UDC 004.738.5:81’322
DOI https://doi.org/10.32782/tps2663-4880/2026.45.2.40

# MODERN INFORMATION RETRIEVAL MODELS AT THE INTERSECTION OF LINGUISTICS AND ARTIFICIAL INTELLIGENCE

# СУЧАСНІ МОДЕЛІ ІНФОРМАЦІЙНОГО ПОШУКУ НА ПЕРЕТИНІ ЛІНГВІСТИКИ ТА ШТУЧНОГО ІНТЕЛЕКТУ

**Heydarova M.I.,**
*orcid.org/0000-0003-4892-9380*
*PhD student in Philology*
*Linguistics Institute named after I.Nasimi of Azerbaijan National Academt of Sciences*

In an era of rapidly growing data in the digital environment, there is an increasing need to optimize information retrieval and query systems. Traditional keyword-based search methods rely on lexical matching between queries and documents, which gives rise to linguistic problems such as polysemy, synonymy, morphological variation, and context dependence. These issues reduce the effectiveness of search results.

This article examines information retrieval and query systems within the context of artificial intelligence challenges, analyzes recent research, and highlights the advantages of semantic search methods. It is noted that the application of machine learning, deep learning, and transformer-based large language models (such as BERT, GPT, RoBERTa, etc.) enables search systems to achieve more effective results in terms of query understanding. Key components of modern search systems–such as dialogue-based search, query reformulation, clarification question generation, and answer generation mechanisms–are also discussed.

The aim of this article is to analyze the impact of the development of artificial intelligence models on semantic search systems, to identify the technologies applied in this field, and to evaluate the effectiveness of the obtained results. The article also examines semantic search approaches based on knowledge graphs, ontologies, neural networks, as well as vector-based and hybrid models. The advantages and limitations of these approaches are analyzed comparatively. In addition, it is emphasized that hybrid approaches, which combine lexical, semantic, and knowledge-based methods, have the capacity to generate more accurate, effective, and context-aware search results. In conclusion, the integration of artificial intelligence technologies into search systems is regarded as a key direction for improving the accuracy and efficiency of information retrieval, and scientific foundations for future research in this area are outlined.

**Key words:** information retrieval, artificial intelligence, semantic search, dialogue search systems, large language models.

У добу стрімкого зростання обсягів даних у цифровому середовищі дедалі більше зростає потреба в оптимізації систем інформаційного пошуку та запитів. Традиційні методи пошуку, засновані на ключових словах, покладаються на лексичну відповідність між запитами та документами, що породжує лінгвістичні проблеми, такі як багатозначність, синонімія, морфологічна варіативність і залежність від контексту. Ці проблеми знижують ефективність результатів пошуку.

У статті розглядаються системи інформаційного пошуку та запитів у контексті викликів штучного інтелекту, аналізуються новітні дослідження та висвітлюються переваги семантичних методів пошуку. Зазначається, що застосування машинного навчання, глибокого навчання та великих мовних моделей на основі трансформерів (таких як BERT, GPT, RoBERTa тощо) дозволяє системам пошуку досягати більш ефективних результатів у розумінні запитів. Також розглядаються ключові компоненти сучасних систем пошуку – діалоговий пошук, реформулювання запитів, генерація уточнювальних питань і механізми формування відповідей.

Метою статті є аналіз впливу розвитку моделей штучного інтелекту на семантичні системи пошуку, визначення технологій, що застосовуються в цій сфері, та оцінка ефективності отриманих результатів. У статті також досліджуються підходи до семантичного пошуку, засновані на графах знань, онтологіях, нейронних мережах, а також векторних і гібридних моделях. Порівняльно аналізуються переваги та обмеження цих підходів. Крім того, підкреслюється, що гібридні підходи, які поєднують лексичні, семантичні та знаннєві методи, мають потенціал для формування більш точних, ефективних і контекстуально обґрунтованих результатів пошуку. У висновках інтеграція технологій штучного інтелекту в системи пошуку розглядається як ключовий напрямок підвищення точності та ефективності інформаційного пошуку, а також окреслюються наукові засади для подальших досліджень у цій сфері.

**Ключові слова:** інформаційний пошук, штучний інтелект, семантичний пошук, діалогові системи пошуку, великі мовні моделі.

**Introduction.** In the modern era of rapidly developing information technologies, information retrieval systems, question–answering systems, chatbots, digital libraries, and similar technologies have become an integral part of everyday life by enabling the retrieval of structured or unstructured data from large-scale information collections. In this process, the main responsibility lies with information retrieval systems, as users obtain the required information by submitting queries to these systems.

Traditional search systems, namely keyword-based search, mainly rely on term matching between queries

and documents. Such systems face a number of problems, including polysemy, synonymy, fixed expressions, and morphological variations between queries and documents [9, p. 203].

In recent years, significant advances in information technologies and the availability of large-scale databases have played an important role in optimizing information retrieval systems. In particular, the rapid development of artificial intelligence (AI) has had a profound impact on information retrieval systems, fundamentally changing methods of accessing, processing, and using large-scale data resources. By enabling the use of deep learning methods, AI has also exerted a strong influence on the field of natural language processing (NLP – Natural Language Processing). As a rule, in traditional information retrieval based on keyword matching, although lexical correspondence between query terms and documents is achieved, the semantic depth of the query is not taken into account. Therefore, in recent years, the development of semantic search systems–systems capable of understanding user intent and the content of queries–has come to the forefront. Thanks to advances in artificial intelligence technologies, the joint application of semantic search, natural language processing, machine learning (ML), and deep learning (DL) has led to more effective results. In particular, transformer-based large language models (LLMs), such as BERT, GPT, RoBERTa, and others, retrieve and present information that is semantically related to the user's query based on the contextual meaning of the text.

LLMs are artificial intelligence systems that understand, predict, and generate human language, trained on large-scale text corpora using deep learning algorithms. They are developed on the basis of massive datasets containing billions of words from diverse sources such as books, articles, blogs, tweets, and others. Through artificial intelligence algorithms, these models learn statistical patterns in data, including word frequency, order, agreement, and meaning.

LLMs are widely used in text generation, machine translation, summarization, question–answering systems, and the enhancement of search systems. Most modern LLMs are based on transformer architectures. One of the key features of the transformer architecture is the self-attention mechanism, which allows the model to assess the importance of different words within a sentence, thereby effectively capturing contextual nuances and semantic meanings. Pre-trained transformer models, trained on large-scale text corpora, deeply internalize the semantics and syntax of language, enabling high performance across a wide range of language tasks [13, p. 1890].

**Literature review.** From studies conducted in the field of optimizing and improving the efficiency of information retrieval systems, several key points can be identified. In her work, Y. V. Rogushina examines the modeling of an intelligent interaction system between information resources and information consumers through the use of external and internal knowledge bases. She also proposes an ontological model that formally describes the interaction between users and web-based information resources during semantic search, outlining its elements and the relationships between them, and suggests sources and methods for extending this model [19, p. 336].

In their work, K.A.Hambarde and H.Proença discuss state-of-the-art models that cover a wide range of methods and approaches in the field of information retrieval. The authors examine approaches ranging from term-based methods to semantic search techniques and neural models, as well as the relationships between them [8]. Bola Abimbola emphasizes that search systems based on ontological concepts improve the quality of search results [1]. M.Rovatsos and R.Filgueira present a methodology for developing lightweight semantic search systems that employ modern technologies to support users in exploring large volumes of data. The authors provide an extensive study to demonstrate the benefits of this methodology [20, p. 71].

At present, large language models, which play a crucial role in the development of information retrieval systems, are opening new possibilities by moving beyond traditional keyword-based queries and ranked result lists. The retrieval-augmented generation (RAG) paradigm based on large language models represents a more dynamic and interactive search system that integrates question–answering, information extraction, and other information access mechanisms between the user and the search system. This approach enables the use of richer semantic representations through advanced language models [3].

**Methods and Materials.** This research employs a theoretical-analytical methodology based on a comprehensive review of current scientific literature in the fields of linguistics, information retrieval, and artificial intelligence. The study synthesizes findings from interdisciplinary research involving natural language processing, semantic search technologies, and large language models.

Materials used include peer-reviewed articles and conference papers on semantic search, transformer-based models, and query optimization; technical documentation and experimental results from open-access machine learning platforms; ontologies and structured knowledge graphs utilized in contemporary search systems; case studies and system architectures from existing dialogue-based and retrieval-augmented gen-

eration (RAG) systems; as well as official publications and implementations from major AI frameworks such as BERT, GPT, T5, RoBERTa, and related tools.

Data for the analysis were collected from open scientific repositories, including ACM Digital Library, IEEE Xplore, arXiv, and Google Scholar. Evaluation of system performance, model behavior, and semantic accuracy was conducted based on secondary data analysis from published experimental studies.

The study does not involve the development of new software but instead evaluates and compares existing AI-driven semantic search models in terms of linguistic effectiveness, adaptability to user queries, and integration potential into practical information retrieval environments.

**Results.** One of the main functions of modern information systems (IS) is the retrieval of data that meet specific criteria. At the present stage, the large volume of archival documents and resources within information systems creates serious challenges for information retrieval and its management. Traditional methods of information retrieval, especially when working with large-scale collections, give rise to a number of difficulties in obtaining the required information. This problem becomes even more pronounced due to the diversity of document formats and during searches within unstructured data.

The specificity of the methodological and technological problems that arise in organizing such searches is determined by several factors. First of all, this is related to the nature of the content contained in the information resources entering the system. Although textual content still predominates in modern information systems, multi-format resources that include multimedia content (graphic, audio, and video information, etc.) are becoming increasingly widespread. In addition, various forms of content structuring are used to improve operational efficiency. As a result of structuring, information is divided into factual data, metadata describing their structure, and even "meta-metadata" that define different variants of data structures. In this context, the following can be identified as specific cases of the information retrieval problem: object and image recognition, speech recognition, semantic analysis, and translation of natural language texts.

Since the task of information retrieval systems is to effectively manage and provide access to such large volumes of data, existing traditional methods are unable to keep pace with the rapidly growing data volume. This leads to problems related to accuracy, speed, and system scalability.

At this stage, artificial intelligence technologies provide significant support by enhancing the functionality of information retrieval systems through the use of machine learning algorithms, deep learning models, and natural language processing techniques.

– *Machine learning algorithms* enable systems to learn from data patterns, thereby improving the relevance and accuracy of search results.

– *Deep learning models* strengthen the system's ability to understand complex relationships within data.

– *NLP technologies* make it possible to capture semantic meaning, perform context analysis, and identify user intent [2, p. 3].

The application of these techniques in search systems results in more effective and personalized information retrieval for users.

In general, considering content characteristics in search systems–either explicitly or implicitly–the following trends can be identified for organizing effective information retrieval within relevant resources:

The aim here is to understand natural language questions and provide dialogue-like responses. For this purpose, natural language processing (NLP) and dialogue management technologies are employed, enabling users to interact with search systems in a more convenient and efficient manner [24, p. 925]. This approach facilitates complex information retrieval through interactive dialogues conducted in natural language. For some time now, voice assistants and chatbots have been widely used by users. However, the information retrieval capabilities of these systems remain relatively limited and are often restricted to answering simple questions [22, p. 73]. In recent years, the development of simple question–answering systems has evolved into the emergence of large language models capable of supporting more complex dialogues, and the integration of these models into search systems has gained significant momentum.

Unlike traditional keyword-based search, dialogue-based search systems support multi-turn interactions, allowing users to gradually refine their queries and explore the information space in a step-by-step manner [28, p. 1402].

Researchers have examined the key components considered essential for dialogue-based search systems built on large language models. These include query reformulation, search clarification, and response generation [6, p. 2].

Query reformulation has long been regarded as one of the primary mechanisms for addressing the problem of lexical mismatch in information retrieval. This can be achieved, for example, by expanding queries with related terms or by generating paraphrases – different expressions that convey the same meaning [27, p. 1].

Recent studies show that large language models have become one of the most effective tools for refor-

mulating queries, including dialogue-based queries [16]. However, queries generated by LLMs, like those rewritten by humans, still require additional optimization for effective integration into search systems. In this regard, several approaches have been proposed by researchers, among which the **zero-shot** approach has attracted particular attention. Compared to traditional methods, this approach can provide more flexible and powerful results in query reformulation.

The main goal of query reformulation is to fully or partially replace the terms in the original query. At the same time, this process may also include the following operations:

– adding new words to the query (expansion),

– removing certain words from the query (optimization),

– adjusting the weights of words according to their importance within the query, or applying all of these methods together [15, p. 135].

In zero-shot query reformulation, LLMs understand the original form of the user query without additional examples and infer the user's intent. In this process, the model may add relevant keywords, paraphrases, and contextual information to the query. For example, a user query: *"Restaurants in Baku."* Since this query is relatively vague, an LLM can reformulate it, for example, as *famous restaurants in Baku*, *outdoor restaurants*, or *family-friendly restaurants*. In addition, the query reformulation process also includes issues related to query structure or format, such as stemming, removal of stop words, acronym expansion [10, p. 77], spelling correction, and the refinement or expansion of terms. The next stage that helps users formulate their queries more precisely involves the search system presenting clarification questions in cases where it detects user uncertainty or ambiguity.

Search clarification enables dialogue-based search systems to interact more actively with users, especially when queries are ambiguous or multi-faceted, that is, when they can be interpreted in multiple ways. Instead of directly presenting results, such systems ask clarification questions to better understand user intent. For example, when a user submits the query "Apple" in English, the search system asks a clarifying question to determine whether the query refers to the company or the fruit: *"What do you mean by 'Apple'?"*

At the same time, the system may offer possible response options, such as:

– *Information about the apple fruit*;

– *Products of Apple Inc. (iPhone, Mac, iPad, etc.)*;

– *Latest news related to Apple*.

After the user selects the desired option, the search system ranks and presents the corresponding results. The analysis of user interactions involving clarification questions can help to gain a deeper understanding of the clarification process in search and enable researchers to identify which queries require clarification, as well as which types of clarification questions users tend to prefer [29, p. 1].

Since dialogue-based search queries may take different directions, search clarification is divided into four main categories:

1. Clarification for conversational document retrieval
2. Web search clarification
3. Search clarification for question answering systems
4. Domain-specific search clarification [6, p. 7].

The goal of conversational document retrieval is to find the most relevant and appropriate document based on the user's query. For example, a user may submit the query: *"Find recent research on environmental changes."* To refine the query, the system may ask a clarification question such as: *"Are you looking for scientific articles on climate change, or general information about its impacts on nature?"* Based on the user's response, the system then presents articles from scientific journals or ecological blogs. To model such dialogues, datasets such as OR-QuAC [4] or CAsT (Conversational Assistance Track) [5] are commonly used.

In some cases, users submit queries consisting of only one or two keywords, which results in a large number of possible answers. The purpose of web search clarification is to disambiguate vague or short queries. For instance, if a user enters the query *"puma"*, the system may ask whether *"puma"* refers to the animal or the company. After determining the intended meaning based on the user's response, the system displays the appropriate results. Such clarification mechanisms are implemented in applications like Google and Bing in the form of *"Did you mean…"* prompts or automatic suggestions (auto-suggest).

In question–answering systems, the aim of search clarification is to identify the correct answer by clarifying the context when a user's question is general or has multiple possible interpretations. Artificial intelligence systems, particularly large language models, refine such questions at the factual level, thereby providing more precise and accurate answers.

During domain-specific searches (e.g., in various fields of science, medicine, programming, etc.), terms are often specialized and frequently ambiguous. In such cases, terminological and contextual clarification is required. To perform this type of clarification, domain-specific large language models (LLMs) such as LegalBERT, CodeBERT, and BioGPT are used.

For example, a user query: *"Can I create a branch for a bug fix?"* The system may ask a clarification

question such as: *"Are you referring to creating a branch in Git, or to a fix branch in a CI/CD system?"*

In traditional search systems (Google, Yandex, Bing, etc.), during response generation, a list of web pages or relevant documents matching the query is returned and ranked. The current improved versions of these systems consist of two components: a retrieval module and a response generation module [14; 21]. The retrieval module identifies relevant information and facts (indexed documents, web pages) using models such as BM25 and dense retrieval approaches (BERT, DPR, ColBERT, etc.).

The generator, in turn, uses large language models (GPT, T5, DeepSeek, etc.) to produce syntactically and semantically coherent new texts based on the retrieved information. One of the most widely used approaches in such systems is Retrieval-Augmented Generation (RAG). Using this approach, large language models access an external knowledge base composed of documents, retrieve relevant information through semantic similarity calculations, and generate new texts that correspond to the user query. This approach is particularly useful for applications such as question–answering systems, summarization, and dialogue agents.

The RAG approach improves the accuracy and reliability of generated texts by incorporating knowledge retrieved from external databases, while also enabling continuous knowledge updates and the integration of domain-specific information [21, p. 155]. Text-based RAG models are primarily built on transformer architectures (e.g., BERT and T5). These models use self-attention mechanisms to capture contextual relationships within text, thereby improving both retrieval accuracy and the fluency of text generation. An example of how LLMs operate using the RAG approach can be illustrated as follows:

User question: *"What is morphology in linguistics?"*

The LLM retrieves semantically relevant texts from external knowledge bases, text corpora, web pages, social networks, and other sources, and then generates a new text by leveraging multiple contexts. For example:

1. In linguistics, morphology studies parts of speech and their rules of change. It examines the rules governing linguistic facts within the internal structure of words [17].

2. Morphology studies word forms. The main subject of morphology is parts of speech. In morphology, words are studied as parts of speech, and their structure and rules of change are analyzed [18].

3. Morphology is the set of rules governing how words change and take different forms in a sentence, as well as the discipline that studies these rules [12, p. 24].

4. Morphology is a branch of grammar. Its object of study is the word. Morphology primarily focuses on various grammatical forms that express grammatical meanings. It covers all parts of speech and examines their similarities and differences [11, p. 5].

Using texts retrieved from various sources by the search module of an LLM (e.g., ChatGPT), the system may generate a synthesized response such as:

*Morphology is a branch of grammar that studies words and their grammatical patterns of change, grammatical forms, as well as the structure and characteristics of parts of speech. It focuses on the internal structure of words, their grammatical forms, the grammatical meanings expressed by these forms, and the similarities and differences among parts of speech.*

*In other words, the main objects of study in morphology include:*

*– word forms and rules of inflection,*
*– parts of speech and their grammatical features,*
*– the internal structure and composition of words,*
*– means of expressing grammatical meanings.*

In general, it should be noted that the RAG approach is primarily used in large language models (LLMs). It was developed to reduce the hallucination problem of LLMs and to enable the retrieval of factual information from external data sources. In search systems, however, RAG-like architectures have been adopted only since 2023. For example, in Google this approach is referred to as Search Generative Experience (SGE). Unlike LLMs, search systems generate shorter, more constrained, and fact-based responses, explicitly grounded in indexed web pages.

Semantic search is a technique that provides more accurate answers by understanding the intent and contextual meaning of a user's query. Unlike traditional keyword-based search methods, semantic search evaluates the relationships between words and expressions and delivers more precise information based on the query. This approach takes into account various factors such as synonyms, different variations of expressions, and the broader context of the query. Semantic search makes it possible to retrieve documents that may not explicitly contain the query terms but are semantically related to them. In this process, the goal of artificial intelligence systems is not merely to match keywords, but to understand semantic relationships between words, identify patterns, and even predict the user's potential next steps.

Compared to traditional search methods, semantic search has several key advantages. First, by considering the context of the query and the relationships between words, it enables higher accuracy and relevance. This approach allows for deeper query under-

standing and, as a result, provides more appropriate answers. In addition, it is capable of extracting knowledge from large volumes of unstructured data (documents and texts).

The operating principle of semantic search is based on natural language processing (NLP), machine learning, and knowledge graphs. The process begins with the analysis of the query to understand the user's intent. This stage is referred to as query understanding, during which the language, syntax (sentence structure), and semantics (meaning) of the query are identified and processed using NLP techniques.

The content stored in the system's database is indexed using methods such as latent semantic analysis. Latent semantic indexing (LSI) is a search engine algorithm that analyzes web pages by considering not only the keywords and expressions present on a page, but also the presence of their synonyms and thematically related terms. The algorithm aims to ensure that documents that best answer the search query appear at the top of the search results. Latent semantic indexing is a specific application of latent semantic analysis in search engines. These methods map similar concepts, words, or expressions into a closely connected semantic network, allowing the system to retrieve relevant results even when different terms or synonyms are used in the query.

The process that improves the accuracy of search results and promotes documents with stronger contextual relevance between the query and content–by using deep learning models – is known as semantic ranking. Semantic ranking refers to ordering search results based on semantic information. To perform semantic ranking, machine learning is applied in several stages. In the first stage, potential data sources that can answer the query are selected. In the second stage, each data source generates candidate answers using learning-to-rank methods. In the third and final stage, these sources are ranked, user intents are selected and ordered, and the answers within each intent (e.g., news results) are presented to the user as a final, aggregated result.

Thus, in semantic search, queries may yield results that:

– use different words conveying the same meaning,

– are based on synonyms,

– present contextually related results from different perspectives,

– provide direct, specific answers, and so on.

For example, if the query *"How can headaches be reduced?"* is submitted to a search system, the following types of results may be returned:

1. *How to relieve a headache?*
2. *What helps with headaches?*
3. *Ways to reduce pain without painkillers.*
4. *Causes of migraine headaches and methods of prevention*, etc.

To understand user intent in semantic search systems, approaches based on knowledge graphs, neural information retrieval, vector-based models, hybrid approaches, and ontology-based methods are employed. Knowledge graphs and Neural Information Retrieval (Neural IR) are two complementary approaches that together enhance the effectiveness of semantic search systems.

Knowledge graphs are databases that store structured information about objects, concepts, and the relationships between them. By using knowledge graphs, search systems can better understand user queries and present more relevant results by linking related concepts [24].

The main advantage of using knowledge graphs in semantic search systems lies in providing a more accurate interpretation of a query's meaning by analyzing the concepts within the query in the context of explicit semantic relationships. The application of knowledge graphs enables more efficient execution of tasks such as identifying entities within a query (identifiers, attributes, and relations), disambiguating multiple possible meanings of the same word or expression based on context, and expanding the semantic relationships of a query. At the same time, building and continuously updating such graphs requires significant resources.

Neural information retrieval (Neural IR) is a research area that uses neural networks to simulate information retrieval tasks such as query understanding and relevance ranking. This approach aims to improve the accuracy of search results by considering complex relationships between queries, documents, and users [24, p. 926].

Neural-based models learn distributed vector representations of queries and document texts, capturing their contextual and semantic properties. Rather than relying solely on lexical matching of query terms, these models account for linguistic features such as synonymy, paraphrasing, and context dependence, enabling the identification of the implicit intent of a query. In particular, transformer-based models such as BERT and its variants enhance information retrieval effectiveness by capturing long-range semantic dependencies and measuring semantic similarity [26, p. 401].

In vector-based approaches, each query and document is represented in a high-dimensional vector space, where semantically similar objects are positioned closer to each other. As a result, semantic relationships between query intent and document content are evaluated using mathematical similarity measures.

Compared to traditional keyword-based methods, this enables more accurate and context-aware matching [23, p. 1].

Ontology-based approaches are widely used in semantic search systems to represent and interlink knowledge in a structured manner. An ontology is a tool for modeling conceptual structures and expressing information at the knowledge and semantic levels. Through their hierarchical organization, ontologies enable effective structuring of knowledge for semantic search and describe semantics by leveraging relationships between concepts. Ontologies provide methods for representing and processing both knowledge and queries. They are designed to describe the semantics of information within a specific domain and to address the problem of conceptual ambiguity [25]. An ontology defines a shared vocabulary for information exchange within a domain, including core concepts and machine-interpretable representations of the relationships between them.

A hybrid approach in information retrieval is created by combining multiple methods, with the aim of improving semantic relevance and search accuracy. In this approach, the strengths of traditional keyword-based lexical matching, vector-based methods, knowledge graph–based methods, and ontology-based methods are integrated [7, p. 1].

The main advantage of the hybrid approach is that the weaknesses of one method are compensated by the strengths of others. For example, neural-based models provide high accuracy in contextual semantic matching during information retrieval, but their explainability is relatively limited. In contrast, knowledge graph– and ontology-based approaches offer clearer explanations through structured data and explicit conceptual relationships. By combining these methods, hybrid approaches yield more effective results.

Hybrid approaches are primarily applied to tasks such as understanding query intent, disambiguation – i.e., determining the contextually appropriate meaning among multiple possible meanings of a word or expression – semantic ranking, and information enrichment. Nevertheless, the effectiveness of these approaches is closely related to the quality of internet resources, as well as the model's learning capacity and computational capabilities.

**Conclusion.** The conducted research shows that information retrieval systems have developed significantly with the application of artificial intelligence technologies. Since traditional keyword-based search methods are insufficient for retrieving query-relevant information from large-scale and unstructured data, the development of semantic search systems has come to the forefront.

The integration of large language models into information retrieval systems has enabled more optimal and effective performance of functions such as dialogue-based search, query reformulation, clarification question generation, and response generation. In particular, the RAG approach used by LLMs enhances the accuracy and reliability of generated responses by leveraging external knowledge bases, while also partially mitigating the hallucination problem.

Various approaches applied in semantic search systems demonstrate high effectiveness in understanding user intent and identifying semantic relevance. Hybrid approaches, by combining the strengths of these methods, make it possible to achieve more effective results in terms of both accuracy and explainability.

In conclusion, the future development of information retrieval and query systems is closely linked to artificial intelligence technologies – especially large language models – integrated within hybrid semantic search frameworks.

**REFERENCES:**

1. Abimbola B. Intelligent Information Retrieval System. International Journal of Artificial Intelligence and Machine Learning. 2022. 2(1). P. 71-74. DOI: 10.51483/IJAIML.2.1.2022.71-74

2. Adelakun, Olawale N. Exploring the Impact of Artificial Intelligence on Information Retrieval Systems. Information Matters. 2024. 4(5). P. 1-5. DOI: 10.2139/ssrn.4834942

3. Breuer T., Frihat, S., Fuhr, N. et al. Large Language Models for Information Retrieval: Challenges and Chances. Datenbank Spektrum. 2025. 25. P. 71-81. DOI: 10.1007/s13222-025-00503-x

4. Chen Q., Liu Y., Cen Ch. Et.al. Open-Retrieval Conversational Question Answering. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25-30, 2020. Virtual Event, China. ACM, New York, NY, USA, 10 p. DOI: 10.1145/3397271.3401110

5. Dalton J., Xiong C., Callan J. TREC CAsT 2019: The Conversational Assistance Track Overview. SIGIR '20: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020. P. 1985 – 1988. DOI: 10.1145/3397271.340120.

6. Fengran M., Kelong M., Ziliang Z. et. al. A Survey of Conversational Search. ACM Trans. Inf. Syst. Vol. 43. № 6. Article 167. 2025. 50 p. DOI: 10.1145/3759453

7. Godinez, A. HySemRAG: A Hybrid Semantic Retrieval-Augmented Generation Framework for Automated Literature Synthesis and Methodological Gap Analysis. 2025. 47 p. (Available at: ArXiv, abs/2508.05666.

8. Hambarde K. A., Proença H. Information Retrieval: Recent Advances and Beyond. in IEEE Access. 2023. 11. P. 76581-76604. 2023. DOI: 10.1109/ACCESS.2023.3295776.

9. Heydarova M. Characteristics of the semantic search system. Terminology issues. 2025. 1. P. 202-207. DOI: 10.59849/2663-8967.2024.1.202

10. Huang J., Efthimis N. Efthimiadis. Analyzing and evaluating query reformulation strategies in web search logs. Proceedings of the 18th ACM conference on Information and knowledge management 2009. P. 77-86. DOI: 10.1145/1645953.1645966

11. Hüseynzadə M. Müasir Azərbaycan dili. III hissə. Morfologiya. Bakı: "Şərq-Qərb", 2007. 280 s.

12. Kazımov Q.Ş. Müasir Azərbaycan dili. Morfologiya. Bakı: "Nurlan", 2010. 400 s.

13. Kumar D., Singh Sh. Advancements in transformer architectures for large language model: from bert to gpt-3 and beyond. International Research Journal of Modernization in Engineering Technology and Science (Peer-Reviewed, Open Access, Fully Refereed International Journal). 2024. 6(5). DOI: 10.56726/IRJMETS55985

14. Lewis P.; Perez E.; Piktus A. et.al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. 2005. (Available at: arXiv preprint arXiv:2005.11401).

15. M. Mosbah. Query Refinement into Information Retrieval Systems: An Overview. J. inf. organ. sci. (Online). 2023. 47(1). P. 133-151.

16. Ma X., Gong Y., He P., et. Al. Query Rewriting for Retrieval-Augmented Large Language Models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. P. 5303–5315 December 6-10, 2023. Association for Computational Linguistics. URL: ArXiv, abs/2305.14283.

17. Morfologiya (dilçilik). Vikipediya Azad Ensiklopediya (Available at: https://az.wikipedia.org/wiki/Morfologiya_(dil%C3%A7ilik).

18. Morfologiya. (Available at: https://kayzen.az/blog/Az%C9%99rbaycan-dili/469/morfologiya.html).

19. Rogushina. J.V. Application of the Ontological Model for Semantic Search of the Information Objects. Ontology of Designing. 2015. 3(17). P. 336-356. DOI: 10.18287/2223-9537-2015-5-3-336-356

20. Rovatsos M., Filgueira R. Building Lightweight Semantic Search Engines. In Proceedings of the 19th IEEE Conference on eScience. (International Conference on e-Science (e-Science)). Institute of Electrical and Electronics Engineers. 2023. P. 1-10. DOI: 10.1109/e-Science58273.2023.10254874

21. Sawarkar K., Mangal A., Solanki S.R. Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers. 7th International Conference on Multimedia Information Processing and Retrieval (MIPR). 2024. P. 155-161.

22. Schneider Ph., Poelman W., Rovatsos M. et al. Engineering Conversational Search Systems: A Review of Applications, Architectures, and Functional Components. In Proceedings of the 6th Workshop on NLP for Conversational AI (NLP4ConvAI 2024). Bangkok, Thailand. 2024. P. 73-88.

23. Seyed Monir S., Lau I., Yang, S. et.al. VectorSearch: Enhancing Document Retrieval with Semantic Embeddings and Optimized Search. 2024. 10 p. URL: ArXiv, abs/2409.17383

24. Sharma S., Panda S.P. Efficient information retrieval model: overcoming challenges in search engines-an overview. Indonesian Journal of Electrical Engineering and Computer Science. 2023. 32(2). P. 925-932.

25. Shunli Q., Wenxia J. Research on the Construction of Semantic Information Retrieval Model Based on Logistics Ontology. In Proceedings of the 2024 4th International Symposium on Big Data and Artificial Intelligence (ISBDAI '24). Association for Computing Machinery, New York, NY, USA. 2025. P. 247-252 DOI: 10.1145/3723366.3723405

26. Trabelsi M., Chen Z., Davison B.D. et al. Neural ranking models for document retrieval. Inf Retrieval J. 2021. 24. P. 400–444. DOI: 10.1007/s10791-021-09398-0

27. Wang X., Macdonald C., Ounis I. Deep Reinforced Query Reformulation for Information Retrieval. Virtual Event, China. DRL4IR, July 30, 2020. 10 p. (Available at: ArXiv, abs/2007.07987).

28. White A. Synthesizing Human-Like Conversational Search Interactions with Large Language Models. Preprints 2025. P. 1401-1412. DOI: 10.20944/preprints202503.1401.v1

29. Zamani H., Mitra B., Chen E. et. al. Analyzing and Learning from User Interactions for Search Clarification. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. In (SIGIR '20), July 25–30, 2020. 10 p. DOI: 10.1145/3397271.3401160



*Дата першого надходження статті до видання: 20.02.2026*
*Дата прийняття статті до друку після рецензування: 30.03.2026*
*Дата публікації (оприлюднення) статті: 07.05.2026*