Please quote as: Li, M. M., Nikishina, I., Sevgili, Ö. & Semmann, M. (2024). Wiping out the limitations of Large Language Models - A Taxonomy for Retrieval Augmented Generation. , . doi: 10.48550/arXiv.2408.02854

Wiping out the limitations of Large Language Models - A Taxonomy for Retrieval Augmented Generation

Mahei Manhai Li[1], Irina Nikishina[2], Özge Sevgili[2,3], Martin Semmann[3]

[1]*Information Systems, University of Kassel, Kassel, Germany*

[2]*Language Technology Group, University of Hamburg, Hamburg, Germany*

[3]*HCDS Group, University of Hamburg, Hamburg, Germany*

Introduction

The rise of new modes of interaction with artificial intelligence (AI) has significantly increased its popularity and broadened its applicability. Despite these advancements, the conceptual integration of AI in organizational settings remains limited, with a notable lack of systematic application (Uba et al. 2023). Among various AI applications, Retrieval-Augmented Generation (RAG) stands out due to its potential to transform information retrieval and content generation (Shuster et al. 2021). This was particularly evident following the public unveiling of OpenAI's models like ChatGPT in November 2022, where generative AI (genAI) has garnered much attention in both academic (Böhmann et al. 2023; Wessel et al. 2023) and industry sectors (McGrath 2024; McKinsey 2023). Recent studies on the potential of genAI, which largely rely on large language model (LLM) systems, ranging from automatization (Engel et al. 2023) to improving knowledge work (Anthony et al., 2023; Dell'Acqua et al., 2023) to creating novel business models (Kanbach et al. 2023).

However, LLMs are not without flaws. In recent studies, Large Language Models (LLMs) have been identified to have several core limitations. These include a tendency to generate incorrect or misleading information (hallucinations) (Blom 2010), poor arithmetic capabilities, a lack of interpretative power, the high costs associated with model revisions, limitations in handling less popular or low-resource concepts and entities, and an inability to reference sources accurately (Barnett et al. 2024; Soudani et al. 2024; Zhao et al. 2024). Several approaches have been developed to mitigate the limitations of, while retrieval augmented generation (RAG) is as of now deemed as most promising (Gao et al. 2024). RAG primarily enhances LLMs by incorporating contextual information during the retrieval process, significantly improving the generated content's accuracy and consistency. Consequently, RAG improves LLM tasks and applications in various ways, as evidenced by recent studies (Asai et al. 2023; Jiang et al. 2023; Martino et al. 2023). Given its potential, this paper aims to develop a conceptualization for RAG applications, illustrating how RAG can be systematically implemented to improve LLM tasks and applications across various domains. Recent studies, such as those by Asai et al. (2023), Jiang et al. (2023), and Martino et al. (2023), have shown various ways in which RAG can enhance LLMs, underscoring its attributes as 'explainable, scalable, and adaptable in nature (Siriwardhana et al. 2023, P. 1).

While RAG has rapidly emerged as a focal point of research among computational linguists (e.g. (Shuster et al. 2021)) and computer scientists (e.g. (Chen et al. 2024)), the field of information systems has not yet embraced this stream of research. A search in the AIS Electronic Library (AISeL) using the term 'retrieval augmented generation' yields no results, highlighting this gap.

Current research on RAGs is distributed across various disciplines, and since the technology is evolving very quickly, its unit of analysis is mostly on technological innovations , rather than applications in business contexts. For example, while computer linguists study RAGs to improve tasks, such as text summarization (Zhao et al. 2024), studying the practical deployment and impact of these functionalities often resides within the realm of information systems research (Ågerfalk 2020). As such, the current state of RAG research's accessibility is limited by specialized terminology, study scope, and their level aggregation. This is particularly evident in technical reviews known as surveys within these disciplines, which tend to

emphasize technological aspects over practical applications (Zhao et al. 2024). Thus, we aim to bridge this divide by providing IS researchers with an initial taxonomy for RAG applications. This taxonomy aims to conceptualize a comprehensive overview of the constituting characteristics that define RAG applications, facilitating the adoption of this technology in the IS community. To the best of our knowledge, there have not been any RAG application taxonomies. Thus, our research question is as follows:

How can RAG applications be conceptualized in a taxonomy?

The following chapters begin with a brief introduction to the theoretical foundations, aimed at establishing a common understanding of the essential concepts and technologies. We then describe our methodology for developing the taxonomy, which includes the criteria for selecting papers, an explanation of our rationale for employing a Large Language Model (LLM)-supported approach to extract and identify initial characteristics, and a concise overview of our systematic process for conceptualizing the taxonomy. Our systematic taxonomy development process includes four iterative phases, each designed to refine and enhance our understanding and presentation of RAG's core dimensions. Next, we outline the dimensions that form the core of our study. We have developed a total of five meta-dimensions and sixteen dimensions to comprehensively capture the concept of Retrieval-Augmented Generation (RAG) applications. In the concluding section, we discuss our findings, detailing specific research areas and posing key research questions. This discussion is designed to guide future IS researchers as they explore the emerging topics of RAG systems.

Terminology and Definitions

This section comprises relevant definitions and concepts further used in the paper. As the RAG concept emerges mainly from the body of knowledge in computational linguistics and related fields, we incorporate those research streams throughout the remainder of the paper.

Information Retrieval: Information Retrieval (IR) is to find required information (in an unstructured nature) from large collections, as defined by (Manning et al. 2008) (Manning et al., 2008). Thus, the goal is to retrieve relevant information to user queries in textual form (Zhu et al. 2024). A basic form of IR as utilized in information systems can be seen in traditional storage of data in databases accessible via specific languages, i.e., SQL, to enable a systematic approach to retrieve data and allow basic operations based on the dataset.

Large Language Model: Language modelling is a task for language understanding and generation, with the goal to predict future (or missing) words/tokens (Zhao et al. 2024) . There have been many techniques developed, and recently Large Language Models (LLMs) are quite prominent due to their abilities in solving complex tasks with their large-sized models (Zhao et al., 2024). In contrast to traditional extractive tasks, LLMs are often employed for generative tasks. Technically, the models are mostly based on a Transformer architecture (Vaswani et al. 2017). These LLMs are pre-trained systems, often on large datasets, and fine-tuned via reinforcement learning and human-in-the-loop to provide a human-like conversational interface (e.g. (Ouyang et al. 2022). Fine-tuning a language model to the intended purpose by providing the necessary contextual information is often done via supervised instructions and chats to leverage high-quality, low-quantity data to improve the model. The contextual information leads to improved models and more useful system output.

Retrieval Augmented Generation: Although LLMs show satisfactory performance on complex tasks, they have some limitations, e.g., hallucination. RAG proposes a solution for such challenges by integrating external knowledge retrieved through semantic similarity of document chunks (Fatemi et al. 2023; Gao et al. 2024). Thus, chains of recursive prompts interacting between both systems can be achieved to reduce hallucinations while acting upon contextual data. They typically work following a retrieve-then-work paradigm, where relevant contextual information is found and retrieved from external sources and is followed by another generation system, which is conditioned on both the retrieved contextual information and the user input to provide the augmented information to the end-user (Karpukhin et al. 2020).

Taxonomy Development

In this section, we describe the methodology used for developing the taxonomy of RAG. We also describe the process of paper selection as well as the iterations of the methodology applied. Moreover, we also

describe the additional approach for identifying the domains where RAG was applied – analysis of the top-50 papers found in Google Scholar using ChatGPT.

Methodology

Classification research and typologies are used for the scientific pursuit of differences theorize about commonalities (Beaulieu et al. 2015). Classification schemes and theories of typologies originate in biology to study and classify species but have since gained widespread adoption in the IS community (Nickerson et al. 2013). While sometimes typologies are synonymous with the term framework, in this paper we will use the term taxonomy to structure the novel technological artifact known as RAGs. Thus, in this section, we describe our methodology to develop an RAG application taxonomy. We follow a systematic approach based on Nickerson et al., (2013).

Following this approach, we defined our meta-characteristic as "*structure and applications of retrieval augmented generation*". In doing so, we included conceptual work and case studies that aimed for specific application domains or use cases. Due to the high dynamic in the research area, we also included pre-print articles. To ensure minimal quality standards, we reviewed each article by two independent GAN researchers. To define a level of saturation, we used the objective and subjective ending conditions as proposed by Nickerson et al. (2013). So, we examined a representative sample of the most recent literature, and those dimensions were stable for an iteration. Thus, no extensions, merges, or splits of characteristics were performed. We also ensured that dimensions have at least one characteristic, and those are directly derived from papers while being unique. Subjective ending conditions were considered as proposed. We aimed for comprehensiveness, robustness, conciseness, extensibility, and explainability. This is reflected in the adaptation of dimensions throughout the iterations as we joined, reorganized, and split categories. As we tackle a recent and ever-changing phenomenon, we conclude that an expanded set of dimensions, namely 16, is still useful for research and practice in the current state. Regarding robustness, we checked for strict separation between dimensions as well as characteristics. Comprehensibility is ensured by our extensive approach. Extensibility and explainability were tackled by repeatedly applying examples to the taxonomy.

To start the development process, we choose a twofold approach. First, we used aspects of conceptual-to-empirical to catalyse the initial iteration (Nickerson et al. 2013). This was done by identifying relevant domains with the help of ChatGPT. Second, we also incorporated an empirical-to-conceptual approach that specifically used surveys of RAG, as those already provide some cumulative knowledge of the field. Still, most are rather new and have not gone through a proper double-blind review process. Thus, in combining both approaches, we propose a new angle to deal with emerging topics. Afterwards, we strictly follow the empirical-to-conceptual regime.

Paper selection

As the object of interest is relatively novel, we employed a naïve approach for identifying and selecting papers. As a first step, we started with the search string "RAG & application" including IS journals and affiliated AIS conferences on AIS eLibrary[1]. This lead to no results at all. Thus, we broadened our search pattern and searched with the given search string at Google Scholar[2]. The results encompassed several articles of the Association of Computational Linguistics[3] that is a driver of the general development of LLMs as well as RAG. Thus, we deemed the general search strategy as useful. Despite high quality results, there are also most papers that are in pre-print status and have not been part of a thorough peer review. Still, we includes those, after quality checks by the author team. If articles are perceived as questionable by one of the researchers, we employed a cross-check by at least another researcher of the team. Also, we excluded published work with publishers that do not comply with minimal standards of our research community.

[1] https://aisel.aisnet.org

[2] https://www.scholar.google.com

[3] https://www.aclweb.org

Due to the iterative approach taken, we reviewed twenty-eight papers in the taxonomy development process. While eight papers have been reviewed and published in several venues, twenty paper are still in pre-print phase.

ChatGPT Domain Identification

ChatGPT (OpenAI 2023) has become a helpful assistant not only for the broader community (McGeorge 2023), but also for the scientific research, especially, in Natural Language Processing (Gilardi et al. 2023). To facilitate the time-consuming process of paper analysis, we decided to apply it for identifying the application domains and further compare the outcome to see, whether such automated technique is applicable for the Taxonomy Construction for Information Systems.

First, we created two queries for Google Scholar – a large search system for academic papers. The first query "rag system" for the papers dating from 2023 returned ninety-seven results. The "application of rag" query from 2023 and later returned twenty-two papers.

Then, inspired by (Rafailov et al. 2023), we created a prompt to ChatGPT to cluster the extracted papers into domains. We formulated the prompt as follows:

> *You are a scientific assistant writing a survey. Here below is a list of paper names. You task is to cluster those papers into domains. Name those domains (it might be something like NLP, medicine).*

After the prompt we pasted ninety-seven papers names from the first query separated with the line separator. Based on the titles provided, the ChatGPT model identified the following 8 classes: Artificial Intelligence and (1) Natural Language Processing (AI/NLP); (2) Legal and Judicial Applications; (3) Healthcare and Medical Science; (4) Education and Pedagogy; (5) Financial Analysis and Stock Prediction; (6) Technology and Computing; (7) Environmental Science and Sustainability; (8) Criminal Investigation and Forensics.

When providing ChatGPT 22 papers from the "Application of RAG", the output was as follows: (1) Natural Language Processing (NLP) and AI; (2) Technology and AI Applications; (3) Healthcare and Medicine; (4) Engineering and Construction; (5) Research Methodologies and Surveys;

It is also important to emphasize that in addition to the class names, the model returned papers examples for each class, therefore, we were able to primarily check the correctness of the identified classes. Furthermore, during iterations, we expected to use manual paper check to prove the ChatGPT clustering efficiency. We will compare the obtained results in the Discussion section.

Iterations

We performed 4 Iterations to build our initial RAG application taxonomy. Overall, we analyzed 28 papers, including 5 surveys on RAG, that already performed the extensive analysis and generalization of previous works (Gao et al. 2024; Lewis et al. 2021; Li et al. 2022; Zhao et al. 2024). Those papers already comprise 2188 citations, resulting in more than twenty-eight papers involved to our study in total. Moreover, a major part of the dimensions was added during the first iteration where three survey papers were analyzed. Therefore, we consider this number of iterations reasonable, which was also confirmed by meeting the following objective condition – no new dimensions or characteristics were added, merged, or split in the iteration.

Our iterative development process for the RAG application taxonomy, as illustrated by Fig. 1 started with an initial set of eleven dimensions, mapped across five meta-dimensions. In this *first iteration*, the key dimensions of RAG Phase, Application Domain, Application Task, Rag Process, Paradigm, Retrieval type, Retrieval Process, RAG Role, Modality, Evaluation Metrics and Failures of RAG were established , each annotated with a specific number of characteristics indicated by the numbers in parentheses. In the *second iteration*, we enriched the taxonomy by analyzing seven additional papers, leading to the introduction of four new dimensions: LLM Status, Granularity, Dataset, and Limitations. Concurrently, we refined several existing dimensions by either merging or adding new characteristics, reflecting deeper insights and broader coverage. By the *third iteration*, only two new dimensions were necessary: Application Stack and Future Directions, indicating a nearing saturation in the scope of the taxonomy. Adjustments were made to five

dimensions during this phase, demonstrating a trend toward the stabilization of the taxonomy's structure. The *fourth iteration* confirmed the saturation, as no new changes were made to the taxonomy, suggesting that the existing structure sufficiently captured the relevant aspects of RAG applications as evidenced by the literature." Across all iterations, the taxonomy has evolved to accommodate and anticipate the dynamic nature of RAG applications, ensuring its relevance and utility in future research endeavors.

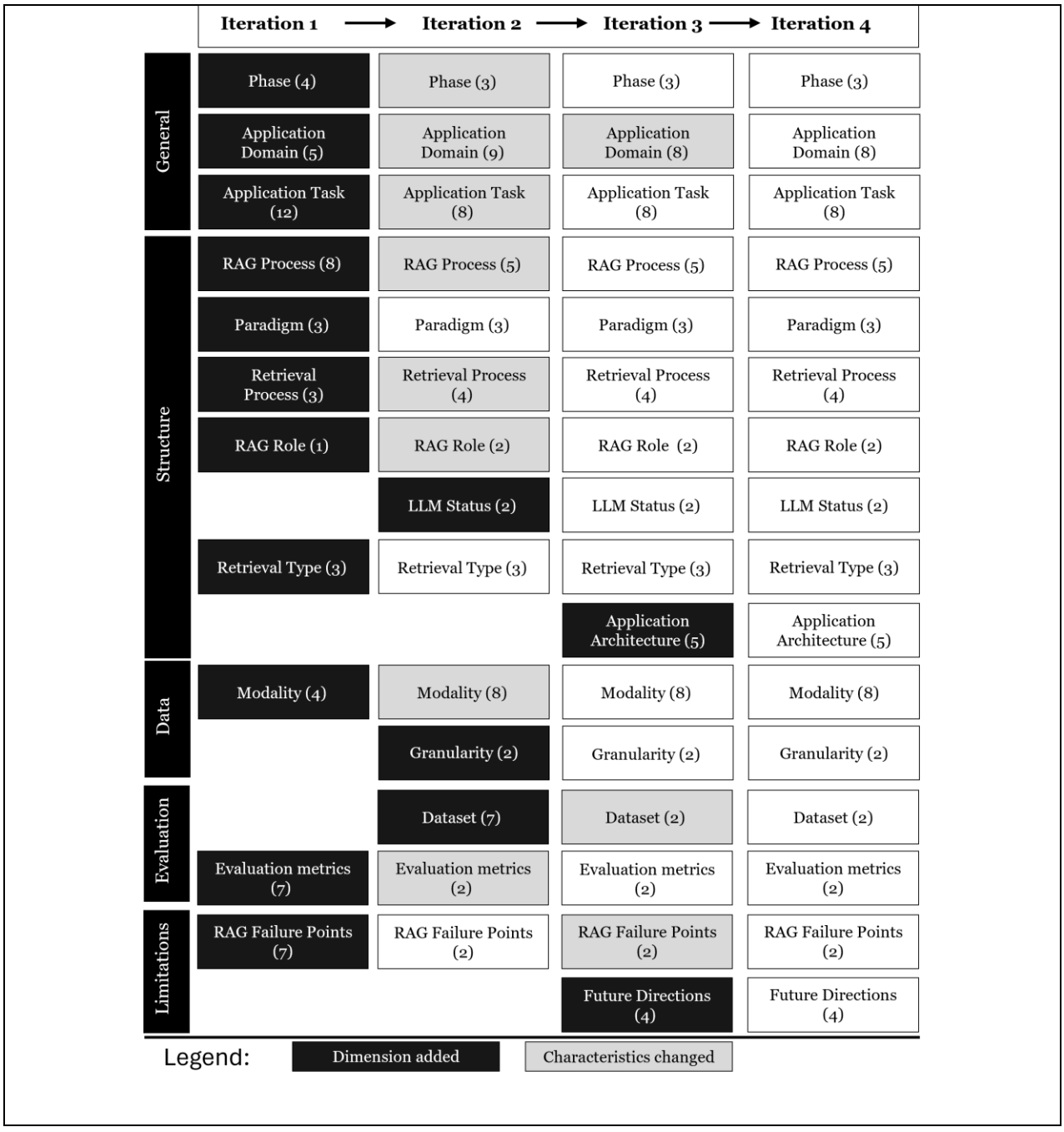

Fig. 1 – Development of taxonomy dimensions and characteristics (adapted from (Bräker et al. 2022; Remane et al. 2016))

Results

In this Section we present the final RAG taxonomy created from twenty-eight papers in four iterations. **Error! Reference source not found.** illustrates the entire taxonomy divided into five main components highlighted with color. In the following subsections, we describe each component in more detail.

Table 1: RAG Taxonomy created from twenty-eight papers within four iterations.

	Dimension	Characteristics							
General	Phase	*pre-training*			*inference*			*fine-tuning*	
	Application Domain	health	law	biology	General AI and NLP	ecology	education	research	media
	Application Task	*QA*	*IE*	*dialog*	*reasoning*	*slot filling*	*Mach. transl.*	*summarization*	*others*
Structure	Retrieval Process	*single retrieval*			*multiple retrieval*			*adaptive retrieval*	
	Paradigm	*naive RAG*			*advanced RAG*			*modular RAG*	
	RAG role	*complete system*				*subsystem*			
	LLM status	*not used*				*used*			
	RAG Process	*pre-retrieval*	*retrieval*		*post-retrieval*		*generator*		*post-generation*
	Retrieval Type	*sparse-vector retrieval*			*dense-vector retrieval*		*task-specific retrieval*		
	Application Architecture	*web app*	*local server*		*database*		*testing*		*configuration*
Data	Modalities	*nat. language*	*image*		*code*	*structured knowledge*	*audio*	*video*	*pdf*	*html*
	Granularity	*text – fine to coarse (e.g. document, chunk, sentence, phrase, token, proposition)*				*entity, triplet, sub-graph*			
Evaluation	Dataset	*publicly available*				*proprietary datasets*			
	Evaluation Metrics	*retrieval evaluation*				*generation evaluation*			
Limitations	RAG Failure Points	*internal limitations*				*component interaction limitations*			
	Future Directions	*more advanced research*		*efficient deployment and processing*		*incorporating long-tail and real-time knowledge*		*combination with other techniques*	

General

The first group of dimensions is devoted to general aspects of RAG systems regardless of specific structural aspects. Within this group, we identified three dimensions to represent the general aim and motivation of applied RAG.

D_1 Phase: The dimension subsumes the focus of RAG in place. It also relates to the evolving discourse – despite being a novel phenomenon. Research shows, that there are three primary areas of application for RAG. The resulting characteristics are *pre-training*, *inference*, and *fine-tuning* (Gao et al. 2024).

D_2 Application Domain: In total eight application domains are found in the discourse. The most frequent are *health*, *law*, *biology*, *general AI and NLP*, as well as *ecology* (Gao et al. 2024). Further application domains are *education* and *research* (Barnett et al. 2024) as well as *media* (Siriwardhana et al. 2023).

D_3 Application Task: RAG methods can be applied to different application or downstream tasks. In this dimension, the characteristics are the prominent tasks (Gao et al. 2024). We consider eight characteristics, i.e. *Question-Answering (QA)*, *Information Extraction (IE)*, *Dialog*, *Reasoning*, *Slot Filling* (Glass et al. 2021), *Machine Translation* (Li et al. 2022), *Summarization* (Zhao et al. 2024), *others*. QA has several different types, e.g. open domain QA, abstractive QA (Lewis et al. 2021), GraphQA (He et al. 2024), etc. Others characteristics contains some other tasks, for example Fact Checking/Verification, Question Generation (Lewis et al. 2021), Code search (Gao et al. 2024), and many more.

Structure

This includes an examination of the underlying technologies that form the architecture of the RAG application, determining whether the RAG acts as the principal system or merely a component within a larger system. We further delineate the structure of RAG systems by analyzing different RAG paradigms—such as naive, advanced, and modular RAG—which reflect varying levels of complexity and integration. Additionally, we consider the specific contexts or processes where the RAG retrieval is realized. In total, the structure includes key characteristics that distinguish RAG systems.

D_4 Retrieval Process: The retrieval process represents to what extent the RAG uses retrieval. *Single retrieval*, *multiple retrieval*, and *adaptive retrieval* are the identified characteristics (Gao et al. 2024). Single retrieval thus solely relies on a single retrieval sequence in an RAG, while multiple retrieval is an iterative or sequential approach. Adaptive retrieval is the most contextual approach as it integrates results of prior retrieval to adapt the next iteration of retrieval.

D_5 Paradigm: Gao et al., (2024) categorize RAG research paradigm into three *Naive RAG*, *Advanced RAG*, and *Modular RAG*. In this dimension, these three categories are the characteristics. In Naive RAG, there are three parts, i.e., indexing, retrieval, and generation. Advance RAG also includes pre-retrieval and post-retrieval parts before and after retrieval. Modular RAG provides flexibility with different modules, e.g., search module, memory module, etc.

D_6 RAG Role: Within the application landscape, the role of RAG systems can vary significantly: they can operate as dedicated, monolithic systems or as modular components integrated within other application systems (Zhao et al. 2024). According to the authors, subsystems can be part of larger architectures that employ multiple frameworks, i.e. RetDream for 3D Generation (Seo et al. 2024), R-ConvED for video captioning (Chen et al. 2023), and kNN-TRANX (Zhang et al. 2023) for text-to-code tasks. In the above-mentioned systems, RAG is used as an additional step to the pipeline, enchancing generation with the retrieved data.

D_7 LLM Status: This dimension is binary, it checks for the adaptability of the LLM in place. So it can either be *not used*, meaning no further approach is taken to improve the LLM performance, or it can be *used* (Cheng et al. 2024). Used thus lead to different forms. It can be trainable to be adjusted in each context or it can be looped as a specification of the paradigm *modular RAG*.

D_8 RAG Process: In this dimension, processes in RAG models are discussed mostly based on the information by Gao et al., (2024). Five characteristics are considered, i.e., pre-retrieval, retrieval, post-retrieval, generation, and post-generation. Pre-retrieval step involves some techniques applied before retrieval step, for instance, chunking, vectorizing, indexing, and some other strategies to e.g., optimize indexing, enhance user input, etc. In retrieval step, the relevant information to the user input is retrieved. Post-retrieval includes methods to improve the retrieved information during an integration with user input, e.g. re-rank the information or subgraph construction (He et al. 2024). In generation step, LLM provide a response for the prompt that contains the retrieved information and user input. Post-generation contains strategies that can be applied after generation, e.g. output rewrite (Zhao et al. 2024). Note that, there exist various modules to enhance different components (see Gao et al., (2024) for more information).

D_9 Retrieval Type: There are different types of retrieval augmentation methods (Li et al. 2022). In this dimension, three characteristics are considered i.e., sparse-vector retrieval, dense-vector retrieval, and task-specific retrieval (Li et al. 2022). Sparse-vector retrieval involves methods, e.g., TF-IDF, BM25, etc.

Dense-vector retrieval contains models based mostly on low-dimensional dense vectors, e.g., BERT-encoders and relying often on vector databases. In the task-specific retrieval, the retrieval module is based on the task (Li et al. 2022) and might comprise a database (Radeva et al. 2024). Some research works "directly use the edit distance between natural language texts (Hayati et al. 2018) or abstract syntax trees (AST) of code snippets (Poesia et al. 2021).

D10 Application Architecture: When developing a RAG system, in addition to the RAG structure, we also need to consider the structure of the final application and the interaction of the components. Radeva et al. (2023) present a web application RAG system that consists of the "local or server-based installation", "web application", "vector storage" (database), as well as the testing and configurations. Therefore, this dimension comprises web app, local server, database, testing, and configuration characteristics.

Data

In this group, dimensions are to discuss the data that RAG model might work with. Two dimensions are considered in this group, modalities, and granularity.

D_{11} Modalities: Although the concept of RAG was originally developed for text-based generation, its use has been adapted for a variety of other generation modalities (Chen et al. 2024; Gao et al. 2024; Lewis et al. 2020; Zhao et al. 2024). This includes programming code, audio, visual content, such as images and videos, 3D models, and other knowledge structures. The latter can include table structures, higher-level modeling languages, graphs, textual graph (He et al. 2024), or knowledge graphs. The fundamental principles of RAG remain similar across these different modalities, even though slight modifications to the augmentation methods are sometimes required.

D_{12} Granularity: This dimension is for different granularities of retrieved data based on the information by Gao et al., (2024). The modality can be natural language (or text), yet still the retrieved granularity might vary from fine to coarse, e.g., document, chunk, sentence, proposition, etc. (Gao et al., 2024). Similarly, there exists several granularities in structured data, e.g., sub-graph, triplet, entity, etc. (Gao et al., 2024).

Evaluation

The current group of dimensions tackles the problem of RAG Evaluation. We consider two dimensions: datasets and evaluation metrics that are crucial for the RAG system assessment.

D_{13} Dataset: Regarding the datasets used for RAG, most RAG surveys consistently list the datasets used regardless the application task, the RAG step to be evaluated on as well as the dataset availability. Thus, we focused on the matter of availability and consider two characteristics, i.e., publicly available, and proprietary datasets. Some examples for publicly available dataset are e.g. FEVER (Thorne et al. 2018), SQuAD (Rajpurkar et al. 2016) etc., and the dataset, e.g. by (Bondarenko et al. 2020), is an example for proprietary datasets.

D_{14} Evaluation Metrics: When reviewing papers discussing separate models and architectures, we can see that the authors mostly use task-specific metrics (Thakur and Vashisth 2024) or the generation output quality only (Chen et al. 2024). However, Gao et al. (2024) split evaluation metrics in two groups: retrieval evaluation and generation evaluation metrics, which are the base parts of RAG. The first group evaluates the relevance of the retrieved data to the query and is mostly represented with the ranking evaluation metrics: Precision@k, Recall@k, F@1, MRR, MAP (Gao et al. 2024). Second group involves generation evaluation metrics, such as BLEU, METEOR, ROUGE, PPL (Radeva et al., 2024) and Accuracy, Rejection Rate, Error Detection Rate, Error Correction Rate (Chen et al. 2024).

Limitation

Despite multiple advantages and ubiquitous application, Zhao et al. (2024) outline limitations and possible directions of RAG. We describe the last two dimensions in more detail, also considering failures from Barnett, S. et al. (2024).

D_{15} RAG Failure Points: RAG limitations can be divided into two groups: internal (related to the system components efficiency) and integration (related to the problems of RAG components iteraction). Here we discuss each type separately.

The most evident and the most frequent failure point for RAG is the retrieval step. Noises in retrieval results or missing the relevant content may drastically decrease the final performance as the information provided to the generator may contain irrelevant objects or misleading information. Barnett, S. et al. (2024) also state that the reason for that might be the missing content, e.g., "when asking a question that cannot be answered from the available documents". The next failure point is called "not in context" (Barnett, S. et al. 2024). In this case, the extracted documents were not correctly consolidated during the post-retreival process. The last three failure points relate to the generated output: the incorrected format of the output, incorrect specificity ("not specific enough or is too specific to address the user's need") and incomplete output that misses essential information even though being extracted by retriever.

When combining RAG with another system, the most common limitation is extra overhead: additional retrieval and interaction processes lead to increased latency of the system. Moreover, speed time also depends on the gap between retrievers and generators: the integration process and increaed system complexity might be other bottlenecks that should be considered. When applying RAG to LLMs or other generators with the limited context size, lengthy context might become a problem: the models might not be able to accept the whole retrieved data as input and the generation process will take much more time than expected.

D_{16} Future Directions: The last dimension outlines future directions for the RAG systems based on the findings of Zhao et al. (2024). The most straightforward directions is further development of RAG methodologies, enhancements, and applications. This might include new iteractions between the retriever and generator, various modular RAG architectures with looping, and more advanced pre- and post-processing steps. Another direction is efficient deployment and processing. When discussing limitations, most of the integration limitations were related to the efficiency and latency. Hopefully, future research on RAG capacities will allow shorter system response time and easier deployment. Another important research direction is the incorporation of long-tail and real-time knowledge. With the rapid growth of the data it is extremely difficult to constantly update large retrievals in RAG. Many existing works apply a static database for knowledge retrieval, which require re-indexing and/or computing additional representations. Zhao et al. (2024) expect newer techniques to solve the issue as well as provide better retrieval of less commonly referenced data. Lastly, the combination of other techniques might be also seen as a promising direction, e.g. integration of RAG with the new state space model architecture like Mamba (Gu and Dao 2023) or RWKV (Peng et al. 2023).

Discussion

Our taxonomy offers multiple contributions. Primarily, it frames the emerging digital phenomenon of RAG applications within a structured taxonomy, providing an initial conceptual foundation and a common understanding critical for advancing future research. The subsequent section will delve into the implications of this taxonomy, outline a detailed research agenda, and address potential limitations.

Research Stream	**Example Research Question**
Implications for Human-AI Collaboration and Decision Making	*What are the impacts of RAG applications on Human-AI collaboration?; How are RAGs influencing LLM-based agentic systems? How are RAGs improving human decision-making? How does RAG shift roles of human workers?; How does RAGs foster trust in humans?*
Developing Domain-specific applications	*How will RAGs affect Marketing/Innovation Management/Education/ and other applications? How will RAGs affect enterprise systems? How will RAGs enable conceptual modeling? How can DSR represent RAG design knowledge?*
Business Value of RAG Applications	*How do RAG applications improve efficiency in business processes? What is the impact on knowledge-intense, person-oriented work arrangements? How do RAGs perform compared to human workforce? Does and to what extend does human labor efficiency increases while applying RAG?*
Ethical, Legal, and Social Implications	*How can RAGs be used for addressing copyright issues inherent in current LLMs? How are RAGs impacting carbon footprints in LLM*

	applications? What is the impact of RAG on work conditions? How can safeguards be designed to guarantee an intended use?
Digital transformation and RAG implementation	*What are critical success factors for deploying RAG systems in ongoing application system? How can RAGs be implemented systematically in work environments? How are RAG-enabled orchestrations implementations affecting digital transformation?*

Table 2. Research Agenda for Socio-technical RAG Application Systems

Overall, our RAG applications taxonomy shows its close alignment with the core objectives of Information Systems, as highlighted by (Ågerfalk 2020), which was originally defined by (Buckingham et al. 1987, P. 18): "a system which assembles, stores, processes, and delivers information relevant to an organization (or to society), in such a way that the information is accessible and useful to those who wish to use it." This foundational definition of IS resonates with the essential functions of RAG applications, which are designed to make information not only more accessible but also more actionable and pertinent for users. By enhancing the retrieval and processing of information, RAG applications support information systems, ensuring that the delivered information is not just available but tailored to the users' needs, thereby increasing its utility and relevance.

RAG application systems are an emerging technology that has received considerable attention outside the IS community, which addresses the limitations of LLM applications (Gao et al. 2024; Leiser et al. 2024). While recent IS studies have begun to address both the applications of LLMs and methods to mitigate their shortcomings (e.g. (Benz et al. 2024; Drori et al. 2024; Schwartz and Te'eni 2024; Watson et al. 2024)), RAGs have not yet been fully recognized in the IS community or explored as a potential solution to these limitations. Thus, our RAG application taxonomy provides a basis for applying RAGs as an emerging technology for novel fields of application. Based on our taxonomy we see that the IS community can engage in a socio-technical perspective for guiding future RAG applications. Our research agenda is illustrated in Table 2.

LLM and genAI have sparked research into studying human-AI collaboration mechanisms and forms of interactions (e.g., (Benz et al. 2024; Drori et al. 2024; Feuerriegel et al. 2024)). Our taxonomy shows that the human interaction was not part of RAG applications, which provides areas for future research. We see the potential in implications on RAGs improving existing genAI systems and thus spark new questions around its relation towards agency (Park et al. 2023), with understanding how certain RAG characteristics (e.g. retrieval types) have a stronger impact on perceived agency between human and AI systems. LLMs are also increasingly used for supporting decision making (Storey et al. 2024), such as work delegation (Baird and Maruping 2021), yet the implications of RAGs might move the need for human-in-the-loop (Dellermann et al. 2019) away from the assumption of human primacy. Also, genAI can create both trust and mistrust (e.g. trough hallucinations) (Banh and Strobel 2023; Feuerriegel et al. 2024), and RAG applications can have a significant impact on human-AI trust making, with research on trust being an established field of research (Söllner et al. 2016).

Domain-specific applications: Our taxonomy shows that different mediums of generation are gaining interest, including conceptual modeling approaches (Baumann et al. 2024). For example, process modeling is already leveraging generative AI for improving and (re-) designing organizational processes (van Dun et al. 2023). RAGs appear to make such application systems much more viable, as our taxonomy provides us an example, where knowledge structures already incorporate conceptual process modeling types (Baumann et al. 2024). Thus, we see potential in incorporating RAGs into design science research endeavors (Hevner et al. 2004; Peffers et al. 2007; Teixeira et al. 2019) to address a variety of domain-specific applications. In our analyses, we identified proof-of-concepts and have seen little research on practical RAG applications. Thus, we call for future research to address this lack of proof-of-value (Nunamaker et al. 2015).

Business Value of RAG Applications: The impacts of AI have largely been due to increasing business value following business processes (Davenport and Ronanki 2018). Research into conversational agents has shown that they can lead to tangible business value (Kull et al. 2021; Mariani et al. 2023; McLean et al. 2021), whereas the assessment of LLM-based impact of business value remains under-studied(Storey et al. 2024). This could be attributed to current restraints of LLMs. However, with RAGs, there might be legitimate potential to create tangible business value, be it by also addressing knowledge- and labor-intensive services or improving work conditions.

Ethical, legal, social implications: Information systems is moving towards sustainability, including calls for studying social value (Nunamaker et al. 2015) and putting the need for reflecting on values in IS designs (e.g.: (Bednar and Spiekermann 2023; Friedman et al. 2013)). With LLMs already being discussed with ecological inefficiencies due to their potential high carbon emissions, integrating RAGs to improve LLM-applications might have unforeseen consequences. Similarly, we see ongoing discussions on the societal implications of improving works systems, leading to potential job losses (e.g. (Brynjolfsson et al. 2023)) or legal disputes about leveraging copyrighted content to generate new content (Golatkar et al. 2024; Samuelson 2023). Integrating RAGs can improve either, yet its ethical, legal, and social implications require careful consideration, exemplifying its socio-technical nature.

Digital transformation and RAG implementations: The challenges of adopting technologies, ,including AI (Grønsund and Aanestad 2020), firms are facing include high resistance to change and organizational barriers (Vial 2019). Since RAG applications address traditionally knowledge-intensive tasks, such as analysis and interpretation of data, aiming to outperform human capabilities, we see potential that RAG applications can lead to increased organizational resistances, especially when integrated into existing Enterprise application. RAGs can play a considerable role as an orchestrators of enterprise application systems (Böhmann et al. 2014) to call each functionality as part of its retrieval and generate respective outputs. This increasing complexity of heterogeneous applications might lead to new challenges for digital transformation.

Due to the fast publication cycle of RAG advances, we use the meta-search engine Google Scholar to collect an initial dataset for the RAG application taxonomy development while avoiding predatory outlets and allowing Arxiv as a paper repository. To ensure paper quality, we screened all existing RAG papers in Arxiv, and the included papers were merged with premier conference publications from related disciplines (e.g., Association of Computational Linguistics), thus our decision for a multidisciplinary taxonomy development team, which coincides with a discussion on the needs for multidisciplinary research (Nunamaker, Jr. et al. 2013). Our taxonomy is the first approach to conceptualizing RAG applications, and our novel approach to utilize LLMs as part of the taxonomy development also provides some methodological impacts toward computer-aided taxonomy development.

Furthermore, we adapted the taxonomy development process to our needs as we deal with a research topic that is skyrocketing in attention throughout many scientific communities. While still being among the first researchers within IS – remember, AIS eLibrary has no results – other domains with more pre-print-oriented cultures are already flourishing. Thus, we took advantage of that and incorporated ChatGPT to identify domains of interest for an initial set of dimensions and characteristics. This was accompanied by starting with focal survey papers to have an encompassing perspective on the research field. This guided our conceptualization of RAGs but also showed shortcomings in the completeness of those surveys. From our IS perspective, we could contribute a socio-technical perspective that helped to expand the taxonomy towards dimensions that are critical to organizational application.

Thus, the taxonomy is broad with sixteen dimensions. As the field is still evolving, we deem this initial breadth beneficial to shape our communities understanding. While the field is settling and maturing, a narrowed down taxonomy could be a next step, to further increase the conciseness and applicability, especially for practice. As of now, still expert knowledge is needed to assess several details within the taxonomy.

Our taxonomy development approach has several limitations, which can be attributed to the novelty of our phenomenon of interest. First, our data collection focuses on non-IS papers, primarily from computer linguistics, computer science, and engineering disciplines, which required deliberation among the research team. While the team members' backgrounds reflected each discipline to accommodate for this data collection shift, the sensemaking sessions were especially important in the first iteration to create a shared understanding (Brennan 1991). For example, an application for an RAG can either be understood as a genome sequencer or text summarizer, which shows that the synonymous use of concepts happens on different levels of abstraction. Nonetheless, this allows for a level of subjectivity that should be considered for future research. Additionally, while dealing with pre-prints in a fast moving research field, papers get updated while working with them, leading to inconsistencies for the research team, that need to be reworked afterwards. Furthermore, we do not make a claim for completeness, as the field is quickly progressing and we aim to capture an initial view of the emerging phenomenon, calling for future taxonomy extensions.

Conclusion

Our RAG taxonomy provides a structured way to categorize and analyze the diverse approaches, system features, and technologies that constitute RAG applications. Thus, we contribute to a clearer understanding of its components and their interactions. The taxonomy has five meta-dimensions, sixteen dimensions and sixty-six characteristics, reflecting the inherent complexities of current RAGs. This systematic classification is essential for IS researchers and practitioners to identify gaps in the current technology, facilitate research and development efforts, and identify potential use cases for real-world applications. Based on our taxonomy, we also present several avenues for future research, accommodating the RAG characteristics for different application types. Overall, the taxonomy not only enriches the academic discourse by providing a foundational framework for study and discussion but also guides practical implementations and innovations within the field.

References

Ågerfalk, P. J. 2020. "Artificial Intelligence as Digital Agency," *European Journal of Information Systems* (29:1), pp. 1–8.

Anthony, C., Bechky, B. A., and Fayard, A.-L. 2023. "'Collaborating' with AI: Taking a System View to Explore the Future of Work," *Organization Science* (34:5), INFORMS, pp. 1672–1694..

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. 2023. *Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection*, arXiv.

Baird, A., and Maruping, L. M. 2021. "The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts," *MIS Quarterly* (45:1), pp. 315–341.

Banh, L., and Strobel, G. 2023. "Generative Artificial Intelligence," *Electronic Markets* (33:1), p. 63.

Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., and Abdelrazek, M. 2024. *Seven Failure Points When Engineering a Retrieval Augmented Generation System*, arXiv.

Baumann, N., Diaz, J. S., Michael, J., Netz, L., Nqiri, H., Reimer, J., and Rumpe, B. 2024. "Combining Retrieval-Augmented Generation and Few-Shot Learning for Model Synthesis of Uncommon DSLs," in *Modellierung 2024 Satellite Events*, Gesellschaft für Informatik eV.

Beaulieu, T., Sarker, Suprateek, and Sarker, Saonee. 2015. "A Conceptual Framework for Understanding Crowdfunding," *Communications of the Association of Information Systems* (37).

Bednar, K., and Spiekermann, S. 2023. "The Power of Ethics: Uncovering Technology Risks and Positive Value Potentials in IT Innovation Planning," *Business & Information Systems Engineering*.

Benz, C., Riefle, L., and Satzger, G. 2024. "User Engagement and Beyond: A Conceptual Framework for Engagement in Information Systems Research," *Communications of the Association for Information Systems* (54), pp. 331–359.

Blom, J. D. 2010. *A Dictionary of Hallucinations*, New York, NY: Springer.

Böhmann, T., Leimeister, J. M., and Möslein, K. 2014. "Service Systems Engineering," *Business & Information Systems Engineering* (6:2), pp. 73–79.

Böhmann, T., Tuunanen, T., Maglio, P., and Fehrer, J. A. 2023. "Call for Papers - Special Issue on GenAl Service," *Journal of Service Research*.

Bondarenko, A., Braslavski, P., Völske, M., Aly, R., Fröbe, M., Panchenko, A., Biemann, C., Stein, B., and Hagen, M. 2020. "Comparative Web Search Questions," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, Houston TX USA: ACM, January 20, pp. 52–60.

Bräker, J., Hertel, J., and Semmann, M. 2022. "Conceptualizing Interactions of Augmented Reality Solutions," in *Proceedings of the 55th Hawaii International Conference on System Sciences*, , January 4.

Brennan, S. E. 1991. "Conversation with and through Computers," *User Modeling and User-Adapted Interaction* (1:1), pp. 67–86.

Brynjolfsson, E., Li, D., and Raymond, L. 2023. "Generative AI at Work," No. w31161, Cambridge, MA: National Bureau of Economic Research, April

Buckingham, R., Hirschheim, R., Land, F., and Tully, C. (eds.). 1987. *Information Systems Education: Recommendations and Implementation*, The British Computer Society Monographs in Informatics, Cambridge ; New York: Cambridge University Press.

Chen, J., Lin, H., Han, X., and Sun, L. 2024. "Benchmarking Large Language Models in Retrieval-Augmented Generation," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38), pp. 17754–17762.

Chen, J., Pan, Y., Li, Y., Yao, T., Chao, H., and Mei, T. 2023. "Retrieval Augmented Convolutional Encoder-Decoder Networks for Video Captioning," *ACM Transactions on Multimedia Computing, Communications, and Applications* (19:1s), 48:1-48:24.

Cheng, X., Luo, D., Chen, X., Liu, L., Zhao, D., and Yan, R. 2024. "Lift Yourself up: Retrieval-Augmented Text Generation with Self-Memory," *Advances in Neural Information Processing Systems* (36

Davenport, T. H., and Ronanki, R. 2018. "Artificial Intelligence for the Real World," *Harvard Business Review* (96:1), pp. 108–116.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. 2023. *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality*, SSRN Scholarly Paper

Dellermann, D., Ebel, P., Söllner, M., and Leimeister, J. M. 2019. "Hybrid Intelligence," *Business & Information Systems Engineering* (61:5), pp. 637–643.

Drori, I., Boston University / Columbia University, Te'eni, D., and Tel Aviv University. 2024. "Human-in-the-Loop AI Reviewing: Feasibility, Opportunities, and Risks," *Journal of the Association for Information Systems* (25:1), pp. 98–109.

van Dun, C., Moder, L., Kratsch, W., and Röglinger, M. 2023. "ProcessGAN: Supporting the Creation of Business Process Improvement Ideas through Generative Machine Learning," *Decision Support Systems* (165

Engel, C., Elshan, E., Ebel, P., and Leimeister, J. M. 2023. "Stairway to Heaven or Highway to Hell: A Model for Assessing Cognitive Automation Use Cases," *Journal of Information Technology*, SAGE Publications

Fatemi, B., Halcrow, J., and Perozzi, B. 2023. *Talk like a Graph: Encoding Graphs for Large Language Models*, arXiv.

Feuerriegel, S., Hartmann, J., Janiesch, C., and Zschech, P. 2024. "Generative AI," *Business & Information Systems Engineering* (66:1), pp. 111–126.

Friedman, B., Kahn, P. H., Borning, A., and Huldtgren, A. 2013. "Value Sensitive Design and Information Systems," in *Early Engagement and New Technologies: Opening up the Laboratory*, N. Doorn, D. Schuurbiers, I. van de Poel, and M. E. Gorman (eds.), Dordrecht: Springer Netherlands, pp. 55–95.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., and Wang, H. 2024. *Retrieval-Augmented Generation for Large Language Models: A Survey*, arXiv.

Gilardi, F., Alizadeh, M., and Kubli, M. 2023. "ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks," *Proceedings of the National Academy of Sciences* (120:30).

Glass, M., Rossiello, G., Chowdhury, M. F. M., and Gliozzo, A. 2021. *Robust Retrieval Augmented Generation for Zero-Shot Slot Filling*, arXiv.

Golatkar, A., Achille, A., Zancato, L., Wang, Y.-X., Swaminathan, A., and Soatto, S. 2024. *CPR: Retrieval Augmented Generation for Copyright Protection*, arXiv.

Grønsund, T., and Aanestad, M. 2020. "Augmenting the Algorithm: Emerging Human-in-the-Loop Work Configurations," *The Journal of Strategic Information Systems* (29:2), p. 101614.

Gu, A., and Dao, T. 2023. *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*, arXiv.

Hayati, S. A., Olivier, R., Avvaru, P., Yin, P., Tomasic, A., and Neubig, G. 2018. "Retrieval-Based Neural Code Generation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii (eds.), Brussels, Belgium: Association for Computational Linguistics, October, pp. 925–930.

He, X., Tian, Y., Sun, Y., Chawla, N. V., Laurent, T., LeCun, Y., Bresson, X., and Hooi, B. 2024. *G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering*, arXiv.

Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research," *Mis Quarterly* (28:1), pp. 75–105.

Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., and Neubig, G. 2023. *Active Retrieval Augmented Generation*, arXiv.

Kanbach, D. K., Heiduk, L., Blueher, G., Schreiter, M., and Lahmann, A. 2023. "The GenAI Is out of the Bottle: Generative Artificial Intelligence from a Business Model Innovation Perspective," *Review of Managerial Science*.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W. 2020. "Dense Passage Retrieval for Open-Domain Question Answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, pp. 6769–6781.

Kull, A. J., Romero, M., and Monahan, L. 2021. "How May I Help You? Driving Brand Engagement through the Warmth of an Initial Chatbot Message," *Journal of Business Research* (135), pp. 840–850.

Leiser, F., Eckhardt, S., Leuthe, V., Knaeble, M., Maedche, A., Schwabe, G., and Sunyaev, A. 2024. *HILL: A Hallucination Identifier for Large Language Models*, arXiv.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., and Rocktäschel, T. 2020. "Retrieval-Augmented Generation for Knowledge-Intensive Nlp Tasks," *Advances in Neural Information Processing Systems* (33), pp. 9459–9474.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. 2021. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, arXiv.

Li, H., Su, Y., Cai, D., Wang, Y., and Liu, L. 2022. *A Survey on Retrieval-Augmented Text Generation*, arXiv.

Manning, C. D., Raghavan, P., and Schütze, H. 2008. *Introduction to Information Retrieval*, New York: Cambridge University Press.

Mariani, M. M., Hashemi, N., and Wirtz, J. 2023. "Artificial Intelligence Empowered Conversational Agents: A Systematic Literature Review and Research Agenda," *Journal of Business Research* (161), p. 113838.

Martino, A., Iannelli, M., and Truong, C. 2023. "Knowledge Injection to Counter Large Language Model (LLM) Hallucination," in *The Semantic Web: ESWC 2023 Satellite Events*, Lecture Notes in Computer Science, C. Pesquita, H. Skaf-Molli, V. Efthymiou, S. Kirrane, A. Ngonga, D. Collarana, R. Cerqueira, M. Alam, C. Trojahn, and S. Hertling (eds.), Cham: Springer Nature Switzerland, pp. 182–185.

McGeorge, D. 2023. *The ChatGPT Revolution: How to Simplify Your Work and Life Admin with AI*, Wiley.

McGrath, Q. 2024. "Responding to the Sharp Rise in AI in the 2023 SIM IT Trends Survey," *MIS Quarterly Executive* (23:1).

McKinsey. 2023. "What Is ChatGPT, DALL-E, and Generative AI? | McKinsey," , November 1. (https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai, accessed November 3, 2023).

McLean, G., Osei-Frimpong, K., and Barhorst, J. 2021. "Alexa, Do Voice Assistants Influence Consumer Brand Engagement? – Examining the Role of AI Powered Voice Assistants in Influencing Consumer Brand Engagement," *Journal of Business Research* (124), pp. 312–328.

Nickerson, R. C., Varshney, U., and Muntermann, J. 2013. "A Method for Taxonomy Development and Its Application in Information Systems," *European Journal of Information Systems* (22:3), pp. 336–359.

Nunamaker, J. F., Briggs, R. O., Derrick, D. C., and Schwabe, G. 2015. "The Last Research Mile: Achieving Both Rigor and Relevance in Information Systems Research," *Journal of Management Information Systems* (32:3), pp. 10–47.

Nunamaker, Jr., J., Twyman, N., and Giboney, J. 2013. "Breaking out of the Design Science Box: High-Value Impact Through Multidisciplinary Design Science Programs of Research," *AMCIS 2013 Proceedings*.

OpenAI. 2023. "Introducing GPTs," , November 6. (https://openai.com/blog/introducing-gpts, accessed November 10, 2023).

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. 2022. *Training Language Models to Follow Instructions with Human Feedback*, arXiv.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. 2023. "Generative Agents: Interactive Simulacra of Human Behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA: Association for Computing Machinery, October 29, pp. 1–22.

Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. 2007. "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems* (24:3), pp. 45–77.

Peng, B., Alcaide, E., Anthony, Q., Albalak, A., Arcadinho, S., Biderman, S., Cao, H., Cheng, X., Chung, M., Derczynski, L., Du, X., Grella, M., Gv, K., He, X., Hou, H., Kazienko, P., Kocon, J., Kong, J., Koptyra, B., Lau, H., Lin, J., Mantri, K. S. I., Mom, F., Saito, A., Song, G., Tang, X., Wind, J., Woźniak, S., Zhang, Z., Zhou, Q., Zhu, J., and Zhu, R.-J. 2023. "RWKV: Reinventing RNNs for the Transformer Era," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali (eds.), Singapore: Association for Computational Linguistics, December, pp. 14048–14077.

Poesia, G., Polozov, A., Le, V., Tiwari, A., Soares, G., Meek, C., and Gulwani, S. 2021. *Synchromesh: Reliable Code Generation from Pre-Trained Language Models*, presented at the International Conference on Learning Representations, , October 6.

Radeva, I., Popchev, I., Doukovska, L., and Dimitrova, M. 2024. *Web Application for Retrieval-Augmented Generation: Implementation and Testing*, Preprints.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. 2023. "Direct Preference Optimization: Your Language Model Is Secretly a Reward Model," *Advances in Neural Information Processing Systems* (36), pp. 53728–53741.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. 2016. "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, J. Su, K. Duh, and X. Carreras (eds.), Austin, Texas: Association for Computational Linguistics, November, pp. 2383–2392.

Remane, G., Nickerson, R., Hanelt, A., Tesch, J., and Kolbe, L. 2016. "A Taxonomy of Carsharing Business Models," *ICIS 2016 Proceedings*.

Samuelson, P. 2023. "Generative AI Meets Copyright," *Science* (381:6654), pp. 158–161..

Schwartz, D., and Te'eni, D. 2024. "AI for Knowledge Creation, Curation, and Consumption in Context," *Journal of the Association for Information Systems* (25:1), pp. 37–47.

Seo, J., Hong, S., Jang, W., Kim, I. H., Kwak, M., Lee, D., and Kim, S. 2024. *Retrieval-Augmented Score Distillation for Text-to-3D Generation*, arXiv.

Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J. 2021. "Retrieval Augmentation Reduces Hallucination in Conversation," in *Findings of the Association for Computational Linguistics: EMNLP*

2021, M.-F. Moens, X. Huang, L. Specia, and S. W. Yih (eds.), Punta Cana, Dominican Republic: Association for Computational Linguistics, November, pp. 3784–3803.

Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rana, R., and Nanayakkara, S. 2023. "Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering," *Transactions of the Association for Computational Linguistics* (11), MIT Press One Broadway, 12th Floor, Cambridge, Massachusetts 02142, USA ..., pp. 1–17.

Söllner, M., Benbasat, I., Gefen, D., Leimeister, J. M., and Pavlou, P. 2016. "Trust - Research Curation," *MIS Quarterly*.

Soudani, H., Kanoulas, E., and Hasibi, F. 2024. *Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge*, arXiv.

Storey, V. C., Hevner, A. R., and Yoon, V. 2024. "The Design of Human-Artificial Intelligence Systems in Decision Sciences: A Look Back and Directions Forward," *Decision Support Systems*, p. 114230.

Teixeira, J. G., Patrício, L., and Tuunanen, T. 2019. "Advancing Service Design Research with Design Science Research," *Journal of Service Management* (30:5), pp. 577–592.

Thakur, A., and Vashisth, R. 2024. *Loops On Retrieval Augmented Generation (LoRAG)*, arXiv..

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. 2018. "FEVER: A Large-Scale Dataset for Fact Extraction and VERification," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, M. Walker, H. Ji, and A. Stent (eds.), New Orleans, Louisiana: Association for Computational Linguistics, June, pp. 809–819.

Uba, C., Lewandowski, T., and Böhmann, T. 2023. "The AI-Based Transformation of Organizations: The 3D-Model for Guiding Enterprise-Wide AI Change," in *Proceedings of the 56th Hawaii International Conference in System Sciences*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. 2017. "Attention Is All You Need," in *Advances in Neural Information Processing Systems* (Vol. 30), Curran Associates, Inc.

Vial, G. 2019. "Understanding Digital Transformation: A Review and a Research Agenda," *The Journal of Strategic Information Systems* (28:2), pp. 118–144..

Watson, R. T., Song, Y. (April), Zhao, X., and Webster, J. 2024. "Extending the Foresight of Phillip Ein-Dor: Causal Knowledge Analytics," *Journal of the Association for Information Systems* (25:1), pp. 145–157.

Wessel, M., Adam, M., Benlian, A., Majchrzak, A., and Thies, F. 2023. "Generative AI and Its Transformative Value for Digital Platforms," *Journal of Management Information Systems*.

Zhang, X., Zhou, Y., Yang, G., and Chen, T. 2023. "Syntax-Aware Retrieval Augmented Code Generation," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali (eds.), Singapore: Association for Computational Linguistics, December, pp. 1291–1302

Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., and Cui, B. 2024. *Retrieval-Augmented Generation for AI-Generated Content: A Survey*, arXiv

Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Chen, H., Dou, Z., and Wen, J.-R. 2024. *Large Language Models for Information Retrieval: A Survey*, arXiv