AfriEconQA: A Benchmark Dataset for African Economic Analysis based on World Bank Reports

Edward Ajayi
Carnegie Mellon University Africa
Kigali, Rwanda
eaajayi@andrew.cmu.edu

Peace Bakare
Carnegie Mellon University Africa
Kigali, Rwanda
pbakare@andrew.cmu.edu

Abstract

We introduce **AfriEconQA**, a specialized benchmark dataset for African economic analysis grounded in a comprehensive corpus of 236 World Bank reports. The task of AfriEconQA is to answer complex economic queries that require high-precision numerical reasoning and temporal disambiguation from specialized institutional documents. The dataset consists of **8,937 curated QA instances**, rigorously filtered from a pool of 10,018 synthetic questions to ensure high-quality evidence-answer alignment. Each instance is composed of: (1) a **question** requiring reasoning over economic indicators, (2) the corresponding **evidence** retrieved from the corpus, (3) a verified ground-truth **answer**, and (4) **source metadata** (e.g., URL and publication date) to ensure temporal provenance. AfriEconQA is the first benchmark focused specifically on African economic analysis, providing a unique challenge for Information Retrieval (IR) systems as the data is largely absent from the pretraining corpora of current Large Language Models (LLMs). We operationalize this dataset through an 11-experiment matrix, benchmarking a zero-shot baseline (**GPT-5 Mini**) against RAG configurations using **GPT-4o** and **Qwen 32B** across five distinct embedding and ranking strategies. Our results demonstrate a severe "parametric knowledge gap," where zero-shot models fail to answer over 90% of queries, and even state-of-the-art RAG pipelines struggle to achieve high precision. This confirms AfriEconQA as a robust and challenging benchmark for the next generation of domain-specific IR and RAG systems.

The AfriEconQA dataset and code will be made publicly available upon publication.

CCS Concepts

• **Information systems** → **Question answering**; **Retrieval models and ranking**; • **Computing methodologies** → *Natural language processing*; • **Applied computing** → *Economics.*

Keywords

Question Answering, Retrieval-Augmented Generation, Dataset, Economic Analysis

ACM Reference Format:
Edward Ajayi and Peace Bakare. 2026. AfriEconQA: A Benchmark Dataset for African Economic Analysis based on World Bank Reports. In . ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Large Language Models (LLMs) have achieved impressive performance on general-knowledge benchmarks [17], yet their capabilities often falter in specialized domains where accuracy and precision are paramount. Economic analysis - a field critical to institutional decision-making and sustainable development [29] exemplifies this challenge. In the domain, practitioners must reason over technical reports dense with statistics, temporal nuances, and domain-specific terminology. Despite the importance of this domain, evaluation datasets that reflect these real-world complexities remain scarce, particularly for emerging markets in Africa.

The core difficulty lies in the nature of closed-domain settings [6]. Institutional reports from organizations like the World Bank contain intricate tables, carefully worded policy explanations, and time-sensitive data that distinguish historical facts from future projections. Models must answer questions by reasoning over these specialized, institutionally curated documents rather than relying on parametric knowledge acquired during pretraining. Evaluating retrieval-augmented generation (RAG) systems on such long-form, technical reports rich in statistics and temporal nuances requires specialized benchmarks.

To address this gap, we introduce **AfriEconQA**, a specialized benchmark dataset constructed from a comprehensive corpus of 236 World Bank reports focused on African economies. The primary goal of AfriEconQA is to provide a real-world, hard-to-memorize testbed that allows researchers to evaluate RAG performance on lengthy, technical, and statistics-heavy documents. These reports are unlikely to appear in the pretraining corpora of current LLMs and contain diverse, multi-modal content, including narrative policy explanations, complex tables, numerical indicators, and enumerated lists.

AfriEconQA consists of **8,937 curated QA instances**, meticulously filtered from an initial pool of over 10,000 synthetic questions to ensure high-fidelity evidence-answer alignment. By providing precise metadata—including publication dates and source universal resource locators (URL), the dataset challenges information retrieval (IR) systems to perform "temporal disambiguation," distinguishing between historical actuals and future projections. We demonstrate that state-of-the-art RAG systems struggle with these temporal and

This work received no external funding. No prior presentations of this material have been made. Author's current address: Carnegie Mellon University Africa, Kigali, Rwanda.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

CURRENT QUESTION

What is the projected percentage of Malawi's population that will have access to on-grid electricity by 2030?

PROPOSED ANSWER

30 percent

FACTUAL EVIDENCE

```
Malawi is working rapidly to close large gaps in electricity access, with plans to achieve 50 percent access by 2030 through a combination of on-grid and off-grid options. The Government has set a goal of increasing access to on-grid electricity to 30 percent by 2030, and to reach another 20 percent of the population with off-grid solutions.
```

Figure 1: ***Example of a factoid question-answer pair from the dataset with supporting evidence.***

statistical reasoning challenges, highlighting the need for improved retrieval and reasoning capabilities in domain-specific contexts.

In this study, we operationalize **AfriEconQA** to characterize the current state of domain-specific retrieval and reasoning. We frame our evaluation around three primary research questions:

- **RQ1:** To what extent do state-of-the-art models rely on parametric memory versus external retrieval for niche economic data?
- **RQ2:** How do different retrieval paradigms (Sparse, Dense, and Hybrid) handle the "terminological variance" and numerical density of institutional economic reports?
- **RQ3:** Can open-weight models achieve parity with proprietary APIs in high-precision extractive tasks within this specialized domain?

2 Related Work

Question answering has evolved from systems designed for broad, open-domain queries to specialized architectures that reason over domain-specific document collections. While Large Language Models (LLMs) have achieved strong performance on general-knowledge benchmarks [8], their effectiveness diminishes in specialized domains requiring high-precision retrieval and reasoning over technical documents. This has driven growing interest in domain-focused QA systems that combine retrieval mechanisms with language models to ground answers in authoritative sources.

2.1 Open-Domain Question Answering

Open-domain question answering systems aim to answer queries across diverse topics using large-scale corpora. Popular benchmarks include the Stanford Question Answering Dataset (SQuAD) [20], QAMPARI [2], XQA [15], and more recent datasets [16, 25, 26, 30], all developed to evaluate general-purpose QA models [9].

These systems typically rely on a retriever model [9] to extract relevant passages from large document collections. The retrieved passages are then fed into a QA model to generate answers, allowing models to handle a wide variety of topics without explicit training on every possible question.

However, a major challenge for open-domain QA is the need for frequent knowledge updates [31]. Since information changes rapidly, models relying solely on pretrained knowledge may provide outdated or incorrect answers [24].

2.2 Closed-Domain Question Answering

In contrast to open-domain systems, closed-domain question answering focuses on retrieving and reasoning over specific document collections. These systems operate on limited, well-defined vocabularies and contexts, making the task fundamentally different [3].

Closed-domain QA has found particular application in institutional settings [7, 13, 22] and specialized domains where general-purpose models fail to provide adequate precision. The combination of Large Language Models (LLMs) with retrieval-augmented generation (RAG) has accelerated development in this area [23], enabling organizations to build domain-focused systems that answer questions unlikely to be addressed by LLMs alone.

These systems have become especially prominent in organizational chatbots [7, 22], where users require accurate, context-grounded responses from internal knowledge bases. Closed-domain QA thus bridges the gap between general-purpose LLMs and domain-specific knowledge, enabling more reliable information retrieval.

2.3 Economic and Financial Question Answering

Closed-domain QA has been extensively studied in institutional [13], biomedical [1, 12], and legal [21] contexts. The financial and economic domains, however, present unique challenges requiring joint reasoning over structured tables and unstructured narratives, as well as high-precision extraction of domain-specific terminology [11, 19].

Most existing financial QA systems focus on corporate analysis using company filings, stock market data, and regulatory documents [10]. Economic development reports—such as those published by the World Bank [28]—offer a more complex environment, weaving together policy narratives, longitudinal data, and multi-year macroeconomic projections. Despite their importance, datasets in this sector remain limited, and existing financial benchmarks largely center on general market terminology or Western corporate data.

AfriEconQA addresses this gap by focusing on macroeconomic policy and sovereign economic analysis for African nations. To the best of our knowledge, this is the first open-source benchmark specifically dedicated to African economies, grounded in the dense statistical and policy narratives characteristic of World Bank institutional reporting.

3 Dataset

3.1 Source Documents

The **AfriEconQA** dataset is curated from **236** distinct World Bank economic reports focusing on the African continent [28], using the search term *"African economic analysis"*. These documents cover a wide range of topics, including macroeconomic updates, digital transformation strategies, and climate resilience reports. The corpus was explicitly selected to challenge retrieval systems with:

(1) **Temporal Ambiguity:** Multiple reports (e.g., "Rwanda Economic Update") are published at intervals of months or years. If a given answer appears in more than one report, we select the **most recent** version to ensure temporal consistency.
(2) **High Specificity:** The texts contain precise floating-point figures (e.g., "10.8 percent") and specialized acronyms.

3.2 Preprocessing Pipeline

To prepare the unstructured text for retrieval, we implemented a robust preprocessing pipeline:

- **Metadata Extraction:** We utilized the Google Custom Search API to recover the URL for each document, ensuring ground truth provenance.
- **Chunking Strategy:** Documents were segmented using a Recursive Character Splitter with a chunk size of 1,000 characters and an overlap of 200 characters. This yielded a total of 64892 chunks, balancing context preservation with granular retrieval.

3.3 Question Curation and Taxonomy

We generated an initial pool of 10,018 unique Question-Answer (QA) pairs using a synthetic pipeline grounded in an economic-expert persona. The pipeline was designed to ensure that the queries reflect the analytical depth required for institutional economic monitoring. To provide a comprehensive evaluation of information retrieval and reasoning, we enforced a diverse taxonomy across the following five categories:

- **Factoid (3,237 samples):** High-precision tasks requiring the extraction of specific macroeconomic indicators or policy instruments (e.g., *"What was Tunisia's merchandise trade deficit as a percentage of GDP in 2023?"*).
- **List (1,705 samples):** Synthesis tasks requiring the identification of multiple discrete factors, such as economic risk drivers.
- **Comparison (1,525 samples):** Reasoning across temporal or regional dimensions, requiring the system to quantify differences between historical outturns and projections.
- **Multiple Choice (1,265 samples):** Discrimination tasks where the system must identify correct statements among plausible distractors drawn from the same document context.
- **Synthesis (1,205 samples):** Complex queries focused on economic transmission mechanisms and causal chains explicitly described in the policy narratives.

A key feature of our curation process was temporal anchoring. We implemented logic to convert relative temporal terms found in the text (e.g., "this year" or "currently") into absolute calendar or fiscal years based on the document's publication metadata. This ensures that the questions are self-contained and evaluate a system's ability to perform temporal disambiguation.

3.4 Quality Control and Filtering

To ensure the reliability of the benchmark, each generated record underwent a secondary verification and filtering process. We utilized a specialized auditing layer to enforce strict factual and structural constraints:

- **Numerical Verification:** We enforced a mandatory verbatim constraint. Every percentage, fiscal year, and monetary value in the answer was required to appear character-for-character within the provided evidence snippet. Any records containing rounded figures or hallucinated metrics were discarded.
- **Entity-Based Decontextualization:** To ensure queries were standalone and solvable in a retrieval-augmented environment, we performed entity-anchoring. Generic references (e.g., "the government" or "the ministry") were replaced with specific names (e.g., "the Nigerian Ministry of Finance") derived from the document metadata.
- **Removal of Meta-Language:** To ensure a realistic query profile, we stripped all meta-commentary and instructional artifacts, such as "According to the report" or "In the provided text." This forces the system to rely on semantic understanding rather than linguistic patterns common in synthetic data.
- **Substance Filtering:** The pipeline filtered for economic relevance, discarding non-economic content such as technical software instructions or internal file-naming conventions.

This multi-stage verification process resulted in a 10.8% attrition rate, filtering the initial pool down to a final curated set of 8,937 high-quality samples. This ensures that the final dataset is factually grounded in verbatim institutional evidence and professionally phrased for economic analysis.

3.5 Dataset Statistics

Table 1 summarizes the scale and structural properties of the **AfriEconQA** benchmark. The corpus is characterized by high information density, designed to evaluate retrieval precision across a large-scale collection of technical reports.

The corpus was segmented using a recursive character text splitter, resulting in 64,892 chunks. The inclusion of a 200-character overlap ensures that semantic context is preserved across chunk boundaries, preventing the loss of critical economic data points that may span multiple segments. With approximately 38 curated questions per document, the benchmark provides a high-density environment that requires both effective retrieval and robust reasoning to resolve.

4 Method

4.1 Indexing and Preprocessing

To enable efficient retrieval, we implement a dual-index architecture over the 64,892 document chunks:

Table 1: Summary statistics of the AfriEconQA benchmark.

Statistic	Value
Total Source Documents (World Bank Reports)	236
Total Retrieval Units (Chunks)	64,892
Chunk Size	1,000 chars
Chunk Overlap	200 chars
Total Curated QA Pairs	8,937
Reasoning Taxonomy Breakdown:	
– Factoid	3,237
– List	1,705
– Comparison	1,525
– Multiple Choice	1,265
– Synthesis	1,205

(1) **Sparse Index:** We tokenize the corpus and build an inverted index using BM25, which excels at exact lexical matching of entities and numerical figures.
(2) **Dense Index:** We encode all 64,892 chunks into vector embeddings. We experiment with two embedding models: *Google GenAI Embeddings* (API-based) and *BAAI/BGE-m3* (Open-Weight).

Each chunk is indexed with its source metadata (publication date and URL), which is returned alongside retrieval results to enable temporal disambiguation during answer generation.

4.2 Hybrid Retrieval via Reciprocal Rank Fusion

While Dense Retrieval captures semantic meaning, it often struggles with the precise numbers common in economic queries. To mitigate this, we implement a Hybrid Retrieval strategy using Reciprocal Rank Fusion (RRF).

For a given query q, we retrieve the top-k documents from both the Sparse (D_{sparse}) and Dense (D_{dense}) indices. The final RRF score for a document d is calculated as:

$$RRF(d) = \sum_{r \in R} \frac{1}{\eta + rank_r(d)} \quad (1)$$

where $R = \{D_{sparse}, D_{dense}\}$ and η is a constant. This method normalizes the disparate scoring scales of BM25 and Vector Similarity, effectively promoting documents that appear in both result sets.

5 Experimental Setup

To characterize the utility of the AfriEconQA benchmark, we evaluate a matrix of state-of-the-art Large Language Models (LLMs) across diverse retrieval architectures.

5.1 System Configurations

We define 11 experimental configurations based on the interplay between three generators and four retrieval strategies:

- **Generators:** We utilize **GPT-4o** as the proprietary industry standard and **Qwen 2.5 32B** as a high-performance open-weight baseline. To establish a "Parametric Knowledge" floor, we include **GPT-5 Mini** in a zero-shot, no-retrieval configuration.
- **Retrievers:** We evaluate four retrieval paradigms at a fixed depth of $k = 3$:
  – **Sparse (BM25):** Lexical matching using Okapi BM25 [27].
  – **Dense-Local (BGE):** High-performance local embeddings (*bge-base-en-v1.5*) [4].
  – **Dense-Cloud (Google):** State of the Art (SOTA) cloud-based embeddings (*text-embedding-004*).
  – **Hybrid (RRF):** Reciprocal Rank Fusion [5] combining Sparse and Dense indices to evaluate the benefits of fused ranking.

5.2 Evaluation Metrics

As *AfriEconQA* is an Information Retrieval resource, we report a dual suite of metrics:

(1) **Ranking Quality:** Precision@k, Recall@k, and Mean Reciprocal Rank (MRR) to evaluate the retriever's ability to surface the correct 1,000-character chunk from the 64,892-chunk corpus.
(2) **Answer Quality:** Exact Match (EM), F1-score, BLEU [18], and ROUGE-L [14] for deterministic similarity. Additionally, we employ **LLM-as-a-Judge** [32] (using GPT-4o-mini) to provide a semantic evaluation (0.0–1.0) of factual consistency between the prediction and the gold-standard answer.

6 Results and Analysis

We evaluate all 11 experimental configurations on a stratified random sample of 300 questions from the full AfriEconQA benchmark. Table 2 summarizes the overall performance across retrieval and generation metrics. Our analysis focuses on three critical dimensions: the failure of parametric memory, the precision-coverage tradeoff in retrieval, and the extractive discipline of open-weight models.

6.1 The Parametric Knowledge Vacuum

The performance of **GPT-5 Mini** (Zero-Shot) establishes a critical baseline for AfriEconQA. With an LLM-Judge score of 0.081 and an EM of 0.053, it is evident that World Bank economic indicators for African nations are virtually absent from the pre-training corpora of current LLMs. As shown in Table 4, the zero-shot model completely fails on *Comparison* (0.00%) and *Factoid* (1.69%) queries. This "parametric vacuum" underscores that African economic analysis cannot rely on model weights alone; retrieval-augmentation is a fundamental requirement for accuracy in this domain.

6.2 Retrieval Analysis: The Precision-Coverage Tradeoff

A significant IR finding is the performance tradeoff between **Dense** and **Hybrid** retrieval. Pure **Google Dense** retrieval achieved the highest MRR (0.763), but when fused with BM25, the MRR dropped to 0.722. However, **Hybrid Google + GPT-4o achieved the highest LLM-Judge score (0.512)**, suggesting that the slight ranking degradation is offset by improved answer quality.

Table 2: Benchmark Results for Retriever + LLM Combinations ($k = 3$) on stratified sample ($n = 300$). Retriever: Sparse (BM25), Dense (BGE/Google), Hybrid (RRF). Metrics: EM (Exact Match), F1, BLEU, ROUGE1/2/L, LLM-Judge (Semantic Accuracy), P@3, R@3, MRR, MAP. GPT-5 Mini represents the zero-shot baseline.

Retriever	Model	k	EM	F1	BLEU	R1	R2	RL	LLM-J	P@3	R@3	MRR	MAP
BGE	GPT-4o	3	0.177	0.395	0.153	0.412	0.301	0.386	0.471	0.266	0.797	0.711	0.711
BGE	Qwen 32B	3	0.220	0.377	0.146	0.394	0.294	0.377	0.451	0.266	0.797	0.711	0.711
Google	GPT-4o	3	0.213	0.427	0.173	0.445	0.330	0.416	0.498	0.280	0.840	0.763	0.763
Google	Qwen 32B	3	0.243	0.404	0.158	0.421	0.318	0.398	0.487	0.280	0.840	0.763	0.763
Hybrid BGE	GPT-4o	3	0.187	0.404	0.165	0.423	0.310	0.398	0.471	0.247	0.740	0.677	0.677
Hybrid BGE	Qwen 32B	3	0.240	0.395	0.168	0.414	0.314	0.397	0.470	0.247	0.740	0.677	0.677
Hybrid Google	GPT-4o	3	0.223	0.438	0.184	0.452	0.349	0.428	**0.512**	0.261	0.783	0.722	0.722
Hybrid Google	Qwen 32B	3	**0.270**	0.436	0.187	0.451	0.351	0.432	0.508	0.261	0.783	0.722	0.722
None	GPT-5 Mini	3	0.053	0.061	0.045	0.064	0.051	0.061	0.081	0.000	0.000	0.000	0.000
BM25	GPT-4o	3	0.137	0.238	0.112	0.245	0.196	0.231	0.288	0.146	0.437	0.373	0.373
BM25	Qwen 32B	3	0.163	0.238	0.119	0.245	0.208	0.236	0.281	0.146	0.437	0.373	0.373

We attribute the MRR reduction to *lexical collisions* within economic reports. Institutional documents frequently reuse technical keywords (e.g., "GDP," "growth," "fiscal year") across disparate sections. Sparse retrieval (BM25) struggles to disambiguate these terms, introducing "ranking noise" into the hybrid pipeline. However, the diversity of retrieved contexts appears to benefit the generation phase, where multiple perspectives aid in producing semantically accurate answers.

6.3 Generation: Extractive Discipline vs. Semantic Paraphrasing

The interaction between generators and retrievers reveals a compelling trade-off between open-weight and proprietary models:

- **Extractive Discipline: Qwen 32B** consistently achieved higher Exact Match (EM) scores than GPT-4o (e.g., 0.270 vs 0.223 on Hybrid Google). Qwen demonstrates a "literalist" approach, faithfully extracting numerical figures and verbatim phrases from the evidence.
- **Semantic Robustness: GPT-4o** maintained the lead in the *LLM-Judge* metric (0.512), which evaluates semantic accuracy. GPT-4o often paraphrases the evidence for better readability, which penalizes its EM score but preserves the underlying economic truth.

This suggests that for tasks requiring high-fidelity numerical extraction—where "close enough" is a failure—highly capable open-weight models like Qwen 32B may be the superior architectural choice.

6.4 Difficulty by Reasoning Taxonomy

Table 3 highlights the "reasoning ceiling" of the current benchmark. **Multiple Choice** is the best-performing category (58.98%) as the model can rely on recognition. However, **List** (20.88%) and **Comparison** (23.30%) queries remain significantly underserved. These tasks require the retriever to surface multiple, often non-contiguous, chunks of data and the generator to synthesize them across different time horizons. The low accuracy in these categories suggests that single-pass retrieval architectures struggle with multi-faceted economic queries, indicating potential benefits from agentic or multi-hop retrieval strategies in future work.

Table 3: Average Accuracy by Question Type Across All Models (Stratified Sample, $n = 300$)

Question Type	Count	Avg Accuracy (%)
List	54	20.88
Comparison	48	23.30
Synthesis	39	42.19
Factoid	118	46.92
Multiple Choice	41	58.98

7 Limitations

While **AfriEconQA** provides a specialized testbed for African economic analysis, it has several limitations:

- **Source Homogeneity:** The dataset is grounded exclusively in a corpus of 236 World Bank reports. While these are high-authority documents, they represent a specific institutional perspective that may not capture the full diversity of African economic discourse found in local news or alternative policy papers.
- **Verbatim Numerical Constraint:** The automated quality control enforces a mandatory verbatim constraint for all percentages, fiscal years, and monetary values. While this ensures high fidelity, it may limit the evaluation of systems on abstractive reasoning tasks where the answer is not a direct string match from the evidence.
- **Temporal Staticity:** The benchmark reflects economic data and projections at a specific point in time. As macroeconomic conditions in emerging markets shift, the "ground truth" for projections will eventually require updates to maintain relevance for real-time IR systems.
- **Evaluation Scope:** Due to computational constraints, we evaluate all system configurations on a stratified sample of 300 questions rather than the full 8,937-question benchmark. While this subset is carefully stratified to maintain the distribution of question types, future work should validate these findings on the complete dataset at scale.

Table 4: Per-Model Accuracy by Question Type (%). Models evaluated on stratified sample of 300 questions with LLM-Judge threshold ≥ 0.7. Bold indicates highest accuracy per question type.

Retriever	Generator	Comparison	Factoid	List	Multiple Choice	Synthesis
Hybrid Google	GPT-4o	22.92	**61.02**	25.93	73.17	53.85
Hybrid Google	Qwen 32B	29.17	57.63	25.93	73.17	51.28
Google	GPT-4o	18.75	57.63	22.22	**75.61**	46.15
Google	Qwen 32B	25.00	54.24	**29.63**	68.29	51.28
Hybrid BGE	GPT-4o	29.17	58.47	24.07	56.10	43.59
Hybrid BGE	Qwen 32B	**31.25**	55.93	25.93	60.98	48.72
BGE	GPT-4o	33.33	53.39	25.93	53.66	**53.85**
BGE	Qwen 32B	29.17	50.85	24.07	56.10	46.15
BM25	GPT-4o	16.67	33.05	11.11	48.78	30.77
BM25	Qwen 32B	20.83	32.20	11.11	48.78	25.64
None	GPT-5 Mini	0.00	1.69	3.70	34.15	12.82

8 Ethical Considerations

The development and release of **AfriEconQA** adhere to the following ethical principles:

- **Data Provenance and Attribution:** All source documents are publicly available institutional reports. We provide full metadata, including canonical URLs and official publication dates, to ensure transparency and respect for source provenance.
- **Addressing Representation Gaps:** By demonstrating that current language models possess minimal parametric knowledge of African economic data (GPT-5 Mini: 8.1% LLM-Judge accuracy), this benchmark highlights the critical need for retrieval-augmented systems in specialized domains and discourages reliance on potentially hallucinated parametric knowledge for sensitive economic contexts.
- **Open Resource Access:** To support equitable development, the dataset is released as a public resource to advance research in high-precision information retrieval for regions often underrepresented in large-scale pre-training corpora.

9 Data Generation and Prompts

The dataset was constructed using a synthetic pipeline grounded in an **economic-expert persona**. Due to space constraints, the full set of prompt configurations is available in our public repository. Our prompting strategy enforced several key constraints:

- **Temporal Anchoring:** Prompts included logic to convert relative terms (e.g., "this year") into absolute calendar or fiscal years based on document metadata.
- **Entity-Based Decontextualization:** Instructions required replacing generic references like "the government" with specific names (e.g., "the Nigerian Ministry of Finance") to ensure queries are standalone and solvable via RAG.
- **Taxonomy Enforcement:** Unique prompt templates were utilized to generate five distinct reasoning categories: Factoid, List, Comparison, Multiple Choice, and Synthesis.

10 Conclusion

We introduce **AfriEconQA**, a specialized benchmark dataset for African economic analysis comprising 8,937 curated QA pairs grounded in 236 World Bank reports. Our evaluation on a stratified sample of 300 questions reveals three key findings: (1) current language models possess virtually no parametric knowledge of African economic indicators, demonstrating the necessity of retrieval-augmented architectures for this domain; (2) while dense retrieval achieves superior ranking precision, hybrid retrieval with reciprocal rank fusion produces higher-quality answers by diversifying retrieved contexts; and (3) open-weight models like Qwen 32B excel at extractive precision for numerical data, offering a viable alternative to proprietary APIs for institutional applications requiring strict fidelity. We release AfriEconQA as an open resource to advance research in domain-specific, temporally-sensitive information retrieval for underrepresented regions.

References

[1] Husan Almusawi. 2025. *BIOMED QA: KNOWLEDGE GRAPH REASONING WITH PRE-AND POST-RETRIEVAL GRAPH RAG FOR BIOMEDICAL QUESTION ANSWERING SYSTEM*. Master's thesis. Cleveland State University.

[2] Samuel Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig, and Jonathan Berant. 2023. QAMPARI: A benchmark for open-domain questions with many answers. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*. 97–110.

[3] Haniel G Cavalcante, Jéferson N Soares, and José EB Maia. 2021. Question expansion in a question-answering system in a closed-domain system. *Int. J. Comput. Appl* 183, 23 (2021), 1–5.

[4] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216* (2024).

[5] Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 758–759.

[6] Xavier Daull, Patrice Bellot, Emmanuel Bruno, Vincent Martin, and Elisabeth Murisasco. 2025. Complex QA and language models hybrid architectures, Survey. arXiv:2302.09051 [cs.CL] https://arxiv.org/abs/2302.09051

[7] Caner Derici, Yiğit Aydin, Çiğdem Yenialaca, Nihal Yağmur Aydin, Günizi Kartal, Arzucan Özgür, and Tunga Güngör. 2018. A closed-domain question answering framework using reliable resources to assist students. *Natural Language Engineering* 24, 5 (2018), 725–762.

[8] Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or Memorization: Data Contamination and Trustworthy Evaluation for Large Language Models. arXiv:2402.15938 [cs.CL] https://arxiv.org/abs/2402.15938

[9] Rujun Han, Peng Qi, Yuhao Zhang, Lan Liu, Juliette Burger, William Yang Wang, Zhiheng Huang, Bing Xiang, and Dan Roth. 2023. RobustQA: Benchmarking the Robustness of Domain Adaptation for Open-Domain Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 4294–4311. doi:10.18653/v1/2023.findings-acl.263

[10] Ivan Iaroshev, Ramalingam Pillai, Leandro Vaglietti, and Thomas Hanne. 2024. Evaluating Retrieval-Augmented Generation Models for Financial Report Question and Answering. *Applied Sciences (2076-3417)* 14, 20 (2024).

[11] Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944* (2023).

[12] Seonok Kim. 2025. Medbiolm: Optimizing medical and biological qa with fine-tuned large language models and retrieval-augmented generation. *arXiv preprint arXiv:2502.03004* (2025).

[13] Gustavo Kuratomi, Paulo Pirozelli, Fabio G Cozman, and Sarajane M Peres. 2025. A RAG-Based Institutional Assistant. *arXiv preprint arXiv:2501.13880* (2025).

[14] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[15] Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019. XQA: A cross-lingual open-domain question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2358–2368.

[16] Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics* 9 (2021), 1389–1406.

[17] Shiwen Ni, Guhong Chen, Shuaimin Li, Xuanang Chen, Siyi Li, Bingli Wang, Qiyao Wang, Xingjian Wang, Yifan Zhang, Liyang Fan, Chengming Li, Ruifeng Xu, Le Sun, and Min Yang. 2025. A Survey on Large Language Model Benchmarks. arXiv:2508.15361 [cs.CL] https://arxiv.org/abs/2508.15361

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[19] Yinzhu Quan and Zefang Liu. 2024. EconLogicQA: A Question-Answering Benchmark for Evaluating Large Language Models in Economic Sequential Reasoning. arXiv:2405.07938 [cs.CL] https://arxiv.org/abs/2405.07938

[20] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* (2016).

[21] Cheol Ryu, Seolhwa Lee, Subeen Pang, Chanyeol Choi, Hojun Choi, Myeonggee Min, and Jy-Yong Sohn. 2023. Retrieval-based evaluation for LLMs: a case study in Korean legal QA. In *Proceedings of the Natural Legal Language Processing Workshop 2023*. 132–137.

[22] Matthew Saad and Zakariya Qawaqneh. 2024. Closed Domain Question-Answering Techniques in an Institutional Chatbot. In *2024 4th International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*. 1–8. doi:10.1109/ICECCME62383.2024.10796881

[23] Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics* 11 (2023), 1–17.

[24] Wei Tang, Yixin Cao, Yang Deng, Jiahao Ying, Bo Wang, Yizhe Yang, Yuyue Zhao, Qi Zhang, Xuan-Jing Huang, Yu-Gang Jiang, et al. 2025. Evowiki: Evaluating llms on evolving knowledge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 948–964.

[25] Jiexin Wang, Adam Jatowt, and Masatoshi Yoshikawa. 2022. Archivalqa: A large-scale benchmark dataset for open-domain question answering over historical news collections. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3025–3035.

[26] Zhen Wang. 2022. Modern question answering datasets and benchmarks: A survey. *arXiv preprint arXiv:2206.15030* (2022).

[27] John S Whissell and Charles LA Clarke. 2011. Improving document clustering using Okapi BM25 feature weighting. *Information retrieval* 14, 5 (2011), 466–487.

[28] World Bank. [n. d.]. Documents & Reports — All Documents | The World Bank. https://documents.worldbank.org/en/publication/documents-reports. Accessed: 2025-12-04.

[29] World Bank. 2023. *Financial & Economic Analysis: Differences and Similarities*. World Bank, EEX Global Knowledge Unit (IEEGK). https://documents1.worldbank.org/curated/en/099093024122526452/pdf/P179103-ce55a292-c282-4136-b911-b0b1e66d7ad5.pdf Prepared by EEX Global Knowledge Unit (IEEGK), December 2023..

[30] Wenhao Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. 2023. Ifqa: A dataset for open-domain question answering under counterfactual presuppositions. *arXiv preprint arXiv:2305.14010* (2023).

[31] Zixuan Zhang, Revanth Gangi Reddy, Kevin Small, Tong Zhang, and Heng Ji. 2024. Towards Better Generalization in Open-Domain Question Answering by Mitigating Context Memorization. arXiv:2404.01652 [cs.CL] https://arxiv.org/abs/2404.01652

[32] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* 36 (2023), 46595–46623.