DiverseRAG: Multi-Source Retrieval Augmented Generation for Multilingual and Multidialectal Question Answering with LLMs

Anonymous ACL submission

Abstract

The field of question answering (QA) has been significantly transformed by the emergence of Large Language Models (LLMs). However, their performance in domain-specific QA, such as in e-government applications, is limited by their access to external, real-time, and highly specific knowledge. To address this, we introduce DiverseRAG, a novel framework that combines Retrieval-Augmented Generation (RAG) with LLMs, emphasizing a multi-source and multi-grained retrieval process to enhance response accuracy and relevance. Our approach employs a multi-source RAG strategy, drawing from diverse data types such as web pages and legal texts, and a multi-grained retrieval process that operates on sentence and multi-sentence levels to ensure both precision and contextual depth in addressing questions. This approach ensures comprehensive coverage of government-related questions. To test DiverseRAG, we curated an English-Arabic dataset from UAE government websites and further extend the questions into 4 Arabic dialects: Egyptian, Iraqi, Lebanese, and Emirati. Our results demonstrate that DiverseRAG substantially boosts performance of LLMs for English, MSA, and dialectal Arabic queries in the government domain, achieving over 10% improvement in metrics such as F-1 score, BertScore, ROUGE and Context Precision compared to conventional RAG approach in the best case.

1 Introduction

The emergence of Large Language Models (LLMs) (Brown et al., 2020; Ouyang et al., 2022) has marked a significant leap forward in the capabilities of question answering (QA) systems, achieving new state-of-the-art in natural language understanding and response generation (Bang et al., 2023; OpenAI, 2023). These models, trained on vast corpora of text, have shown remarkable proficiency in generating coherent and contextually appropriate answers across a broad spectrum of

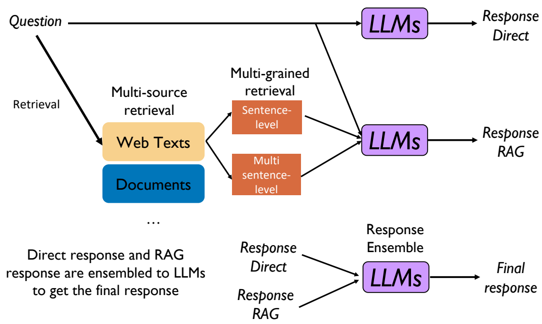

Figure 1: An overview of the ensemble framework of our `DiverseRAG` approach, illustrating the integration of multi-grained retrieval and diverse knowledge sources as well as how to ensemble the final response.

general knowledge questions. However, the utility of LLMs in specialized domains presents a unique set of challenges, particularly when the questions involve complex, domain-specific knowledge that goes beyond the general information contained in the training data (Lai et al., 2023; Yang et al., 2023).

Some QA domains, such as government or law require greater emphasis on accuracy and specificity where standard open-domain approaches often fall short. In such domains, where the information is often not stored within the model's parametric knowledge, a retrieval-based solution, often integrating multiple sources, is necessary. In this work, we explore and improve LLMs' QA capabilities in retrieval-augmented generation, focusing on government-domain QA.

We propose a `DiverseRAG` approach integrated with an ensemble framework that extends the capabilities of LLMs by incorporating a dynamic retrieval component. This approach leverages the intrinsic generative capabilities of LLMs while enhancing them with targeted, contextually relevant data retrieved from diverse knowledge sources. The multi-source retrieval component dynamically aggregates information from various domains, including web texts and legal documents, ensuring a comprehensive coverage of the knowledge landscape necessary for government-domain QA. Our framework is further designed to perform multi-grained retrieval processes (Chen et al., 2023) that effectively harness information from varied data formats and granularities—from sentence-level to multi-sentence level. This multi-grained approach allows for more comprehensive extraction of crucial data as well as broader contextual information. The system can efficiently identify and retrieve the most relevant pieces of information, which are then fed into enriched prompts for the LLM. An illustration of our approach is shown in Figure 1. This ensures that the generated responses are not only accurate but also adequately contextualized, addressing the complexity and specificity of government queries. By integrating these capabilities, the DiverseRAG framework provides a robust solution tailored specifically to the requirements of government-domain QA.

To support our framework, we conducted extensive data collection, crucial for developing and evaluating our QA system. This includes two primary datasets: a comprehensive knowledge base of web pages and legal texts for the RAG process, and a parallel English-MSA dataset of over 2,000 FAQs. The knowledge base provides relevant context for accurate government domain answers, while the EN-MSA dataset evaluates system performance in both languages. Additionally, we translated a subset of these FAQs into four Arabic dialects—Emirati, Levantine, Egyptian, and Iraqi—to address linguistic diversity and ensure effectiveness across dialects. This diverse data collection is essential for testing and refining our framework, ensuring robustness in real-world government-domain queries.

Our research distinguishes itself through its focus on the integration of these multi-grained retrieval processes. This method supports a more nuanced combination of information by dynamically integrating various levels of granularity, which significantly enhances the relevance and accuracy of the responses from LLMs. The multilingual nature of our corpus, encompassing both English (EN), Arabic (MSA), and dialects, adds further complexity and broadens the impact value of our research, making it relevant in a diverse range of governmental contexts.

- We propose `DiverseRAG`, a novel approach that enhances LLMs with multi-source, multi-grained retrieval to improve accuracy and relevance in government-domain QA.
- We construct a new multilingual dataset of government-related QA to enable domain-specific question answering, encompassing diverse sources from UAE government web pages, legal documents, and FAQs.
- We further translate the question into four Arabic dialects-Emirati, Levantine, Egyptian, and Iraqi, which is useful for cross-dialect benchmarking.
- We demonstrate the effectiveness of `DiverseRAG` through rigorous experimentation across various LLMs, significantly outperforming traditional zero-shot methods in English and Arabic, paving the way for future evaluations with advanced models.

Next, we present related work in Section 2 and detail our methodology in Section 3. In Section 4 we present our data collection effort; and we discuss our experimental results in Section 5.

2 Related Work

The intersection of LLMs with domain-specific QA as been an area of active research, with several studies addressing the limitations and potential enhancements of LLMs in this context.

2.1 Domain-Specific QA with LLMs

The limitations of LLMs in domain-specific contexts (Chen et al., 2024; Sadat et al., 2023), have been noted by Shuster et al. (2021), who argue that while LLMs are proficient in general QA, their performance drops significantly when answering domain-specific questions. This is particularly evident in government-domain QA, where the need for up-to-date and precise information is paramount (Cui et al., 2023). Our work contributes to this area by specifically addressing the challenges posed by government-domain questions.

2.2 Retrieval-Augmented Generation (RAG)

The concept of RAG, introduced by Lewis et al. (2020) has been a significant advancement in the field (Gao et al., 2023). It combines the generative power of LLMs with a retrieval component to provide more accurate, up-to-data, and relevant answers (Lee et al., 2019; Goldfarb-Tarrant et al., 2024; Arefeen et al., 2024) while alleviating the need to train or finetune a model for that specific domain. Our approach extends this work by implementing a multi-grained retrieval process that leverages a wider array of diverse knowledge sources, including legal PDF articles and web texts.

2.3 Use of Diverse Knowledge Sources

The integration of diverse knowledge sources into QA systems has been explored in various capacities (Dinan et al., 2019; Peng et al., 2023). Karpukhin et al. (2020) introduced a dense vector retrieval method, namely DPR, that significantly improved the retrieval of relevant documents for open-domain QA (Liu et al., 2021). This aspect is crucial in the government-domain QA scenario since the information (e.g. tweets, web announcements, newly-passed laws, regulations, . . . , etc.) shared by the government can occur using different media. Our framework builds upon this by not only retrieving but also effectively integrating information from these diverse sources into the generative process of LLMs.

2.4 Multilingual/Low-resource Language QA

The challenges of multilingual QA, especially in low-resource languages (Longpre et al., 2021; Nguyen et al., 2023), have been highlighted by Asai et al. (2021a,b), who introduced a cross-lingual open-retrieval QA dataset. Our research addresses this gap by incorporating MSA and Arabic dialects into our RAG framework, thus enhancing its applicability in multilingual government contexts.

2.5 Evaluation Metrics for QA

The development of robust evaluation metrics remains crucial for advancing QA systems. Some studies (Zhang et al., 2020; Sellam et al., 2020) have proposed metrics based on contextual embeddings that offer more nuanced assessments of model outputs. More recently, the use of LLMs in evaluation (Fu et al., 2023) provides a comprehensive measure of the quality of generated responses.

3 Methodology

The methodology of this work focuses on integrating LLMs with our proposed DiverseRAG approach within an ensemble framework. DiverseRAG leverages multiple types of knowledge sources and employs a multi-grained retrieval process. It accesses web texts and legal texts to compile a comprehensive set of documents for each query. The framework utilizes various levels of granularity, such as sentence-level and multi-sentence-level retrieval, to provide LLMs with the necessary context for generating accurate answers. The ensemble framework enhances this process by combining direct LLM responses with the context-enriched outputs from DiverseRAG, ensuring that the final answers are both comprehensive and precise.

3.1 DiverseRAG Approach

The core of our methodology is the `DiverseRAG` approach, designed to utilize a wide range of knowledge sources and detailed retrieval techniques to enhance the contextual understanding of LLMs.

Diverse Knowledge Sources: The `DiverseRAG` framework taps into a variety of knowledge sources to ensure thorough and reliable information retrieval. These include:

- **Web Texts:** A collection of over 19,000 web pages from various ministry websites provides a broad spectrum of current governmental information.
- **Legal Documents:** Approximately 611 legal documents offer detailed insights into statutory and regulatory frameworks.

By combining information from these sources, `DiverseRAG` aims to make sure that the retrieved content is both relevant and comprehensive.

Document Retrieval: For a given query q, the document retrieval process is initiated to gather a relevant set of documents $\boldsymbol{D} = \{\boldsymbol{d_1}, \boldsymbol{d_2}, \ldots, \boldsymbol{d_n}\}$ from the knowledge base. This retrieval is based on the relevance of each document to the query, maximizing the cosine similarity between encoded representations:

$$\boldsymbol{C} = \arg\max_{\boldsymbol{d}_i \in \boldsymbol{D}}{}^{k} \left(\frac{F_e(\boldsymbol{q}) \cdot F_e(\boldsymbol{d}_i)^T}{\|F_e(\boldsymbol{q})\| \|F_e(\boldsymbol{d}_i)\|} \right) \quad (1)$$

where F_e is our encoder model for encoding query and documents $\boldsymbol{q}$, $F_e(\boldsymbol{q}) \in \mathbf{R}^{1\times h}$ and $F_e(\boldsymbol{d}_i) \in \mathbf{R}^{1\times h}$. k indicates that we sample the top-k documents from $\boldsymbol{Q}$ that maximize the cosine similarity with $\boldsymbol{q}$.

Multi-grained Retrieval: To address the complex nature of government queries, our retrieval process operates on multiple granularities:

- **Sentence-Level Retrieval:** Focuses on extracting precise, sentence-level information crucial for addressing specific aspects of the query.
- **Multi-Sentence-Level Retrieval:** Provides broader contextual information by retrieving

documents consisting of multiple sentences, thereby enhancing the comprehensiveness of the response.

This multi-grained approach, denoted as $\boldsymbol{D}_g$ where g indicates the granularity level.

3.2 Ensemble Framework

Building on our proposed `DiverseRAG` framework, we introduce an ensemble framework that combines both direct and contextually enriched responses to produce the final response. Our ensemble framework is motivated by the need to ensure accuracy and reliability in the final response. It balances the quality of direct LLM responses with the precision and depth provided by `DiverseRAG`-enhanced responses. By integrating both responses, the ensemble guarantees a baseline quality, ensuring that at least the direct LLM response can be relied upon if the RAG-enhanced response is not satisfactory.

Upon receiving a query q, the ensemble framework uses two parallel generation pathways:

(1) **Direct LLM Generation:** Utilizes the inherent capabilities of the LLM to generate a response R_{direct} directly from the query:

$$R_{\text{direct}} = \text{LLM}(q) \quad (2)$$

(2) **DiverseRAG-Enhanced Generation:** Engages the `DiverseRAG` framework to retrieve context C from the knowledge sources, forming an enriched prompt P for the LLM:

$$P = \text{"Context: "} + C + \text{"Question: "} + q \quad (3)$$

The LLM then generates a contextually informed response R_{RAG}:

$$R_{\text{RAG}} = \text{LLM}(P) \quad (4)$$

Ensemble Integration: The final step involves merging the two responses, R_{direct} and R_{RAG}, through an ensemble mechanism. This integration evaluates and combines the responses, resulting in a final answer R_{final} that is both comprehensive and contextually accurate:

$$R_{\text{final}} = \text{LLM}(R_{\text{direct}}, R_{\text{RAG}}) \quad (5)$$

This ensemble approach leverages the distinct advantages of each generation pathway while mitigating their individual limitations.

4 Data Collection

In the domain of government-domain question answering, there is a notable absence of standard QA datasets especially evaluation benchmark and knowledge bases for LLMs, motivating us to conduct our own extensive data collection. In this section, we give the details of the collection and construction of the data resources in this work. Specifically, we provide corresponding details of the collection of FAQs in Section 4.1, the process of crawling web and legal documents in Section 4.2 and the details of building dialectal translation of FAQs in Section 4.3.

4.1 Crawling Government Website FAQs

UAE government websites provide useful information to citizens, residents and visitors on a wide range of topics. We selected 15 of those sites that represent various government domains. Each of these sites contains bilingual information and a dedicated FAQ section. The first step of our data collection process therefore involved scraping FAQ sections from all these government sites. The full list is shown in Table 8 in Appendix. We crawled 2,134 EN FAQs and 2,205 MSA FAQs. We then carried out an alignment of this bilingual set, from which we curated a parallel set of 360 FAQs in both EN and MSA and translated the questions to four Arabic dialects: Egyptian, Lebanese, Iraqi and Emirati (Section 4.3). This combination represented a realistic QA use-case test set for UAE government domain content. We split the parallel dataset into two parts; `Mixed domain`, which includes 200 FAQs representing 40 FAQs each from 5 different websites (MOHRE, UAE Government Portal, MOEC, FTA and MOE) and `Single domain`, which includes 160 FAQs from the MOHRE site only.

4.2 Collection of Web Texts and Legal Documents

To facilitate our DiverseRAG government-domain QA system, we conduct collection of various knowledge bases. Specifically, we aim at the texts of government websites for up-to-date information and legal documents for reliable legal references and accurate legal information.

	Web Pages	Articles
# of Web pages	9, 404	611
# of Sentences	247, 406	13, 621
# of Words	4, 569, 452	93, 347
# of tokens	19, 823, 377	2,434,115

Table 1: Descriptive statistics for the collected web pages and PDF/Docx articles in English and Arabic.

4.2.1 Collection of Web Texts

The web crawling process was implemented using the Selenium [1] package, specifically utilizing the Safari browser to navigate and extract textual content from selected UAE governmental websites. The crawler targeted a list of URLs and was set to extract text within paragraph (<p>) tags, ensuring that each segment contained at least two words to retain meaningful content. A recursive link-following strategy was employed, gathering hyperlinks within anchor (<a>) tags to expand the scope of the crawl, while avoiding links to non-informative pages such as login screens and downloadable files. Specific handling of website interactions was incorporated to manage common obstacles like pop-ups and cookie consent forms, with adjustments made for certain sites where initial user actions were necessary to access the main content. The process was controlled to limit the crawl to 100,000 pages per base URL, with a delay between page loads to simulate human browsing behavior and adhere to website policies.

The collected web texts encompass a wide array of topics and are represented by the statistics available in Table 1.

4.2.2 Collection of Legal Documents

Alongside web texts, we compiled a diverse collection of legal texts from government websites in various formats (PDF, DOCX) and languages (English and Arabic). The statistics for these documents are detailed in 1. This collection of web and legal documents forms the knowledge bases of our DiverseRAG system.

4.3 FAQs Translation into Arabic Dialects

The UAE is a multicultural society that is home to a diverse population that includes both locals (Emiratis) and expatriates from around the globe. Expatriates represent over 80% of the population, including speakers of dialects from neighbouring Arab countries such as Gulf (Saudi Arabia, Bahrain, Kuwait, Qatar), Levantine (Lebanon, Jordan, Palestine, Syria), Egyptian and Iraqi.

To make our system applicable in a realistic setting, we therefore translated the question-side of 360 FAQs (`Mixed domain` and `Single domain`) into four Arabic dialects, specifically Gulf (Emirati), Levantine (Lebanese), Egyptian (Cairo) and Iraqi (Baghdad). We outsourced this translation task to a professional language service provider. For the translation task set-up, we shared only the EN question and EN answer with the translators. We chose not to share the MSA version of the questions so as not to prime or bias the translators with specific choices of terminology of linguistic structures. The translations were then reviewed by our own in-house language experts to assess the quality of the dataset.

Our internal reviews revealed a number of issues arose with the Arabic dialect translations. Firstly, the translator for Emirati dialect chose an informal register, which differed greatly to the formal register of the source text. E.g. زق *zq* translates as 'hit up', instead of 'call'; يتطرش *ytTrš* translates as 'blast off' instead of 'send' and عسب *ςsb* translates as the equivalent of 'cuz' instead of 'because'. Additionally, some translations did not reflect the specific terminology typically used in UAE government domains. These were updated to reflect official use: e.g. استقطاب 'recruitment', اعفاء *AςfA'* 'exemption', الهيكل التنظيمي *Alhykl Altnðymy* 'organizational structure', القانون الاتحادي *AlqAnwn AlAtHAdy* 'Federal Law', الإمارات السبع *AlǍmArAt Alsbς* 'The Seven Emirates', صدور القرار *Sdwr AlqrAr* 'issuance of the decision', and النزاع العمالي الجماعي *AlnzAς AlςmAly AljmAςy* 'Collective Labor Dispute'. A full revision of the register resulted in two versions of the Emirati question set; formal and informal. Secondly, the Iraqi translations presented a number of issues with the feedback from the reviewer stating that the professionally translated text "was useful but not accurate" and that it did not reflect translations by an Iraqi speaker but someone familiar with the dialect. Specific terms were modified to reflect common usage such as: using the "ch" sound جان *jAn* instead of كان *kAn*, misuse of فصل *fSl* which is used for tribal issues

[1] `https://pypi.org/project/selenium/`

and not legal issues, and merging شنو هية *šnw hyħ* 'what's this[?]'into شنية *šnyħ*, for example.

5 Experiments

5.1 Experimental Setup

We employed several state-of-the-art LLMs in our experiments, including Llama (Touvron et al., 2023a,b; Dubey et al., 2024), Mistral (Jiang et al., 2023), AceGPT (Huang et al., 2024), and Jais (Sengupta et al., 2023), which vary in model sizes ranging from 7 billion to 70 billion parameters.

Datasets: The dataset consisted of 2,134 English FAQs and 2,205 Arabic FAQs collected from 15 UAE government ministries, providing a comprehensive set of questions typical in government-domain applications.

Knowledge Sources: The knowledge base for the RAG included 9, 404 web pages and 611 legal documents, we split them into sentence-level and multi-sentence-level to facilitate multi-granularity retrieval.

Retrieval Setup: We employ Multi-grained retrieval fully utilize information at both sentence-level and multi-sentence-level, enriching the LLMs' context. We firstly use BM25 to retrieve top-1000 documents for each granularity and then use Sentence-BERT to pick the top-k documents.

5.2 Evaluation Metric

To evaluate the enhancements brought by our proposed RAG approach for LLMs on gov-domain QA, we employ various metrics including BertScore (Zhang et al., 2020), BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and F-1 score. We also employ a widely used evaluation framework for RAG named RAGAS (Es et al., 2023) including six metrics: *Faithfulness*, *Answer Relevancy*, *Context Precision*, *Context Recall*, *Answer Similarity*, *Answer Correctness*, as well as GPTScore (Fu et al., 2023), which all assess the relevance and accuracy of the generated responses using LLMs.

5.3 Results

Comparison between w/ RAG and w/o RAG In our experiments, we evaluated the effectiveness of our `DiverseRAG` approach in enhancing the performance of various LLMs on our English and MSA FAQs, the results of multiple LLMs (with RAG and without RAG) on conventional metrics including BertScore, BLEU, ROUGE and F-1 score are shown in Table 2 and Table 3. We further show the results evaluated by RAGAS (Es et al., 2023) in Table 4 including *Faithfulness, Answer Relevance, Answer Correctness*. As demonstrated by the results (indicated by the bold numbers in brackets), our proposed `DiverseRAG` approach showed substantial improvements over counterpart LLMs with non-RAG methods across all metrics. For English FAQs, the integration of retrieval-augmented generation consistently enhanced BertScore, BLEU, ROUGE and F1 scores as well as LLMs-based metrics *Faithfulness, Answer Relevance, Answer Correctness* among all tested LLMs, including Vicuna-7B and Llama-2-7B which got substantial increases in BLEU scores with the help of our approach. This indicates that `DiverseRAG` effectively harnesses diverse knowledge sources to improve the quality of generated answers. Similarly, the results on MSA FAQs revealed substantial performance gains, with Llama-3-8B and AceGPT-13B showing improvements in F1 scores (+12.1 and +6.4, respectively) and Llama-3-8B's improvement on *Faithfulness* for English and MSA FAQs. These highlight the framework's ability in handling MSA questions.

Effect of Multi-grained Retrieval We study the effect of integrating a multi-grained retrieval method within our `DiverseRAG` approach. We compare its performance to the baseline vanilla RAG approach (only pick the top-1 sentence) across six metrics by RAGAS (Es et al., 2023) using Vicuna-7B, LLama3-8B, and Jais-13B models for both MSA and English FAQs. The results are shown in Table 5.

Our `DiverseRAG` approach consistently enhances performance of LLMs on Gov-domain QA, with substantial improvements noted in *Faithfulness* and *Context Recall*, particularly for Jais-13B, which saw increases up to 8.9 and 15.8 points, respectively. Importantly, the improvements in *Context Recall* and *Context Precision* across all models highlight the effectiveness of the multi-grained retrieval in accurately sourcing and utilizing relevant data. These enhancements in retrieval precision are crucial for generating more coherent and contextually aligned responses. These findings show the effectiveness of the multi-grained retrieval in refining the retrieval and generation quality of the models.

Model	BERTScore			BLEU				ROUGE-L			F1 Score
	Precision	Recall	F1	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Precision	Recall	F1	
Vicuna-7B	81.7 (**+3.5**)	84.9 (**+2.0**)	83.2 (**+2.8**)	7.5 (**+8.8**)	3.0 (**+5.6**)	1.2 (**+4.1**)	0.4 (**+3.1**)	9.4 (**+8.8**)	33.2 (**-1.1**)	13.2 (**+7.2**)	28.6 (**+3.9**)
Llama-2-7B	81.0 (**+3.7**)	84.8 (**+2.2**)	82.8 (**+3.0**)	5.3 (**+8.1**)	2.3 (**+4.8**)	1.0 (**+3.2**)	0.4 (**+2.3**)	8.3 (**+7.2**)	37.8 (**-1.8**)	12.4 (**+6.8**)	30.3 (**+3.3**)
Llama-2-13B	81.7 (**+1.6**)	85.6 (**+0.8**)	83.6 (**+1.2**)	4.7 (**+4.3**)	2.0 (**+2.5**)	0.8 (**+1.6**)	0.3 (**+1.1**)	7.9 (**+3.9**)	38.2 (**-0.9**)	12.0 (**+4.1**)	30.4 (**+2.7**)
Llama-3-8B	77.6 (**+3.3**)	84.9 (**+1.8**)	81.1 (**+2.6**)	0.2 (**+4.1**)	0.1 (**+2.1**)	0.1 (**+1.1**)	0.1 (**+0.7**)	2.8 (**+5.6**)	47.8 (**+4.6**)	5.0 (**+7.6**)	23.8 (**+5.9**)
Mistral-7B	83.2 (**+2.2**)	85.6 (**+1.6**)	84.4 (**+1.9**)	9.3 (**+7.0**)	3.9 (**+5.7**)	1.6 (**+4.9**)	0.7 (**+4.2**)	11.0 (**+7.2**)	32.9 (**+4.1**)	14.8 (**+6.9**)	30.5 (**+5.5**)
Mixtral-8x7B	82.5 (**+2.5**)	86.1 (**+1.1**)	84.2 (**+1.8**)	7.7 (**+7.6**)	3.3 (**+5.1**)	1.5 (**+3.8**)	0.7 (**+2.8**)	10.0 (**+7.0**)	35.6 (**+0.3**)	13.9 (**+6.5**)	29.2 (**+4.2**)
AceGPT-13B	82.8 (**+1.6**)	86.0 (**+0.7**)	84.4 (**+1.2**)	8.4 (**+5.6**)	3.3 (**+3.9**)	1.3 (**+3.0**)	0.5 (**+2.2**)	9.9 (**+5.8**)	30.8 (**+1.2**)	13.3 (**+5.1**)	26.9 (**+2.9**)
Jais-13B	84.3 (**+1.6**)	85.3 (**+1.0**)	84.8 (**+1.3**)	14.6 (**+5.4**)	5.0 (**+6.7**)	1.8 (**+6.6**)	0.8 (**+6.2**)	16.6 (**+8.2**)	19.8 (**+7.5**)	15.4 (**+7.4**)	24.8 (**+7.5**)

Table 2: Comparison of LLMs without RAG and with our DiverseRAG approaches evaluated on English FAQs. The numbers shown in this table are the performance of non-RAG approach, the bold numbers within the brackets are the improvements given by our proposed DiverseRAG approach.

Model	BERTScore			BLEU				ROUGE-L			F1 Score
	Precision	Recall	F1	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Precision	Recall	F1	
Vicuna-7B	63.6 (**+2.1**)	66.8 (**+0.7**)	65.0 (**+1.4**)	6.9 (**+1.7**)	2.8 (**+0.7**)	1.5 (**+0.3**)	0.8 (**+0.1**)	2.3 (**+3.4**)	3.1 (**+6.2**)	2.0 (**+3.6**)	13.7 (**+4.2**)
Llama-2-7B	59.2 (**+3.7**)	62.5 (**+3.0**)	60.7 (**+3.4**)	2.4 (**+3.5**)	0.8 (**+1.5**)	0.4 (**+0.9**)	0.2 (**+0.6**)	0.9 (**+7.5**)	5.8 (**+6.6**)	0.7 (**+8.0**)	4.7 (**+9.7**)
Llama-2-13B	62.0 (**+2.2**)	66.1 (**+0.4**)	63.9 (**+1.4**)	4.9 (**+2.7**)	2.1 (**+1.0**)	1.1 (**+0.5**)	0.6 (**+0.3**)	1.7 (**+8.5**)	5.2 (**+8.8**)	1.9 (**+8.6**)	13.3 (**+3.9**)
Llama-3-8B	58.7 (**+7.0**)	68.3 (**+1.2**)	63.1 (**+4.5**)	1.6 (**+6.8**)	0.6 (**+2.6**)	0.3 (**+1.1**)	0.1 (**+0.5**)	1.1 (**+9.1**)	8.8 (**+12.0**)	1.2 (**+11.0**)	8.5 (**+12.1**)
Mistral-7B	58.9 (**+4.4**)	64.8 (**+1.0**)	61.4 (**+3.0**)	5.8 (**+1.2**)	2.8 (**+0.2**)	1.7 (**+0.1**)	1.0 (**+0.1**)	2.1 (**+3.3**)	2.2 (**+4.1**)	1.6 (**+2.4**)	12.6 (**+3.7**)
Mixtral-8x7B	66.8 (**+1.0**)	68.3 (**+1.0**)	67.4 (**+1.1**)	9.8 (**+1.7**)	4.3 (**+0.4**)	2.2 (**+0.1**)	1.2 (**+0.1**)	3.6 (**+3.7**)	5.9 (**+2.9**)	3.4 (**+4.5**)	14.8 (**+7.1**)
AceGPT-13B	63.3 (**+4.3**)	69.4 (**+0.3**)	66.1 (**+2.4**)	4.9 (**+6.5**)	2.0 (**+2.7**)	0.9 (**+1.3**)	0.4 (**+0.7**)	3.8 (**+10.1**)	7.0 (**+12.5**)	4.1 (**+10.5**)	16.5 (**+6.4**)
Jais-13B	66.1 (**+1.6**)	68.8 (**+0.1**)	67.3 (**+0.8**)	9.0 (**+1.5**)	3.2 (**+0.6**)	1.3 (**+0.2**)	0.5 (**+0.2**)	3.5 (**+6.9**)	4.5 (**+5.6**)	2.9 (**+4.2**)	13.6 (**+5.0**)

Table 3: Comparison of LLMs without RAG and with our DiverseRAG approaches evaluated on MSA FAQs. The numbers shown in this table are the performance of non-RAG approach.

Model	Faithful.	Ans. Rel.	Ans. Corr.
Vicuna-7B (MSA)	59.8 (**+5.5**)	63.2 (**+2.7**)	61.1 (**+7.3**)
LLama3-8B (MSA)	56.4 (**+8.8**)	61.9 (**+5.0**)	58.3 (**+6.0**)
Jais-13B (MSA)	61.9 (**+8.2**)	68.5 (**+2.2**)	65.6 (**+6.2**)
Vicuna-7B (En)	62.1 (**+6.8**)	66.0 (**+3.2**)	64.5 (**+3.8**)
LLama3-8B (En)	59.3 (**+5.7**)	64.8 (**+5.3**)	61.7 (**+6.6**)
Jais-13B (En)	64.8 (**+9.6**)	71.5 (**+1.5**)	69.2 (**+7.0**)

Table 4: Evaluation results using RAGAS for three LLMs without RAG and with our DiverseRAG. The bold numbers within the brackets are the improvements given by DiverseRAG.

Effect of Knowledge Bases We further examine how different knowledge base utilisation affect the Jais-13B model's performance in MSA and English, results shown in Table 6. The tested setups include using both Web and Document sources, Web-only, Document-only, and no external knowledge base. The combination of Web and Document sources achieved the highest performance in both languages. For MSA, this setup achieves an F-1 Score of 18.6, while in English, it reaches 32.3, alongside the best scores in Answer Relevance and Correctness. This indicates that accessing diverse knowledge sources enhances the model's capacity to produce relevant and accurate responses. In contrast, using solely Web or Document sources results in slightly worse performance, and the absence of a knowledge base leads to the lowest scores. These results show the importance of integrating multiple knowledge bases to improve LLMs outputs.

5.4 Evaluating Dialectal Variation for Gov-domain QA

We further conduct experiments to examine the performance of various LLMs when encountering dialects of Arabic such as Egyptian, Iraqi, etc, which are commonly used in real life. Specifically, we translate a subset (360 questions) of the full set of the FAQs in MSA to four Arabic dialects including Egyptian, Lebanese, Iraqi and Emirati, the experimental results are shown in Table 7. This study evaluates Vicuna-7B, LLama3-8B, and Jais-13B on Egyptian, Lebanese, Iraqi, and Emirati dialects. We assess *Faithfulness*, *Answer Relevance*, and *Answer Correctness*, comparing DiverseRAG with vanilla RAG. DiverseRAG consistently enhances performance of LLMs. Jais-13B shows the largest *Faithfulness* gains with 7.1 points in Lebanese. Vicuna-7B and LLama3-8B also improve, with Vicuna-7B gaining 5.3 points in *Answer Correctness* for Iraqi, and LLama3-8B gaining 7.2 points in *Faithfulness* for Emirati. Results demonstrate the DiverseRAG approach effectively handles linguistic variations, improving both the relevance and accuracy of responses across Arabic dialects.

5.5 Qualitative Error Analysis

We conduct errors analysis on a set of 22 randomly selected MSA samples including question, context,

Model	Faithfulness	Answer Relevancy	Context Precision	Context Recall	Answer Similarity	Answer Correctness
Vicuna-7B (MSA)	61.5 (**+3.8**)	65.0 (**+1.0**)	63.7 (**+9.6**)	50.1 (**+10.5**)	78.2 (**+6.5**)	63.3 (**+5.1**)
LLama3-8B (MSA)	58.2 (**+7.0**)	63.8 (**+3.1**)	63.3 (**+8.2**)	53.0 (**+15.1**)	73.6 (**+7.5**)	60.9 (**+3.4**)
Jais-13B (MSA)	63.8 (**+8.9**)	70.6 (**+2.5**)	65.4 (**+9.3**)	53.5 (**+14.9**)	82.9 (**+5.4**)	67.3 (**+6.7**)
Vicuna-7B (En)	64.2 (**+4.7**)	67.9 (**+1.3**)	67.1 (**+10.1**)	54.3 (**+11.9**)	81.0 (**+7.3**)	66.8 (**+1.5**)
LLama3-8B (En)	61.4 (**+3.6**)	66.3 (**+3.8**)	66.0 (**+9.8**)	55.4 (**+16.4**)	76.9 (**+8.2**)	64.7 (**+3.6**)
Jais-13B (En)	67.5 (**+6.9**)	72.4 (**+0.6**)	68.8 (**+10.4**)	57.2 (**+15.8**)	86.7 (**+6.0**)	71.2 (**+5.6**)

Table 5: Evaluation results of the comparison between vanilla RAG approach and our `DiverseRAG` approach with multi-grained retrieval, the numbers are the performance of LLMs with vanilla RAG approach measured by RAGAS, the bold numbers in brackets are improvements given by the multi-grained retrieval method in our proposed `DiverseRAG` approach.

Language	Configuration	F-1 Score	Ans. Rel.	Ans. Corr.
MSA	Web & Doc	18.6	73.1	74.1
	Web	13.9	70.5	70.8
	Doc	15.6	71.0	71.7
	None	13.6	70.6	67.3
English	Web & Doc	32.3	77.5	78.7
	Web	28.7	75.1	71.8
	Doc	30.5	75.0	73.4
	None	24.8	74.4	71.2

Table 6: Ablation Study Results for the effect of knowledge base of Jais-13B.

Model	Dialects	Faithful.	Ans. Rel.	Ans. Corr.
Vicuna-7B	Egyptian	57.5 (**+3.8**)	61.0 (**+1.0**)	58.8 (**+4.6**)
	Lebanese	59.8 (**+4.5**)	63.6 (**+1.5**)	62.1 (**+2.0**)
	Iraqi	58.2 (**+3.5**)	61.8 (**+1.3**)	59.6 (**+5.3**)
	Emirati	58.9 (**+4.0**)	63.0 (**+1.6**)	60.5 (**+4.8**)
LLama3-8B	Egyptian	54.1 (**+6.5**)	59.5 (**+2.8**)	56.0 (**+3.7**)
	Lebanese	56.9 (**+3.4**)	62.4 (**+3.0**)	59.3 (**+4.4**)
	Iraqi	54.8 (**+6.8**)	60.2 (**+3.0**)	56.7 (**+4.0**)
	Emirati	55.7 (**+7.2**)	61.2 (**+3.3**)	57.6 (**+4.5**)
Jais-13B	Egyptian	59.5 (**+5.8**)	66.2 (**+0.2**)	63.3 (**+4.0**)
	Lebanese	62.3 (**+7.1**)	69.2 (**+0.4**)	66.9 (**+4.8**)
	Iraqi	60.6 (**+6.1**)	66.7 (**+0.4**)	64.2 (**+4.2**)
	Emirati	61.7 (**+6.7**)	67.7 (**+0.7**)	65.2 (**+4.5**)

Table 7: Evaluation results of three LLMs on our test data translated into four Arabic dialects, we compare the performance of these LLMs on dialectal data with vanilla RAG approach with our `DiverseRAG` approach, the bold numbers in brackets are the improvements by our approach.

and answer from Jais-13B (our best model), as shown in Table 3 and 4. While the system shows competitive results in terms of evaluation metrics, yet it demonstrates some weaknesses. The most prominent reason for the incorrect answers is its frequent failure to leverage the provided context to answer questions accurately often ignoring relevant information within the context or retrieving information unrelated to the query. Another significant proportion of responses exhibits poor alignment with the posed questions, focusing on tangentially related information or peripheral aspects rather than addressing the core query intent, potentially limiting their usefulness for users seeking comprehensive answers. A third notable source of errors can be attributed to the model's occasional hallucination generating factually incorrect or contradictory information, particularly problematic in domains involving legal or procedural content where precision is crucial. This suggests some limitations in the model's knowledge representation or retrieval mechanisms, leading to the production of responses that, while coherent, contain inaccuracies not supported by the provided context or general factual knowledge. Lastly, the system demonstrates a tendency to present oversimplified representations of intricate processes and regulatory frameworks. This reductionist approach can potentially lead users to develop incomplete or distorted understandings of critical information. This errors analysis highlights key improvements: context, alignment, accuracy, and depth. Addressing these enhances reliability for precise responses.

6 Conclusion and Future Work

In this paper, we introduced `DiverseRAG`, a RAG framework for leveraging diverse knowledge sources such as web pages and legal documents, and employing multi-grained retrieval. We curated a new benchmark of English and MSA government FAQs, crawled web and legal texts for knowledge bases, and provided dialectal translations into four Arabic dialects. Our experiments with bilingual datasets from UAE government websites demonstrate that `DiverseRAG` largely enhances response accuracy and relevance in English, MSA, and various Arabic dialects. This shows its effectiveness in domain-specific and linguistically diverse QA tasks. Future research will focus on extending `DiverseRAG` to other domains such as healthcare and finance, where domain-specific knowledge is crucial. We also plan to refine the model's handling of dialectal variations and expand its language support to enhance generalizability and impact.

Limitations

While `DiverseRAG` demonstrates considerable improvements in domain-specific question answering, several limitations remain. Firstly, our proposed approach relies on pre-existing knowledge sources means that the system's accuracy can be affected by outdated or incomplete information, reading the need for frequent updates to maintain relevance. Additionally, although the framework is capable of dealing with multiple dialects, its performance may vary with less common dialectal variations not covered by the current dataset. Furthermore, the integration of multi-grained retrieval methods introduces computational complexity, which might impact efficiency and scalability when applied to larger datasets or in real-time applications. Finally, while our experiments focus on the e-government domain, further validation across diverse domains is necessary to fully understand the framework's generalizability and adaptability to different types of queries and document structures.

References

Md Adnan Arefeen, Biplob Debnath, and Srimat Chakradhar. 2024. Leancontext: Cost-efficient domain-specific question answering using llms. *Natural Language Processing Journal*, 7:100065.

Akari Asai, Jungo Kasai, Jonathan H Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021a. Xor qa: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hanna Hajishirzi. 2021b. One question answering model for many languages with cross-lingual dense passage retrieval. *Advances in Neural Information Processing Systems*, 34:7547–7560.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2023. Dense x retrieval: What retrieval granularity should we use? *Preprint*, arXiv:2312.06648.

Xi Chen, Li Wang, MingKe You, WeiZhi Liu, Yu Fu, Jie Xu, Shaiting Zhang, Gang Chen, Kang Li, and Jian Li. 2024. Evaluating and enhancing large language models' performance in domain-specific medicine: Explainable llm with docoa. *Journal of medical Internet research*.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *Preprint*, arXiv:2306.16092.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia,

Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li,

Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. *Preprint*, arXiv:2309.15217.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *Preprint*, arXiv:2302.04166.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Seraphina Goldfarb-Tarrant, Pedro Rodriguez, Jane Dwivedi-Yu, and Patrick Lewis. 2024. Multicontrievers: Analysis of dense retrieval representations. *arXiv preprint arXiv:2402.15925*.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Juncai He, Ziche Liu, Zhiyi Zhang, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. Acegpt, localizing large language models in arabic. *Preprint*, arXiv:2309.12053.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ye Liu, Kazuma Hashimoto, Yingbo Zhou, Semih Yavuz, Caiming Xiong, and S Yu Philip. 2021. Dense hierarchical retrieval for open-domain question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 188–200.

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

Minh Thuan Nguyen, Khanh Tung Tran, Nhu Van Nguyen, and Xuan-Son Vu. 2023. Vigptqa-state-of-the-art llms for vietnamese question answering: system overview, core models training, and evaluations. In *Proceedings of the 2023 conference on empirical methods in natural language processing: industry track*, pages 754–764.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Mobashir Sadat, Zhengyu Zhou, Lukas Lange, Jun Araki, Arsalan Gundroo, Bingqing Wang, Rakesh Menon, Md Parvez, and Zhe Feng. 2023. DelucionQA: Detecting hallucinations in domain-specific question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 822–835, Singapore. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Alham Fikri Aji, Zhengzhong Liu, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Fangkai Yang, Pu Zhao, Zezhong Wang, Lu Wang, Jue Zhang, Mohit Garg, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. 2023. Empower large language model to perform better on industrial domain-specific question answering. *Preprint*, arXiv:2305.11541.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

7 Appendix

7.1 Evaluation Metrics

BERTScore: BERTScore (Zhang et al., 2020) utilizes BERT embeddings (Devlin et al., 2019) to measure the semantic similarity between generated text and the reference text. It is calculated as:

$$R_{BERT} = \frac{\sum_{y \in Y_x} \max_{z \in Z_x} \cos(\mathbf{y}, \mathbf{z})}{|Y_x|} \tag{6}$$

where Y_x and Z_x are the token embeddings of the reference and generated text for example x, respectively, and $\cos$ denotes the cosine similarity.

BLEU Score: BLEU (Papineni et al., 2002) compares n-grams of the machine-generated text to n-grams of the reference text, calculating the precision for each, and then applying a brevity penalty to penalize short translations:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{7}$$

where p_n is the precision of n-grams, w_n are weights summing to 1, BP is the brevity penalty, and N is typically 4.

ROUGE-L: ROUGE-L (Lin, 2004) measures the longest common subsequence (LCS) between the generated text and the reference, focusing on the sequence order:

$$\text{ROUGE-L} = \frac{(1+\beta^2) \cdot \text{Precision}_{LCS} \cdot \text{Recall}_{LCS}}{\text{Recall}_{LCS} + \beta^2 \cdot \text{Precision}_{LCS}} \tag{8}$$

where β is typically set to prioritize recall more than precision.

F-1 Score The F-1 Score, as used in evaluating question answering systems like SQuAD (Rajpurkar et al., 2016, 2018), quantifies the overlap between the predicted and reference answers:

$$F\text{-}1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

where Precision is the ratio of overlapping words in the predicted answer to the total number of words in the predicted answer, and Recall is the ratio of overlapping words in the predicted answer to the total number of words in the reference answer.

Table 8: Selected UAE Government Websites with FAQ Sections

Website	Description
UAE Government Portal	The official portal of the UAE government, providing a wide range of general information and services about UAE. www.u.ae
MOFA	The Ministry of Foreign Affairs, offering information on international relations, consular services, and diplomatic missions. www.mofa.gov.ae
GPSSA	The General Pension and Social Security Authority, providing details on pension schemes and social security benefits. www.gpssa.gov.ae
MOHRE	The Ministry of Human Resources and Emiratisation, covering labor laws, employment services, and workforce regulations. www.mohre.gov.ae
MOJ	The Ministry of Justice, offering legal information, judicial services, and legislative updates. www.moj.gov.ae
FTA	The Federal Tax Authority, providing guidelines on tax regulations, compliance, and e-services. www.tax.gov.ae
MOEC	The Ministry of Economy, including information on economic policies, business regulations, and trade. www.moec.gov.ae
MOEC Investment	Investment Section from the Ministry of Economy, offering insights into investment opportunities and regulations. https://www.moec.gov.ae/en/investment-faqs
MOE	The Ministry of Education, covering educational policies, school regulations, and academic services. www.moe.gov.ae
ESE	The Emirates Schools Establishment, focusing on school management, educational resources, and student services. www.ese.gov.ae
FAHR	The Federal Authority for Government Human Resources, providing information on HR policies, employee services, and training programs. www.fahr.gov.ae
EHS	The Emirates Health Services, offering healthcare services, medical guidelines, and public health information. www.ehs.gov.ae
MOIAT	The Ministry of Industry and Advanced Technology, including information on industrial policies, technological advancements, and innovation. www.moiat.gov.ae
MOEI	The Ministry of Energy and Infrastructure, covering energy policies, infrastructure projects, and sustainability initiatives. www.moei.gov.ae
MOCCAE	The Ministry of Climate Change and Environment, providing information on environmental policies, climate initiatives, and agricultural services. www.moccae.gov.ae