Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 246 (2024) 3781–3790

Procedia
Computer Science

www.elsevier.com/locate/procedia

28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2024)

A Survey on RAG with LLMs

Muhammad Arslan[a]*, Hussam Ghanem[a], Saba Munawar[b] and Christophe Cruz[a]

[a]*Laboratoire Interdisciplinaire Carnot de Bourgogne (ICB), Dijon, France*
[b]*National University of Computer and Emerging Sciences (NUCES), Islamabad, Pakistan*

Abstract

In the fast-paced realm of digital transformation, businesses are increasingly pressured to innovate and boost efficiency to remain competitive and foster growth. Large Language Models (LLMs) have emerged as game-changers across industries, revolutionizing various sectors by harnessing extensive text data to analyze and generate human-like text. Despite their impressive capabilities, LLMs often encounter challenges when dealing with domain-specific queries, potentially leading to inaccuracies in their outputs. In response, Retrieval-Augmented Generation (RAG) has emerged as a viable solution. By seamlessly integrating external data retrieval into text generation processes, RAG aims to enhance the accuracy and relevance of the generated content. However, existing literature reviews tend to focus primarily on the technological advancements of RAG, overlooking a comprehensive exploration of its applications. This paper seeks to address this gap by providing a thorough review of RAG applications, encompassing both task-specific and discipline-specific studies, while also outlining potential avenues for future research. By shedding light on current RAG research and outlining future directions, this review aims to catalyze further exploration and development in this dynamic field, thereby contributing to ongoing digital transformation efforts.

Keywords: Large Language Models (LLMs); Natural Language Processing (NLP); Retrieval-Augmented Generation (RAG); Text generation; Digital transformation.

1. Introduction

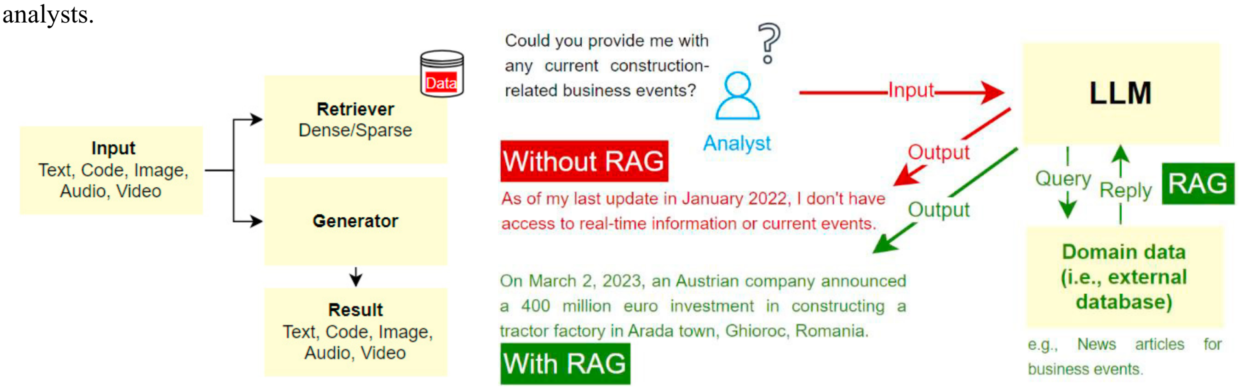
Digital transformation signifies the incorporation of digital technology across different facets of a business, reshaping its operations and value delivery to customers [1]. At the forefront of driving such transformative practices are Large Language Models (LLMs), advanced machine learning models trained extensively on textual data to comprehend and produce human-like text [1]. LLMs, such as the Generative Pre-training Transformer (GPT)

* Corresponding author. Tel.: +33 03 80 39 50 00; fax: +33 03 80 39 50 69.
E-mail address: muhammad.arslan@u-bourgogne.fr

10.1016/j.procs.2024.09.178

series [2, 3] and others, have demonstrated remarkable capabilities in NLP tasks [4]. However, these models face challenges when dealing with domain-specific queries, often generating inaccurate or irrelevant information, commonly referred to as "hallucinations", particularly when data is sparse [5]. This limitation makes deploying LLMs in real-world settings impractical, as the generated output may not be reliable [4].

In the middle of 2020, Lewis et al. [6] introduced RAG, a significant advancement in the field of LLMs for improving generative tasks (see Fig. 1 (a)). RAG incorporates an initial step where LLMs search an external data source to retrieve relevant information before producing text or answering questions. RAG addresses these limitations by integrating external data retrieval into the generative process, thereby enhancing the accuracy and relevance of the generated output. By dynamically retrieving information from knowledge bases during inference, RAG provides a more informed and evidence-based approach to language generation, significantly reducing the risk of hallucinations and improving the overall quality of the generated text [4, 6]. This approach has the potential to make LLMs more practical for real-world applications, as it ensures that the generated output is grounded in retrieved evidence, leading to more reliable and accurate results. Fig. 1 (b) showcases how real-time business systems can leverage the RAG with LLM architecture. As an example, without RAG, the system lacks access to real-time or updated information. However, with RAG integration, leveraging external data sources such as news articles, the system can respond to current business events, presenting opportunities for business intelligence analysts.


Input
Text, Code, Image, Audio, Video
Retriever
Dense/Sparse
Data
Generator
Result
Text, Code, Image, Audio, Video
Analyst
Input
LLM
Output
Output
Without RAG
Query
Reply
RAG
Domain data (i.e., external database)
e.g., News articles for business events.
With RAG

Could you provide me with any current construction-related business events?

As of my last update in January 2022, I don't have access to real-time information or current events.

On March 2, 2023, an Austrian company announced a 400 million euro investment in constructing a tractor factory in Arada town, Ghioroc, Romania.

(a) (b)

Fig. 1. (a) A generic RAG architecture, where users' queries, potentially in different modalities (e.g., text, code, image, etc.), are inputted into both the retriever and the generator. The retriever scans for relevant data sources in storage, while the generator engages with the retrieval outcomes, ultimately generating results across various modalities [6]; Fig. 1. (b) illustrates how RAG integration with the LLM handles queries that fall outside the scope of the LLM's training data.

While the field of RAG has seen substantial growth, several online surveys [4, 7, 8, 9] have explored technological advancements in RAG. Although these surveys provide valuable insights and references, they offer only a limited overview of RAG applications. To address this gap, this paper aims to provide an exhaustive overview of RAG applications, including both task-specific and discipline-specific studies, as well as future directions. By highlighting the current state of RAG research and its potential future directions, this review aims to inspire further investigation and development in this exciting field.

The paper's structure is as follows: Section 2 presents the adopted research methodology for this survey. In Section 3, we provide an overview of RAG applications, followed by a detailed discussion in Section 4. The paper concludes in Section 5, summarizing the key findings and implications of the study.

2. Background

The research method (see Fig. 2) employed in this paper involves a thorough review and analysis of research publications related to RAG. The main objective is to identify and categorize its applications across various NLP tasks and disciplines. The paper begins by collecting research publications specific to RAG, focusing on their applications. Since the RAG with LLM domain is relatively new and emerging, with many studies available as preprints online, limiting the search to platforms such as Scopus or IEEE would greatly reduce the number of studies. Therefore, Google Scholar was utilized to access the studies on RAG. However, in cases where both pre-print and published versions of a study were available, the published version was chosen to cover the maximum number of peer-reviewed studies. Each study underwent manual review to assess its comprehensiveness and depth, excluding short studies. It is important to note that the purpose of the survey is not to cover the most optimal studies, but rather to provide an overview of how this field has attained significant attention in a short period, with researchers exploring diverse application scenarios.

The keywords used to collect research publications included “retrieval augmented generation”, “RAG applications”, “generative models with retrieval”, “external data retrieval in text generation”, “enhancing text generation with retrieval”, “integrating retrieval into generative models”, “external knowledge in text generation”, “retrieval-based text generation”, “information retrieval for text generation”, and “contextualized retrieval in language models”. These publications are then classified into two principal categories: task-based classification and discipline-based classification. Task-based classification focuses on categorizing RAG studies according to their execution of information processing tasks, particularly within NLP. Conversely, discipline-based classification categorizes studies based on their application to specific domains. Under the task-based classification, the publications are further subdivided into categories such as Question Answering (QA), Text Generation and Summarization, Information Retrieval and Extraction, Text Analysis and Processing, Software Development and Maintenance (SDM), Decision Making and Applications, and Other Categories. Similarly, under the discipline-based classification, the publications are further subdivided into categories such as Medical/Biomedical, Financial, Educational, Technology and Software Development, Social and Communication, Literature, and Other Categories. These categories are selected based on an understanding of the context of the studies and the underlying problems they address. Within both classification methods, “software development” stands out as a common category. It involves programming information processing tasks under task-based classification and encompasses systems for developing various applications across different domains under discipline-based classification. Figure 3 illustrates the number of publications related to RAG applications from 2020 to February 2024. Specifically, there was a single publication found in 2020, 6 publications in 2022, 28 publications in 2023, and 16 publications until February 2024, indicating a growing interest and research activity in the field of RAG applications.

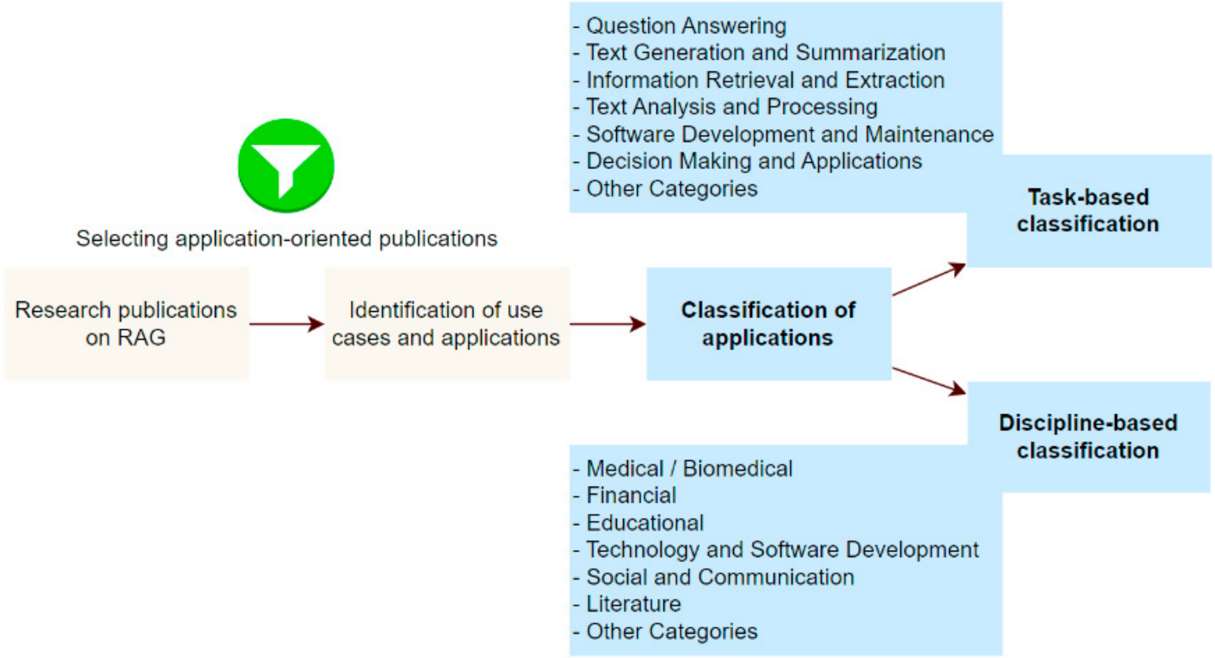

- Question Answering
- Text Generation and Summarization
- Information Retrieval and Extraction
- Text Analysis and Processing
- Software Development and Maintenance
- Decision Making and Applications
- Other Categories
Selecting application-oriented publications
Task-based classification
Research publications on RAG
Identification of use cases and applications
Classification of applications
Discipline-based classification
- Medical / Biomedical
- Financial
- Educational
- Technology and Software Development
- Social and Communication
- Literature
- Other Categories

Fig. 2. Research Method

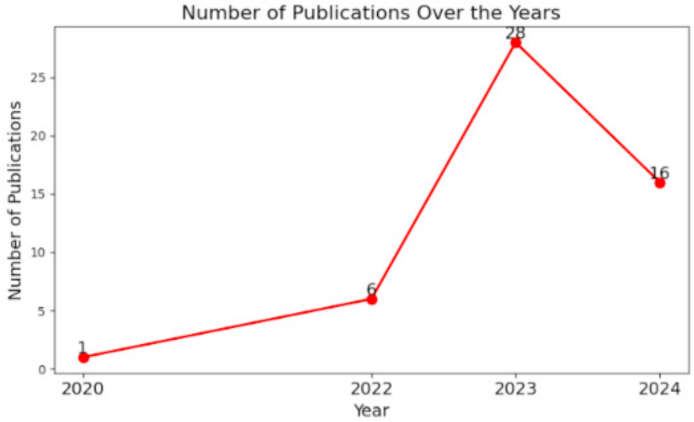


Fig. 3. Evolution of Research Publications on RAG Applications

3. Applications of RAG with LLMs

Upon thorough examination of the selected papers focusing on RAG applications, we uncovered a vast array of diverse applications. These findings are distilled into a comprehensive table format (see Table 1), detailing three crucial aspects: 1) Use case with RAG, 2) Used datasets/benchmarks, and 3) Application area. Noteworthy applications span various domains, including biomedical, financial, and medical inquiries, alongside text summarization and book review generation. RAG's versatility extends to commonsense QA, table-based queries, and clinical decision-making, among others. It further encompasses educational decision making, textbook question answering, and enterprise search functionalities. RAG is instrumental in sentiments classification, health education, and generating biomedical explanations, while also enhancing user writing accuracy and speed. Its utility spans humanitarian assistance, generating informative dialogues, crafting realistic images and intricate plotlines, and much more.

Additionally, RAG aids in natural language QA, disease identification, and information extraction. It handles decision-making tasks, hashtag management, hate speech detection, and scientific document classification. RAG excels in entity description generation, text correction, and SQL translation, while also enhancing open-domain QA and professional knowledge inquiries. Also, it extends the capabilities of machine translation tasks beyond text-to-SQL, such as neural text re-ranking [6]. Moreover, it supports multicultural enterprise queries, e-commerce searches, and personalized dialogue systems. Furthermore, RAG facilitates event argument extraction, intelligence report generation, short-form QA, automated transactions, and private data handling. Lastly, it contributes to science QA, clinical writing, and pharmaceutical regulatory compliance inquiries.

After compiling all the applications of RAG, the subsequent step involves categorizing them based on the specific nature of the NLP tasks they tackle (see Table 2 and Fig. 4). From the compiled publications, it was observed that 20 studies were dedicated to QA, 6 to Text Generation and Summarization, 6 to Information Retrieval and Extraction, 5 to Text Analysis and Processing, 4 to SDM, and 5 to Decision Making and Applications, while the remaining 6 studies were classified under "Other Categories." This classification is significant as it helps in understanding the distribution and focus of RAG applications across different NLP tasks. Additionally, since RAG applications span various disciplines, further classification (see Table 3 and Fig. 5) reveals that 9 publications were related to Medical/Biomedical, 2 to Financial, 2 to Educational, 9 to Technology and Software Development, 7 to Social and Communication, and 3 to Literature, with the remaining falling into "Other Categories".

Table 1. Applications of RAG

No.	Use case with RAG	Used datasets / benchmarks	Application area
1	MIRAGE: Medical information RAG [10]	Medical QA datasets	Biomedical QA
2	RAG for improved context accuracy [11]	Financial reports	Financial QA
3	Retrieval-augmented Electrocardiography (ECG) [12]	Cardiac symptoms and sleep apnea diagnosis	Medical QA
4	Representative Vector Summarization (RVS) [13]	PDFs, text documents, spreadsheets, etc.	Medical text summarization
5	Retrieval-augmented controllable reviews [14]	Amazon book reviews	Book review generation

6	Retrieval-augmented knowledge graph reasoning [15]	Commonsense QA and OpenBookQA.	Commonsense QA
7	Answers from table corpus via RAG [16]	Wikipedia data	Table QA
8	LiVersa: a liver disease specific LLM using RAG [17]	Liver Diseases	Medical QA
9	Almanac: RAG for clinical medicine [18]	Guidelines and treatment recommendations.	Clinical decision-making
10	Assessment of tutoring practices [19]	Dialogue transcripts from a middle-school.	Educational decision making
11	Handling out of domain scenarios [20]	Life science, earth science, etc. lessons.	Textbook QA
12	Automated form filling [21]	Request forms for IT projects	Enterprise search
13	Financial sentiment analysis [22]	Twitter financial news and FiQA datasets	Sentiments classification
14	Frontline health worker capacity building [23]	Pregnancy-related guidelines	Health education QA
15	Self-BioRAG: a framework for biomedical text [24]	Biomedical instruction sets	Biomedical Informatics
16	Hybrid RAG for real-time composition assistance [25]	WikiText-103, Enron Emails, etc.	Writing speed and accuracy
17	RAG-Fusion to obtain product information [26]	Product datasheets	Technical information QA
18	Commit message generation for code intelligence [27]	MCMD dataset	SDM
19	FloodBrain: Flood disaster reporting [28]	ReliefWeb reports	Humanitarian assistance
20	Rich answer encoding [29]	MSMARCO QA and WoW dataset.	Generative QA
21	Text-to-image generator [30]	COCO and WikiImages datasets.	Realistic images generation
22	Code completion framework [31]	CodeXGLUE and CodeNet datasets.	SDM
23	Complex story generation framework [32]	IMDB movie details dataset	Generate stories
24	TRAC: Trustworthy retrieval augmented chatbot [33]	Natural Question dataset	Natural QA
25	Clinfo.ai using scientific literature [34]	PubMed dataset	Medical QA
26	RealGen for controllable traffic scenarios [35]	nuScenes dataset	Critical traffic scenarios
27	Zero-shot disease phenotyping [36]	Clinical notes	Identifying diseases
28	RAP-Gen for automatic program repair [37]	TFix, Defects4J, etc. datasets	SDM
29	Code4UIE : retrieval-augmented code generation [38]	ACE04, ACE05, CoNLL03, etc. datasets	Information extraction
30	RAP: retrieval-augmented planning [39]	ALFWorld, Webshop, etc. datasets	Decision-making
31	RIGHT for mainstream hashtag recommendation [40]	Twitter and Weibo data.	Retrieval-enhanced hashtags
32	RAUCG for counter narrative generation for hate speech [41]	MultitargetCONAN dataset	Combating hate speech
33	Weakly-supervised scientific document classification [42]	AGNews and MeSH datasets.	Scientific documents classification
34	rT5 for Chinese entity description generation [43]	XunZi and MengZi datasets.	Entity description generation
35	RSpell: domain adaptive Chinese spelling check [44]	CSC dataset	Text error correction
36	XRICL: cross-lingual retrieval-augmented in-context learning for cross-lingual text-to-SQL semantic parsing [45]	XSPIDER and XKAGGLE-DBQA datasets.	Text-to-SQL translation
37	SELF-RAG: learning to retrieve, generate, and critique through self-reflection [46]	Open-Instruct processed data.	Open-domain QA and fact verification
38	ChatDOC with enhanced PDF structure recognition [47]	Academic papers, financial reports, textbooks, and legislative materials	Professional knowledge QA
39	G-Retriever for textual graph understanding [48]	GraphQA (ExplaGraphs, SceneGraphs and WebQSP)	Chat with graphs
40	Enhancing multilingual information retrieval in mixed Human Resources (HR) environments [49]	HR standard operating procedures and Quality Assurance (QA) documents	Multicultural enterprise QA
41	Differentiable RAG [50]	User-clicked logs	E-commerce search (query intent classification)
42	RAG to elevate low-code developer skills [51]	Caspio and Power automate data	SDM
43	UniMS-RAG: a unified multi-source RAG [52]	DuLeMon and KBP datasets	Personalized dialogue systems
44	RAG QA for event argument extraction [53]	ACE 2005 and WikiEvent datasets	Event argument (answer) extraction
45	FABULA: retrieval-augmented narrative construction [54]	OntoNotes and Pile datasets	Intelligence report generation
46	Time-Aware Adaptive Retrieval (TA-ARE) [55]	RetrievalQA dataset	Short-form open-domain QA
47	Cash transaction booking via RAG [56]	Cash Management Software (CMS) transactions.	Automated cash transaction booking
48	Retrieval-Augmented Thought Process (RATP) [57]	Boolq and emrQA datasets.	QA with private data
49	ATLANTIC for interdisciplinary science [58]	S2ORC dataset	Science QA and scientific document classification
50	Writing documents for clinical trials [59]	FDA guidance database, ClinicalTrials.gov, and AACT database.	Clinical-related writing
51	QA RAG model [60]	FDA Q&A datasets	Pharma industry regulatory compliance QA

4. Discussion

The classification of RAG applications according to the specific NLP tasks they target holds significant importance for several reasons. Firstly, it offers valuable insights into the distribution and focus of RAG applications across various tasks within the field of NLP. By quantifying the number of studies dedicated to each task, researchers gain a deeper understanding of where efforts and resources are predominantly concentrated within the RAG domain. By analyzing the distribution of RAG applications, researchers can discern prevailing trends in research interest and identify emerging areas of importance. The classification of RAG applications based on discipline offers valuable insights into its widespread adoption across various domains. This classification not only provides a comprehensive understanding of RAG's applicability but also underscores its potential to revolutionize various domains, thereby contributing significantly to the advancement of NLP technologies.

While this survey offers a comprehensive overview of RAG applications across various NLP tasks and disciplines, it also has its limitations. 1) Given that RAG technology is still emerging, the majority of RAG-based studies are available in pre-print formats on platforms like arXiv, lacking peer review. This raises questions about their authenticity. 2) Additionally, the survey overlooks the technical implementation details and challenges associated with using RAG technology alongside open-source LLMs. Organizations may find RAG implementation costly if they do not opt for open-source LLM architectures, especially considering the expense of querying the LLM via Application Programming Interface (API). 3) Furthermore, the performance of RAG concerning the volume and variety of datasets has not been discussed. Deploying RAG with large datasets of varying structures (e.g., structured, semi-structured, or non-structured) may lead to processing delays, warranting further exploration before selecting a RAG with LLM integrated solution for organizational deployment.

4) Additionally, this survey did not cover the diverse range of RAG architectures and technologies available for integration with different LLMs. Future work should delve into these options to discuss how various RAG solutions can be adapted with LLMs for different NLP tasks and applications. 5) Furthermore, the survey did not address the accuracy of information obtained from RAG with LLM solutions. It is essential to explore the reliability of these systems and assess the organizations' dependency on their generated responses. LLMs often generate responses with high confidence, making it challenging to evaluate the accuracy of the information provided. 6) While the survey primarily focuses on task-based and discipline-based applications of RAG, there is a need for further research to explore ethical considerations associated with its usage, especially when dealing with sensitive datasets. For example, in the biomedical domain, RAG has the potential to accidentally expose private information to analysts, raising concerns about data privacy and security. Additionally, in the legal domain, RAG may mistakeably reveal privileged information during document analysis, potentially violating client confidentiality and attorney-client privilege. Therefore, future studies should delve deeper into these ethical implications to ensure responsible and ethical use of RAG technology across various domains.

5. Conclusion

This article offers a thorough examination of the applications of RAG with LLMs, showcasing their potential to drive digital transformation across diverse industries. Initially, it gathers the latest publications on RAG from online repositories. These publications are then classified based on task-oriented and discipline-oriented criteria. A notable trend observed is the increasing number of research papers on RAG deposited in open-access sources, particularly since 2023. However, many works remain unpublished or are in the preprint stage, awaiting review by various journals. A significant portion of these studies primarily focus on the task of QA in NLP. Conversely, there is a noticeable gap in research exploring Entity Linking, an essential NLP task that contributes to knowledge graph development. Addressing this gap could unlock numerous applications in the realm of linked data. Regarding disciplines, the majority of research applications are concentrated in the fields of Medical/Biomedical and Technology and Software Development. In contrast, disciplines such as Business and Agriculture receive comparatively less attention. Future research endeavors should aim to bridge this gap by addressing the specific needs of these underrepresented disciplines.

Table 2. Task-based classification of RAG applications. The detailed categories are derived from the "Application area" column of Table 1. These categories are assigned based on a thorough comprehension of the study's context.

1) Question Answering (QA)

- Biomedical QA [1]
- Financial QA [2]
- Medical QA [3]
- Commonsense QA [6]
- Textbook QA [11]
- Health education QA [14]
- Technical product information QA [17]
- Natural QA [24]
- Professional knowledge QA [38]
- Multicultural enterprise QA [40]
- Open-domain QA and fact verification [46]
- Short-form open-domain QA [46]
- Generative QA and informative conversations [29]
- Pharma industry regulatory compliance QA [51]
- Science QA and document classification [49]
- Clinical-related writing [50]
- Personalized dialogue systems [43]

2) Text Generation and Summarization

- Medical text summarization [4]
- Book review generation [5]
- Biomedical Informatics [15]
- Generate stories with complex plots [23]
- Generate realistic and faithful images [21]
- Entity description generation [34]

3) Information Retrieval and Extraction

- Table QA [7]
- Enterprise search [12]
- Retrieval-enhanced hashtags [31]
- Information extraction [29]
- Event argument (answer) extraction [44]
- E-commerce search (query intent classification) [41]

4) Text Analysis and Processing

- Sentiments classification [13]
- Text error correction [35]
- Text-to-SQL translation [36]
- Scientific documents classification [33]
- Combating online hate speech [32]

5) Software Development and Maintenance

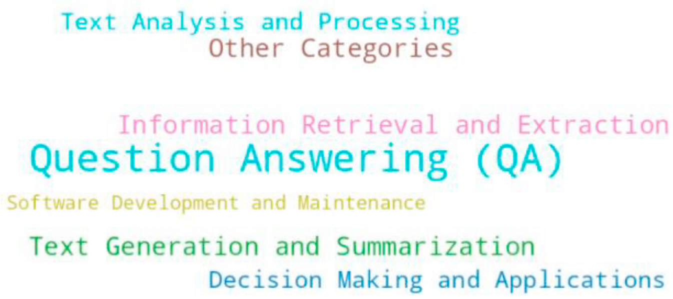
- Code intelligence [18]
- Code completion [22]
- Automatic program repair [28]
- Elevate low-code developer skills [42]

6) Decision Making and Applications

- Clinical decision-making [9]
- Educational decision making [10]
- Decision-making applications [30]
- Automated cash transaction booking [47]
- Intelligence report generation [45]

7) Other Categories:

- Editing and crafting diverse behaviors, including critical traffic scenarios [26]
- Identifying diseases [27]
- Chat with graphs [39]



Task: (Count of Publications)

Question Answering (QA): (20)

Text Generation and Summarization: (6)

Information Retrieval and Extraction: (6)

Text Analysis and Processing: (5)

Software Development and Maintenance: (4)

Decision Making and Applications: (5)

Other Categories: (6)

Fig. 4. Task-based classification of RAG applications with count of publications. The word cloud is generated based on the publication counts listed under various headings in Table 2.

Table 3. Discipline-based classification of RAG applications. The detailed categories are derived from the "Application area" column of Table 1. These categories are assigned based on a thorough comprehension of the study's context.

1) Medical / Biomedical

- Biomedical QA [1]
- Medical QA [3]
- Medical text summarization [4]
- Health education QA [14]
- Identifying diseases [27]
- Clinical decision-making [9]
- Clinical-related writing [50]
- Science QA and scientific document classification [49]
- Pharma industry regulatory compliance QA [51]

2) Financial

- Financial QA [2]
- Automated cash transaction booking [47]

3) Educational

- Educational decision making [10]
- Textbook QA [11]

4) Technology and Software Development

- Table QA [7]
- Technical product information QA [17]
- Software development and maintenance [18, 22, 28, 42]
- Generative QA and informative conversations [20]
- Information extraction [29]
- Text error correction [35]
- Text-to-SQL translation [36]
- Personalized dialogue systems [43]
- Event argument (answer) extraction [44]

5) Social and Communication

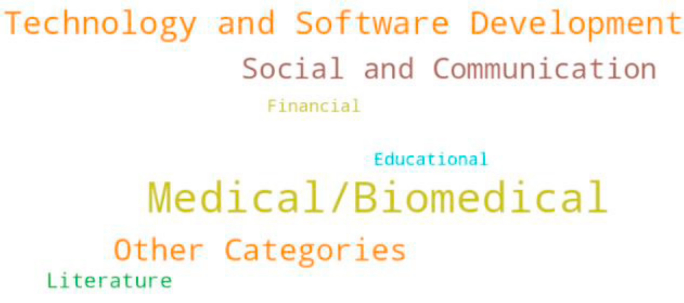
- Commonsense QA [6]
- Sentiments classification [13]
- Combating online hate speech [32]
- Retrieval-enhanced hashtags [31]
- Humanitarian assistance [19]
- Chat with graphs [39]
- Multicultural enterprise QA [40]

6) Literature

- Book review generation guided by reference documents [5]
- Enhance user writing speed and accuracy [16]
- Generate stories with complex plots [23]

7) Other Categories

- Enterprise search [12]
- Generate realistic and faithful images [21]
- Decision-making applications [30]
- Open-domain question answering and fact verification [37]
- Professional knowledge QA [38]
- Intelligence report generation [45]
- Short-form open-domain QA [46]
- Question answering with private data [48]



Discipline:	**(Count of Publications)**
Medical / Biomedical:	(9)
Financial:	(2)
Educational:	(2)
Technology and Software Development:	(9)
Social and Communication:	(7)
Literature	(3)
Other Categories:	(8)

Fig. 5. Discipline-based classification of RAG applications with count of publications. The word cloud is generated based on the publication counts listed under various headings in Table 3.

Acknowledgements

The authors thank the French Government and the National Research Agency (ANR) for their funding.

References

[1] Roumeliotis KI, Tselikas ND, & Nasiopoulos DK. (2024). "LLMs in e-commerce: a comparative analysis of GPT and LLaMA models in product review evaluation," *Natural Language Processing Journal*:**1-6**:100056.

[2] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems* **33** (NeurIPS 2020).

[3] OpenAI, R. (2023). "Gpt-4 technical report," arxiv 2303.08774. View in Article: **2**(**5**).

[4] 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:2312.10997.

[5] Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023). “Large language models struggle to learn long-tail knowledge,” In International Conference on Machine Learning. PMLR: 5696-15707.

[6] 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.

[7] Li, H., Su, Y., Cai, D., Wang, Y., & Liu, L. (2022). “A survey on retrieval-augmented text generation,” arXiv preprint arXiv:2202.01110.

[8] Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., ... & Scialom, T. (2023). “Augmented language models: a survey,” arXiv preprint arXiv:2302.07842.

[9] Zhao, R., Chen, H., Wang, W., Jiao, F., Do, X. L., Qin, C., ... & Joty, S. (2023). “Retrieving multimodal information for augmented generation: A survey,” arXiv preprint arXiv:2303.10868.

[10] Xiong, G., Jin, Q., Lu, Z., & Zhang, A. (2024). “Benchmarking retrieval-augmented generation for medicine,” arXiv preprint arXiv:2402.13178.

[11] 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.

[12] Yu, H., Guo, P., & Sano, A. (2023). “Zero-Shot ECG Diagnosis with Large Language Models and Retrieval-Augmented Generation,” In *Machine Learning for Health (ML4H)* PMLR: 650-663.

[13] 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:2308.00479.

[14] Kim, J., Choi, S., Amplayo, R. K., & Hwang, S. W. (2020). “Retrieval-augmented controllable review generation,” In Proceedings of the 28th International Conference on Computational Linguistics: 2284-2295.

[15] 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.

[16] Pan, F., Canim, M., Glass, M., Gliozzo, A., & Hendler, J. (2022). “End-to-End Table Question Answering via Retrieval-Augmented Generation,” arXiv preprint arXiv:2203.16714.

[17] 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.

[18] Zakka, C., Shad, R., Chaurasia, A., Dalal, A. R., Kim, J. L., Moor, M., ... & Hiesinger, W. (2024). “Almanac—retrieval-augmented language models for clinical medicine,” *NEJM AI* **1**(**2**), AIoa2300068.

[19] Han, Z. FeiFei, 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:2402.14594.

[20] 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:2402.05128.

[21] Bucur, M. (2023). “Exploring Large Language Models and Retrieval Augmented Generation for Automated Form Filling,” (Bachelor's thesis, University of Twente).

[22] Zhang, B., Yang, H., Zhou, T., Ali Babar, M., & Liu, X. Y. (2023). “Enhancing financial sentiment analysis via retrieval augmented large language models,” In Proceedings of the Fourth ACM International Conference on AI in Finance: 349-356.

[23] 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.

[24] Jeong, M., Sohn, J., Sung, M., & Kang, J. (2024). “Improving Medical Reasoning through Retrieval and Self-Reflection with Retrieval-Augmented Large Language Models,” arXiv preprint arXiv:2401.15269.

[25] 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:2308.04215.

[26] Rackauckas, Z. (2024). “RAG-Fusion: A New Take on Retrieval-Augmented Generation,” arXiv preprint arXiv:2402.03367.

[27] Shi, E., Wang, Y., Tao, W., Du, L., Zhang, H., Han, S., ... & Sun, H. (2022). “RACE: Retrieval-Augmented Commit Message Generation,” arXiv preprint arXiv:2203.02700.

[28] 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:2311.02597.

[29] Huang, W., Lapata, M., Vougiouklis, P., Papasarantopoulos, N., & Pan, J. (2023). “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): 1012-1025.

[30] Chen, W., Hu, H., Saharia, C., & Cohen, W. W. (2022). “Re-imagen: Retrieval-augmented text-to-image generator,” arXiv preprint arXiv:2209.14491.

[31] Lu, S., Duan, N., Han, H., Guo, D., Hwang, S. W., & Svyatkovskiy, A. (2022). “Reacc: A retrieval-augmented code completion framework,” arXiv preprint arXiv:2203.07722.

[32] 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:2310.05388.

[33] Li, S., Park, S., Lee, I., & Bastani, O. (2023). "TRAC: Trustworthy Retrieval Augmented Chatbot," arXiv preprint arXiv:2307.04642.

[34] 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: 8-23.

[35] Ding, W., Cao, Y., Zhao, D., Xiao, C., & Pavone, M. (2023). "RealGen: Retrieval Augmented Generation for Controllable Traffic Scenarios," arXiv preprint arXiv:2312.13303.

[36] 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:2312.06457.

[37] Wang, W., Wang, Y., Joty, S., & Hoi, S. C. (2023). "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: 146-158.

[38] 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:2311.02962.

[39] 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:2402.03610.

[40] Fan, R. Z., Fan, Y., Chen, J., Guo, J., Zhang, R., & Cheng, X. (2023). "RIGHT: Retrieval-augmented Generation for Mainstream Hashtag Recommendation," arXiv preprint arXiv:2312.10466.

[41] 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:2310.05650.

[42] Xu, R., Yu, Y., Ho, J., & Yang, C. (2023). "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: 2501-2505.

[43] Hu, M., Zhao, X., Wei, J., Wu, J., Sun, X., Li, Z., ... & Zhang, Y. (2023). "rT5: A Retrieval-Augmented Pre-trained Model for Ancient Chinese Entity Description Generation," In International Conference on NLP and Chinese Computing. Cham: Springer: 736-748.

[44] Song, S., Lv, Q., Geng, L., Cao, Z., & Fu, G. (2023). "RSpell: Retrieval-augmented Framework for Domain Adaptive Chinese Spelling Check," In CCF International Conference on Natural Language Processing and Chinese Computing. Cham: Springer: 551-562.

[45] 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:2210.13693.

[46] 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:2310.11511.

[47] Lin, D. (2024). "Revolutionizing Retrieval-Augmented Generation with Enhanced PDF Structure Recognition," arXiv preprint arXiv:2401.12599.

[48] 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:2402.07630.

[49] Ahmad, S. R. (2024). "Enhancing Multilingual Information Retrieval in Mixed Human Resources Environments: A RAG Model Implementation for Multicultural Enterprise," arXiv preprint arXiv:2401.01511.

[50] Zhao, C., Jiang, Y., Qiu, Y., Zhang, H., & Yang, W. Y. (2023). "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: 4445-4449.

[51] Nakhod, o. Using retrieval-augmented generation to elevate low-code developer skills. https://doi.org/10.15407/jai2023.03.126

[52] 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:2401.13256.

[53] Du, X., & Ji, H. (2022). "Retrieval-augmented generative question answering for event argument extraction," arXiv preprint arXiv:2211.07067.

[54] Ranade, P., & Joshi, A. (2023). "FABULA: Intelligence Report Generation Using Retrieval-Augmented Narrative Construction," arXiv preprint arXiv:2310.13848.

[55] Zhang, Z., Fang, M., & Chen, L. (2024). "RetrievalQA: Assessing Adaptive Retrieval-Augmented Generation for Short-form Open-Domain Question Answering," arXiv preprint arXiv:2402.16457.

[56] Zhang, S., Yadav, D., & Jin, T. (2023). "Cash transaction booking via retrieval augmented LLM. KDD 2023 Workshop on Robust NLP for Finance (RobustFin)," https://www.amazon.science/publications/cash-transaction-booking-via-retrieval-augmented-llm

[57] Pouplin, T., Sun, H., Holt, S., & Van der Schaar, M. (2024). "Retrieval-Augmented Thought Process as Sequential Decision Making," arXiv preprint arXiv:2402.07812.

[58] Munikoti, S., Acharya, A., Wagle, S., & Horawalavithana, S. (2023). "ATLANTIC: Structure-Aware Retrieval-Augmented Language Model for Interdisciplinary Science," arXiv preprint arXiv:2311.12289.

[59] 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:2402.16406.

[60] Kim, J., & Min, M. (2024). "From RAG to QA-RAG: Integrating Generative AI for Pharmaceutical Regulatory Compliance Process," arXiv preprint arXiv:2402.01717.