JAISCR, 2025, Vol. 15, No. 2, pp. 115 – 146
10.2478/jaiscr-2025-0007

A COMPREHENSIVE SURVEY OF RETRIEVAL-AUGMENTED LARGE LANGUAGE MODELS FOR DECISION MAKING IN AGRICULTURE: UNSOLVED PROBLEMS AND RESEARCH OPPORTUNITIES

Artem Vizniuk[1], Grygorii Diachenko[2], Ivan Laktionov[2,*],
Agnieszka Siwocha[3], Min Xiao[4], Jacek Smoląg[5]

[1] *CLOUD FLOW LLC,*
Str. Khotkevycha Hnata, 12, UA02094 Kyiv, Ukraine

[2] *Dnipro University of Technology,*
av. Dmytra Yavornytskoho, 19, Dnipro, UA49005, Ukraine

[3] *Information Technology Institute, SAN University,*
90-113, Łódź, Poland

[4] *Nanjing University of Posts and Telecommunications,*
College of Automation & College of Artificial Intelligence, Nanjing, 210003, China

[5] *Częstochowa University of Technology, Department of Artificial Intelligent,*
Al. Armii Krajowej, 36, Częstochowa, 42-201, Poland

[*] *E-mail: Laktionov.I.S@nmu.one*

Submitted: 9th September 2024; Accepted: 1st December 2024

Abstract

The breakthrough in developing large language models (LLMs) over the past few years has led to their widespread implementation in various areas of industry, business, and agriculture. The aim of this article is to critically analyse and generalise the known results and research directions on approaches to the development and utilisation of LLMs, with a particular focus on their functional characteristics when integrated into decision support systems (DSSs) for agricultural monitoring. The subject of the research is approaches to the development and integration of LLMs into DSSs for agrotechnical monitoring. The main scientific and applied results of the article are as follows: the world experience of using LLMs to improve agricultural processes has been analysed; a critical analysis of the functional characteristics of LLMs has been carried out, and the areas of application of their architectures have been identified; the necessity of focusing on retrieval-augmented generation (RAG) as an approach to solving one of the main limitations of LLMs, which is the limited knowledge base of training data, has been established; the characteristics and prospects of using LLMs for DSSs in agriculture have been analysed to highlight trustworthiness, explainability and bias reduction as priority areas of research; the potential socio-economic effect from the implementation of LLMs and RAG in the agricultural sector is substantiated.

Keywords: large language models, agriculture, decision-making, retrieval-augmented generation.

1 Introduction

1.1 Relevance of the topic

Agriculture plays a key role in the global economy, contributing significantly to food security, employment, and trade. Over the past two decades, the agricultural sector has undergone significant transformation driven by technological advances and increasing production requirements. According to the Food and Agricultural Organisation (FAO), global agricultural production has grown steadily, with volumes of major agricultural goods increasing by 54% from 2000 to 2021, reaching 9.5 billion tonnes in 2021. This surge was crucial to feed a growing population, which grew by 29% over the same period [1].

As agriculture continues to face challenges such as climate change and resource depletion, there is an increasing need for improved agricultural practices. Modern information and digital technologies can significantly optimise agricultural processes by increasing yields, reducing the use of pesticides, and others [2]. Numerous national [3-7] and international programmes [8-15] (see Figure 1) emphasise the importance of integrating IT into agriculture to drive innovation, ensure food security, and combat climate change. For example, in Ukraine, agriculture not only forms the basis of the national economy but also significantly contributes to the global food supply [16].

In line with the global digital transformation trend, Ukraine has made significant progress in implementing IT solutions in agriculture. However, with the growing complexity of agricultural systems, there is an increasing need for advanced artificial intelligence technologies, such as LLMs, to implement the conceptual framework of automated decision support leveraging big data. World practice has proven that the involvement of LLMs can potentially increase the efficiency of known information and digital technologies for agricultural purposes [17].

1.2 Historical context

The development of the LLMs can be broadly divided into three stages [17, 18]:

1. Statistical language models (SLMs) that use statistical approaches, such as n-gram, which calculates the probability of a character/word occurring depending on the previous n-1 characters/words. The probability is calculated depending on the frequency and order of occurrence of words in the text. Such models are used for code autocompletion, spell-checking, and more. Disadvantages: limited context, poor scalability when the parameter n increases (requires more computations and memory), and lack of semantics as it builds statistics based on word frequencies.

2. Neural language models (NLMs) have improved the understanding of semantics compared to n-gram models. Feed-forward neural networks have made it possible to study the vector representation of words depending on their context in the text (semantically similar words correspond to vectors close to each other in the vector space) [19]. Over time, the development of the recurrent neural network architecture, namely long-short term memory (LSTM), has improved the processing of sequential data and the understanding of long-term dependencies [19].

3. The emergence of the Transformer architecture. In 2017, the Google research team proposed the Transformer architecture [20], which became the foundation for modern LLM architectures (such as GPT). The architecture of the transformers contributed to faster learning and scalability of these models due to parallelisation, and the self-attention mechanism used improved the un-

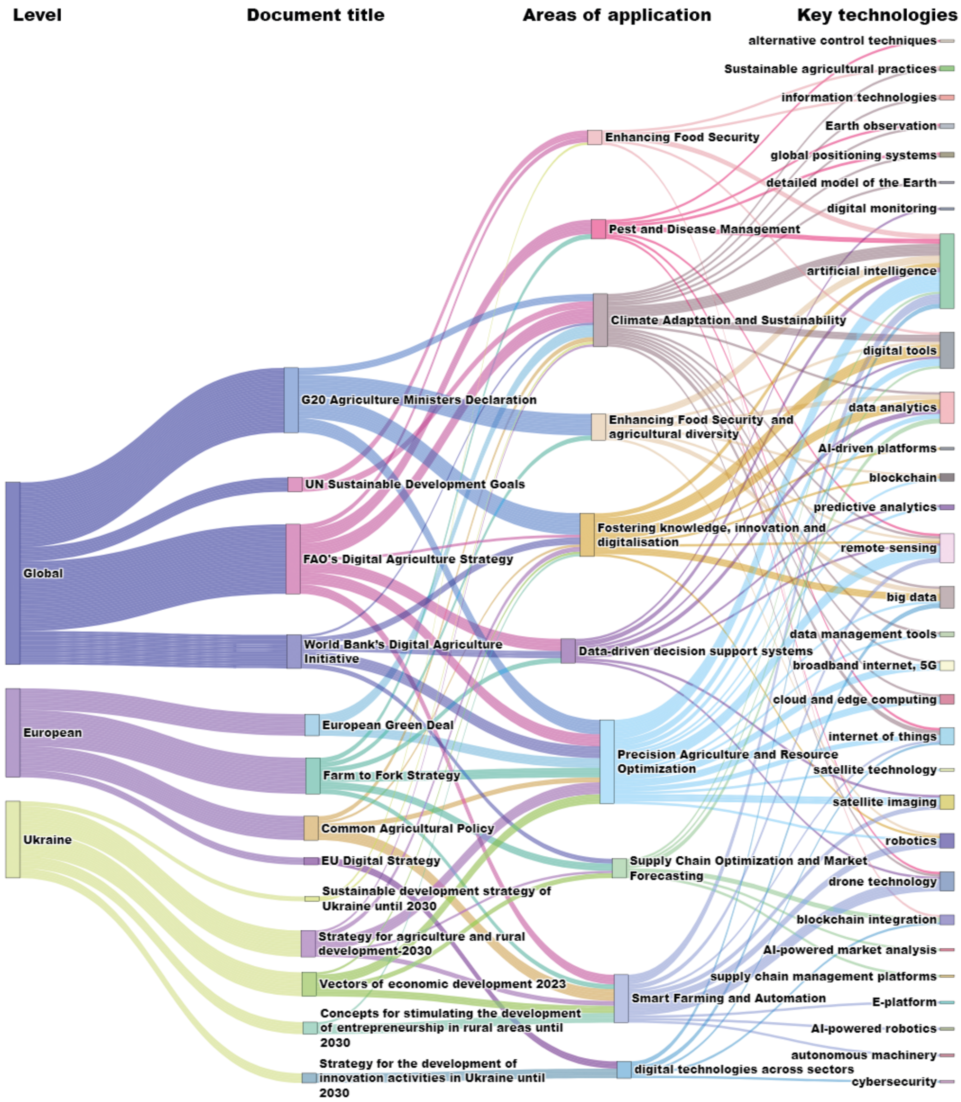


Figure 1. Sankey diagram of prioritised objectives and state-of-the-art technologies in agriculture according to actual strategies, concepts and policies at European, Ukraine and global levels

derstanding of long-term dependencies compared to RNN counterparts.

Despite these achievements, the application of LLMs in agriculture remains in its early stages and requires further research. Given the complexity of decision-making in agriculture, from yield forecasting to managing resource allocation, LLMs have the potential to become a powerful tool to support farmers and stakeholders in making more informed and timely decisions through their ability to process and interpret large amounts of data, provide contextualised recommendations and assist in decision-making, which can have a significant impact on agricultural productivity and sustainability. Therefore, research on the integration of LLMs into agriculture should be a priority for this technology to reach its full potential and be practically oriented in the global context of agricultural activities.

1.3 Real-world examples of leveraging LLM for various domains

The relevance of the development, research, and implementation of applied LLMs is confirmed by the scale of their application in various industries that require understanding, generation, and analysis of natural language. The most common applied tasks and a list of LLM application areas [21-31] are demonstrated in Figure 2. It should be noted that this list is not exhaustive since every year, a significant number of publications appear at international symposia, conferences, and scientific journals devoted to the development and validation of new algorithms and applications of LLMs. According to IEEE alone, the total number of articles on the search query "Large Language Models" is more than twenty-four thousand (as of September 26, 2024), and with the development of technology, their role in various industries will continue to expand [32-52].

What distinguishes LLMs from previous virtual assistants such as Siri (Service Interface for Real-time Information) [26] and Alexa [27] is their versatility. LLMs can perform a wide range of tasks related to language learning. In addition to the top 6 use cases in Figure 2, it is worth noting that LLMs perfectly understand the nuances of human language and can also translate text from one language to another with extreme accuracy. In addition, they can serve as highly personalised virtual assistants to help with a wide range of tasks in various fields. Real-world examples of LLM implementation include:

- Conversational search Engines [53-57]. Bing is a search engine from Microsoft based on GPT models that works in a dialogue mode. Users can send complex multi-part queries, and LLM generates detailed answers with links to source pages in the search interface based on the history of messages. This approach has significantly improved user experience and search quality [36].
- Finance [58-61]. COIN (contract intelligence) at JPMorgan Chase is a natural language processing (NLP) model that significantly improves business efficiency by automating the verification of commercial credit agreements, which previously required 360000 hours of manual work annually. It reduces operational costs, improves accuracy and ensures consistency in data extraction. COIN is also efficiently scalable, allowing more contracts to be processed without increasing the number of employees [62].
- Healthcare [63-68]. The MedPaLM-2 model is trained on medical data and aims to provide high-quality answers to medical inquiries. MedPaLM-2 is the first model to achieve an expert level in answering questions on the United States Medical Licensing Examination (USMLE) [69].
- Cybersecurity and privacy. In this area, LLMs are used to create automated reports by processing security-related data. These reports help organisations identify potential cybersecurity vulnerabilities and threats faster and more accurately so that appropriate action can be taken to mitigate them. LLMs can also detect vulnerabilities and bugs in existing code and generate more secure code. The use of LLMs can help detect potential cyberattacks and anomalies by

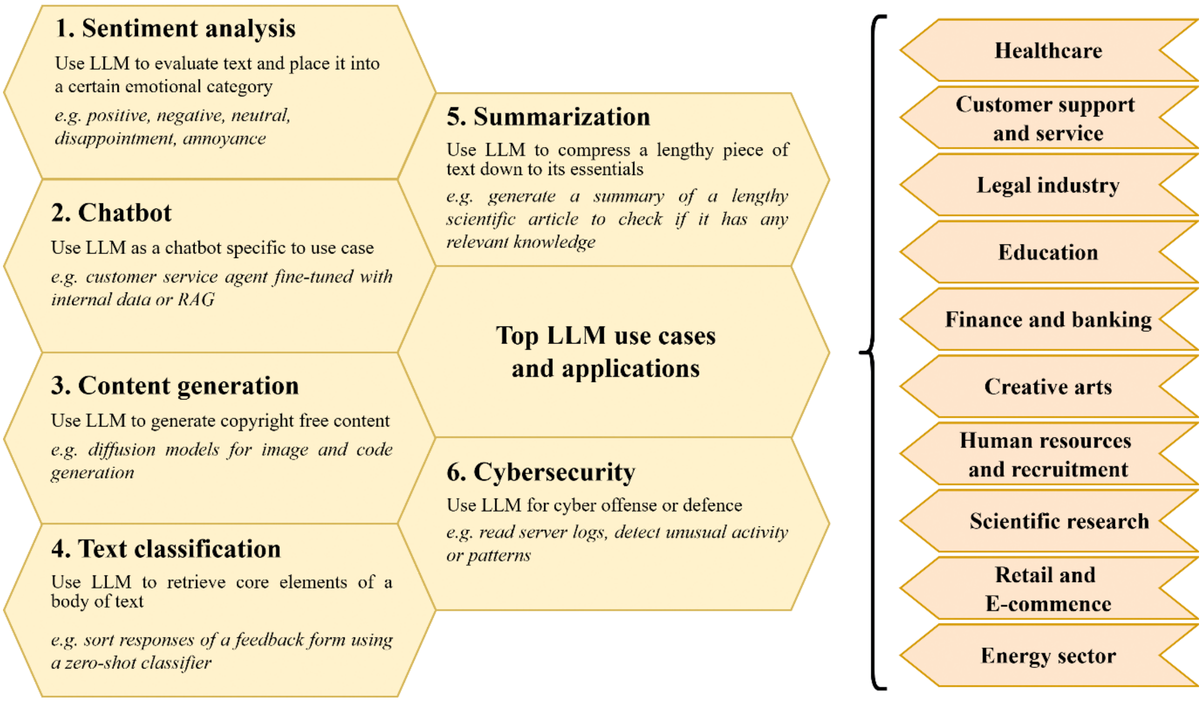


Figure 2. Application areas of LLM

analysing Internet logs, security event alerts, and other security-related data [43-47].

- Legal industry [70-72]. LLMs are transforming legal research, contract analysis, and document drafting, significantly speeding up processes that traditionally required significant human effort. Example: ROSS Intelligence [29] is a tool for legal research. The developed LLM allows lawyers to enter queries in natural language and provides relevant legal precedents and court materials, which significantly reduces research time.

- Education [73-77]. LLMs are used to create educational content, provide tutoring, and automate administrative tasks. Example: Duolingo uses LLMs [30] to improve language learning by creating personalised language exercises and providing real-time feedback with natural language explanations.

LLMs are also used in such industries as Creative arts [78-80], Human resources and recruitment [81-84], Retail and E-commerce [85-87], Energy sector [88-91] and Scientific research [92-95].

This widespread adoption of LLMs across diverse industries highlights their potential for transformative applications in other fields, including agriculture. The possible use-cases of LLMs in agronomic practice are shown in Figure 3 [96-98].

Real-world examples of successful implementation of LLMs in various industries and businesses, including agriculture [99, 100], prove the prospects of their application in building intelligent systems and technologies for automated decision support for agronomists to improve the efficiency of planning and implementation of agrotechnical measures for growing crops, provided that additional research is carried out.

1.4 Localisation of the expected scientific and applied effect

As stated above, to date, there have been significant theoretical and applied advances in the development and utilisation of LLMs in various sectors. However, it is important to note that the global processes of intellectualisation and digitalisation of agriculture require additional research to substantiate approaches to ensure the functional compatibility of LLMs

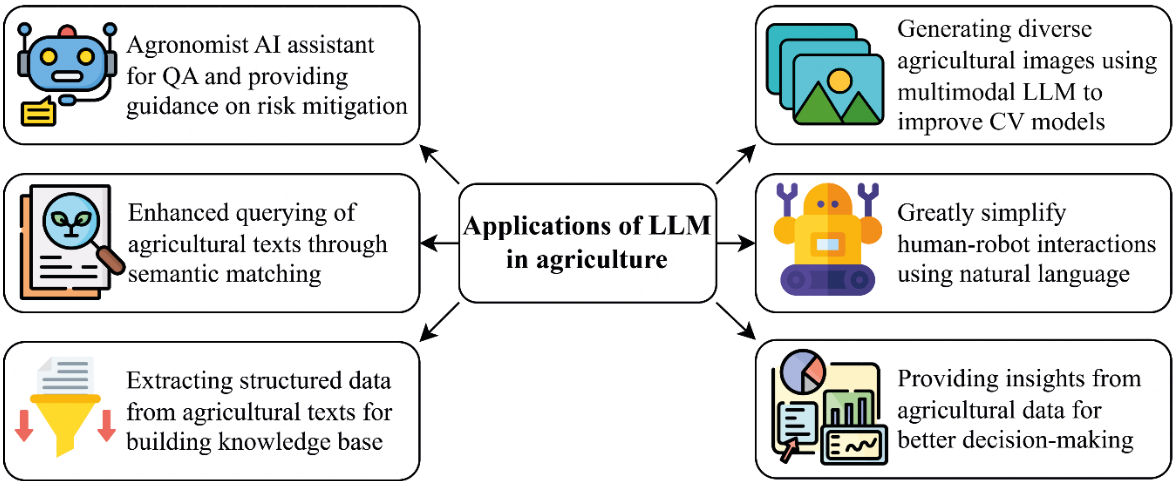


Figure 3. Applications of LLM in agriculture

when they are integrated into intelligent information technologies for automated decision support.

Therefore, based on the aforesaid, it can be asserted that the main expected scientific and applied effect of the research of this article is to systematise the current state of functional characteristics of LLMs and to substantiate the vector of promising research on the development of approaches to their integration into information technologies for decision support for agrotechnical purposes, which, in turn, will optimise agricultural processes for growing crops in changing agroclimatic conditions and, as a result, positively affect the long-term sustainability and investment attractiveness of agricultural enterprises.

1.5 Aim, object, and subject of the research

Based on the analysis and synthesis of the known literature sources in the previous subsections of the article, the following has been established. Today, LLMs play a key role in optimising the processes of analysing and interpreting large amounts of data, supporting decision-making and managing agrotechnical processes, which significantly increases the efficiency of the work of agrarians and farmers. The main functional purpose of LLMs is to assist agronomists in determining rational approaches to growing crops in changing agro-climatic conditions with further planning of an algorithm of actions to ensure the stress resistance of crops. Consequently, the main motivation for conducting the research in this article is to disclose the current state of LLMs with a focus on their functional suitability and existing limitations in their use in the agricultural sector, which will allow substantiating promising areas for further research to improve the practical aspects of using LLMs as functional components of decision support information technologies in the agricultural sector.

The aim of this article is to critically analyse and generalise the known results and research directions on approaches to the development and utilisation of LLMs, with a particular focus on their functional characteristics when integrated into DSSs for agricultural monitoring.

The object of research is information processes of intellectual processing of agrotechnical monitoring results based on LLMs.

The subject of research is approaches to the development and integration of LLMs into DSSs for agrotechnical monitoring.

1.6 Structure and organisation of the article

This article is divided into five sections: introduction, methods, results, discussion and conclusions. The first section (introduction) justifies the relevance of the topic and the importance of conducting research on the implementation and application of LLM in the agricultural sector. The second section (methods) contains a description of the methodology used to search for sources of information. The third section (results) contains information on the critical analysis and summarisation of existing research on LLMs, RAG and their application in agriculture. The fourth section is devoted to a discussion of the results of the analysis, including promising areas for LLM and RAG research in the agricultural sector. The fifth section contains the main conclusions of the conducted research.

2 Methods

2.1 Information sources and search strategy

This research is devoted to a comprehensive analysis and generalisation of the latest research on the development and use of LLMs and RAG, as well as to identifying the prospects for their implementation and application in the agricultural sector and their integration into DSSs. Therefore, the main approaches used in the research in this article are: information and analytical search, critical analysis and logical generalisation of the latest research in the areas of LLMs, RAG and their application to the agricultural sector in the creation of DSSs. The criteria and characteristics of the search and analysis of known research and development results used in this article are given in Table 1.

The research in this article was carried out in four stages: in the first stage, the importance of using IT in the field of agriculture to ensure food supply in the world, in particular, LLM, was substantiated by the example of real systems in other business areas that actively use them; at the second stage, generalised analysis of the current state of development of LLMs and RAG was carried out; at the third stage, the analysis of LLMs and RAG research in the context of their use in the field of agriculture, in particular, implementation in DSS, was carried out; in the fourth stage, promising directions for further research were substantiated.

2.2 Data items

Here is the revised text with no extra line breaks:

This article contains 157 references to sources, which include scientific articles in periodicals, conference materials, and analytical and information resources. The criteria used for the search and the general characteristics of the primary sources of the article are given in Table 1.

As can be seen from the characteristics of the information and analytical search presented in Table 1, this article highlights the most relevant publications (for the last 5 years) related to the principles of development and utilisation of LLMs and RAG, as well as the specific features of their application and development prospects in the agricultural sector. In addition, during the search and analysis of well-known scientific works, special attention is given to existing review articles on the use of LLMs in various areas of agriculture.

A world map with shaded regions of the countries of the authors of the cited articles and the corresponding number of references for each region is shown in Figure 4. A graphical interpretation of the results of the statistical analysis of the reviewed scientific and analytical sources by type and year of publication is also presented in Figure 5.

As can be seen from the analysis of graphical interpretation (see Figure 4), the largest number of papers are by authors from such countries as the United States (38 references), China (27 references), and India (14 references). This explains the fact that these countries are leaders in terms of the scale of agricultural activities and the introduction of modern information technologies in the agricultural sector and, as a result, are among the largest exporters of agricultural products in the world.

3 Results

3.1 General analysis of LLMs

Thanks to the invention of the transformer architecture and self-attention mecha-

Table 1. Characteristics of search and analysis of known scientific information sources

Category	**The criteria and characteristics of information-analytical search and analysis applied**
Primary range of years of publication activity	From 2019 to 2024
Extended range of years of publication activity	From 1997 to 2024
Scientometric databases	Scopus, Web of Science
Main digital libraries	arXiv, IEEE Xplore, ScienceDirect, MDPI
Analytical databases	FAOSTAT Database
Types of literature sources	Scientific articles in periodicals, peer-reviewed publications, international conference proceedings, analytical and information resources
Primary language	English
Additional language	Ukrainian
Primary subject areas	Agriculture, LLM, information retrieval
Primary search queries	LLM in agriculture (review), LLM review, RAG in agriculture, RAG review, agronomist AI assistant
Additional keywords	Survey, embedding, sparse & denser encoders, information retrieval or search, prompt engineering, rerankers, semantic caching, chunking, preprocessing, indexing, multimodal LLM or RAG, query expansion or refinement, LLM for decision-making

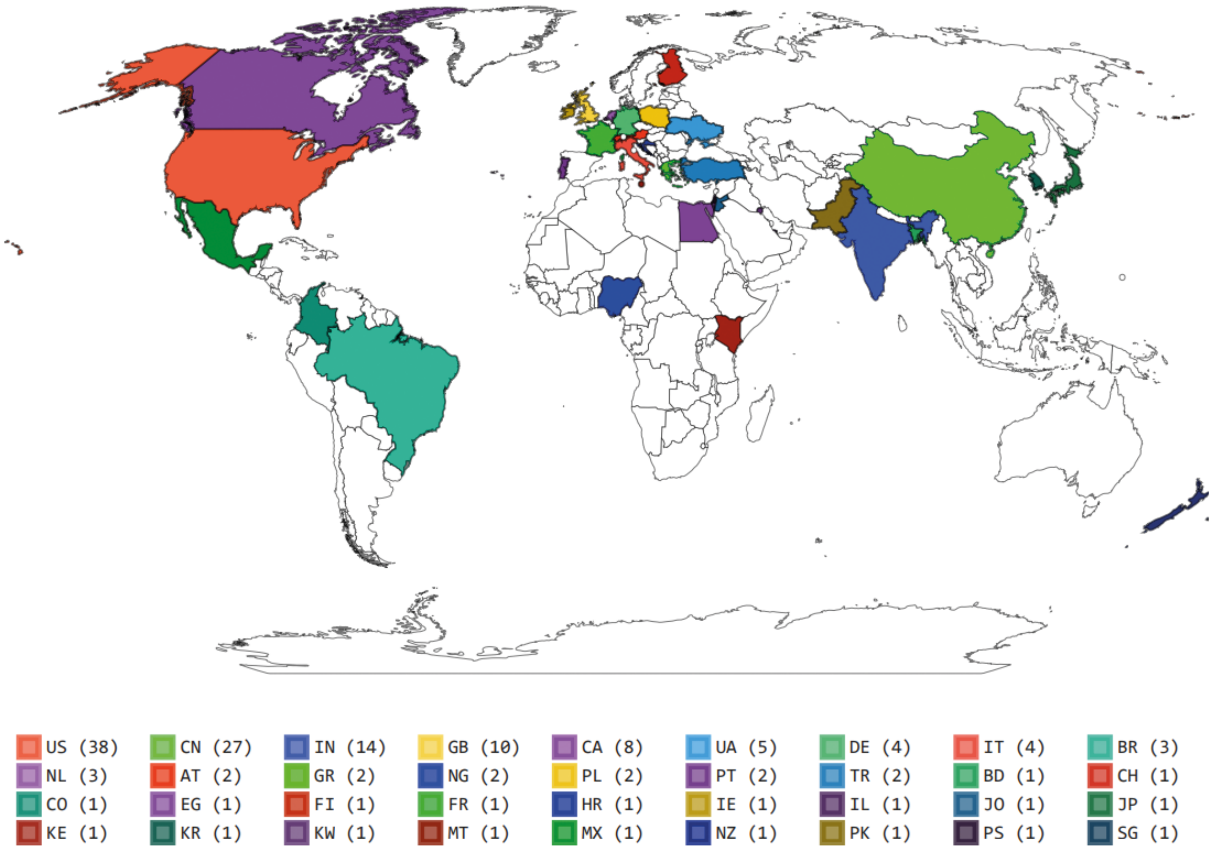
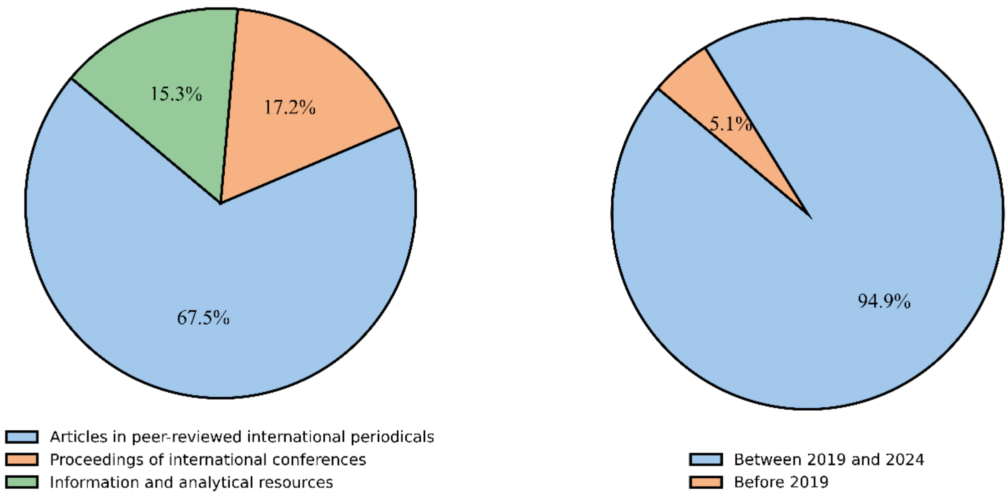


Figure 4. World map of references per country



(a) Types of literature sources (b) Years of publication activity

Figure 5. Statistics of the reviewed sources of information

nism, LLMs have excelled in NLP tasks, demonstrating impressive reasoning capabilities of generating human-like text based on a given prompt [20, 101].

LLMs can be broadly divided into three main categories: Encoder-only, Decoder-only, and Encoder-Decoder models. The key feature of the Encoder-only models is their bi-directional nature, allowing them to take into account both the left and right context of each token when encoding it. This bi-directionality helps Encoder-only models better understand the meaning of words in context, which is crucial for tasks like sentiment analysis, text classification, semantic search, and others. The primary focus of such models is on text comprehension rather than text generation. Bidirectional encoder representations from transformers (BERT) is an example of Encoder-only architecture. In contrast, Decoder-only models generate text in a left-to-right autoregressive fashion. For example, GPT models predict the next token in a sequence based on the context provided by the previous tokens. Their architecture makes them effective for language generation, code generation, and creative writing tasks. Encoder-Decoder models, such as Text-To-Text Transfer Transformer (T5), utilise both the encoder (for processing the input sequence to capture its meaning) and decoder (for generating the output sequence based on the encoded information). Such models are used for tasks like machine translation, summarisation, and conversational response generation [102].

The complexity of LLMs (GPT-4o boasts over 200 billion parameters), the intricacies of their architecture [20], and training on large and diverse text corpora (web pages, research papers, books, programming code, and others) allow them to develop their own internal knowledge base. The quality of the knowledge base is determined by the volume, diversity and trustworthiness of the training data.

The internal knowledge base of LLMs is limited by the training data. As a result, it can become outdated, leading to the generation of stale information. One way to address this issue is by fine-tuning a pre-trained LLM on more recent, custom data. However, this approach has significant drawbacks, such as the necessity of retraining a model each time the knowledge base needs updating, as well as the time and resources spent on both model retraining and dataset preparation. To address this limitation, Retrieval-Augmented Generation (RAG) was developed. RAG enhances LLMs by dynamically integrating external information retrieval mechanisms during inference, enabling the system to search for relevant, up-to-date, or domain-specific knowledge and inject it into a prompt with instructions for the LLM to generate a response based solely on the retrieved information.

To build a DSS based on LLMs and RAG, several LLM architectures can be employed, depending on the stage of the pipeline: Encoder-only – for semantic search of relevant documents, Decoder-only or Encoder-Decoder – for text generation and question answering.

3.2 RAG pipeline

As mentioned in the previous subsection of this article, RAG overcomes one of the main drawbacks of using LLMs to answer questions that require access to new data not included in their training. The RAG framework consists of three main components:

1. Retrieval: extracting relevant information from external data sources.
2. Augmentation: injecting the retrieved information into the LLM's input prompt.
3. Generation: producing a final response based on the augmented prompt.

Figure 6 illustrates a typical RAG inference pipeline, showcasing the framework's modular structure and the flow of information from user input to response generation [102, 103].

The main steps in the RAG inference pipeline are as follows:

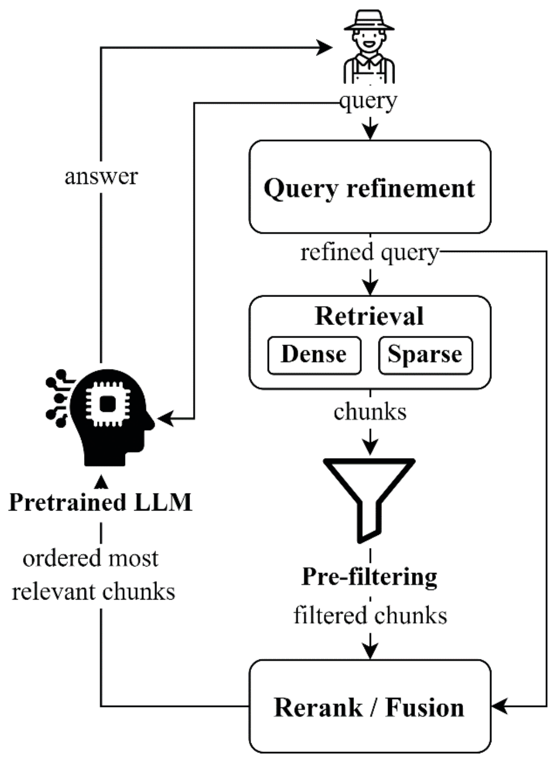


Figure 6. Rag inference pipeline

1. Receive a user request. The process begins with a user submitting a query. This query may represent a specific question or task requiring contextualised information.

2. Query refinement (optional). This procedure enhances the input query to improve the accuracy of information retrieval. It may involve correcting grammatical errors, expanding the query with related terms, or clarifying ambiguous phrases.

3. Information retrieval of the most relevant chunks from external sources to the query. Hybrid retrieval, combining vector search (dense retrieval) and term-based search (sparse retrieval) is a widely adopted approach.

4. Pre-filtering (optional). In this step, irrelevant data chunks are removed before further processing. Certain studies have shown that LLM-based pre-filtering can significantly improve the quality of information retrieval.

5. Rerank/Fusion (optional). Retrieved data is ranked to prioritise the most relevant information (chunks). The following methods are commonly used for ranking chunks by relevance:

   – Reciprocal Rank Fusion (RRF). It considers only the rank position of chunks in their respective search results and does not rely on the input query. It is fast, efficient, and easy to implement.

   – Cross-Encoder model. Computes a similarity score between the query and each chunk. While this method is more precise, it is also slower and more resource-intensive.

6. Augmentation and generation. The ranked data chunks are injected into the LLM's input prompt as context. The LLM then generates a final response based on both the user query and the retrieved information.

Search methods (the third step) can be broadly categorised into two types: dense and sparse retrieval. Sparse retrieval methods encode a query as a sparse vector, where non-zero values correspond to terms in a vocabulary. Examples of sparse retrieval include Bag-of-Words methods such as TF-IDF and BM25 [104]. These methods excel when exact term matches are present in documents. However, their limitation lies in the inability to capture semantic meaning. An alternative to traditional sparse retrieval methods like BM25 is the Sparse Lexical and Expansion Model (SPLADE) [105]. SPLADE is based on the BERT [106] architecture and expands the search query with related terms to address the vocabulary mismatch problem. The research has shown that SPLADE achieves superior metric scores compared to BM25 on the BEIR benchmark [105]. In contrast, dense retrieval methods make use of an Encoder-only architecture to encode text as dense vectors. The primary advantage of dense retrieval is the ability to capture semantic relationships. Even if two texts use different terms, their vectors will remain close in vector space if they are semantically similar. Common distance metrics employed in dense retrieval are cosine similarity, L2 distance, and dot product. Dense retrieval may not be optimal for cases where documents with exact term matches need to be found. As a result, dense and sparse retrievals are often used together, as they complement each other's weaknesses.

Query refinement (the second step) in RAG aims to enhance information retrieval by modifying/expanding the original input query. The existing query refinement approaches include the following:

– Query expansion. Expands an input query with related information and terms. Examples are the Pseudo Relevance Feedback (PRF) and Query2doc methods [107]. The latter uses LLMs to generate a pseudo-document /passage to the given query. A new query is formed by concatenating the input query with the generated pseudo-document. Studies have shown that this approach improves the performance of denser and sparse retrievals.

– Query rewrite [108]. Input queries are often suboptimal for searching, as they may contain spelling errors, slang, excessive length, or ambiguity (such as unclear domain areas or abbreviations). To improve search accuracy, the original query can be divided into several more specific subqueries for targeted information retrieval. LLMs can also be utilised in this process.

The fifth step of the RAG inference pipeline reranks the search results from different search methods. The Reciprocal Rank Fusion method ranks the documents according to the following equation [109]:

$$RRF(d) = \sum_{i=1}^{N} \frac{1}{k + r_i(d)} \tag{1}$$

where k is a constant (defaults to 60), N represents the number of search methods and $r_i(d)$ denotes the position of document d in the search results of the i-th search method.

After scoring each document, they are sorted in descending order by the assigned score value.

While intuitive and easy to implement, this method considers only the relative positions of documents within each search result.

A more accurate reranking method leverages the Cross-Encoder architecture, which takes two texts as input (a query and a chunk) and outputs a similarity score between them. Similarly, once the scores are calculated, the documents are sorted in descending order.

The final step of the RAG inference pipeline involves invoking an LLM with a prompt that includes the retrieved relevant chunks and instructions to answer a user request using only the information in those chunks. This significantly reduces the likelihood of hallucinations and the generation of irrelevant or outdated information [110].

Figure 7 below shows the extract, transform, load (ETL) pipeline for document processing:

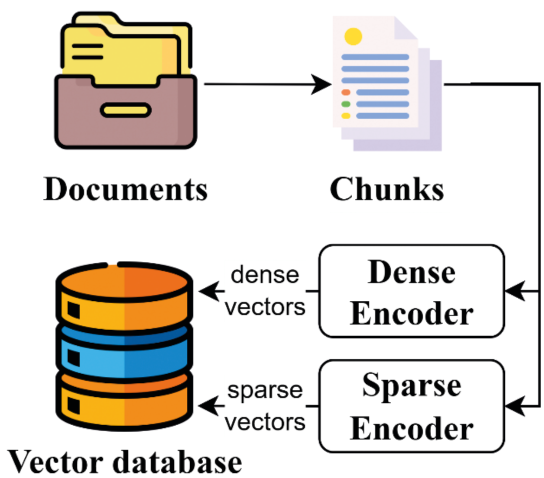


Figure 7. RAG ETL pipeline

Before the RAG system can query the uploaded documents, they must be indexed. This involves calculating dense and sparse vectors (depending on the chosen search methods) – and storing them in a vector database. Since LLMs have a limited context length and excessive contextual information can degrade the quality of responses, large documents should be divided into separate semantic blocks (chunks) prior to indexing. The choice of chunking strategy significantly affects the efficiency of RAG.

The existing chunking strategies include [102, 103]:

- Fixed-size chunking: splits a document into chunks of a fixed size (e.g., 256, 512, or 1024 characters). This approach may lead to truncation within sentences which results in a loss of contextual information and response quality degradation.

- Recursive character split: a more sophisticated approach that tries to preserve the integrity of paragraphs, sentences, and words as those are typically the strongest semantically related pieces of text. It splits text in the following order until the chunk size limit is reached: first by paragraph, then by sentence, and finally by word.

- Document-based chunking: considers the document structure and file formats (.txt, .docx, .pdf, .md, and others) during the chunking process. Tables within a document are transformed into formats readable by LLMs, such as HTML or Markdown. To enhance search efficiency, summaries generated from these tables are indexed instead of the tables themselves. If a table summary is recognised as relevant, the corresponding table content is injected into a prompt instead. Similarly, a multimodal LLM generates summaries for images, which are then used for searching.

In practice, implementing RAG systems on the server side typically involves deploying the following components [111]:

- Vector databases with sharding: enable scalable, low-latency retrieval in RAG systems.

- LLMs and embedding models: utilise third-party APIs, such as OpenAI or Azure OpenAI, or self-hosted solutions with load-balancing and GPU-accelerated infrastructure for low-latency inference.

- In-memory caching solutions: tools like Redis store frequently accessed vectors, reducing database load, lowering costs, and improving response times.

- RAG APIs: include streaming capabilities, context-aware endpoints, document processing, and robust error handling for production resilience.

In summary, Retrieval-Augmented Generation bridges the gap between pre-trained LLMs and real-world decision-making. By integrating query refinement, advanced retrieval methods, and effective reranking techniques, RAG ensures that LLM-generated responses are accurate, relevant, and up-to-date.

3.3 LLM for decision support systems in agriculture

Figure 8 below shows how an LLM can be integrated into an app to enhance user experience by executing app functions through text messages sent to the LLM.

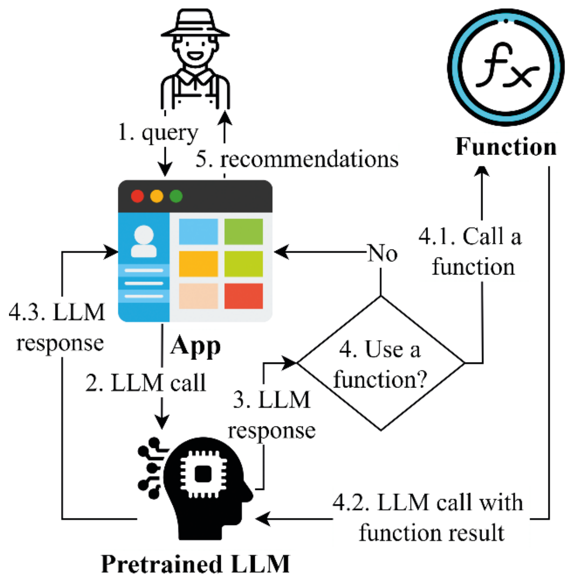


Figure 8. LLM pipeline with function calling

Upon receiving a user request, the system calls an LLM, which returns either a generated response or a description of a function call with its arguments. If a function call is returned, the system executes the corresponding function using the specified arguments and then returns the result for a second LLM call, generating the final response. This approach allows users to interact with the AI assistant to perform various system functions, greatly enhancing the user experience [112].

LLMs exhibit significant potential in DSSs due to their capabilities in text comprehension and generating human-like responses [113].

The proposed architecture of a DSS for agricultural crop cultivation based on the analysis of existing research is shown in Figure 9:

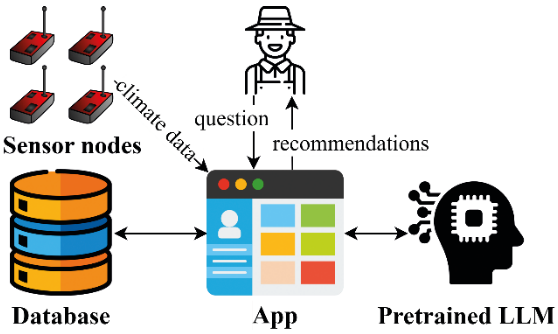


Figure 9. LLM-based DSS for agriculture

The DSS receives climate data from soil and climate sensors as input, which is processed and analysed before being stored in a database. The system relies on LLMs to generate recommendations on crop cultivation. Essential features of any DSS include transparency, explainability (the ability to clarify the reasoning behind decisions), and trustworthiness (the confidence in LLM-generated results to reduce the risk of false or harmful outcomes) [113].

Additionally, research studies on self-induced bias in recommendation algorithms highlight the risk of reinforcing past errors, potentially leading to cycles of flawed recommendations [114]. This is particularly critical in agricultural DSSs, where past inaccuracies or incorrect actions can perpetuate errors in future recommendations, negatively affecting crop production.

Given the critical role of LLMs in agricultural DSSs and the associated risks of self-induced bias, robust evaluation methodologies are essential to ensure these systems' reliability and impact. Assessing the performance of LLMs in agriculture requires tailored validation approaches to ensure their effectiveness and reliability in real-world scenarios. Traditional NLP metrics such as BLEU, ROUGE, or perplexity [115, 116] are insufficient to capture the domain-specific impact of these systems. Instead, agricultural applications demand more targeted metrics, including:

- Yield increase. Assessing the extent to which LLM recommendations, such as irrigation schedules or pest management strategies, improve overall yield [117, 118].
- Resource efficiency. Measuring reductions in water, fertilizer, or pesticide usage achieved through AI-driven decision support [119, 120].
- Decision accuracy. Comparing the accuracy of LLM outputs with those of human experts, particularly in tasks such as disease diagnosis [121] or soil fertility assessment [122].
- Economic impact: Quantifying cost savings or revenue increases attributable to AI-driven interventions in farm operations [123-125].

To enhance this evaluation, case studies provide valuable insights. For instance, a recent study demonstrated that integrating RAG into an LLM pipeline to generate high-quality, agriculture-specific questions improved answer similarity from 47% to 72% [126]. Another case study showed how LLMs augmented with RAG allowed users to significantly reduce the barriers to utilising a farm-to-fork traceability blockchain solution, which increased trust in the food supply chain and enhanced transparency [127].

Therefore, it has been established that LLMs can be integrated into information systems for automated decision support for agronomists. However, further research is required to enhance their trustworthiness and explainability and reduce bias.

3.4 Critical analysis and generalisation of the recent RAG research studies

The research study by G. Lu, S. Li, G. Mai, J. Sun, D. Zhu, L. Chai, and others on potential future applications areas of LLMs in agriculture has identified the following promising areas for their use [98]:

– Information extraction. This involves extracting structured data from unstructured agricultural documents. Structured data facilitates faster information querying, efficient filtration, and the construction of a knowledge base.

– Search and Q&A systems. LLMs can enhance search engines and generate responses to common agricultural questions due to their capabilities in understanding semantics and producing text that directly addresses the core of a question.

– Training data generation. Multimodal LLMs trained on agricultural data can synthesise images and videos from text descriptions, enhancing the training of computer vision (CV) models. This addresses the key limitation in applying CV models in agriculture: the scarcity of visual data for training.

– Human-robot interaction. LLMs can significantly simplify the interaction between humans and robots, enabling agronomists to control robots using natural language instead of typing out commands. The multimodality of LLMs will enhance their usefulness across various agricultural stages, including planting, weeding, fertilising, irrigation and harvesting.

Researchers S. Yang, Z. Yua, S. Li, R. Peng, K. Liu and P. Yang investigated the problem of pest identification by leveraging embedding-based retrieval and LLMs to extract structured information from unstructured agricultural documents. They proposed a four-stage LLM-based approach for extracting structured information in JSON format [97]:

– Stage 1: identify descriptive words in the retrieved text.

– Stage 2: convert descriptive words from stage 1 into attributes (e.g., white – colour).

– Stage 3: extract physical entities, particularly those with description, from text (e.g., larvae, adults, and others).

– Stage 4: match the entities and attributes and generate a JSON document with attribute types and values.

The performance of the GPT-3.5-turbo model under zero-shot conditions was manually assessed and is given in Table 2.

Table 2. LLM performance for each stage

	Stage 2	**Stage 3**	**Stage 4**
Precison	90%	89%	90%
Recall	76%	84%	89%

One of the problems the researchers encountered was the ambiguity of certain descriptive words when converting them into attributes using an LLM. For example, terms like “dark spots” could be identified as “colour” and “pattern,” causing the accuracy of the entire pipeline to fluctuate.

In terms of future research, the researchers emphasised the need to adjust the prompts and the pipeline structure for more precise information extraction to resolve the above issues. They also noted the importance of testing with more LLMs and datasets to determine the stability of this pipeline.

The research study by B. Silva, L. Nunes, R. Estevao, V. Aski and R. Chandra presented a comprehensive evaluation of popular LLMs, such as GPT-3.5, GPT-4, Llama-2-13B, and Llama2-70B, on their ability to pass agriculture-related exams [96].

To evaluate the LLMs, the agriculture exams and benchmark datasets were selected from three of the largest agriculture production countries: the USA, Brazil, and India.

The results demonstrated that the GPT-4 model consistently outperforms other models across different datasets and question types, achieving an impressive 93% accuracy on the US Certified Crop Adviser (CCA) exam, which is sufficient for renewing an agronomist certification. Furthermore, the inclusion of a preamble (contextual information such as the exam name and/or location), RAG, and ensemble refinement (ER) significantly improved the accuracy of LLM responses, with RAG demonstrating a notably greater enhancement in performance compared to the other techniques.

The primary contributions of the paper include the establishment of performance baselines for LLMs on agriculture-related problems, enabling researchers and practitioners to compare their results with the current state of LLM performance on these problems.

The researchers suggest that future research should focus on mitigating the risks associated with erroneous generations by exploring richer prompt strategies while also maximising the benefits of AI integration in agriculture.

The researchers R. Jagerman, H. Zhuang, Z. Qin, X. Wang, and M. Bendersky, as detailed in [128], proposed an LLM-based approach for expanding search queries with related terms to improve information retrieval results. Throughout the study, various sizes of Flan open-source models were utilised, including flan-t5-small (60M), flan-t5-base (220M), flan-t5-large (770M), flan-t5-xl (3B), flan-t5-xxl (11B), and flan-ul2 (20B), all fine-tuned to follow instructions.

The proposed LLM-based approach for search query expansion was compared to classical query expansion algorithms based on PRF, such as Bo1, Bo2, and KL. BM25 was employed as the search algorithm. The effectiveness of BM25 with these query expansion methods was evaluated on the MS-MARCO and BEIR datasets using metrics such as Recall@1K, MRR@10, and NDCG@10.

The findings indicate that the LLM-based query expansion method consistently outperformed the classical approaches across the evaluated datasets. Among the prompt engineering strategies, including zero-shot, few-shot, and Chain-of-Thoughts (CoT), the CoT prompts proved to be the most effective. The largest model, flan-ul2, achieved a Recall@1K score of 90.61 on the MS-MARCO dataset.

Regarding future improvements, the researchers plan to explore more LLM models for query expansion, refine prompt templates, and address production deployment, along with the costs involved.

The researchers W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua and Q. Li [102] comprehensively reviewed existing research studies in RAG, covering their architecture, training strategies, and applications. They identified the following research directions that can be explored in the future in the field of RAG [102]:

- Trustworthy RAG. As RAG can be maliciously or unintentionally manipulated into making unreliable decisions or harming humans, the ideal trustworthy RAG systems should possess the following characteristics: robustness, fairness, explainability and privacy. Robustness means a trustworthy RAG system should be able to withstand malicious changes caused by attackers. Fairness indicates that a system should avoid discrimination during a decision-making process. Explainability requires a complete understanding of the internal workings of RAG systems so that the predictions are explainable and transparent. Privacy involves protecting personal information within RAG systems and preventing the leakage of sensitive data.

- Multilingual RAG. The ability of RAG systems to process information sources in different languages. For example, users from countries where less common languages are spoken can leverage the rich English and Chinese corpora for knowledge retrieval, improving the performance of LLMs in subsequent tasks.

- Multimodal RAG. Multimodality enables RAG systems to query and process documents of various formats, such as images, videos, and audio. This benefits a wide range of domains, including healthcare, drug discovery, molecular analysis, and others.

- Quality of external knowledge. This involves employing methods to filter out low-quality or unreliable information, thereby enhancing the quality of the external knowledge base. Consequently, RAG systems can generate more accurate and reliable outputs, improving their effectiveness in real-world applications.

The researchers B. Nouriinanloo and M. Lamothe investigated the use of an LLM-based pre-filtering step to improve reranking in information retrieval [129].

A prompt using Chain-of-Thought (CoT) and Plan-and-Solve (PS) methods was designed. After the information retrieval, an LLM assigns a score between 0 and 1 to each chunk. Based on a chosen threshold, each chunk receives a value of 1 if the score meets or exceeds the threshold, and 0 if it falls below. Chunks valued at 0 are considered irrelevant and filtered out. The threshold is determined by the highest F1 score.

The effectiveness of the reranking results was evaluated on the benchmark datasets TREC and BEIR using the NDCG@10 metric with the Mixtral-8x7B-Instruct-v0.1 model. The research findings revealed that incorporating the pre-filtering step increased the nDCG@10 score by an average of 7.2% across all datasets compared to the baseline approach without it.

One limitation of the proposed approach noted by the researchers is the need to empirically determine a threshold value, as it may vary depending on the dataset.

R. Mohandoss, in his study [130], proposed an alternative solution to the GPTCache library for implementing a semantic cache for LLM systems that takes into account user contextual data, such as geolocation, department, employee role, identification number, and others. The proposed approach can work with both context-free queries and context-sensitive queries. An example of a query that does not require contextual data is “What is the capital of Canada?”, and one that does is “What is the capital of my country?”.

The proposed solution consisted of five core components:

1. Cognitive gateway (receives a user query, interacts with other system components and returns a user response).
2. Context extractor (receives a user query from cognitive gateway, decides which context attributes are needed for response generation, and passes the result back to cognitive gateway).
3. Context store (receives the context attributes from cognitive gateway, retrieves and returns the values of those attributes to Cognitive Gateway).
4. Cache manager (receives a user query from cognitive gateway, searches for cache matches by the defined threshold, and returns the result back to cognitive gateway; if no cache matches are found, it is called for the second time to cache an LLM response).
5. LLM (calls an LLM to generate a response for Cognitive Gateway in cases when no cache matches are found).

During cache evaluation, to empirically determine the optimal threshold value, the priority was to minimise the percentage of false cache hits. For a threshold value of 0.95, a 0% false cache hit rate was achieved, along with the highest percentage of true cache hits.

Using the proposed caching approach, the average time for response generation dropped from 3262 ms to 580 ms.

The researcher mentions storing a modified version of the original query with embedded context values in the cache as a possible extension to the proposed approach. For example, the cache could store “Good Thai restaurants in my hometown” along with “Good restaurants in Milwaukee”, where “Milwaukee” serves as the contextual value for the user’s “hometown”.

Based on the analysis of the components of the RAG pipeline and the research papers discussed above, the existing studies have been classified according to their focus on improving the performance of RAG systems, as shown in Table 3.

4 Discussion and future work

In this research article, dedicated to the identification of promising areas of research in

Table 3. Classification of the research papers by areas of research and improvements in RAG

Area of research	**Goal**	**Examples**	**Papers**
Pre-retrieval enhancement	Improve the RAG performance (recall) prior to information search	Query refinement: query rewrite, query expansion, sub-queries generation, and others.	[107, 108], [128]
		Indexing optimisation: better chunking strategy, metadata attachment, hierarchical index structure, table and image parsing, and others.	[103], [131, 132]
Post-retrieval enhancement	Improve the RAG precision after the information search to remove excess context for less noisy LLM generation	Pre-filtering chunks with LLM prior to reranking	[129]
		Reciprocal Rank Fusion (RRF)	[109]
		Rerankers (cross-encoder models)	[103], [133–135]
Search methods improvements	Improve the information search performance (precision and recall)	Dense and sparse encoder models (BERT, SPLADE and others). Retriever fine-tuning.	[20], [105, 106], [138]
LLM instruction tuning	Improve pretrained LLMs responses quality, reduce hallucinations.	Prompt refinement (richer prompt strategies like CoT, Few-shot, One-shot, and others). LLM fine-tuning.	[96], [110], [136], [138]
Efficiency of RAG	Decrease cost, reduce latency	Semantic caching, context-based caching	[130]
Multimodality	Add support or improve processing of multiple data types for a better user experience	Support for various file formats (images, videos and others), not just text.	[103], [98], [139, 140]
Domain-specific pipeline improvements	Improve the RAG pipeline performance for a specific domain (agriculture and others.)	Improving the pipeline for answering agriculture-related questions through prompt engineering, and others.	[96, 97], [137]
Decision-making	Improve the LLM and RAG capabilities in decision support systems.	Enhance LLM reasoning capabilities, trustworthiness, transparency and explainability with prompt engineering.	[112, 113], [141, 142]

the field of LLM and RAG for the agricultural sector, the current research on the improvement and application of LLM and RAG both in general and in the field of agricultural technologies has been analysed and systematised. A summary of the current state of research and achievements in RAG and LLM has made it possible to identify key areas for further research and improvement both in the general and agricultural context, namely:

1. Generating visual training data (images, videos) that have a similar distribution to the original data utilizing multimodal LLMs to improve the quality of CV models in agriculture. The input to such multimodal LLMs is a textual description of what is in the image, and the output is the image itself. The technical effect is the improved training results of CV models due to greater diversity in the training data. Steps to solve the problem include preparing an agricultural dataset, where the input values are image descriptions and the output values are the images themselves, preparing hardware (GPUs), and using Pytorch/Tensorflow frameworks to fine-tune the model on agricultural data.

2. Enhancing the RAG pipeline for agriculture. Technical effect is the improvement of the quality of RAG responses for agricultural documents. Steps to solve the problem include improving existing chunking strategies and indexing methods for agricultural documents, taking into acoount their internal structure and content; tailoring prompts to the agriculture domain to minimise hallucinations and improve the quality of LLM responses; expanding queries with agriculture-related terms to improve information retrieval; and others.

3. Developing and testing models of interaction between user applications for climate data processing and GenAI in machine-to-machine mode to promptly form an algorithm of actions for practising agronomists to increase the stress resistance of crops. The technical effect is the ability to execute the remote code of other system services through LLM interaction. Steps to implement: writing code to interact with external APIs via the HTTPS protocol and integration testing of the created code, creating prompts for calling functions for LLMs that supports function calling.

4. Improving the RAG and LLM for DSSs, namely: transparency and explainability, as well as trustworthiness. Technical effect is the generation of more reliable LLM responses, with detailed explanations of how conclusions and decisions are derived. Steps to solve the problem include the utilization of prompt engineering approaches such as CoT, prompt chaining, and others.

The expected social and economical effect, given that additional research is conducted on the research areas mentioned above, are as follows:

1. Earlier and more accurate detection of plant diseases, effective crop monitoring, which helps to reduce costs and increase yields.

2. More accurate and convenient search for information on agricultural data.

3. Simplified interaction with system components and functions in a dialogue mode via an LLM.

4. Improved decision support for agriculture due to thorough explanations of LLM responses that can be verified by humans.

As with any research, this article takes a particular perspective on the given question, focusing on certain aspects and omitting others. In particular, the aim was not to cover all possible aspects in a comprehensive manner. Given the considerable amount of literature on LLM and RAG, the focus is primarily on aspects related to the application of these technologies in agriculture. It was also decided to avoid a detailed review of all existing methods of explainable AI (XAI) and Generative AI (GenAI), focusing on aspects directly related to the use of LLMs for DSSs in agriculture.

Compared to recent publications [17, 98], this article clearly focuses on agriculture and the optimisation of information support. The authors of the review article [17] investigate the current status of LLMs for the transition to smart and unmanned agriculture. They emphasise the need for reliable and correct models to avoid risks for farmers and the environment. The article [98] focuses on artificial general intelligence (AGI) to enhance automation, robotics, precision agriculture, and other applications to increase productivity and sustainability in the agricultural sector. What distinguishes our work is the emphasis on indexing source documents and improving the components of the RAG pipeline, as this affects the ability of systems to provide accurate answers to questions on relevant documents, as well as researching applied aspects of LLM integration as functional components of DSSs in agriculture.

A graphical interpretation of the promising areas of LLM research in agriculture, ways of solving and/or further developing them, and the technical and socio-economic effects of their solution and/or further development are shown in Figure 10.

Thus, conducting research in the abovementioned promising areas will allow solving the topical scientific and applied task of increasing the efficiency, software and hardware compatibility of LLMs in implementing agromonitoring processes with automated decision support.

When applied to agriculture, where accuracy, specificity and contextual awareness are paramount, LLMs face several challenges despite their promise [143]. One critical limitation is the lack of domain-specific datasets for fine-tuning. Unlike general-purpose tasks, agricultural applications often require localised knowledge of crop varieties, soil conditions, pest behaviour and climate variability, which are rarely included in publicly available datasets [144, 145]. This results in outputs that may not be contextually appropriate or actionable for agronomists. For example, while an LLM may understand general concepts about "wheat diseases", it may not distinguish between regional variations of diseases or the specific pesticides approved for use in different jurisdictions [146, 147].

Another significant issue is the risk of hallucinations discussed in section 3.4, a phenomenon where LLMs generate plausible but factually incorrect responses [148-150]. In agriculture, such errors could lead to incorrect diagnoses of crop diseases, inappropriate fertiliser recommendations or sub-optimal irrigation schedules, potentially causing economic loss or environmental damage.

There are also ethical concerns, particularly around data privacy and security. Many agricultural systems collect sensitive data from farms, such as geographic locations, yield metrics, and proprietary practices. Integrating LLMs into such systems raises questions about data ownership, consent and protection from misuse [151-154].

Mitigation strategies include incorporating expert validation to cross-check model outputs, developing hybrid decision support systems that combine LLMs with traditional rule-based models, and fine-tuning LLMs on curated agricultural datasets. Collaboration between researchers, policymakers and practitioners is essential to address these challenges and create robust and reliable AI tools for agriculture [155-157].

The main scientific and practical effect of the research results of this article is a critical analysis and systematisation of the known methods and approaches to the design of automated decision support systems for agrotechnical purposes by substantiating the possibilities of using LLMs as functional elements of the interpretation of measurement information and generating recommendations for improving agrotechnical procedures for growing crops. This has allowed us to localise relevant research tasks and directions for their solution for the further development of applied intelligent decision support technologies based on the integration of generative artificial intelligence software tools.

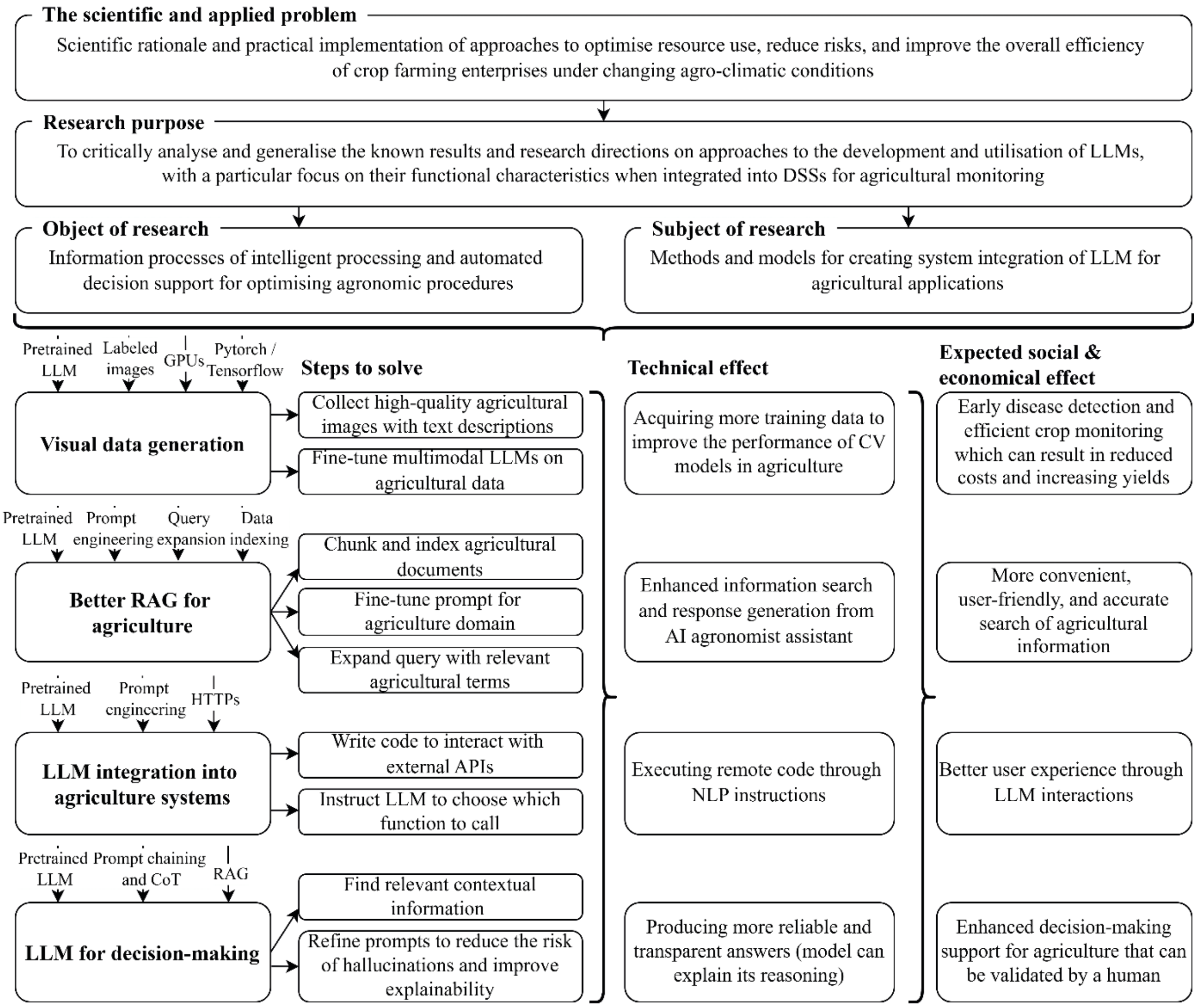


Figure 10. Graphical representation of the concept of solving the declared scientific and applied problems

5 Conclusions

In this article, an actual scientific and applied problem has been solved, which is devoted to the critical analysis and generalisation of known results and research directions regarding the approaches and functional characteristics of LLMs when implemented in DSSs for agrotechnical monitoring. The main results are as follows:

1. The pace of development and growth of agricultural products in the world and the existing promising areas of application of information technologies, particularly LLMs, to improve agricultural processes have been analysed. The existing world examples of successful implementation of LLMs in various industries and businesses have been considered, which allowed to prove the prospects and applied efficiency of LLMs in the construction of applied information technologies for automated decision support in substantiating the directions of optimisation of agrotechnical processes of growing crops.

2. A generalised analysis of LLMs has been carried out, including an analysis of the types of architectures, as well as their strengths and limitations. This allowed us to identify the areas of application of LLM architectures, as well as to focus on RAG as the main approach to solving one of the main limitations of LLMs, which is the limited knowledge base of training data.

3. An analysis of the utilisation of LLMs for DSSs has been carried out, which allowed to highlight trustworthiness, explainability and bias reduction as promising areas of research when integrating them into such systems for the agricultural sector.

4. A critical analysis and synthesis of the latest research on RAG has been carried out, which allowed the identification of promising areas of research to improve the work of LLMs and RAG in the agricultural sector, possible ways to solve them, as well as technical and socio-economic effects of implementation.

6 Acknowledgements

This research was carried out as part of the scientific project "Development of software and hardware of intelligent technologies for sustainable crop production in wartime and post-war" funded by the Ministry of Education and Science of Ukraine at the expense of the state budget (state registration number 0124U000289).

7 List of abbreviations

AGI	– Artificial general intelligence
BERT	– Bidirectional encoder representations from transformers
CoT	– Chain-of-thought
CV	– Computer vision
DSS	– Decision support system
ETL	– Extract, transform, and load
FAO	– Food and Agricultural Organization
GenAI	– Generative artificial intelligence
LLM	– Large language model
LSTM	– Long-short term memory
NLM	– Neural language models
NLP	– Natural language processing
PRF	– Pseudo Relevance Feedback
RAG	– Retrieval-augmented generation
RNN	– Recurrent neural network
SLM	– Statistical language models
SPLADE	– Sparse lexical and expansion model
XAI	– Explainable artificial intelligence

References

[1] FAOSTAT ANALYTICAL BRIEF 60: Agricultural production statistics 2000-2021. Available at: https://openknowledge.fao.org/server/api/core/ bitstreams/58971ed8-c831-4ee6-ab0a-e47ea66a7e6a/content. [Accessed 01 August 2024].

[2] I. Laktionov, G. Diachenko, V. Kashtan, A. Vizniuk, V. Gorev, K. Khabarlak, Y. Shedlovska, A comprehensive review of recent

approaches and Hardware-Software technologies for digitalisation and intellectualisation of Open-Field crop Production: Ukrainian case study in the global context, Computers and Electronics in Agriculture, 225, 2024, pp. 1−31. doi.org/10.1016/j.compag.2024.109326.

[3] Strategy of Agriculture and Rural development of Ukraine - 2030. Available at: https://www.agroberichtenbuitenland.nl/documenten/publicaties/2024/06/07/ua-strategy-agro-and-rural-development [Accessed 02 August 2024].

[4] Ministry of Agrarian Policy and Food of Ukraine: On Approval of the Concept of Stimulating the Development of Entrepreneurship in Rural Areas until 2030. Available at: https://minagro.gov.ua/npa/pro-shvalennya-koncepciyi-stimulyuvannya-rozvitku-pidpriyemnictva-na-silskih-teritoriyah-do-2030-roku [Accessed 02 August 2024] (in Ukrainian).

[5] On the Sustainable Development Strategy of Ukraine until 2030. Available at: https://ips.ligazakon.net/document/JH6YF00A?an=332 [Accessed 02 August 2024] (in Ukrainian).

[6] Ministry of Digital Transformation of Ukraine: Strategy for the Development of Innovation Activities of Ukraine until 2030. Available at: https://thedigital.gov.ua/regulations/strategiya-rozvitku-innovacijnoyi-diyalnosti-ukrayini-na-period-do-2030-roku%20 [Accessed 03 August 2024] (in Ukrainian).

[7] Cabinet of Ministers of Ukraine: Vectors of Economic Development 2030. Available at: https://nes2030.org.ua/docs/doc-vector.pdf [Accessed 03 August 2024] (in Ukrainian).

[8] C. Fetting, The European Green Deal, ESDN Report, Office, Vienna, December 2020. Available at: https://www.esdn.eu/fileadmin/ESDN_Reports/ESDN_Report_2_2020.pdf [Accessed 03 August 2024].

[9] Approved 28 CAP Strategic Plans (2023-2027), Summary overview for 27 Member States Facts and figures. Available at: https://agriculture.ec.europa.eu/document/download/7b3a0485-c335-4e1b-a53a-9fe3733ca48f_en?filename=approved-28-cap-strategic-plans-2023-27.pdf [Accessed 23 September 2024].

[10] EU Digital Strategy. EU4Digital. Available at: https://eufordigital.eu/discover-eu/eu-digital-strategy/ [Accessed 05 August 2024].

[11] European Commission: For a fair, healthy and environmentally-friendly food system Farm to Fork Strategy. Available at: https://food.ec.europa.eu/system/files/2020-05/f2f_action-plan_2020_strategy-info_en.pdf [Accessed 05 August 2024].

[12] FAO The role of innovation and digitalization in the sustainable use of natural resources to accelerate the implementation of climate-resilient and low-emission pathways in agrifood systems - ERC/24/2, 34th Session of the Regional Conference for Europe, Rome, Italy, 14–17 May 2024, URL: https://openknowledge.fao.org/server/api/core/bitstreams/019ae381-8546-4cce-b224-b04a761bd57e/content.

[13] Sustainable Development Goals: 17 Goals to Transform our World. United Nations. Available at: https://www.un.org/en/exhibits/page/sdgs-17-goals-transform-world [Accessed 07 August 2024].

[14] G20 Agriculture Ministers Declaration. Available at: https://www.g20.org/en/tracks/sherpa-track/agriculture [Accessed 07 August 2024].

[15] World Bank's Digital Agriculture Initiative. Available at: https://documents1. worldbank.org/curated/en/417641615957226621/pdf/Whats-Cooking-Digital-Transformation-of-the-Agrifood-System.pdf [Accessed 07 August 2024].

[16] S.K. Routray, A. Javali, K.P. Sharmila, M.K. Jha, M. Pappa, M. Singh, Large Language Models (LLMs): Hypes and Realities, 2023 International Conference on Computer Science and Emerging Technologies (CSET), Bangalore, India, 2023, pp. 1−6, doi.org/10.1109/CSET58993.2023.10346621.

[17] H. Zhu, S. Qin, M. Su, C. Lin, A. Li, J. Gao, Harnessing Large Vision and Language Models in Agriculture: A Review. arXiv preprint arXiv:2407.19679, 2024, pp. 1−54. doi.org/10.48550/arXiv.2407.19679.

[18] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, J. Ca, J. Kandola, T. Hofmann, T. Poggio, J. Shawe-Taylor, A Neural Probabilistic Language Model, Journal of Machine Learning Research, 3, 2003, pp. 1137–1155. URL: https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf.

[19] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Computation, 9 (8), 1997, pp. 1735–1780. doi.org/10.1162/neco.1997.9.8.1735.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, I. Polosukhin, Attention Is All You Need, arXiv preprint arXiv:1706.03762, 2017, pp. 1−15. doi.org/10.48550/arXiv.1706.03762.

[21] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805, 2018, pp. 1−16. doi.org/10.48550/arXiv.1810.04805.

[22] A.Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D.S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L.R. Lavaud, M.-A. Lachaux, P. Stock, T.L. Scao, T. Lavril, T. Wang, T. Lacroix, W.E. Sayed, Mistral 7B, arXiv preprint arXiv:2310.06825, 2023, pp. 1−9. doi.org/10.48550/arXiv.2310.06825.

[23] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, M.-H. Yang, Diffusion Models: A Comprehensive Survey of Methods and Applications, arXiv preprint arXiv:2209.00796, 2024, pp. 1−54. doi.org/10.48550/arXiv.2209.00796.

[24] M. Gupta, C. Akiri, K. Aryal, E. Parker, L. Praharaj, From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy, IEEE Access, 11, 2023, pp. 80218−80245. doi.org/10.1109/ACCESS.2023.3300381.

[25] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, A. Ahmed, Big Bird: Transformers for Longer Sequences, arXiv preprint arXiv:2007.14062, 2021, pp. 1−42. doi.org/10.48550/arXiv.2007.14062.

[26] S. Sayago, M. Ribera, Apple Siri (input) + Voice Over (output) = a de facto marriage, 9th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion, New York, NY, USA, 2021, pp. 6−10. doi.org/10.1145/3439231.3440603.

[27] M. Ford, W. Palmer, Alexa are you listening to me? An analysis of Alexa voice service network traffic, Pers Ubiquit Comput, 23, 2019, pp. 67–79. doi.org/10.1007/s00779-018-1174-x.

[28] M. Aggarwal, M. Madhukar, IBM's Watson Analytics for Health Care, Cloud Computing Systems and Applications in Healthcare, 2017, pp. 117–134. doi.org/10.4018/978-1-5225-1002-4.ch007.

[29] L. Schwartz-croft, Effects of ROSS Intelligence and NDAS, highlighting the need for AI regulation, SSRN Electronic Journal, 2024. doi.org/10.2139/ssrn.4727662.

[30] A. Caines, L. Benedetto, S. Taslimipoor, C. Davis, Y. Gao, Ø. Andersen, Z. Yuan, M. Elliott, R. Moore, C. Bryant, M. Rei, H. Yannakoudakis, A. Mullooly, D. Nicholls, P. Buttery, On the application of Large Language Models for language teaching and assessment technology, arXiv preprint arXiv:2307.08393v1, 2023, pp. 1−25. doi.org/10.48550/arXiv.2307.08393.

[31] J. Su, C. Jiang, X. Jin, Y. Qiao, T. Xiao, H. Ma, R. Wei, Z. Jing, J. Xu, J. Lin, Large Language Models for Forecasting and Anomaly Detection: A Systematic Literature Review, arXiv preprint arXiv:2402.10350v1, 2024, pp. 1−56. doi.org/10.48550/arXiv.2402.10350.

[32] B. Zhang, H. Yang, X.-Y. Liu, Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-Purpose Large Language Models, arXiv preprint arXiv:2306.12659, 2023, pp. 1−7. doi.org/10.48550/arXiv.2306.12659.

[33] J.O. Krugmann, J. Hartmann, Sentiment Analysis in the Age of Generative AI, Cust. Need. and Solut, 11 (3), 2024, pp. 1−19. doi.org/10.1007/s40547-024-00143-4.

[34] J. Fields, K. Chovanec and P. Madiraju, A Survey of Text Classification With Transformers: How Wide? How Large? How Long? How Accurate? How Expensive? How Safe?, IEEE Access, 12, 2024, pp. 6518−6531. doi.org/10.1109/ACCESS.2024.3349952.

[35] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging LLM-as-a-judge with MT-Bench and Chatbot Arena, arXiv preprint arXiv:2306.05685, 2023, pp. 1−29. doi.org/10.48550/arXiv.2306.05685.

[36] H. Tamoyan, H. Schuff, I. Gurevych, LLM Roleplay: Simulating Human-Chatbot Interaction, arXiv preprint arXiv:2407.03974, 2024, pp. 1−26. doi.org/10.48550/arXiv.2407.03974.

[37] S. Vakayil, D. S. Juliet, A. J and S. Vakayil, RAG-Based LLM Chatbot Using Llama-2, 2024 7th International Conference on Devices, Circuits and Systems

(ICDCS), Coimbatore, India, 2024, pp. 1−5. doi.org/10.1109/ICDCS59278.2024.10561020.

[38] K. S. John, G. A. Roy and B. P. S, LLM Based 3D Avatar Assistant, 2024 1st International Conference on Trends in Engineering Systems and Technologies (ICTEST), Kochi, India, 2024, pp. 1−5. doi.org/10.1109/ICTEST60614.2024.10576146.

[39] L. Ramaul, P. Ritala, M. Ruokonen, Creational and conversational AI affordances: How the new breed of chatbots is revolutionizing knowledge industries, Business Horizons, 67 (5), 2024, pp. 615−627. doi.org/10.1016/j.bushor.2024.05.006.

[40] D. Leiker, S. Finnigan, A. R. Gyllen, M. Cukurova, Prototyping the use of Large Language Models (LLMs) for adult learning content creation at scale, arXiv preprint arXiv:2306.01815, 2023, pp. 1−5. doi.org/10.48550/arXiv.2306.01815.

[41] R. Gallotta, G. Todd, M. Zammit, S. Earle, A. Liapis, J. Togelius, G. N. Yannakakis, Large Language Models and Games: A Survey and Roadmap, IEEE Transactions on Games, 2024, pp. 1−18. doi.org/10.1109/TG.2024.3461510.

[42] D. Barman, Z. Guo, O. Conlan, The Dark Side of Language Models: Exploring the Potential of LLMs in Multimedia Disinformation Generation and Dissemination, Machine Learning with Applications, 16, 2024, pp. 1−17. doi.org/10.1016/j.mlwa.2024.100545.

[43] O. D. Okey, E. U. Udo, R. L. Rosa, D. Z. Rodríguez, J. H. Kleinschmidt, Investigating ChatGPT and cybersecurity: A perspective on topic modeling and sentiment analysis, Computers & Security, 135, 2023. doi.org/10.1016/j.cose.2023.103476.

[44] A. Zaboli, S. L. Choi, T.-J. Song, J. Hong, ChatGPT and Other Large Language Models for Cybersecurity of Smart Grid Applications, arXiv preprint arXiv:2311.05462, 2024, pp. 1−5. doi.org/10.48550/arXiv.2311.05462.

[45] M. Guastalla, Y. Li, A. Hekmati, B. Krishnamachari, Application of Large Language Models to DDoS Attack Detection. In: Chen, Y., Lin, CW., Chen, B., Zhu, Q. (eds) Security and Privacy in Cyber-Physical Systems and Smart Vehicles. SmartSP 2023. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Springer, Cham, 552, 2024, pp. 83−99. doi.org/10.1007/978-3-031-51630-6_6.

[46] Y. Chen, M. Cui, D. Wang, Y. Cao, P. Yang, B. Jiang, Z. Lu, B. Liu, A survey of large language models for cyber threat detection, Computers & Security, 145, 2024, pp. 104016. doi.org/10.1016/j.cose.2024.104016.

[47] N. Capodieci, C. Sanchez-Adames, J. Harris and U. Tatar, The Impact of Generative AI and LLMs on the Cybersecurity Profession, 2024 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 2024, pp. 448−453. doi.org/10.1109/SIEDS61124.2024.10534674.

[48] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, S. Azam, A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges, IEEE Access, 12, 2024, pp. 26839−26874. doi.org/10.1109/ACCESS.2024.3365742.

[49] Z. Liu, Y. Tang, X. Luo, Y. Zhou, L.F. Zhang, No Need to Lift a Finger Anymore? Assessing the Quality of Code Generation by ChatGPT, IEEE Transactions on Software Engineering, 50 (6), 2024, pp. 1548−1584. doi.org/10.1109/TSE.2024.3392499.

[50] A. Onan, H. A. Alhumyani, DeepExtract: Semantic-driven extractive text summarization framework using LLMs and hierarchical positional encoding, Journal of King Saud University - Computer and Information Sciences, 36 (8), 2024, pp. 1−19. doi.org/10.1016/j.jksuci.2024.102178.

[51] P. Laban, W. Kryscinski, D. Agarwal, A. Fabbri, C. Xiong, S. Joty, C.-S. Wu, SummEdits: Measuring LLM Ability at Factual Reasoning Through The Lens of Summarization, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 9662−9676. doi.org/10.18653/v1/2023.emnlp-main.600.

[52] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, T. B. Hashimoto, Benchmarking Large Language Models for News Summarization, Transactions of the Association for Computational Linguistics, 12, 2024, pp. 39−57. doi.org/10.1162/tacl_a_00632.

[53] K. Pandya, M. Holia, Automating Customer Service using LangChain: Building custom open-source GPT Chatbot for organizations, arXiv preprint arXiv:2310.05421, 2023, 1-4. doi.org/10.48550/arXiv.2310.05421.

[54] Z. Xu, M. J. Cruz, M. Guevara, T. Wang, M. Deshpande, X. Wang, Z. Li, Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Washington DC, USA, 2024, pp. 2905−2909. doi.org/10.1145/3626772.3661370.

[55] J. J. Bird, A. Lotfi, Customer service chatbot enhancement with attention-based transfer learning, Knowledge-Based Systems, 301, 2024, pp. 1−12. doi.org/10.1016/j.knosys.2024.112293.

[56] A. Ishtiaq, K. Munir, A. Raza, N.A. Samee, M.M. Jamjoom and Z. Ullah, Product Helpfulness Detection With Novel Transformer Based BERT Embedding and Class Probability Features, IEEE Access, 12, 2024, pp. 55905−55917. doi.org/10.1109/ACCESS.2024.3390605.

[57] Y. Mehdi, Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web, The Official Microsoft Blog, Feb. 07, 2023. Available at: https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/ [Accessed 19 October 2024].

[58] Y. Li, S. Wang, H. Ding, H. Chen, Large Language Models in Finance: A Survey, in: Proceedings of the Fourth ACM International Conference on AI in Finance, ACM, Brooklyn, NY, USA, 2023, pp. 374−382. doi.org/10.1145/3604237.3626869.

[59] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, G. Mann, BloombergGPT: A Large Language Model for Finance, arXiv preprint arXiv:2303.17564, 2023, pp. 1−76. doi.org/10.48550/arXiv.2303.17564.

[60] Q. Xie, W. Han, X. Zhang, Y. Lai, M. Peng, A. Lopez-Lira, J. Huang, PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance, arXiv preprint arXiv:2306.05443, 2023, pp. 1−12. doi.org/10.48550/arXiv.2306.05443.

[61] M. S. Khan, H. Umer, ChatGPT in finance: Applications, challenges, and solutions, Heliyon, 10(2), 2024, pp. 1−8. doi.org/10.1016/j.heliyon.2024.e24890.

[62] How JPMorgan Chase's COIN is Revolutionizing Financial Operations with AI, Medium, Available at: https://medium.com/@the_AI_ZONE/how-jpmorgan-chases-coin-is-revolutionizing-financial-operations-with-ai-120a2938dab7 [Accessed 19 October 2024].

[63] A.J. Thirunavukarasu, D.S.J. Ting, K. Elangovan, L. Gutierrez, T.F. Tan, D.S.W. Ting, Large language models in medicine. Nat Med, 29, 2023, pp. 1930–1940. doi.org/10.1038/s41591-023-02448-8.

[64] M. Sallam, ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns, Healthcare, 11 (6), 2023, pp. 1−20. doi.org/10.3390/healthcare11060887.

[65] M. Cascella, J. Montomoli, V. Bellini et al., Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios, J. Med. Syst., 47 (33), 2023, pp. 1−5. doi.org/10.1007/s10916-023-01925-4.

[66] S. Pashangpour, G. Nejat, The Future of Intelligent Healthcare: A Systematic Analysis and Discussion on the Integration and Impact of Robots Using Large Language Models for Healthcare, Robotics, 13 (8), 2024, pp. 1−43. doi.org/10.3390/robotics13080112.

[67] C. Peng, X. Yang, A. Chen et al., A study of generative large language model for medical research and healthcare, npj Digit. Med., 6 (210), 2023, pp. 1−10. doi.org/10.1038/s41746-023-00958-w.

[68] X. Wu, B. Zhang, ChatGPT promotes healthcare: current applications and potential challenges, Int. J. Surg., 110 (1), 2024, pp. 606–608. doi.org/10.1097/JS9.0000000000000802.

[69] Med-PaLM. Available at: https://sites.research.google/med-palm [Accessed 02 August 2024].

[70] L. Martin, N. Whitehouse, S. Yiu, L. Catterson, R. Perera, Better Call GPT: Comparing Large Language Models Against Lawyers, arXiv preprint arXiv:2401.16212, 2024, p. 1-16. doi.org/10.48550/arXiv.2401.16212

[71] W. Han et al., Human-Centered and AI-Empowered Machine to Enhance Court Productivity and Legal Assistance, Information Sciences, 2024, pp. 121052−121052. doi.org/10.1016/j.ins.2024.121052

[72] P. Sarzaeim, Q. H. Mahmoud, A. Azim, A Framework for LLM-Assisted Smart Policing System, IEEE Access, 12, 2024, pp. 74915–74929. doi.org/10.1109/ACCESS.2024.3404862

[73] I.C. Peláez-Sánchez, D. Velarde-Camaqui, L.D. Glasserman-Morales, The Impact of Large Language Models on Higher Education: Exploring the Connection Between AI and Education 4.0, Front. Educ., 9, 2024, pp. 1–21. doi.org/10.3389/feduc.2024.1392091

[74] E. Waisberg, J. Ong, M. Masalkhi, A.G. Lee, Large Language Model (LLM)-Driven Chatbots for Neuro-Ophthalmic Medical Education, Eye, 38, 2024, pp. 639–641. doi.org/10.1038/s41433-023-02759-7

[75] J. Jeon, S. Lee, Large Language Models in Education: A Focus on the Complementary Relationship Between Human Teachers and ChatGPT, Educ Inf Technol, 28, 2023, pp. 15873–15892. doi.org/10.1007/s10639-023-11834-1

[76] M. Hosseini, C.A. Gao, D.M. Liebovitz, A.M. Carvalho, F.S. Ahmad, Y. Luo, N. MacDonald, K.L. Holmes, A. Kho, An Exploratory Survey About Using ChatGPT in Education, Healthcare, and Research, PLoS ONE, 18 (10), 2023, pp. 1–14. doi.org/10.1371/journal.pone.0292216.

[77] B. Alsafari, E. Atwell, A. Walker, M. Callaghan, Towards Effective Teaching Assistants: From Intent-Based Chatbots to LLM-Powered Teaching Assistants, Natural Language Processing Journal, 2024, pp. 100101–100101. doi.org/10.1016/j.nlp.2024.100101.

[78] Z. Epstein, A. Hertzmann, Art and the Science of Generative AI, Science, 380, 2023, pp.1110–1111. doi.org/10.1126/science.adh4451.

[79] J. Tsao, C. Nogues, Beyond the Author: Artificial Intelligence, Creative Writing and Intellectual Emancipation, Poetics, 102, 2024, pp. 1–12. doi.org/10.1016/j.poetic.2024.101865.

[80] S. Zhu, Z. Wang, Y. Zhuang, Y. Jiang, M. Guo, X. Zhang, Z. Gao, Exploring the Impact of ChatGPT on Art Creation and Collaboration: Benefits, Challenges and Ethical Implications, Telematics and Informatics Reports, 14, 2024, pp. 100138–100138. doi.org/10.1016/j.teler.2024.100138.

[81] C. Gan, Q. Zhang, T. Mori, Application of LLM Agents in Recruitment: A Novel Framework for Resume Screening, arXiv preprint arXiv:2401.08315, 2024, pp. 1–18. doi.org/10.48550/arXiv.2401.08315.

[82] R. J. Sunico, S. Pachchigar, V. Kumar, I. Shah, J. Wang, I. Song, Resume Building Application based on LLM (Large Language Model), 2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 2023, pp. 486–492. doi.org/10.1109/ICCCIS60361. 2023.10425602.

[83] G. Vagale, S. Y. Bhat, P. P. P. Dharishini, P. GK, ProspectCV: LLM-Based Advanced CV-JD Evaluation Platform, 2024 IEEE Students Conference on Engineering and Systems (SCES), Prayagraj, India, 2024, pp. 1–6. doi.org/10.1109/SCES61914.2024.10652548.

[84] S. Vijayakumar, F. Louis, Revolutionizing Staffing and Recruiting with Contextual Knowledge Graphs and QNLP: An End-to-End Quantum Training Paradigm, 2023 IEEE International Conference on Knowledge Graph (ICKG), Shanghai, China, 2023, pp. 45–51. doi.org/10.1109/ICKG59574.2023.00011.

[85] D. Gao, K. Chen, B. Chen, H. Dai, L. Jin, W. Jiang, W. Ning, S. Yu, Q. Xuan, X. Cai, L. Yang, Z. Wang, LLMs-based Machine Translation for E-commerce, Expert Systems with Applications, 258, 2024, pp. 125087–125087. doi.org/10.1016/j.eswa.2024.125087.

[86] K. I. Roumeliotis, N. D. Tselikas, D. K. Nasiopoulos, LLMs in E-commerce: A Comparative Analysis of GPT and LLaMA Models in Product Review Evaluation, Natural Language Processing Journal, 6, 2024, pp. 1–15. doi.org/10.1016/j.nlp.2024.100056.

[87] A. Mari, A. Mandelli, R. Algesheimer, Empathic Voice Assistants: Enhancing Consumer Responses in Voice Commerce, Journal of Business Research, 175, 2024, pp. 114566–114566. doi.org/10.1016/j.jbusres.2024.114566.

[88] A. Sharma et al., Automatic Data Transformation Using Large Language Model - An Experimental Study on Building Energy Data, 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 2023, pp. 1824–1834. doi.org/10.1109/BigData59044. 2023.10386931.

[89] S. Majumder, L. Dong, F. Doudi, Y. Cai, C. Tian, D. Kalathil, K. Ding, A. A. Thatte, N. Li, L. Xie, Exploring the Capabilities and Limitations of Large Language Models in the Electric Energy Sector, Joule, 8 (6), 2024, pp. 1544–1549. doi.org/10.1016/j.joule.2024.05.009.

[90] G. Jiang, Z. Ma, L. Zhang, J. Chen, EPlus-LLM: A Large Language Model-Based Computing Platform for Automated Building Energy Modeling, Applied Energy, 367, 2024, pp. 123431−123431. doi.org/10.1016/j.apenergy.2024.123431.

[91] J. Lu, X. Tian, C. Zhang, Y. Zhao, J. Zhang, W. Zhang, C. Feng, J. He, J. Wang, F. He, Evaluation of Large Language Models (LLMs) on the Mastery of Knowledge and Skills in the Heating, Ventilation and Air Conditioning (HVAC) Industry, Energy and Built Environment, 2024, pp. 1−18. doi.org/10.1016/j.enbenv.2024.03.010.

[92] N. Rane, A. Tawde, S. Choudhary, J. Rane, Contribution and Performance of ChatGPT and Other Large Language Models (LLM) for Scientific and Research Advancements: A Double-Edged Sword, International Research Journal of Modern Engineering and Technology, 5, 2023, pp. 875−899. doi.org/10.56726/IRJMETS45312.

[93] S. Jiang, D. Evans-Yamamoto, D. Bersenev, S. K. Palaniappan, A. Yachie-Kinoshita, ProtoCode: Leveraging Large Language Models (LLMs) for Automated Generation of Machine-Readable PCR Protocols from Scientific Publications, SLAS Technology, 29 (3), 2024, pp. 1−6. doi.org/10.1016/j.slast.2024.100134.

[94] T.A. Mohamed, M.H. Khafgy, A.B. Elsedawy, A.S. Ismail, A Proposed Model for Distinguishing Between Human-Based and ChatGPT Content in Scientific Articles, IEEE Access, 12, 2024, pp. 121251−121260. doi.org/10.1109/ACCESS.2024.3448315.

[95] B. Wang, X. Zhang, S. Li, Y. Wang, The Practice of Enhancing Learning and Scientific Innovative Abilities Using LLM-Based AI Tools, 2024 6th International Conference on Computer Science and Technologies in Education (CSTE), Xi'an, China, 2024, pp. 166−170. doi.org/10.1109/CSTE62025.2024.00038.

[96] B. Silva, L. Nunes, R. Estevão, V. Aski, R. Chandra, GPT-4 as an Agronomist Assistant? Answering Agriculture Exams Using Large Language Models, arXiv preprint arXiv:2310.06225, 2023, pp. 1−15. doi.org/10.48550/arXiv.2310.06225.

[97] R. Peng, K. Liu, P. Yang, Z. Yuan, S. Li, Embedding-Based Retrieval with LLM for Effective Agriculture Information Extracting from Unstructured Data, arXiv preprint arXiv:2308.03107, 2023, pp. 1−6. doi.org/10.48550/arXiv.2308.03107.

[98] G. Lu, S. Li, G. Mai, J. Sun, D. Zhu, L. Chai, H. Sun, X. Wang, H. Dai, N. Liu, R. Xu, D. Petti, C. Li, T. Liu, AGI for Agriculture, arXiv preprint arXiv:2304.06136, 2023, pp. 1−18. doi.org/10.48550/arXiv.2304.06136.

[99] X. Zhao, B. Chen, M. Ji, X. Wang, Y. Yan, J. Zhang, S. Liu, M. Ye, C. Lv, Implementation of Large Language Models and Agricultural Knowledge Graphs for Efficient Plant Disease Detection, Agriculture, 14 (8), 2024, pp. 1–24. doi.org/10.3390/agriculture14081359.

[100] P. Yu, B. Lin, A Framework for Agricultural Intelligent Analysis Based on a Visual Language Large Model, Applied Sciences, 14 (18), 2024, pp. 1–15. doi.org/10.3390/app14188350.

[101] M. Trzcinski, S. Łukasik, A.H. Gandomi, Optimizing the Structures of Transformer Neural Networks Using Parallel Simulated Annealing, JAISCR, 14 (3), 2024, pp. 267–282. doi.org/10.2478/jaiscr-2024-0015.

[102] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, Q. Li, A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models, arXiv preprint arXiv:2405.06211, 2024, pp. 1−18. doi.org/10.48550/arXiv.2405.06211.

[103] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retrieval-Augmented Generation for Large Language Models: A Survey, arXiv preprint arXiv:2312.10997, 2024, pp. 1−21. doi.org/10.48550/arXiv.2312.10997.

[104] X. Chen, S. Wiseman, BM25 Query Augmentation Learned End-to-End, arXiv preprint arXiv:2305.14087, 2023, pp. 1−6. doi.org/10.48550/arXiv.2305.14087.

[105] T. Formal, C. Lassance, B. Piwowarski, S. Clinchant, SPLADE v2: Sparse Lexical and Expansion Model for Information Retrieval, arXiv preprint arXiv:2109.10086, 2021, pp. 1−6. doi.org/10.48550/arXiv.2109.10086.

[106] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805, 2019, pp. 1−16. doi.org/10.48550/arXiv.1810.04805.

[107] L. Wang, N. Yang, F. Wei, Query2doc: Query Expansion with Large Language Models, arXiv preprint arXiv:2303.07678, 2023, pp. 1−10. doi.org/10.48550/arXiv.2303.07678.

[108] X. Ma, Y. Gong, P. He, H. Zhao, N. Duan, Query Rewriting for Retrieval-Augmented Large Language Models, arXiv preprint arXiv:2305.14283, 2023, pp. 1−13. doi.org/10.48550/arXiv.2305.14283.

[109] G. V. Cormack, C. L. A. Clarke, S. Buettcher, Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods, in Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, NY, USA, 2009, pp. 758–759. doi.org/10.1145/1571941.1572114.

[110] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, A. Chadha, A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications, arXiv preprint arXiv:2402.07927, 2024, pp. 1−9. doi.org/10.48550/arXiv.2402.07927.

[111] RAG in Production: Deployment Strategies and Practical Considerations. Available at: https://www.aporia.com/learn/rag-in-production/ [Accessed 18 November 2024].

[112] OpenAI Platform: Function calling. Available at: https://platform.openai.com/docs/guides/function-calling [Accessed 30 August 2024].

[113] E. Eigner, T. Händler, Determinants of LLM-assisted Decision-Making, arXiv preprint arXiv:2402.17385, 2024, pp. 1−44. doi.org/10.48550/arXiv.2402.17385.

[114] J. Pawłowska, K. Rydzewska, A. Wierzbicki, Using Cognitive Models to Understand and Counteract the Effect of Self-Induced Bias on Recommendation Algorithms, JAISCR, 13 (2), 2023, pp. 73−94. doi.org/10.2478/jaiscr-2023-0008.

[115] H. Chatoui, O. Ata, Automated Evaluation of the Virtual Assistant in BLEU and ROUGE Scores, Proceedings of the 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 2021, pp. 1–6. doi.org/10.1109/HORA52670.2021.9461351.

[116] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, Z. Liu, Evaluation of Retrieval-Augmented Generation: A Survey, arXiv preprint arXiv:2405.07437, 2024, pp. 1−21. doi.org/10.48550/arXiv.2405.07437.

[117] E. Elbasi, N. Mostafa, C. Zaki, Z. AlArnaout, A. E. Topcu, L. Saker, Optimizing Agricultural Data Analysis Techniques through AI-Powered Decision-Making Processes, Applied Sciences, 14 (17), 2024, pp. 1−26. doi.org/10.3390/app14178018.

[118] M. Mirali, Generative AI and Agriculture: A New Era of Farming Efficiency, Grain Data Solutions Inc., 2024. Available at: https://graindatasolutions.com/generative-ai-agriculture-farming-efficiency/ [Accessed 23 November 2024].

[119] M. A. Hamed, M. F. El-Habib, R. Z. Sababa, M. M. Al-Hanjor, B. S. Abunasser, S. S. Abu-Naser, Artificial Intelligence in Agriculture: Enhancing Productivity and Sustainability, International Journal of Engineering and Information Systems (IJEAIS), 8 (8), 2024, pp. 1–8. URL: https://philarchive.org/archive/HAMAII-2.

[120] D. R. Kale, J. Nalvade, P. S. Randive, S. Hirve, Artificial Intelligence in Sustainable Agriculture: Enhancing Efficiency and Reducing Environmental Impact, Industrial Engineering Journal, 53 (9), 2024, pp. 103−109. URL: https://www.researchgate.net/publication/382949284_Artificial_Intelligence_In_Sustainable_Agriculture_Enhancing_Efficiency_and_Reducing_Environmental_Impact

[121] D. Tirkey, K. K. Singh, S. Tripathi, Performance analysis of AI-based solutions for crop disease identification, detection, and classification, Smart Agricultural Technology, 5, 2023, pp. 1−13. doi.org/10.1016/j.atech.2023.100238.

[122] A. Sarangi, S. K. Raula, S. Ghoshal, S. Kumar, C. S. Kumar, N. Padhy, Enhancing Process Control in Agriculture: Leveraging Machine Learning for Soil Fertility Assessment, Engineering Proceedings, 67 (1), 2024, pp. 1−11. doi.org/10.3390/engproc2024067031.

[123] W. Geng, L. Liu, J. Zhao, X. Kang, W. Wang, Digital Technologies Adoption and Economic Benefits in Agriculture: A Mixed-Methods Approach, Sustainability, 16 (11), 2024, pp. 1−24. doi.org/10.3390/su16114431.

[124] V. Varriale, A. Cammarano, F. Michelino, M. Caputo, Critical analysis of the impact of artificial intelligence integration with cutting-edge technologies for production systems, Journal of Intelligent Manufacturing, 2023, pp. 1−33. doi.org/10.1007/s10845-023-02244-8.

[125] Y. Qin, Z. Xu, X. Wang, M. Skare, Artificial Intelligence and Economic Development: An Evolutionary Investigation and

Systematic Review, Journal of the Knowledge Economy, 15 (1), 2024, pp. 1736–1770. doi.org/10.1007/s13132-023-01183-2.

[126] A. Balaguer, V. Benara, R. L. de Freitas Cunha, R. de M. Estevão Filho, T. Hendry, D. Holstein, J. Marsman, N. Mecklenburg, S. Malvar, L. O. Nunes, R. Padilha, M. Sharp, B. Silva, S. Sharma, V. Aski, R. Chandra, RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture, arXiv preprint arXiv:2401.08406, 2024, pp. 1−33. doi.org/10.48550/arXiv.2401.08406.

[127] J. Benzinho, J. Ferreira, J. Batista, L. Pereira, M. Maximiano, V. Távora, R. Gomes, O. Remédios, LLM Based Chatbot for Farm-to-Fork Blockchain Traceability Platform, Applied Sciences, 14 (19), 2024, pp. 1−15. doi.org/10.3390/app14198856.

[128] R. Jagerman, H. Zhuang, Z. Qin, X. Wang, M. Bendersky, Query Expansion by Prompting Large Language Models, arXiv preprint arXiv:2305.03653, 2023, pp. 1−7. doi.org/10.48550/arXiv.2305.03653.

[129] B. Nouriinanloo, M. Lamothe, Re-Ranking Step by Step: Investigating Pre-Filtering for Re-Ranking with Large Language Models, arXiv preprint arXiv:2406.18740, 2024, pp. 1−10. doi.org/10.48550/arXiv.2406.18740.

[130] R. Mohandoss, Context-based Semantic Caching for LLM Applications, 2024 IEEE Conference on Artificial Intelligence (CAI), Singapore, Singapore, 2024, pp. 371−376. doi.org/10.1109/CAI59869.2024.00075.

[131] Recursively split by character. LangChain. Available at: https://python.langchain.com/v0.1/docs/modules/data_connection/document_transformers/recursive_text_splitter/ [Accessed 31 August 2024].

[132] Y. Wang, N. Lipka, R.A. Rossi, A. Siu, R. Zhang, T. Derr, Knowledge Graph Prompting for Multi-Document Question Answering, arXiv preprint arXiv:2308.11730, 2023, pp. 1−22. doi.org/10.48550/arXiv.2308.11730.

[133] Cohere: Rerank Overview. Available at: https://docs.cohere.com/docs/overview [Accessed 30 August 2024].

[134] Voyage AI: Rerankers. Available at: https://docs.voyageai.com/docs/reranker [Accessed 30 August 2024].

[135] W. Sun, L. Yan, X. Ma, S. Wang, P. Ren, Z. Chen, D. Yin, Z. Ren, Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents, arXiv preprint arXiv:2304.09542, 2023, pp. 1−20. doi.org/10.48550/arXiv.2304.09542.

[136] J.-j. Park, S.-j. Choi, LLMs for Enhanced Agricultural Meteorological Recommendations, arXiv preprint arXiv:2408.04640, 2024, pp. 1−10. doi.org/10.48550/arXiv.2408.04640.

[137] T. Wang, N. Wang, Y. Cui, J. Liu. Agricultural Technology Knowledge Intelligent Question-Answering System Based on Large Language Model, Smart Agriculture, 5 (4), 2023, pp. 105−116. doi.org/10.12133/j.smartag.SA202311005.

[138] X. V. Lin, X. Chen, M. Chen, W. Shi, M. Lomeli, R. James, P. Rodriguez, J. Kahn, G. Szilvasy, M. Lewis, L. Zettlemoyer, S. Yih, RA-DIT: Retrieval-Augmented Dual Instruction Tuning, arXiv preprint arXiv:2310.01352, 2024, pp. 1−25. doi.org/10.48550/arXiv.2310.01352

[139] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, C. Li, Y. Xu, H. Chen, J. Tian, Q. Qian, J. Zhang, F. Huang, J. Zhou, mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality, arXiv preprint arXiv:2304.14178, 2024, pp. 1−21. doi.org/10.48550/arXiv.2304.14178.

[140] P. Qi, Movie Visual and Speech Analysis Through Multi-Modal LLM for Recommendation Systems, IEEE Access, 12, 2024, pp. 145686−145702. doi.org/10.1109/ACCESS.2024.3471568.

[141] L. Chen, L. Wang, H. Dong, Y. Du, J. Yan, F. Yang, S. Li, P. Zhao, S. Qin, S. Rajmohan, Q. Lin, D. Zhang, Introspective Tips: Large Language Model for In-Context Decision Making, arXiv preprint arXiv:2305.11598, 2023, pp. 1−22. doi.org/10.48550/arXiv.2305.11598.

[142] M. Chen, Z. Tao, W. Tang, T. Qin, R. Yang, C. Zhu, Enhancing emergency decision-making with knowledge graphs and large language models, International Journal of Disaster Risk Reduction, 113, 2024, pp. 104804. doi.org/10.1016/j.ijdrr.2024.104804.

[143] D. De Clercq, E. Nehring, H. Mayne, A. Mahdi, Large language models can help boost food production, but be mindful of their risks, Frontiers in Artificial Intelligence, 7, 2024, pp. 1−11. doi.org/10.3389/frai.2024.1326153.

[144] K. Gikunda, Harnessing Artificial Intelligence for Sustainable Agricultural Development in

Africa: Opportunities, Challenges, and Impact, arXiv preprint arXiv:2401.06171, 2024, pp. 1−8. doi.org/10.48550/arXiv.2401.06171.

[145] M. Gardezi, B. Joshi, D. M. Rizzo, M. Ryan, E. Prutzer, S. Brugler, A. Dadkhah, Artificial Intelligence in Farming: Challenges and Opportunities for Building Trust, Agronomy Journal, 116 (3), 2024, pp. 1217−1228. doi.org/10.1002/agj2.21353.

[146] S. Kumar S, A. K. M. Ajmal Khan, I. A. Banday, M. Gada, V. V. Shanbhag, Overcoming LLM Challenges Using RAG-Driven Precision in Coffee Leaf Disease Remediation, arXiv preprint arXiv:2405.01310, 2024, pp. 1−6. doi.org/10.48550/arXiv.2405.01310.

[147] J. Li, M. Xu, L. Xiang, D. Chen, W. Zhuang, X. Yin, Z. Li, Large Language Models and Foundation Models in Smart Agriculture: Basics, Opportunities, and Challenges, arXiv preprint arXiv:2308.06668, 2024, pp. 1−18. doi.org/10.48550/arXiv.2308.06668.

[148] A. Mishra, A. Asai, V. Balachandran, Y. Wang, G. Neubig, Y. Tsvetkov, H. Hajishirzi, Fine-grained Hallucination Detection and Editing for Language Models, arXiv preprint arXiv:2401.06855, 2024, pp. 1−23. doi.org/10.48550/arXiv.2401.06855.

[149] G. Perković, A. Drobnjak, I. Botički, Hallucinations in LLMs: Understanding and Addressing Challenges, in Proceedings of the 2024 47th MIPRO ICT and Electronics Convention, 2024, pp. 2084−2088. doi.org/10.1109/MIPRO60963.2024.10569238.

[150] W. de Almeida da Silva, L. C. Costa Fonseca, S. Labidi, J. C. Lima Pacheco, Mitigation of Hallucinations in Language Models in Education: A New Approach of Comparative and Cross-Verification, in Proceedings of the 2024 IEEE International Conference on Advanced Learning Technologies (ICALT), 2024, pp. 207−209. doi.org/10.1109/ICALT61570.2024.00066.

[151] R. Mark, Ethics of Using AI and Big Data in Agriculture: The Case of a Large Agriculture Multinational, The ORBIT Journal, 2 (2), 2019, pp. 1−27. doi.org/10.29297/orbit.v2i2.109.

[152] P. B. Falola, A. E. Adeniyi, O. A. Madamidola, J. B. Awotunde, O. A. Olukiran, S. O. Akinola, Artificial Intelligence in Agriculture: The Potential for Efficiency and Sustainability, With Ethical Considerations. In H. Kannan, R. Rodriguez, Z. Paprika, & A. Ade-Ibijola (Eds.), Exploring Ethical Dimensions of Environmental Sustainability and Use of AI, IGI Global Scientific Publishing, 2024, pp. 307−329. doi.org/10.4018/979-8-3693-0892-9.ch015.

[153] P. Karkhile, V. Kavade, P. Bahalkar, Use of Ethical AI in Agriculture, International Journal for Multidisciplinary Research (IJFMR), 6 (3), 2024, pp. 1−12. URL: https://www.ijfmr.com/papers/2024/3/20356.pdf.

[154] M. Uddin, A. Chowdhury, M. A. Kabir, Legal and ethical aspects of deploying artificial intelligence in climate-smart agriculture, AI & Society, 39 (1), 2024, pp. 221−234. doi:10.1007/s00146-022-01421-2.

[155] B. Kisliuk, J. C. Krause, H. Meemken, J. C. Saborío Morales, H. Müller, J. Hertzberg, AI in Current and Future Agriculture: An Introductory Overview, KI - Künstliche Intelligenz, 37 (2), 2023, pp. 117−132. doi:10.1007/s13218-023-00826-5.

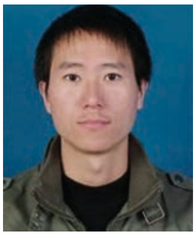
[156] F. Assimakopoulos, C. Vassilakis, D. Margaris, K. Kotis, D. Spiliotopoulos, Artificial Intelligence Tools for the Agriculture Value Chain: Status and Prospects, Electronics, 13 (22), 2024, pp. 1−36. doi:10.3390/electronics13224362.

[157] O. B. Akintuyi, AI in Agriculture: A Comparative Review of Developments in the USA and Africa, Open Access Research Journal of Science and Technology, 10 (2), 2024, pp. 60–70. doi.org/10.53022/oarjst.2024.10.2.0051.

Artem Vizniuk is a deep learning engineer at CLOUD FLOW LLC. He received his master's degree in software engineering from Dnipro University of Technology in 2023. His research interests include machine learning, NLP and large language models.
https://orcid.org/0000-0002-0594-2633

Grygorii Diachenko is an Associate Professor in the Department of Electric Drive at Dnipro University of Technology (Dnipro, Ukraine). He received his Ph.D. degree in 2021 majoring in electrotechnical complexes and systems. His research interests include machine learning, mechatronics, control theory and IoT. He is the author of more than 30 scientific works, including 11 publications Scopus and WoS.
https://orcid.org/0000-0001-9105-1951

Ivan Laktionov is a Professor in the Department of Computer Systems Software at Dnipro University of Technology (Dnipro, Ukraine). He received his Doctor of Science degree in 2021 majoring in computer systems and components. He specialises in advanced computer, sensor, infocommunication and microprocessor technologies of physical and chemical parameters monitoring and control systems. He is the author of more than 90 scientific works, including more than 30 articles in the world's authoritative publications Scopus and WoS.
https://orcid.org/0000-0001-7857-6382



Min Xiao (Member, IEEE) received the B.S. degree in mathematics and the M.S. degree in fundamental mathematics from Nanjing Normal University, Nanjing, China, in 1998 and 2001, respectively, and the Ph.D. degree in applied mathematics from Southeast University, Nanjing, in 2007., He was a Postdoctoral Researcher or a Visiting Researcher with Southeast University; The City University of Hong Kong, Hong Kong; and Western Sydney University (Sydney Campus), Sydney, NSW, Australia. He is currently a Professor with the College of Automation and the College of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing. His current research interests include information security, anomalous diffusion systems, networked control systems, tipping and control, and cyber–physical systems.
https://orcid.org/0000-0002-8992-153X

Agnieszka Siwocha received M.Sc. the degree from Lodz University of Technology, Faculty of Technical Physics, Computer Science and Applied Mathematics, and her a Ph.D. in 2015 in the field of computer science, computer graphics at the University of Social Sciences, Łódź, Poland. She is currently an Assistant Professor at the Social Academy of Sciences in Łódź. She has a title of Adobe Certified Expert, and provides training on Adobe software and computer graphics. Author of over 20 publications related to various problems of computer science, computer graphics and IT applications. Her present research interests include fractal coding, compression, and the quality of digital images, computer graphics, machine learning, and multifractal analysis.
https://orcid.org/0000-0003-2562-236X

Jacek Smoląg received the M.Sc. degree in electrical engineering from Częstochowa University of Technology in 1991 and Ph.D. degree in electronics from Lodz University of Technology in 1998, Lodz, Poland. He is an Assistant Professor in the Institute of Computational Intelligence at Częstochowa University of Technology, Częstochowa, Poland. He is engaged in research in parallel processing focusing on mapping algorithms into parallel computers, parallel systolic processing and parallel architectures for artificial intelligence. His current research interests include design and implementation of neural network learning algorithms in hardware. He is a member of the Polish Neural Network Society.
https://orcid.org/0000-0002-1326-3374