IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 January 2025, accepted 16 February 2025, date of publication 21 February 2025, date of current version 7 March 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3544408

RESEARCH ARTICLE

ICCA-RAG: Intelligent Customs Clearance Assistant Using Retrieval-Augmented Generation (RAG)

RONG HU[1], SEN LIU[2], PANPAN QI[3], JINGYI LIU[4], AND FENGYUAN LI[5]

[1]Customs and Public Management College, Shanghai Customs University, Shanghai 201204, China
[2]Department of Electronic Information, Shanghai Dianji University, Shanghai 201306, China
[3]Information Department, Xinglin College, Nantong University, Nantong 226236, China
[4]School of Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[5]Xi'an Jiaotong University, Xi'an 710049, China

Corresponding author: Rong Hu (hurong@shcc.edu.cn)

This work was supported in part by the General Administration of Customs under the Project titled "Research on the Customs Supervision System of China-Europe Railway Express from the Perspective of International Supply Chain Hubs" under Project 2021HK248.

ABSTRACT Document processing and query generation tasks in customs declaration scenarios face key challenges such as the complexity of multimodal data, adaptability to dynamic regulations, and ambiguity in query semantics. This study proposes a Retrieval-Augmented Generation system (ICCA-RAG) that addresses the core issues of processing complex customs documents and dynamically generating queries through multimodal document parsing, sparse-dense hybrid storage, and context-driven large language model generation. In terms of multimodal document parsing, the system supports comprehensive parsing of PDFs, images, tables, and text, which are uniformly transformed into semantic vectors and keyword indices for hybrid storage. By combining the retrieval and generation modules, the ICCA-RAG system achieves significant improvements in contextual relevance and generation accuracy. Compared to traditional methods, the ICCA-RAG system demonstrates a 20.1% increase in answer correctness, a 15.3% increase in answer relevancy, and an 18.7% increase in the faithfulness of generated content, with outstanding performance in noisy query scenarios. The research findings validate the ICCA-RAG system's advancement and applicability in handling complex document processing and professional domain question-answering tasks, while also providing a transferable technical framework for other fields, such as law and healthcare.

INDEX TERMS Customs declaration assistance, dynamic regulation adaptation, intelligent question-answering system, large language model (LLM), multimodal document parsing, retrieval-augmented generation (RAG), semantic retrieval.

I. INTRODUCTION

Foreign trade occupies an indispensable position in global economic development, serving as a crucial link and driving force for economic connections between nations. With the continuous deepening of globalization, international trade has become a key factor in enhancing economic strength and improving people's living standards worldwide. Customs clearance, as a core aspect of foreign trade activities, refers to the process of preparing and submitting necessary documents to facilitate the import and export of goods through customs. It involves not only verifying the accuracy of trade-related documents and determining applicable tariffs but also ensuring compliance with customs regulations and handling related duties and fees. The efficiency and accuracy of customs clearance directly affect the smoothness of international trade, playing a pivotal role in the stability of global supply chains and economic growth.

The customs clearance process generally includes several steps, such as preparing customs declarations, submitting supporting documents to customs authorities, undergoing customs reviews, inspections, paying related taxes and duties,

The associate editor coordinating the review of this manuscript and approving it for publication was Anandakumar Haldorai.

and obtaining clearance for goods [1], [2]. Due to the involvement of complex customs regulations, commodity classification and valuation, and origin determination for tariff calculations, the entire process is intricate and time-consuming [3]. For small and medium-sized enterprises (SMEs), the complexity of customs clearance procedures often becomes a major obstacle to exporting goods. Many companies, especially smaller manufacturers and traders, often lack the professional knowledge, skills, and resources required for customs clearance and therefore rely on freight forwarders to handle these procedures [4]. While general freight forwarders may be knowledgeable about routine customs operations, they often face challenges when dealing with increasingly complex trade patterns, emerging goods, and upgrades to customs facilitation measures and free trade agreements. Leading freight forwarders with expertise in customs regulations can provide efficient document preparation and logistics support, but their services are often prohibitively expensive for smaller enterprises. This makes it challenging for these businesses to complete customs clearance independently. Moreover, the quality of customs services provided by standard freight forwarders is difficult to control and monitor.

In addition, customs declarations and related documents are the primary records required for export tax rebates. Any errors or deficiencies in these documents can directly impact the tax rebate benefits of enterprises, resulting in unfavorable outcomes, such as downgrades in their credit ratings with customs authorities. Under the framework of the Authorized Economic Operator (AEO) mutual recognition system between Chinese and foreign customs, such downgrades can negatively affect a company's credit rating with foreign customs authorities, further influencing its customs clearance convenience both domestically and internationally. In more severe cases, document-related issues may lead to smuggling investigations and even criminal cases [5]. In summary, due to the complexity of customs operations and the lack of skills and knowledge among SMEs, these businesses face a series of challenges, including high service fees, tax rebate risks, credit downgrades, and even anti-smuggling investigations. Fig. 1 illustrates the customs clearance process.

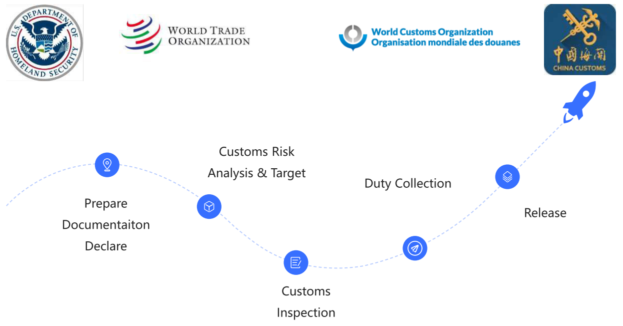

WORLD TRADE ORGANIZATION
World Customs Organization
Organisation mondiale des douanes
中国海关
CHINA CUSTOMS
Customs Risk Analysis & Target
Duty Collection
Release
Prepare Documentaiton Declare
Customs Inspection

FIGURE 1. **The customs clearance process, including duty collection, risk analysis, documentation preparation, declaration, inspection, and release.**

The complexity of customs clearance primarily stems from three key aspects [5], [6]: the diversity of document sources, the high demand for compliance expertise, and the cumbersome nature of the procedures. These factors collectively pose significant challenges in international trade. First, the wide range of required documents for customs clearance encompasses tariffs, trade regulations, inspection and quarantine ordinances, and customs regulations of various countries, with information scattered across multiple organizations. For instance, in the United States, inspection and quarantine regulations are managed separately by the Department of Agriculture and the Department of Health and Human Services, making it difficult for businesses to obtain comprehensive supply chain compliance information through a single channel [7]. Moreover, the complexity and frequent updates of these documents further increase the difficulty of integration. For example, an exporter of frozen foods must invest considerable resources to meet the requirements of food safety standards, cold chain transportation regulations, and specific tariff reduction policies in the target country.

Dual-use items, which can serve both civilian and military purposes, present another example. These goods and technologies are subject to strict export controls due to their potential use in producing weapons of mass destruction (WMD) or other military applications. Exporters must comply with these regulations, which often require specialized legal knowledge. Misclassification of dual-use items can lead to significant legal risks, including criminal liability, underscoring the necessity for careful compliance.

Second, the accurate completion of customs declarations demands extensive compliance expertise, particularly in the precise classification of goods using Harmonized System (HS) codes. As a globally accepted commodity classification system, the complexity of HS codes makes the classification process highly challenging. For instance, certain electronic components may fall into multiple categories, and even minor errors can result in incorrect tariffs or goods being detained. For example, the specific HS code for exporting LED lamps varies depending on their intended use, requiring experience and expertise for accurate classification.

In 2015, the United States witnessed a case involving the misclassification of hazardous goods. A chemical company attempted to export a shipment of chemicals classified by the International Civil Aviation Organization (ICAO) and the International Maritime Organization (IMO) as hazardous materials. However, the company incorrectly declared these chemicals as non-hazardous in its export documentation and failed to comply with the stringent transportation and packaging requirements for hazardous materials. As a result of this misclassification, the shipment lacked the necessary safety measures during transportation, significantly increasing the risk of accidents. Fortunately, the shipment was flagged and intercepted by customs officers during a routine inspection before exportation. The company faced severe legal consequences, including hefty fines and the seizure of the goods. Moreover, the company's executives were subjected to criminal charges for violating federal hazardous materials transportation regulations.

This case highlights the critical importance of accurate classification and declaration of hazardous goods, as well as the severe consequences of non-compliance. It underscores that proper product classification is not only essential for trade compliance but also for ensuring public safety and environmental protection. Therefore, exporters must ensure the accuracy of product classifications to avoid potential repercussions, including legal liabilities and safety risks.

Beyond HS codes, technical barriers to trade [5] (TBT) present another critical aspect of compliance. TBTs refer to regulatory measures designed to protect national security, public health and safety, the lives and health of animals and plants, and the environment. These barriers, implemented through technical regulations, standards, and conformity assessment procedures, create obstacles for foreign goods and services entering domestic markets. TBTs often take the form of technical requirements, and a lack of in-depth understanding of these regulations can result in severe consequences for customs clearance.

The complexity and evolving nature of TBTs significantly increase uncertainty during customs clearance. Businesses may struggle to predict the outcomes of their clearance processes due to changing technical requirements. Additionally, compliance with TBTs necessitates the submission of extensive documentation, such as proof that products meet specific safety and quality standards. Companies also face heightened compliance risks, as any failure to meet TBT requirements could result in the detention or return of their goods.

Finally, the customs clearance process itself is highly cumbersome, involving steps such as document preparation, submission, customs review, inspection, tax payment, and goods release. Each step can encounter obstacles due to minor details. For example, a small furniture manufacturer might face delays in goods release due to incomplete proof of the origin of the wood used, leading to disruptions in delivery timelines and potential damage to the company's reputation.

To address these challenges, we propose an Intelligent Customs Clearance Assistant (ICCA-RAG) that combines large language models (LLMs) and Retrieval-Augmented Generation (RAG) technologies to provide comprehensive customs clearance support for enterprises. LLMs, based on deep learning, excel in understanding and generating natural language text with outstanding semantic comprehension capabilities, making them well-suited for complex language tasks. Meanwhile, RAG integrates document retrieval with generation models, enabling the precise extraction of relevant information from extensive documents and generating highly context-relevant answers.

The ICCA-RAG system leverages RAG technology to semantically retrieve information from various documents (such as tariffs, inspection and quarantine ordinances, and customs regulations), efficiently consolidating and extracting key information. By harnessing the generative capabilities of LLMs, it provides users with comprehensive guidance on customs clearance procedures and suggestions for completing documents. The system not only reduces the technical barriers of the customs clearance process but also significantly enhances operational efficiency, offering an intelligent and user-friendly solution for SMEs involved in international trade.

The main contributions of this study are as follows:

- This study is the first to apply large language models and RAG technologies to customs clearance, pioneering the exploration of intelligent solutions in this field.
- This paper proposes ICCA-RAG, which integrates multimodal document parsing with a retrieval-augmented generation approach to efficiently process complex customs documents and dynamically generate queries.
- It systematically organizes and open-sources a complete dataset of Chinese customs materials, including tariff chapters, customs laws, and inspection and quarantine ordinances, providing a high-quality data foundation for future research.

II. RELATED WORKS

This section reviews the studies relevant to this research, including retrieval-augmented generation (RAG) techniques, the applications of large language models (LLMs) in specialized domains, and the development of intelligent customs clearance support systems.

A. APPLICATIONS OF LARGE LANGUAGE MODELS (LLMS) IN SPECIALIZED DOMAINS

LLMs [8] represent an advanced evolution of language models, typically based on Transformer architectures and comprising billions or even hundreds of billions of parameters [8], [9]. A robust LLM generally possesses several core capabilities [10]: natural language understanding, natural language generation, contextual awareness, and instruction-following ability. In recent years, various LLMs have been released, achieving widespread adoption [11]. Representative models include OpenAI's ChatGPT [12], Meta AI's LLama [13], and Databricks' Dolly [14]. The rapid development of LLMs has demonstrated their exceptional capabilities in a variety of fields, such as search engines [15], [16], customer service [17], and machine translation—domains [18], [19] directly related to natural language processing. Additionally, LLMs have found broad applications in other general-purpose scenarios, such as code generation [20], [21], healthcare [22], [23], finance [24], [25], and education [26].

B. RETRIEVAL-AUGMENTED GENERATION (RAG) TECHNIQUES

While LLMs have shown remarkable capabilities across diverse domains, they also exhibit significant limitations [27], especially when addressing tasks in specialized or knowledge-intensive fields. These limitations arise primarily when problems exceed the scope of their training data or require up-to-date domain-specific information, leading to a decline in output accuracy. This phenomenon is commonly referred to as "hallucination." [12], [28], [29] Although increasing model size can alleviate this issue to some extent, it cannot fundamentally eliminate hallucinations [29], [30]. RAG techniques [31] address these challenges effectively by incorporating external domain-specific knowledge bases and performing retrieval tasks to supplement the generation process. This approach significantly reduces factual inaccuracies in the outputs of LLMs [32].

A typical RAG system consists of three key components [33]: indexing, retrieval, and generation. By following this standardized workflow, RAG systems mitigate hallucinations to some degree. However, they often suffer from limitations such as low precision and redundant information [34]. To overcome these drawbacks, advanced RAG systems have been developed, which improve performance in several ways. During the indexing phase, they optimize data embeddings to enhance the quality of indexed content [35]. In the retrieval phase, they determine relevant content by calculating the similarity between query content and text blocks, involving techniques such as fine-tuning embedding models [36] and learning dynamic embeddings for various content types [37]. In the generation phase, retrieved content is merged with the user query and used as prompts for the LLM, with mechanisms to reorder relevant content based on relevance or compress prompts to address context window limitations [37], [38], [39], [40].

Additionally, Asai et al. proposed a Self-RAG mode [41] that introduces a new module to assess whether retrieval is necessary and evaluate the relevance of retrieved content. This innovation enhances the meaningfulness of generated content [41]. Modular RAG systems differ structurally from traditional RAG by integrating external modules such as search modules [42], memory modules [43], [44], adjustment modules [45], [46], and task adapters to improve performance. Further research has optimized retrieval module performance by adopting dense vector retrieval methods [35], [46] (e.g., DPR [37] and FAISS [47], [48]) instead of traditional sparse methods [49] like BM25 [50], [51], [52]. These advancements have significantly enhanced the effectiveness of RAG systems in tasks such as document summarization, multi-turn dialogue, and intelligent search [27], [53].

C. INTELLIGENT CUSTOMS CLEARANCE SUPPORT SYSTEMS

With the increasing complexity of global trade and the standardization of customs processes, intelligent customs clearance systems have emerged as a research focus. Traditional customs systems typically rely on static knowledge bases or expert systems to support tariff calculations, HS code classification, and document compliance. However, these systems face notable limitations, particularly in handling dynamically updated regulations and multimodal data efficiently [54].

In recent years, customs clearance support systems based on machine learning have gained prominence. For example, some studies have utilized deep learning techniques to predict and classify HS codes, significantly improving the efficiency and accuracy of complex product categorization. Other research efforts have focused on optimizing customs processes by leveraging automation tools to reduce the complexity of document handling. However, these systems still exhibit critical shortcomings. First, most systems only support single-text-type processing, lacking comprehensive capabilities for multimodal document analysis. Second, existing systems fail to provide context-sensitive generation support in dynamic query scenarios.

In this study, we combine LLMs with RAG techniques to explore novel applications in the customs domain, pioneering intelligent solutions for this field.

III. PRELIMINARIES

A. PROBLEM DEFINITION

Customs clearance, as a critical step in international trade, poses significant challenges for small and medium-sized enterprises (SMEs) due to the complexity of information sources, the high demand for professional knowledge, and the cumbersome nature of the processes. On one hand, the required customs documents are dispersed across various institutions and frequently updated, making information integration challenging. On the other hand, tasks such as HS code classification require extensive expertise, where even minor errors can result in incorrect tariffs or shipment delays. Furthermore, the customs clearance process consists of multiple steps, each vulnerable to disruptions caused by small mistakes. These challenges leave SMEs unable to afford professional freight forwarding services while lacking the capacity to independently manage customs clearance, thereby severely impacting their competitiveness in international markets.

B. DESIGN RATIONALE

The complexity of customs clearance is primarily rooted in three areas: (1) dispersed and frequently updated information, (2) the difficulty of HS code classification, and (3) the error-prone nature of multistep processes. To address these challenges, this paper introduces an Intelligent Customs Clearance Assistant system that employs innovative technologies in semantic processing, semantic retrieval for commodity classification, and comprehensive process guidance. The following subsections elaborate on these components.

1) INFORMATION DISPERSAL AND DYNAMIC UPDATES

Customs-related information originates from diverse sources, including tariffs, inspection and quarantine ordinances, and customs regulations. These sources often present unstructured data distributed across multiple institutions, lacking unified standards. Additionally, frequent updates to this information further complicate access and usability for enterprises. Traditional solutions typically rely on static knowledge bases or manual updates, which are inadequate for dynamic regulatory environments. To overcome these limitations, this study constructs a multimodal semantic data pipeline. By leveraging multimodal document understanding techniques, scattered information is transformed into searchable semantic vector representations and stored in a unified semantic database. This system efficiently retrieves the latest, most relevant information from extensive datasets, providing dynamic support for enterprises.

2) HS CODE CLASSIFICATION CHALLENGES

HS code classification, a cornerstone of customs clearance, involves complex rules dependent on the characteristics and applications of specific goods. Traditional rule-based methods and manual operations often fail to meet the precision required for accurate classification, leading to tariff calculation errors or shipment delays. To address this, this study develops a semantic retrieval method. By semantically parsing user-input commodity information, the system extracts key features and matches them with HS code entries in the semantic vector database. This method intelligently recommends the most appropriate codes, significantly enhancing classification accuracy and efficiency.

3) CUMBERSOME AND ERROR-PRONE PROCESSES

The customs clearance process involves multiple steps, including document preparation, submission, payment of duties and taxes, and goods release. Errors or unfamiliarity with the process can lead to delays or detentions at any stage. Existing systems often focus on specific steps without offering holistic, end-to-end support. To address this, this study implements an intent recognition mechanism based on large language models (LLMs). This mechanism identifies the user's current stage in the customs clearance process and provides personalized guidance, such as required documents, key considerations, and operational suggestions, enabling users to efficiently complete tasks.

In summary, the proposed system integrates a multimodal semantic pipeline for information consolidation, semantic retrieval for enhanced commodity classification, and end-to-end guidance using intent recognition. These innovations collectively create a dynamic, efficient, and intelligent customs support system, reducing technical barriers and empowering SMEs to participate effectively in international trade.

C. SYSTEM OVERVIEW AND ARCHITECTURE

The architecture of the proposed Intelligent Customs Clearance Assistant system is depicted in Fig. 2. It demonstrates the multimodal document vectorization and query-driven contextual retrieval pipelines.

The system employs two main pipelines to facilitate multimodal document processing and query response generation.

1) PIPELINE 1: MULTIMODAL DOCUMENT VECTORIZATION

The first pipeline handles the vectorization of multimodal documents. Supporting various document formats (e.g., DOC, PDF, and images), the system extracts content using OCR or document parsing tools (e.g., LibreOffice and PyMuPDF), converting it into text. The text is then logically segmented and transformed into high-dimensional embeddings using semantic vectorization models. These embeddings, along with their corresponding text, are stored in a hybrid storage module. Dense storage preserves the vectorized representations for semantic retrieval, while sparse storage retains the original text and tokenized results for keyword searches. This pipeline outputs a structured semantic storage system, enabling efficient query processing in subsequent steps.

2) PIPELINE 2: CONTEXT-BASED QUERY RETRIEVAL AND RESPONSE GENERATION

The second pipeline focuses on context-driven query retrieval and response generation. User queries are first semantically rewritten by a large language model to create optimized expressions for retrieval. The rewritten query is then matched with vectorized representations in dense storage to identify the most relevant text blocks. Sparse storage supplements these results with keyword-based information to enrich the retrieved content. The retrieved texts are structured into a context prompt, which, along with the user's query, forms the input for the large language model to generate responses. This process seamlessly integrates user queries with document content, achieving accurate and context-aware responses.

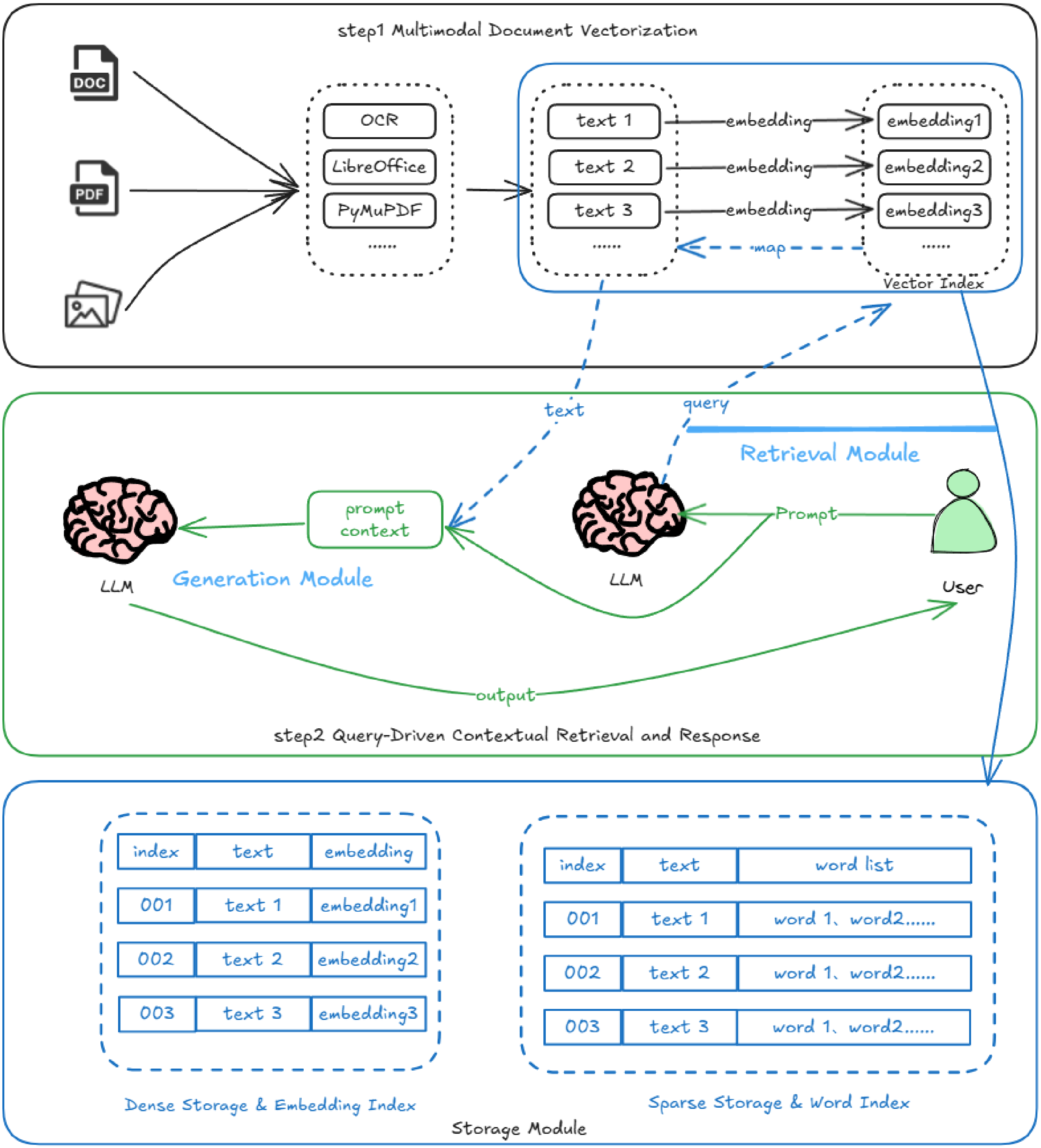
IV. INTELLIGENT CUSTOMS CLEARANCE ASSISTANT USING RETRIEVAL-AUGMENTED GENERATION

A. MULTIMODAL DOCUMENT VECTORIZATION

1) MULTIMODAL DOCUMENT PARSING AND PREPROCESSING

This study designs a unified parsing framework tailored to the diversity of multimodal documents, capable of processing various file formats, including DOCX, PDF, images (IMG), and Excel spreadsheets. Parsing methods are specifically designed for each format, based on their structural characteristics and content storage approaches, to extract structured text and assemble a collection suitable for subsequent processing.

The logic of the entire parsing process is clearly defined in Algorithm 1. The pseudocode provides details on selecting appropriate tools and methods based on document types (DOCX, PDF, IMG, or Excel). For instance, `python-docx` is utilized to parse paragraphs, headings, and table data from DOCX files; `PyMuPDF` is applied to extract textual content and layout structures from PDFs; `Tesseract OCR` processes images to extract embedded text; and Excel tables are converted into Markdown format for unified management. All parsing results are further segmented: long



index	text	embedding
001	text 1	embedding1
002	text 2	embedding2
003	text 3	embedding3

index	text	word list
001	text 1	word 1、word2......
002	text 2	word 1、word2......
003	text 3	word 1、word2......

FIGURE 2. System architecture of the intelligent customs clearance Assistant (ICCA-RAG). The architecture includes two main pipelines: multimodal document vectorization and query-driven contextual retrieval and response. The system employs dense and sparse storage modules to enable efficient semantic and keyword-based retrieval.

documents are divided into sections or paragraphs, while short documents are stored as single text units.

Finally, all parsing results are consolidated into a text collection T, defined as:

$$T = \{t_{i,j} \mid d_i \in D,\ t_{i,j} \text{ is the } j\text{-th section or paragraph of } d_i\} \quad (1)$$

where $D = \{d_1, d_2, \ldots, d_m\}$ represents the set of input documents, and $t_{i,j}$ refers to the j-th textual unit (e.g., section or paragraph) extracted from document d_i. If d_i is a short document, then $t_{i,j} = d_i$. This text collection serves as a unified input for subsequent semantic vectorization modules.

Algorithm 1 Unified Multimodal Parsing Workflow

```
Require: Document $D$ (DOCX, PDF, IMG, Excel)
Ensure: Text collection $T = \{t_1, t_2, \ldots, t_n\}$
Initialize $T \leftarrow \emptyset$
Identify document type: type $\leftarrow$ GetDocumentType($D$)
if type = "DOCX" then
    content $\leftarrow$ ParseDocx($D$) {Use python-docx to extract content}
    for all elements in content do
        if element is "Paragraph" or "Title" then
            $t \leftarrow$ ExtractText($element$)
            Append $t$ to $T$
        else if element is "Table" then
            table_data $\leftarrow$ ParseTable($element$)
            $t \leftarrow$ FormatTableAsMarkdown($table\_data$)
            Append $t$ to $T$
        else if element is "Image" then
            img_text $\leftarrow$ OCR($element$) {Use Tesseract OCR}
            Append img_text to $T$
        end if
    end for
else if type = "PDF" then
    content $\leftarrow$ ParsePdf($D$) {Use PyMuPDF to extract text and structure}
    for all pages in content do
        page_text $\leftarrow$ ExtractText($page$)
        Append page_text to $T$
        for all images in page do
            img_text $\leftarrow$ OCR($image$) {Use Tesseract OCR}
            Append img_text to $T$
        end for
    end for
else if type = "IMG" then
    img_text $\leftarrow$ OCR($D$) {Use Tesseract OCR to extract text}
    Append img_text to $T$
else if type = "Excel" then
    table_data $\leftarrow$ ParseExcel($D$) {Use pandas to extract table data}
    $t \leftarrow$ FormatTableAsMarkdown($table\_data$)
    Append $t$ to $T$
end if
if Length($T$) > Threshold then
    $T \leftarrow$ SplitTextBySections($T$) {Split long text into sections or paragraphs}
end if
return $T$
```

2) SEMANTIC INDEXING AND HYBRID STORAGE ARCHITECTURE

After completing the parsing and segmentation of multimodal documents, each text unit in the collection is stored using both sparse and dense storage methods to enable efficient retrieval. Sparse storage tokenizes the text and constructs inverted indexes for keyword-based searches, while dense storage leverages pre-trained semantic models to generate high-dimensional embeddings, facilitating semantic similarity-based retrieval.

a: SPARSE STORAGE

Sparse storage employs an inverted index architecture to support efficient keyword-based retrieval. It consists of three main components: keyword extraction and normalization, statistical information generation, and inverted index construction.

Keyword Extraction and Normalization: Text units are processed for keyword extraction and standardization. This includes removing stopwords, normalizing case, and performing stemming. During this process, each text unit is decomposed into a set of normalized keywords, which form the basis of the inverted index. Special symbols such as numbers, proper nouns, and other contextually relevant tokens are retained to support diverse query scenarios.

Statistical Information Generation: For each keyword, the system calculates its term frequency (TF) within text units and document frequency (DF) across the collection. Additionally, the length of each text unit is recorded for use in retrieval algorithms like BM25. These statistical metrics form the foundation for constructing the inverted index and support subsequent retrieval scoring.

Inverted Index Construction: Based on the processed results, an inverted index is constructed. This index maps keywords to the text units in which they appear and records metadata such as term frequency (TF) and text length. For example, if the keyword "data" appears in text units t_1 and t_3, the corresponding inverted index entry would be:

```
``data'': [
{``id'': t_1, ``tf'': 3, ``length'': 100},
{``id'': t_3, ``tf'': 1, ``length'': 150}
]
```

The sparse storage module consists of two core components: the inverted index and a document metadata table. The inverted index maps keywords to text units for fast retrieval, while the metadata table stores global information about each text unit, such as document sources and section numbers. During retrieval, user queries undergo the same tokenization and preprocessing steps to generate a set of keywords, which the system uses to quickly locate relevant text units and provide candidate results.

b: DENSE STORAGE

Dense storage is designed to handle semantic-level retrieval by generating and storing high-dimensional embeddings for each text unit. The pre-trained Bge-m3 model is employed to generate fixed-dimensional embeddings for the parsed text units. Specifically, for each text unit $t_{i,j}$, the semantic embedding is generated as follows:

$$v_{i,j} = f_{\text{Bge-m3}}(t_{i,j}) \tag{2}$$

These embeddings preserve the deep semantic information of the text, ensuring that semantically related text units are closer in vector space (e.g. measured by Euclidean distance) or exhibit higher similarity scores (e.g., cosine similarity). The Bge-m3 model was selected for its superior performance in multi-language and multi-domain text embedding tasks, ensuring both accuracy and generalization.

The generated embeddings are stored in a vector storage module implemented using FAISS (Facebook AI Similarity Search). FAISS is an efficient open-source library designed for handling high-dimensional vector search tasks. In this study, FAISS is used to construct approximate nearest neighbor (ANN) indexes, enabling fast retrieval from large-scale collections of semantic embeddings. Specifically, this study adopts the Hierarchical Navigable Small World (HNSW) algorithm for index construction (see Algorithm 2). HNSW achieves a balance between retrieval speed and accuracy by constructing hierarchical graph structures that efficiently identify the vectors that are the most similar to a query vector in a high-dimensional space.

Dense storage also maintains a mapping between embeddings and their corresponding text units, enabling quick access to the textual content upon retrieval. Additionally, document metadata (e.g., document ID and section number) is stored to support contextual backtracking and result ranking.

Algorithm 2 HNSW Index Construction

```
Require: Set of dense embeddings $V = \{v_1, v_2, \ldots, v_n\}$, maximum level $L_{\max}$, maximum neighbors per layer $M$
Ensure: HNSW Index
Initialize an empty graph $G$ with $L_{\max}$ layers
Set entry point $ep \leftarrow$ NULL
for each embedding $v_i$ in $V$ do
    Randomly assign a maximum layer level $l_i$ for $v_i$: $l_i \leftarrow$ RandomLevel($L_{\max}$)
    if $ep =$ NULL then
        Set $ep \leftarrow v_i$ {Assign the first point as the entry point}
        Add $v_i$ to all layers of $G$ up to level $l_i$
        continue
    end if
    for $l = L_{\max}$ to 0 do
        if $l > l_i$ then
            continue {Skip higher layers for this point}
        end if
        NearestNeighbors $\leftarrow$ SearchKNN($ep, v_i, l, M$) {Find nearest neighbors in the current layer}
        Connect $v_i$ to the nearest neighbors in layer $l$
        Update $ep$ with the nearest neighbor found
    end for
end for
return $G$
```

B. QUERY-DRIVEN CONTEXTUAL RETRIEVAL AND RESPONSE

1) QUERY REFORMULATION USING LARGE LANGUAGE MODELS (LLMS)

User queries are typically expressed in natural language, which may contain semantic ambiguities, verbose expressions, or insufficient keywords. Directly using such queries for retrieval may negatively impact relevance and recall rates. To enhance retrieval performance, this study employs pre-trained Large Language Models (LLMs) to refine user queries. By optimizing the semantic representation of the queries, high-quality queries better suited for retrieval modules are generated. The reformulated queries are semantically more precise and exhibit higher adaptability to retrieval systems.

The query reformulation process can be formally described as follows: Given an initial query q_{input}, the LLM model f_{LLM} outputs an optimized query q_{query}:

$$Q = \left\{ \begin{array}{l} \text{``documents'', ``required'', ``customs'',} \\ \text{``clearance'', ``electronic'', ``imported''} \end{array} \right\} \tag{3}$$

Here, q_{query} is a concise phrase or sentence resulting from semantic refinement and optimization. It retains the core semantic information of the user query and is tailored to both sparse and dense retrieval modules.

To achieve high-quality query reformulation, this study designs a specific prompt engineering strategy that fully leverages the semantic understanding and generation capabilities of LLMs. The prompts explicitly define the task requirements and output format, guiding the model to produce queries aligned with retrieval objectives. After generating q_{query}, lightweight post-processing is applied to normalize the output, such as reducing the use of stop-words, standardizing case, and cleaning special characters.

Through query reformulation, ambiguities and redundancies in the original queries are effectively eliminated. The resulting q_{query} significantly improves semantic clarity and retrieval adaptability. For example, the user query *"What documents are required for customs clearance of electronic products imported from China?"* can be reformulated into *"Required documents for customs clearance of electronic imports from China."* This reformulation retains semantic completeness while optimizing conciseness and keyword matchability.

a: EXAMPLE PROMPT

- **Task Description:** Rewrite the following query to make it concise and suitable for retrieval. Retain the core semantic meaning while removing redundant expressions or ambiguous terms. The output should be a single sentence focusing on the key concepts in the query.

- **Input Query:** *"What are the challenges of combining sparse and dense storage in a retrieval system?"*
- **Output Format:** *"Challenges of combining sparse and dense storage in retrieval systems."*

2) CONTEXTUAL RETRIEVAL FROM SPARSE AND DENSE STORAGE

To efficiently retrieve content relevant to user queries from large-scale storage, this study integrates sparse and dense retrieval methods. Sparse retrieval enables precise keyword matching, while dense retrieval leverages semantic vector similarity to capture semantically related but lexically different content. By combining these two approaches, the system constructs richer and more relevant contexts to support downstream generative models.

a: SPARSE RETRIEVAL

Sparse retrieval relies on the inverted index structure within the sparse storage module to achieve efficient keyword matching. Sparse storage records keyword distribution information for text units, including term frequency (TF) within a text unit and document frequency (DF) across the collection. This structure enables the system to quickly locate text units containing query keywords and rank them based on relevance.

To evaluate the relevance of a text unit to a query, the BM25 retrieval algorithm is employed. BM25 combines term frequency, inverse document frequency (IDF), and length normalization. The relevance score is calculated as:

$$\text{Score}(q,d)=\sum_{t\in q}\text{IDF}(t)\cdot\frac{\text{TF}(t,d)\cdot(k_1+1)}{\text{TF}(t,d)+k_1\cdot\left(1-b+b\cdot\frac{|d|}{\text{avgdl}}\right)},\tag{4}$$

where q represents the query, d is a text unit, $\text{TF}(t,d)$ is the term frequency of keyword t in d, and $\text{IDF}(t)$ denotes the inverse document frequency of t. Parameters k_1 and b control the balance between term frequency and length normalization.

For example, given the query *"What documents are required for customs clearance of imported electronic products?"*, sparse retrieval identifies the keyword set:

$$Q=\left\{\begin{matrix}\text{“documents”, “required”, “customs”,}\\ \text{“clearance”, “electronic”, “imported”}\end{matrix}\right\}\tag{5}$$

Using the inverted index, the system retrieves and ranks text units. For instance:

- Text unit t_1: *"Documents required for customs clearance of imported goods"* (Score = 4.3)
- Text unit t_2: *"Electronic products customs clearance process"* (Score = 3.8)

Sparse retrieval selects the highest-scoring text units as candidates for contextual construction, leveraging the efficiency of inverted indexes and BM25 scoring for fast and accurate results.

b: DENSE RETRIEVAL

Dense retrieval is based on semantic embeddings stored in the dense storage module. This approach evaluates semantic-level similarity between user queries and text units. The process begins by encoding the user query into a semantic embedding v_q using the pre-trained `Bge-m3` model. Similarly, each text unit in the dense storage module is represented as a high-dimensional vector v_i. The semantic similarity between the query and a text unit is computed using cosine similarity:

$$\text{Similarity}(v_q,v_i)=\frac{v_q\cdot v_i}{\|v_q\|\|v_i\|}.\tag{6}$$

Dense retrieval ranks text units by similarity scores and returns the top k results. Unlike sparse retrieval, which relies on exact keyword matches, dense retrieval captures deep semantic relationships, making it suitable for queries and documents with lexical differences but semantic overlap.

Due to the large size of the semantic embedding collection in dense storage, direct comparisons with every vector would be computationally expensive. To address this, the study employs FAISS (Facebook AI Similarity Search) to build approximate nearest neighbor (ANN) indexes using the Hierarchical Navigable Small World (HNSW) algorithm. HNSW efficiently manages high-dimensional vectors by constructing a layered graph structure, balancing retrieval speed and accuracy.

c: PRACTICAL EXAMPLE

For the query *"Procedures for customs clearance of machinery imports"*, the system generates a query embedding v_q using the `Bge-m3` model. The query embedding is then compared to the semantic embeddings in dense storage. The retrieved text units might include:

- Text unit t_1: *"Steps for customs clearance of imported machinery"* (Similarity Score = 0.92)
- Text unit t_2: *"Required documents for customs clearance of machinery imports"* (Similarity Score = 0.88)

These results effectively capture the core semantic meaning of the query, even if there is no direct lexical match between the query and the text.

d: INTEGRATION OF SPARSE AND DENSE RETRIEVAL

Dense retrieval complements sparse retrieval by providing semantically related text units, even when keyword overlap is minimal. By combining sparse and dense retrieval results, the system ensures a more comprehensive and contextually relevant candidate set for downstream tasks. This hybrid approach enhances the quality of contextual information used for subsequent generation processes.

3) CONTEXT CONSTRUCTION AND PROMPT ENGINEERING

In retrieval-augmented generation tasks, prompt construction serves as a critical bridge between the retrieval and generation modules. The primary objective of the prompt is to integrate user queries with retrieval results into a structured input that explicitly defines the generation task's goal. Simultaneously, the prompt ensures that the generation model effectively leverages the contextual information retrieved. This study proposes a retrieval-based prompt design method, focusing on how to extract relevant content from sparse and dense retrieval results and present it to the generation model in an appropriate structure.

To achieve high-quality contextual construction, this study selects the top 5 most relevant text units from both sparse and dense retrieval results, forming a candidate set. Sparse retrieval results are primarily based on keyword matching, capturing content directly related to the user query. Dense retrieval results, on the other hand, rely on semantic vector matching, providing broader semantic relevance. When integrating the two types of retrieval results, a redundancy removal strategy is employed to retain the higher-scoring version of semantically similar text units. The final text unit set retains the complementary advantages of sparse and dense retrieval while significantly enhancing multidimensional relevance.

The design of the prompt revolves around the user query and incorporates the following components in a structured manner:

1) **User Query:** This serves as the starting point for the generation model and clearly defines the task objective.
2) **Sparse and Dense Retrieval Results:** The results are listed separately and ranked in descending order of relevance, forming a layered contextual content set.
3) **Instructions:** Explicit instructions direct the generation model to produce responses solely based on the retrieval results, avoiding the inclusion of external knowledge or assumptions.

Such a design ensures that the generation model adheres to the provided context and produces high-quality responses strictly aligned with the retrieved information.

For instance, for the query *"What documents are required for customs clearance of machinery imports?"*, the constructed prompt is as follows:

> **Query:** What documents are required for customs clearance of machinery imports?
> **Instruction:** Use the retrieved results below to generate a concise and accurate response to the query. Only use the provided results as the basis for your response. Do not add external information or assumptions.
> **Sparse Retrieval Results:**
> 1) Required documents for machinery imports.
> 2) Documents needed for customs processing.
> 3) Customs clearance required documentation for machinery.
> 4) Necessary forms for machinery imports.
> 5) Guidelines for customs clearance documents.
>
> **Dense Retrieval Results:**
> 1) Steps for customs clearance of imported machinery.
> 2) Required documents for customs clearance of machinery imports.
> 3) Customs forms for machinery imports.
> 4) Documentation requirements for customs processing.
> 5) Procedures for customs clearance of machinery.

In this prompt, the user query defines the generation objective, and the retrieval results are categorized into sparse and dense sources, ranked by relevance. The instruction section constrains the generation model's behavior, ensuring that responses are based solely on the retrieval content. This structured design significantly reduces the risk of the generation model deviating from the context.

Based on the above prompt, the generation model effectively combines the retrieval results to produce a response. For example, given the above prompt, the generation model might generate the following response:

> **Response:** The documents required for customs clearance of machinery imports include necessary forms, customs clearance guidelines, and documentation for customs processing.

This response is accurate and highly relevant, directly derived from the retrieval results. It avoids errors caused by a lack of context or speculative reasoning by the generation model.

Through the integration of retrieval results and the refinement of prompt design, this study achieves an efficient connection between contextual construction and the generation model. The combination of sparse and dense retrieval results ensures comprehensive context coverage, while the structured prompt design further optimizes the generation model's understanding of the context and response generation process, thereby improving the accuracy and relevance of the generated output.

V. RESULTS

The goal of the Retrieval-Augmented Generation (RAG) system is to improve the accuracy, contextual relevance, and robustness of generated content by combining the strengths of retrieval and generation modules. This section provides a comprehensive evaluation of the system's performance through experiments involving proprietary and open-source models in both standard and noisy query scenarios.

The experiments were divided into two major tasks. First, the improvement in generation tasks using the RAG system was evaluated, including the correctness of answers and the ability to utilize context. Second, the system's robustness in noisy environments was tested, particularly its performance with inputs containing misspellings, incomplete expressions, or ambiguous phrasing.

To ensure comprehensive and comparative evaluation, the experiments included multiple versions of proprietary models (e.g., GPT and Claude series) and open-source models (e.g., LLaMA and Mistral series). The performance was compared under two conditions: without retrieval augmentation ("Basic") and with retrieval augmentation ("Ours"). This section highlights the experimental results and provides an in-depth analysis from the perspectives of accuracy, relevance, and efficiency.

A. EXPERIMENTAL SETUP

The experiments were conducted in a high-performance computing environment. The hardware configuration included four NVIDIA A100 GPUs, each with 80 GB of memory, to accelerate semantic vector generation and inference for the generation models. The central processor was a 64-core Intel Xeon Platinum 8358 CPU, supporting retrieval index construction and concurrent queries. The server was equipped with 256 GB of RAM and 4 TB of NVMe SSD storage to enable high-throughput semantic vector access.

For the software environment, the experiments were developed using Python 3.10. Semantic vector generation utilized the `bge-m3` model implemented with PyTorch. The retrieval module employed FAISS for efficient dense vector storage and retrieval, with HNSW indexing to optimize query efficiency. The generation module evaluated both proprietary and open-source models, including the GPT, Claude, LLaMA, and Mistral series.

The system performance was evaluated using metrics that covered both retrieval and generation aspects. For retrieval, metrics such as Precision@k, Recall@k, and Mean Reciprocal Rank (MRR) were used to assess the relevance and coverage of retrieval results. For generation, semantic-level metrics such as Answer Correctness, Answer Relevancy, and Faithfulness were used, along with traditional quality measures such as BLEU and ROUGE. To test robustness, the system was evaluated under noisy query conditions, such as inputs with misspellings and ambiguous expressions.

B. RESULTS AND ANALYSIS

1) PERFORMANCE OF PROPRIETARY MODELS

This experiment evaluated the performance of multiple proprietary models, including the GPT and Claude series, under two distinct conditions: without retrieval augmentation ("Basic") and with retrieval augmentation ("Ours"). The results clearly demonstrate that retrieval augmentation significantly improves generation quality, particularly in terms of answer correctness, relevance, and contextual utilization. The detailed experimental outcomes are presented in Figure 3.

a: ANSWER CORRECTNESS

The correctness of generated answers improved consistently across all models. For example, GPT-3.5-turbo and GPT-4-turbo showed increases in correctness scores from 0.75 and 0.73 to 0.81 and 0.79, respectively. Similarly, Claude-3-haiku improved from 0.95 to 0.97. This indicates that retrieval augmentation effectively provides accurate contextual information, addressing the limitations of the models in understanding complex queries.

b: ANSWER RELEVANCY

Retrieval augmentation improved the alignment between generated content and the provided context. Claude-3-sonnet exhibited particularly strong performance, with its relevancy score increasing from 0.96 to 0.97, while GPT-4-mini improved from 0.87 to 0.90. These results suggest that high-relevance contextual information introduced by retrieval augmentation helps models better align query intent with generated content.

c: CONTEXTUAL UTILIZATION

Significant improvements were observed in the ability to utilize context, as evidenced by metrics such as Context Precision and Context Recall. For instance, Claude-3-opus achieved a context recall of 0.99 after retrieval augmentation, while GPT-4o improved its context precision from 0.83 to 0.88. These findings validate the retrieval module's effectiveness in filtering and providing targeted contextual information, enabling the generation module to leverage more relevant inputs.

d: FAITHFULNESS

Retrieval augmentation also reduced the occurrence of hallucinated or unsupported information in generated content. Claude-3-haiku's faithfulness improved from 0.95 to 0.98, and GPT-3.5-turbo showed a similar trend, with its score increasing from 0.91 to 0.96. This enhancement is particularly crucial for tasks requiring precise reasoning based on complex contexts.

Overall, the Claude series demonstrated superior performance across all metrics, achieving near-optimal results due to its advanced semantic understanding and contextual processing capabilities. In contrast, smaller models like GPT-4-mini exhibited larger relative improvements, highlighting the effectiveness of retrieval augmentation in enhancing models with limited baseline capabilities.

e: SUMMARY

The RAG system significantly improved the generation performance of proprietary models by providing high-quality contextual input. These improvements were more pronounced in smaller models, further demonstrating the versatility and scalability of the retrieval augmentation strategy.

2) PERFORMANCE OF OPEN-SOURCE MODELS IN NOISY ENVIRONMENTS

To evaluate the impact of the Retrieval-Augmented Generation (RAG) system on the performance of open-source models in noisy environments, experiments were conducted on several open-source models, including the Mistral series, LLaMA series, and GLM4 series. The evaluation compared performance under two conditions: without retrieval augmentation ("Basic") and with retrieval augmentation ("Ours"). The experiments assessed the quality of generation, contextual utilization, and faithfulness of generated content in noisy

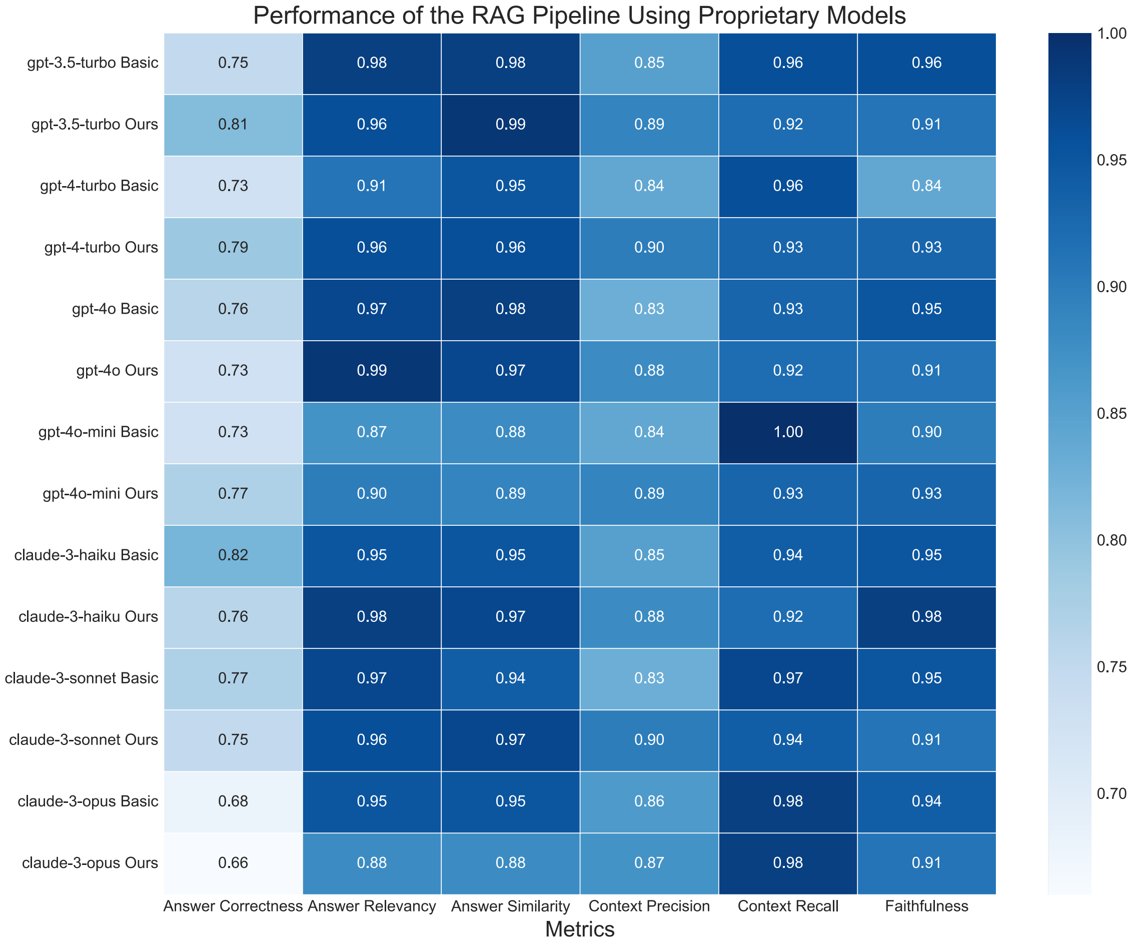


FIGURE 3. **Performance comparison of proprietary models under two conditions: "Basic" (without retrieval augmentation) and "Ours" (with retrieval augmentation). Metrics include answer correctness, relevance, and contextual utilization, highlighting the significant improvements achieved through retrieval augmentation.**

query conditions, such as misspellings, ambiguous phrasing, and incomplete expressions.

As illustrated in Figure 4, the results demonstrate that retrieval augmentation significantly enhances the performance of open-source models in addressing complex queries. These improvements are particularly notable for smaller-scale models and those with weaker baseline performance, highlighting the effectiveness of retrieval augmentation in providing essential contextual information.

a: ANSWER CORRECTNESS

Retrieval augmentation exhibited substantial improvements in the correctness of generated answers across various models. For example, the correctness score of Mistral-8 × 22b increased from 0.886 to 0.903, while LLaMA2-7b improved from 0.707 to 0.723. Larger gains were observed for smaller-scale models, such as LLaMA-8b, which saw an increase from 0.693 to 0.828. These results highlight the ability of retrieval augmentation to provide critical contextual information, enabling models to address complex or ambiguous queries with higher accuracy.

b: ANSWER RELEVANCY AND SIMILARITY

The RAG system demonstrated remarkable improvements in the relevancy and similarity of generated content to the query context. For instance, the relevancy scores of Mistral-nemo and Mistral-large-2 increased from 0.912 and 0.952 to 0.955 and 0.926, respectively. Similarly, LLaMA-13b and GLM4-9b achieved relevancy scores of 0.925 and 0.892, respectively, under retrieval augmentation. These results emphasize that retrieval augmentation can provide high-relevance contextual information, enabling models to generate responses that align more closely with the semantic intent of the query.

c: CONTEXTUAL UTILIZATION

Context precision and recall metrics showed significant improvements, particularly for models with weaker baseline performance. For example, the context recall of LLaMA-70b increased from 0.876 to 0.914, while Mistral-8 × 7b saw an increase in context precision from 0.841 to 0.903. Smaller-scale models such as LLaMA-8b and LLaMA-70b exhibited context recall improvements of 0.080 and 0.107, respectively.

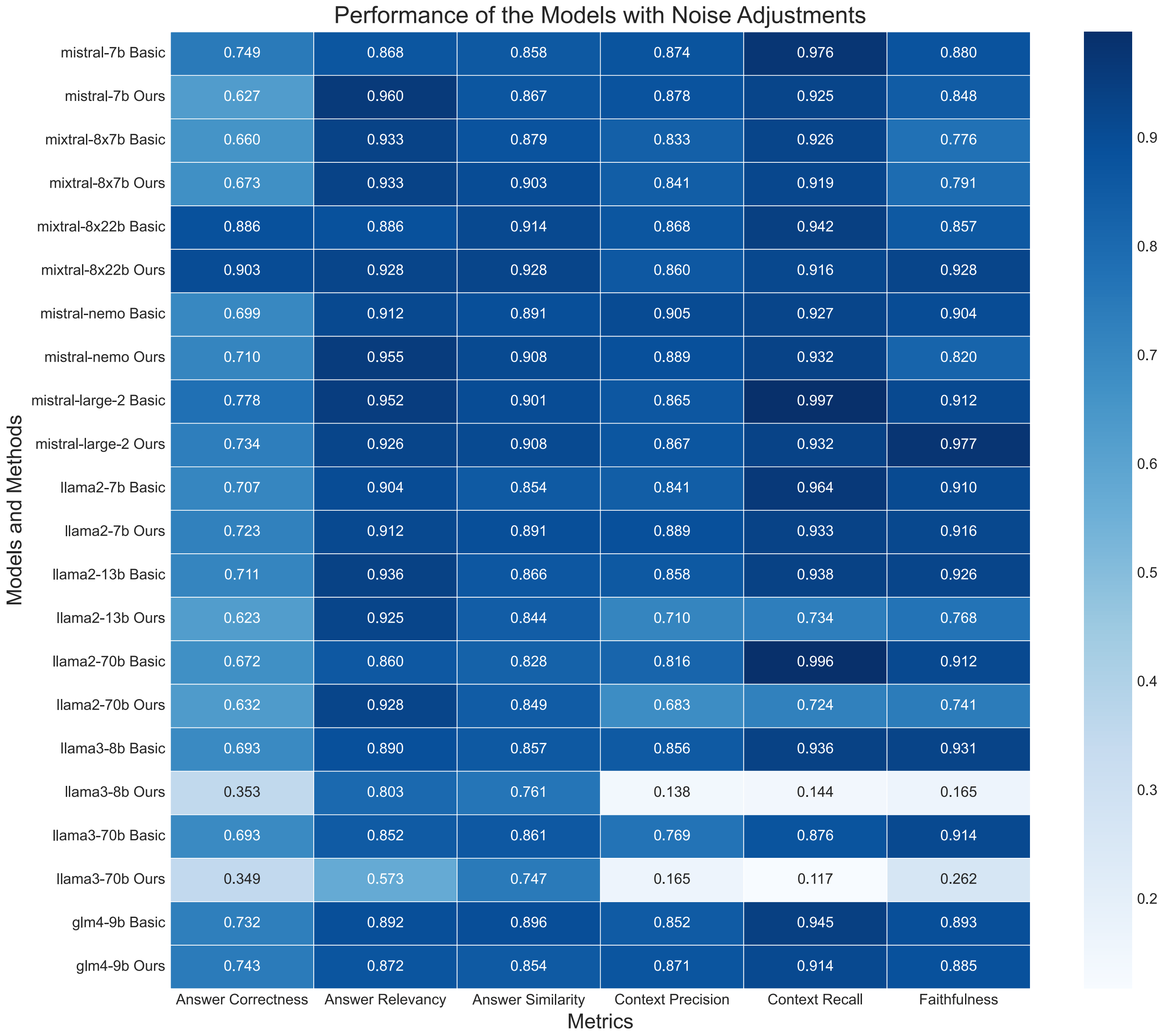


FIGURE 4. Performance improvement of open-source models with retrieval augmentation. The figure highlights the notable gains achieved by smaller-scale models and models with weaker baseline performance when handling complex queries.

These findings demonstrate that retrieval augmentation can comprehensively optimize the utilization of contextual information, particularly in challenging scenarios.

d: FAITHFULNESS

The RAG system effectively reduced unsupported or hallucinated content in noisy environments. For example, the faithfulness score of Mistral-large-2 increased from 0.912 to 0.977, while LLaMA-8b showed an improvement of 0.070. This enhancement is particularly crucial in noisy conditions, where unreliable context can lead to inaccurate responses. By providing reliable contextual information, retrieval augmentation significantly mitigates this risk.

e: MODEL-SPECIFIC OBSERVATIONS

The experiments also revealed differences in performance across models of varying scales. Large-scale models, such as Mistral-8 × 22b and LLaMA-70b, exhibited relatively stable baseline performance without retrieval augmentation but still benefited from improvements in contextual utilization and faithfulness when retrieval augmentation was applied. In contrast, smaller-scale models, such as LLaMA-8b and Mistral-7b, showed more substantial performance gains, particularly in correctness and relevancy metrics. This indicates that retrieval augmentation plays a more pronounced role in enhancing the capabilities of smaller models.

Overall, the RAG system demonstrated its robustness in noisy environments by significantly improving the quality of generation and contextual utilization for open-source models. The results validate the effectiveness of retrieval augmentation, particularly in addressing ambiguous or incomplete queries, with a more pronounced impact on models with weaker baseline performance.

3) COMPARATIVE ANALYSIS OF PROPRIETARY AND OPEN-SOURCE MODELS

A comparative analysis of proprietary and open-source models highlights the advantages of the RAG system in improving generation quality, contextual utilization, and robustness to noisy environments. However, the degree of improvement varies depending on the type and scale of the model.

a: GENERATION QUALITY

Proprietary models demonstrated higher baseline performance compared to open-source models. In the "Basic" condition (without retrieval augmentation), the correctness and relevancy scores of GPT and Claude series models ranged between 0.75 and 0.95, while open-source models, particularly smaller-scale ones, scored lower, with some models achieving only 0.66 to 0.69. However, retrieval augmentation led to larger improvements for open-source models. For example, Mistral-8 × 22b's correctness increased from 0.886 to 0.903, and LLaMA-8b improved significantly from 0.693 to 0.828. In contrast, proprietary models such as Claude-3-haiku showed smaller improvements, increasing from 0.95 to 0.97. These results suggest that retrieval augmentation is particularly effective in addressing the limitations of open-source models while further optimizing proprietary models.

b: CONTEXTUAL UTILIZATION

Proprietary models exhibited stable contextual utilization without retrieval augmentation, with Claude series achieving a context recall of 0.95. Open-source models, however, displayed significant gains after retrieval augmentation. For instance, LLaMA-8b's context precision improved from 0.856 to 0.933, and Mistral-large-2's context recall increased from 0.932 to 0.977. While the improvements for proprietary models were more modest, they still demonstrated marginal gains in context precision and recall.

c: ROBUSTNESS IN NOISY ENVIRONMENTS

Retrieval augmentation proved particularly effective for open-source models under noisy conditions. For example, LLaMA2-13b's faithfulness improved from 0.710 to 0.926, and Mistral-8 × 22b increased from 0.857 to 0.928. Proprietary models, such as Claude-3-haiku and GPT-3.5-turbo, showed smaller improvements, with faithfulness scores increasing from 0.95 to 0.98 and 0.96 to 0.91, respectively. These findings indicate that retrieval augmentation significantly reduces the risk of generating distorted or unsupported content in challenging scenarios while providing performance stabilization for proprietary models.

d: OVERALL COMPARISON

The results demonstrate that the RAG system benefits both proprietary and open-source models, with larger relative improvements observed for open-source models, particularly smaller-scale ones. Proprietary models, leveraging their strong semantic understanding and robust baseline performance, exhibited marginal but valuable optimizations. In contrast, open-source models achieved significant gains in generation quality, contextual utilization, and robustness, particularly under noisy conditions. This underscores the versatility of retrieval augmentation in addressing the limitations of resource-constrained models while enhancing the overall quality and reliability of generated content.

VI. CONCLUSION

This study proposed a Retrieval-Augmented Generation (RAG) system tailored for customs clearance scenarios, referred to as ICCA-RAG. By integrating multimodal document parsing, hybrid retrieval and storage, and context-driven generation models, the system effectively addresses key challenges in customs clearance tasks, including document complexity, query ambiguity, and dynamic regulatory adaptation. Experimental results demonstrate that the ICCA-RAG system significantly improves answer accuracy, contextual relevance, and faithfulness, with exceptional performance in open-source models and under noisy query conditions.

Compared with existing methods, this study achieves several innovative breakthroughs:

1) **Efficient Handling of Multimodal Customs Documents:** Through multimodal document parsing and a hybrid sparse-dense retrieval storage architecture, the system processes diverse customs-related documents efficiently.
2) **Enhanced Contextual Utilization:** A retrieval-augmented semantic vector strategy significantly improves the generation model's ability to leverage context in complex tasks.
3) **Robustness to Noisy Queries:** The system demonstrates exceptional robustness in noisy query scenarios by dynamically constructing context and optimizing prompts.

These results validate the applicability and advanced nature of the ICCA-RAG system in customs clearance scenarios.

Despite these achievements, the study has certain limitations. First, the real-time performance of the retrieval module in handling large-scale document collections requires further optimization. Second, the accuracy of the generation model for certain specific queries heavily depends on the quality of retrieved content. Future work may explore integrating domain knowledge graphs or reinforcement learning techniques to enhance system performance. Additionally, while this study focuses on the customs domain, the proposed methods and framework have broad potential for application in other complex document processing scenarios.

Future research will focus on the following directions:

- **Optimizing Retrieval Efficiency:** Enhancing the storage and retrieval efficiency to support larger-scale datasets.
- **Semantic Augmentation Using Knowledge Graphs:** Improving the generation model's accuracy by incorporating domain-specific semantic enhancements.
- **Expanding Application Domains:** Extending the system's applicability to other professional fields, such as legal, medical, and financial domains.

Through further research and practical exploration, the ICCA-RAG system holds the potential to provide more generalized and practical solutions for intelligent document processing and generation in complex domains.

ACKNOWLEDGMENT

(Rong Hu and Sen Liu contributed equally to this work.)

REFERENCES

[1] M. R. Mpekethu, "Effects of customs procedures on import business performance for small and medium enterprises in Kenya," Tech. Rep., 2018.

[2] M. Goh, "Issues facing Asian SMEs and their supply chains," in *Asian Cases on Supply Chain Management for SME*, vol. 35, 2002.

[3] S. Karklina-Admine, A. Cevers, A. Kovalenko, and A. Auzins, "Challenges for customs risk management today: A literature review," *J. Risk Financial Manage.*, vol. 17, no. 8, p. 321, Jul. 2024.

[4] G. Munyoro, B. Chiinze, and Y. M. Dzapasi, "The role of customs and excise duties on small enterprises: A case study of women cross border traders," *ADRRI J. (Multidisciplinary)*, vol. 25, no. 10, pp. 25–48, Aug. 2016.

[5] Y. Li and J. C. Beghin, "A meta-analysis of estimates of the impact of technical barriers to trade," *J. Policy Model.*, vol. 34, no. 3, pp. 497–511, May 2012.

[6] A. M. Rbehat and H. B. Marafi, "The role of customs process in facilitating international trade," *Saudi J. Bus. Manage. Stud.*, vol. 9, no. 1, pp. 7–14, Jan. 2024.

[7] D. R. R. dos Santos, "Efficiency and protection in international trade: Real-time risk assessment in the Brazilian customs clearance process," *Available at SSRN 4908093*.

[8] W. X. Zhao et al., "A survey of large language models," 2023, *arXiv:2303.18223*.

[9] R. Li, D. Fu, C. Shi, Z. Huang, and G. Lu, "Efficient LLMs training and inference: An introduction," *IEEE Access*, early access, Nov. 18, 2024, doi: 10.1109/ACCESS.2024.3501358.

[10] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, and X. Hu, "Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond," *ACM Trans. Knowl. Discovery from Data*, vol. 18, no. 6, pp. 1–32, Jul. 2024.

[11] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly," *High-Confidence Comput.*, vol. 4, no. 2, Jun. 2024, Art. no. 100211.

[12] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[13] M. Ai, "Introducing llama: A foundational, 65-billionparameter language model," Tech. Rep., 2023.

[14] M. Conover, M. Hayes, A. Mathur, J. Xie, J. Wan, S. Shah, A. Ghodsi, P. Wendell, M. Zaharia, and R. Xin, "Free dolly: Introducing the world's first truly open instruction-tuned LLM," *Company Blog Databricks*, Apr. 2023.

[15] N. Ziems, W. Yu, Z. Zhang, and M. Jiang, "Large language models are built-in autoregressive search engines," 2023, *arXiv:2305.09612*.

[16] B. B. Arcila, "Is it a platform? Is it a search engine? It's ChatGPT! The European liability regime for large language models," *J. Free Speech L.*, vol. 3, p. 455, Jan. 2023.

[17] S. Eleni Spatharioti, D. M. Rothschild, D. G. Goldstein, and J. M. Hofman, "Comparing traditional and LLM-based search for consumer choice: A randomized experiment," 2023, *arXiv:2307.03744*.

[18] B. Yao, M. Jiang, T. Bobinac, D. Yang, and J. Hu, "Benchmarking machine translation with cultural awareness," 2023, *arXiv:2305.14328*.

[19] M. Karpinska and M. Iyyer, "Large language models effectively leverage document-level context for literary translation, but critical errors persist," 2023, *arXiv:2304.03245*.

[20] R. Jain, N. Gervasoni, M. Ndhlovu, and S. Rawat, "A code centric evaluation of C/C++ vulnerability datasets for deep learning based vulnerability detection techniques," in *Proc. 16th Innov. Softw. Eng. Conf.*, Feb. 2023, pp. 1–10.

[21] M. Mudassar Yamin, E. Hashmi, M. Ullah, and B. Katt, "Applications of LLMs for generating cyber security exercise scenarios," *IEEE Access*, vol. 12, pp. 143806–143822, 2024.

[22] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutiérrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Med.*, vol. 29, no. 8, pp. 1930–1940, Jul. 2023.

[23] K. Singhal et al., "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, Jul. 2023.

[24] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "BloombergGPT: A large language model for finance," 2023, *arXiv:2303.17564*.

[25] L. Réka Ónozó, F. Viktor Arthur, and B. Gyires-Tóth, "Leveraging LLMs for financial news analysis and macroeconomic indicator nowcasting," *IEEE Access*, vol. 12, pp. 160529–160547, 2024.

[26] A. B. Mbakwe, I. Lourentzou, L. A. Celi, O. J. Mechanic, and A. Dagan, "ChatGPT passing USMLE shines a spotlight on the flaws of medical education," *PLOS Digit. Health*, vol. 2, no. 2, Feb. 2023, Art. no. e0000205.

[27] B. Saha, U. Saha, and M. Z. Malik, "Quim-rag: Advancing retrieval-augmented generation with inverted question matching for enhanced QA performance," *IEEE Access*, vol. 12, pp. 185401–185410, 2024.

[28] N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel, "Large language models struggle to learn long-tail knowledge," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2022, pp. 15696–15707.

[29] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, L. Wang, A. Tuan Luu, W. Bi, F. Shi, and S. Shi, "Siren's song in the AI ocean: A survey on hallucination in large language models," 2023, *arXiv:2309.01219*.

[30] X. Wang, Y. Yan, L. Huang, X. Zheng, and X. Huang, "Hallucination detection for generative large language models by Bayesian sequential estimation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2023, pp. 15361–15371.

[31] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9459–9474.

[32] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," 2023, *arXiv:2312.10997*.

[33] X. Ma, Y. Gong, P. He, H. Zhao, and N. Duan, "Query rewriting for retrieval-augmented large language models," 2023, *arXiv:2305.14283*.

[34] Z. Wang, S. X. Teo, J. Ouyang, Y. Xu, and W. Shi, "M-RAG: Reinforcing large language model performance through retrieval-augmented generation with multiple partitions," 2024, *arXiv:2405.16420*.

[35] X. Li, Z. Liu, C. Xiong, S. Yu, Y. Gu, Z. Liu, and G. Yu, "Structure-aware language model pretraining improves dense retrieval on structured data," 2023, *arXiv:2305.19912*.

[36] S. Xiao, Z. Liu, P. Zhang, and X. Xing, "LM-cocktail: Resilient tuning of language models via model merging," 2023, *arXiv:2311.13534*.

[37] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-T. Yih, "Dense passage retrieval for open-domain question answering," 2020, *arXiv:2004.04906*.

[38] S. Zhuang, B. Liu, B. Koopman, and G. Zuccon, "Open-source large language models are strong zero-shot query likelihood models for document ranking," 2023, *arXiv:2310.13243*.

[39] R. Litman, O. Anschel, S. Tsiper, R. Litman, S. Mazor, and R. Manmatha, "SCATTER: Selective context attentional scene text recognizer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11962–11972.

[40] F. Xu, W. Shi, and E. Choi, "RECOMP: Improving retrieval-augmented LMs with compression and selective augmentation," 2023, *arXiv:2310.04408*.

[41] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," 2023, *arXiv:2310.11511*.

[42] X. Wang, Q. Yang, Y. Qiu, J. Liang, Q. He, Z. Gu, Y. Xiao, and W. Wang, "KnowledGPT: Enhancing large language models with retrieval and storage access on knowledge bases," 2023, *arXiv:2308.11761*.

[43] S. Wang, Y. Xu, Y. Fang, Y. Liu, S. Sun, R. Xu, C. Zhu, and M. Zeng, "Training data is more valuable than you think: A simple and effective method by retrieving from training data," 2022, *arXiv:2203.08773*.

[44] X. Cheng, D. Luo, X. Chen, L. Liu, D. Zhao, and R. Yan, "Lift yourself up: Retrieval-augmented text generation with self-memory," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024.

[45] D. Cheng, S. Huang, J. Bi, Y. Zhan, J. Liu, Y. Wang, H. Sun, F. Wei, D. Deng, and Q. Zhang, "UPRISE: Universal prompt retrieval for improving zero-shot evaluation," 2023, *arXiv:2303.08518*.

[46] Z. Dai, V. Y. Zhao, J. Ma, Y. Luan, J. Ni, J. Lu, A. Bakalov, K. Guu, K. B. Hall, and M.-W. Chang, "Promptagator: Few-shot dense retrieval from 8 examples," 2022, *arXiv:2209.11755*.

[47] D. Danopoulos, C. Kachris, and D. Soudris, "Approximate similarity search with FAISS framework using FPGAs on the cloud," in *Proc. Int. Conf. Embedded Comput. Syst.* Springer, Jan. 2019, pp. 373–386.

[48] L. D. Krisnawati, A. W. Mahastama, S.-C. Haw, K.-W. Ng, and P. Naveen, "Indonesian-english textual similarity detection using universal sentence encoder (USE) and Facebook AI similarity search (FAISS)," *CommIT (Commun. Inf. Technol.) J.*, vol. 18, no. 2, pp. 183–195, Sep. 2024.

[49] Y. Liu, T.-P. Tan, and X. Zhan, "Iterative self-supervised learning for legal similar case retrieval," *IEEE Access*, vol. 12, pp. 17231–17241, 2024.

[50] M. Murata, H. Nagano, R. Mukai, K. Kashino, and S. Satoh, "BM25 with exponential IDF for instance search," *IEEE Trans. Multimedia*, vol. 16, no. 6, pp. 1690–1699, Oct. 2014.

[51] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Found. Trends Inf. Retr.*, vol. 3, no. 4, pp. 333–389, 2009.

[52] M. Kim and Y. Ko, "Multitask fine-tuning for passage re-ranking using BM25 and pseudo relevance feedback," *IEEE Access*, vol. 10, pp. 54254–54262, 2022.

[53] Y. Ahn, S.-G. Lee, J. Shim, and J. Park, "Retrieval-augmented response generation for knowledge-grounded conversation in the wild," *IEEE Access*, vol. 10, pp. 131374–131385, 2022.

[54] Z. Zeyu, W. Hao, Z. Zibo, L. Yueyan, and Z. Xiaoqin, "Construction and application of GCN model for text classification with associated information," *Data Anal. Knowl. Discovery*, vol. 5, no. 9, pp. 31–41, 2021.

RONG HU is currently an Associate Professor with Shanghai Customs University and the Head of China Customs control study studio, a "think tank" of China Customs. She is also the Director of Shanghai Customs University Postgraduate Department and a Visiting Scholar with Delft University of Technology, Delft. She is also the Representative of China Delegation in RKC comprehensive review. Her research interests include customs control, trade facilitation and security, risk management, and related subjects. Before joining China Customs, she worked for almost ten years with the Maersk Group and gained extensive experience in supply chain management.

SEN LIU was born in Henan, in 2001. He is currently pursuing the master's degree with the Department of Electronic Information, Shanghai Dianji University.

PANPAN QI was born in Huai'an, Jiangsu, China, in 2001. She received the Bachelor of Engineering degree from the Xinglin College, Nantong University, in 2024. During her school years, she received multiple scholarships and was honored with the title of "Outstanding Student." She also won the Third Prize in the Zhongruan International Excellence Cup.

JINGYI LIU is currently a junior student with the Department of Information Engineering, Xi'an Jiaotong University, Xi'an, China. Her research interest includes image processing. She was awarded the Silver Prize in the 2024 Intel Cup Undergraduate Electronic Design Contest—Embedded System Design Invitational Contest.

FENGYUAN LI was born in Futian, Shenzhen, Guangdong, China, in 2004. He is currently pursuing the bachelor's degree in information engineering with Xi'an Jiaotong University.