

ChatMIT: A university recruitment chatbot based on zero-shot self–retrieval augmented generation

Michał Błaszczkowski¹, Magdalena Jastrzębska¹,
Magdalena Paśnikowska-Łukaszuk^{2*}

¹ Department of Applied Mathematics, Faculty of Mathematics and Information Technology,
Lublin University of Technology, Nadbystrzycka 38, 20-618 Lublin, Poland

² Department of Technical Computer Science, Faculty of Mathematics and Information Technology,
Lublin University of Technology, Nadbystrzycka 38, 20-618 Lublin, Poland

* Corresponding author's e-mail: m.pasnikowska-lukaszuk@pollub.pl

ABSTRACT

This study presents ChatMIT, a domain-specific conversational system designed to support undergraduate recruitment at the Faculty of Mathematics and Information Technology at Lublin University of Technology. Three Retrieval-Augmented Generation architectures, namely Naive RAG, Hybrid RAG, and Self-RAG, were implemented and evaluated across three levels of question complexity using three large language models of varying parameter scales: Gemma 3 (27B), Llama 3 (70B), and gpt-oss (120B). Response quality was assessed by independent domain experts according to two criteria: factual correctness and linguistic quality, each evaluated using a five-point scale developed by the authors. The results indicate that the choice of RAG architecture is the primary factor influencing response quality, exerting a greater impact than model scale. Self-RAG substantially outperformed both Hybrid RAG and Naive RAG across all levels of question complexity, achieving mean factual correctness scores of 4.67/5, 2.77/5, and 2.33/5, as well as linguistic quality scores of 4.90/5, 4.37/5, and 4.33/5, respectively. Furthermore, the adoption of a zero-shot variant of Self-RAG constitutes a methodological contribution of this study, demonstrating that the adaptive properties of the original architecture can be preserved while significantly reducing deployment complexity, without the need for supervised training of a dedicated critic model. The findings confirm that RAG-based conversational systems represent an effective approach to automating information access in academic environments and offer considerable potential for deployment across diverse institutional domains and faculties.

Keywords: natural language processing, retrieval-augmented generation, large language models, conversational systems, chatbots.

INTRODUCTION

The dynamic development of information technologies, along with the rapid growth in the volume of available data, has made effective information retrieval one of the key challenges of modern computer science. In the era of the information society, users expect fast access to reliable and precise information, often formulating their queries in natural language. A significant portion of available resources is unstructured in nature,

primarily including textual documents, articles, advertisements, and various forms of content published on the Internet. Information retrieval (IR) refers to the process of identifying documents or text fragments that best match a user query within large collections of data.

In contrast to traditional data retrieval from structured databases, where queries have a strictly defined form and lead to unambiguous results, information retrieval is inherently probabilistic. IR systems estimate the relevance of documents

with respect to a given query by considering keyword matching or semantic similarity of content. With the increasing volume and heterogeneity of available data, the development of effective methods for information retrieval and selection has become one of the main research areas in natural language processing and artificial intelligence.

The history of information retrieval systems reflects the long-term evolution of text representation models and content matching methods. Early approaches were based on the Boolean model, in which documents were retrieved according to logical conditions concerning the presence or absence of specific terms in the text [1]. Over time, statistical approaches emerged that represented words or entire documents as numerical vectors, such as the Bag-of-Words model and the TF-IDF weighting scheme, which enabled the comparison of texts using similarity measures in a vector space [2, 3]. Subsequently, probabilistic models gained importance, among which the widely used BM25 algorithm occupies a prominent position [4, 5]. A significant breakthrough was also the development of static and contextual word embeddings, such as Word2Vec, GloVe, and BERT, which enabled the modeling of semantic relationships between words and sentences and significantly improved the effectiveness of information retrieval in large document collections [6–8].

An important direction of development in the field of natural language processing and information retrieval has been the emergence of large language models (LLMs), which enable the generation of coherent and contextually relevant responses to user queries. Despite their impressive generative capabilities, these models exhibit certain limitations. Their knowledge is derived exclusively from the data used during the training process, which is inherently static and does not always reflect the current state of knowledge. Therefore, the generated responses may be imprecise or lack references to reliable sources of information [9, 10].

A response to the limitations is the retrieval-augmented generation (RAG) approach, which combines information retrieval mechanisms with generative language models. In systems based on this architecture, the generative model does not rely solely on knowledge acquired during training, but also leverages documents from external knowledge sources through a retrieval mechanism. Given a query, the system first retrieves

the most relevant documents or their fragments, which are then used as contextual input for the language model responsible for generating the final response. This approach increases the reliability and timeliness of generated outputs while enabling the system to be adapted to a specific domain of knowledge [11].

Various variants of RAG architectures have been identified in the literature, differing in their information retrieval methods and in the way retrieved results are integrated into the response generation process [11]. In this study, three representative approaches are considered: Naive RAG, Hybrid RAG, and Self-RAG. Naive RAG represents the basic form of the architecture, in which documents are retrieved based on semantic similarity and provided to the generative model as contextual input. The Hybrid RAG architecture extends this approach by incorporating keyword-based retrieval, thereby combining semantic and lexical matching. In contrast, Self-RAG introduces additional self-evaluation mechanisms that enable the assessment of the coherence of the retrieved context and the quality of the generated responses.

The application of RAG architectures is of particular importance in chatbot systems, which are conversational tools designed to simulate human communication, capable of operating across various digital environments, including websites, mobile applications, and instant messaging platforms, making them widely applicable in customer service, digital platforms, and recruitment processes. One prominent application of such systems is human resource management, where chatbots support recruitment by providing information about job vacancies, answering candidate questions, collecting application data, and conducting initial candidate screening. Artificial intelligence technologies can analyse candidate experience, education, and competencies, comparing them with job requirements to identify the most suitable applicants efficiently [12]. Research indicates that job seekers perceive chatbots as moderately effective, particularly in providing recruitment-related information and scheduling interviews, which has led to their increasing adoption by organizations seeking to emulate human intelligence in recruitment-related activities [13].

Chatbots are also increasingly used by higher education institutions to support student recruitment. Prospective students often seek information regarding study programmes, admission

requirements, required documentation, and application deadlines, yet traditional information dissemination methods require them to independently navigate large volumes of content. The implementation of a conversational system based on advanced information retrieval mechanisms and language models can significantly improve this process by enabling users to ask questions in natural language and receive comprehensive, direct, and personalized responses, while simultaneously handling a large number of inquiries. An example is the chatbot Pounce, implemented at Georgia State University, (Figure 1), which contributed to reducing the „summer melt” phenomenon, where admitted students ultimately fail to enrol at the university.

Similar solutions are also implemented at other higher education institutions, including Nanyang Technological University, where a chatbot supports the handling of applicant inquiries concerning educational programmes and admissions procedures (Figure 2).

In this paper, a conversational system, ChatMIT, designed to support the recruitment process for undergraduate studies at the Faculty of Mathematics and Information Technology of Lublin University of Technology, is developed and evaluated. The system employs Retrieval-Augmented Generation architectures to improve the quality of generated responses and ensure their consistency with information contained in the recruitment knowledge base. Three RAG architectures are compared in terms of the factual correctness and linguistic quality of the generated responses, using three language models of varying parameter sizes. Figure 3 presents the user panel view of the recruitment chatbot ChatMIT.

Overall, artificial intelligence-based chatbots are becoming an increasingly significant component of modern recruitment and admissions processes. They enable the automation of numerous stages of candidate screening and selection, while simultaneously accelerating communication with applicants, both prospective students and individuals seeking information regarding academic programmes prior to making decisions concerning higher education enrolment. Despite their numerous advantages, the effectiveness of such systems remains dependent upon the quality of the implemented algorithms and the appropriate integration of technological solutions with the conventional work performed by recruitment and admissions committees.

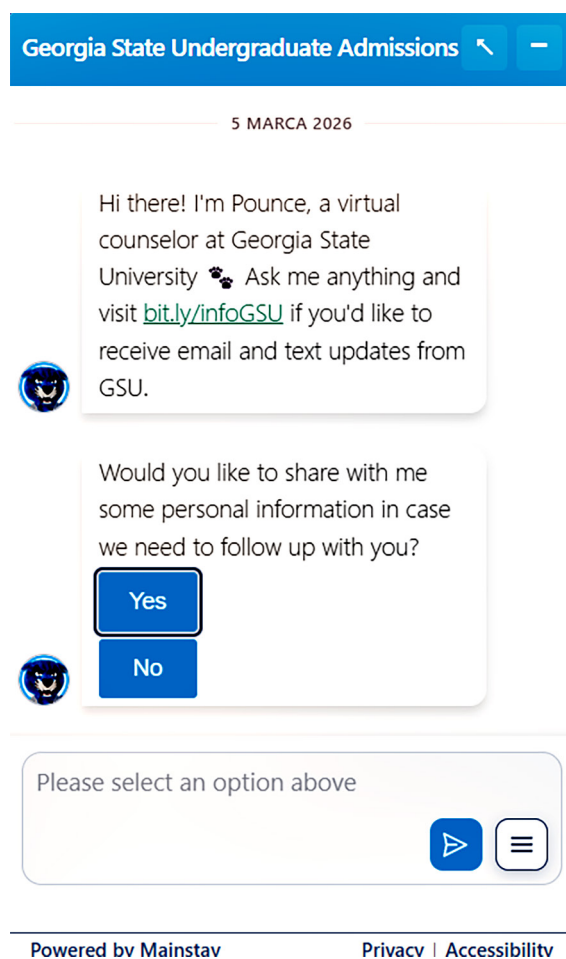


Figure 1. Chatbot Pounce, implemented at Georgia State University [14]

MATERIALS AND METHODS

Knowledge base and data preparation

The ChatMIT system is based on a knowledge base constructed from official documents published by the Faculty of Mathematics and Information Technology at Lublin University of Technology. The collected materials included regulations and resolutions governing the recruitment process, recruitment schedules, requirements regarding candidate documentation, descriptions of study programmes together with learning outcomes, as well as organizational information from the faculty website. In all cases, the most up-to-date version of each source was selected to ensure the currency of the knowledge base. The test question dataset was developed based on real email correspondence of the Faculty Recruitment Committee from the academic years 2023/2024 and 2024/2025, covering two academic years in order to capture a diverse range of prospective

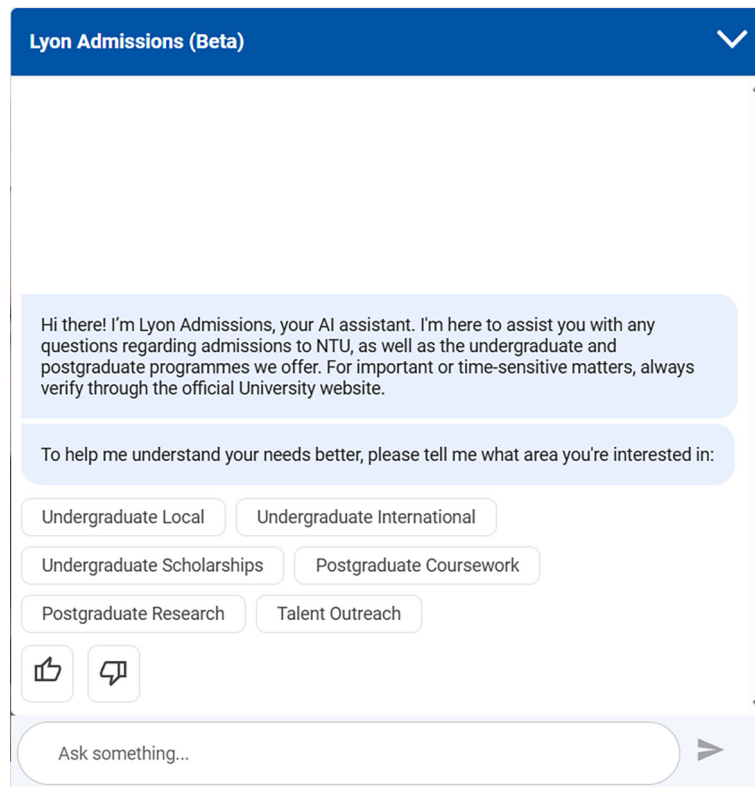


Figure 2. Chatbot implemented at Nanyang Technological University [15]

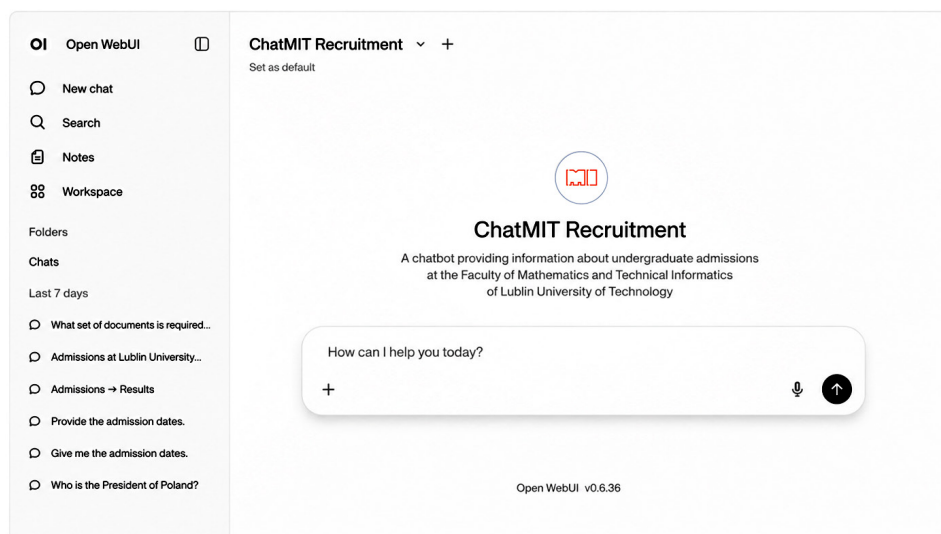


Figure 3. Chatbot ChatMIT implemented at Lublin University of Technology

student inquiries. The correspondence was used exclusively to identify recurring question types and topics, while responses were generated based on the most current documents available in the knowledge base. In accordance with data protection requirements, all personal information was removed from the correspondence, retaining only content related to the recruitment process.

The document preparation stage played a crucial role in the overall effectiveness of the system and enabled approximately a twentyfold reduction in the size of the knowledge base while preserving its full informational value. The preprocessing steps were performed manually and included the removal of elements irrelevant to text analysis, such as signatures, stamps, and logos, and

the transformation of tabular data into continuous text while preserving relationships between fields. Additionally, complex terms characteristic of the academic environment were reformulated according to plain language principles and redundant content was identified and removed, using a suitably prompted large language model (GPT-4o), in order to improve the quality of semantic matching between user queries and document content. All modifications were subsequently verified by a university expert to ensure that the original meaning was preserved while domain-specific language was appropriately simplified.

Deployment environment

The ChatMIT system was implemented and deployed on the OpenWebUI platform, which enables the integration of language models with retrieval mechanisms based on RAG architectures, as well as testing the system under conditions similar to real-world usage. The platform provides integrated management of language models, the knowledge base, and system parameters without the need to implement a dedicated user interface, which significantly reduced deployment time and enabled the research effort to be directed towards the core experimental aspects of the study.

In the case of Naive RAG and Hybrid RAG architectures, it was possible to directly utilize the built-in mechanisms of the OpenWebUI platform, with only selected retrieval parameters requiring modification. In contrast, the Self-RAG architecture required the development of a dedicated external function implementing its operational mechanism, including the logic of reflection tokens and iterative control of the retrieval process, followed by its full integration with the deployment environment of the platform.

The computational environment was deployed on a server equipped with two NVIDIA RTX A6000 GPUs (48 GB VRAM each), based on the Ampere architecture, running NVIDIA drivers version 550.163.01 and the CUDA 12.4 environment. To ensure comparability across the analyzed architectures, uniform text segmentation parameters were applied in all configurations. Based on preliminary experiments, the analyzed documents were divided into chunks of 1000 characters, with an overlap of 100 characters between adjacent segments. This configuration yielded satisfactory retrieval effectiveness within the studied domain,

although optimal parameters may vary depending on the document structure, query complexity, and application context, even within the same domain. In all architectures, the initial context provided to the generative model consisted of the top three segments ($k = 3$). In the case of Self-RAG, this value defines the maximum candidate pool per retrieval call, as the model autonomously determines both the necessity of retrieval and the relevance of individual segments. This ensured that all architectures were evaluated under identical input conditions, enabling a direct and fair comparison of results.

Before evaluating the individual RAG architectures in the system, an initial assessment was conducted on two candidate embedding models for the retrieval module, namely sentence-transformers/all-MiniLM-L6-v2 and text-embedding-3-large (OpenAI), to determine the most suitable option. Based on comparative experiments, no significant differences were observed between the models in terms of the effectiveness of context retrieval, while a high degree of overlap between the retrieved text segments was maintained. For practical reasons, sentence-transformers/all-MiniLM-L6-v2 was selected as the final solution, as it is an open-source model that does not incur operational costs associated with the use of commercial models. In the generation module, three generative large language models of varying parameter scales were used for evaluation: Gemma 3 (27B), Llama 3 (70B), and gpt-oss (120B), enabling the analysis of the impact of model scale on response quality.

System architecture: RAG approaches

The core component of the ChatMIT system is the RAG architecture, which can be formally defined as a generative model conditioned on both the query x and the set of retrieved documents z such that:

$$p_{\theta}(y|x) = \sum_{z \in \text{top-}k} p_{\eta}(z|x) \cdot p_{\theta}(y|x, z) \quad (1)$$

where: y – denotes the generated response,
 $p_{\eta}(z|x)$ – the retrieval model assigning the probability of relevance to a document z with respect to the query x ,
 $p_{\theta}(y|x, z)$ – the generative model producing a response based on the query and the retrieved context.

The three analyzed architectures differ primarily in the formulation of the retrieval component $p_\eta(z|x)$, as well as in the strategies used to control the generation process [16, 17].

Naive RAG constitutes the fundamental implementation of the retrieval-augmented generation architecture. In this approach, the context retrieval stage is performed only once for a given user query, and the retrieved fragments are not subject to further iterative updates. The retrieval of relevant fragments is carried out by comparing the vector representation of the query with the vectors of text fragments in the embedding space. For this purpose, any similarity or distance measure may be applied, though in practice the most commonly used is cosine distance, defined as:

$$d_{\cos}(\vec{q}, \vec{v}) = 1 - \frac{\vec{q} \cdot \vec{v}}{\|\vec{q}\| \cdot \|\vec{v}\|} \quad (2)$$

where: $\vec{v}$ – denotes a vector representing a text fragment in the embedding space, and $\vec{q}$ – the query vector.

This metric takes values within the interval $[0, 2]$, where a value of 0 corresponds to the highest semantic similarity. Subsequently, the k fragments with the lowest distance values are selected, forming the context that is directly passed to the generative model together with the user query in order to generate a response [18,19]. The architecture of Naive RAG is illustrated in Figure 4.

Hybrid RAG extends the Naive RAG approach by incorporating an additional lexical retrieval component, most commonly implemented using the BM25 algorithm:

$$BM25(D, Q) = \sum_{i=1}^n \frac{IDF(q_i) \cdot f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{avgdl}\right)} \quad (3)$$

where: $f(q_i, D)$ – denotes the frequency of the i -th query term in document D ,

$IDF(q_i)$ – the inverse document frequency of the term q_i , k_1 is a term frequency saturation parameter, b – controls the influence of document length, $|D|$ – the length of the document, and $avgdl$ – the average document length in the corpus [20].

The semantic and lexical retrieval scores are subsequently combined through a weighted linear combination:

$$Score_{hybrid}(d) = \lambda \cdot Score_{lexical}(d) + (1 - \lambda) \cdot Score_{semantic}(d), \dots \quad (4)$$

where: $\lambda \in [0.1]$ – determines the contribution proportion of both methods.

In the conducted study, $\lambda = 0.5$ was adopted, reflecting equal weighting of both components, a choice motivated by the domain-specific nature of the analyzed documents, where neither retrieval method could be assumed to be universally superior. The value of this parameter was not subject to further tuning, as the primary objective was to incorporate the lexical retrieval component rather than to optimize its relative contribution. The hybrid approach is particularly effective in scenarios involving specialized, rare, or domain-specific vocabulary [21,22]. In such cases, general-purpose embedding models may struggle to accurately capture the meaning of terms in the vector space, which reduces the effectiveness of semantic retrieval. Incorporating a lexical component enables exact keyword matching, thereby improving retrieval relevance in specialized and technical contexts. The architecture of Hybrid RAG is illustrated in Figure 5.

Self-RAG extends classical RAG approaches by introducing an adaptive and sophisticated

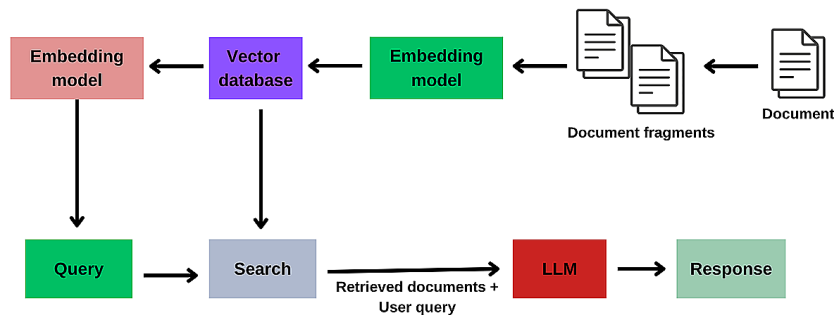


Figure 4. Architecture of the naive RAG system

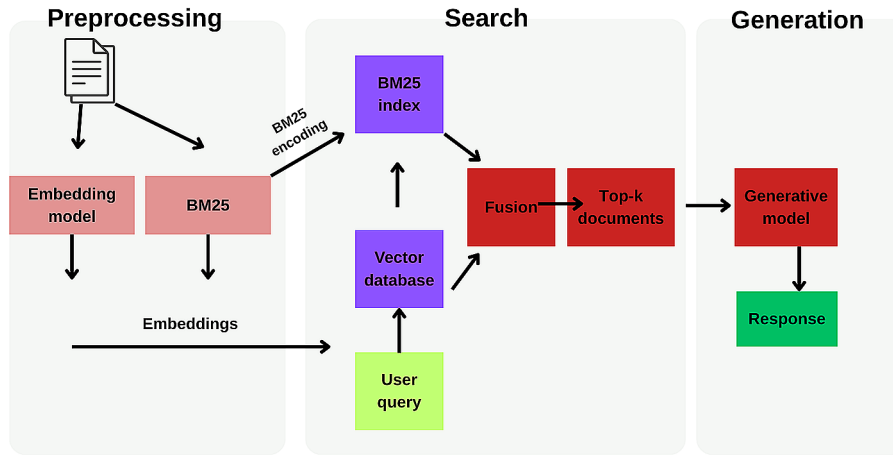


Figure 5. Architecture of the hybrid RAG system

mechanism for controlling both the retrieval process and response generation. Unlike static approaches, the model not only utilizes the provided context but also actively evaluates its relevance as well as the quality of the generated response. This mechanism is implemented through special reflection tokens, such as Retrieve, IsREL, IsSUP, and IsUSE, which enable the model to make decisions regarding the need to retrieve additional information, assess its relevance, and determine the extent to which it should be utilized during the generation process [23]. Specifically, Retrieve triggers an additional retrieval call when the model determines that the current context is insufficient; IsREL assesses whether the retrieved passage is relevant to the query; IsSUP evaluates whether the generated response is supported by the retrieved context; and IsUSE determines the overall utility of the retrieved passage for the final response. The critic model C is trained by maximizing:

$$\max_C \mathbb{E}_{(x,y,r) \sim \mathcal{D}_{critic}} \log p_C(r|x,y) \quad (5)$$

where: x – denotes the input query, y – the generated response, r – the reflection signal or label, $p_C(r|x,y)$ – the probability assigned by the critic model C to the reflection signal given the query and the generated response.

Based on the trained critic model, a dataset is constructed for training the generative model M , which optimizes:

$$\max_M \mathbb{E}_{(x,y,r) \sim \mathcal{D}_{gen}} \log p_M(y,r|x) \quad (6)$$

where: x – denotes the input query, y – the generated response, r – the reflection signal, $p_M(y,r|x)$ – the probability assigned by the

generative model M to the joint occurrence of the response and reflection signal given the query.

During inference, a total score is computed for each generation path:

$$f(y_t, d, Critique) = p(y_t|x, d, y_{<t}) + S(Critique) \quad (7)$$

where: $S(Critique)$ – is a weighted sum of the scores derived from individual reflection tokens.

As a result, the model not only generates a response but also simultaneously evaluates its correctness and consistency with the provided context. The application of the beam search algorithm enables the parallel expansion of multiple generation paths, from which the response with the highest overall score is ultimately selected [24].

The above formulation describes the original Self-RAG architecture, in which both the critic model C and the generative model M are trained with reflection tokens. In the present study, a simplified zero-shot variant was adopted, in which reflection-based evaluation is performed by the generative model itself through separate inference calls. Concretely, the model assesses at inference time both the relevance of the retrieved context with respect to the input query, and the degree to which the generated response is grounded in and supported by that context. In this approach, reflection labels are not embedded in the model weights through training but are generated dynamically as model outputs in response to dedicated evaluation prompts. The labels generated at both the retrieval

and generation stages were reviewed by domain experts, whose evaluation confirmed a sufficiently high level of correctness to consider the zero-shot approach viable. Where systematic errors were identified, the evaluation prompts were iteratively adjusted to improve the accuracy of reflection token assignment. Consequently, it was determined that training a dedicated critic model was not required, as the zero-shot variant was found to preserve the core adaptive properties of the Self-RAG architecture, including dynamic retrieval control and response quality assessment [25].

In the implemented zero-shot variant, reflection-based evaluation was performed through independent inference calls using the same model as the one responsible for response generation, albeit with a distinct system prompt describing all four reflection tokens (Retrieve, IsREL, IsSUP, IsUSE), their binary values, and the criteria governing their assignment. Each model evaluated its own reflection tokens independently, meaning that Gemma 3 (27B), Llama 3 (70B), and gpt-oss (120B) each operated as two separate instances: one dedicated to reflection token assessment and one to response generation, with no shared state between them. In cases where conflicting token assessments were obtained, the system initiated an additional retrieval iteration rather than proceeding with ambiguous context. The retrieval loop was permitted a maximum of five iterations per query. If the IsUSE token indicated that the retrieved context remained insufficient after five attempts, the system explicitly informed the user that a reliable answer could not be

provided, rather than generating a response based on inadequate context.

These adaptive mechanisms collectively enable more controlled and context-aware response generation, enhancing both the relevance and reliability of the final output. The architecture of the original Self-RAG is illustrated in Figure 6. The zero-shot variant implemented in this study follows an analogous structure, with the distinction that the response evaluation stage is performed by an independent inference call to the same generative model, guided by a dedicated prompt, rather than by a separately trained critic model.

System prompt and evaluation methodology

For all analyzed architectures and language models, the same specially designed system prompt was applied, serving as a higher-level control layer governing the system behavior regardless of the RAG architecture used. The prompt specified the thematic scope of the system, limited exclusively to undergraduate admissions at WMIT, defined the strategy to follow in cases where relevant information was not available in the knowledge base, and established guidelines for formulating responses in accordance with plain language principles. Key principles included the use of clear and understandable vocabulary, constructing short sentences conveying a single idea, employing the active voice, and visually structuring the content. The model was instructed to base its responses solely on the provided context and to explicitly indicate when there was insufficient

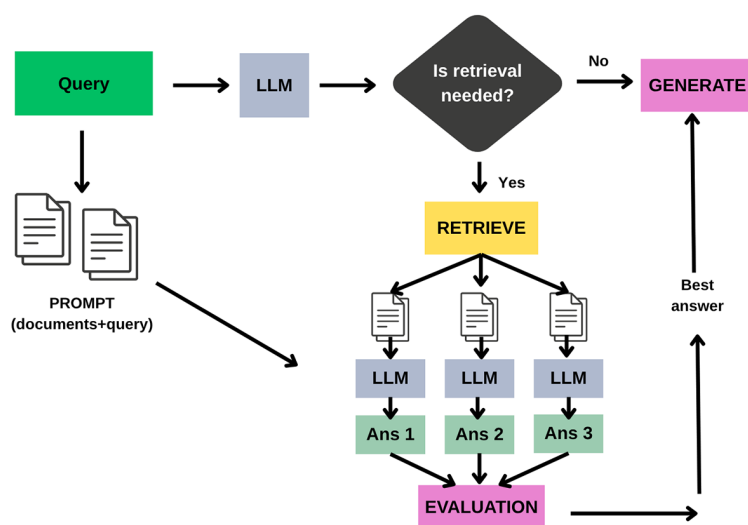


Figure 6. Architecture of the self-RAG system

information to provide an answer. Additionally, the prompt defined response control guidelines whereby answers were to remain concise and not expanded beyond what was necessary, with no elaboration unless explicitly requested by the user. Each response was to conclude with a single follow-up question to verify whether the user required additional detail. For general queries, the model was instructed to provide only the key steps or most essential information in a condensed form.

To systematically assess the quality of responses generated by the analyzed architectures, a dedicated evaluation framework was developed. A set of 70 test questions based on real recruitment correspondence was divided into three levels of complexity: 33 easy, 27 medium, and 10 difficult. The dataset was derived through a preliminary selection process from a larger pool of correspondence items, in which numerous questions addressed identical topics with varied phrasing. Consolidation into thematically distinct questions reduced the pool to the final 70 items used in the evaluation. This distribution reflects the natural and repetitive structure of prospective student inquiries, in which factual single-source questions predominate, while complex multi-document queries are considerably less frequent. Easy questions exclusively required the extraction of information from a single source and assumed a short, factual response (e.g., Is an up-to-date photograph required for the recruitment process?). Medium questions involved the analysis and synthesis of information derived from multiple documents (e.g., By what deadline and in what form must I submit the original of my secondary school leaving certificate?). In contrast, difficult questions were based on information dispersed across multiple documents and required the model to demonstrate reasoning capabilities and the integration of information from multiple sources (e.g., Calculate my recruitment coefficient based on the provided exam results and determine which degree programmes I could be admitted to...). Due to an insufficient number of difficult questions in the original dataset, this category was supplemented with 4 additional items prepared by specialists responsible for the recruitment process at the Faculty, bringing the total number of difficult questions to 10. All queries were processed independently, without taking into account the conversation history, and

the system generated a single response per question, which was then evaluated according to the established criteria.

The quality of the generated responses was assessed using two complementary evaluation criteria, each rated on a five-point scale. The first criterion addressed the factual correctness and completeness of the response with respect to the information available in the knowledge base. The second criterion evaluated the linguistic quality of the response, including its clarity, coherence, and adherence to plain language principles. The evaluation criteria are presented in Table 1.

The evaluation was conducted by two independent domain experts, specialists in undergraduate admissions at WMiIT of the Lublin University of Technology. For each question, the scores assigned by both evaluators were averaged using a simple arithmetic mean. The involvement of independent domain experts as evaluators constitutes a key element of the adopted methodology. It ensures objectivity of the assessments and strong grounding in institutional practice, clearly distinguishing the conducted evaluation from approaches based solely on automatic metrics.

RESULTS AND DISCUSSION

The following section presents the results of the comparative evaluation of three RAG architectures (Naive RAG, Hybrid RAG, and Self-RAG) implemented within the ChatMIT conversational system. The evaluation was conducted using three generative language models of varying parameter scales: Gemma 3 (27B), Llama 3 (70B), and gpt-oss (120B), across three levels of question complexity. The results are reported according to the five-point evaluation scale and are summarized in Table 2 and Table 3.

The Naive RAG architecture achieved satisfactory results exclusively for easy questions, with mean factual correctness scores ranging from 4.2 to 4.3 across all three models. For medium-difficulty questions, the scores dropped substantially to the range of 1.7 to 1.8, and for difficult questions, all three models received the minimum score of 1.0, which in most cases resulted from the system returning no substantive response rather than generating an incorrect one, cf. Table 2. For easy questions, linguistic quality scores ranged from 4.7 to 4.8, while for medium and difficult questions they did not exceed 4.2 and

Table 1. The evaluation criteria of the factual correctness and linguistic quality of the response

Scale for assessing factual correctness	Scale for assessing language quality
1. Information completely incorrect, inconsistent with the factual state, introducing the user into error and preventing proper understanding of the issue.	1. Response incomprehensible, inconsistent, containing numerous language errors that prevent proper reading and interpretation of the content.
2. Information largely incorrect, containing isolated elements consistent with the facts, but the overall message remains incorrect.	2. Response with low language quality, containing numerous stylistic or grammatical errors that significantly hinder understanding of the content.
3. Information partially correct, but incomplete or containing significant gaps or omissions that may lead to ambiguity in interpretation.	3. Response understandable, but containing noticeable stylistic or linguistic issues that reduce the clarity of the message.
4. Information mostly factually correct, containing minor inaccuracies or omissions that do not significantly affect the overall correctness.	4. Linguistically correct and coherent response, fully understandable to the recipient, containing only minor stylistic or visual inconsistencies.
5. All provided information is complete, unambiguous, and fully factually correct, compliant with applicable rules and the current state of knowledge.	5. Clear, coherent, and linguistically correct response, fully adapted to the target audience and using plain language principles appropriately.

Table 2. Average values of the factual correctness assessment of responses for different question difficulty levels

RAG architecture	Model	Easy questions	Medium questions	Difficult questions	Overall average
Naive RAG	27B	4.2	1.8	1.0	2.3
	70B	4.3	1.7	1.0	2.3
	120B	4.3	1.8	1.0	2.4
Hybrid RAG	27B	4.3	2.1	1.8	2.7
	70B	4.4	2.2	1.9	2.8
	120B	4.4	2.2	1.9	2.8
Self-RAG	27B	4.8	4.7	3.6	4.4
	70B	4.8	4.8	4.3	4.6
	120B	5.0	5.0	5.0	5.0

Table 3. Average values of the linguistic quality assessment of responses across different question difficulty level

RAG architecture	Model	Easy questions	Medium questions	Difficult questions	Overall average
Naive RAG	27B	4.7	4.1	4.0	4.3
	70B	4.7	4.2	4.0	4.3
	120B	4.8	4.2	4.0	4.3
Hybrid RAG	27B	4.7	4.1	4.0	4.3
	70B	4.8	4.3	4.0	4.4
	120B	4.8	4.2	4.1	4.4
Self-RAG	27B	4.9	4.8	4.7	4.8
	70B	5.0	4.9	4.7	4.9
	120B	5.0	5.0	5.0	5.0

4.0 respectively, cf. Table 3. Specifically, when the system returned a refusal message instead of a substantive response, a linguistic quality score of 4.0 was assigned, as the brevity of such outputs precluded a comprehensive linguistic evaluation. It is worth noting that since the test set was deliberately constructed to include only questions with answers available in the knowledge base, every

refusal response represented a retrieval failure rather than appropriate system behavior.

The Hybrid RAG architecture produced marginally higher factual correctness scores compared to Naive RAG across all difficulty levels. For easy questions, scores ranged from 4.3 to 4.4, while for medium questions, values of 2.1 to 2.2 were observed. For difficult questions, scores of

1.8 to 1.9 were recorded, slightly above those of Naive RAG, yet still indicating largely incorrect or misleading responses, cf. Table 2. It is worth emphasizing that in the case of medium and difficult questions, the marginal improvement over Naive RAG should be interpreted with caution. While Hybrid RAG more frequently attempted to provide a substantive response, the generated answers were often partially correct or introduced factual inaccuracies, which from the perspective of the end user may be considered more problematic than an explicit refusal to answer. Linguistic quality scores for Hybrid RAG were comparable to those of Naive RAG, ranging from 4.7 to 4.8 for easy questions, 4.1 to 4.3 for medium questions, and 4.0 to 4.1 for difficult questions, cf. Table 3. Analogously to Naive RAG, when the system returned a refusal message, a linguistic quality score of 4.0 was assigned due to the limited scope of such outputs.

Self-RAG demonstrated a marked and consistent performance advantage over both static architectures across all evaluated dimensions. For easy questions, factual correctness scores ranged from 4.8 to 5.0 depending on the model, and for medium questions, from 4.7 to 5.0. Notably, Self-RAG was the only architecture capable of generating correct responses to difficult questions, with scores of 3.6, 4.3, and 5.0 for Gemma 3 (27B), Llama 3 (70B), and gpt-oss (120B), respectively, cf. Table 2. Linguistic quality scores were consistently high across all difficulty levels, ranging from 4.7 to 5.0, reflecting both the general fluency of the models and the positive impact of richer, more precisely targeted context on response coherence and structure, cf. Table 3. Among all evaluated configurations, the combination of the Self-RAG architecture with the gpt-oss (120B) model achieved the highest scores on both evaluation criteria across all question complexity levels, confirming its superiority as the most capable configuration examined in this study. The scores assigned by the two independent evaluators demonstrated a high degree of consistency throughout the evaluation. The level of agreement between the evaluators was quantified using the kappa coefficient [26], yielding a value of $\kappa = 0.82$, which indicates almost perfect agreement between the evaluators. Furthermore, no individual assessment differed by more than one point on the adopted five-point scale, confirming the reliability and stability of the evaluation process.

Regarding the influence of model parameter scale on factual correctness, differences were relatively minor for easy and medium questions across all architectures, where larger models demonstrated only a marginal advantage. For difficult questions, however, and particularly within the Self-RAG architecture, the advantage of larger models became substantially more pronounced. In this context, the gpt-oss (120B) model outperformed its smaller counterparts by a considerable margin, suggesting that tasks requiring multi-step reasoning and the integration of dispersed information benefit significantly from greater model capacity. In terms of linguistic quality, differences between models were similarly small across all difficulty levels. The primary determinant of response quality in this regard was the visual structuring of the output and the clear highlighting of key information, as textual coherence was maintained consistently across all models. Nevertheless, larger models demonstrated a slightly superior ability to organize and present information clearly, particularly for more complex queries. An illustrative example of a response to a medium question related to recruitment documents is presented in Figure 7.

The most consistent finding of this study is the substantial performance advantage of Self-RAG over both Naive RAG and Hybrid RAG across all evaluated configurations. This advantage is most clearly attributable to the adaptive retrieval and generation control mechanisms introduced by the Self-RAG architecture. In contrast to static approaches, in which a fixed number of context fragments is retrieved once per query and passed directly to the generative model, Self-RAG iteratively evaluates the sufficiency of the retrieved context and initiates additional retrieval calls when necessary. This dynamic behavior proved particularly consequential for medium and difficult questions, where a single retrieval pass consistently failed to assemble the dispersed information required for a complete response.

A notable observation is that the inability to adapt retrieval to varying informational demands constitutes a structural limitation inherent to static architectures, one that cannot be compensated by simply increasing model capacity, as the bottleneck remained the insufficiency of the retrieved context rather than the reasoning capacity of the model. Larger models (gpt-oss 120B) did not yield substantially better results than smaller ones (Gemma 3 27B and Llama 3 70B) within

Table of documents required in the recruitment process

ChatMIT Recruitment Today at 22:03
Thought for 13 seconds

Table – Documents required in the recruitment process for undergraduate (full-time) studies at the Faculty of Mathematics and Computer Science, Lublin University of Technology (academic year 2025/2026)

NO.	DOCUMENT	IS IT MANDATORY?*	APPLIES TO (–Y)	FORM / NOTES
1	Personal questionnaire (signed)	✓	All applicants	Generated in the electronic registration system; must be signed and attached to the documents.
2	Secondary school certificate (original)	✓	All applicants	For review – the original must be presented, and a copy placed in the recruitment folder.
3	Copy of secondary school certificate	✓	All applicants	Copy (xerox) attached to the original.
4	Diploma / certificate (for vocational secondary school graduates)	✓ (only for vocational school graduates)	All applicants (when applicant is a vocational school graduate)	Original diploma or certificate confirming vocational qualifications.
5	RODO information clause	✓	All applicants	Generated together with the questionnaire in the recruitment system; required for data processing.
6	Photograph (electronic)	✓	All applicants	Photo with a plain background (size, e.g. headshot). Uploaded via the recruitment system.
7	Medical certificate – no contraindications to study	✓ (only for Computer Science and Information Technology education)	Computer Science and Information Technology education	Examination assesses fitness for intensive study work (e.g. 4th and 5th year), information retrieval and voting.
8	Document confirming entitlements (e.g. laureate/ finalist of olympiads, competitions)	● (optional)	All applicants (where applicable)	Submit a diploma, certificate or other official proof of entitlement.
9	Power of attorney (if documents are submitted by a third party)	● (if applicable)	All applicants	Written power of attorney with candidate's signature.
10	Recruitment fee (confirmation of payment)	✓	All applicants	Proof of payment of PLN 85 to the individual student account (max. 2 days after completing registration).
11	Original and copy of ID document (or other identity document)	● (if identity is in doubt)	All applicants	For review when submitting personal documents.

*** Legend:**

- ✓ – mandatory document in every case.
- – required only in specific situations (e.g. competition entitlements, submission by proxy).

Figure 7. Example Self-RAG response using gpt-oss (120B) for a medium question

the Naive RAG and Hybrid RAG configurations, despite the considerable difference in parameter count. Unlike standalone language models that can fall back on parametric knowledge, RAG systems are fundamentally dependent on the quality and completeness of the retrieved context at inference time, making a well-designed retrieval module the core determinant of system reliability.

Conversely, within the Self-RAG architecture, where sufficient context was assembled through iterative retrieval, model scale became a decisive factor specifically for difficult questions. The gpt-oss (120B) model achieved perfect scores on both evaluation criteria, while smaller models produced partially incorrect or incomplete responses. This pattern suggests that once

the retrieval bottleneck is resolved, advanced reasoning capabilities become a necessary complement to effective retrieval control, particularly when the retrieved context is extensive and demands sophisticated interpretation. These two factors, effective retrieval control and sufficient model capacity, are jointly necessary for reliable performance on complex and demanding information retrieval tasks.

With respect to linguistic quality, differences across all evaluated architectures and models were comparatively minor, reflecting the general ability of modern language models to produce coherent and fluent responses regardless of the underlying retrieval mechanism. For difficult questions, however, larger models demonstrated

superior ability to organize and present complex information clearly, which was reflected in higher linguistic quality scores for the Self-RAG architecture combined with the gpt-oss (120B) model.

A noteworthy methodological contribution of this study is the demonstration of the viability of the zero-shot variant of Self-RAG. In the original formulation, reflection tokens are embedded into the model weights through supervised training on annotated data, requiring the construction of a dedicated labeled dataset and a separate critic model. In the present work, reflection-based evaluation was performed by the generative model itself at inference time through dedicated prompts, without any domain-specific fine-tuning.

The evaluation results confirm that this simplified variant preserved the core adaptive properties of the Self-RAG architecture, including dynamic retrieval control and context quality assessment, at a level sufficient to yield substantially superior performance compared to static RAG approaches. These results demonstrate that the Self-RAG paradigm can be deployed effectively without the substantial annotation effort and computational cost associated with training a dedicated critic model, thereby reducing the barrier to adoption and increasing deployment flexibility.

The reliability of the zero-shot reflection mechanism is however conditional on the generative model correctly assigning reflection tokens in response to evaluation prompts. The model may, for instance, incorrectly mark an irrelevant passage as relevant, or fail to trigger retrieval for a query that clearly requires it. Regular expert review of token assignment quality is therefore a necessary component of any production deployment, and where systematic errors are identified, iterative adjustment of the reflection prompts provides an effective corrective measure. More broadly, these findings reflect a wider trend towards agentic AI architectures, in which systems autonomously manage complex multi-step tasks with minimal human intervention. The increasing maturity of self-correcting and self-evaluating mechanisms suggests a future direction in which human oversight becomes less burdensome, with experts serving primarily as high-level supervisors rather than active participants in each individual decision.

The results of this study also carry a direct practical implication, namely that the selection of a RAG architecture should be guided by the specific requirements of the target application rather

than by a default preference for more complex solutions. For applications in which the majority of user queries are factual and can be answered from a single document fragment, Naive RAG or Hybrid RAG may provide a sufficient and considerably more resource-efficient solution. In the context of the ChatMIT system, the presence of medium and difficult recruitment queries made the adoption of Self-RAG clearly justified, as the additional latency overhead is acceptable when response correctness and completeness take precedence over speed. Notably, even for the most complex queries, response times did not exceed a few seconds within the ChatMIT deployment environment. In other deployment scenarios, careful empirical assessment of query complexity and retrieval requirements should nonetheless precede architectural decisions. Moreover, the highly repetitive and predictable nature of recruitment inquiries creates near-ideal conditions for conversational AI automation, making RAG-based systems particularly well-suited for institutional deployment at scale.

Beyond the specific context of this study, the superiority of Self-RAG reflects a wider trend in which the critical evaluation of information is becoming a fundamental requirement for AI systems deployed in information-sensitive contexts, driven by the exponential growth of available information and the increasing presence of AI-generated content. The key challenge in this regard is not whether to incorporate human-in-the-loop oversight, but how to design workflows that make such involvement practical and scalable. One promising direction is AI-assisted human review, a human-in-the-loop framework in which an automated mechanism performs an initial assessment of response quality and flags cases requiring attention, while a human expert serves as the final verifier, reducing the cognitive burden on reviewers without sacrificing oversight.

Limitations

The evaluation was conducted using a set of 70 test questions, with difficult questions under-represented in the original dataset and partially supplemented by recruitment specialists. While this allowed for a meaningful assessment within the studied domain, the overall size of the test set limits the statistical robustness of the conclusions, particularly for the difficult question category. Furthermore, the system was evaluated

exclusively within a single institutional domain, namely undergraduate admissions at one faculty.

Retrieval parameters, including chunk size, overlap, and the number of retrieved fragments, were deliberately fixed across all evaluated configurations to ensure comparability of results. Although preliminary experiments indicated that parameter modifications did not yield meaningful improvements in retrieval effectiveness, adaptive chunking strategies represent a promising direction for future work. Dynamically adjusting chunk size and granularity based on query characteristics or document structure could further improve retrieval effectiveness, with hierarchical indexing approaches representing a particularly promising option for queries requiring both broad contextual understanding and precise factual extraction.

The system prompt was intentionally unified across all architectures and language models to maintain experimental consistency, which precluded architecture-specific prompt optimization. Although preliminary tests indicated that such adjustments did not produce substantial improvements in response quality, this aspect was not systematically investigated and represents a limitation of the present study. More targeted prompt engineering, particularly for the Self-RAG architecture given its distinct generation control mechanisms, therefore constitutes a promising direction for future research. Furthermore, the integration of significantly larger language models via API represents a natural extension of this work, which is feasible given that the knowledge base consists exclusively of publicly available institutional documents and does not contain any sensitive data.

CONCLUSIONS

The results demonstrate that the choice of RAG architecture is the primary determinant of response quality, exerting a stronger influence than the scale of the language model, particularly for questions of low to medium complexity. Naive RAG and Hybrid RAG achieved satisfactory performance exclusively for easy questions, with both architectures failing consistently on medium and difficult queries due to the structural limitation of static, single-pass context retrieval. The marginal improvement of Hybrid RAG over Naive RAG, while observable, should be interpreted with caution, as partially correct responses may mislead users more than an explicit refusal to answer.

Self-RAG demonstrated substantially superior performance across all evaluated dimensions and was the only architecture capable of generating correct responses to difficult questions. Its superiority stems from the iterative retrieval mechanism and reflection-based generation control, enabling the system to dynamically adapt to the varying informational demands of different query types. These findings confirm that the critical evaluation of retrieved context is a decisive factor in RAG system design. While architectural choice proved to be the primary determinant of response quality, model scale emerged as a decisive complementary factor within the Self-RAG architecture, particularly for questions of higher complexity. The combination of Self-RAG with the gpt-oss (120B) model achieved perfect scores on both evaluation criteria across all question complexity levels, confirming its suitability for production deployment in information-sensitive conversational systems.

A key methodological finding of this study is that the zero-shot variant of Self-RAG retains the core adaptive properties of the original architecture, including dynamic retrieval control and context quality assessment, at a substantially reduced deployment complexity. This finding has implications for the broader adoption of Self-RAG in domain-specific applications, as it eliminates the need for supervised training of a dedicated critic model and allows reflection-based evaluation to be applied to any capable generative model without additional fine-tuning. The reliability of this approach remains however conditional on regular expert review of reflection token assignments, with iterative prompt adjustment serving as an effective corrective measure where systematic errors are identified.

The findings of this study underscore that architectural decision-making in RAG-based conversational systems should be guided by a careful assessment of query complexity distribution, the informational structure of the knowledge base, and available computational resources, rather than by a default preference for either simplicity or sophistication.

REFERENCES

1. Manning CD. An Introduction to Information Retrieval 2004;53:462–3.
2. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of Tricks for Efficient Text Classification BT

- Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: 2017; 2: 427–31.
3. Zhixiang, Xu, Chen M, Weinberger KQ, Sha F. An alternative text representation to TF-IDF and Bag-of-Words 2013.
4. Askari A, Abolghasemi A, Pasi G, Kraaij W, Verberne S. Injecting the BM25 Score as Text Improves BERT-Based Re-rankers. *Lect Notes Comput Sci (Including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 2023;13980 LNCS:66–83. https://doi.org/10.1007/978-3-031-28244-7_5
5. Mitra B, Craswell N. An introduction to neural information retrieval. *Found Trends Inf Retr* 2018;13:1–129. <https://doi.org/10.1561/15000000061>
6. Rong X. word2vec Parameter Learning Explained 2016.
7. Pennington J, Socher R, Manning CD. GloVe: Global vectors for word representation. *EMNLP 2014 - 2014 Conf Empir Methods Nat Lang Process Proc Conf* 2014:1532–43. <https://doi.org/10.3115/v1/d14-1162>
8. Devlin J, Chang M-W, Lee K, Google KT, Language AI. BERT: Pre-training of Deep bidirectional transformers for language understanding. *Naacl-Hlt 2019* 2018:4171–86.
9. Zhou K, Junyi L, Tianyi T, Hou Y, Zhao WX, Li J, Tang T, Wang X, Min Y, Zhang B, Zhang J, Dong Z, Du Y, Yang C, Chen Y, Chen Z, Jiang J, Ren R, Li Y, Tang X, Liu Z, Liu P, Nie J-Y, Wen J-R. A survey of large language models. *ArXiv Prepr ArXiv230318223* 2023;1:1–125.
10. Zhou Z, Li C, Chen X, Wang S, Chao Y, Li Z, Wang H, An R, Shi Q, Tan Z, Han X, Shi X, Liu Z, Sun M. LLMs\$times\$MapReduce: Simplified long-sequence processing using large language models. *Computer Science*, 2024. <https://doi.org/10.48550/arXiv.2410.09342>
11. Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, Dai Y, Sun J, Wang H, Wang H. Retrieval-augmented generation for large language models: A survey. *ArXiv Prepr ArXiv231210997* 2023;2:32.
12. Nawaz N, Gomes AM. Artificial intelligence chatbots are new recruiters. *Int J Adv Comput Sci Appl* 2019;10:1–5. <https://doi.org/10.14569/ijacsa.2019.0100901>
13. Winkler R, Soellner M. Unleashing the potential of chatbots in education: A state-of-the-art analysis. *Acad Manag Proc* 2018;2018:15903. <https://doi.org/10.5465/ambpp.2018.15903abstract>
14. Nurshatayeva A, Page LC, White CC, Gehlbach H. Are artificially intelligent conversational chatbots uniformly effective in reducing summer melt? Evidence from a Randomized Controlled Trial. *Res High Educ* 2021;62:392–402. <https://doi.org/10.1007/s11162-021-09633-z>
15. Okonkwo CW, Ade-Ibijola A. Chatbots applications in education: A systematic review. *Comput Educ Artif Intell* 2021;2. <https://doi.org/10.1016/j.caeai.2021.100033>
16. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih WT, Rocktäschel T, Riedel S, Kiela D. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv Neural Inf Process Syst* 2020;2020-Decem.
17. Arora D, Kini A, Chowdhury SR, Natarajan N, Sinha G, Sharma A. GAR-meets-RAG paradigm for zero-shot information retrieval. *Computer Science* 2023.
18. Ma X, Gong Y, He P, Zhao H, Duan N. Query Rewriting for Retrieval-Augmented Large Language Models. *EMNLP 2023 - 2023 Conf Empir Methods Nat Lang Process Proc* 2023:5303–15. <https://doi.org/10.18653/v1/2023.emnlp-main.322>
19. Borgeaud S, Mensch A, Hoffmann J, Cai T, Rutherford E, Millican K, van den Driessche G, Lespiau JB, Damoc B, Clark A, de Las Casas D, Guy A, Menick J, Ring R, Hennigan T, Huang S, Maggiore L, Jones C, Cassirer A, Brock A, Paganini M, Irving G, Vinyals O, Osindero S, Simonyan K, Rae JW, Elsen E, Sifre L. Improving language models by retrieving from trillions of tokens. *Proc Mach Learn Res* 2022;162:2206–40.
20. Sun Z, Wang X, Tay Y, Yang Y, Zhou D. Recitation-Augmented Language Models. *11th Int Conf Learn Represent ICLR 2023* 2023.
21. Xia M, Zhang X, Couturier C, Zheng G, Rajmohan S, Rühle V. Hybrid-RACA: Hybrid Retrieval-Augmented Composition Assistance for Real-time Text Prediction. *EMNLP 2024 - 2024 Conf Empir Methods Nat Lang Process Proc Ind Track* 2024:120–31. <https://doi.org/10.18653/v1/2024.emnlp-industry.11>.
22. Doan NN, Härmä A, Celebi R, Gottardo V. A hybrid retrieval approach for advancing retrieval-augmented generation systems. *Assoc Comput Linguist* 2024:397–409.
23. Asai A, Wu Z, Wang Y, Sil A, Hajishirzi H. Self-Rag: Learning To Retrieve, Generate, and Critique Through Self-Reflection. *12th Int Conf Learn Represent ICLR 2024* 2024.
24. Huang Y, Huang J. A survey on retrieval-augmented text generation for large language models *ACM Computing Surveys*, 2024;58(12):1–38. <https://doi.org/10.48550/arXiv.2404.10981>
25. Yan Z-Q, Shi H, Chen Z, Yan M, Chen J, Li X-Y, Tu K. Corrective retrieval augmented generation. *ArXiv Prepr ArXiv240115884* 2024:1–18.