Journal of Decision Systems

ISSN: 1246-0125 (Print) 2116-7052 (Online) Journal homepage: www.tandfonline.com/journals/tjds20

Business insights using RAG–LLMs: a review and case study

Muhammad Arslan, Saba Munawar & Christophe Cruz

To cite this article: Muhammad Arslan, Saba Munawar & Christophe Cruz (03 Oct 2024): Business insights using RAG–LLMs: a review and case study, Journal of Decision Systems, DOI: 10.1080/12460125.2024.2410040

To link to this article: https://doi.org/10.1080/12460125.2024.2410040

Published online: 03 Oct 2024.

JOURNAL OF DECISION SYSTEMS
https://doi.org/10.1080/12460125.2024.2410040

Taylor & Francis
Taylor & Francis Group

Business insights using RAG–LLMs: a review and case study

Muhammad Arslan [a,b], Saba Munawar [c] and Christophe Cruz [b]

[a]School of Architecture and Environment, University of the West of England, Bristol, UK; [b]Laboratoire Interdisciplinaire Carnot de Bourgogne (ICB), Université de Bourgogne, Dijon, France; [c]Electrical (Telecommunication) Engineering Department, National University of Computer and Emerging Sciences (NUCES), Islamabad, Pakistan

ABSTRACT

As organizations increasingly rely on diverse data sources like invoices and surveys, efficient Information Extraction (IE) is crucial. Natural Language Processing (NLP) enhances IE through tasks such as Named Entity Recognition (NER), Relation Extraction (RE), Event Extraction (EE), Term Extraction (TE), and Topic Modeling (TM). However, implementing these methods requires significant expertise, which smaller organizations often lack. Large Language Models (LLMs), powered by Generative Artificial Intelligence (GenAI), can address this by performing multiple IE tasks without extensive development costs. However, LLMs may struggle with domain-specific accuracy. Integrating Retrieval-Augmented Generation (RAG) with LLMs improves precision by incorporating external data. Despite the potential, research on RAG-LLM applications in the business domain is limited. This article reviews Business IE systems, explores RAG-LLM applications across disciplines, and presents a case study demonstrating how RAG-LLMs can enhance business insights, offering scalable, cost-effective solutions.

ARTICLE HISTORY
Received 3 July 2024
Accepted 21 September 2024

KEYWORDS
Information extraction; natural language processing; large language models; generative artificial intelligence; retrieval-augmented generation

1. Introduction

Organisations rely on diverse information sources like invoices, customer surveys, legal documents, and banking records to support business activities (Al-Okaily et al., 2023). As these data, both structured and unstructured, grows in volume efficient Information Extraction (IE) methods become essential for informed decision-making (Abdullah et al., 2023). IE is crucial for extracting entities, relations, events, terms, and topics needed for business analysis (Cunningham, 2005; Martinez-Rodriguez et al., 2020). However, many organisations still use manual extraction methods, which are resource-intensive, prone to errors, and lead to delays (Adnan & Akbar, 2019). Manual processes also require costly and time-consuming verification to ensure information accuracy, further complicating and slowing down business operations. Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI) focused on enabling machines to understand and interpret human languages (de Almeida Bordignon et al., 2018). It

CONTACT Muhammad Arslan muhammad.arslan@uwe.ac.uk; muhammad.arslan@u-bourgogne.fr School of Architecture and Environment, University of the West of England, Coldharbour Lane, Bristol BS16 1QY, UK

has revolutionised IE by automating tasks like Named Entity Recognition (NER), which identifies entities such as people and organisations in text (Abdullah et al., 2023). Other NLP-based IE tasks include Relation Extraction (RE) to identify relationships between entities, Event Extraction (EE) to extract event details, Term Extraction (TE) to identify key phrases, and Topic Modelling (TM) to uncover themes in text. These tasks enhance business operations by providing valuable insights for decision-making, improving efficiency, reducing costs, and minimising errors across various sectors like finance, healthcare, and manufacturing.

Numerous studies (Abdullah et al., 2023; Martinez-Rodriguez et al., 2020) have explored methods like NER, RE, EE, TE, and TM to extract valuable business information. These methods often require domain expertise, extensive rule creation, or large datasets for training machine learning models, posing challenges for smaller organisations with limited Research and Development (R&D) resources. Instead of separate methods for each task, LLMs powered by GenAI offer a more efficient solution. GenAI, which creates diverse content like text and synthetic data, has become popular due to its user-friendly interfaces (Feuerriegel et al., 2024; Mariani & Dwivedi, 2024). Trained on vast datasets, they quickly generate human-like text, enabling the creation of AI-powered applications that efficiently extract key information, streamline business processes, and reduce manual labour and development costs (Feuerriegel et al., 2024). LLMs exhibit potential for IE tasks but face challenges with domain-specific queries, often generating inaccurate or misleading information, known as 'hallucinations' (Kandpal et al., 2023). These hallucinations can arise from an overload of data, a lack of contextual relevance, or both, compromising the reliability of LLMs in practical applications (Y. Gao et al., 2023).

While the integration of RAG with LLMs has achieved notable technological advancements (Y. Gao et al., 2023; H. Li et al., 2022; Mialon et al., 2023; C. Zhao et al., 2023), the existing literature primarily focuses on these innovations themselves, overlooking their practical applications in the business sector. Specifically, there is a lack of comprehensive classification of RAG–LLM studies across various tasks and disciplines, which is crucial for businesses aiming to adopt these technologies effectively. A detailed classification would enable organisations to pinpoint the most effective RAG–LLM approaches for specific business IE tasks. This would provide actionable insights into how these advancements can be leveraged to enhance efficiency, reduce costs, and improve decision-making processes. Before implementing RAG–LLM-based approaches, it is essential to understand the existing background on Business IE systems and associated NLP tasks. This review examines how current Business IE systems operate across diverse use cases and identifies opportunities for enhancement through RAG–LLM solutions. By mapping traditional systems to potential RAG–LLM-driven upgrades, the paper offers valuable guidance for businesses looking to modernise their IE capabilities. The paper makes the following key contributions:

(1) A review of RAG–LLM applications is provided, organising them by specific NLP tasks (such as Question Answering, Text Generation, and Summarisation) and disciplines. These fundamental NLP tasks span numerous fields, making the findings particularly relevant for the business community. The review reveals how RAG–LLM integration has revolutionised traditional IE processes, offering businesses actionable insights to adopt these advancements. Additionally, it uncovers a significant gap in the literature, indicating that the application of RAG–LLMs in

Business IE remains largely unexplored, thus highlighting a crucial area for future R&D.

(2) A novel case study is presented, showcasing the practical application of RAG with LLMs in developing a Business IE solution. This case study not only demonstrates the potential of this integrated approach to significantly enhance business insights but also serves as a catalyst for further R&D in the field. It provides a tangible example of how RAG–LLM can be leveraged to create cost-effective, scalable, and highly efficient Business IE systems, thereby inspiring businesses and researchers to explore this promising avenue further.

The paper is structured as follows: Section 2 provides a comprehensive background on Business IE systems and pertinent NLP tasks. Section 3 outlines the research methodology in detail. Section 4 reviews the literature on RAG with LLMs, categorising their applications by task and discipline. In Section 5, we present a novel case study showcasing the application of RAG with LLMs for Business IE and evaluate the effectiveness of the system used. Section 6 discusses the advantages and limitations of the system. Finally, Section 7 concludes the paper with a summary of key findings and implications.

2. Background

IE involves deriving structured data from sources like text documents, emails, web pages, and databases (Martinez-Rodriguez et al., 2020). This structured data includes entities, relationships, events, and other relevant details, obtained through techniques in NLP, machine learning, and computational linguistics. Business IE, a specialised form of IE, caters to the specific needs of businesses by extracting information from sources such as financial reports, customer feedback, market research, and legal documents (de Almeida Bordignon et al., 2018). The extracted data may include key performance indicators, market trends, customer sentiments, competitive analysis, and regulatory compliance information, all critical for business operations and decision-making. Existing studies (Al-Okaily et al., 2023; de Almeida Bordignon et al., 2018) have highlighted five key IE tasks in business contexts: NER, RE, EE, TE, and TM (see Figure 1). NER identifies and classifies named entities, such as company names and financial figures, within texts. RE detects semantic relationships between entities, such as connections between a company and its acquisitions. EE focuses on specific events, including details like participants, time, and location, such as merger and acquisition events. TE extracts key terms or keywords from text, like performance metrics in financial reports. TM uncovers underlying themes in document collections, such as recurring topics in customer reviews. These techniques enable businesses to gain deeper insights into various aspects of their operations and customer preferences.

In the domain of Business IE, LLMs have significantly advanced the process. Leveraging GenAI, these models represent a major leap in NLP. Notable examples include Meta's Llama2 (Touvron et al., 2023), OpenAI's GPT-4.0 (Achiam et al., 2023), and Anthropic's Claude 2.0 (Anthropic, 2023). These sophisticated machine learning models are trained on extensive text corpora, enabling them to generate human-like text by deeply understanding context and filtering out irrelevant information. Their capabilities enhance text coherence and contextual accuracy, thus

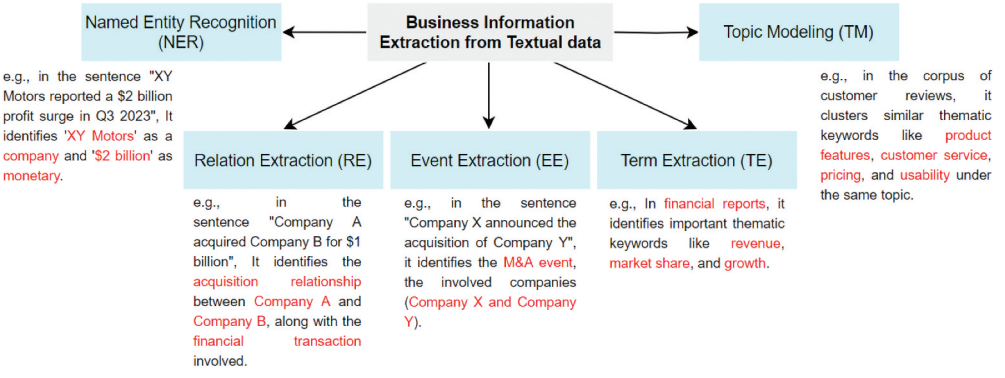

Figure 1. Various tasks in business IE from textual data.

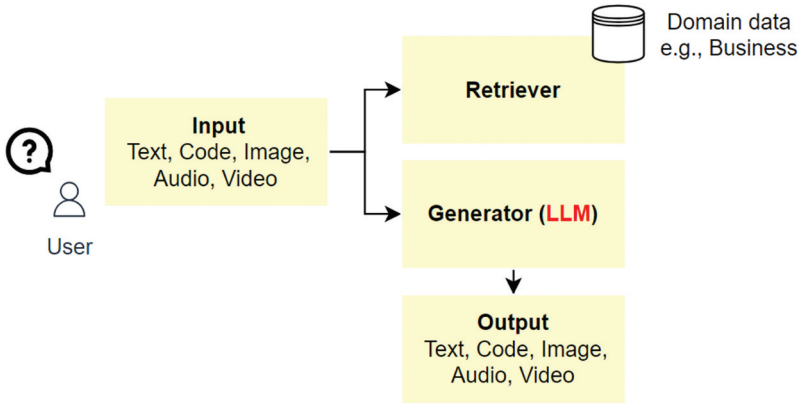

Figure 2. Generic RAG architecture integrating LLMs.

improving the efficiency of extracting valuable insights from large datasets. Despite these strengths, LLMs often encounter challenges with domain-specific queries, where their outputs may be inaccurate or not sufficiently tailored to specific business needs. To address these limitations, the RAG framework has been introduced. As illustrated in Figure 2, RAG enhances LLM performance by incorporating an additional retrieval step: before generating text, LLMs first retrieve relevant information from external sources (Lewis et al., 2020). This step involves sourcing pertinent data to provide a more accurate and contextually relevant foundation for text generation. By integrating external data retrieval, RAG significantly improves the accuracy and relevance of LLM outputs. This enhancement reduces the risk of errors and elevates the quality of extracted information, making it more reliable for business decision-making. Consequently, businesses can leverage RAG-enhanced LLMs to obtain precise, relevant, and evidence-based insights tailored to their needs, thus improving decision-making and strategic planning. This combination of LLMs with RAG offers a more effective and practical tool for extracting and utilising business insights in real-world scenarios.

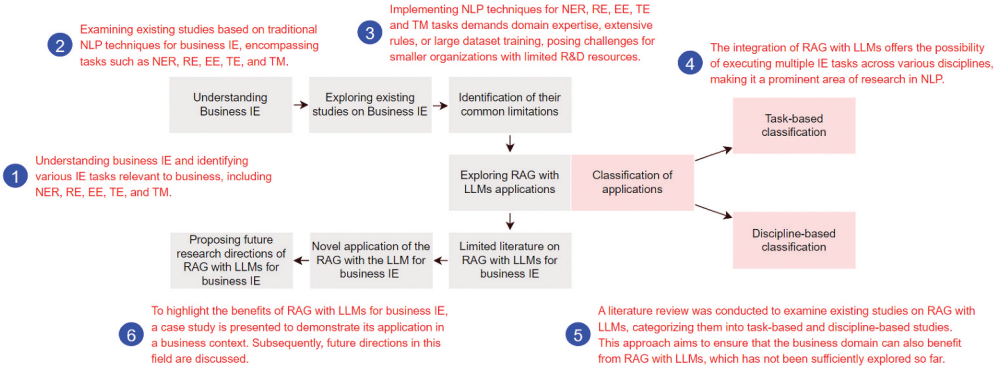

Figure 3. Literature analysis process for identifying the need for RAG–LLM integrated systems in the business domain.

3. Research method

A comprehensive literature review (see Figure 3) was conducted to evaluate the current landscape of Business IE systems and related NLP tasks. The primary databases utilised for this search included Scopus, IEEE Xplore, ACM, Google Scholar, and ScienceDirect. Keywords used for the search were: *Business IE*, *Business Data Extraction*, *Business IE Systems*, *Enterprise IE*, *Business NLP*, *Automated Data Extraction in Business*, *Financial IE*, *Business Intelligence (BI) Automation*, among others. This initial search resulted in 124 documents. After eliminating duplicates and selecting only studies that provided implementation details and validation, 36 studies relevant to Business IE were retained. Through this review, we identified key IE tasks essential for business operations, such as NER, RE, EE, TE, and TM. During data analysis, the selected studies were categorised based on the specific IE tasks addressed and the NLP techniques employed. The review highlighted challenges in implementing these techniques, particularly the need for domain expertise, complex rule creation, or large datasets for training, obstacles that are especially difficult for smaller organisations with limited R&D resources. In addition to reviewing traditional IE methods, we extended our analysis to explore the latest advancements in the field, particularly focusing on LLMs. Recognising that general-purpose LLMs often struggle with domain-specific IE tasks, we examined how the integration of RAG with LLMs could address these limitations. The same databases were used to search for literature on RAG–LLM integration, employing keywords such as *RAG*, *LLMs*, *External Knowledge Retrieval*, *Hybrid Retrieval-Generation*, *Knowledge-Enhanced LLMs*, and *Context-Aware NLP*. A total of 75 articles were found, many of which were preprints available through platforms like arXiv. After a thorough review, 51 articles were selected based on the quality and depth of their analysis, with particular attention to the fact that many were preprints. These studies were categorised into task-based and discipline-specific classifications.

This comprehensive review illustrates how RAG–LLM integration has revolutionised traditional IE tasks across multiple domains, including Medical, Financial, Education, Technology/Software, and Open-Domain Applications. Additionally, the studies cover a range of NLP tasks such as QA (answering questions based on given text), Text

Generation and Summarisation (creating new text and summarising long documents into concise summaries), Information Retrieval and Extraction (locating and pulling out specific information from large datasets), Text Analysis and Processing (analysing text for insights and processing it for various uses), Software Development (designing, coding, and maintaining software applications), Decision-Making (assisting in making data-driven decisions), and related applications. These advancements provide valuable insights for the business community to enhance their operational strategies and leverage cutting-edge technology. Despite the relevance of these IE tasks in the business context, the review identified a significant gap in the literature, with few studies exploring RAG–LLM applications in Business IE, signalling a critical area for future R&D. To demonstrate the practical application of RAG–LLMs in enhancing Business IE processes such as NER, RE, EE, TE, and TM, we conducted a case study.

The study centred on a digital transformation project designed to automate the handling and classification of business information for a company. Digital transformation, in this context, refers to the strategic implementation of digital technologies to streamline processes, improve decision-making, and ultimately deliver greater value to customers (Kraus et al., 2021). To effectively tailor the project to the company's needs, we conducted a series of focus group discussions with the BI team. These discussions were crucial for understanding the specific types of information the team required from their datasets. Based on the insights gained, we developed a RAG with LLM-based system for Business IE. The performance of the RAG–LLM system was evaluated using state-of-the-art metrics, which provided concrete evidence of its effectiveness in improving Business IE processes. This approach ensured that the system met the BI team's specific needs and demonstrated its practical value in enhancing business insights.

4. Literature review

Business IE tasks employ a wide array of techniques, such as Bidirectional Encoder Representations from Transformers (BERT)-based models (Devlin et al., 2018), ontology-based methods (Arendarenko & Kakkonen, 2012), machine learning algorithms (Jacobs & Hoste, 2022), syntactic and semantic rules (X. Gao et al., 2005), Term Frequency-Inverse Document Frequency (TF-IDF) approaches (Sul & Cho, 2024), statistical techniques (Bzhalava et al., 2024), rule-based approaches, deep learning models (Yan et al., 2019), and Markov logic networks (Yamamoto et al., 2017). Table 1 provides a comprehensive overview of existing studies focused on Business IE. It presents key details across several categories, including the use cases addressed, the types of IE performed, the models and techniques employed, the datasets used, and the specific types of information extracted. Each study is mapped to a particular business application, outlining whether it involves NER, RE, EE, or other types of IE tasks. The table also lists the models and techniques applied. The datasets used in these studies vary, encompassing structured and unstructured business data, while the extracted information ranges from named entities and relationships between entities to events and other key business insights. Despite their advantages, these studies reveal notable limitations:

Table 1. Existing studies on business ie.

No.	Use case	IE type	Used models/techniques	Datasets	Extracted information
1	Improving IE on business documents Douzon et al. (2022)	NER	BERT-based	Invoices, purchase orders and expense receipts.	Company, address, date, total amount and purchase order number.
2	Few-exemplar IE for business documents Esser et al. (2014)	NER	One-shot learning (model unknown)	Business documents from DocuWare.	Document type, sender, recipient, and date.
3	IE using multi-modal transformers Geletka et al. (2022)	NER	BERT-based	Born-digital invoices	Total amount, invoice number, invoice date, bank code, account number, IBAN, etc.
4	Gather company intelligence and country/region information for BI Saggion et al. (2007)	NER	Ontology-based, and General Architecture for Text Engineering (GATE) system.	Company websites, company reports and newspapers.	Company name, activities, employee count, directors, and regional details.
5	Open-domain IE from business news Arslan & Cruz (2024)	NER	Ontology-based	News articles	People, organisations, locations, etc.
6	IE of business documents captured with smartphones and tablets Esser et al. (2013)	NER	Intellix (i.e. a commercial quality extraction system)	Business documents	Document type, recipient, sender, date, amount, document number, and subject.
7	IE from invoices Hamdi et al. (2021)	NER	BERT-based	French and English invoices.	Dates, document numbers, types, amounts, etc.
8	IE for domain-specific business documents with limited data Nguyen et al. (2020)	NER	BERT-based	Benefit pension plan documents	Payee, payer, and the deadline for applying qualification.
9	Business IE from semi-structured webpages Sung and Chang (2004)	NER	Inductive learning using MS-SQL server	Shopping malls data	Name of the business, business license number, address, telephone number, etc.
10	Extraction of structured IE from business documents Sage et al. (2020)	NER	Pointer-generator networks	Purchase orders	Product ID and qualtity.
11	Automated extraction of insurance policy information (Hedberg & Furberg, 2023)	NER	BERT- based	Insurance policy documents	Company name, etc.
12	IE on domain-specific business documents R. Zhang et al. (2020)	NER	BERT- based	Regulatory filings and property lease agreements.	Shareholder name, initials of the shareholder, etc.
13	Automatic IE in business documents Moreno Acevedo (2023)	NER	BERT- based	News wire articles, registration documents, court decision documents and legal dispositions.	Activity, date person, organisation, etc.
14	Enabling supervised IE of company-specific events Jacobs and Hoste (2022)	EE	BERT-based	Economic and financial news articles.	Events e.g. acquisition, investment, etc.

(*Continued*)

Table 1. (Continued).

No.	Use case	IE type	Used models/techniques	Datasets	Extracted information
15	Semantic based IE of financial information Simmons and Conlon (2013)	EE	Content Analysis and INformation Extraction System (CAINES)	Online business reports	Financial activities e.g., mergers, acquisitions, and new business segments.
16	An event-extraction approach for business analysis S. Han et al. (2018)	EE	Machine learning models and word embeddings.	News articles	Analyzing industry trends
17	Financial events using weak supervision Dor et al. (2019)	EE	BERT- based	News articles and Wikipedia	Wikipedia events in the news
18	Extraction of keyterms for business information retrieval X. Gao et al. (2005)	TE	Syntactic rules, geographic layout of document, occurrence of terms and co-occurrence of related terms.	Web documents	Key terms and their semantic relationships.
19	Extracting keywords from open-ended business survey questions McGillivray et al. (2020)	TE	TF-IDF approach	Free-text survey data	Business terms
20	Prediction of news popularity Pugachev et al. (2021)	TE	BERT- based	News articles	Business keywords
21	Keyword extraction for indexing in multilingual set-up Piskorski et al. (2021)	TE	TF-IDF, RAKE, KPMiner, YAKE, KeyBERT, and variants of TextRank-based.	News articles	Business keywords
22	Entity relation extraction for competitive intelligence Reyes et al. (2021)	RE	BERT- based	News articles	Semantic relationships between named entities.
23	Company relations for analyzing industry structure Yamamoto et al. (2017)	RE	Markov logic network	News articles	Cooperative and competitive company relations.
24	Relations of enterprises in credit risk management Yan et al. (2019)	RE	Neural Networks	Corporate news articles	Business relations
25	Multilevel entity-informed business relations Khaldi et al. (2021)	RE	BERT- based	BizRel dataset	Business relations
26	Semantic relation extraction in regulatory documents (Korger & Baumeister, 2021)	RE	Rule-based	Public events data	Relations between incidents and measures
27	Relation extraction from financial reports Sun (2022)	RE	BERT- based	Financial reports	Semantic relations
28	Identifying digital traces for business marketing Luo et al. (2015)	TM	LDA-based	Company's micro-blog posts	Digital traces

(*Continued*)

Table 1. (Continued).

No.	Use case	IE type	Used models/techniques	Datasets	Extracted information
29	Identifying business opportunities Pournemat and Weiss (2021)	TM	LDA-based	Customer comments on the social media.	Business opportunities
30	Complex events matching to the user query La Fleur et al. (2015)	NER and EE	AlchemyAPI and Esper	DBpedia dataset	Business events
31	Relation extraction of business products with entities Schön et al. (2020)	NER and RE	Annotation-based	Web pages from business news portals, company home page, etc.	Product entity and company-product relations.
32	KPI-BERT for financial reports Hillebrand et al. (2022)	NER and RE	BERT- based	Financial reports	Linking key performance indicators (KPIs)
33	Extracting business process entities and relations from text Bellan et al. (2022)	NER and RE	GPT-3-based	Process description documents	Activity, participant, and the relation.
34	Extracting business insights Arslan and Cruz (2022)	NER and TM	BERT- based	News articles	Business topic, location, and person.
35	Ontology-based IE and EE for BI Arendarenko and Kakkonen (2012)	NER and EE	Ontology-based	News stories	Company and product information and events.
36	Digital business foresight Bzhalava et al. (2024)	TE and TM	TF-IDF and ANOVA	CrunchBase metadata	Link weak/strong signals with business concepts.

- Business IE tasks like NER, RE, EE, TE, and TM typically require distinct systems or models, creating challenges for organisations needing to perform multiple tasks simultaneously.
- The wide array of available systems complicates selection and validation, making it difficult for businesses to find solutions that align with their specific needs.
- Many systems lack accessibility (as seen in Table 1), and organisations often do not have the proprietary data required to train these models from scratch, further hindering implementation.

These challenges demand significant time and resources, which smaller organisations may not have for dedicated R&D efforts. To address these challenges, we explored the potential of RAG integrated with LLMs based on advanced NLP principles to execute multiple Business IE tasks efficiently. Table 2, adapted from our previous research (Arslan et al., 2024), provides an overview of the diverse applications of RAG integrated with LLMs, highlighting various use cases across different domains.

It details the specific scenarios in which RAG has been employed, the datasets or benchmarks used in these studies, and the associated application areas. Based on this information, we can categorise RAG with LLM-based studies into two primary types: task-specific applications, which concentrate on particular NLP tasks, and discipline-specific applications, which target specific fields or industries. Table 3 presents a task-based classification of RAG applications, showing the distribution of publications across various

Table 2. Applications of RAG with LLMs.

No.	Use case with RAG	Used datasets/benchmarks	Application area
1	MIRAGE: Medical information RAG Xiong et al, (2024)	Medical QA datasets	Biomedical QA
2	RAG through financial report chunking for improved context and information accuracy Jimeno Yepes et al. (2024)	Financial reports	Financial QA
3	Retrieval-augmented Electrocardiography (ECG) analysis model Yu et al. (2023)	Domain knowledge of cardiac symptoms and sleep apnea diagnosis	Medical QA
4	Enhancing Representative Vector Summarization (RVS) with RAG-assisted abstractive-extractive workflows Manathunga and Illangasekara (2023)	PDFs, text documents, spreadsheets, and slide presentations	Medical text summarization
5	Retrieval-augmented controllable review generation Kim et al. (2020)	Amazon book reviews	Book review generation guided by reference documents
6	Retrieval-augmented knowledge graph reasoning Sha et al. (2023)	CommonsenseQA and OpenBookQA datasets	Commonsense QA
7	Extract answers from table corpus via RAG Pan et al. (2022)	Wikipedia data	Table QA
8	LiVersa: a liver disease specific LLM using RAG Ge et al. (2023)	Data by Association for the Study of Liver Diseases (AASLD)	Medical QA
9	Almanac: Retrieval-augmented language Model for clinical medicine Zakka et al. (2024)	Medical resources (guidelines and treatment recommendations)	Clinical decision-making
10	Assessment of tutoring practices using RAG Z. F. Han et al. (2024)	Tutoring dialogue transcripts from a middle-school	Educational decision making
11	Handling out of domain scenarios using RAG Alawwad et al. (2024)	Lessons covering life science, earth science, and physical science	Textbook QA
12	RAG for automated form filling Bucur (2023)	Web page files of request forms for IT projects	Enterprise search
13	Financial sentiment analysis via RAG B. Zhang et al. (2023)	Twitter financial news and FiQA datasets	Sentiments classification
14	Frontline health worker capacity building using RAG Al Ghadban et al. (2023)	Pregnancy-related guidelines	Health education QA
15	Self-BioRAG: a framework for biomedical text Jeong et al. (2024)	Biomedical instruction sets	Generate biomedical explanations and retrieve domain-specific documents.
16	Hybrid RAG for real-time composition assistance Xia et al. (2023)	WikiText-103, Enron Emails, HackerNews, NIH ExPorter, and Youtube Subtitles	Enhance user writing speed and accuracy
17	RAG-Fusion to obtain product information Rackauckas (2024)	Product datasheets	Technical product information QA
18	RACE: Retrieval-augmented commit message generation for code intelligence E. Shi et al. (2022)	MCMD dataset with five programming languages (PLs): Java, C#, C++, Python and JavaScript.	Software development and maintenance
19	FloodBrain: Flood disaster reporting via RAG Colverd et al. (2023)	ReliefWeb reports	Humanitarian assistance
20	RAG with rich answer encoding Huang et al. (2023)	MSMARCO QA dataset, Wizard of Wikipedia (WoW) dataset	Generative QA and informative conversations
21	Retrieval-augmented text-to-image generator Chen et al. (2022)	COCO and WikiImages datasets	Generate realistic and faithful images
22	ReACC: a retrieval-augmented code completion framework Lu et al. (2022)	CodeXGLUE and CodeNet datasets	Software development and maintenance
23	GROVE: a retrieval-augmented complex story generation framework Wen et al. (2023)	IMDB movie details dataset	Generate stories with complex plots

(*Continued*)

Table 2. (Continued).

No.	Use case with RAG	Used datasets/benchmarks	Application area
24	TRAC: Trustworthy retrieval augmented chatbot S. Li et al. (2023)	Natural Question dataset	Natural QA
25	Clinfo.ai: an open-source retrieval-augmented system using scientific literature Lozano et al. (2023)	PubMed dataset	Medical QA
26	RealGen: RAG for controllable traffic scenarios Ding et al. (2023)	nuScenes dataset	Editing and crafting diverse behaviours, including critical traffic scenarios
27	RAG for zero-shot disease phenotyping Thompson et al. (2023)	Clinical notes	Identifying diseases
28	RAP-Gen: retrieval-augmented patch generation for automatic program repair W. Wang et al. (2023)	TFix, Code Refinement and Defects4J datasets	Software development and maintenance
29	Code4UIE : retrieval-augmented code generation Guo et al. (2023)	ACE04, ACE05, CoNLL03, ADE, CoNLL04, NYT, ACE05 and CASIE datasets	Information extraction
30	RAP: retrieval-augmented planning with contextual memory Kagaya et al. (2024)	ALFWorld, Webshop, FrankaKitchen and MetaWorld datasets	Decision-making applications
31	RIGHT: RAG for mainstream hashtag recommendation Fan et al. (2023)	English Twitter (THG) and Chinese Weibo (WHG) datasets	Retrieval-enhanced hashtags
32	RAUCG: retrieval-augmented unsupervised counter narrative generation for hate speech Jiang et al. (2023)	MultitargetCONAN dataset	Combating online hate speech
33	Weakly-supervised scientific document classification via retrieval-augmented multi-stage training Xu et al. (2023)	AGNews and MeSH datasets	Scientific documents classification
34	rT5: a retrieval-augmented model for ancient Chinese entity description generation Hu et al. (2023)	XunZi and MengZi datasets	Entity description generation
35	RSpell: retrieval-augmented framework for domain adaptive Chinese spelling check Song et al. (2023)	CSC dataset	Text error correction
36	XRICL: cross-lingual retrieval-augmented in-context learning for cross-lingual text-to-SQL semantic Parsing E. Shi et al. (2022)	XSPIDER and XKAGGLE-DBQA datasets	Text-to-SQL translation
37	SELF-RAG: learning to retrieve, generate, and critique through self-reflection Asai et al. (2023)	Open-Instruct processed data	Open-domain question answering and fact verification
38	ChatDOC with enhanced PDF structure recognition Lin (2024)	Academic papers, financial reports, textbooks, and legislative materials	Professional knowledge QA
39	G-Retriever: retrieval-augmented generation for textual graph understanding He et al. (2024)	GraphQA (ExplaGraphs, SceneGraphs and WebQSP)	Chat with graphs
40	Enhancing multilingual information retrieval in mixed Human Resources (HR) environments Ahmad (2024)	HR standard operating procedures (SOPs) and Quality Assurance (QA) documents	Multicultural enterprise QA
41	Differentiable retrieval augmentation via generative language modelling C. Zhao et al. (2023)	User-clicked logs	E-commerce search (query intent classification)
42	RAG to elevate low-code developer skills Nakhod (2023)	Caspio and Power automate documentation data	Software development and maintenance
43	UniMS-RAG: a unified multi-source RAG H. Wang et al. (2024)	DuLeMon and KBP datasets	Personalised dialogue systems
44	RAG question answering for event argument extraction Du and Ji (2022)	ACE 2005 and WikiEvent datasets	Event argument (answer) extraction
45	FABULA: retrieval-augmented narrative construction Ranade and Joshi (2023)	OntoNotes and Pile datasets	Intelligence report generation
46	Time-Aware Adaptive Retrieval (TA-ARE) Z. Zhang et al. (2024)	RetrievalQA dataset	Short-form open-domain QA

(*Continued*)

Table 2. (Continued).

No.	Use case with RAG	Used datasets/benchmarks	Application area
47	Cash transaction booking via RAG B. Zhang et al. (2023)	Cash Management Software (CMS) transactions dataset	Automated cash transaction booking
48	Retrieval-Augmented Thought Process (RATP) Pouplin et al. (2024)	Boolq and emrQA datasets	Question answering with private data
49	ATLANTIC: structure-aware RAG for interdisciplinary science Munikoti et al. (2023)	S2ORC dataset	Science QA and scientific document classification
50	Writing documents for clinical trials Markey et al. (2024)	FDA guidance database, ClinicalTrials.gov, and AACT database	Clinical-related writing
51	Question and Answer Retrieval Augmented Generation (QARAG) model Kim and Min (2024)	FDA Q&A datasets	Pharma industry regulatory compliance QA

Table 3. Task-based classification of RAG applications with count of publications.

No.	Task	Count of Publications
1	Question Answering (QA)	20
2	Text Generation and Summarisation	6
3	Information Retrieval and Extraction	6
4	Text Analysis and Processing	5
5	Software Development and Maintenance	3
6	Decision Making and Applications	5
7	Other	6

tasks. The most addressed task is QA, with 20 publications, highlighting the prominence of RAG in enhancing QA systems. Text Generation and Summarisation, along with Information Retrieval and Extraction, both have 6 publications, indicating a strong focus on improving these critical NLP tasks. Text Analysis and Processing and Decision Making and Applications each have 5 publications, reflecting their growing importance in leveraging RAG for more advanced decision-making capabilities. Software Development and Maintenance is represented by 3 publications, while 6 papers fall under the 'Other' category, covering miscellaneous applications of RAG across different domains.

Table 4 provides a discipline-based classification of RAG applications, detailing the number of publications across various fields. The Medical and Technology/Software disciplines are the most extensively covered, each with 12 publications, underscoring the significant research focus on applying RAG in these areas. General and Open-Domain Applications follow with 8 publications, indicating a broad interest in versatile

Table 4. Discipline-based classification of RAG applications with count of publications.

No.	Discipline	Count of Publications
1	Medical	12
2	Financial	5
3	Education	5
4	Technology/Software	12
5	General and Open-Domain Applications	8
6	Sentiment and Social Applications	2
7	Humanitarian and Assistance Applications	1
8	Creative and Content Generation	6

RAG applications. The fields of Financial and Education each have 5 publications, reflecting a moderate level of research in these sectors. Creative and Content Generation applications are explored in 6 publications, while Sentiment and Social Applications and Humanitarian and Assistance Applications are less represented, with only 2 and 1 publication, respectively. From this review, business analysts gain a comprehensive understanding of how the RAG with LLM-based approach has transformed the IE process across various domains. However, it is evident that the application of RAG with LLMs in the business domain remains underexplored. This gap underscores a critical opportunity for further exploration and innovation within the business sector. To address this gap, the next section will present a detailed case study that illustrates the digital transformation of a company's IE process using RAG with LLMs. This case study will provide practical insights into how RAG–LLM integration can be utilised to streamline and enhance Business IE, demonstrating its potential benefits and applications in a real-world business context.

5. Case study: business IE for insights using RAG and LLM

This section presents a case study that demonstrates the digital transformation of a company's IE process. The case study was selected to highlight how the integration of advanced technologies, particularly RAG with LLMs, can enhance the efficiency and accuracy of business data extraction. The rationale for this case study stems from the increasing need for businesses to modernise their operations by adopting AI-driven solutions, a necessity driven by the rapidly evolving digital landscape. The case study focuses on a mid-sized company seeking to automate its IE tasks, which were previously manual and inefficient. It provides a detailed look at how GenAI models, specifically RAG integrated with LLMs, can be utilised to transform business operations, reduce operational costs, and improve data-driven insights. The following sections will outline the specific business challenges addressed, followed by an exploration of the GenAI model's application, illustrating the practical benefits of integrating RAG with LLMs in this real-world scenario.

5.1. Context and requirements acquisition, mapped to IE tasks

A news aggregation company aimed to extract valuable business opportunities from online articles for its clients. For instance, articles about personnel or machinery acquisitions could provide leads for clients in related industries. BI experts manually reviewed these news articles, identifying key topics from the company's taxonomy (i.e. a structured list of relevant topics). In addition, they extracted important details such as company names, locations, timestamps, financial figures, and other relevant information. The experts then condensed these lengthy articles into 5–6 sentence summaries that encapsulated critical details for business clients. These summaries, referred to as 'business opportunities', included insights about companies, their activities, geographical reach, financial data, personnel changes, industry trends, and regulatory updates (see Table 5). The BI team, comprised of six experts specialising in different geographic regions, processed an average of 15 to 20 articles daily. However, the manual nature of this process presented several limitations;

Table 5. Information required for business opportunity understanding.

No.	Key information	Purpose of extraction
1	Company Information	News articles provides details about different companies, covering their names, industries, locations, size (in terms of employees), and key personnel. This data aids in identifying potential business opportunities, partnerships, or areas for growth. Business clients are typically selective, often favouring business opportunities from larger companies with ample manpower or higher revenue generation.
2	Business Activities and Developments across Geographic Locations	News articles provides details on new initiatives, expansions, acquisitions, construction projects, fundraising activities, product launches, and other relevant developments within specific regions or countries. This information helps identify potential business opportunities, market trends, and areas for business growth or investment.
3	Financial Information	Turnover figures, growth projections or targets for companies, market share data, investment announcements, and any other relevant financial indicators mentioned in the news articles enable clients to make informed decisions about investments, partnerships, and expansions, and to assess the financial health of companies and industries.
4	Personnel and Leadership Changes	From news articles, information about new hires, appointments, resignations, promotions, and changes in leadership roles within companies. This information is valuable for business clients as it allows them to stay updated on key personnel changes within their industry and competitor organisations.
5	Industry Trends and Market Dynamics	News articles provide information about emerging trends, technological advancements, and competitive landscapes. This information aids business clients in staying informed about developments that affect their operations and competitiveness. By analysing these trends, clients can identify growth opportunities, anticipate challenges, and adjust business strategies.
6	Regulatory and Compliance Issues	News articles provide updates on regulatory changes, compliance requirements, legal challenges, and government policies affecting businesses. This information is essential for businesses to stay informed about changes in regulations, ensure compliance, mitigate risks, and maintain a positive reputatio

- Reading each news article and extracting key details is a time-consuming process that requires a significant number of company resources.
- Additionally, the company relies on a taxonomy to categorise business opportunities under specific topics. However, this taxonomy was created years ago and has not been updated since, potentially causing the omission of numerous opportunities for their clients.

The company is undergoing a digital transformation to automate the extraction of business opportunities from news articles. Seeking a tailored solution, they approached us for assistance. After a series of meetings and discussions aimed at defining the desired information for their automated Business Dashboard, which would facilitate the exploration of business opportunities, the specific requirements for extracting key business information were finalised. For instance, each news article contains one or several contextual keywords, such as *'building'*, *'infrastructure'*, *'architecture'*, *'engineering'*, *'construction site'*, *'project management'*, *'contractors'*, etc. These keywords enable the categorisation of articles into specific topics (e.g. Construction, Healthcare, Education, etc.). Within these topics, various business activities occur or are anticipated in the future.

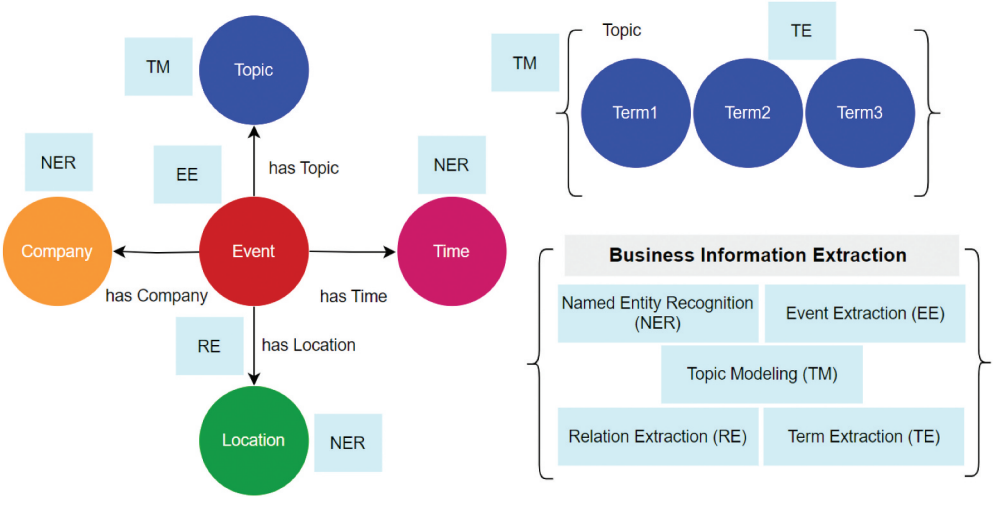

Figure 4. Business information requirements mapped to IE tasks.

These activities require details such as location, time, and persons involved. After identifying the essential business information requiring extraction, we mapped IE tasks, drawn from the literature review, to fulfill each of these requirements (see Figure 4). Contextual keyword identification was considered as TE, while topic extraction was performed using TM. Business activities were identified as events (EE), necessitating the extraction of corresponding named entities (NER) such as location and person, linked by RE (e.g. 'has Location', 'has Company').

5.2. Implementing RAG and LLM for extracting identified IE tasks

After identifying the necessary information for constructing business opportunities and determining the relevant IE tasks, it is time to leverage the RAG with LLM model. This system allows us to exploit multiple IE tasks simultaneously, eliminating the need for building Business IE systems from scratch or extensive training with business datasets. To facilitate interactive exploration of business opportunities, we opted to design a Business Chatbot Assistant based on insights gathered from the existing literature. This choice allows seamless interaction with business users, enhancing their ability to explore information effectively. LLMs, renowned for their expertise in complex reasoning tasks across various domains, offered a promising avenue in constructing a Business Chatbot Assistant. We explored several publicly available LLM releases such as BLOOM (Le Scao et al., 2022), Falcon (Almazrouei et al., 2023), GPT-4 (Achiam et al., 2023), Llama2 (Touvron et al., 2023), and Chinchilla (Hoffmann et al., 2022). Among these, we opted for Llama2, an updated version trained on diverse data sources, renowned for its enhanced performance compared to existing open-source models (Touvron et al., 2023). However, to customise it for business-specific tasks and effectively utilise business data extracted from news articles, we integrated RAG technology into our solution (see Figure 5). This strategic integration allows the Chatbot Assistant to better comprehend and respond to business-related queries with the extracted information. The range of business-related queries that

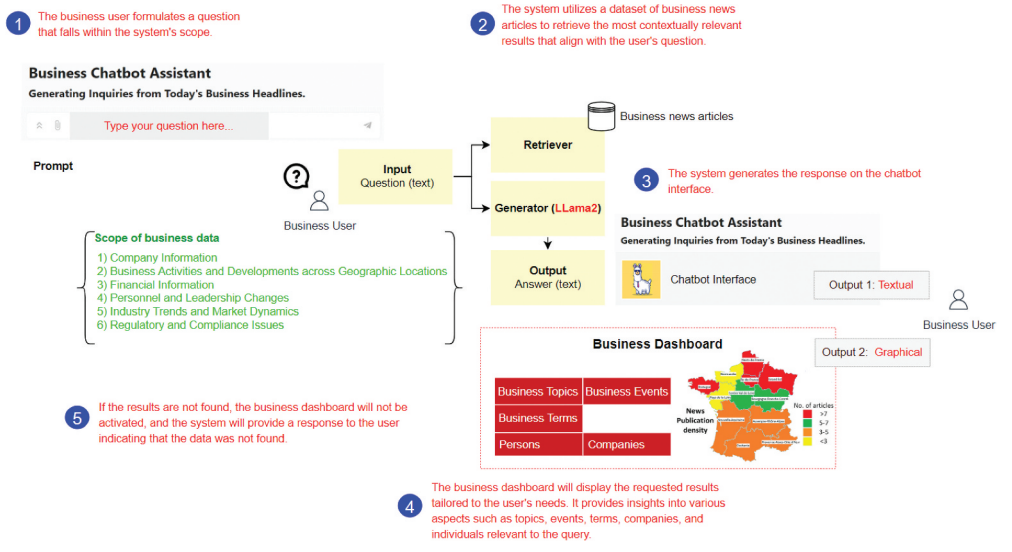

Figure 5. Graphical representation of a scenario in our case study: a business user queries the RAG with llm-based system, leveraging external business news articles dataset to generate a response.

the Chatbot Assistant can handle depends on the quality and comprehensiveness of the provided business dataset. In this case, the company provided us with a dataset comprising 2,845 news articles published on online platforms in March 2023. Each article contains crucial business entities such as individuals, locations, and more (see Figure 6). The highlighting is purely for reader comprehension; no such tags are present in the system. This dataset holds details about companies, their operations across diverse geographic regions, financial metrics, personnel shifts, industry trends, and regulatory updates.

Thales ORG will recruit 450 people in Brittany GPE in 2023. Thales (Courbevoie NORP , 92) is a specialist in radiocommunication, satellite communication, cybersecurity and electronic warfare. It will recruit 450 people in Brittany GPE in 2023, where it has sites in Étrelles GPE (35), currently being expanded, and in Brest (29). Thales is hiring in particular in R&D, supply chain, electronic production, mechanical production, etc. Publication date: 2023-03-01

"The German Ionity continues its development in France GPE with the deployment of two charging stations in Mornas Les Adrets and Mornas Villageionity (Munich GPE , Germany GPE) specializes in high-intensity charging stations. The company is continuing its development in France GPE with the deployment of two stations in Mornas Les Adrets ORG (84) and Mornas Village (84). The company plans to double its number of stations in order to reach the milestone of 1,000 sites in Europe LOC . It is in France GPE that Ionity ORG has the largest share of its charging stations, with 120 stations, including 10 under construction, and 576 active charging points. Publication date: 2023-03-01

Figure 6. Sample records from a dataset of news articles containing information on business events: each business news text includes details of organisations (ORG), locations (GPE and NORP), and people (PER) involved.Spacy Tool

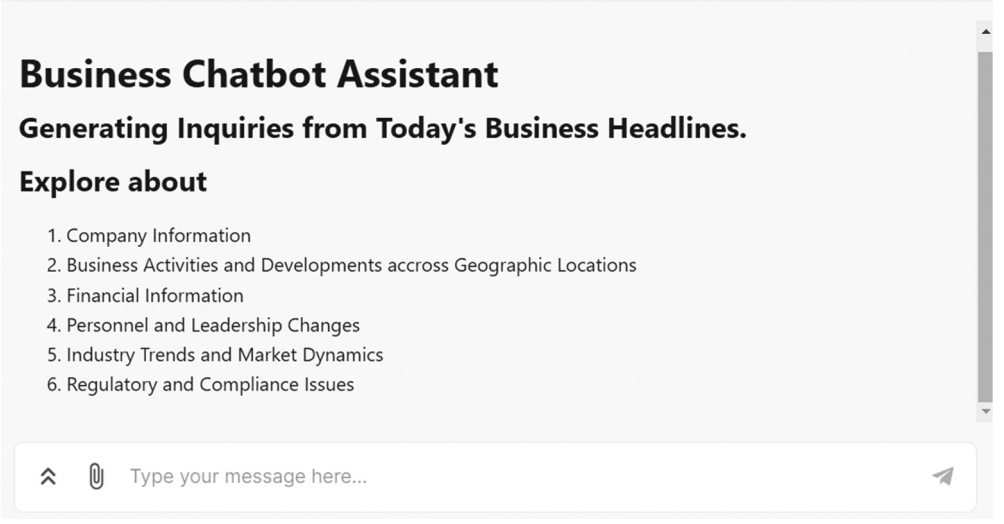

Figure 7. Design of a business chatbot assistant to enable business users to ask questions related to news articles.

To create a Business Chatbot Assistant, as depicted in Figure 7, we utilised the scenario outlined in Figure 5 as a foundation. This involved implementing the Algorithm 1 that links LLM (specifically, Llama2 in our case) with a business dataset, enabling the development of a RAG application. The goal is to utilise Llama2 to extract business information and respond to queries. The process begins with the installation of necessary packages and the importation of required modules, including prompts for Llama2, and logging in to Hugging Face (https://huggingface.co/) for model access. Initialisation of HuggingFaceLLM and LangchainEmbedding (https://python.langchain.com/) with ServiceContext facilitates the management of embeddings and components. News articles are then loaded, embeddings generated, and an index constructed. Queries are executed against the index using the query engine, which leverages the Llama2 model and embeddings to retrieve pertinent answers. The resulting response, containing extracted business information, is further processed, or displayed as needed, ultimately furnishing a text-based answer to the inquiry. In this case study, textual answers are generated via the Business Chatbot Assistant terminal (see Figure 7), while the Business Dashboard offers graphical visualisations to facilitate user interaction. For example, Figure 8 presents a general view showcasing insights into trending and dormant topics, key business events, popular individuals, trending terms, and notable companies extracted from the news dataset. The specific view further drills down into detailed business opportunities linked to particular topics, events, and terms. Additionally, a geographical map highlights the regional distribution density of news articles within the country. Throughout, the Chatbot Assistant terminal (see Figure 7) remains active for continuous user engagement, ensuring seamless interaction for business users.

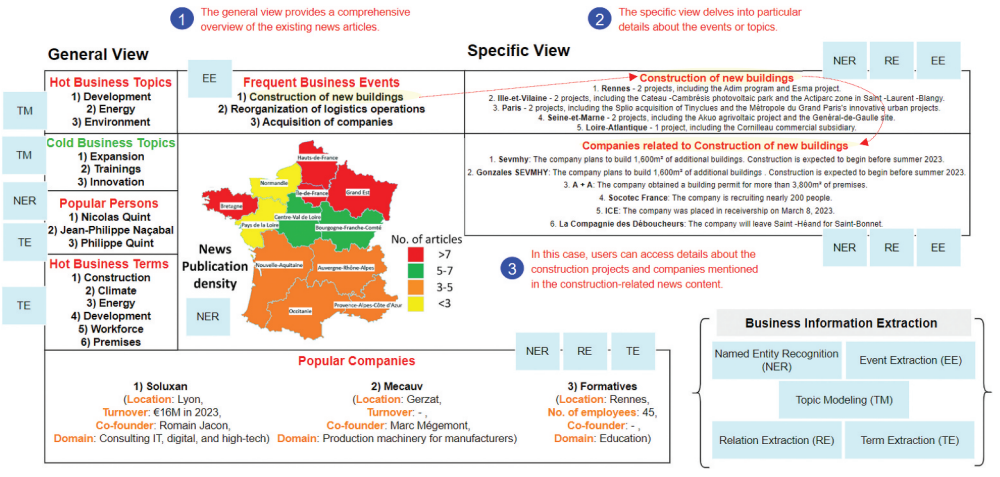

Figure 8. A prototype business dashboard designed to execute various business IE tasks, including NER, RE, EE, TE, and TM, for extracting information across two distinct views.

5.3. System evaluation

The system, designed using Algorithm 1 for Business IE through RAG with LLM, was evaluated to measure its performance. Precision, Recall, and Accuracy were chosen as the key metrics for this evaluation, as they are standard indicators for assessing LLM-based systems (Sajjadi et al., 2018). These metrics are crucial for gauging the quality of information provided by the Chatbot, which hinges on both the relevance and completeness of the retrieved data. The evaluation specifically targeted the EE tasks within the RAG with LLM framework, encompassing NER, RE, TE, and TM. Given that each event involves NER through RE and includes detailed topics, evaluating at the event level is essential. Information Components (ICs) such as organisations, individuals, monetary amounts, topics, actions, locations, and dates were defined to systematically assess performance. Precision measures the retrieval process's effectiveness by calculating the proportion of relevant ICs among all retrieved components, highlighting the system's ability to filter out irrelevant data (see Equation 1). Recall evaluates the proportion of relevant ICs successfully retrieved from the total available, reflecting the system's capacity to capture all pertinent information (see Equation 2). Accuracy provides an overall view of the system's performance by determining the proportion of correctly identified ICs, both relevant and irrelevant, out of all evaluated ICs, thus reflecting its effectiveness in retrieval and classification (see Equation 3). For testing, a dataset of 120 questions related to business events was manually curated, with a subset shown in Table 6. Table 7 summarises the evaluation metrics for the Chatbot using the RAG with Llama2 model, showing a Precision of 89%, indicating effective identification of relevant information among positive predictions. The Recall of 84.5% demonstrates the model's efficiency in capturing a significant portion of relevant instances, while the overall Accuracy of 85% confirms its reliable performance in delivering accurate results across all classifications.

Table 6. Sample records of user queries and system generated answers for RAG with LLM evaluation.

No.	Question	Involved IE tasks	Answer
1	Which companies exhibit the highest level of interest in construction-related projects?	**NER**: Identifies company names, and; **TE**: Recognises construction-related terms.	1. FIAGEO GROUPE DELBOS (Loire) 2. A + A (Lot -et-Garonne) 3. SOCOTEC FRANCE (Aquitaine) 4. LA COMPAGNIE DES DEBOUCHEURS (Loire)
2	In which locations are recruitment activities most concentrated?	**NER**: Identifies locations, and; **RE**: Extracts the relationship between recruitment activities and locations.	1. Nord 2. Haute-Garonne 3. Essonne 4. Val-d'Oise 5. Ain 6. Haute-Savoie
3	Who are the individuals demonstrating the greatest involvement in environmental projects?	**NER**: Identifies individual names, and; **TE**: Identifies environmental project-related terms.	1. Ms. Anne Lassman -Trappier, representative of France Nature Environnement. 2. Ms. Hélène Valade – environmental development director at LVMH. 3. Ms. Laetitia Vasseur – representative of Stop Planned Obsolescence.
4	What was the geographic location of Sourdillon, the company recently liquidated by the court of Tours?	**NER**: Recognises the company name, and; **TE**: Identifies terms related to geographic locations.	The geographic location of Sourdillon, the company recently liquidated by the court of Tours, was Maine-et-Loire.
5	How many employees did the Metz start-up 45-8 Energy have at the time of fundraising?	**NER**: Recognises the company name, and; **TE**: Identifies numerical values representing the number of employees.	45-8 Energy had 15 employees at the time of fundraising.
6	What is the primary focus of 45-8 Energy?	**TM**: Determines the main focus areas of the company based on the text content.	45-8 Energy is focused on developing renewable energy projects, particularly hydrogen, to support the European Union's goal of reducing greenhouse gas emissions. They are also involved in the construction of photovoltaic parks and the installation of heat pumps and photovoltaic solar panels for individual homes.
7	Which company is acquiring land in Fay-aux-Loges?	**NER**: Recognises the company name, and; **TE**: Identifies terms related to land acquisition.	The company acquiring land in Fay-aux-Loges is NEOEN.
8	When was the decision to liquidate Sourdillon pronounced by the commercial court of Tours?	**NER**: Identifies the company name, and; **RE**: Extracts the relationship between the company and the decision announcement.	The decision to liquidate Sourdillon was pronounced by the commercial court of Tours on March 8, 2023.
9	What new activities is U-Space, planning to undertake in its 1,000 m^2 site?	**NER**: Identifies the company name, and; **TE**: Identifies terms related to new activities.	U-Space is planning to undertake new activities in its 1,000 m^2 site, including the development of an outdoor park and the creation of a multipurpose room with a catering offer. Additionally, the company is planning to launch a call to residents with specifications drawn up jointly with the city, with the aim of accommodating between 250 and 300 residents in the first wave, planned for spring 2023.

(*Continued*)

Table 6. (Continued).

No.	Question	Involved IE tasks	Answer
10	What is the primary focus of U-Space's contract with the company Grasp de Bachy?	**NER**: Recognises company names, and; **TE**: Identifies terms related to the contract focus.	U-Space's contract with Grasp de Bachy focuses on the development of algorithms and the processing of data from satellite surveys to improve global monitoring of aerosols and their distribution in cities, particularly in relation to weather conditions.
11	How many members is Burger King looking to recruit for its future restaurant in Douvres-la-Délivrande?	**NER**: Identifies the company name, and; **TE**: Identify numerical values representing the number of recruits.	Burger King is recruiting 50 people for its new restaurant in Gujan -Mestras.
12	Have there been any recent mergers within the agriculture sector?	**TM**: Identifies topics related to mergers within the agriculture sector based on the text content.	Yes, recent mergers within the agriculture sector are evident. Scobat, a cooperative specialising in structural work and civil engineering, is extending its Trégueux site and seeking land for a production and recycling platform. Questembert, a citizen energy project, submitted statutes for their energy ambition, aiming for wind or photovoltaic projects. Enedis, an electricity network manager, seeks 63 employees in Picardy. These developments suggest mergers with a focus on innovation and sustainability.

Table 7. Evaluation of chatbot performance using RAG with Llama2.

Model	Precision (%)	Recall (%)	Accuracy (%)
Chatbot using RAG with Llama2	89	84.5	85

$$Precision = \frac{Number\ of\ Relevant\ ICs\ Retrieved}{Total\ Number\ of\ ICs\ Retrieved} \tag{1}$$

$$Recall = \frac{Number\ of\ Relevant\ ICs\ Retrieved}{Total\ Number\ of\ Relevant\ ICs\ Available} \tag{2}$$

$$Accuracy = \frac{Number\ of\ Correctly\ Identified\ ICs}{Total\ Number\ of\ ICs\ Evaluated} \tag{3}$$

Algorithm 1: Implementing RAG with LLM for Business IE

Objective: Business IE and Querying using LLama2 and Hugging Face Models.
1) Input: Business news articles, question
2) Process:
2.1 Install necessary packages:
Transformers, langchain, sentence_transformers, llama_index, llama-index-llms-huggingface, llama-index-embeddings-langchain
2.2 Import required modules and packages:
VectorStoreIndex, HuggingFaceLLM, SimpleInputPrompt, VectorStoreIndex, SimpleDirectoryReader
2.3 Define system_prompt and query_wrapper_prompt for LLama2. These prompts provide context and structure for the language model.
2.4 Log in to Hugging Face using huggingface-cli to access models from the Hugging Face model hub.
2.5 Initialise HuggingFaceLLM.
2.6 Initialise LangchainEmbedding and create ServiceContext:
- Choose an embedding model (e.g. HuggingFaceEmbeddings) to convert text into vectors.
- Create a ServiceContext to manage embeddings, LLama2, and other components.
2.7 Create VectorStoreIndex from documents:
- Load documents from a specified directory using SimpleDirectoryReader.
- Utilise the ServiceContext to generate embeddings for each document and build the index.
2.8 Query the index with a specific question:
- Use the query_engine to search for relevant information based on the provided question.
- The query engine leverages the LLama2 model and embeddings to retrieve accurate answers.
2.9 Retrieve the response produced by the query engine, which includes the relevant information extracted from the documents.
2.10 Process or present the response as required for the application or user interface.
3) Output: The text generated as the answer to the question posed.

6. Discussion

Extracting relevant business information from diverse textual datasets, which span structured, semi-structured, and unstructured formats, poses a significant challenge (Netzer et al., 2012; Sarawagi, 2008; Sprague, 2004). This complexity arises from the need to handle varying information structures and apply a range of IE tasks, including NER, RE, EE, TE, and TM, tailored to the specific needs of the organisation. Typically, implementing these IE tasks requires selecting and deploying separate systems from a broad array of available tools, which can be intricate, resource-intensive, and time-consuming. GenAI presents a compelling solution by leveraging LLMs that are trained on extensive datasets (Touvron et al., 2023). However, these models are often not specifically tailored for domain-specific tasks. To address this, the literature explores how contextualising LLMs through the RAG framework can effectively adapt them for domain-specific NLP tasks. While LLMs are technologically advanced and versatile, their effectiveness in specialised domains hinges on the careful selection of relevant datasets and their integration via RAG. NLP tasks such as QA, Text Generation and Summarisation, Information Retrieval and Extraction, Text Analysis and Processing, and Decision Making are integral across various disciplines. The effectiveness of these tasks in specific fields relies on adapting LLMs to the unique requirements of each domain through precise data integration.

For the business community, while state-of-the-art RAG with LLM solutions offer considerable potential, their application in the business domain remains relatively underexplored. This gap presents a significant opportunity for further R&D. By applying insights from existing RAG–LLM studies and adapting them to business-specific contexts,

organisations can greatly enhance their NLP capabilities, improve decision-making, and achieve greater operational efficiency. We validated this approach through a practical case study that focused on a company's digital transformation initiative. The goal was to automate the extraction of business opportunities from news articles. Our case study demonstrated the effectiveness of integrating the RAG with LLM model, which enabled the system to provide tailored, business-specific responses to user inquiries. This functionality is delivered through two key methods: textual views available via the Chatbot Assistant terminal for user interaction, and graphical views on the Business Dashboard, which offer enriched data visualisation. While the Dashboard serves as a proof-of-concept, its primary aim is to present clients with an intuitive and interactive display of relevant information. The approach offers several compelling advantages, including:

- **Versatility and Adaptability**: The system's design allows for easy adaptation to various business applications and disciplines. By simply swapping the dataset repository to include domain-specific information, the system can rapidly deploy customised IE solutions across different domains with minimal adjustments.
- **Efficient Deployment**: Utilising the existing RAG and LLM infrastructure significantly reduces the time, cost, and manpower required for implementation. This streamlined approach enables businesses to quickly set up tailored IE solutions without extensive development efforts.
- **Dynamic Information Updates and Scalability**: The RAG approach provides continuous access to the latest data, ensuring responses are current and relevant. Additionally, the system's scalability accommodates increasing data volumes and evolving business needs, maintaining effectiveness and efficiency as requirements expand.

While our proposed system offers numerous benefits, it also has several limitations that should be addressed for future improvements.

- **Summarisation Process**: The system currently generates insights by providing summarised versions of business articles in response to inquiries about business opportunities. While this helps in delivering concise information, it may impact the comprehensiveness of the information provided to users. Further research is needed to optimise the summarisation process to enhance both accuracy and completeness. Our study did not include an evaluation of the text summarisation component, which could be a valuable area for future investigation.
- **Evaluation Metrics**: Our evaluation primarily relied on standard metrics such as Precision, Recall, and Accuracy. However, these metrics alone may not fully capture the performance nuances of the RAG–LLM system. Future research should involve extensive datasets and incorporate comprehensive evaluation metrics (De Stefano et al., 2024; Y. Gao et al., 2023) to better assess and validate the performance of RAG–LLM systems in diverse scenarios.
- **Model Limitation**: Our case study used a single LLM, Llama2, chosen for its open-source availability and demonstrated performance in the literature. However, relying on one model limits the scope of our findings. Future studies should explore other LLMs to compare performance and identify which models are most effective for business applications.

- **Retrieval Time Impact**: Increasing the size of the dataset can lead to longer retrieval times due to the greater volume of documents the LLM needs to process. Our study did not address how the retrieval time is affected by dataset size. Future research should examine the impact of larger datasets on response times and explore potential optimisations to manage this challenge.

Addressing these limitations will be crucial for advancing the effectiveness and efficiency of RAG with LLM-based systems in business contexts.

7. Conclusion

The integration of RAG with LLMs presents a promising approach to enhancing IE tasks, particularly in the business domain. This integration capitalises on the capabilities of LLMs grounded in NLP principles, offering a cost-effective and efficient solution for performing multiple IE tasks simultaneously. Although the integration of RAG with LLMs has experienced significant growth in recent years, the current literature lacks a comprehensive classification of RAG with LLMs studies based on NLP tasks and disciplines. Such a classification could provide valuable insights and guidance for applications in the business sector. Moreover, there is a scarcity of research focusing on developing applications using RAG with LLMs for Business-related IE domains. To bridge these gaps, this paper reviews existing studies on Business IE and associated tasks, providing insights into the potential applications of RAG with LLMs. Additionally, it offers a novel real-world case study demonstrating the practical implementation of this integration in developing a Business IE application. The primary objective is to showcase the versatility and effectiveness of RAG with LLMs in the business discipline, fostering further exploration and development in this field of research.

Acknowledgments

The authors gratefully acknowledge the financial support provided by the University of the West of England, Bristol, and the French National Research Agency (ANR).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the Agence Nationale de la Recherche [Salary]; University of the West of England [Salary].

ORCID

Muhammad Arslan http://orcid.org/0000-0003-3682-7002
Saba Munawar http://orcid.org/0009-0007-1255-9425
Christophe Cruz http://orcid.org/0000-0002-5611-9479

Dataset availability

https://drive.google.com/file/d/18UB-TamXvCFpq0edfH7EVPjBl9Ec34dC/view?usp=sharing

References

Abdullah, M.H.A., Aziz, N., Abdulkadir, S.J., Alhussian, H.S.A., & Talpur, N. (2023). Systematic literature review of information extraction from textual data: Recent methods, applications, trends, and challenges. *Institute of Electrical and Electronics Engineers Access*, *11*, 10535–10562. https://doi.org/10.1109/ACCESS.2023.3240898

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:230308774.

Adnan, K., & Akbar, R. (2019). Limitations of information extraction methods and techniques for heterogeneous unstructured big data. *International Journal of Engineering Business Management*, *11*, 1847979019890771. https://doi.org/10.1177/1847979019890771

Ahmad, S.R. (2024). Enhancing multilingual information retrieval in mixed human resources environments: A RAG model implementation for multicultural enterprise. *arXiv Preprint arXiv: 240101511*. https://doi.org/10.48550/arXiv.2401.01511

Alawwad, H.A., Alhothali, A., Naseem, U., Alkhathlan, A., & Jamal, A. (2024). Enhancing textbook question answering task with large language models and retrieval augmented generation. *arXiv Preprint arXiv: 240205128*. https://doi.org/10.48550/arXiv.2402.05128

Al Ghadban, Y., Lu, H.Y., Adavi, U., Sharma, A., Gara, S., Das, N., … & Hirst, J.E. (2023). Transforming healthcare education: Harnessing large language models for frontline health worker capacity building using retrieval-augmented generation. *medRxiv*, 2023–12. https://doi.org/10.1101/2023.12.15.23300009

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., & Penedo, G. (2023). *The falcon series of open language models*. arXiv preprint arXiv:2311.16867. https://huggingface.co/tiiuae/

Al-Okaily, A., Teoh, A.P., & Al-Okaily, M. (2023). Evaluation of data analytics-oriented business intelligence technology effectiveness: An enterprise-level analysis. *Business Process Management Journal*, *29*(3), 777–800. https://doi.org/10.1108/BPMJ-10-2022-0546

Anthropic. (2023). Retrieved March 28, 2024, from https://www.anthropic.com/news/introducing-claude

Arendarenko, E., & Kakkonen, T. (2012). Ontology-based information and event extraction for business intelligence. In A. Ramsay & G. Agre (Eds.), Artificial Intelligence: Methodology, Systems, and Applications. AIMSA 2012. Lecture Notes in Computer Science (Vol. 7557, pp. 89–102). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-33185-5_10

Arslan, M., & Cruz, C. (2022). Extracting business insights through dynamic topic modeling and NER. In Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2022) - Volume 2: KEOD, 215–222. https://doi.org/10.5220/0011552900003335

Arslan, M., & Cruz, C. (2024). Business-RAG: Information Extraction for Business Insights. In *ICSBT* (p. 88). Dijon, France. https://doi.org/10.5220/0012812800003764

Arslan, M., Ghanem, H., Munawar, S., & Cruz, C. (2024). A survey of RAG with LLMs. In Proceedings of the 28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES), Seville, Spain (pp. 11–13).

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv Preprint arXiv: 231011511*. https://doi.org/10.48550/arXiv.2310.11511

Bellan, P., Dragoni, M., & Ghidini, C. (2022, September). Extracting business process entities and relations from text using pre-trained language models and in-context learning. In International Conference on Enterprise Design, Operations, and Computing (pp. 182–199). Springer International Publishing, Cham.

Bucur, M. (2023). *Exploring large language models and retrieval augmented generation for automated form filling* [Bachelor's thesis]. University of Twente.

Bzhalava, L., Kaivo-Oja, J., & Hassan, S.S. (2024). Digital business foresight: Keyword-based analysis and CorEx topic modeling. *Futures*, *155*, 103303. https://doi.org/10.1016/j.futures.2023.103303

Chen, W., Hu, H., Saharia, C., & Cohen, W.W. (2022). Re-imagen: Retrieval-augmented text-to-image generator. *arXiv Preprint arXiv: 220914491*. https://doi.org/10.48550/arXiv.2209.14491

Colverd, G., Darm, P., Silverberg, L., & Kasmanoff, N. (2023). FloodBrain: Flood disaster reporting by web-based retrieval augmented generation with an LLM. *arXiv Preprint arXiv: 231102597*. https://doi.org/10.48550/arXiv.2311.02597

Cunningham, H. (2005). Information extraction, automatic. *Encyclopedia of Language and Linguistics*, *3*(8), 10.

de Almeida Bordignon, A.C., Thom, L.H., Silva, T.S., Dani, V.S., Fantinato, M., & Ferreira, R.C.B. (2018, June). Natural language processing in business process identification and modeling: A systematic literature review. In SBSI'18: XIV Brazilian Symposium on Information Systems Caxias do Sul Brazil June 4 - 8, 2018 (pp. 1–8).

De Stefano, G., Pellegrino, G., & Schönherr, L. (2024). Rag and roll: An end-to-end evaluation of indirect prompt manipulations in llm-based application frameworks. *arXiv Preprint arXiv: 240805025*. https://doi.org/10.48550/arXiv.2408.05025

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv Preprint arXiv: 181004805*. https://doi.org/10.48550/arXiv.1810.04805

Ding, W., Cao, Y., Zhao, D., Xiao, C., & Pavone, M. (2023). RealGen: Retrieval augmented generation for controllable traffic scenarios. *arXiv Preprint arXiv: 231213303*. https://doi.org/10.48550/arXiv.2312.13303

Dor, L.E., Gera, A., Toledo-Ronen, O., Halfon, A., Sznajder, B., Dankin, L., & Slonim, N. (2019, November). Financial event extraction using Wikipedia-based weak supervision. In Proceedings of the Second Workshop on Economics and Natural Language Processing (pp. 10–15). https://doi.org/10.48550/arXiv.1911.10783

Douzon, T., Duffner, S., Garcia, C., & Espinas, J. (2022, May). Improving information extraction on business documents with specific pre-training tasks. In *International workshop on document analysis systems* (pp. 111–125). Springer International Publishing.

Du, X., & Ji, H. (2022). Retrieval-augmented generative question answering for event argument extraction. *arXiv Preprint arXiv: 221107067*. https://doi.org/10.48550/arXiv.2211.07067

Esser, D., Muthmann, K., & Schuster, D. (2013, September). Information extraction efficiency of business documents captured with smartphones and tablets. In DocEng '13: ACM Symposium on Document Engineering 2013 Florence Italy September 10 - 13, 2013 (pp. 111–114).

Esser, D., Schuster, D., Muthmann, K., & Schill, A. (2014, April). Few-exemplar information extraction for business documents. In Proceedings of the 16th International Conference on Enterprise Information Systems - Volume 3: ICEIS, 293-298, 2014 , Lisbon, Portugal (Vol. 2. pp. 293–298). SCITEPRESS.

Fan, R.Z., Fan, Y., Chen, J., Guo, J., Zhang, R., & Cheng, X. (2023). RIGHT: Retrieval-augmented generation for mainstream hashtag recommendation. In European Conference on Information Retrieval (pp. 39–55). Cham: Springer Nature Switzerland.

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative ai. *Business & Information Systems Engineering*, *66*(1), 111–126. https://doi.org/10.1007/s12599-023-00834-7

Gao, X., Murugesan, S., & Lo, B. (2005, October). Extraction of keyterms by simple text mining for business information retrieval. In IEEE International Conference on e-Business Engineering (ICEBE'05), Beijing (pp. 332–339). IEEE.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv Preprint arXiv: 231210997*. https://doi.org/10.48550/arXiv.2312.10997

Ge, J., Sun, S., Owens, J., Galvez, V., Gologorskaya, O., Lai, J.C., & Lai, K. (2023). Development of a liver disease-specific large language Model chat interface using retrieval augmented generation. *medRxiv*. https://doi.org/10.1101/2023.11.10.23298364

Geletka, M., Bankovic, M., Melus, D., Scavnická, S., Stefánik, M., & Sojka, P. (2022). Information extraction from business documents: A case study. In Proceedings of Recent Advances in Slavonic Natural Language (pp. 35–46). https://nlp.fi.muni.cz/raslan/2022/paper18.pdf

Guo, Y., Li, Z., Jin, X., Liu, Y., Zeng, Y., Liu, W., & Cheng, X. (2023). Retrieval-augmented code generation for universal information extraction. *arXiv Preprint arXiv: 231102962*. https://doi.org/10.48550/arXiv.2311.02962

Hamdi, A., Carel, E., Joseph, A., Coustaty, M., & Doucet, A. (2021, September). Information extraction from invoices. In International Conference on Document Analysis and Recognition (pp. 699–714). Springer International Publishing, Cham.

Han, S., Hao, X., & Huang, H. (2018). An event-extraction approach for business analysis from online Chinese news. *Electronic Commerce Research and Applications*, *28*, 244–260. https://doi.org/10.1016/j.elerap.2018.02.006

Han, Z.F., Lin, J., Gurung, A., Thomas, D.R., Chen, E., Borchers, C., Gupta, S., & Koedinger, K.R. (2024). Improving assessment of tutoring practices using retrieval-augmented generation. *arXiv Preprint arXiv: 240214594*. https://doi.org/10.48550/arXiv.2402.14594

He, X., Tian, Y., Sun, Y., Chawla, N.V., Laurent, T., LeCun, Y., & Hooi, B. (2024). G-Retriever: Retrieval-augmented generation for textual graph understanding and question answering. *arXiv Preprint arXiv: 240207630*. https://doi.org/10.48550/arXiv.2402.07630

Hedberg, J., & Furberg, E. (2023). Automated extraction of insurance policy information: Natural language processing techniques to automate the process of extracting information about the insurance coverage from unstructured insurance policy documents.

Hillebrand, L., Deußer, T., Dilmaghani, T., Kliem, B., Loitz, R., Bauckhage, C., & Sifa, R. (2022, August). Kpi-bert: A joint named entity recognition and relation extraction model for financial reports. In 26th International Conference on Pattern Recognition (ICPR), Montréal Québec, August 21-25, 2022 (pp. 606–612). IEEE.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., & Sifre, L. (2022). Training compute-optimal large language models. *arXiv Preprint arXiv: 220315556*. https://doi.org/10.48550/arXiv.2203.15556

Hu, M., Zhao, X., Wei, J., Wu, J., Sun, X., Li, Z., & Zhang, Y. (2023, October). rT5: A retrieval-augmented pre-trained Model for ancient Chinese entity description generation. In CCF International Conference on Natural Language Processing and Chinese Computing (pp. 736–748). Springer Nature Switzerland, Cham.

Huang, W., Lapata, M., Vougiouklis, P., Papasarantopoulos, N., & Pan, J. (2023, November). Retrieval augmented generation with rich answer encoding. In Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Nusa Dua, Bali (pp. 1012–1025).

Jacobs, G., & Hoste, V. (2022). SENTiVENT: Enabling supervised information extraction of company-specific events in economic and financial news. *Language Resources and Evaluation*, *56*(1), 225–257. https://doi.org/10.1007/s10579-021-09562-4

Jeong, M., Sohn, J., Sung, M., & Kang, J. (2024). Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, *40* (Supplement_1), i119–i129. https://doi.org/10.1093/bioinformatics/btae238

Jiang, S., Tang, W., Chen, X., Tanga, R., Wang, H., & Wang, W. (2023). Raucg: Retrieval-augmented unsupervised counter narrative generation for hate speech. *arXiv Preprint arXiv: 231005650*. https://doi.org/10.48550/arXiv.2310.05650

Jimeno Yepes, A., You, Y., Milczek, J., Laverde, S., & Li, L. (2024). Financial report chunking for effective retrieval augmented generation. *arXiv E-Prints, arXiv-2402*. https://doi.org/10.48550/arXiv.2402.05131

Kagaya, T., Yuan, T.J., Lou, Y., Karlekar, J., Pranata, S., Kinose, A., & You, Y. (2024). RAP: Retrieval-augmented planning with contextual memory for multimodal LLM agents. *arXiv Preprint arXiv: 240203610*. https://doi.org/10.48550/arXiv.2402.03610

Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023, July). Large language models struggle to learn long-tail knowledge. In ICML'23: International Conference on Machine

Learning Honolulu Hawaii USA (pp. 15696–15707). PMLR. https://doi.org/10.5555/3618408.3619049

Khaldi, H., Benamara, F., Abdaoui, A., Aussenac-Gilles, N., & Kang, E. (2021, June). Multilevel entity-informed business relation extraction. In International Conference on Applications of Natural Language to Information Systems (pp. 105–118). Springer International Publishing, Cham.

Kim, J., Choi, S., Amplayo, R.K., & Hwang, S.W. (2020, December). Retrieval-augmented controllable review generation. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (pp. 2284–2295).

Kim, J., & Min, M. (2024). From RAG to QA-RAG: Integrating generative AI for pharmaceutical regulatory compliance process. *arXiv Preprint arXiv: 240201717*. https://doi.org/10.48550/arXiv.2402.01717

Korger, A., & Baumeister, J. (2021, September). Rule-based semantic relation extraction in regulatory documents. *LWDA*, 26–37. https://ceur-ws.org/Vol-2993/paper-03.pdf

Kraus, S., Jones, P., Kailer, N., Weinmann, A., Chaparro-Banegas, N., & Roig-Tierno, N. (2021). Digital transformation: An overview of the current state of the art of research. *SAGE Open*, *11*(3), 21582440211047576. https://doi.org/10.1177/21582440211047576

La Fleur, A., Teymourian, K., & Paschke, A. (2015, September). Complex event extraction from real-time news streams. In SEMANTiCS '15: 11th International Conference on Semantic Systems Vienna Austria September 16 - 17, 2015 (pp. 9–16).

Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., & Al-Shaibani, M.S. (2022). Bloom: A 176b-parameter open-access multilingual language model.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N. Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

Li, H., Su, Y., Cai, D., Wang, Y., & Liu, L. (2022). A survey on retrieval-augmented text generation. *arXiv Preprint arXiv: 220201110*. https://doi.org/10.48550/arXiv.2202.01110

Li, S., Park, S., Lee, I., & Bastani, O. (2023). TRAC: Trustworthy retrieval augmented chatbot. *arXiv Preprint arXiv: 230704642*. https://doi.org/10.48550/arXiv.2307.04642

Lin, D. (2024). Revolutionizing retrieval-augmented generation with enhanced PDF structure recognition. *arXiv Preprint arXiv: 240112599*. https://doi.org/10.48550/arXiv.2401.12599

Lozano, A., Fleming, S.L., Chiang, C.C., & Shah, N. (2023). Clinfo. ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature. In PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024, Fairmont Orchid - Hawaii, Puako, Hawaï, USA (pp. 8–23).

Lu, S., Duan, N., Han, H., Guo, D., Hwang, S.W., & Svyatkovskiy, A. (2022). Reacc: A retrieval-augmented code completion framework. *arXiv Preprint arXiv: 220307722*. https://doi.org/10.48550/arXiv.2203.07722

Luo, J., Pan, X., & Zhu, X. (2015). Identifying digital traces for business marketing through topic probabilistic model. *Technology Analysis & Strategic Management*, *27*(10), 1176–1192. https://doi.org/10.1080/09537325.2015.1061118

Manathunga, S.S., & Illangasekara, Y.A. (2023). Retrieval augmented generation and Representative vector summarization for large unstructured textual data in medical education. *arXiv Preprint arXiv: 230800479*. https://doi.org/10.48550/arXiv.2308.00479

Mariani, M., & Dwivedi, Y.K. (2024). Generative artificial intelligence in innovation management: A preview of future research developments. *Journal of Business Research*, *175*, 114542. https://doi.org/10.1016/j.jbusres.2024.114542

Markey, N., El-Mansouri, I., Rensonnet, G., van Langen, C., & Meier, C. (2024). From RAGs to riches: Using large language models to write documents for clinical trials. *arXiv Preprint arXiv: 240216406*. https://doi.org/10.48550/arXiv.2402.16406

Martinez-Rodriguez, J.L., Hogan, A., Lopez-Arevalo, I., & Hotho, A. (2020). Information extraction meets the semantic web: A survey. *Semantic Web*, *11*(2), 255–335. https://doi.org/10.3233/SW-180333

McGillivray, B., Jenset, G., & Heil, D. (2020). Extracting keywords from open-ended business survey questions. *Journal of Data Mining & Digital Humanities*, *2020*(Project). https://doi.org/10.46298/jdmdh.5077

Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., and Scialom, T. (2023). Augmented language models: A survey. *arXiv Preprint arXiv: 230207842*. https://doi.org/10.48550/arXiv.2302.07842

Moreno Acevedo, S.A. (2023). Automatic information extraction in business document. https://bibliotecadigital.udea.edu.co/bitstream/10495/37581/1/MorenoSantiago_2023_InformationExtractionDeepLearningNaturalLanguageProcessing.pdf

Munikoti, S., Acharya, A., Wagle, S., & Horawalavithana, S. (2023). ATLANTIC: Structure-aware retrieval-augmented language Model for interdisciplinary science. *arXiv Preprint arXiv: 231112289*. https://doi.org/10.48550/arXiv.2311.12289

Nakhod, O. (2023). Using retrieval-augmented generation to elevate low-code developer skills. *Artificial Intelligence*, *28*(3), 126–130. https://doi.org/10.15407/jai2023.03.126

Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, *31*(3), 521–543. https://doi.org/10.1287/mksc.1120.0713

Nguyen, M.T., Le, D.T., Linh, L.T., Hong Son, N., Duong, D.H.T., Cong Minh, B., & Huu Hiep, N. (2020, October). Aurora: An information extraction system of domain-specific business documents with limited data. In CIKM '20: The 29th ACM International Conference on Information and Knowledge Management Virtual Event Ireland October 19 - 23, 2020 (pp. 3437–3440). https://doi.org/10.1145/3340531.3417434

Pan, F., Canim, M., Glass, M., Gliozzo, A., & Hendler, J. (2022). End-to-end table question answering via retrieval-augmented generation. *arXiv Preprint arXiv: 220316714*. https://doi.org/10.48550/arXiv.2203.16714

Piskorski, J., Stefanovitch, N., Jacquet, G., & Podavini, A. (2021, April). Exploring linguistically-lightweight keyword extraction techniques for indexing news articles in a multilingual set-up. In Proceedings of the EACL Hackashop on news media content analysis and automated report generation (pp. 35–44). Association for Computational Linguistics.

Pouplin, T., Sun, H., Holt, S., & Van der Schaar, M. (2024). Retrieval-augmented thought process for private data handling in healthcare. *arXiv Preprint arXiv: 240207812*. https://arxiv.org/abs/2402.07812

Pournemat, M., & Weiss, M. (2021). Identifying business opportunities using topic modeling and chance discovery. In The ISPIM Innovation Conference – Innovating Our Common Future, Berlin, Germany on 20–23 June 2021. (pp. 1–11). The International Society for Professional Innovation Management (ISPIM).

Pugachev, A., Voronov, A., & Makarov, I. (2021). Prediction of news popularity via keywords extraction and trends tracking. In Recent Trends in Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Skolkovo, Moscow, Russia, October 15–16, 2020 Revised Supplementary Proceedings 9, Moscow, Russia (pp. 37–51). Springer International Publishing.

Rackauckas, Z. (2024). RAG-Fusion: A New take on retrieval-augmented generation. *International Journal on Natural Language Computing*, *13*(1), 37–47. https://doi.org/10.5121/ijnlc.2024.13103

Ranade, P., & Joshi, A. (2023, November). Fabula: Intelligence report generation using retrieval-augmented narrative construction. In Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (pp. 603–610). https://doi.org/10.1145/3625007.3627505

Reyes, D.D.L., Trajano, D., Manssour, I.H., Vieira, R., & Bordini, R.H. (2021, November). Entity relation extraction from news articles in portuguese for competitive intelligence based on bert. In Brazilian Conference on Intelligent Systems (pp. 449–464). Springer International Publishing, Cham.

Sage, C., Aussem, A., Eglin, V., Elghazel, H., & Espinas, J. (2020, November). End-to-end extraction of structured information from business documents with pointer-generator networks. In Proceedings of the fourth workshop on structured prediction for NLP (pp. 43–52). https://doi.org/10.18653/v1/2020.spnlp-1.6

Saggion, H., Funk, A., Maynard, D., & Bontcheva, K. (2007). Ontology-based information extraction for business intelligence. In The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings (pp. 843–856). Springer, Berlin Heidelberg.

Sajjadi, M.S., Bachem, O., Lucic, M., Bousquet, O., & Gelly, S. (2018). Assessing generative models via precision and recall. *Advances in Neural Information Processing Systems*, 31. https://doi.org/10.48550/arXiv.1806.00035

Sarawagi, S. (2008). Information extraction. *Foundations and Trends® in Databases*, *1*(3), 261–377. https://doi.org/10.1561/1900000003

Schön, S., Mironova, V., Gabryszak, A., & Hennig, L. (2020). A corpus study and annotation schema for named entity recognition and relation extraction of business products. *arXiv Preprint arXiv: 200403287*. https://doi.org/10.48550/arXiv.2004.03287

Sha, Y., Feng, Y., He, M., Liu, S., & Ji, Y. (2023). Retrieval-augmented knowledge graph reasoning for commonsense question answering. *Mathematics*, *11*(15), 3269. https://doi.org/10.3390/math11153269

Shi, E., Wang, Y., Tao, W., Du, L., Zhang, H., Han, S., & Sun, H. (2022). RACE: Retrieval-augmented commit message generation. *arXiv Preprint arXiv: 220302700*. https://doi.org/10.48550/arXiv.2203.02700

Shi, P., Zhang, R., Bai, H., & Lin, J. (2022). Xricl: Cross-lingual retrieval-augmented in-context learning for cross-lingual text-to-sql semantic parsing. *arXiv Preprint arXiv: 221013693*.

Simmons, L.L., & Conlon, S.J. (2013). Extraction of financial information from online business reports. *ACM SIGMIS Database: The DATABASE for Advances in Information Systems*, *44*(3), 34–48. https://doi.org/10.1145/2516955.2516958

Song, S., Lv, Q., Geng, L., Cao, Z., & Fu, G. (2023, October). Rspell: Retrieval-augmented framework for domain adaptive Chinese spelling check. In CCF International Conference on Natural Language Processing and Chinese Computing (pp. 551–562). Springer Nature Switzerland, Cham.

Sprague, R.H. (2004). Document mining for DSS. *Journal of Decision Systems*, *13*(2), 173–182. https://doi.org/10.3166/jds.13.173-182

Sul, S., & Cho, S.B. (2024). Understanding people's attitudes in IoT systems using wellness probes and TF-IDF data analysis. *Multimedia Tools & Applications*, 1–20. https://doi.org/10.1007/s11042-024-18830-8

Sun, T. (2022). *Relation extraction from financial reports* [Doctoral dissertation]. University of York.

Sung, N.H., & Chang, Y.S. (2004). Business information extraction from semi-structured webpages. *Expert Systems with Applications*, *26*(4), 575–582. https://doi.org/10.1016/j.eswa.2003.12.008

Thompson, W.E., Vidmar, D.M., De Freitas, J.K., Pfeifer, J.M., Fornwalt, B.K., Chen, R., & Miotto, R. (2023). Large language models with retrieval-augmented generation for zero-shot disease phenotyping. *arXiv Preprint arXiv: 231206457*. https://doi.org/10.48550/arXiv.2312.06457

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv Preprint arXiv: 230709288*. https://doi.org/10.48550/arXiv.2307.09288

Wang, H., Huang, W., Deng, Y., Wang, R., Wang, Z., Wang, Y., & Wong, K.F. (2024). UniMS-rag: A unified multi-source retrieval-augmented generation for personalized dialogue systems. *arXiv Preprint arXiv: 240113256*. https://doi.org/10.48550/arXiv.2401.13256

Wang, W., Wang, Y., Joty, S., & Hoi, S.C. (2023, November). Rap-gen: Retrieval-augmented patch generation with codet5 for automatic program repair. In Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (pp. 146–158). https://doi.org/10.1145/3611643.3616256

Wen, Z., Tian, Z., Wu, W., Yang, Y., Shi, Y., Huang, Z., & Li, D. (2023). Grove: A retrieval-augmented complex story generation framework with a forest of evidence. *arXiv Preprint arXiv: 231005388*. https://doi.org/10.48550/arXiv.2310.05388

Xia, M., Zhang, X., Couturier, C., Zheng, G., Rajmohan, S., & Ruhle, V. (2023). Hybrid retrieval-augmented generation for real-time composition assistance. *arXiv Preprint arXiv: 230804215*. https://doi.org/10.48550/arXiv.2308.04215

Xiong, G., Jin, Q., Lu, Z., & Zhang, A. (2024). Benchmarking retrieval-augmented generation for medicine. *arXiv Preprint arXiv: 240213178*. https://doi.org/10.48550/arXiv.2402.13178

Xu, R., Yu, Y., Ho, J., & Yang, C. (2023, July). Weakly-supervised scientific document classification via retrieval-augmented multi-stage training. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, Taipei Taiwan (pp. 2501–2505). https://doi.org/10.1145/3539618.3592085

Yamamoto, A., Miyamura, Y., Nakata, K., & Okamoto, M. (2017, January). Company relation extraction from web news articles for analyzing industry structure. In 2017 IEEE 11th International Conference on Semantic Computing (ICSC), San Diego, CA, USA (pp. 89–92). IEEE.

Yan, C., Fu, X., Wu, W., Lu, S., & Wu, J. (2019, February). Neural network based relation extraction of enterprises in credit risk management. In 2019 IEEE International Conference on Big Data and Smart Computing (BigComp), Kyoto, Japan (pp. 1–6). IEEE.

Yu, H., Guo, P., & Sano, A. (2023, December). Zero-shot ECG diagnosis with large language models and retrieval-augmented generation. In *Machine learning for health (ML4H)* (pp. 650–663). New Orleans, USA: PMLR.

Zakka, C., Shad, R., Chaurasia, A., Dalal, A.R., Kim, J.L., Moor, M., Fong, R., Phillips, C., Alexander, K., Ashley, E., Boyd, J., Boyd, K., Hirsch, K., Langlotz, C., Lee, R., Melia, J., Nelson, J., Sallam, K., Tullis, S., & Cunningham, J.P.. . . Hiesinger, W. (2024). Almanac—retrieval-augmented language models for clinical medicine. *Nejm Ai*, *1*(2), Aloa2300068. https://doi.org/10.1056/Aloa2300068

Zhang, B., Yang, H., Zhou, T., Ali Babar, M., & Liu, X.Y. (2023, November). Enhancing financial sentiment analysis via retrieval augmented large language models. In ICAIF '23: 4th ACM International Conference on AI in Finance Brooklyn NY USA (pp. 349–356). Association for Computing Machinery: New York, NY, USA. https://doi.org/10.1145/3604237.3626866

Zhang, R., Yang, W., Lin, L., Tu, Z., Xie, Y., Fu, Z., & Lin, J. (2020). Rapid adaptation of bert for information extraction on domain-specific business documents. *arXiv Preprint arXiv: 200201861*. https://doi.org/10.48550/arXiv.2002.01861

Zhang, Z., Fang, M., & Chen, L. (2024). RetrievalQA: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. *arXiv Preprint arXiv: 240216457*. https://doi.org/10.48550/arXiv.2402.16457

Zhao, C., Jiang, Y., Qiu, Y., Zhang, H., & Yang, W.Y. (2023, October). Differentiable retrieval augmentation via generative language modeling for E-commerce query intent classification. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, Birmingham, UK (pp. 4445–4449). https://doi.org/10.1145/3583780.3615210