering a promising approach for more accurate and verifiable information access. This technological evolution addresses critical challenges in modern information access, including processing vast amounts of data, working across different languages, and adapting to specialized domains. In this context, our focus on English and Italian languages is strategically motivated by several compelling factors: (i) Linguistic Diversity: Italian represents a morphologically rich Romance language with complex verbal systems and agreement patterns, providing an excellent test case for model robustness compared to English's relatively simpler morphological structure. (ii) Research Gap: While English dominates NLP research, Italian, despite being spoken by approximately 67 million people[1] worldwide and being a major European language, remains underrepresented in large-scale NLP evaluations. This creates an important opportunity to assess model generalization. (iii) Industrial Relevance: Italy's significant technological sector and growing AI industry make Italian language support crucial for practical applications. The country's diverse industrial domains, from manufacturing to healthcare, from finance to tourism, present unique challenges for domain-specific IR and QA systems. (iv) Cross-family Evaluation: The comparison between Germanic (English) and Romance (Italian) language families offers insights into the cross-linguistic transfer capabilities of modern language models.

The current state-of-the-art in IR and QA reflects rapid technological advancement, particularly in multilingual capabilities. Recent industry developments have introduced models like Mistral-Nemo and Gemma, which specifically target the performance gap between high-resource and lower-resource languages. This evolution is driven by growing market demands for efficient multilingual solutions that can serve diverse markets without language-specific models while ensuring factual accuracy through retrieval-augmented approaches. These industry needs have shaped three key technological trends. First, transformer-based architectures, particularly BERT and its variants, have revolutionized the field by capturing sophisticated semantic relationships across languages. Second, dense retrieval methods have emerged as superior alternatives to traditional term-based approaches, significantly improving IR task performance. Third, the integration of LLMs with Retrieval-Augmented Generation (RAG) has enhanced QA systems by combining neural information retrieval with context-aware text generation, enabling more accurate and nuanced responses through external knowledge integration. However, despite these advances, significant challenges persist. The effectiveness of these models varies considerably across languages and domains, with performance patterns not yet fully understood. Critical questions remain about the trade-offs between model size, computational efficiency, and multilingual performance. Furthermore, ethical considerations, particularly regarding bias and fairness in cross-lingual information access, require deeper investigation.

1.1. Research Questions

Building on the current state-of-the-art and identified challenges, our study investigates four fundamental questions at the intersection of IR, QA, and language technologies:

1. **Embedding Effectiveness:** How do state-of-the-art embedding techniques perform across English and Italian IR tasks, and what factors influence their cross-lingual effectiveness?
2. **LLM Impact:** What are the quantitative and qualitative effects of integrating LLMs into RAG pipelines for multilingual QA tasks, particularly regarding answer accuracy and factuality?
3. **Cross-domain and Cross-language Generalization:** To what extent do current models maintain performance across domains and languages in zero-shot scenarios, and what patterns emerge in their generalization capabilities?
4. **Evaluation Methodology:** How can we effectively assess multilingual IR and QA systems, and what complementary insights do traditional and LLM-based metrics provide?

1.2. Contributions

Our research makes five significant contributions to the field of IR and QA:

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

1. **Comprehensive Performance Analysis:** A systematic evaluation of embedding techniques and LLMs across multiple IR and QA tasks, revealing key patterns in cross-lingual and cross-domain effectiveness. Our analysis encompasses 12 embedding models and 4 LLMs.
2. **Cross-lingual Insights:** An in-depth investigation of English-Italian language pair dynamics, offering valuable insights into the challenges and opportunities in bridging high-resource and lower-resource European languages.
3. **Evaluation Framework:** Development and application of a comprehensive evaluation methodology that combines traditional IR metrics with LLM-based assessments, enabling a more nuanced understanding of model performance across languages and domains.
4. **RAG Pipeline Insights:** We offer detailed insights into the effectiveness of integrating LLMs into RAG pipelines for QA tasks, highlighting both the potential and limitations of this approach.
5. **Practical Implications:** Our findings provide valuable guidance for practitioners in selecting appropriate models and techniques for specific IR and QA applications, considering factors such as language, domain, and computational resources.

These contributions advance both theoretical understanding and practical implementation of multilingual IR and QA systems. Our findings have direct applications in developing more effective search engines, cross-lingual information systems, and domain-specific QA tools, while also identifying promising directions for future research in multilingual language technologies.

1.3. Paper Organization

The remainder of this paper is organized as follows: Section 2 provides a comprehensive review of related work. Section 3 details our methodology, including the datasets, models, and evaluation metrics used. Section 4 presents our experimental results and analysis. Section 5 discusses the implications of our findings and their broader impact on the field. Finally, Section 6 concludes the paper and outlines directions for future research.

2. Related Work

Recent advances in Information Retrieval (IR) and Question Answering (QA) have been driven by two major technological shifts: the emergence of sophisticated embedding techniques and the development of large language models (LLMs). To systematically analyze these developments and position our research, we structure our review around five interconnected themes:

1. Evolution of IR and QA Systems - A Survey Landscape: Recent surveys and benchmark frameworks that have shaped our understanding of modern IR and QA systems.
2. Embedding Models for Information Retrieval: Embedding models specifically designed for IR tasks.
3. LLM Integration in Question Answering: The transformation of QA systems through large language models.
4. RAG Architecture: The development of retrieval-augmented generation (RAG) systems.
5. Evaluation Methodologies: The assessment metrics and methodologies for modern IR and QA systems.

This structure allows us to systematically examine the current state-of-the-art, identify existing gaps, and position our research within the broader landscape of IR and QA advancements.

2.1. Evolution of IR and QA Systems - A Survey Landscape

The rapid evolution of IR and QA technologies has spawned comprehensive surveys and benchmark frameworks addressing three critical aspects: system architectures, interpretability, and performance benchmarking.

From an architectural perspective, Hambarde and Proença [1] provide a systematic categorization of IR approaches, tracing the progression from traditional statistical methods approaches to modern deep learning methods, through discrete, dense, and hybrid retrieval techniques.

About explainable IR, Anand et al. [2] explored various approaches to make IR systems more interpretable, introducing the concept of Explainable Information Retrieval (ExIR). They identify three fundamental approaches to interpretability: (i) Post-hoc interpretability: Techniques for explaining trained model decisions. (ii) Interpretability by design: Architectures with inherent explanatory capabilities. (iii) IR principle grounding: Methods verifying adherence to established IR fundamentals.

Performance evaluation has been significantly advanced through several benchmark frameworks. Thakur et al. [3] introduced BEIR, a comprehensive zero-shot evaluation framework spanning 18 diverse domains, establishing new standards for assessing model generalization. Building on this foundation, Muennighoff et al. [4] developed MTEB, expanding evaluation to eight distinct embedding tasks across multiple languages and providing a valuable performance leaderboard[2].

Recent specialized frameworks have addressed emerging challenges in modern IR and QA systems. Tang et al. [5] focus on evaluating document-level retrieval and reasoning in RAG pipelines, while Zhang et al. [6] examine adaptive retrieval for open-domain QA. Gao et al. [7] contribute valuable insights into LLM-based evaluation methodologies, particularly exploring human-LLM collaboration in assessment.

While these frameworks have advanced our understanding of IR and QA systems, they leave a critical gap in comprehensive multilingual evaluation. Our study addresses this limitation by offering an in-depth analysis spanning English and Italian, building upon and extending frameworks like BEIR and MTEB. This approach enables a nuanced assessment of how embedding techniques and LLMs perform across linguistic boundaries and diverse domains, providing crucial insights for developing more effective multilingual IR and QA systems, with a focus on English and Italian.

2.2. Embedding Models for Information Retrieval

The landscape of information retrieval has undergone a fundamental transformation, shifting from traditional term-based methods to sophisticated neural approaches. At the core of this evolution are dense retrieval methods, which represent both documents and queries as dense vectors in a shared semantic space. Unlike traditional term-frequency approaches, these methods excel at capturing complex semantic relationships, enabling more nuanced retrieval for sophisticated queries.

Notable examples of dense retrieval methods include DPR [8], ColBERT [9], and ANCE [10]. The development of powerful embedding models has been crucial in advancing these capabilities, with models like BERT [11] and its variants being widely adopted and adapted for IR tasks. Some research has focused on enhancing retrieval effectiveness through specialized architectures and training approaches. Nogueira et al. [12] introduced doc2query, a method that expands documents with predicted queries, enhancing retrieval performance even with traditional methods like BM25. More recent works have focused on creating specialized embedding models for IR. Gao and Callan [13] proposed Condenser, a pre-training architecture designed specifically for dense retrieval. Wang et al. [14] introduced E5, a family of text embedding models trained on a diverse range of tasks and languages.

The challenge of multilingual information retrieval has sparked significant innovations. Xiao et al. [15] introduced BGE (BAAI General Embeddings)[3], demonstrating robust performance across multiple languages and retrieval tasks. These models leverage RetroMAE [16,17] pre-training on large-scale paired data through contrastive learning. Complementing these multilingual approaches, language-specific models like BERTino [18], an Italian DistilBERT variant, have emerged to address unique linguistic characteristics.

While significant advancements have been made in developing embedding models for IR, three critical gaps remain in the field: (1) comprehensive cross-lingual evaluation, particularly for morphologically rich languages like Italian, (2) systematic assessment of domain adaptation capabilities, and (3) comparative analysis of language-specific versus multilingual models. Our study addresses these gaps by providing a rigorous evaluation framework across linguistic and domain-specific boundaries.

[2] MTEB Leaderboard https://huggingface.co/spaces/mteb/leaderboard
[3] https://github.com/FlagOpen/FlagEmbedding

It assesses state-of-the-art embedding models in English and Italian contexts, specifically designed for IR tasks, offering insights into the adaptability and robustness of contemporary embedding models.

2.3. LLM Integration in Question Answering

The integration of large language models (LLMs) has fundamentally transformed question-answering systems, moving beyond traditional information extraction methods to enable sophisticated contextual understanding, multi-step reasoning, and the generation of natural, human-like responses.

A pivotal development in this evolution was the introduction of Retrieval-Augmented Generation (RAG) by Lewis et al. [19]. By combining neural retrievers with generation models, RAG established a powerful framework for knowledge-intensive NLP tasks, enabling QA systems to dynamically access and integrate external knowledge. This approach has proven particularly valuable in practical applications where both accuracy and contextual understanding are essential, allowing models to access external knowledge dynamically.

Brown et al. [20] transformed the landscape of question answering and demonstrated GPT-3's remarkable few-shot learning capabilities across various NLP tasks. This breakthrough revealed that large-scale language models could achieve sophisticated reasoning and response generation with minimal task-specific training, establishing new benchmarks for what was possible in automated question answering.

However, recent research has identified important challenges in LLM applications. Liu et al. [21] revealed a significant limitation dubbed the "Lost in the Middle" problem, where models struggle to maintain attention across long input contexts. This finding has crucial implications for QA systems that must process extensive documents or integrate information from multiple sources, highlighting the need for careful system design and implementation strategies.

The effectiveness of LLMs across different languages and specialized domains remains an active area of investigation. Our study addresses this critical research gap by evaluating state-of-the-art models in multilingual QA scenarios, including GPT4o, Llama 3.1, Mistral-Nemo, and Gemma2. Our findings contribute to a deeper understanding of how these powerful models can be effectively deployed in real-world, multilingual QA applications while acknowledging and addressing their current constraints.

2.4. RAG Architecture

Retrieval-augmented generation (RAG) has emerged as a pivotal architecture in modern information systems, with diverse implementations addressing different aspects of knowledge integration and generation [22–26]. Modern RAG architectures incorporate several critical components that work in concert. These include advanced document-splitting mechanisms that preserve semantic coherence, intelligent chunking strategies that optimize information density, and sophisticated retrieval mechanisms that leverage state-of-the-art embedding models. The integration of these components with powerful language generation models has created systems capable of producing more accurate and contextually appropriate responses.

Our research contributes to the current understanding of RAG systems through a systematic evaluation across multiple dimensions. We assess various RAG configurations in monolingual and cross-lingual settings, particularly in English and Italian. This comprehensive evaluation is motivated by two critical factors in modern AI system development: First, cross-lingual knowledge transfer capabilities are essential for developing truly multilingual AI systems. Second, domain adaptation flexibility is crucial for real-world deployments.

Our approach distinguishes itself from existing implementations through three key innovations: (1) systematic assessment of cross-lingual performance with focused attention on English-Italian language pairs, (2) comprehensive evaluation of domain adaptability across various sectors, and (3) integration of cutting-edge LLMs within RAG pipelines. This multifaceted evaluation provides valuable insights for both researchers and practitioners working to develop more robust and versatile information systems.

2.5. Evaluation Methodologies

The evolution of evaluation methodologies for IR and QA systems reflects the increasing sophistication of neural models and LLMs, necessitating a multi-faceted approach to performance assessment. This evolution spans traditional metrics, semantic evaluation approaches, and emerging LLM-based frameworks.

While traditional metrics [27] such as precision, recall, and F1 score continue to be relevant, they are often insufficient for capturing the nuanced performance of modern systems, particularly in assessing the quality and relevance of generated responses. For QA tasks, metrics like BLEU [28] and ROUGE [29] have been widely used to evaluate the quality of generated answers.

A significant advancement in evaluation methodology came with the introduction of BERTScore by Zhang et al. [30]. This approach leveraged contextual embeddings to capture semantic similarities, marking a shift toward more sophisticated evaluation techniques that better align with human judgments of text quality. This development proved particularly valuable for assessing systems that generate diverse yet semantically correct responses.

In recent years, specialized evaluation frameworks have emerged that are designed specifically for modern IR and QA architectures. Es et al. [31] introduced RAGAS, a framework for assessing retrieval-augmented generation systems. RAGAS innovatively addresses the unique challenges of evaluating LLM-based QA systems by incorporating multiple dimensions of assessment, including answer relevance and contextual alignment with retrieved information. Two significant contributions have further enriched the evaluation landscape. Katranidis et al. [32] developed FAAF, an approach to the fact verification task that leverages the function-calling capabilities of LMs. Additionally, Saad-Falcon et al. [33] introduced ARES, an automated evaluation system that assesses RAG systems across three critical dimensions: context relevance, answer faithfulness, and answer relevance.

Our research synthesizes these various evaluation approaches, combining traditional metrics with contemporary LLM-based assessment techniques to provide a comprehensive evaluation framework. This integrated approach enables us to assess multiple aspects of system performance, including: (i) Cross-lingual effectiveness across language boundaries. (ii) Adaptation capabilities across diverse domains. (iii) Quality and relevance of generated responses. (iv) Retrieval precision and efficiency metrics.

Through this holistic evaluation methodology, we offer insights into both the technical performance and practical applicability of modern IR and QA systems, contributing to a more nuanced understanding of their capabilities and limitations.

2.6. Research Gaps and Our Contributions

While significant progress has been made in IR and QA technologies, our comprehensive literature review reveals several critical gaps that currently limit the effectiveness of multilingual information systems. Most notably, there remains a significant lack of comprehensive studies evaluating system performance across linguistically diverse languages, particularly for morphologically rich languages like Italian. This gap is especially critical as the global deployment of these systems increases, yet our understanding of their behavior across different linguistic contexts remains limited. The challenge of domain adaptation presents another crucial area requiring investigation. While effective in general contexts, current systems often struggle to maintain consistent performance when dealing with specialized domains. This limitation becomes particularly evident in professional sectors such as healthcare, legal, and technical fields, where domain-specific terminology and reasoning patterns demand sophisticated adaptation mechanisms. The integration of external knowledge with LLM capabilities adds another layer of complexity, especially in multilingual settings, where RAG systems face significant challenges in maintaining consistency and accuracy across language barriers. Furthermore, current evaluation methodologies often fail to capture the full complexity of modern IR and QA systems. Traditional metrics, while valuable, may not adequately reflect real-world utility and reliability across different languages and use cases. This limitation is compounded by ethical considerations regarding the deployment of LLM-based systems across different languages and cultures, raising important questions about bias, fairness, and representation that require systematic investigation.

Our research addresses these challenges through two major contributions: First, we provide a comprehensive evaluation framework spanning both English and Italian, offering insights into model performance across linguistic boundaries. Second, we develop an evaluation methodology that combines traditional metrics with LLM-based assessment techniques.

Our research establishes a foundation for future developments of more effective, adaptable, and equitable multilingual IR and QA systems through these contributions. The insights and methodologies we present contribute to the ongoing effort to create more robust and inclusive language technologies that can effectively serve diverse linguistic communities and specialized domains.

3. Methodology and Evaluation Framework

This section presents our methodology for evaluating embedding techniques and large language models (LLMs) in Information Retrieval (IR) and Question Answering (QA) tasks. We describe the frameworks, datasets, models, and evaluation metrics employed in our study, as well as the rationale behind our choices and the potential limitations of our approach.

3.1. Overview of Approach

Our study encompasses a comprehensive evaluation of state-of-the-art embedding techniques and LLMs for enhancing IR and QA tasks, with a focus on English and Italian languages. The key components of our methodology include:

1. A diverse set of datasets across different domains and languages
2. A Retrieval-Augmented Generation (RAG) pipeline for QA tasks
3. A range of embedding models and LLMs
4. A comprehensive evaluation framework combining traditional and LLM-based metrics

3.2. RAG Pipeline

Our RAG pipeline consists of four main phases: ingestion, retrieval, generation, and evaluation, as illustrated in Figure 1. Each phase serves a specific function in the pipeline.

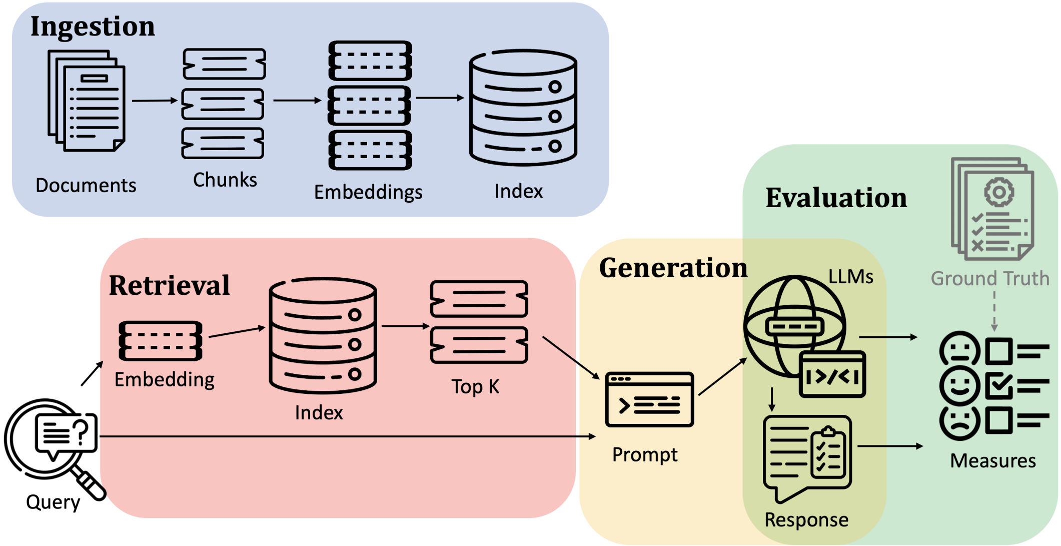

Figure 1. Architecture of the implemented RAG system showing the four main components: ingestion, retrieval, generation, and evaluation.

Ingestion. The initial phase processes input documents to create manageable and searchable chunks. This is achieved by segmenting the documents into smaller parts, referred to as "chunks". Different chunking strategies can be implemented, and for visual-oriented input documents like

PDFs, we exploit Document Layout Analysis to recognize more significant splitting of the document. These chunks are then embedded. The embedding step transforms the textual information into high-dimensional vectors that capture the semantic essence of each chunk. Following the embedding, these vector representations are ingested into a vector store such as Pinecone [4], Weaviate[5], and *Milvus*[6]. These vector databases are designed for efficient similarity search operations. The embedding and indexing process is critical for facilitating rapid and accurate retrieval of information relevant to user queries.

Retrieval. Upon receiving a query, the system employs the same embedding model to convert the query into its vector form. This query vector undergoes a similarity search within the vector store to identify the k most similar embeddings corresponding to previously indexed chunks. The similarity search leverages the vector space to find chunks whose content is most relevant to the query, thereby ensuring that the information retrieved is pertinent and comprehensive. This step is pivotal in narrowing down the vast amount of available information to the most relevant chunks for answer generation.

Generation. In this phase, a large language model (LLM) processes the query enriched with retrieved context to generate the final answer. The system first formats retrieved chunks into structured prompts, which are combined with the original query. The LLM synthesizes this information to construct a coherent and informative response, leveraging its ability to understand context and generate natural language answers.

Evaluation. The final phase of the system involves evaluating the quality of the generated answers. We employ both ground-truth dependent and independent metrics. Ground-truth-dependent metrics require a set of pre-defined correct answers against which the system's outputs are compared, allowing for the assessment of correctness. In contrast, ground-truth independent metrics evaluate the responses based on the answer's relevance to the question and are independent of a predefined answer set. This dual evaluation approach enables a comprehensive assessment of the system's performance, providing insights into both its correctness in relation to known answers and the overall quality of its generated text. In addition, the system can receive human evaluation of question-answer pairs as input and use it to evaluate metrics reliability and correspondence to expectations.

3.3. Datasets for Information Retrieval and Question Answering

We utilize a diverse set of datasets to evaluate models across different languages, domains, and task types. This diversity allows us to assess the models' generalization capabilities and domain adaptability. Table 1 provides an overview of the key datasets used in this study, and Figure 2 illustrates their distribution.

[4] Pinecone https://www.pinecone.io/

[5] Weaviate https://weaviate.io/

[6] Milvus https://milvus.io/

Table 1. Overview of datasets used for evaluating IR and QA performance. RC = Reading comprehension

Dataset	Task	Domain	Language	Used Samples	Retrieval unit granularity
SQuAD-en	IR and QA	Open	English	150 of 10.6k tuples (Dev-set = sample of SQuAD-it)	Entire paragraphs
SQuAD-it	IR and QA	Open	Italian	150 of 7.6k tuples (Test-set, random seed 433)	Entire paragraphs
DICE	IR - News Retrieval	Crime News	Italian	All 10.3k tuples	Single chunk with truncation
SciFact	IR - Fact checking	Scientific literature	English	5k test-set tuples (300 queries)	Single chunk with truncation
ArguAna	IR - Argument Retrieval	Misc. Arguments	English	1.4k queries (Test set, corpus of 8.6k docs)	Single chunk with truncation
NFCorpus	IR	Bio-Medical	English	323 queries (Test-set) (3.6k docs)	Single chunk with truncation
CovidQA	QA	Medical	English	All 124 tuples (27 questions, and 85 unique articles)	Chunks of 512 tokens of CORD19 documents
NarrativeQA	QA - RC	Narrative books and Movie scripts	English	100 queries (50 books, 50 movies)	Chunks of 512 tokens
NarrativeQA-translated	QA - RC	Narrative books and Movie scripts	Cross-lingual (En docs, It QA)	Same as NarrativeQA	Chunks of 512 tokens

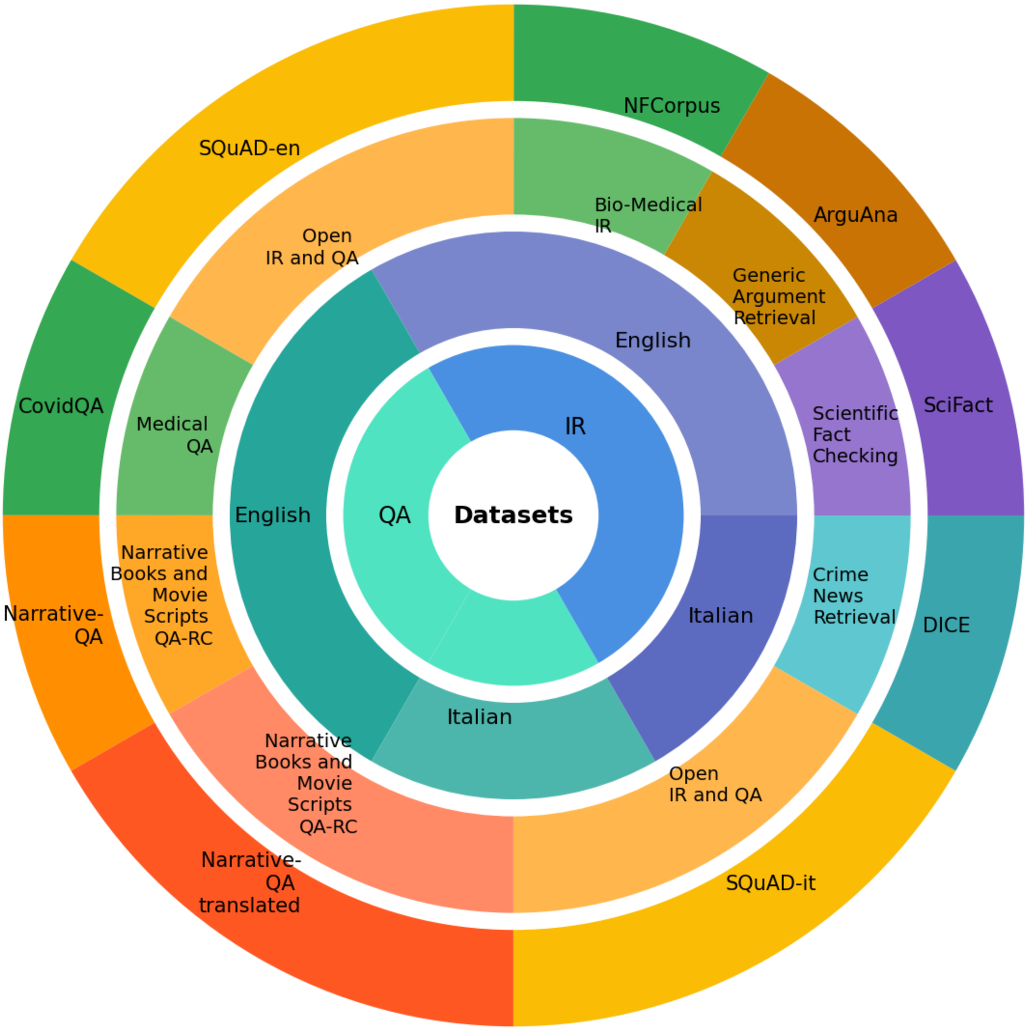

Figure 2. Distribution of datasets across different tasks and languages. The circular visualization shows the hierarchical organization of datasets used in the study, including IR and QA tasks in both English and Italian.

Below, we provide detailed descriptions of each dataset, including their specific characteristics and how they are used in our study:

SQuAD-en

SQuAD (Stanford Question Answering Dataset)[7] is a benchmark dataset focused on reading comprehension for Question Answering and Passage Retrieval tasks. The initial release, SQuAD 1.1 [34], comprises over 100K question-answer pairs about passages from 536 articles. These pairs were created through crowdsourcing, with each query linked to both its answer and the source passage. A subsequent release, SQuAD 2.0 [35], introduced an additional 50K unanswerable questions designed to evaluate systems' ability to identify when no answer exists in the given passage. SQuAD Open was developed for passage retrieval based on SQuAD 1.1 [36,37]. This variant uses the original crowdsourced questions but enables open-domain search across Wikipedia content dump. Each SQuAD entry contains four key elements:

(i) id: Unique entry identifier
(ii) title: Wikipedia article title
(iii) context: Source passage containing the answer
(iv) answers: Gold-standard answers with context position indices

Our study used SQuAD 1.1 for both IR and QA tasks, selecting 150 tuples from the validation set of 10.6k entries (1.5%) due to resource constraints. We ensured these selections matched the corresponding SQuAD-it samples to enable direct cross-lingual comparison. For IR, we processed the documents by splitting them into paragraphs and generating embeddings for each paragraph. We used the same splits for QA to evaluate our RAG pipeline's ability to generate answers.

SQuAD-it

The SQuAD 1.1 dataset has been translated into several languages, including Italian and Spanish. SQuAD-it[8] [38], the Italian version of SQuAD 1.1, contains over 60K question-answer pairs translated from the original English dataset. For our evaluation of both Italian IR and QA capabilities, we selected 150 tuples from the test set of 7.6k entries (1.9% of the test set), using random seed 433 for reproducibility and to work with limited resources. These samples directly correspond to the selected English SQuAD tuples, enabling parallel evaluation across languages. As with the English version, we processed the documents for IR by splitting them into paragraphs and generating embeddings for each segment, while using the same splits for QA evaluation.

DICE

Dataset of Italian Crime Event news (DICE)[9] [39] is a specialized corpus for Italian NLP tasks, containing 10.3k online crime news articles from Gazzetta di Modena. The dataset includes automatically annotated information for each article. Each entry contains the following key fields:

(i) id: Unique document identifier
(ii) url: Article URL
(iii) title: Article title
(iv) subtitle: Article subtitle
(v) publication date: Article publication date
(vi) event date: Date of the reported crime event
(vii) newspaper: Source newspaper name

We used DICE to evaluate IR performance in the specific domain of Italian crime news. In our experimental setting, we used the complete dataset (10.3k articles), with article titles serving as queries

[7] SQuAD Explorer https://github.com/rajpurkar/SQuAD-explorer https://rajpurkar.github.io/SQuAD-explorer/
[8] SQuAD-it https://github.com/crux82/squad-it http://sag.art.uniroma2.it/demo-software/squadit/
[9] DICE https://github.com/federicarollo/Italian-Crime-News

and their corresponding full texts as the retrieval corpus. The task involves retrieving the complete article text given its title, creating a one-to-one correspondence between queries and passages.

SciFact

SciFact[10] [40] is a dataset designed for scientific fact-checking, containing 1.4K expert-written scientific claims paired with evidence from research abstracts. In the retrieval task, claims serve as queries to find supporting evidence from scientific literature. The complete dataset contains 5,183 research abstracts, with multiple abstracts potentially supporting each claim. For our evaluation, we used the BEIR version[11]of the dataset, which preserves all passages from the original collection. We specifically used 300 queries from the original test set. Each corpus entry contains:

(i) id: Unique text identifier
(ii) title: Scientific article title
(iii) text: Article abstract

ArguAna

ArguAna[12] [41] is a dataset of argument-counterargument pairs collected from the online debate platform iDebate[13]. The corpus contains 8,674 passages, comprising 4,299 arguments and 4,375 counterarguments. The dataset is designed to evaluate retrieval systems' ability to find relevant counterarguments for given arguments. The evaluation set consists of 1,406 arguments serving as queries, each paired with a corresponding counterargument. The dataset is accessible through the BEIR datasets loader[14]. Each corpus entry contains:

(i) id: Unique argument identifier
(ii) title: Argument title
(iii) text: Argument content

NFCorpus

NFCorpus [42] is a dataset designed for evaluating the retrieval of scientific nutrition information from PubMed. The dataset comprises 3,244 natural language queries in non-technical English, collected from NutritionFacts.org[15]. These queries are paired with 169,756 automatically generated relevance judgments across 9,964 medical documents. For our evaluation, we used the BEIR version of the dataset, containing 3,633 passages and 323 queries selected from the original set. The dataset allows multiple relevant passages per query. Each corpus entry contains:

(i) id: Unique document identifier
(ii) title: Document title
(iii) text: Document content

CovidQA

CovidQA[16] [43] is a manually curated question answering dataset focused on COVID-19 research, built from Kaggle's COVID-19 Open Research Dataset Challenge (CORD-19)[17] [44]. While too small for training purposes, the dataset is valuable for evaluating models' zero-shot capabilities in the COVID-19 domain. The dataset contains 124 question-answer pairs referring to 27 questions across 85 unique research articles. Each query includes:

[10] Available in BEIR datasets: https://public.ukp.informatik.tu-darmstadt.de/thakur/BEIR/datasets/
[11] SciFact https://huggingface.co/datasets/BeIR/scifact
[12] ArguAna http://argumentation.bplaced.net/arguana/data
[13] Idebate FKA idebate.org https://idebate.net/
[14] BEIR datasets https://github.com/beir-cellar/beir/tree/main/beir/datasets
[15] NutritionFacts website https://nutritionfacts.org/
[16] CovidQA https://huggingface.co/datasets/castorini/covid_qa_castorini
[17] CORD-19 https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge

(i) category: Semantic category
(ii) subcategory: Specific subcategory
(iii) query: Keyword-based query
(iv) question: Natural language question form

Each answer entry contains:

(i) id: Answer identifier
(ii) title: Source document title
(iii) answer: Answer text

In our evaluation, we used the complete CovidQA dataset to assess domain-specific QA capabilities. For each query, which is associated with a set of potentially relevant paper titles, our system retrieves chunks of 512 tokens from the vector store and generates answers. Since multiple answers are generated for a query (one for each title), we compute the mean value of evaluation metrics per query. Due to slight variations in paper titles between CovidQA and CORD-19, we matched documents using Jaccard similarity with a 0.7 threshold.

NarrativeQA

NarrativeQA[18] [45] is an English dataset for question answering over long narrative texts, including books and movie scripts. The dataset spans diverse genres and styles, testing models' ability to comprehend and respond to complex queries about extended narratives. NarrativeQA training set contains 1102 documents divided into 548 books and 552 movie scripts, it also contains over 32k question-answer pairs. The test set contains 355 documents divided into 177 books and 178 movie scripts, it also contains over 10k question-answer pairs. Each entry contains:

(i) document: Source book or movie script
(ii) question: Query to be answered
(iii) answers: List of valid answers

For our evaluation, we used a balanced subsample of the test set (1%, 100 pairs total), consisting of 50 questions from books (covering 41 unique books) and 50 questions from movie scripts (covering 42 unique scripts). Using random seed 42 for reproducibility, this sampling strategy was chosen to manage OpenAI API costs while maintaining representation across both narrative types. We processed documents using 512-token chunks, retrieving relevant segments from the source document for each query.

NarrativeQA-cross-lingual

To evaluate cross-lingual capabilities, we created an Italian version of the NarrativeQA test set by maintaining the original English documents but translating the question-answer pairs into Italian. This approach allows us to assess how well LLMs can bridge the language gap between source documents and queries.

3.4. Models

3.4.1. Models Used for Information Retrieval

We evaluate a diverse set of embedding models, focusing on their performance in both English and Italian. All models were used with their default pretrained weights without additional fine-tuning. Table 2 provides an overview of these models.

[18] NarrativeQA https://huggingface.co/datasets/deepmind/narrativeqa

Table 2. Embedding Model Configurations

Model	Parameters	Max Input Length	Language
GTE-base[19] [46]	109M	512	English
GTE-large[20] [46]	335M	512	English
BGE-base-en-v1.5[21] [15]	109M	512	English
BGE-large-en-v1.5[22] [15]	335M	512	English
multilingual-E5-base[23] [14]	278M	512	Multilingual
multilingual-E5-large[24] [14]	560M	512	Multilingual
text-embedding-ada-002 (OpenAI)[25]	Not disclosed	8192	Multilingual
embed-multilingual-v2.0 (Cohere)[26] [27]	Not disclosed	256	Multilingual
embed-multilingual-v3.0 (Cohere)	Not disclosed	512	Multilingual
sentence-bert-base[28]	109M	512	Italian
BERTino[29] [18]	65M	512	Italian
BERTino v2[30]	65M	512	Italian

Rationale: This selection covers both language-specific and multilingual models, enabling us to assess cross-lingual performance and the effectiveness of specialized versus general-purpose embeddings.

3.4.2. Large Language Models for Question Answering

For QA tasks, we focus on retrieval-augmented generation (RAG) pipelines, integrating dense retrieval with LLMs for answer generation. For the retrieval component of our RAG pipeline, we selected the Cohere embed-multilingual-v3.0 model based on its superior performance in our IR experiments. This model achieved the highest consistent $nDCG@10$ scores across both English (0.90) and Italian (0.86) tasks, making it ideal for cross-lingual retrieval. We configured it to retrieve the top 10 passages for each query, balancing comprehensive context capture with computational efficiency. We tested different LLMs for answer generation and compared a widely used commercial API model with open-source alternatives. Table 3 provides an overview of the LLMs used in our study.

Table 3. Large Language Model Configurations. Because the chosen QA task requires a short answer, we set the response max tokens to 100.

Model	Company	API / Open-source	Param.	Context Win.	Language
GPT-4o[31]	OpenAI	API-based	>175B	128,000	Multilingual
Llama 3.1 8b[32] [47]	Meta	Open-source	8.03B	8,192[33]	Multilingual
Mistral-Nemo [34][35]	MistralAI	Open-source	12.2B	128,000	Multilingual
Gemma2b[36] [48,49]	Google	Open-source	2.51B	128,000	English

Rationale: This selection of LLMs represents a range of model sizes and architectures, allowing us to assess the impact of these factors on QA performance. The number of parameters for the OpenAI GPT-4o model is not disclosed but is likely greater than 175 billion.

3.5. Evaluation Metrics

We employ a set of evaluation metrics to assess both IR and QA performance. We focus on NDCG for IR tasks and a combination of reference-based (e.g., BERTScore, ROUGE) and reference-free metrics (e.g., Answer Relevance, Groundedness) for QA tasks. This diverse set of metrics allows for a multifaceted evaluation, capturing different aspects of model performance.

3.5.1. IR Evaluation Metric

For IR tasks, we primarily use the Normalized Discounted Cumulative Gain (NDCG) metric:

Normalized Discounted Cumulative Gain (NDCG@k) [27]:

- **Definition:** A ranking quality metric comparing rankings to an ideal order where relevant items are at the top.
- **Formula:** $NDCG@k = \frac{DCG@k}{IDCG@k}$ where $DCG@k$ is the Discounted Cumulative Gain at k, and $IDCG@k$ is the Ideal DCG at k, with k is a chosen cutoff point. DCG measures the total item relevance in a list with a discount that helps address the diminishing value of items further down the list.
- **Range:** 0 to 1, where 1 indicates a perfect match with the ideal order.
- **Use:** Primary metric for evaluating ranking quality, with k typically set to 10. $NDCG$ is used for experimental evaluation in different IR works such as [50] and [51].
- **Rationale:** NDCG is chosen as our sole IR metric because it effectively captures the quality of ranking, considering both the relevance and position of retrieved items. It's particularly useful for evaluating systems where the order of results matters, making it well-suited for assessing the performance of our embedding models in retrieval tasks.
- **Implementation:** Available in PyTorch, TensorFlow, and the BEIR framework.

3.5.2. QA Evaluation Metrics

For QA tasks, we employ both reference-based and reference-free metrics. Reference-based metrics use provided gold answers and may focus on either word overlap or semantic similarity. Reference-free metrics do not require gold answers, instead using LLMs to evaluate candidate answers along different dimensions.

Reference-based metrics:

- **BERTScore [30]:** Measures semantic similarity using contextual embeddings. BERTScore is a language generation evaluation metric based on pre-trained BERT contextual embeddings [11]. It computes the similarity of two sentences as a sum of cosine similarities between their tokens' embeddings. This metric can handle such cases where two sentences are semantically similar but differ in form. This evaluation method is used in many papers like [52] and [53]. This metric is often used in question-answering, Summarization, and translation. This metric can be implemented using different libraries, including TensorFlow and HuggingFace.
- **BEM** (BERT-based Evaluation Metric) [54]: Uses a fine-tuned BERT trained to assess answer equivalence. This model receives a question, a candidate answer, and a reference answer as input and returns a score quantifying the similarity between the candidate and the reference answers. This evaluation method is used in some recent papers like [55] and [56]. This metric can be implemented using TensorFlow. The model trained to perform the answer equivalence task is available on the TensorFlow hub.
- **ROUGE [29]:** Evaluates n-gram overlap between generated and reference answers. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) evaluates the overlap of n-grams between generated and reference answers. More in detail, it is a set of different metrics (ROUGE-1, ROUGE-2, ROUGE-L) used to evaluate text summarization and machine comprehension systems:
  1. ROUGE-N: Is defined as a n-gram recall between a predicted text and a ground truth text: ROUGE-N $= \frac{\sum_{S \in examples} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S' \in examples} \sum_{gram'_n \in S'} Count(gram'_n)}$ where $Count_{match}(gram_n)$ is the maximum number of n-grams of size n co-occurring in a candidate text and the ground truth text. The denominator is the total sum of the number of n-grams occurring in the ground truth text.
  2. ROUGE-L: Calculates an F-measure using the Longest Common Subsequence (LCS); the idea is that the longer the LCS of two texts is, the more similar the two summaries are. Given two texts, the ground truth X of length m and the prediction Y of length n, the formal definition is: ROUGE-L $= \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}}$ where:
     $R_{lcs} = \frac{LCS(X,Y)}{m}$ and $P_{lcs} = \frac{LCS(X,Y)}{n}$.

ROUGE metrics are very popular in Natural Language Processing specific tasks involving text generation like Summarization and Question Answering [57]. The advantage of ROUGE is that it allows us to estimate the quality of a generative model's output in common NLP tasks without dependencies from language. The disadvantages are:

1. It doesn't consider words semantic.
2. It's sensitive to the words choice and to the structure of the sentences.

Rouge metrics are implemented in PyTorch, TensorFlow, and Huggingface.

- **F1 Score:** Harmonic mean of precision and recall of word overlap. The F1 score is defined as the harmonic mean of precision and recall of word overlap between generated and reference answers. $F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$. This score summarizes the information on both aspects of a classification problem, focusing on precision and recall. F1 score is a very popular metric to evaluate performances of Artificial Intelligence and Machine Learning systems on classification tasks [58]. In question answering two popular benchmark datasets that use F1 as one of the metrics for evaluation are SQuAD [34] and TriviaQA [59]. The advantages of the F1 score are:

  1. It can handle unbalanced classes well.
  2. It captures and summarizes in a single metric, both the aspects of Precision and Recall.

  The main disadvantage is that, if left alone, the F1 score can be harder to interpret. The F1 score could be used in both Information Extraction and Question Answering settings. F1 score is implemented in all the popular libraries of Machine/Deep Learning and Data Analysis, such as Scikit-learn, PyTorch, and TensorFlow.

Reference-free metrics:

Then, we analyze reference-free LLM-based metrics: Context Relevance, Groundedness, and Answer Relevance. All of these are implemented using TruLens, which calls LLM GPT-3.5-turbo. These metrics are implemented using standard libraries and custom scripts, ensuring a comprehensive evaluation of our models across various IR and QA performance aspects. The combination of traditional IR metrics, reference-based QA metrics, and novel reference-free metrics provides a holistic view of model capabilities, allowing for nuanced comparisons across different approaches and datasets. LLM-based metrics are recent, mentioned in a few recent papers like [60]. Retrieval Augmented Generation Assessment (RAGAs) is an evaluation framework introduced in [31] which uses Large Language Models to test RAG pipelines. These metrics are implemented into ARES [37][33] and into a library called TruLens. We used TruLens with GPT-3.5-turbo in this paper.

- **Context Relevance:** Evaluates retrieved-context relevance to the question. It assesses if the passage returned is relevant for answering the given query. Therefore, this measure is useful for evaluating IR, after obtaining the answer.
- **Groundedness or Faithfulness:** Assesses the degree to which the generated answer is supported by retrieved documents obtained in a RAG pipeline. Therefore it measures if the generated answer is faithful to the retrieved passage or if it contains hallucinated or extrapolated statements beyond the passage.
- **Answer Relevance:** Measures the relevance of the generated answer to the query and retrieved passage.

Metric Classification:

We can classify all the previous metrics into two categories based on their capabilities to evaluate the answer, to exploit pure syntactic or also semantic aspects:

- **Syntactic metrics** evaluate formal response aspects, including BLEU [61], ROUGE [29], Precision, Recall, F1, and Exact Match [34]. These focus on text properties rather than semantic meaning.

[37] ARES https://github.com/stanford-futuredata/ARES

These metrics are generally considered less indicative of the semantic value of the generated responses. This is due to their focus on the text's formal properties rather than its content or inherent meaning.

- **Semantic metrics** evaluate response meaning, including BERTScore [30] and BEM score [54]. The BEM score is preferred to BERTScore for its correlation with human evaluations as reported in the original study we refer to and because we empirically found that the BERTScore tends to take values in a very short subset of values in the $(0, 1)$ range. The LLM-based metrics also belong to this group.

Manual Evaluation:

We conduct manual evaluations using a 5-point Likert scale. This method is not so popular because it requires high costs in terms of both money and time. Indeed, a lot of work done by human experts is required. We use manual evaluation principally to verify the reliability of automated evaluation metrics. Three independent human annotators with domain expertise evaluated the generated answers. For each evaluation session, the annotators were presented with: the original question, the RAG system's generated answer, and the ground truth from the dataset or customer answers. Annotators used a 5-point Likert scale to assess the quality of the generated answer in relation to the posed question, considering relevance, accuracy, and coherence. The criteria for scoring were as follows:

1. **Very Poor**: The generated answer is totally incorrect or irrelevant to the question. This case indicates a failure of the system to comprehend the query or retrieve pertinent information.
2. **Poor**: The generated answer is predominantly incorrect but with glimpses of relevance suggesting some level of understanding or appropriate retrieval.
3. **Neither**: The generated answer mixes relevant and irrelevant information almost equally, showcasing the system's partial success in addressing the query.
4. **Good**: The generated answer is largely correct but includes minor inaccuracies or irrelevant details, demonstrating a strong understanding and response to the question.
5. **Very Good**: Reserved for completely correct and fully relevant answers, reflecting an ideal outcome where the system accurately understood and responded to the query.

The annotators conducted their assessments independently to ensure unbiased evaluations. Upon completion, the scores for each question-answer pair were collected and compared. In cases of discrepancy, a consensus discussion was initiated among the annotators to agree on the most accurate score. This consensus process allowed for mitigating individual bias and considering different perspectives in evaluating the quality of the generated answers. This manual evaluation process helps particularly in assessing the reliability and validity of our system's automated evaluation metrics.

Inter-metric Correlation:

We use **Spearman Rank Correlation** [62] to assess automated metrics' reliability against human evaluation. This non-parametric measure evaluates the statistical dependence between rankings of two variables through a monotonic function. Computed on ranked data, it enables ordinal and continuous variables analysis. The correlation coefficient (ρ) ranges from -1 to 1, where 1 indicates perfect positive correlation, 0 indicates no correlation, and -1 indicates perfect negative correlation.

3.6. Experimental Design

Our experimental methodology aims to comprehensively evaluate embedding models and LLMs across multiple dimensions of IR and QA tasks. We structure our investigation around two complementary areas: (i) Information Retrieval performance, evaluating embedding models across domains and languages, and (ii) Question-answering capabilities, assessing LLM performance in RAG pipelines.

In the IR domain, we first evaluate embedding model performance across different domains, using datasets that span from general knowledge (SQuAD) to specialized scientific and medical content

(SciFact, ArguAna, and NFCorpus). We complement this with cross-language evaluation using Italian datasets (SQuAD-it and DICE) to assess how language-specific and multilingual models perform in non-English contexts. Additionally, we analyze the impact of retrieval size by varying the number of retrieved documents ($k \in \{1, 5, 10, 20, 50, 100\}$), with particular attention to recall metrics.

For QA tasks, our evaluation encompasses several dimensions. We assess LLM performance using both reference-based metrics (ROUGE-L, F1, BERTScore, BEM) and reference-free metrics (Answer Relevance, Context Relevance, Groundedness). We specifically test system capabilities in general domains using SQuAD for both English and Italian languages, in specialized domains using CovidQA for medical knowledge and NarrativeQA for narrative understanding. The cross-lingual aspect is explored using NarrativeQA with English documents and Italian queries, allowing us to measure the effectiveness of language transfer in QA contexts.

Across both IR and QA domains, we examine the relationship between model size and performance to understand scaling effects. Additionally, we perform manual assessments of system outputs to validate automated metrics and understand real-world effectiveness.

To ensure systematic evaluation, we implement the experiments following a structured methodology:

1. Dataset preparation: We preprocess and embed each dataset using the relevant embedding model
2. IR evaluation: For retrieval tasks, we implement top-k document retrieval ($k = 10$) and evaluate using NDCG@10
3. QA pipeline: For question answering, we implement the complete RAG pipeline and generate answers using multiple LLMs
4. Metric application: We apply our comprehensive set of evaluation metrics, including both reference-based and reference-free measures
5. Validation: We conduct the manual evaluation on carefully selected result subsets and analyze correlation with automated metrics

3.6.1. Hardware and Software Specifications

We conducted our experiments using the Google Colab platform[38]. Our implementation uses Python with the following key components: (i) Langchain framework for RAG pipeline implementation, (ii) Milvus vector store for efficient similarity search, (iii) HuggingFace endpoints, OpenAI and Cohere APIs for embedding models, (iii) OpenAI and HuggingFace endpoints for large language models.

3.6.2. Procedure

For IR activities, we followed this procedure:

1. **Data Preparation**:
   (a) Indexed all documents in the corpus using each embedding model
   (b) For documents exceeding the maximum token limit, we considered single-chunk truncation following BEIR settings
2. **Query Processing**: Encoded each query using the corresponding embedding model
3. **Retrieval**:
   (a) Used Milvus for efficient similarity search
   (b) Retrieved top-k documents for each query ($k \in \{10, 20, 50, 100\}$), with extensive experiments reported for $k = 10$
4. **Evaluation**:
   (a) Computed nDCG@10, MAP@10, Recall@10, and Precision@10 for each model on each dataset, focusing on nDCG@10 as the primary metric

[38] Google Colab https://colab.google/

(b) Used existing relevance judgments where available; for datasets without explicit judgments (e.g., DICE), considered documents relevant if matching the ground truth

For QA tasks, we employed the following protocol:

1. **Data Preparation**:
   (a) Indexed documents using Cohere embed-multilingual-v3.0 (best-performing IR model based on nDCG@10)
   (b) Split documents into passages of 512 tokens without sliding windows, balancing semantic integrity with information relevance
2. **Query Processing**: Encoded each query using the corresponding embedding model
3. **Retrieval Stage**: Used Cohere embed-multilingual-v3.0 to retrieve top-10 passages
4. **Answer Generation**:
   (a) Constructed bilingual prompts combining questions and retrieved passages
   (b) Applied consistent prompt templates across all models and datasets
   (c) Generated answers using each LLM

   During generation, we employed the following prompt structure for both English and Italian tasks:

Table 4. Standardized prompts used for English and Italian QA tasks

You are a Question Answering system that is rewarded if the response is short, concise and straight to the point, use the following pieces of context to answer the question at the end. If the context doesn't provide the required information simply respond <no answer>.
Context: {retrieved_passages}
Question: {human_question}
Answer:

Sei un sistema in grado di rispondere a domande e che viene premiato se la risposta è breve, concisa e dritta al punto, utilizza i seguenti pezzi di contesto per rispondere alla domanda alla fine. Se il contesto non fornisce le informazioni richieste, rispondi semplicemente <nessuna risposta>.
Context: {retrieved_passages}
Question: {human_question}
Answer:

   This prompt structure provides explicit instructions and context to the language model while encouraging concise and truthful answers without fabrication.
5. **Evaluation**:
   (a) Computed reference-based metrics (BERTScore, BEM, ROUGE, BLEU, EM, F1) using generated answers and ground truth
   (b) Used GPT-3.5-turbo to compute reference-free metrics (Answer Relevance, Groundedness, Context Relevance) through prompted evaluation

3.7. Cross-lingual and Domain Adaptation Experiments

To assess cross-lingual and domain adaptation capabilities:

1. **Cross-lingual**:
   (a) Evaluated multilingual models (e.g., E5, BGE) on both English and Italian datasets without fine-tuning
   (b) Compared performance against monolingual models (BERTino for Italian)
2. **Domain Adaptation**:
   (a) Tested models trained on general domain data (e.g., SQuAD) on specialized datasets (e.g., SciFact, NFCorpus)
   (b) Analyzed performance changes when moving to domain-specific tasks

3.8. Reproducibility Measures

We implemented several measures to ensure experimental reproducibility:

- **Randomization Control**: Fixed random seeds for all processes requiring randomization
- **Data Management**:
  - Used standard dataset splits where available
  - Selected representative subsets for efficiency:
    - 150 tuples from SQuAD-en validation set (1.5% of dev set)
    - 150 tuples from SQuAD-it test set (1.9% of test set)
    - 100 tuples from NarrativeQA (1% of test set, balanced between books and movies scripts)
- **Model Configuration**:
  - Used default pretrained weights without fine-tuning
  - Standardized parameters (e.g., 512-token chunk size)
- **Implementation Environment**:
  - Google Colab platform
  - Python with Langchain framework
  - Milvus vector store
  - Standardized evaluation protocols and thresholds

All configurations, datasets (including NarrativeQA-translated), and detailed protocols are available in our public repository[39][40].

3.9. Ethical Considerations

In conducting our experiments, we prioritized responsible research practices by carefully paying attention to ethical guidelines. We ensured strict compliance with dataset licenses and usage agreements while maintaining complete transparency regarding our data sources and processing methods. This commitment to data rights and transparency forms the foundation of reproducible and ethical research.

For model deployment, we paid particular attention to the ethical use of API-based models like GPT-4o, adhering strictly to providers' usage policies and rate limits. We thoroughly documented model limitations and potential biases in outputs, ensuring transparency about system capabilities and constraints. This documentation serves both to support reproducibility and to help future researchers understand the boundaries of these systems.

3.10. Limitations and Potential Biases

While our methodology strives for comprehensive evaluation, several important limitations warrant careful consideration when interpreting our results.

Our dataset selection, though diverse, cannot fully capture the complexity of real-world IR and QA scenarios. Despite including both general and specialized datasets, our coverage represents only a fraction of potential use cases across languages and domains. While our datasets span general knowledge (SQuAD), scientific content (SciFact), and specialized domains (NFCorpus), they cannot encompass the full breadth of linguistic variations and domain-specific applications.

Model accessibility imposed significant constraints on our evaluation scope. Access limitations to proprietary models and computational resource constraints prevented exhaustive experimentation with larger models.

The current state of evaluation metrics presents another important limitation. Though we employed both traditional metrics and specialized metrics, these measurements may not capture all nuanced aspects of model performance. This limitation becomes particularly apparent in complex

[39] URL to be added

[40] ToDo

QA tasks requiring sophisticated context understanding and reasoning capabilities. The challenge of quantifying aspects like answer relevance and factual accuracy remains an active area of research.

Practical resource constraints necessitated trade-offs in our experimental design. These limitations influenced our choices regarding sample sizes and the number of evaluation runs, though we worked to maximize the utility of available resources. For instance, our use of selected subsets from larger datasets (1.5-1.9% of original data) represents a necessary compromise between comprehensive evaluation and computational feasibility.

The temporal nature of our findings presents a final important consideration. Given the rapid evolution of NLP technology, our results represent a snapshot of model capabilities at a specific point in time. Future developments may shift the relative performance characteristics we observed, particularly as new models and architectures emerge.

These limitations collectively affect the generalization of our results. To address these constraints, we have: (i) Maintained complete transparency in our experimental setup. (ii) Documented all assumptions and methodological choices. (iii) Employed diverse evaluation metrics where possible. (iv) Provided detailed documentation of our implementation choices

While our findings have limitations, this approach ensures that they provide valuable insights within their defined scope and contribute meaningfully to the field's understanding of multilingual IR and QA systems.

4. Results

Our investigation into the capabilities of embedding techniques and large language models (LLMs) reveals a complex landscape of performance patterns across Information Retrieval (IR) and Question Answering (QA) tasks. Through systematic evaluation, we uncovered several intriguing findings that challenge common assumptions about multilingual model performance.

4.1. Information Retrieval Performance

The effectiveness of embedding models proves to be highly nuanced, with performance varying significantly across languages and domains. Our zero-shot evaluation reveals that while models generally maintain strong performance across languages, the degree of success depends heavily on the specific task and domain context.

Perhaps most surprisingly, multilingual models demonstrate remarkable adaptability, often matching or exceeding the performance of language-specific alternatives. This finding challenges the conventional wisdom that specialized monolingual models are necessary for optimal performance in specific languages. The embed-multilingual-v3.0 model, in particular, achieves impressive results across both English and Italian tasks, suggesting that recent advances in multilingual architectures are closing the historical gap between language-specific and multilingual models.

The relationship between model size and performance emerged as particularly intriguing. Our analysis reveals that architectural design often proves more crucial than raw parameter count, as evidenced by the performance patterns of base and large model variants.

Table 5. nDCG@10 scores for English and Italian, and different domain datasets

Model	SQuAD-en	SQuAD-it	DICE	SciFact	ArguAna	NFCorpus
GTE-base	0.87	_	_	0.74	0.56	0.37
GTE-large	0.87	_	_	0.74	0.57	**0.38**
bge-base-en-v1.5	0.86	_	_	0.74	**0.64**	0.37
bge-large-en-v1.5	0.89	_	_	**0.75**	**0.64**	**0.38**
multilingual-e5-base	0.90	0.85	0.56	0.69	0.51	0.32
multilingual-e5-large	**0.91**	**0.86**	0.64	0.70	0.54	0.34
text-embedding-ada-002 (OpenAI)	0.86	0.79	0.54	0.71	0.55	0.37
embed-multilingual-v2.0 (Cohere)	0.84	0.79	0.64	0.66	0.55	0.32
embed-multilingual-v3.0 (Cohere)	0.90	**0.86**	**0.72**	0.70	0.55	0.36
sentence-bert-base	_	0.52	0.22	_	_	_
BERTino	_	0.57	0.33	_	_	_
BERTino v2	_	0.64	0.40	_	_	_

As shown in Table 5, multilingual models demonstrate remarkable consistency across tasks. The embed-multilingual-v3.0 model achieves particularly noteworthy results, maintaining strong performance not only in general tasks (nDCG@10 scores of 0.90 and 0.86 for English and Italian SQuAD respectively) but also in specialized domains like DICE (0.72). This robust cross-domain performance suggests that recent architectural advances are successfully addressing the historical challenges of multilingual modeling. Interestingly, the comparison between the base and large model variants suggests that architectural design choices may have more impact than model size alone, as larger models don't consistently outperform their smaller counterparts.

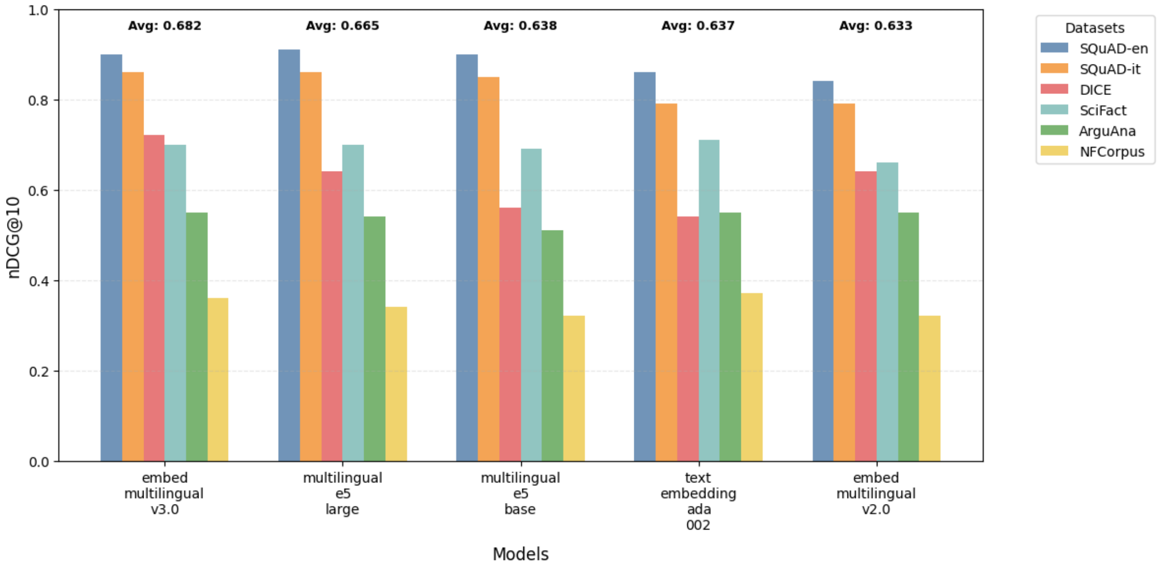

Figure 3. Comparison of nDCG@10 scores across different embedding models and datasets. The visualization shows the performance of multilingual models on various IR tasks.

Figure 3 provides a comprehensive visualization of how different models perform across tasks and domains, revealing several key patterns in information retrieval performance. The visualization demonstrates the general superiority of multilingual models, particularly evident in their consistently strong performance on SQuAD-type tasks. However, it also illustrates an important performance gradient: while models excel in general-domain tasks, their effectiveness tends to decrease when handling specialized domains. This performance drop in specialized areas suggests a crucial direction for future research and improvements in model development, especially for domain-specific applications.

4.1.1. Cross Domain Results

The conducted tests on English datasets compared state-of-the-art embedding models across SQuAD, SciFact, ArguAna, and NFCorpus datasets to evaluate cross-domain effectiveness. Table 5 presents the performance results measured by nDCG@10 across these diverse domains. Key observations from cross-domain evaluation:

- Performance varies significantly by domain, with no single model achieving universal superiority across all tasks.
- Multilingual-e5-large achieves the highest performance on general domain tasks, with nDCG@10 of 0.91 on SQuAD-en.
- BGE models demonstrate particular strength in specialized content, achieving top performance on ArguAna (0.64) and SciFact (0.75).
- GTE and BGE architectures show robust adaptability to scientific and medical domains, maintaining strong performance across SciFact and NFCorpus datasets.

4.1.2. Cross Language Results

We evaluated cross-lingual capabilities through benchmark tests comparing multilingual and Italian-specific models using two datasets: (i) The Italian translation of SQuAD for general domain assessment, and (ii) the DICE dataset (Italian Crime Event news) for domain-specific evaluation For DICE evaluation, we used news titles as queries to retrieve relevant corpus documents.

Table 5 presents comparative results between multilingual and Italian-specific models.

Key findings from cross-lingual analysis:

- Multilingual models consistently outperform Italian-specific models (e.g., BERTino) across both datasets.
- Multilingual-e5-large achieves top performance on SQuAD-it (nDCG@10: 0.86).
- Embed-multilingual-v3.0 demonstrates exceptional versatility, excelling in both SQuAD-it (0.86) and DICE (0.72).
- The performance gap between multilingual and monolingual models suggests superior domain adaptation capabilities in larger multilingual architectures.

4.1.3. Retrieval Size Impact

We systematically analyzed how retrieval size affects model performance, using multilingual-e5-large on the DICE dataset as our test case. Table 6 and Figure 4 present Recall@k scores across different retrieval sizes (k).

Table 6. Recall@k for e5-multilingual-large on DICE.

k	Recall@k
1	0.335
5	0.535
10	0.611
20	0.680
50	0.767
100	0.827

The data reveals three distinct performance phases: (i) Rapid growth (k=1 to k=20): Recall more than doubles from 0.335 to 0.680. (ii) Moderate improvement (k=20 to k=50): Recall increases by 0.087. (iii) Diminishing returns (k>50): Marginal improvements decrease significantly.

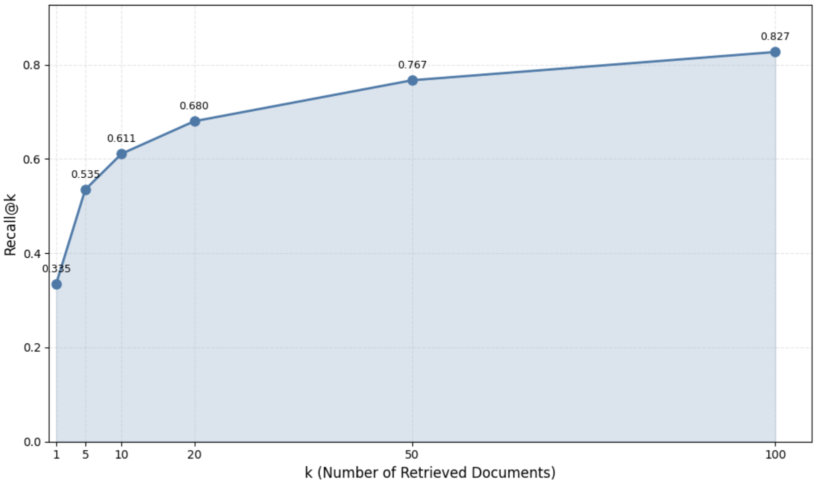

Figure 4. Evolution of Recall@k for increasing values of k on the DICE dataset. The plot demonstrates how retrieval performance improves with larger k values.

Figure 4 transforms this data into a clear visual pattern, revealing three distinct phases in recall improvement: a steep initial climb (k=1 to k=20), a moderate growth period (k=20 to k=50), and a plateau phase (beyond k=50). This characteristic logarithmic curve helps visualize the diminishing returns phenomenon in retrieval system performance. This non-linear relationship is particularly valuable for system designers, as it shows that while increasing retrieval size consistently improves performance, the marginal benefits diminish substantially after certain thresholds. Understanding this pattern helps practitioners make informed decisions about system configuration, balancing the desire for higher recall against computational costs and response time requirements.

This analysis has important practical implications for system design. While recall continues to improve up to k=100, where it reaches 80%, the diminishing returns pattern suggests that smaller retrieval sizes might be more practical. A promising approach would be to use a moderate initial retrieval size (around k=50) and then apply more sophisticated re-ranking techniques to the retrieved passages, balancing computational efficiency with retrieval effectiveness.

4.2. Question Answering Performance

4.2.1. Model Performance Across Tasks and Languages

We evaluated different LLMs within a Retrieval-Augmented Generation (RAG) pipeline, utilizing Cohere embed-multilingual-v3.0 for the retrieval phase based on its superior performance in embedding evaluations. Our analysis spans multiple datasets: SQuAD-en, SQuAD-it, CovidQA, NarrativeQA (books and movie scripts), and their translated versions (NarrativeQA-translated). We assessed performance through three complementary perspectives: syntactic accuracy, semantic similarity, and reference-free evaluation.

Tables 7, 8, and 9 present the performance of different LLMs on the various datasets considering syntactic, semantics, and LLM-based groud-truth free metrics respectively.

Table 7. (Syntactic) Results for English and Italian, and different domain datasets. ROUGE-L, F1 score

Model	SQuAD-en	SQuAD-it	CovidQA	NaQA-B	NaQA-M	NaQA-B-tran	NaQA-M-tran
GPT-4o	0.26 | 0.25	0.21 | 0.18	0.21 | 0.13	0.14 | 0.12	0.16 | 0.12	**0.13** | **0.13**	**0.13** | **0.13**
Llama 3.1 8b	**0.72** | **0.69**	**0.57** | **0.54**	0.22 | 0.15	0.12 | 0.11	0.13 | 0.11	0.09 | 0.09	0.09 | 0.09
Mistral-Nemo	0.43 | 0.41	0.27 | 0.25	**0.27** | **0.17**	**0.23** | **0.21**	**0.30** | **0.25**	0.10 | 0.06	0.05 | 0.04
Gemma2b	0.40 | 0.39	_	0.24 | 0.16	0.15 | 0.11	0.17 | 0.13	_	_

The syntactic evaluation results in Table 7 reveal notable performance variations across models and tasks. Llama 3.1 8b demonstrates superior performance on general question-answering tasks

(SQuAD-en: 0.72/0.69, SQuAD-it: 0.57/0.54), while Mistral-Nemo shows stronger capabilities in specialized domains (CovidQA: 0.27/0.17, NaQA-B: 0.23/0.21). Figure 5 visualizes the variation in syntactic metric performance across different models and datasets.

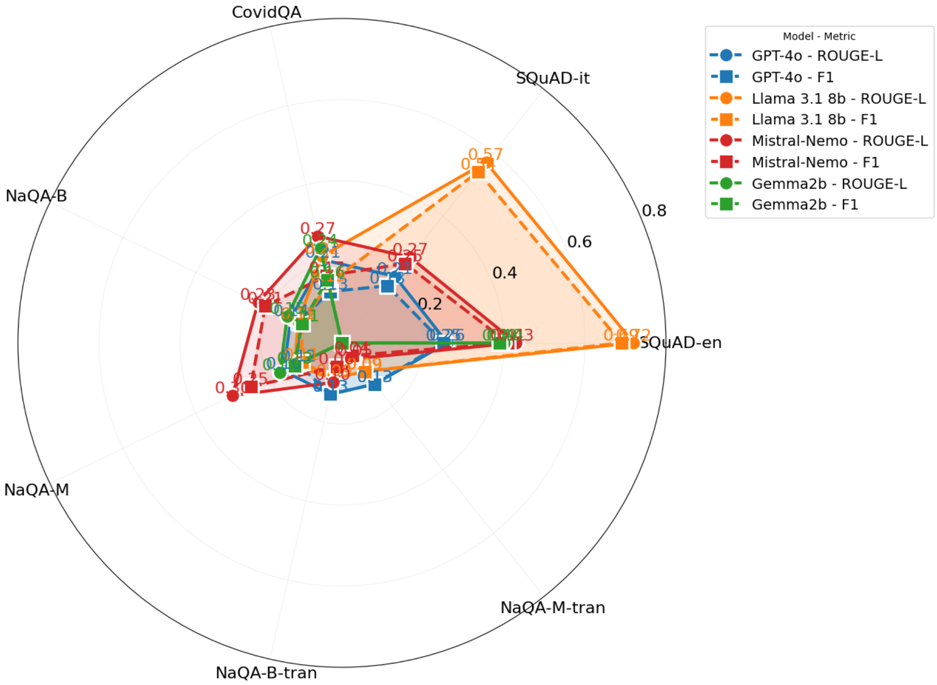

Figure 5. Syntactic metrics comparison across models and datasets. The radar chart visualizes ROUGE-L and F1 scores for each model, showing the performance patterns across different datasets.

Table 8. Results for English and Italian, and different domain datasets. BERTscore f1, BEM score

Model	SQuAD-en	SQuAD-it	CovidQA	NaQA-B	NaQA-M	NaQA-B-tran	NaQA-M-tran
GPT-4o	0.85 ∣ 0.93	0.81 ∣ **0.92**	0.85 ∣ 0.61	0.85 ∣ 0.50	0.85 ∣ 0.46	**0.83** ∣ **0.47**	**0.83** ∣ **0.45**
Llama 3.1 8b	**0.92** ∣ 0.90	**0.90** ∣ 0.79	0.85 ∣ 0.61	0.85 ∣ 0.45	0.85 ∣ 0.47	0.81 ∣ 0.44	0.82 ∣ 0.43
Mistral-Nemo	0.88 ∣ **0.94**	0.83 ∣ 0.82	**0.86** ∣ **0.62**	**0.87** ∣ **0.60**	**0.88** ∣ **0.51**	0.83 ∣ 0.25	0.82 ∣ 0.18
Gemma2b	0.88 ∣ 0.77	_	0.85 ∣ 0.43	0.85 ∣ 0.38	0.86 ∣ 0.32	_	_

Semantic evaluation results (Table 8) consistently show higher scores compared to syntactic metrics, particularly in BERTScore values. This pattern suggests models often generate semantically appropriate answers even when they deviate from reference answers lexically. Figure 6 visualizes the variation in semantic metric performance across different models and datasets.

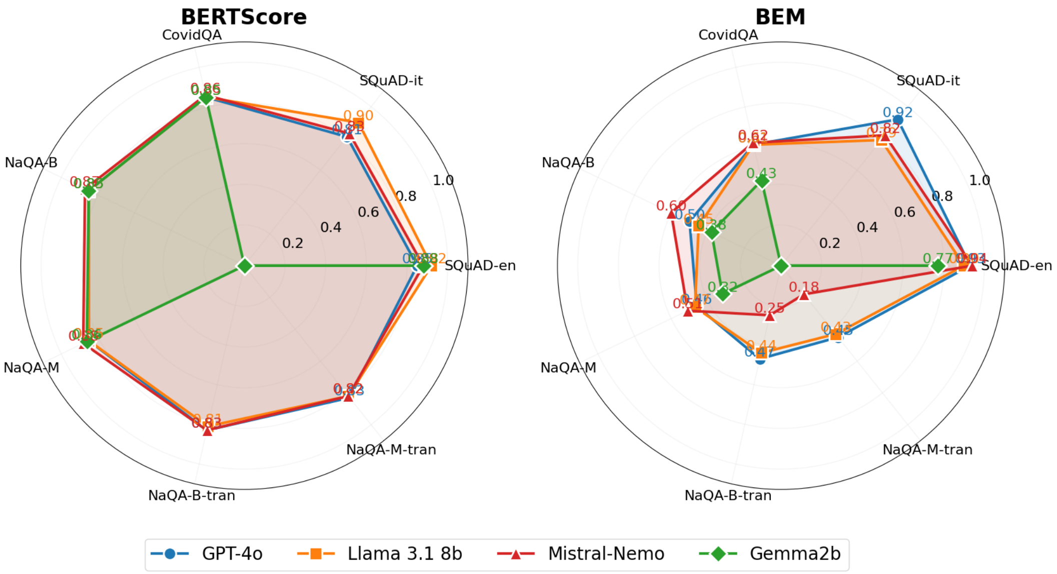

Figure 6. Semantic metrics comparison across models and datasets. The visualization shows BERTScore and BEM scores, providing insights into the semantic quality of model outputs.

Table 9. Ground-truth-free results for English and Italian, and different domain datasets. Answer relevace/Context relevance/Groundness

Model	SQuAD-en	SQuAD-it	CovidQA	NaQA-B	NaQA-M	NaQA-B-tran	NaQA-M-tran
GPT-4o	**1.0** | **0.90** | **0.79**	**0.99** | 0.80 | **0.81**	0.89 | 0.82 | 0.61	0.89 | 0.58 | **0.58**	0.91 | 0.59 | **0.39**	0.96 | 0.55 | 0.45	0.94 | 0.49 | 0.31
Llama 3.1 8b	1.0 | 0.89 | 0.67	0.99 | 0.80 | 0.71	0.86 | 0.82 | 0.62	0.95 | 0.58 | 0.53	0.95 | 0.59 | 0.33	0.93 | **0.56** | 0.40	0.91 | **0.50** | **0.33**
Mistral-Nemo	1.0 | 0.89 | 0.78	0.98 | **0.81** | 0.78	**0.91** | **0.82** | **0.64**	**1.0** | **0.59** | 0.52	**0.96** | **0.59** | 0.37	**0.99** | 0.55 | **0.47**	**0.94** | 0.49 | 0.30
Gemma2b	0.98 | 0.90 | 0.67	–	0.77 | 0.82 | 0.51	0.91 | 0.57| 0.56	0.87 | 0.59 | 0.31	–	–

Our reference-free evaluation (Table 9) reveals several key patterns: (i) Models consistently achieve higher scores in answer relevance compared to groundedness. (ii) GPT-4o excels in cross-lingual scenarios, particularly on translated narrative tasks. (iii) Mistral-Nemo demonstrates strong performance across domains while maintaining reasonable groundedness. (iv) Complex narratives pose greater challenges for maintaining factual accuracy. Figure 7 visualizes the variation in LLM-based metrics performance across different models and datasets.

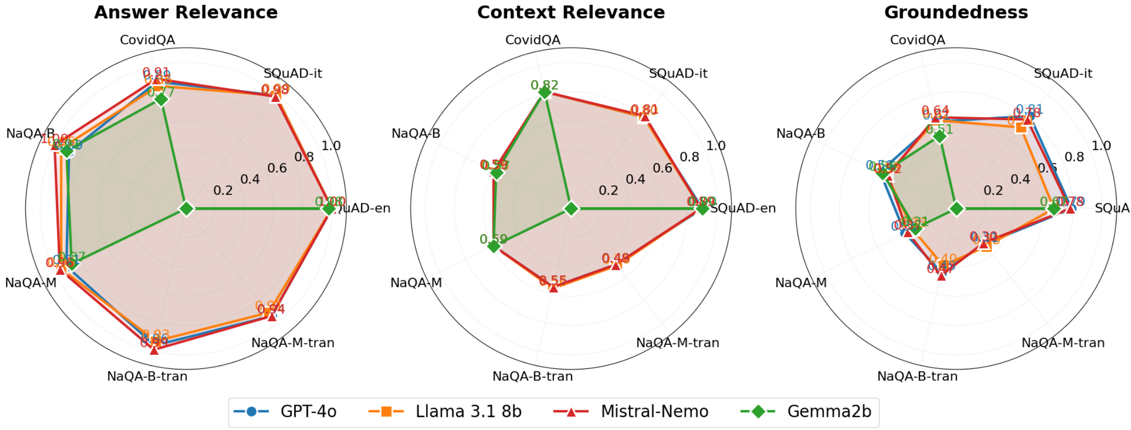

Figure 7. LLM-based metrics comparison across models and datasets. The radar charts display Answer Relevance, Context Relevance, and Groundedness scores for each model.

Key findings from our comprehensive evaluation include:

Table 10. Spearman correlations on NarrativeQA books and movies subsample

	NarrativeQA books			NarrativeQA movies		
Metrics	Human Judgement	BEM	AR TruLens	Human Judgement	BEM	AR TruLens
Human Judgement	1.000	**0.735**	0.436	1.000	**0.704**	0.565
BEM	**0.735**	1.000	0.185	**0.704**	1.000	0.522
AR TruLens gpt-3.5-turbo	**0.436**	0.185	1.000	**0.565**	0.522	1.000

1. Model Specialization: (i) Llama 3.1 8b excels in syntactic accuracy on general domain tasks. (ii) GPT-4o demonstrates superior cross-lingual capabilities. (iii) Mistral-Nemo achieves consistent performance across diverse tasks.
2. Performance Patterns: (i) BERTScores indicate strong semantic understanding across all models. (ii) Groundedness scores decrease in complex domains. (iii) Semantic metrics consistently outperform syntactic measures.
3. Domain Effects: (i) Factual domains (CovidQA) show higher groundedness scores. (ii) Narrative domains pose greater challenges for factual accuracy. (iii) Cross-lingual performance remains robust in structured tasks.

4.2.2. Metrics effectiveness versus human evaluation

Analysis of human judgment correlation on NarrativeQA samples reveals stronger alignment with BEM scores (correlation: 0.735 for books, 0.704 for movies) compared to reference-free metrics. This suggests that ground-truth-based evaluation remains more reliable for assessing answer quality, particularly in complex narrative contexts.

5. Discussion

Our comprehensive evaluation of embedding models and large language models reveals a complex landscape of capabilities and limitations in multilingual information retrieval and question-answering. The results show how these systems perform across different languages and domains, challenging some common assumptions while reinforcing others. The findings offer important insights for both theoretical understanding and practical applications while highlighting critical areas for future development.

5.1. The Domain Specialization Challenge

Our analysis reveals distinct patterns of domain specialization impact across both Information Retrieval and question-answering tasks.

Looking at IR performance (Table 5), we observe a clear degradation pattern as tasks become more specialized. The embed-multilingual-v3.0 model demonstrates this trend clearly in English tasks: achieving 0.90 nDCG@10 on general domain (SQuAD), dropping to 0.70 on scientific literature (SciFact), further declining to 0.55 on argument retrieval (ArguAna), and reaching its lowest performance of 0.36 on medical domain tasks (NFCorpus), see Figure 8. Similar patterns are observed across other models, with multilingual-E5-large showing comparable degradation: 0.91 (SQuAD), 0.70 (SciFact), 0.54 (ArguAna), and 0.34 (NFCorpus).

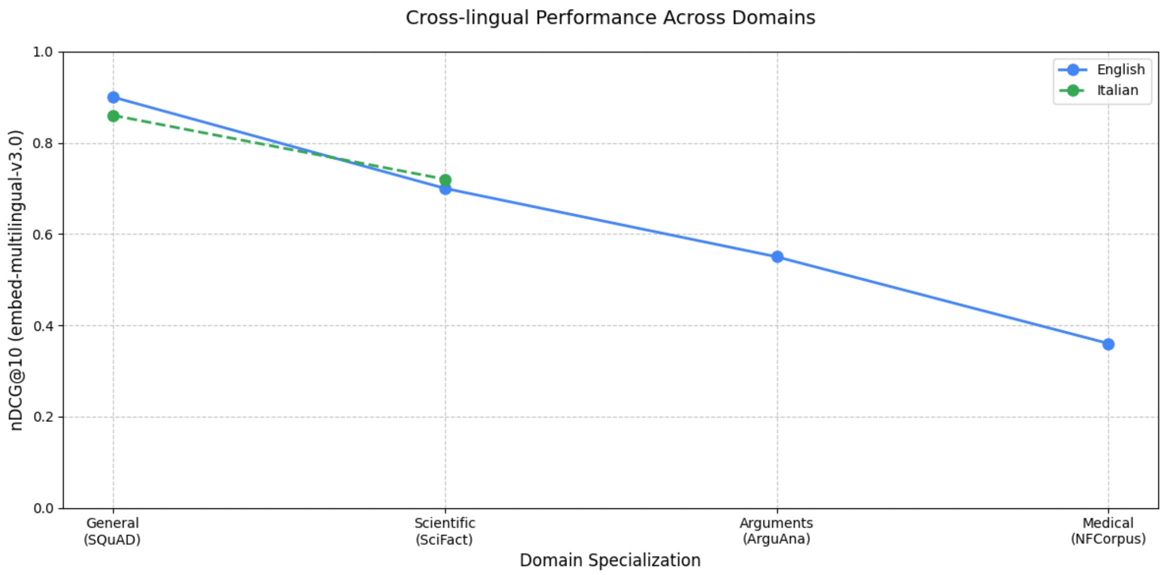

Figure 8. Cross-domain performance comparison showing nDCG@10 scores for embed-multilingual-v3.0 model across different domains. The plot demonstrates the performance degradation pattern from general to specialized domains in both English (solid line) and Italian (dashed line) tasks.

In Italian IR tasks, while we have fewer domain-specific datasets, the pattern persists as shown in Table 5 and illustrated in Figure 8. The embed-multilingual-v3.0 model achieves 0.86 nDCG@10 on the general domain (SQuAD-it) and 0.72 on the specialized news domain (DICE). Language-specific models like BERTino show more pronounced degradation, with performance dropping from 0.64 on SQuAD-it to 0.40 on DICE.

For Question Answering tasks, the domain specialization effect is evident across different evaluation metrics. Looking at syntactic metrics (Table 7), Llama 3.1 8b shows strong general domain performance (ROUGE-L: 0.72/0.69 on SQuAD-en) but drops significantly on specialized medical content (CovidQA: 0.22/0.15). Mistral-Nemo follows a similar pattern, declining from 0.43/0.41 on SQuAD-en to 0.27/0.17 on CovidQA.

Semantic metrics (Table 8) show more stability across domains but still reflect the specialization challenge. BERTScore results for GPT-4o decrease from 0.85 in the general domain to 0.85 in the medical domain, while BEM scores show a more pronounced drop from 0.93 to 0.61. This pattern is consistent across models, with Llama 3.1 8b showing similar degradation (BERTScore: 0.92 to 0.85, BEM: 0.90 to 0.61).

The reference-free metrics (Table 9) provide additional insight into domain adaptation challenges. While answer relevance remains relatively high across domains (ranging from 0.89 to 1.0), groundedness scores show significant degradation when moving from general to specialized domains. For instance, Mistral-Nemo's groundedness drops from 0.78 on SQuAD to 0.64 on CovidQA, while GPT-4o shows a decline from 0.79 to 0.61.

These results demonstrate a consistent pattern: while models perform well in general domains, their effectiveness decreases substantially as domain specificity increases, regardless of the evaluation metric or language used. This degradation is particularly pronounced in medical and technical domains, suggesting that current approaches face significant challenges in handling specialized knowledge. This gradient reveals fundamental challenges in domain adaptation that persist across all models, regardless of their size or architectural sophistication. This suggests that current pre-training approaches might not sufficiently capture domain-specific nuances across languages.

5.2. Cross-lingual Performance: A Tale of Two Languages

Our analysis reveals distinct cross-lingual performance patterns across both IR and QA tasks. In Information Retrieval, the results from Table 5 show that multilingual models achieve competitive performance across languages. The embed-multilingual-v3.0 model maintains strong performance with nDCG@10 scores of 0.90 for English and 0.86 for Italian on SQuAD tasks. Similar patterns are seen with multilingual-e5-large, achieving 0.91 and 0.86 for English and Italian respectively. In contrast, language-specific models like BERTino show limited performance (0.64 on SQuAD-it), suggesting that multilingual architectures have become more effective than language-specific approaches.

In Question Answering tasks, the cross-lingual performance shows more variation across different metrics. For syntactic measures (Table 7), we see larger gaps between languages: Llama 3.1 8b achieves ROUGE-L scores of 0.72/0.69 for English but drops to 0.57/0.54 for Italian, while Mistral-Nemo shows scores of 0.43/0.41 for English reducing to 0.27/0.25 for Italian. Semantic evaluation metrics (Table 8) reveal more stable cross-lingual performance. BERTScore results show closer parity between languages, with scores ranging from 0.85-0.92 for English and 0.81-0.90 for Italian across models. GPT-4o maintains relatively consistent performance (BERTScore: 0.85 English, 0.81 Italian), while Llama 3.1 8b achieves 0.92 for English and 0.90 for Italian. The ground-truth-free metrics (Table 9) provide additional insights into cross-lingual capabilities. Answer relevance remains high across languages (0.98-1.0 for both), but groundedness shows interesting variations. Mistral-Nemo achieves comparable groundedness scores in both languages (0.78 English, 0.78 Italian), while GPT-4o shows a slight variation (0.79 English, 0.81 Italian).

These patterns suggest that while modern architectures have made significant progress in bridging the cross-lingual gap, particularly in IR tasks and semantic understanding, challenges remain in maintaining consistent syntactic quality across languages in QA tasks. These findings have important implications for the deployment of multilingual IR and QA systems. While current models show promising cross-lingual capabilities in general domains, practitioners should carefully consider domain-specific requirements, particularly when working with non-English languages in specialized fields. Future research should focus on developing techniques to better preserve performance across both linguistic and domain boundaries, possibly through more effective pre-training strategies or domain adaptation methods.

5.3. The Architecture vs. Scale Debate

Our results prove that architectural efficiency and type of training matter more than raw parameter count challenging the common assumption that larger models necessarily perform better.

In IR tasks (Table 5), comparing architectures of different sizes reveals interesting patterns. When comparing the multilingual-E5 base (278M parameters) and large (560M parameters) variants, we find minimal performance differences of just 0.01-0.02 nDCG@10 points, GTE-base is similar to GTE-large indicating that model size alone does not guarantee superior performance.

In English QA tasks (Table 7), we observe varied performance patterns across different model sizes. The 8B parameter Llama 3.1 achieves the highest ROUGE-L scores (0.72/0.69) on SQuAD-en, outperforming both the larger GPT-4o (0.26/0.25) and the 12.2B parameter Mistral-Nemo (0.43/0.41). However, this advantage doesn't hold consistently across all tasks—on CovidQA, Llama 3.1 8b's performance (0.22/0.15) is comparable to that of other models. The semantic evaluation metrics (Table 8) show a different pattern. While Llama 3.1 8b maintains strong BERTScore performance (0.92/0.90), GPT-4o and Mistral-Nemo show competitive results (0.85/0.93 and 0.88/0.94 respectively) despite their architectural differences. Looking at ground-truth-free metrics (Table 9), we see consistent answer relevance scores across architectures (ranging from 0.98 to 1.0), regardless of model size. Groundedness scores show more variation, with Mistral-Nemo (0.78) and GPT-4o (0.79) performing similarly on SQuAD-en despite their different architectures.

This pattern holds true across different tasks and domains, suggesting that clever design might be more crucial than sheer size. The comparable or sometimes superior performance of smaller models than bigger ones in specific tasks indicates that efficient architectural design and training approaches can effectively compete with larger models.

5.4. Patterns in Model Evaluation Metrics

Our evaluation across different metrics and human judgments reveals distinct patterns in model performance assessment. The discrepancies observed between reference-based and reference-free metrics highlight the importance of using diverse evaluation approaches, especially for complex QA tasks where a single "correct" answer may not exist.

For QA tasks, syntactic metrics (Table 7) show relatively low scores, with ROUGE-L ranging from 0.26 to 0.72 for English SQuAD and 0.21 to 0.57 for Italian SQuAD. These scores decline further on specialized domains, with CovidQA showing ROUGE-L scores between 0.21 and 0.27. Semantic metrics (Table 8) consistently show higher scores across all models. BERTScore ranges from 0.85 to 0.92 for English tasks and 0.81 to 0.90 for Italian tasks. BEM scores show similar patterns but with greater variation, ranging from 0.77 to 0.94 for general domain tasks and dropping to 0.43 to 0.62 for specialized domains. Therefore, we observe a consistent divide between semantic metrics (BERTScore: 0.85-0.92) and syntactic metrics (ROUGE-L: 0.21-0.72). BEM scores show a strong correlation with human evaluation (0.735), suggesting that modern models may be better at capturing meaning than current syntactic evaluation metrics might indicate.

Reference-free metrics (Table 9) reveal a consistent gap between answer relevance and groundedness. Answer relevance scores remain high across all models (0.98-1.0 for SQuAD-en, 0.98-0.99 for SQuAD-it), while groundedness scores are notably lower (0.67-0.79 for SQuAD-en, 0.71-0.81 for SQuAD-it). Lower Groundedness scores compared to Answer Relevance scores across all models highlight a critical challenge in LLM-based QA systems. This gap is remarkably consistent across different models and languages, suggesting a fundamental challenge in maintaining factual accuracy while generating natural responses. Models sometimes generate plausible but unfaithful answers, emphasizing the need for improved mechanisms to ensure answer fidelity to the provided context.

The correlation analysis with human judgments (Table 10) provides crucial insights into metric reliability. On NarrativeQA books, BEM shows a strong correlation with human judgment (0.735) compared to answer relevance metrics (0.436). Similar patterns emerge for NarrativeQA movies, where BEM correlates at 0.704 with human judgment, while answer relevance shows correlation of 0.565. These results suggest that BEM more closely aligns with human assessment of answer quality than reference-free metrics.

Looking at IR results (Table 5), we see that nDCG@10 scores provide yet another perspective on quality assessment, showing clear performance gradients across domains and languages while maintaining consistency within similar task types.

This multi-metric analysis demonstrates that different evaluation approaches capture distinct aspects of model performance.

5.5. Practical Implications and Ethical Considerations

Our comprehensive evaluation reveals critical implications for the practical deployment of these systems, particularly in domains where accuracy directly impacts human welfare. The story told by our empirical results raises important considerations about how these systems should be implemented and monitored in real-world applications.

The journey from general to specialized domains reveals a particularly concerning pattern in our IR results. Looking at the embed-multilingual-v3.0 model's performance in Table 5, we see a dramatic decline in effectiveness as tasks become more specialized. Starting with an impressive nDCG@10 score of 0.90 in general domains, the performance plummets to just 0.36 in medical contexts (NFCorpus). This substantial drop of 0.54 points isn't just a number – it represents a significant degradation in the system's ability to retrieve relevant information in medical contexts, where accuracy can have direct implications for healthcare decisions.

The story becomes even more nuanced when we examine our QA results. Table 9 reveals a fascinating but troubling pattern in how models handle factual accuracy versus answer relevance. Take Mistral-Nemo's performance on CovidQA, for instance. While it achieves an impressive answer relevance score of 0.91, its groundedness score sits much lower at 0.64. We see similar patterns with GPT-4o, which shows a relevance score of 0.89 but a groundedness score of only 0.61 on specialized medical content. This gap between a model's ability to generate plausible-sounding answers and its ability to maintain factual accuracy raises serious concerns, particularly in medical and legal contexts where factual precision is paramount.

The importance of human oversight in these systems is not just a theoretical consideration – it's supported by our empirical findings. Our correlation analysis in Table 10 demonstrates that human judgment remains a crucial component in evaluating system performance. This finding reinforces what the performance gaps have already suggested: while these systems show remarkable capabilities, they cannot be deployed without appropriate human supervision and robust verification mechanisms.

These patterns in our data point to several crucial requirements for responsible system deployment. We need robust fact-checking mechanisms, particularly in specialized domains where performance degradation is most severe. We need clear protocols for human oversight, supported by our correlation analysis findings. Most importantly, we need transparent communication about system limitations, backed by our documented performance patterns.

6. Conclusion

Our analysis of embedding models and large language models across English and Italian has revealed several significant patterns that both advance our understanding and challenge common assumptions about multilingual AI systems. The patterns we observe suggest that while current approaches have made significant strides in bridging linguistic divides, particularly in general domains, substantial work remains in handling specialized knowledge and maintaining factual accuracy across languages.

The empirical results demonstrate clear performance patterns across languages and domains. In IR tasks, embed-multilingual-v3.0 maintains consistent performance across languages with minimal gaps (0.90 vs 0.86 nDCG@10 for English and Italian SQuAD respectively). However, performance degrades significantly in specialized domains, dropping to 0.36 for medical content (NFCorpus) in English tasks. In QA tasks, our results showed varying patterns across different evaluation metrics. Syntactic metrics revealed larger cross-lingual gaps (ROUGE-L scores of 0.72/0.69 vs 0.57/0.54 for Llama 3.1 8b on English vs Italian), while semantic metrics showed more stability (BERTScore ranging from 0.85-0.92 for English and 0.81-0.90 for Italian). The consistent gap between answer relevance (0.91-1.0) and groundedness (0.64-0.78) across models highlights a fundamental challenge in maintaining factual accuracy.

These findings have significant implications for both research and practice:

- The success of well-designed smaller models suggests that focused architectural innovation may be more valuable than simply scaling up existing approaches.
- The consistent pattern of domain-specific performance degradation indicates a need for more sophisticated approaches to specialized knowledge transfer across languages.
- The challenge of maintaining answer groundedness while preserving natural language generation capabilities emerges as a critical area for future work.

Building on these insights, our analysis identifies several critical areas for future research:

1. **Dataset Diversity:** Future work should expand to include a wider range of languages and domains to further validate the cross-lingual and domain adaptation capabilities of these models.
2. **Domain Adaptation:** The documented performance drop from general (0.90 nDCG@10) to specialized domains (0.36 nDCG@10) calls for more sophisticated domain adaptation mechanisms.
3. **Cross-lingual Knowledge Transfer:** The varying performance gaps between English and Italian, particularly in specialized domains, suggest the need for improved cross-lingual transfer methods. Explore methods for leveraging high-resource language models to improve performance on low-resource languages, potentially through zero-shot or few-shot learning approaches.

4. **Improved Groundedness:** Develop techniques to enhance the faithfulness of LLM-generated answers to the provided context, possibly through modified training objectives or architectural changes.
5. **Architectural Innovation:** The comparable performance of different-sized models (e.g., multilingual-E5-base vs large showing minimal differences of 0.01-0.02 nDCG@10 points) indicates that architectural efficiency may be more crucial than model scale. Therefore, developing more efficient architectures that maintain performance across languages without requiring massive computational resources is necessary.
6. **Long-context LLMs for QA:** Exploring the potential of emerging long-context LLMs (e.g., Claude 3, GPT-4 with extended context) in handling complex, multi-hop QA tasks without the need for separate retrieval steps. For addressing long documents, we will compare this approach with a smart selection of chunks through structural document representation (Document Object Model).
7. **Dynamic Retrieval:** Investigate adaptive retrieval methods that can dynamically adjust the number of retrieved passages based on query complexity or ambiguity.
8. **Multimodal IR and QA:** Extend the current work to include multimodal information retrieval and question answering, incorporating text, images, and potentially other modalities.
9. **Evaluation Methodologies:** Advance our understanding of how to assess both technical performance and practical utility in real-world applications. Develop better factual accuracy metrics given the observed groundedness challenges.
10. **Model Updates:** Given the rapid pace of development in NLP, regular re-evaluations with newly released models will be necessary to keep findings current.
11. **Interpretability and Explainability:** Develop methods for better understanding and interpreting the decision-making processes of dense retrievers and LLMs in IR and QA tasks.
12. **Ethical AI in IR and QA:** Further investigation into bias mitigation and fairness across languages and cultures in IR and QA systems. Develop frameworks for ethical deployment of AI-powered systems, including methods for bias detection and mitigation, and strategies for clearly communicating model limitations to end-users.

This comprehensive analysis serves not just as a benchmark of recent capabilities but as evidence-based guidance for future development in this rapidly evolving field. As we move forward, the focus must remain on creating systems that are not only more accurate and efficient, but also ethical, transparent, and truly beneficial to diverse linguistic communities worldwide while maintaining practical applicability.

References

1. Hambarde, K.A.; Proença, H. Information Retrieval: Recent Advances and Beyond. *IEEE Access* **2023**, *11*, 76581–76604. https://doi.org/10.1109/access.2023.3295776.
2. Anand, A.; Lyu, L.; Idahl, M.; Wang, Y.; Wallat, J.; Zhang, Z. Explainable Information Retrieval: A Survey, 2022, [arXiv:cs.IR/2211.02405].
3. Thakur, N.; Reimers, N.; Rücklé, A.; Srivastava, A.; Gurevych, I. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models, 2021, [arXiv:cs.IR/2104.08663].
4. Muennighoff, N.; Tazi, N.; Magne, L.; Reimers, N. MTEB: Massive Text Embedding Benchmark. In Proceedings of the Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics; Vlachos, A.; Augenstein, I., Eds., Dubrovnik, Croatia, 5 2023; pp. 2014–2037. https://doi.org/10.18653/v1/2023.eacl-main.148.
5. Tang, Y.; Yang, Y. MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries, 2024, [arXiv:cs.CL/2401.15391].
6. Zhang, Z.; Fang, M.; Chen, L. RetrievalQA: Assessing Adaptive Retrieval-Augmented Generation for Short-form Open-Domain Question Answering, 2024, [arXiv:cs.CL/2402.16457].
7. Gao, M.; Hu, X.; Ruan, J.; Pu, X.; Wan, X. LLM-based NLG Evaluation: Current Status and Challenges, 2024, [arXiv:cs.CL/2402.01383].

8. Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; Yih, W.t. Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); Webber, B.; Cohn, T.; He, Y.; Liu, Y., Eds., Online, 11 2020; pp. 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550.
9. Khattab, O.; Zaharia, M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In Proceedings of the Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, 2020; SIGIR '20, p. 39–48. https://doi.org/10.1145/3397271.3401075.
10. Xiong, L.; Xiong, C.; Li, Y.; Tang, K.F.; Liu, J.; Bennett, P.; Ahmed, J.; Overwijk, A. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval, 2020, [arXiv:cs.IR/2007.00808].
11. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); Burstein, J.; Doran, C.; Solorio, T., Eds., Minneapolis, Minnesota, 6 2019; pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.
12. Nogueira, R.; Yang, W.; Lin, J.; Cho, K. Document Expansion by Query Prediction, 2019, [arXiv:cs.IR/1904.08375].
13. Gao, L.; Callan, J. Condenser: a Pre-training Architecture for Dense Retrieval. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; Moens, M.F.; Huang, X.; Specia, L.; Yih, S.W.t., Eds., Online and Punta Cana, Dominican Republic, 11 2021; pp. 981–993. https://doi.org/10.18653/v1/2021.emnlp-main.75.
14. Wang, L.; Yang, N.; Huang, X.; Jiao, B.; Yang, L.; Jiang, D.; Majumder, R.; Wei, F. Text Embeddings by Weakly-Supervised Contrastive Pre-training, 2024, [arXiv:cs.CL/2212.03533].
15. Xiao, S.; Liu, Z.; Zhang, P.; Muennighoff, N.; Lian, D.; Nie, J.Y. C-Pack: Packaged Resources To Advance General Chinese Embedding, 2024, [arXiv:cs.CL/2309.07597].
16. Xiao, S.; Liu, Z.; Shao, Y.; Cao, Z. RetroMAE: Pre-Training Retrieval-oriented Language Models Via Masked Auto-Encoder. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; Goldberg, Y.; Kozareva, Z.; Zhang, Y., Eds., Abu Dhabi, United Arab Emirates, 12 2022; pp. 538–548. https://doi.org/10.18653/v1/2022.emnlp-main.35.
17. Liu, Z.; Xiao, S.; Shao, Y.; Cao, Z. RetroMAE-2: Duplex Masked Auto-Encoder For Pre-Training Retrieval-Oriented Language Models. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Rogers, A.; Boyd-Graber, J.; Okazaki, N., Eds., Toronto, Canada, 7 2023; pp. 2635–2648. https://doi.org/10.18653/v1/2023.acl-long.148.
18. Muffo, M.; Bertino, E. BERTino: an Italian DistilBERT model, 2023, [arXiv:cs.CL/2303.18121].
19. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Proceedings of the Proceedings of the 34th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2020; NIPS '20.
20. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. In Proceedings of the Advances in Neural Information Processing Systems; Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; Lin, H., Eds. Curran Associates, Inc., 2020, Vol. 33, pp. 1877–1901.
21. Liu, N.F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; Liang, P. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics* **2024**, *12*, 157–173. https://doi.org/10.1162/tacl_a_00638.
22. Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; Chang, M. Retrieval Augmented Language Model Pre-Training. In Proceedings of the ICML 2020, 13-18 July 2020, Virtual Event. PMLR, 2020, Vol. 119, *Proceedings of Machine Learning Research*, pp. 3929–3938.
23. Khattab, O.; Potts, C.; Zaharia, M. Relevance-guided Supervision for OpenQA with ColBERT, 2021, [arXiv:cs.CL/2007.00814].
24. Shuster, K.; Poff, S.; Chen, M.; Kiela, D.; Weston, J. Retrieval Augmentation Reduces Hallucination in Conversation, 2021, [arXiv:cs.CL/2104.07567].
25. Huo, S.; Arabzadeh, N.; Clarke, C. Retrieving Supporting Evidence for Generative Question Answering. In Proceedings of the SIGIR-AP. ACM, 2023, pp. 11–20. https://doi.org/10.1145/3624918.3625336.

26. Zhang, T.; Patil, S.G.; Jain, N.; Shen, S.; Zaharia, M.; Stoica, I.; Gonzalez, J.E. RAFT: Adapting Language Model to Domain Specific RAG, 2024, [arXiv:cs.CL/2403.10131].
27. Carterette, B.; Voorhees, E.M., Overview of Information Retrieval Evaluation. In *Current Challenges in Patent Information Retrieval*; Springer Berlin Heidelberg: Berlin, Heidelberg, 2011; pp. 69–85. https://doi.org/10.1007/978-3-642-19231-9_3.
28. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics; Isabelle, P.; Charniak, E.; Lin, D., Eds., Philadelphia, Pennsylvania, USA, 7 2002; pp. 311–318. https://doi.org/10.3115/1073083.1073135.
29. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 7 2004; pp. 74–81.
30. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. In Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020.
31. Es, S.; James, J.; Espinosa Anke, L.; Schockaert, S. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In Proceedings of the Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations; Aletras, N.; De Clercq, O., Eds., St. Julians, Malta, 3 2024; pp. 150–158.
32. Katranidis, V.; Barany, G. FaaF: Facts as a Function for the evaluation of RAG systems, 2024, [arXiv:cs.CL/2403.03888].
33. Saad-Falcon, J.; Khattab, O.; Potts, C.; Zaharia, M. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In Proceedings of the Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers); Duh, K.; Gomez, H.; Bethard, S., Eds., Mexico City, Mexico, 6 2024; pp. 338–354. https://doi.org/10.18653/v1/2024.naacl-long.20.
34. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Proceedings of the Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing; Su, J.; Duh, K.; Carreras, X., Eds., Austin, Texas, 11 2016; pp. 2383–2392. https://doi.org/10.18653/v1/D16-1264.
35. Rajpurkar, P.; Jia, R.; Liang, P. Know What You Don't Know: Unanswerable Questions for SQuAD, 2018, [arXiv:cs.CL/1806.03822].
36. Chen, D.; Fisch, A.; Weston, J.; Bordes, A. Reading Wikipedia to Answer Open-Domain Questions. In Proceedings of the Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Barzilay, R.; Kan, M.Y., Eds., Vancouver, Canada, 7 2017; pp. 1870–1879. https://doi.org/10.18653/v1/P17-1171.
37. Zhang, Y.; Nie, P.; Geng, X.; Ramamurthy, A.; Song, L.; Jiang, D. DC-BERT: Decoupling Question and Document for Efficient Contextual Encoding, 2020, [arXiv:cs.CL/2002.12591].
38. Croce, D.; Zelenanska, A.; Basili, R. Neural Learning for Question Answering in Italian. In Proceedings of the AI*IA 2018 – Advances in Artificial Intelligence; Ghidini, C.; Magnini, B.; Passerini, A.; Traverso, P., Eds., Cham, 2018; pp. 389–402.
39. Bonisoli, G.; Di Buono, M.P.; Po, L.; Rollo, F. DICE: a Dataset of Italian Crime Event news. In Proceedings of the Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, 2023; SIGIR '23, p. 2985–2995. https://doi.org/10.1145/3539618.3591904.
40. Wadden, D.; Lin, S.; Lo, K.; Wang, L.L.; van Zuylen, M.; Cohan, A.; Hajishirzi, H. Fact or Fiction: Verifying Scientific Claims. In Proceedings of the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); Webber, B.; Cohn, T.; He, Y.; Liu, Y., Eds., Online, 11 2020; pp. 7534–7550. https://doi.org/10.18653/v1/2020.emnlp-main.609.
41. Wachsmuth, H.; Syed, S.; Stein, B. Retrieval of the Best Counterargument without Prior Topic Knowledge. In Proceedings of the Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Gurevych, I.; Miyao, Y., Eds., Melbourne, Australia, 7 2018; pp. 241–251. https://doi.org/10.18653/v1/P18-1023.
42. Boteva, V.; Ghalandari, D.G.; Sokolov, A.; Riezler, S. A Full-Text Learning to Rank Dataset for Medical Information Retrieval. In Proceedings of the ECIR; Ferro, N.; Crestani, F.; Moens, M.F.; Mothe, J.; Silvestri, F.;

Nunzio, G.M.D.; Hauff, C.; Silvello, G., Eds. Springer, 2016, Vol. 9626, *Lecture Notes in Computer Science*, pp. 716–722.

43. Tang, R.; Nogueira, R.; Zhang, E.; Gupta, N.; Cam, P.; Cho, K.; Lin, J. Rapidly Bootstrapping a Question Answering Dataset for COVID-19, 2020, [arXiv:cs.CL/2004.11339].
44. Wang, L.L.; Lo, K.; Chandrasekhar, Y.; Reas, R.; Yang, J.; Burdick, D.; Eide, D.; Funk, K.; Katsis, Y.; Kinney, R.; et al. CORD-19: The COVID-19 Open Research Dataset, 2020, [arXiv:cs.DL/2004.10706].
45. Kočiský, T.; Schwarz, J.; Blunsom, P.; Dyer, C.; Hermann, K.M.; Melis, G.; Grefenstette, E. The NarrativeQA Reading Comprehension Challenge. *Transactions of the Association for Computational Linguistics* **2018**, *6*, 317–328. https://doi.org/10.1162/tacl_a_00023.
46. Li, Z.; Zhang, X.; Zhang, Y.; Long, D.; Xie, P.; Zhang, M. Towards General Text Embeddings with Multi-stage Contrastive Learning, 2023, [arXiv:cs.CL/2308.03281].
47. et al., A.D. The Llama 3 Herd of Models, 2024, [arXiv:cs.AI/2407.21783].
48. Team, G.; et al., T.M. Gemma: Open Models Based on Gemini Research and Technology, 2024, [arXiv:cs.CL/2403.08295].
49. Team, G.; et al., M.R. Gemma 2: Improving Open Language Models at a Practical Size, 2024, [arXiv:cs.CL/2408.00118].
50. Wrzalik, M.; Krechel, D. CoRT: Complementary Rankings from Transformers, 2021, [arXiv:cs.IR/2010.10252].
51. Tang, H.; Sun, X.; Jin, B.; Wang, J.; Zhang, F.; Wu, W. Improving Document Representations by Generating Pseudo Query Embeddings for Dense Retrieval, 2021, [arXiv:cs.IR/2105.03599].
52. Durmus, E.; He, H.; Diab, M. FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization. In Proceedings of the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.acl-main.454.
53. Goyal, T.; Li, J.J.; Durrett, G. News Summarization and Evaluation in the Era of GPT-3, 2023, [arXiv:cs.CL/2209.12356].
54. Bulian, J.; Buck, C.; Gajewski, W.; Börschinger, B.; Schuster, T. Tomayto, Tomahto. Beyond Token-level Answer Equivalence for Question Answering Evaluation. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; Goldberg, Y.; Kozareva, Z.; Zhang, Y., Eds., Abu Dhabi, United Arab Emirates, 12 2022; pp. 291–305. https://doi.org/10.18653/v1/2022.emnlp-main.20.
55. Kamalloo, E.; Dziri, N.; Clarke, C.L.A.; Rafiei, D. Evaluating Open-Domain Question Answering in the Era of Large Language Models, 2023, [arXiv:cs.CL/2305.06984].
56. Lerner, P.; Ferret, O.; Guinaudeau, C. Cross-modal Retrieval for Knowledge-based Visual Question Answering, 2024, [arXiv:cs.CL/2401.05736].
57. Blagec, K.; Dorffner, G.; Moradi, M.; Ott, S.; Samwald, M. A global analysis of metrics used for measuring performance in natural language processing, 2022, [arXiv:cs.CL/2204.11574].
58. Blagec, K.; Dorffner, G.; Moradi, M.; Samwald, M. A critical analysis of metrics used for measuring progress in artificial intelligence, 2021, [arXiv:cs.AI/2008.02577].
59. Joshi, M.; Choi, E.; Weld, D.S.; Zettlemoyer, L. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension, 2017, [arXiv:cs.CL/1705.03551].
60. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Guo, Q.; Wang, M.; et al. Retrieval-Augmented Generation for Large Language Models: A Survey, 2024, [arXiv:cs.CL/2312.10997].
61. Lin, C.Y.; Hovy, E. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In Proceedings of the Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 2003, pp. 150–157.
62. Weaver, K.F.; Morales, V.; Dunn, S.L.; Godde, K.; Weaver, P.F., Pearson's and Spearman's Correlation. In *An Introduction to Statistical Analysis in Research*; John Wiley and Sons, Ltd, 2017; chapter 10, pp. 435–471. https://doi.org/https://doi.org/10.1002/9781119454205.ch10.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.