

Reducing Hallucinations in Large Language Models through External Knowledge Retrieval

Ansh Mittal

Abstract

Large Language Models (LLMs) have rapidly transformed conversational artificial intelligence by enabling systems to generate fluent, contextually relevant, and human-like responses across a wide range of applications. However, their increasing use in information-intensive domains has highlighted the persistent problem of hallucination, in which models generate inaccurate, unsupported, or fabricated information. This issue is particularly significant in educational and career guidance, where incorrect information concerning university admissions, scholarships, visa regulations, immigration pathways, and career opportunities can influence important student decisions. This secondary review examines the causes of hallucinations in LLMs and evaluates existing approaches for improving factual reliability, with particular emphasis on external knowledge retrieval and Retrieval-Augmented Generation (RAG). The review discusses fine-tuning, Reinforcement Learning from Human Feedback, prompt engineering, semantic search, embedding models, vector databases, knowledge graphs, and hybrid retrieval architectures. It further examines evaluation approaches based on retrieval performance, factual correctness, relevance, faithfulness, and human assessment. The literature indicates that retrieval-based approaches can reduce dependence on the static knowledge encoded within LLM parameters by connecting models with external and potentially updated information sources. However, retrieval does not completely eliminate hallucinations, as response reliability remains dependent on the quality, relevance, completeness, and recency of retrieved information. Based on these findings, the paper proposes a conceptual framework for a domain-specific conversational assistant integrating curated educational and immigration data, document processing, semantic retrieval, vector databases, LLM-based generation, and source-aware user interaction. The review concludes that combining language generation with reliable external knowledge represents a promising direction for developing more accurate, transparent, and trustworthy AI systems for higher education and career decision-making.

Keywords: Large Language Models, hallucination, Retrieval-Augmented Generation, external knowledge retrieval, semantic search, vector databases, knowledge graphs, generative AI, educational technology, conversational AI

1. Introduction

1.1 Background

Large Language Models (LLMs) have transformed artificial intelligence by enabling machines to understand and generate human-like language at an unprecedented scale. Built on the Transformer architecture introduced by Vaswani et al. (2017), these models learn linguistic patterns from massive text datasets and can perform tasks such as question answering, translation, summarization, coding assistance, and conversational interaction. Successive developments, including GPT-3 (Brown et al., 2020), GPT-4, BERT (Devlin et al., 2019), and other advanced language models, have demonstrated remarkable

improvements in language understanding, making generative AI a widely adopted technology across industries.

The rapid adoption of conversational AI has expanded its applications beyond traditional technology sectors into education, healthcare, finance, customer support, research, and government services. In education, AI-powered assistants are increasingly being used for personalized tutoring, academic writing support, college admissions guidance, scholarship discovery, career counseling, and interview preparation (Kasneci et al., 2023). Students seeking higher education abroad frequently rely on AI tools to compare universities, understand admission requirements, explore visa regulations, and evaluate post-study work opportunities. This growing dependence highlights the importance of ensuring that AI-generated information is both accurate and trustworthy.

Unlike conventional search engines that retrieve existing documents, LLMs generate responses by predicting the most probable sequence of words based on patterns learned during training (Brown et al., 2020). While this capability produces fluent and contextually relevant answers, it also introduces a significant limitation: models may produce information that appears convincing despite being factually incorrect. As AI systems become integrated into high-stakes decision-making environments, improving factual reliability has become a major research priority.

1.2 Problem Statement

One of the most significant challenges facing Large Language Models is the phenomenon known as hallucination, where the model generates fabricated, inaccurate, or unsupported information with high confidence (Ji et al., 2023). Hallucinations may involve inventing references, misrepresenting policies, confusing facts across different sources, or presenting outdated information as current knowledge. These errors occur because LLMs generate text probabilistically rather than verifying facts against external knowledge sources during response generation.

The consequences of hallucinations become particularly serious in domains where users make important life decisions based on AI-generated advice. For example, students applying to international universities may receive incorrect information about eligibility criteria, scholarship deadlines, standardized testing requirements, or application procedures. Similarly, inaccurate guidance regarding visa regulations, immigration policies, post-study work permits, or career pathways can lead to financial losses, missed opportunities, and poor decision-making. Unlike casual conversational errors, these inaccuracies can have long-term educational and professional consequences.

Traditional approaches such as prompt engineering, supervised fine-tuning, and Reinforcement Learning from Human Feedback (RLHF) improve model behavior but do not guarantee factual correctness because the underlying knowledge remains static once training is completed (Ouyang et al., 2022). Consequently, researchers have increasingly explored methods that enable language models to retrieve relevant external information before generating responses. These retrieval-based approaches seek to combine the language generation capabilities of LLMs with continuously updated knowledge sources, reducing the likelihood of unsupported claims.

1.3 Research Objectives

This secondary review paper examines current research on reducing hallucinations in Large Language Models through external knowledge retrieval mechanisms. The study specifically aims to:

- Understand the primary causes of hallucinations in Large Language Models.
- Examine external knowledge retrieval techniques used to improve factual accuracy.

- Review Retrieval-Augmented Generation (RAG) architectures, semantic search methods, and vector database technologies.
- Analyze evaluation metrics used to assess hallucination reduction, factual correctness, and response quality.
- Propose a conceptual framework for developing domain-specific educational advisory systems that provide reliable guidance on higher education, scholarships, visa regulations, and career pathways.

1.4 Research Questions

To achieve these objectives, the review addresses the following research questions:

- Why do hallucinations occur in Large Language Models?
- What techniques have been developed to reduce hallucinations?
- How do retrieval-based architectures improve factual accuracy?
- What are the strengths and limitations of existing hallucination mitigation approaches?
- How can retrieval-based AI systems support the development of trustworthy educational question-answering platforms for students seeking higher education guidance?

Addressing these questions provides a foundation for understanding how external knowledge retrieval can enhance the reliability of conversational AI systems. The findings contribute to ongoing discussions on building domain-specific AI assistants that combine advanced language generation with verifiable, up-to-date information, particularly in educational and career counseling environments where factual accuracy is essential.

2. Literature Review

2.1 Evolution of Large Language Models and Generative AI

The development of Large Language Models (LLMs) represents one of the most significant advances in artificial intelligence. A major breakthrough occurred with the introduction of the Transformer architecture, which replaced recurrent neural networks with a self-attention mechanism capable of processing entire sequences simultaneously (Vaswani et al., 2017). This innovation improved computational efficiency and enabled models to capture long-range contextual relationships within text. Subsequent models, including BERT (Devlin et al., 2019) and the GPT series (Brown et al., 2020), demonstrated remarkable improvements in natural language understanding and generation. While BERT focused on bidirectional language understanding, GPT models emphasized autoregressive text generation, enabling conversational AI systems capable of producing coherent and context-aware responses. These developments accelerated the adoption of AI assistants across education, software development, healthcare, and customer support (Bommasani et al., 2021).

Despite their strengths, LLMs remain heavily dependent on the quality, diversity, and recency of their training data. Once training is completed, their internal knowledge becomes static unless updated through additional training or external retrieval mechanisms (Brown et al., 2020).

2.2 Understanding Hallucinations in Large Language Models

Hallucination refers to the generation of information that appears plausible but lacks factual support or contradicts reliable evidence (Ji et al., 2023). Researchers distinguish between intrinsic hallucinations, where responses contradict the provided context, and extrinsic hallucinations, where fabricated information is introduced beyond the available evidence (Ji et al., 2023).

Several factors contribute to hallucinations. First, training datasets may contain incomplete, biased, or outdated information. Second, LLMs generate text by predicting probable word sequences rather than

verifying facts, allowing confident but incorrect statements to emerge (Brown et al., 2020). Third, static knowledge prevents models from incorporating newly published regulations or events. Finally, context-window limitations may cause models to lose important information during lengthy conversations.

Hallucinations are particularly concerning in educational applications. Incorrect advice regarding university admissions, scholarships, or visa policies can directly influence students' decisions. Researchers also identify a trade-off between fluency and factual accuracy, where highly coherent responses may appear trustworthy even when they contain factual errors (Ji et al., 2023).

2.3 Existing Approaches to Hallucination Reduction

Multiple techniques have been proposed to reduce hallucinations in LLMs.

Fine-tuning adapts pretrained models to domain-specific datasets, improving performance in specialized tasks but requiring significant computational resources and repeated updates (Devlin et al., 2019).

Reinforcement Learning from Human Feedback (RLHF) aligns model behavior with human preferences by rewarding desirable responses, improving conversational quality and safety without guaranteeing factual correctness (Ouyang et al., 2022).

Prompt engineering, including chain-of-thought and few-shot prompting, guides models toward better reasoning through carefully designed instructions (Wei et al., 2022). However, prompting cannot fully eliminate fabricated information because the underlying knowledge remains unchanged.

External knowledge retrieval has emerged as one of the most promising solutions. Instead of relying solely on internal model memory, retrieval systems access external documents during response generation, allowing outputs to be grounded in verifiable information (Lewis et al., 2020).

Approach	Advantages	Limitations
Fine-tuning	Better domain performance	Expensive and difficult to update
Prompt engineering	Simple implementation	Limited reliability
RLHF	Improves alignment	Does not guarantee factuality
External retrieval	Dynamic and verifiable	Retrieval quality dependent

Overall, retrieval-based approaches provide greater adaptability than purely model-centric methods.

2.4 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) combines information retrieval with text generation to improve factual accuracy (Lewis et al., 2020). Rather than generating answers solely from model parameters, RAG first retrieves relevant documents before producing a response.

The workflow consists of four stages: the user submits a query, a retriever identifies relevant documents, retrieved content is injected into the model's context, and the generator produces a grounded response. This architecture enables models to reference current and domain-specific information without requiring complete retraining.

Research demonstrates several advantages of RAG, including reduced hallucination, access to updated knowledge, and improved source grounding (Gao et al., 2023). However, retrieval quality remains a critical challenge. Incorrect document selection, increased latency, and limited context windows may still reduce response quality, highlighting the importance of efficient retrieval systems.

2.5 Semantic Search and Embedding Models

Traditional keyword search relies on exact word matching, often failing when users express similar ideas using different vocabulary. Semantic search addresses this limitation by converting text into dense vector representations that capture contextual meaning (Reimers & Gurevych, 2019).

Sentence-BERT (SBERT) significantly improved sentence-level similarity tasks by generating efficient semantic embeddings suitable for retrieval applications (Reimers & Gurevych, 2019). OpenAI embedding models further extend semantic retrieval by representing documents and queries within shared vector spaces.

Feature	Keyword Search	Semantic Search
Exact match	High	Moderate
Context understanding	Low	High
Scalability	High	High
Synonym handling	Weak	Strong

Semantic search therefore plays a central role in modern retrieval-augmented systems by improving document relevance.

2.6 Vector Databases and Information Storage

Efficient retrieval depends on specialized vector databases capable of storing and searching millions of embeddings. Technologies such as FAISS (Johnson et al., 2019), Pinecone, ChromaDB, and Weaviate enable fast similarity search using Approximate Nearest Neighbor (ANN) algorithms.

These systems also support metadata filtering, allowing searches to be restricted by categories such as country, university, or document type. While vector databases provide scalability and low-latency retrieval, their performance depends heavily on embedding quality and requires careful storage management for large datasets.

2.7 Knowledge Graphs and Structured Knowledge Representation

Knowledge graphs organize information through entities and relationships, making them valuable for representing structured educational information. Universities, degree programs, visa policies, and career pathways can be connected through ontologies that improve explainability and reasoning (Hogan et al., 2021).

Compared with purely text-based retrieval, knowledge graphs provide transparent relationship mapping. However, constructing and maintaining comprehensive graphs requires substantial manual effort and continuous updates.

2.8 Hybrid Retrieval Architectures

Recent research increasingly combines multiple retrieval techniques within hybrid architectures. These systems integrate vector search, knowledge graphs, metadata filtering, and LLM generation to improve retrieval quality (Gao et al., 2023).

Hybrid systems offer higher factual accuracy and better interpretability because structured knowledge complements semantic retrieval. Nevertheless, integrating multiple components increases computational complexity and system maintenance requirements.

2.9 Evaluation Metrics for Hallucination Reduction

Researchers evaluate hallucination reduction using both automated and human-centered metrics. Traditional measures such as Precision, Recall, F1-score, Exact Match, BLEU, and ROUGE assess retrieval and text generation performance (Papineni et al., 2002; Lin, 2004).

However, automated metrics alone cannot fully capture factual reliability. Human evaluation remains essential for assessing relevance, faithfulness, correctness, and explainability (Ji et al., 2023). Current research increasingly emphasizes the need for domain-specific evaluation frameworks tailored to educational and advisory applications.

2.10 Applications of Retrieval-Based AI Systems in Education

Retrieval-based AI systems are becoming increasingly important in educational support. Applications include academic advising, university selection, scholarship guidance, visa information, and career pathway recommendations. Platforms such as Duolingo Max use generative AI for language learning, while Khanmigo provides AI-assisted tutoring grounded in educational content.

Many universities have also introduced virtual assistants that answer administrative questions and provide student support. Nevertheless, existing systems often rely on generic knowledge bases and lack specialized repositories containing verified admissions policies, scholarship databases, and immigration regulations. This limitation reinforces the growing need for domain-specific retrieval architectures capable of delivering accurate, explainable, and continuously updated guidance for students pursuing higher education abroad.

3. Research Gap

Recent research has made significant progress in addressing hallucinations in Large Language Models (LLMs), particularly through Retrieval-Augmented Generation (RAG), semantic search, vector databases, and knowledge-grounding techniques (Lewis et al., 2020; Gao et al., 2023). These approaches have demonstrated measurable improvements in factual accuracy by enabling models to retrieve relevant external information before generating responses. Similarly, advances in embedding models, knowledge graphs, and reinforcement learning have contributed to improving the reliability and usability of conversational AI systems across various domains (Reimers & Gurevych, 2019; Ouyang et al., 2022). Despite these developments, several important research gaps remain, particularly in applications related to educational and career guidance.

A major limitation in the existing literature is the limited focus on domain-specific educational advisory systems. Most retrieval-based research evaluates performance using general-purpose benchmarks, open-domain question answering datasets, or encyclopedic knowledge sources (Ji et al., 2023). While these benchmarks provide valuable insights into overall model performance, they do not adequately represent the complex information needs of students seeking guidance on university admissions, scholarships, visa regulations, immigration policies, and post-study career opportunities. These domains involve frequently changing policies, institution-specific requirements, and country-specific regulations that require continuously updated and verifiable information.

Another important gap is the insufficient integration of structured and unstructured educational data within conversational AI systems. Existing RAG implementations primarily retrieve information from unstructured text documents, while structured resources such as university databases, scholarship eligibility criteria, visa categories, and degree-to-career relationships are often managed separately (Hogan et al., 2021). Combining these diverse knowledge sources into a unified retrieval framework remains a

relatively underexplored area. Such integration could improve both factual accuracy and explainability by allowing AI systems to present information alongside its supporting sources and relationships.

Additionally, current evaluation methods place considerable emphasis on automated metrics such as BLEU, ROUGE, and Exact Match, which may not fully capture the reliability required in educational decision-making contexts (Papineni et al., 2002; Lin, 2004). Greater attention is needed toward domain-specific evaluation frameworks that assess factual correctness, source transparency, policy compliance, and user trust.

Therefore, this review seeks to bridge these gaps by synthesizing existing research on hallucination reduction while proposing a conceptual framework for a retrieval-based conversational assistant specifically designed for higher education guidance. By integrating semantic search, vector databases, metadata filtering, and curated educational repositories, such systems have the potential to deliver more accurate, explainable, and trustworthy support for students making important academic and career decisions.

4. Proposed Conceptual Framework

The proposed conceptual framework presents a domain-specific Retrieval-Augmented Generation (RAG) system designed to reduce hallucinations in Large Language Models when providing educational and career-related guidance. The framework combines curated educational data, structured knowledge representation, semantic retrieval, and LLM-based response generation. Its primary purpose is to ensure that responses provided to students are grounded in relevant and current external information rather than being generated exclusively from the model's pretrained knowledge. This approach is particularly relevant for international education, where admission requirements, scholarships, visa regulations, and post-study work policies may differ across countries and change over time.

The framework consists of five interconnected layers: the Data Layer, Knowledge Layer, Retrieval Layer, Generation Layer, and User Interaction Layer. Each layer performs a distinct function while contributing to the overall objective of producing accurate, relevant, and explainable responses.

4.1 Data Layer

The Data Layer forms the foundation of the proposed system. It consists of a curated repository containing information relevant to students making higher education and career decisions. The repository may include university profiles, degree programs, admission requirements, tuition fees, scholarships, visa regulations, immigration pathways, career trajectories, and frequently asked questions.

University information could include institutional rankings, locations, available programs, eligibility requirements, application deadlines, and standardized testing requirements. Course-level information could include curriculum structure, duration, specialization options, and prerequisites. Similarly, scholarship information could contain eligibility criteria, application deadlines, funding levels, and nationality restrictions.

The inclusion of visa and immigration information is particularly important because these policies can change frequently. Therefore, the repository should include information obtained from reliable and authoritative sources, such as university websites and official government portals. Each record should also contain information regarding its source and date of verification. This would allow the system to distinguish between current and potentially outdated information.

The repository could additionally include career trajectories connecting academic qualifications with possible career outcomes. For example, a Computer Science degree could be associated with careers such

as software engineering, cybersecurity, data science, or artificial intelligence. Such relationships would allow the system to provide more meaningful career-oriented recommendations.

4.2 Knowledge Layer

The Knowledge Layer converts the information stored in the Data Layer into a format that can be efficiently processed and retrieved by the AI system. One of the first steps is document chunking, in which lengthy documents are divided into smaller sections. Chunking enables the retrieval system to identify specific passages relevant to a user's question rather than providing an entire document to the language model.

The next step involves generating embeddings. Embedding models convert text into numerical vector representations that capture semantic meaning (Reimers & Gurevych, 2019). Queries and documents can subsequently be compared according to their semantic similarity.

Metadata tagging further improves retrieval accuracy. Information can be tagged according to attributes such as country, university, degree, subject area, document type, and date of publication or verification. For example, a search concerning Computer Science programs in Canada could be restricted to Canadian universities and recent admission documents. This combination of semantic embeddings and metadata can reduce the retrieval of irrelevant information.

4.3 Retrieval Layer

The Retrieval Layer acts as the connection between the student's question and the external knowledge repository. When a user submits a query, the system first interprets its meaning and converts it into an appropriate representation for retrieval.

Semantic search enables the system to identify documents based on meaning rather than relying exclusively on exact keyword matches. The relevant embeddings are compared within a vector database, and the most similar documents or passages are retrieved. Vector databases and similarity-search technologies can efficiently identify relevant information from large collections of documents (Johnson et al., 2019).

Metadata filtering can be incorporated into this stage to improve precision. For instance, if a student asks about engineering programs in Germany, the retrieval process can prioritize documents tagged with Germany, engineering, undergraduate study, and current admission information.

The effectiveness of the entire framework depends substantially on retrieval quality. If irrelevant, incomplete, or outdated information is retrieved, the language model may still generate an inaccurate response. Consequently, retrieval should be evaluated independently in addition to evaluating the final generated answer.

4.4 Generation Layer

The Generation Layer uses an LLM to formulate a response based on the information retrieved from the external repository. Instead of relying entirely on its pretrained parameters, the model receives relevant retrieved passages as contextual information.

This process is central to RAG architectures, which combine a retrieval mechanism with a generative model (Lewis et al., 2020). The retrieved information provides an evidentiary foundation for the response, potentially reducing unsupported claims and improving factual grounding.

For example, if a student asks, *"What are the admission requirements for Computer Science in Canada?"*, the system should retrieve relevant university-specific admission documents before generating a response. The model can then synthesize the retrieved information into a concise explanation while identifying the relevant institution and source.

The generation layer should also be designed to acknowledge uncertainty. If the repository does not contain sufficient information to answer a question, the system should avoid generating an unsupported answer. Instead, it should indicate that the available information is insufficient and direct the student toward an appropriate authoritative source.

4.5 User Interaction Layer

The User Interaction Layer provides the conversational interface through which students interact with the system. A conversational interface allows students to ask questions using natural language rather than requiring them to understand databases, search syntax, or technical terminology.

Potential questions may include:

"What are the admission requirements for Computer Science in Canada?"

"Which countries provide post-study work permits?"

"Compare Engineering programs in Germany and Australia."

The system should be capable of identifying the key components of these questions, including subject area, country, degree level, institution, and information type. It can then use these parameters to guide retrieval and generate an appropriate response.

An important feature of the proposed interface should be **source transparency**. Responses should ideally identify the documents or sources from which important claims were retrieved. This is particularly important for admissions and immigration information because students need to verify critical requirements independently.

4.6 Overall Framework and Expected Contribution

The proposed framework establishes a pipeline in which reliable external information moves from data collection to knowledge processing, retrieval, generation, and user interaction. Its central principle is that the quality of an AI-generated educational response depends not only on the language model but also on the quality, relevance, and currency of the information supplied to it.

The framework therefore addresses several limitations of conventional LLMs. External retrieval can provide access to information beyond the model's original training data, while semantic search can improve the identification of relevant documents. Metadata filtering can increase retrieval precision, and structured information can improve consistency for complex educational queries. Finally, source attribution can provide greater transparency and enable users to verify important claims.

However, the framework does not completely eliminate hallucinations. Errors may still occur because of incomplete repositories, incorrect embeddings, retrieval failures, outdated documents, or inappropriate interpretation of retrieved information. Therefore, continuous data validation, source monitoring, retrieval evaluation, and human oversight would remain essential.

Overall, the proposed conceptual framework provides a theoretical foundation for developing a trustworthy educational conversational AI system. By integrating curated external knowledge with semantic retrieval and LLM-based generation, it offers a potential approach for supporting students in university selection, admissions, scholarships, international mobility, and career planning while reducing dependence on the static knowledge of language models.

5. Future Scope

The rapid development of Large Language Models (LLMs) has created significant opportunities for developing intelligent systems capable of supporting students in educational and career decision-making. However, the findings discussed in this review indicate that the usefulness of such systems depends

strongly on the accuracy, relevance, and currency of the information provided to users. Retrieval-based architectures, particularly Retrieval-Augmented Generation (RAG), provide a promising direction for addressing these limitations by connecting language models with external and verifiable knowledge sources (Lewis et al., 2020; Gao et al., 2023). Building upon the literature reviewed in this study, several areas of future research can be identified for developing a reliable, domain-specific conversational assistant for higher education guidance.

5.1 Development of a Curated Educational Knowledge Repository

A major area for future development is the creation of a comprehensive and continuously updated repository containing reliable educational and immigration information. The repository could include university profiles, undergraduate programs, admission requirements, tuition fees, scholarships, standardized testing requirements, visa regulations, post-study work opportunities, and career pathways. The quality of such a system would depend substantially on the quality of its underlying data. Therefore, information should preferably be collected from authoritative sources, including official university websites, government immigration portals, scholarship providers, and recognized educational institutions. Each document could be accompanied by metadata identifying its source, publication date, geographical relevance, subject area, and last verification date.

A particularly important feature would be automated and human-assisted updating mechanisms. Admission requirements and immigration policies can change periodically, meaning that a static repository could eventually reproduce the same problem of outdated information found in conventional LLMs. Future systems could therefore incorporate mechanisms for detecting changes in source documents and flagging information requiring verification.

5.2 Implementation of Retrieval-Augmented Generation

Future research could move beyond the conceptual framework presented in this review toward the practical implementation of a Retrieval-Augmented Generation architecture. A prototype system could combine a pretrained LLM with a retrieval system capable of identifying relevant educational documents before generating an answer.

Such a system would allow researchers to compare responses generated using only the LLM against responses generated using retrieved external information. This comparison could provide empirical evidence regarding whether retrieval improves factual accuracy and reduces unsupported claims.

Future implementations could also investigate different retrieval strategies, including dense retrieval, sparse retrieval, hybrid retrieval, and reranking. Different approaches may perform differently depending on the type of question. For example, semantic retrieval may be useful for broad career-related questions, while exact keyword or metadata-based retrieval may be more appropriate for identifying specific visa regulations or admission requirements.

5.3 Integration of Vector Databases and Semantic Search

Another important direction is the integration of vector databases and advanced semantic search mechanisms. Embedding-based retrieval allows documents and user queries to be represented as numerical vectors, enabling the system to identify information based on semantic similarity rather than exact word matching (Reimers & Gurevych, 2019).

Future research could compare different embedding models and retrieval configurations to determine which combinations provide the highest retrieval precision for educational queries. Vector databases such as FAISS and other specialized retrieval platforms could also be evaluated according to retrieval speed, scalability, filtering capabilities, and resource requirements (Johnson et al., 2019).

The system could further combine semantic retrieval with metadata filtering. For example, a query regarding undergraduate Computer Science programs could be restricted according to country, degree level, academic requirements, and information recency. Such filtering could reduce irrelevant retrieval and improve the reliability of generated responses.

5.4 Evaluation of Hallucination Reduction

A critical area of future research is the empirical evaluation of whether the proposed architecture actually reduces hallucinations. The system should not be considered reliable solely because it uses external retrieval. Retrieval itself can introduce errors if irrelevant or incorrect documents are selected.

Future studies could therefore establish experimental benchmarks containing realistic educational questions and verified answers. The system could be evaluated using retrieval metrics such as precision, recall, and F1-score, together with measures of response quality and factuality.

Human evaluation would also be valuable. Experts in education, admissions, or international student advising could assess responses according to factual correctness, relevance, completeness, source faithfulness, and clarity. Evaluation should additionally examine whether the system appropriately communicates uncertainty when reliable information is unavailable.

Longitudinal evaluation could provide further insight by testing whether the system remains accurate as university policies, admission requirements, and immigration regulations change.

5.5 Development of Personalized Recommendation Capabilities

Future versions of the proposed system could extend beyond question answering to provide personalized educational recommendations. Students differ considerably in academic performance, subject interests, financial constraints, preferred countries, career goals, and extracurricular experiences. A retrieval-based system could potentially combine these preferences with structured educational data to provide more relevant recommendations.

For example, a student interested in Computer Science and cybersecurity could receive recommendations based on academic eligibility, preferred countries, tuition budget, university offerings, and potential career pathways. However, personalization should not replace user decision-making. Recommendations should be presented as evidence-based options rather than definitive instructions.

Hybrid AI architectures could support this development by combining LLMs with vector retrieval, knowledge graphs, recommendation algorithms, and structured databases. Knowledge graphs could represent relationships between degrees, universities, skills, occupations, and industries, while retrieval systems could provide current supporting information.

5.6 Integration of Explainability and Source Verification

Another promising area of future research is the development of explainable retrieval-based educational AI. Students should be able to understand why a particular answer or recommendation was generated. Providing source references, retrieved passages, and the date on which information was verified could improve transparency and user trust.

This is especially important for high-stakes information such as immigration and visa regulations. The system should clearly distinguish between general educational guidance and information requiring confirmation from official authorities. Future research could therefore examine how source attribution and uncertainty communication influence students' trust and decision-making.

5.7 Ethical and Privacy Considerations

Future implementations must also address ethical considerations surrounding personalization and student data. A personalized educational assistant may process sensitive information such as academic records,

financial preferences, educational aspirations, and application histories. Appropriate data minimization, security, consent, and access controls would therefore be necessary.

The system should also be monitored for potential biases in university or career recommendations. If the underlying repository disproportionately represents certain institutions, countries, or career paths, recommendations could unintentionally favor those options. Future research should therefore evaluate fairness, transparency, and representational balance.

5.8 Overall Future Direction

The long-term objective of this research could be the development of a trustworthy, continuously updated, and domain-specific educational conversational assistant. Such a system would combine curated external knowledge with semantic retrieval, vector databases, knowledge graphs, and LLM-based generation.

Future research should move from conceptual analysis toward controlled experimentation, prototype development, and real-world evaluation. Particular attention should be given to retrieval accuracy, factual grounding, source transparency, information freshness, personalization, and ethical safeguards. If these challenges are addressed effectively, retrieval-based AI could become a valuable supplementary tool for students navigating increasingly complex higher education and career decisions. Nevertheless, such systems should complement rather than replace official sources, qualified counselors, and individual student judgment, particularly when decisions involve legally or financially significant consequences.

6. Conclusion

Large Language Models have emerged as powerful tools for generating and processing natural language, with applications extending from general conversational systems to education, research, and career guidance. However, their increasing use in decision-support contexts has highlighted the problem of hallucination, in which models produce information that is inaccurate, unsupported, or fabricated while presenting it in a fluent and convincing manner (Ji et al., 2023). This issue is particularly significant in educational applications because students may use AI-generated information to make decisions concerning university admissions, scholarships, international study, immigration, and career planning. Therefore, improving the factual reliability of LLMs is an important requirement for their responsible application in education.

This review examined the causes of hallucinations and the major approaches developed to reduce them. Fine-tuning, Reinforcement Learning from Human Feedback, and prompt engineering can improve model performance, alignment, and response quality, but none of these approaches independently guarantees factual accuracy (Ouyang et al., 2022). Their limitations arise partly from the fact that an LLM continues to depend on patterns learned from its training data. If information is incomplete, outdated, or absent from the model's training data, the model may still generate an incorrect response.

External knowledge retrieval provides an alternative approach by connecting language models to information sources outside their internal parameters. Retrieval-Augmented Generation is particularly promising because it combines information retrieval with language generation, allowing relevant documents to be retrieved before an answer is produced (Lewis et al., 2020). This approach can provide access to more current information while allowing responses to be grounded in identifiable sources. However, RAG should not be considered a complete solution to hallucination. The reliability of the final response depends on the quality, relevance, and freshness of the retrieved information. Retrieval failures, poor document segmentation, unsuitable embeddings, and limited context can continue to introduce errors.

The review also highlighted the importance of semantic search and embedding models in improving retrieval quality. Unlike conventional keyword-based approaches, semantic retrieval can identify information according to contextual meaning and similarity (Reimers & Gurevych, 2019). Vector databases further enable efficient storage and retrieval of large collections of embedded information, while knowledge graphs can represent relationships between entities such as universities, degree programs, skills, occupations, and career pathways (Hogan et al., 2021). Combining these technologies through hybrid retrieval architectures may therefore provide a stronger foundation for domain-specific conversational systems.

Based on these findings, the proposed conceptual framework integrates a curated educational knowledge repository, document chunking, embeddings, metadata tagging, semantic retrieval, vector databases, LLM-based generation, and a conversational user interface. Such a framework could support questions concerning university admissions, scholarships, degree selection, visa regulations, post-study work opportunities, and career pathways. Importantly, the system should prioritize source transparency and communicate uncertainty when sufficient evidence is unavailable rather than generating unsupported responses.

The review identifies a significant opportunity for future research in developing retrieval-based AI systems specifically designed for higher education and international student guidance. Existing research has established the technical potential of retrieval-based architectures, but comparatively less attention has been given to integrating diverse educational information sources into a unified, continuously updated conversational framework. Future empirical studies should therefore evaluate such systems using retrieval metrics, factuality measures, human evaluation, and domain-specific benchmarks.

Overall, external knowledge retrieval represents a promising strategy for reducing hallucinations and improving the trustworthiness of Large Language Models. Its effectiveness, however, depends on more than simply adding a retrieval component to an LLM. Reliable data sources, effective retrieval mechanisms, appropriate evaluation, continuous updating, transparency, and human oversight are equally important. For educational applications, retrieval-based AI should therefore be viewed as a **decision-support tool rather than a replacement for official sources or qualified human guidance**. When designed with these principles, retrieval-augmented systems could contribute to more accurate, transparent, and responsible use of generative AI in higher education and career decision-making.

References

1. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Evtimov, I., Fan, Y., Ford, J., Fu, M., Gao, M., Germain, C., Glavaš, G., Goh, G., Grosse, R., Guzman, S., Hou, C., Hsu, J., Ji, Y., Kandre, P., Kapoor, S., Katanforoosh, K., Khattab, O., Koh, P. W., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, T., Li, J., ... Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv*. <https://arxiv.org/abs/2108.07258>
2. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
3. 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* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
4. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A survey. *arXiv*. <https://arxiv.org/abs/2312.10997>
5. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngonga Ngomo, A.-C., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. *ACM Computing Surveys*, 54(4), Article 71. <https://doi.org/10.1145/3447772>
6. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), Article 248. <https://doi.org/10.1145/3571730>
7. Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535–547. <https://doi.org/10.1109/TBDATA.2019.2921572>
8. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
9. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 33 (pp. 9459–9474).
10. Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81). Association for Computational Linguistics.
11. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
12. Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. <https://doi.org/10.3115/1073083.1073135>
13. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 3982–3992). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1410>
14. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. (2020). Learning to summarize from human feedback. *Advances in Neural Information Processing Systems*, 33, 3008–3021.
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

16. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.