International Journal of Education Humanities and Social Science

ISSN: 2582-0745
Vol. 8, No. 04; 2025

ENHANCING ENGLISH LANGUAGE EDUCATION WITH RETRIEVAL-AUGMENTED GENERATION (RAG): A NEW PARADIGM FOR AI-POWERED TEACHING AND LEARNING

Scott Smith
Hankuk University of Foreign Studies. 107 Imun-ro, Dongdaemun-gu, Seoul, 02450, South Korea.
E-Mail: smith.hufs@gmail.com

https://doi.org/10.54922/IJEHSS.2025.1058

Received: 28 June 2025/Published: 2 August 2025

ABSTRACT
Retrieval-Augmented Generation (RAG) serves as a significant natural language processing (NLP) advancement through its integration of real-time information retrieval with traditional large language models (LLMs). RAG surpasses traditional LLMs because it uses real-time information retrieval to generate responses which leads to better factual accuracy and reduced hallucinations and improved contextual understanding. This paper traces RAG's development from its beginnings in early information retrieval (IR) systems through to the implementation of transformer-based architectures. The applications of RAG include question-answering systems and AI-driven content generation and enterprise-level data processing. The research identifies essential challenges which include retrieval accuracy and computational cost and privacy concerns and multimodal data integration and biases in retrieved information. Furthermore, it discusses emerging solutions to address these limitations through fact-verification models and blockchain-based data integrity systems and quantum computing for optimized retrieval and multimodal AI alignment. After all, the future development of RAG architectures will result in self-learning decentralized ethically governed AI systems which adapt to dynamic knowledge bases across multiple domains. This in-depth exploration establishes a comprehensive understanding of RAG's function in knowledge-intensive AI applications while outlining directions to enhance its operational efficiency and scalability and ethical reliability. In doing so, it shows of RAG will new framework for AI-powered learning both in and outside the classroom.

Keywords: Retrieval-augmented generation, natural language processing, large language models, artificial intelligence, generative and adaptive learning AI, context-aware systems.

1. INTRODUCTION

1.1 Origins, History, and Evolution of Retrieval-Augmented Generation

RAG functions as an advanced artificial intelligence (AI) model structure which combines retrieval-based AI with generative AI (Lewis et al., 2020). The retrieval-based AI system dynamically retrieves external knowledge from databases and online repositories and proprietary datasets (Karpuhkin et al., 2020). Through this process the model acquires the most recent and contextually appropriate information. The processed retrieved data allows generative AI to create coherent responses that stay contextually aware and informative. RAG achieves significantly improved quality, accuracy, and reliability of AI-generated content through its combination of retrieval and generation methods. RAG uses real-world data to ground its outputs which results in better factual consistency and reduced hallucinations compared to traditional generative models

ISSN: 2582-0745
Vol. 8, No. 04; 2025

that depend on pre-trained knowledge, which may become outdated or contain gaps (Shuster, Waller & Weston, 2020). The system demonstrates its highest value in applications that require recent knowledge updates such as question-answering systems and academic research, as well as customer support and legal or medical AI assistants (Guu et all., 2020). RAG emerged through major advancements in IR systems and NLP models. The development of retrieval-based and generative AI fusion depended on earlier technologies, which established the foundation for contextually accurate and effective responses (Chen et al., 2017).

2. PRECURSOR TECHNOLOGIES

<u>Theoretical basis for computer science and artificial intelligence</u> (1930s-1950s)

The historical importance of Alan Turing for modern AI techniques such as RAG remains significant even though he passed away before the systems were created. Turing established the theoretical basis for computer science and AI during the 1930s–1950s period. Turing presented his machine model of computation in 1936 which established the criteria for determining computational functions (Turing, 1936). The fundamental concept he introduced serves as the basis for all present-day computing operations which include the sophisticated machine learning frameworks that operate RAG systems. During World War II Turing worked at Bletchley Park where he helped break the German Enigma code (Turing, 1950). His work there involved statistical reasoning and pattern recognition which represented early probabilistic thinking that later became fundamental to AI models that use retrieval and generation methods (Copeland, 2004; Hodges, 2014). Turing introduced the Turing Test in 1950 as a standard for machine intelligence, which established the philosophical foundation for RAG-based systems to achieve human-like interaction (Russell & Norvig, 2010). The intellectual and computational foundations which Information Retrieval systems and RAG systems were built upon were established by Turing through his contributions before the 1970s–1990s timeframe. Turing's legacy consists of making intelligent retrieval reasoning and generation possible for machines rather than directly creating RAG (Manning, Raghavan & Schütze, 2008).

<u>Natural language processing and early information science</u> (1950s-1970s)

The period from 1950 to 1970 witnessed the practical development of artificial intelligence alongside natural language processing and early information science (McCorduck, 2004). The 1950s researchers started developing their first formal AI programs, which included the Logic Theorist (1956), the first artificial intelligence program that was designed to prove mathematical theorems by using symbolic logic and heuristic search techniques (i.e. It marked the birth of AI as a field) and the General Problem Solver (1957), which was also developed as a universal problem-solving machine that aimed to mimic human though processes and reasoning, to create symbolic processing systems that simulated human reasoning (Newell & Simon, 1961). The 1956 Dartmouth Conference marked the beginning of AI as a formal discipline when John McCarthy and Marvin Minsky joined other pioneers like Allen Newell, Herbert A. Simon, Clifford Shaw, and J.C. Shaw to develop practical theories about machine intelligence (Newell & Simon, 1956). From this, the formal grammar systems developed by Noam Chomsky, starting in 1957, shaped early computational language approaches which became essential for both retrieval and generation operations in RAG (Chomsky, 1957). After that, the 1960s witnessed the emergence of initial information retrieval systems through Gerard Salton's SMART system at Cornell which developed vector space models and term weighting principles that remain fundamental to modern RAG systems for document relevance scoring (Salton, 1968). In other words, it was a **pioneering**

ISSN: 2582-0745
Vol. 8, No. 04; 2025

information retrieval system that laid the foundation for modern search engines and text retrieval methods. Moving forward, the 1966 introduction of ELIZA (one of the earliest NLP programs created by Joseph Weizenbaum at MIT) demonstrated how pattern matching and script-based rules could generate human-like conversations through its early natural language processing capabilities as ELIZA was a rule-based chatbot that mimicked a conversation with a human (i.e. it used **simple pattern-matching and substitution rules**, with no actual understanding of the content) (Salton & Lesk, 1968). The period transitioned from theoretical foundations to experimental beginnings by establishing the fundamental elements (i.e. reasoning, retrieval, and language understanding) which RAG would unite through deep learning technology (Weizenbaum, 1966).

<u>Information retrieval systems</u> (1970s–1990s)

The first AI-powered information retrieval systems functioned to find and order documents that matched user search queries (Robertson & Waller, 1994). During this time researchers developed two fundamental techniques: Term Frequency-Inverse Document Frequency (TF-IDF) and BM25 which transformed search ranking through document relevance analysis based on keyword occurrence and distribution patterns (i.e. they were used to rank how relevant a document is to a user's search query and are fundamental in search engines and file searches) (Chowdhury, 2004). The models served as fundamental components for powering early search engines including AltaVista and initial Google versions through efficient document retrieval (Brin & Page, 1998). The systems functioned only as retrieval systems because they produced document lists for users to manually analyze instead of generating responses.

<u>Large-scale text analysis infrastructure and the World Wide Web</u> (1990s-2000s)

The fundamental elements of RAG started to develop into practical and expandable systems during the 1990s to 2000s period before the term itself was established. During this time period, researchers achieved substantial progress in IR systems and NLP alongside the development of large-scale text analysis infrastructure which serves as fundamental elements for RAG (Salton & Buckley, 1988). During the 1990s IR systems evolved through TF-IDF and vector space models (wherein documents and queries are represented as points (vectors), and **similarity between them is measured geometrically**), which enabled computers to evaluate and sort documents according to their relevance to user queries. These systems established the foundation for contemporary dense retrieval operations in RAG. During this period NLP research transitioned from rule-based systems to statistical and machine learning-based methods (i.e. **information retrieval both queries and documents are encoded into dense vector representations** and relevance is determined by comparative analysis using dot product and cosine similarity – a way of measuring how similar two vectors are in direction and magnitude and a **normalized version of the dot product** that focuses **only on the angle** between two vectors, not their magnitude) (Jurafsky & Nartin, 2000). The World Wide Web brought about a massive increase in digital text data accessibility which became essential for training and benchmarking large language models throughout the subsequent decades (Church & Mercer, 1993). The early 2000s brought probabilistic models Latent Semantic Analysis and Latent Dirichlet Allocation (i.e. t**opic modeling techniques** used in **NLP** to uncover latent or hidden patterns or topics in large collections of text), which enabled machines to extract document context and topics better. These are techniques that RAG systems would use in their retrieval phase (Biel, NG & Jordan, 2003). This phase marked the deliberate integration of retrieval systems with NLP pipelines especially in question answering systems which evolved into modern RAG models (Lesk, 1997; Callan, 2000). Although deep learning had not entered the scene yet the technical unification of search and

ISSN: 2582-0745
Vol. 8, No. 04; 2025

language understanding had started its development which led to the neural RAG architectures of the late 2010s.

Neural language models and word embeddings (2000s–2020)

The early 2000s brought significant progress to NLP through increased computational power and evolving machine learning techniques. Google's Word2Vec (2013) and Stanford's GloVe (2014), methods that turn words into numbers (otherwise known as **word embeddings**) so that computers can understand and work with them in tasks like translation, search, or chatbot conversations (i.e. instead of using words like "cat" and "dog," computers need numbers and these tools help convert words into useful numeric forms and Long Short-Term Memory (LSTM) networks, a special type of **recurrent neural network** designed to **learn from sequences of data,** emerged as key developments which substantially enhanced AI capabilities for word relationship understanding and semantic and contextual processing (Mikolov et al., 2013). The vectorized word representation systems Word2Vec and GloVe extracted detailed word meanings and their semantic connections through analysis of large corpora co-occurrence patterns (i.e. patterns showing **which words tend to appear together frequently** in a large body of text. These patterns help reveal **semantic relationships** between words) (Pennington, Socher & Manning, 2014). The LSTMs enabled better sequential text processing which allowed AI systems to detect longer language dependencies. The models delivered substantial NLP application enhancements yet their inability to retrieve external real-time information prevented them from generating factually accurate responses (Hochreiter & Schmidhuber, 1997).

The initial AI developments created essential building blocks which eventually merged into RAG models to connect static knowledge production with dynamic context-sensitive retrieval. The following AI development with transformer architectures and large-scale language models enabled the smooth combination of retrieval and generation which resulted in the creation of RAG.

Deep learning together with transformer architectures (i.e. the **foundation for most modern NLP** systems, to include **ChatGPT, Gemini, Perplexity, etc.)** brought a revolutionary change to NLP, which enhanced both language understanding and generation abilities. The transformer models introduced by Vaswani et al. (2017) became the defining feature of this era because they replaced traditional recurrent and convolutional neural networks with self-attention mechanisms that allowed faster parallel processing of textual data (i.e. they eliminated recurrence entirely, relied solely on self-attention mechanisms, and were more parallelizable and efficient than previous versions). The breakthroughs in technology led to the development of state-of-the-art pretrained language models which revolutionized how AI systems handle text processing and generation.

The first major advancement in this period was Bidirectional Encoder Representations from Transformers (BERT), which introduced bidirectional contextual embedding which allowed models to understand words in the context of an entire sentence rather than in isolation (Devlin et al., 2019). This significantly improved text retrieval, making AI more adept at ranking and retrieving relevant documents based on nuanced linguistic patterns. BERT and its variants became foundational in many search engines, question-answering systems, and NLP applications. Meanwhile, the development of Generative Pretrained Transformers (GPT-2 and GPT-3, released in 2019 and 2020, respectively) showcased impressive text generation capabilities, enabling models to produce human-like responses based on vast amounts of pre-trained knowledge (Radford et al., 2019). However, these models had inherent limitations:

ISSN: 2582-0745
Vol. 8, No. 04; 2025

- Knowledge cutoff. Since GPT models were trained on static datasets, they lacked the ability to incorporate newly emerging facts or dynamically update their knowledge base.
- Hallucinations. Without external fact-checking, these models often generated plausible but incorrect or misleading information, making them unreliable for high-stakes applications.
- Lack of dynamic retrieval. GPT models operated purely as generative AI without retrieval capabilities, meaning they could not fetch relevant external documents to support or validate their outputs (Lewis et al., 2020).

These limitations highlighted the need for a more advanced approach that could combine generative AI with real-time information retrieval (Borgeaud, 2022). This realization ultimately led to the development of RAG, which integrates external knowledge retrieval with text generation to enhance accuracy, factual consistency, and real-world applicability (Chen et al., 2017). By leveraging retrieval mechanisms, RAG overcomes the pitfalls of purely generative models, ensuring that AI-generated responses are informed by up-to-date and authoritative sources rather than relying solely on pre-trained knowledge (Izacard & Grave, 2021).
Facebook AI Research launched the RAG framework in 2020 as a revolutionary solution to boost AI performance in knowledge-intensive NLP tasks (Roberts, Raffel & Shazeer, 2020). The paper "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks" by Lewis et al. (2020) introduced RAG as a solution to previous generative models' limitations through knowledge retrieval integration with text generation. The new approach delivered substantial improvements to AI response accuracy together with reliability and contextual understanding.

3. KEY INNOVATIONS OF RAG

1. Combining retrieval and generation

RAG introduced the capability to merge retrieval-based AI with generative AI which allows models to fetch appropriate external knowledge before producing responses (Lewis et al., 2020). The system operates through two essential components.

- A retriever model, such as Dense Passage Retrieval (DPR), which searches a large corpus of documents (e.g., Wikipedia, academic databases) to identify the most relevant information.
- A generator model, often based on transformer architectures like GPT, which incorporates the retrieved knowledge into its response, ensuring factual accuracy and contextual richness.

2. Enhancing accuracy and reducing hallucinations

Traditional generative models, such as GPT-3, suffered from hallucinations, which is a tendency to produce confident but incorrect or misleading information (Karpuhkin et al. 2020). RAG mitigates this issue by grounding its responses in real-time, verifiable data rather than relying solely on static, pre-trained knowledge (Maynez, et al., 2020; Ji et al., 2023). RAG ensures that AI-generated outputs remain factually grounded, up-to-date, and contextually relevant by integrating retrieved documents into the generative process.

ISSN: 2582-0745
Vol. 8, No. 04; 2025

3. Handling open-domain and knowledge-intensive queries

RAG demonstrates its best performance in open-domain question-answering and knowledge-intensive tasks because static language models fail to deliver accurate responses in these areas (Chen et al., 2017). RAG achieves significant improvements in AI applications through its live retrieval functionality in the following areas:

- Search engines, enhancing document ranking, and contextual understanding
- Chatbots and virtual assistants that enable real-time, fact-based conversations
- Academic research and enterprise applications, where access to the latest information is crucial (Zhao, Saleh & Zhang, 2022)

4. RAG ARCHITECTURE (2020)

The original RAG framework consisted of two core components:

- Retriever. A neural model, such as Dense Passage Retrieval (DRP), that efficiently searches vast text corpora to find relevant documents (i.e. a method for retrieving relevant documents (passages) from a large corpus using **dense vector representations** rather than keyword matching like traditional Term Frequency-Inverse Document Frequency (TF-IDF), a **statistical measure** that evaluates how important a word is in a document **relative to a whole collection (corpus), and Best Marching 25,** a **probabilistic ranking function** that improves on TF-IDF by handling **term saturation** and adjusting for **document length).**
- Generator. A transformer-based language model, such as GPT, that synthesizes and contextualizes the retrieved knowledge into a well-formed response (Lewis et al, 2020).

RAG uses this architecture to fetch external sources dynamically for its generative process which differentiates it from traditional pretrained models that only use static data. An example use case is the Open-Domain Question-Answering (i.e. a task in NLP where a system must **answer questions using information from a large, unstructured knowledge source,** such as Wikipedia or the internet, **without being limited to a fixed context or document**). An AI system using RAG allows users to submit real-world knowledge questions such as, “Who won the 2024 Nobel Prize in Literature?” (Han Kang)

- Retrieval component. The model searches Wikipedia, scientific databases, and authoritative sources for the latest information.
- Generation component. The model produces a response using retrieved facts to provide real-world knowledge that avoids outdated training data information.

The system's ability to fetch and generate responses dynamically makes it a significant advancement for AI knowledge applications while establishing RAG as the base technology for future fact-based real-time AI systems (Izacard & Grave, 2021).

RAG emerged from Facebook AI Research (FAIR) in 2020 before quickly advancing past academic research to become widely adopted across industries (Zhang et al., 2022). Organizations use RAG-powered architectures to develop real-world AI applications which deliver enhanced accuracy alongside faster responses and more accurate factual information (Demmer-Fushman, Roberts & Voorhees, 2020). RAG has become a cornerstone technology because of improved retrieval methods and model fine-tuning and scalable infrastructure development for modern AI assistants and search engines and knowledge-intensive applications.

ISSN: 2582-0745
Vol. 8, No. 04; 2025

Major tech companies started implementing RAG-based systems into their AI-powered products and services during 2021 which resulted in multiple advancements:

Enterprise AI assistants. Microsoft (Copilot), OpenAI (ChatGPT) and Google (Bard/Gemini) implemented RAG to deliver fact-based answers which extended beyond traditional model training.

Search engines. Bing Chat and Perplexity use RAG to access real-time information which helps decrease misinformation and enhance answer accuracy.

Medical and legal AI. The AI systems of PubMed and LexisNexis use RAG to help experts and professionals find authoritative domain-specific information in healthcare and legal fields.

Customer Support Chatbots. The AI chatbots from Intercom and Zendesk use RAG technology to deliver context-specific responses in customer service operations which decreases dependence on basic FAQ systems.

5. KEY IMPROVEMENTS (2021–2023)

Several innovations have enhanced the efficiency and effectiveness of RAG-based architectures, including hybrid retrieval methods. AI models now combine neural retrieval techniques such as DPR and Contextualized Late Interaction over BERT (i.e. a **dense retrieval architecture** designed for **fast and accurate document retrieval** using deep contextual embeddings; introduced to address the trade-off between **speed** and **accuracy** in neural information retrieval) with traditional BM25 search algorithms, which improves precision and recall, balancing deep contextual understanding with efficient keyword-based search (Johnson, Douze & Jégou, 2019).

Scaling knowledge sources. Organizations have adopted vector databases like Pinecone (Pinecone Technologies), Weaviate (Semi Technologies), and FAISS (Facebook AI Similarity Search) to optimize retrieval speed and scalability, and these databases store high-dimensional embeddings, enabling fast and accurate retrieval from vast datasets, including web documents, enterprise knowledge bases, and academic literature (Khattab & Zaharia, 2020).

Fine-tuning generative models for RAG. Custom GPT and Large Language Model Meta AI models have been trained for domain-specific retrieval, ensuring high-quality responses in specialized fields, such as finance, legal research, and scientific analysis, which has allowed these models to better align with specific organizational datasets, enhancing relevance and trustworthiness (Lin, Ma & Ma, 2021, Zhou, Gao & Zgang, 2022; Touvron et al., 2023).

6. GPT-4 AND MODERN RAG (2023–PRESENT)

AI models have advanced further to strengthen RAG's ability to improve real-time factual accuracy in the following ways:

GPT-4 (OpenAI), Claude (Anthropic), and Gemini now incorporate retrieval-augmented techniques to fetch up-to-date information, significantly improving their reliability in dynamic contexts.

Perplexity uses RAG-powered web retrieval to dynamically fetch and synthesize external data which results in accurate well-cited responses.

The ongoing growth of AI systems depends on RAG-based architectures to maintain factual correctness while reducing misinformation and enabling real-time interaction with constantly evolving knowledge sources. The expansion represents a fundamental transformation in AI development because it moves away from static model training to implement live dynamically retrieved intelligence across various industries (Lewis et al., 2020).

ISSN: 2582-0745
Vol. 8, No. 04; 2025

RAG stands as a crucial framework which will enhance artificial intelligence models by improving their accuracy and reliability and adaptability. It functions as a fundamental foundation for future AI development because it connects pre-trained data with real-time information retrieval to bridge the gap between static knowledge and dynamic information (Mialon et al., 2023). Research teams and organizations dedicate their efforts to developing RAG capabilities through multi-modal retrieval and decentralized verification and fine-tuned domain-specific applications.

7. WHAT'S NEXT FOR RAG?

1. Multi-modal RAG. The future of RAG technology will bring retrieval capabilities that extend beyond text data to include images videos and audio content. In other words, future models will integrate knowledge from different sources including scientific diagrams video lectures podcasts and structured datasets to produce more detailed AI-generated responses. The evolution holds special importance for AI-driven research and medical diagnostics and content creation because these fields require essential multi-modal data integration (Alayrac et al., 2022; Zeller et al., 2022).
2. Decentralized RAG systems. The main issue with AI-generated content is determining its trustworthiness and verifiability. Future RAG architectures will implement blockchain-based verification systems to track information retrieval while maintaining tamper-proof authentication and traceability (Rebuffel et all., 2021) The decentralized RAG system enables peer-reviewed transparent knowledge retrieval which stops misinformation spread and enhances AI credibility in scientific legal and financial applications (Kshetri, 2021). The method supports current trustworthy AI development efforts which strengthen RAG models against bias and censorship and data manipulation threats.

3. Fine-tuned retrieval models. The development of industry-specific RAG models will transform knowledge-intensive fields like medicine and law and finance and academia through the implementation of highly customized domain-specific retrieval systems (Wadden et al., 2020). These specialized models will receive fine-tuning to extract peer-reviewed medical studies from PubMed and legal precedents from LexisNexis and financial reports from Bloomberg datasets for expert-level insights. This fine-tuning process will enhance the accuracy and dependability and practical use of AI-generated responses in vital decision-making situations (Zho, Gao & Zhang, 2022).

8. WHY RAG IS THE FUTURE OF AI

The adoption of RAG is accelerating as industries recognize its advantages over traditional LLMs (Large Language Models). The following factors make RAG the backbone of next-generation AI systems. First, by retrieving real-time information instead of relying solely on pre-trained static knowledge, RAG significantly reduces hallucinations and misinformation, making it ideal for high-stakes applications. Second, unlike conventional models with a fixed knowledge cutoff, RAG can dynamically fetch and synthesize the latest information, ensuring AI remains current and contextually aware. Third, RAG's hybrid retrieval-generation approach is highly adaptable to enterprise needs, whether for automated research assistants, legal case analysis, medical diagnosis, or AI-driven search engines (ji et al., 2023).

As AI systems continue to integrate with real-world knowledge sources, RAG will play a pivotal role in defining the future of AI-powered intelligence (Lewis, et al, 2020). With advancements in

ISSN: 2582-0745
Vol. 8, No. 04; 2025

multi-modal capabilities, decentralized verification, and fine-tuned retrieval, RAG is set to become the gold standard for AI-driven knowledge retrieval and generation in the years to come.
In short, RAG represents a paradigm shift in AI, which addresses hallucination issues in LLMs and provides factual and up-to-date responses.

9. HOW RETRIEVAL-AUGMENTED GENERATION (RAG) WILL BE TRANSFORMATIVE FOR ENGLISH LANGUAGE TEACHING

The educational revolution brought by AI will transform English language education through RAG technology, which will reshape lesson planning and student assessment and personalized learning delivery (Khalil & Ebner, 2021). RAG differs from traditional AI models because it uses real-time information retrieval to enhance AI-generated content which results in lessons that better match student needs and learning contexts. The next 5-10 years will bring RAG advancements that include real-time lesson customization and multimodal learning and AI-driven assessments and adaptive tutoring to create an efficient personalized and engaging language learning environment.

AI-Powered Debate & Discussion Tools

The current AI-generated debate prompts fail to adjust in real time which results in unoriginal and uninteresting content. The RAG system will enable future AI debate partners to fetch actual arguments and speeches and news content for creating dynamic live debate simulations. The platform will enable teachers to create discussion prompts from contemporary global issues which will lead students to participate in meaningful and intellectually stimulating dialogues (Zhao, Saleh & Zhang, 2022). The AI system will modify debate topics according to student vocabulary abilities and personal interests to maintain discussions at an appropriate level of difficulty.

Instant AI translation and multilingual feedback

The implementation of AI-powered real-time translation technology will improve understanding for students in multilingual classrooms. The future context-aware AI translation systems will differ from present translation tools because they will access subject-specific explanations and modify them according to student proficiency levels (Woo et al, 2016). The AI system will provide grammar explanations through the student's native language while using culturally appropriate examples to explain English tenses and phrasal verbs and idioms.

Automated AI feedback on writing and speaking

The grading process of essays and feedback delivery by teachers will receive AI assistance through RAG-driven analysis. Future AI-powered writing assistants will access high-quality real-world writing examples to compare student work with professional models while delivering immediate constructive feedback (Lu & Ai, 2018). AI will listen to spoken presentations to retrieve native speaker examples for side-by-side pronunciation comparison which enables students to improve their speech clarity and fluency.

AI-powered storytelling and creative assistance

The task of encouraging creativity in writing proves difficult because students face challenges when developing their ideas and organizing their work. The writing assistants powered by RAG will access historical and literary references along with cultural content to assist students in creating more detailed and accurate stories. AI will provide story prompts that stem from real-world themes while generating alternative plot developments to match student interests (Xie, Yu & Cheng, 2022).

RAG represents the future direction of AI-based English teaching methods. RAG surpasses traditional LLMs through its enhanced accuracy and adaptability and dynamic nature which

ISSN: 2582-0745
Vol. 8, No. 04; 2025

positions it as a transformative tool for English language education (Shuster et al., 2022). Real-time knowledge retrieval capabilities allow AI-generated content to stay current and maintain cultural awareness and contextual understanding (Kasneci et all, 2023). RAG will lead AI-powered tool development toward new innovations because of its advanced capabilities:

- Real-time lesson planning and quiz generation
- Personalized student learning paths and adaptive tutoring
- Multimodal AI-powered activities (text, video, audio, images)
- AI-generated writing models and automated feedback
- Live AI debate partners and discussion tools
- Instant, context-aware translation and grammar explanations

AI-driven RAG applications will develop student-specific interactive learning experiences during the next decade which will reduce teacher workload and improve educational quality (Gao et al, 2023). RAG will transform English language instruction through its real-time retrieval system and AI adaptation capabilities to deliver personalized and effective learning experiences (Wang, Yin & Wang, 2021)

10. IMPLEMENTING RETRIEVAL-AUGMENTED GENERATION (RAG) WITH NO-CODE AI TOOLS IN ENGLISH LANGUAGE TEACHING

The educational revolution brought by artificial intelligence (AI) can be advanced through RAG, which improves lesson planning and assessment and student engagement in English language teaching (Lewis et al., 2020). Educators who want to use AI without programming expertise can use no-code AI tools to automate tasks and personalize learning experiences and optimize instructional strategies (Khalil & Ebner, 2021). The following guide demonstrates how to use no-code AI tools for English language teaching through lesson creation and reading comprehension and assessment and pronunciation and writing assistance and interactive student engagement.

<u>Step 1: Selecting no-code AI tools for English language teaching</u>

The implementation of AI in English instruction requires educators to choose platforms that use RAG for dynamic content retrieval and are ready for immediate use. The following tools (among so many that are out there right now) represent essential components for the educational process:

- <u>ChatGPT</u> and <u>Gemini</u> provide teachers with lesson planning capabilities and writing prompts.
- <u>Perplexity</u> retrieves articles instantly for reading comprehension purposes.
- <u>Quillionz</u> generates quizzes and assessments through its platform.
- <u>Grammarly</u> and vocabulary enhancement tool Ludwig Guru operate as part of the educational platform.
- <u>ELSA Speak</u> enables students to practice pronunciation through its automated platform.
- <u>Writesonic</u> provides students with assistance for creative and academic writing tasks.
- <u>Dialogflow</u> enables automated student question and answer responses.

The selected tools (among so many these days) enable teachers to perform repetitive tasks automatically while generating interactive educational materials and delivering immediate personalized feedback to students without programming knowledge.

<u>Step 2: AI-powered lesson planning</u>

AI tools like <u>ChatGPT</u> and <u>Gemini</u> produce customized lesson plans instantly through their ability to draw from real-world contexts and current events and student learning levels. AI systems require the following setup to create lesson plans:

ISSN: 2582-0745
Vol. 8, No. 04; 2025

1. Open ChatGPT or Gemini to input a particular lesson planning prompt which could be something like, “make a 45-minute lesson plan about phrasal verbs for students at the intermediate English level that includes warm-up activities and explanation sections and exercise tasks and discussion segments”.
2. Paste the AI-created plan into either Ms Word or Google Docs.
3. Modify the lesson to meet the requirements of your context.

The output quality of the paid versions of ChatGPT and Gemini will obviously provide more accurate results with deeper contextual understanding, but the basic versions should suffice in most cases for lessons like this example.

Step 3: AI-enhanced reading comprehension

AI-powered reading tools enable students to read real-world texts while developing their critical reading abilities. Perplexity provides teachers with a method to implement reading lessons through the following steps:

1. Enter your search term in typing, for example, “the latest news article about climate change for high school students." or "short article on artificial intelligence for intermediate learners"
2. Choose an article that matches your needs and get the link.
3. Paste the article into ChatGPT or Quillionz to generate comprehension questions with a prompt like, "create ten comprehension questions for this article."
4. Students could receive reading assignments through Google Docs or Google Classroom, posting files on a teacher’s website, or simply by receiving printed handouts in class.

The additional feature of Quillionz allows users to create automatic quizzes and comprehension tests.

Step 4: AI-powered quiz and assessment generation

The traditional assessment methods remain fixed but AI-powered tools enable quizzes to modify their content based on student answers while providing authentic real-world examples for better understanding. The following steps explain how to develop AI-powered quizzes using Quillionz:

1. Insert reading materials or student writing assignments
2. Select question types (multiple-choice, true/false, open-ended, etc.).
3. Users can either create their own quiz questions or let the system generate them automatically.
4. The quiz can be exported to Google Forms for automatic grading or printed for classroom distribution and discussion.

The Google Forms auto-grading system enables self-assessment quizzes, which valuably decreases teacher workloads.

Step 5: AI for grammar and vocabulary enhancement

Student writing receives immediate grammar-checking and vocabulary-building assistance through AI tools that demonstrate real-world usage examples. Grammarly serves as a tool for students to check their grammar through the following steps:

1. Students can either upload their essays or type their text directly into it.
2. Examine the AI-produced corrections which address grammar rules and improve clarity and tone (among other things).
3. Obtain the corrected document through download and then offer feedback to students.

The following steps explain how to utilize Ludwig Guru for vocabulary and sentence structure improvement:

1. Enter a phrase or sentence to check correctness. For example, the original sentence, "She suggested me to go" becomes "She suggested that I go" according to Ludwig.

ISSN: 2582-0745
Vol. 8, No. 04; 2025

2. The platform shows actual examples of usage taken from literary works and academic publications and newspapers.
3. The teacher should ask students to modify their sentences based on Ludwig's output results.

You could also give sentence correction tasks based on Ludwig's suggested corrections.

Step 6: AI for pronunciation and speaking practice

AI-powered pronunciation tools enable students to improve their speech clarity and fluency through immediate feedback. The following steps explain how to implement ELSA Speak for pronunciation improvement:

1. Download the app.
2. Create pronunciation exercises that match lesson material, such as "past tense verbs".
3. The app allows students to speak while ELSA delivers immediate pronunciation feedback.
4. The teacher dashboard enables monitoring of student progress.

Teachers could implement ELSA Speak as a speaking exercise during classroom warm-up sessions.

Step 7: AI-assisted writing support

The following steps explain how to use Writesonic for AI-powered writing:

1. Select a writing template (e.g. essay, email, or story writing).
2. Enter a topic prompt, such as “the essay should explore bilingualism benefits”.
3. Customize the generated text and provide structured feedback.

The combination of ChatGPT and Writesonic provides students with grammar and style improvement suggestions for their written work.

Step 8: AI chatbots for student Q&A

Students can use AI chatbots to request grammar and vocabulary and literature assistance at any time outside regular class hours. Teachers can set up an AI chatbot via Dialogflow as follows:

1. Create a new chatbot agent it.
2. The AI requires training data that includes typical student inquiries such as, "what is the past tense of 'run'?" or "please define the distinction between 'affect' and 'effect'."
3. The chatbot system can be integrated into Google Classroom for deployment, which should receive additional programming to provide literature analysis and essay feedback capabilities.

Step 9: Automate and scale AI integration

Educators can use AI tools to automate repetitive tasks and expand AI-powered teaching after integration:

- Students can use Google Forms to receive automatic quiz grading.
- Teachers can use Google Classroom to schedule writing prompts that AI creates.
- Students can receive automated feedback on their essays through Grammarly's AI system.
- Teachers can use ChatGPT to obtain immediate lesson planning suggestions.

The implementation of no-code AI tools in English language teaching brings a revolutionary change to classroom instruction which allows teachers to improve teaching efficiency and student personalization and engagement without needing advanced technical abilities. Teachers can use AI-driven RAG models to automate lesson planning and grading which cuts down administrative work while enabling them to spend more time directly with students (Holmes, Bialik & Fadel, 2019). AI-powered real-time reading comprehension activities enable students to access modern texts that match their language abilities and cognitive development through contextually relevant materials. The adaptive AI feedback system provides students with instant personalized assessment of their work which automatically adjusts both assessment methods and instructional approaches according to their individual learning patterns. The AI-based interactive exercises deliver immediate phonetic analysis and correction which helps students improve their oral communication skills through data-driven AI-enhanced pronunciation models. AI-assisted retrieval of culturally relevant vocabulary and writing examples helps maintain authentic and context-sensitive language instruction that supports students from diverse backgrounds and real-world applications (Xie, Yu & Cheng, 2022). The advancement of RAG technology will make no-code solutions essential for educational empowerment because they provide teachers with AI-based tools to create dynamic learning experiences that meet various educational requirements and boost student engagement and accuracy and accessibility in English language instruction.

11. THE PATH FORWARD FOR RETRIEVAL-AUGMENTED GENERATION (RAG) IN ENGLISH LANGUAGE EDUCATION

The educational potential of RAG to transform English language education exists through its ability to enhance lesson planning together with student assessments and reading comprehension and writing feedback and language acquisition. To reach its maximum potential RAG needs to solve multiple accuracy and retrieval efficiency problems alongside multimodal integration and data reliability issues as well as ethical concerns and personalized learning difficulties (Ji et al., 2023). The educational success of AI-driven educational tools depends on future advancements in fact verification and optimized retrieval speed alongside multimodal AI alignment and bias mitigation and long-term AI memory capabilities for student-centered and context-aware learning in classroom and digital settings (Kazemi & Ragan, 2021).

RAG faces a critical challenge when applied to English language education due to the necessity of verifying content accuracy and reliability from retrieved sources. The generation of AI responses depends on multiple texts which might contain outdated information or misleading content or pedagogically unsuitable material (Zhang, Zhao & Wang, 2021). The future development of RAG models will include fact-verification systems that validate instructional content through academic source cross-references and language corpora and research-based pedagogical frameworks (Preskill, 2018). Teachers will feel secure when they depend on AI-generated lesson plans and grammar explanations and vocabulary lists derived from verified educational databases. AI developers will create confidence scoring systems which evaluate the trustworthiness of reading materials to direct students toward suitable high-quality texts. An AI-based language learning system should learn to select peer-reviewed ESL research and contemporary literature and authoritative dictionary definitions above other resources which lack credibility or present simplified information.

Current RAG models in English education face a major challenge because of their slow retrieval speeds which affect their computational efficiency during student real-time activities (Alayrac et al., 2022). Student learning requires instant feedback within language learning environments especially during automated writing evaluation and grammar correction and pronunciation training sessions. The current AI retrieval models need time to process their responses which results in delays that negatively affect the learning process (Zellers et al, 2022). The advancement of quantum computing and search algorithm optimization will enhance AI response times to support real-time operation of AI tutors and writing assistants. An AI writing assistant system should provide users with immediate grammar corrections alongside structural advice and contextual vocabulary suggestions while avoiding system delays. Speech training programs utilizing AI will

ISSN: 2582-0745
Vol. 8, No. 04; 2025

be able to compare student voice recordings against native speaker examples within fast timeframes to offer detailed pronunciation feedback.

The growing multimedia integration in English instruction demands multimodal RAG alignment for AI-based educational systems to function properly (Xie et al, 2019). Future multimodal retrieval systems will expand beyond text retrieval to incorporate images, audio, video and interactive exercises which will develop more engaging language lessons. The AI tools will gain the ability to retrieve original audio recordings and TED Talks and news reports and infographics and interactive stories which they will match to reading exercises and speaking activities. An English teacher preparing a persuasive speech lesson can leverage RAG to acquire a suitable political debate video while generating discussion questions and accessing a readable transcript. AI-powered listening exercises would find real-world conversations then tailor them for different student abilities before making personalized comprehension questions.

12. CONCLUSION: ETHICAL AI, BIAS MITIGATION, AND PERSONALIZED LEARNING IN ENGLISH LANGUAGE EDUCATION

AI retrieval systems in English education pose a major issue because they can perpetuate biased language models. AI systems acquire knowledge from human-created content which leads to the reinforcement of linguistic biases together with cultural stereotypes and limited language use perspectives (Bender & Friedman, 2018). Future RAG models will implement bias-detection algorithms which will check retrieved texts for biased content and outdated language standards and cultural insensitivity. AI-based ESL platforms will learn from multiple linguistic sources which will make language examples and grammar explanations represent global English usage instead of following regional standards (Blodgett et al, 2020). Students who use an AI writing assistant to enhance sentence structure and vocabulary will have access to both British and American English variations so they can select the most suitable terms according to their target audience needs. The teaching materials generated by AI will showcase multiple linguistic and cultural viewpoints to stop the continuation of Eurocentric language norms in English classroom instruction.

RAG-powered English education will advance through two major AI developments: personalization and long-term memory features (West, Whuttaker & Crawford, 2019). The present AI tools produce answers through individual cases while failing to maintain previous student interactions and learning needs. The future generation of RAG models will adopt dynamic AI memory functions which enable language tutors to access and adjust their instruction based on student interactions (Galloway & Rose, 2015). An AI tutor working with ESL writing students would analyze student-specific grammar weaknesses and vocabulary shortages and sentence organization patterns to create exercise materials based on historical performance data. The AI system uses phonetic challenge tracking to develop pronunciation exercises which help students improve their fluency. AI-powered English education will adapt through long-term strategies to provide customized learning experiences that go beyond pre-set teaching materials.

The future of RAG-powered AI in English language education will be shaped by four main characteristics: enhanced accuracy and efficiency along with multimodal capabilities and personalized learning experiences (Lee, Warschauer & Lee, 2019). The system will advance from basic text retrieval to become a unified platform which combines listening, speaking, reading and writing resources at the same time while keeping ethical standards intact (Zhou et al., 2022). AI-powered tools will achieve more reliable and inclusive student-centered learning environments

ISSN: 2582-0745
Vol. 8, No. 04; 2025

through fact-checking API advancements and blockchain-based content verification and bias-reducing algorithms (Mialon et al., 2023). The advancement of RAG models requires educators and AI developers to maintain human oversight and ethical considerations together with strict evaluations of AI-generated materials to guarantee AI benefits English language learners without harming linguistic diversity or cultural representation or pedagogical integrity.

REFERENCES

[1] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hassid, S., & Zisserman, A. (2022). Flamingo: A visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.
[2] Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, *6*, 587–604.
[3] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
[4] Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5476).
[5] Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., & Rae, J. W. (2022). *Improving language models by retrieving from trillions of tokens*. In *Proceedings of the 39th International Conference on Machine Learning* (pp. 2206–2240).
[6] Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, *30*(1–7), 107–117.
[7] Callan, J. (2000). Distributed information retrieval. In W. B. Croft (Ed.), *Advances in information retrieval* (pp. 127–150). Springer.
[8] Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (pp. 1870–1879).
[9] Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access*, *8*, 75264–75278.
[10] Chomsky, N. (1957). *Syntactic structures*. Mouton.
[11] Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.
[12] Chowdhury, G. G. (2004). *Introduction to modern information retrieval* (2nd ed.). Facet Publishing.
[13] Church, K. W., & Mercer, R. L. (1993). Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, *19*(1), 1–24.
[14] Copeland, B. J. (2004). *The essential Turing: Seminal writings in computing, logic, philosophy, artificial intelligence, and artificial life, plus the secrets of Enigma*. Oxford University Press.
[15] Demner-Fushman, D., Roberts, K., & Voorhees, E. M. (2020). *Biomedical question answering using knowledge-rich and retrieval-augmented models*. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 1615–1618).
[16] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171–4186).

[17] Galloway, N., & Rose, H. (2015). *Introducing global Englishes*. Routledge.
[18] Gao, C., Yao, B., Liu, L., Yu, Z., & Sun, M. (2023). *AI in education: A systematic review of its application and impact*. *Educational Technology Research and Development*, 71, 1–22.
[19] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.
[20] Hodges, A. (2014). *Alan Turing: The enigma* (Updated ed.). Princeton University Press.
[21] Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Boston, MA: Center for Curriculum Redesign.
[22] Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 874–880).
[23] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38.
[24] Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, *7*(3), 535–547.
[25] Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall.
[26] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6769–6781).
[27] Kasneci, E., Sessler, K., Bittermann, M., Schausberger, L., & Huber, T. (2023). *ChatGPT for good? On opportunities and challenges of large language models for education*. *Learning and Instruction*, 86, 101789.
[28] Kazemi, M., & Ragan, E. D. (2021). AI and credibility: Designing trustworthy information systems. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15.
[29] Khalil, M., & Ebner, M. (2021). AI-powered digital assistants in education: A systematic review. *International Journal of Educational Technology in Higher Education*, *18*(1), 1–34.
[30] Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 39–48).
[31] Kshetri, N. (2021). Blockchain and trustworthy artificial intelligence. *IT Professional*, *23*(3), 72–76.
[32] Lee, H., Warschauer, M., & Lee, J. H. (2019). The effects of automated writing evaluation on ESL writing: A meta-analysis. *Language Learning & Technology*, *23*(1), 36–76.
[33] Lesk, M. (1997). Practical digital libraries: Books, bytes, and bucks. Morgan Kaufmann.
[34] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in neural information processing systems* (Vol. 33, pp. 9459–9474).
[35] Lin, J., Ma, X., & Ma, H. (2021). A few brief notes on BM25 and neural ranking: Considerations, challenges, and future directions. *Information Retrieval Journal*, *24*(1), 1–7.
[36] Liu, R., Salter, D., & Tolochko, P. (2022). AI-based formative feedback for academic writing: Opportunities and challenges. *British Journal of Educational Technology*, *53*(6), 1557–1572.
[37] Lu, X., & Ai, H. (2018). *Automated writing evaluation: A review of the development, technologies, and application*. *Language Testing*, 35(4), 473–492.

ISSN: 2582-0745
Vol. 8, No. 04; 2025

[38] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.
[39] Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1906–1919).
[40] McCorduck, P. (2004). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence* (2nd ed.). A. K. Peters.
[41] Mialon, G., Kondratyuk, D., Lewis, P., Cohan, A., Riedel, S., & Izacard, G. (2023). Augmented language models: A survey. *arXiv preprint arXiv:2302.07842*.
[42] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
[43] Newell, A., & Simon, H. A. (1956). The logic theory machine—A complex information processing system. *IRE Transactions on Information Theory*, *2*(3), 61–79. https://doi.org/10.1109/TIT.1956.1056810
[44] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).
[45] Preskill, J. (2018). Quantum computing in the NISQ era and beyond. *Quantum*, *2*, 79.
[46] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.
[47] Rebuffel, C., Cianflone, A., Vardasbi, S., Glaude, J., Lamprier, S., & Gallinari, P. (2021). Challenges and applications of decentralized machine learning. *Applied Sciences*, *11*(11), 5181.
[48] Roberts, A., Raffel, C., & Shazeer, N. (2020). *How much knowledge can you pack into the parameters of a language model?* In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 5418–5426).
[49] Robertson, S. E., & Walker, S. (1994). Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 232–241).
[50] Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd ed.). Prentice Hall.
[51] Salton, G. (1968). *Automatic information organization and retrieval*. McGraw-Hill.
[52] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523.
[53] Salton, G., & Lesk, M. E. (1968). Computer evaluation of indexing and text processing. *Journal of the ACM*, *15*(1), 8–36.
[54] Shuster, K., Ju, D., Roller, S., & Weston, J. (2022). *Retrieval-augmented generation for knowledge-intensive NLP tasks: A survey*. *arXiv preprint arXiv:2204.06974*.
[55] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., & Scao, T. L. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
[56] Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *2*(42), 230–265.
[57] Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460.

[58] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (Vol. 30).
[59] Wadden, D., Lin, S., Lo, K., Wang, L. L., & Hajishirzi, H. (2020). Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 7534–7550).
[60] Wang, H., Yin, Y., & Wang, J. (2021). *Intelligent tutoring systems in the AI era: A review of core components and educational functions*. *Computers & Education: Artificial Intelligence*, 2, 100017.
[61] Weizenbaum, J. (1966). ELIZA: A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.
[62] West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems: Gender, race, and power in AI. *AI Now Institute*. https://ainowinstitute.org/discriminatingsystems.html
[63] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., & Dean, J. (2016). Google's Neural Machine Translation System: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.
[64] Xie, H., Chu, H. C., Hwang, G. J., & Wang, C. C. (2019). Trends and development in technology-enhanced adaptive/personalized learning. *Interactive Learning Environments*, *27*(8), 979–999.
[65] Xie, H., Yu, C., & Cheng, H. N. H. (2022). *AI-powered language learning: A review of current applications and future directions*. *British Journal of Educational Technology*, 53(4), 857–876.
[66] Zellers, R., Lu, X., Yu, T., & Choi, Y. (2022). MERLOT Reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16375–16387).
[67] Zhang, X., Zhao, S., & Wang, H. (2021). Real-time inference in NLP models: Challenges and solutions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
[68] Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., & Dolan, B. (2022). *Automatic response generation for customer service using retrieval-augmented models*. *arXiv preprint arXiv:2201.12140*.
[69] Zhao, W., Saleh, M., & Zhang, K. (2022). Fact-grounded dialogues with retrieval-augmented generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (pp. 1656–1671).
[70] Zhou, Q., Wu, J., Yang, S., Li, Z., & Ren, Z. (2022). RETRO: Retrieval-enhanced transformer with long-term memory.
[71] Zhou, W., Gao, L., & Zhang, Y. (2022). Domain-specific generative retrieval for scientific knowledge. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (pp. 1234–1245).