BALNEO RESEARCH Balneo and PRM Research Journal
English Edition
e ISSN 2734 - 8458 p ISSN 2734 - 844X ROMANIAN ASSOCIATION OF BALNEOLOGY

*Research article*

# "A Midsummer Night's Dream" quest for truth: From ChatGPT "hallucinations" to RAG reasoning and ACURAI precision — a scoping review on detection, minimizing, and (almost) complete error elimination and enhancing Large Language Models' reliability

**Aurelian Anghelescu [1,2], Constantin Munteanu [2,3], Lucia Ana-Maria Anghelescu [4], Gelu Onose [1,2]**

**Citation:** Anghelescu A., Munteanu C., Anghelescu L.A.M., Onose G. - "A Midsummer Night's Dream" quest for truth: From ChatGPT "hallucinations" to RAG reasoning and ACURAI precision — a scoping review on detection, minimizing, and (almost) complete error elimination and enhancing Large Language Models' reliability
*Balneo and PRM Research Journal* **2025**, *16(3): 847*

Academic Editor(s):
Constantin Munteanu

Reviewer Officer:
Viorela Bembea

Production Officer:
Camil Filimon

Received: 04.08.2025
Published: 30.09.2025

**Reviewers:**
Corina Sporea
Himena Zippenfening

**Publisher's Note:** Balneo and PRM Research Journal stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



1 Carol Davila University of Medicine and Pharmacy, Bucharest
2. Teaching Emergency Hospital "Bagdasar-Arseni," Bucharest
3. Grigore T. Popa University of Medicine and Pharmacy, Faculty of Bioengineering, Iaşi
4. Master's in psychology, Bucharest

* Correspondence: aurelian.anghelescu@umfcd.ro

**Abstract:** Like A Midsummer Night's Dream, large language models (LLMs) exhibit vast imagination, drawing on massive training datasets. However, they may fabricate or mix information, lacking mechanisms to verify real-world sources. Most commercial LLMs, including those used in medicine, remain prone to hallucinations—plausible but false content. Retrieval-Augmented Generation (RAG) aims to address this by grounding LLM outputs in real-time access to verified sources like scientific databases. A 2023–2025 PubMed search identified 91 papers on RAG and LLM applications across biomedical domains; 78 were useful for our paper, adressing medical domains. RAG techniques significantly reduce hallucinations by ensuring that only validated information informs model outputs. ACURAI, an advanced system based on "phrase dominance and discrete functional units (DFUs)," further enhances LLM accuracy. Tested on a novel "RAG-Truth Dataset Caveats," ACURAI eliminated 91–100% of junk outputs in GPT-3.5 and GPT-4. While LLMs can resemble Puck (creative yet unreliable), ACURAI, aided by RAG, acts more like Theseus, grounding answers in verified data. This framework strengthens the possible role of LLMs in clinical diagnosis, academic writing, and patient education, offering a practical path toward safer and more accurate medical AI. Ultimately, human oversight remains key to interpreting and validating AI-generated outputs.

**Keywords:** Artificial intelligence (AI); Large language models (LLMs); ChatGPT; RAG (Retrieval-Augmented Generation systems); ACURAI

## 1. Introduction

Large generative language models (LLMs) are state-of-the-art artificial intelligence (AI) models that use deep learning techniques to generate human-like responses based on contextual inputs. It functions like a "word wizard," capable of understanding questions, processing information, and producing coherent dialogue.

Commercial LLMs (such as GPT-4, Anthropic Claude 2, and Google Bard/Gemini) are trained on massive datasets from the internet, enabling them to learn the nuances of language, grammar, and human expression.

LLMs exhibit general knowledge and reasoning skills, but often perform poorly on domain-specific tasks such as medical subject-question answering, and can generate misleading responses when prompted on specific subject areas.

LLMs trained with domain-specific knowledge can reduce the generation of fake information (hallucinations) and enhance the precision of LLMs in specialized domains.

Shortly after ChatGPT's launch, one researcher ironically wondered whether GPT was experiencing hallucinations from LSD. [1]

**2. Hypothesis**: If you feed an LLM with only pertinent and factual data, will the results be 100% factual?
**The answer**: The hypothesis is unfounded.

The synergistic combination of retrieval-augmented generation (RAG) for medical question answering addresses a persistent concern: improving the accuracy and reliability of LLMs in medical knowledge.

Facing this tremendous progress of the human mind`s creation (the AI's potential), one emphasized the conflicting emotions of awe and „ecstasy" versus the potential intellectual "agony" caused by the same AI`s misleading and inaccurate responses during scientific use.

Some have compared this mental state to an ambiguous oscillation between the Greek myths of Pandora and Prometheus. [2,3]

Focusing on scholarly research, the rhetorical Hamlet-like question "To be or not to be ChatGPT..." highlighted the importance of the human mind in maintaining analytical rigor when critically assessing AI-generated outputs. [4-6]

The limitations of LLMs include reliance on historical data and a lack of real-time access to verified scientific databases (e.g., PubMed, Scopus, SpringerLink). During its early development, up to 93% of PubMed ID references were inaccurate, 64% had incorrect volume/page numbers, and 60% cited the wrong publication year [4,7].

Numerous attempts to access specialized medical literature yielded irrelevant results or were redirected to other sources.[4,8]

Further investigation revealed that Chat GPT struggles to understand scholarly concepts and provide accurate citations or detailed bibliographic information. Additionally, because it is trained on content from diverse sources, it also absorbs their biases.

An LLM could (sometimes) be compared to mischievous spirits like Puck in the Shakespearean play "A Midsummer Night's Dream", offering creative and fluent answers but without guaranteed veracity. It can create varied and imaginative responses based on statistical combinations of information. However, this process isn't always accurate or verifiable. It can construct fluent and coherent answers, but it lacks a reliable way to verify real-world sources. Thus, its ideas can turn into "hallucinations" as the model might invent or incorrectly combine information.

Here are some examples of responses, obtained from the free variant of GPT-4.0 when prompted on specific subject areas. In February 2025, I questioned the AI on a few anatomical items because I needed some illustrations for my students. The resultant cartoons (figs. 1, 2, and 3) are blatantly incorrect and speak for themselves.

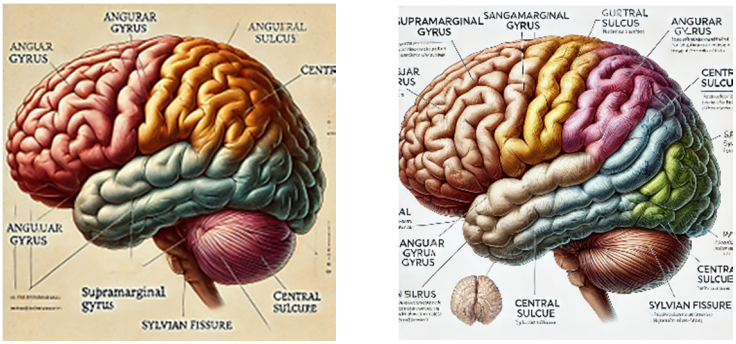


**Fig.1** ChatGPT`s response to my solicitation: "I have generated an anatomical illustration of the angular gyrus and supramarginal gyrus, clearly highlighted in the context of other brain structures. If you need any changes or additional details, let me know!"

The second attempt was focused on the patellar reflex (Fig. 2), and the generated picture was hilarious.

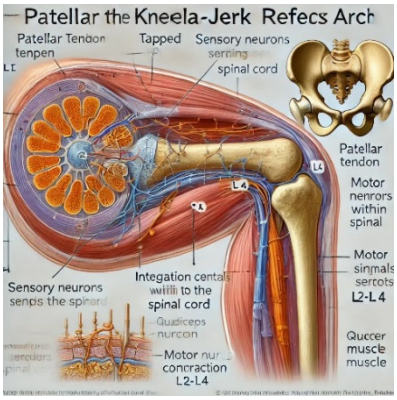


**Fig.2** ChatGPT-4.0: "I will now generate a clear visual diagram for you." The knee-jerk reflex is awe-inspiring in artistic stills but disappointing in terms of scientific accuracy.

Plate 3, which refers to the peroneal nerve, reflects the lack of specialized medical knowledge and the susceptibility to generating inaccurate information (Fig.3). The image is so anatomically inaccurate that it could be a „replica" of a famous painting by Francisco de Goya: "The Sleep of Reason Produces Monsters" (original title in Spanish: "El sueño de la razón produce monstruos")

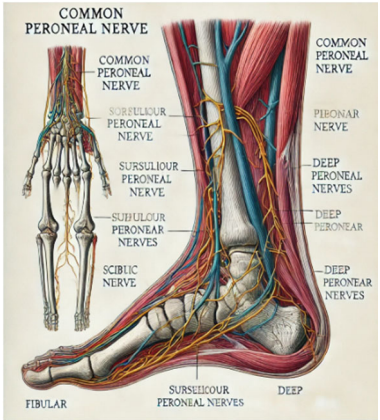


**Fig.3**. The scheme generated by the free GPT-4.0 version is a symbol of irrationality (content, anatomical terms, hallucinatory segmental representation).

To acquire accurate information, one must first understand the sources of errors generated by AI. The primordial ” prehistoric” paradigm in IT: ”Garbage in, garbage out”, means inaccurate or false training data will generate ”hallucinatory” answers. The LLMs may ingest incorrect texts, fake news, speculative posts, or opinions. They cannot always differentiate between high- and low-quality sources, and if a false statement frequently appears, the model may learn it as a probable truth.

Faulty generalizations are another cause of possible „hallucinatory” answers: rather than copying phrases, models learn patterns from data. The LLM occasionally creates statements that appear to be true but are false by combining passages from various contexts.

LLMs do not verify output against a "truth database" because they lack factual database verification. They generate the most statistically likely (fluent) texts, without proper evidence-checking. The LLMs have design-induced creativity and are optimized for generating fluid text even in the absence of comprehensive information. Instead of stating "I don't know," they prefer elegant guesses that may lead to „hallucinations”. Thus, a model like ChatGPT can fabricate incorrect answers, such as nonexistent academic papers, imaginary scientific concepts, or false historical facts. These hallucinations are especially problematic as the model presents its answers in a confident and convincing tone, making users more likely to believe false information. [9,10]

However, AI models are continuously evolving. ChatGPT has undergone significant improvements since its first version launched in November 2022. The benchmark in GPT-3.5 has an estimated 175 billion parameters.[11]

Although the GPT-4.0 version produced more balanced answers than the GPT-3.5 version, academic methodology and human reasoning skills still outperformed them. [6]

Most problems will be fixed by the updated GPT-4.5 and Turbo (officially released in February 2025) and the much-anticipated GPT-5 (expected to launch in the second half of 2025, according to OpenAI).

To get the right answers and address hallucinations in LLMs necessitates strategies to prevent and manage such errors in real-time interactions. For instance, addressing clear, crystal questions and references can reduce ambiguity and semantic collisions. One must use precise labels and contextual framing to prevent confusion between semantically similar terms.

Enhancing LLMs` performance with the RAG strategy improves their reliability and accuracy, providing an updated corpus of knowledge that can be used to generate correct answers.

In the endeavour to eliminate incorrect answers, a specially created dataset called the "Retrieval-Augmented Generation—Truth Dataset" was created. The RAG techniques combine text generation with real-time information extracted from external sources to generate accurate answers. RAG-type systems are a technique utilized in developing AI assistants, sometimes used alongside models like GPT, but not implicitly integrated into the standard (free) version of ChatGPT for everyday use.

As its name suggests, the RAG retrieves relevant information by accessing scientific databases, websites, or documents to verify external sources before generating an answer. It's like AI performs a "Google Search" or browses Wikipedia or another encyclopedia before responding. After finding the relevant information, the AI model generates texts and answers using this previously verified data.

"Corpus" is a term often used in linguistics and AI, and generally means a large collection of texts used for training or evaluation. It serves as the reference basis for comparing the AI model's response.

RAG-Truth Corpus means a specially created data set containing a collection of correct data, from verified sources, which are considered true (Ground Truth Answers). RAG has access to external, actualised data sources for each question, and uses current and accurate information to avoid "hallucinations." It can be customized, in a „niche of interest" if loaded with specific, ultra-specialized documents for a company, institution, or on specific subjects.
In this regard, RAG-Truth Corpus provides better than "standard" LLMs (which don`t have new or updated information).

**3. Methods**

This scoping review was conducted in accordance with the methodological framework for scoping reviews. A systematic search was performed in the **PubMed** database (using the beta version of the PMC search engine), covering the period **January 2023 – July 2025**. The search syntax used was: *("Retrieval-Augmented Generation" OR "RAG") AND ("Large Language Model" OR "LLM") AND (medicine)*. The search was last updated on 15 July 2025 and revealed 91 papers (1 in 2023, 31 in 2024, and 59 in 2025), most of them (n=78) being related to the medical fields/ domains.
**Inclusion criteria** comprised original research articles, reviews, and clinical reports addressing applications of RAG and LLMs in biomedical and healthcare-related fields, published in English. **Exclusion criteria** were non-peer-reviewed papers, editorials, and studies unrelated to biomedical domains.
The selection process followed four steps: (1) identification of records, (2) removal of duplicates, (3) screening of titles and abstracts, and (4) full-text eligibility assessment. Two reviewers independently screened the records, with disagreements resolved by consensus.
From the initial search, 91 papers were identified (1 in 2023, 31 in 2024, and 59 in 2025). Of these, 78 were related to medical applications.

**4. Results**

Refining the specificity of the search by adding different medical specialities, one found different papers in genomics (n=11), radiology (n=7), disease-specific LLM Chat platforms (n=24), applications in emergency departments (n=7), nutrition (n=3), medical education for different specialists for clinical decision support (n=12), and healthcare administration themes (n=7), for the patient`s (n=5) and caregiver`s education (n=2).
Thirteen papers focused on AI in enhancing medical information retrieval from the PubMed database ( medical information retrieval) and were not included in the detailed discussion.
Below is PRISMA flow diagram of the study selection process (fig.4)

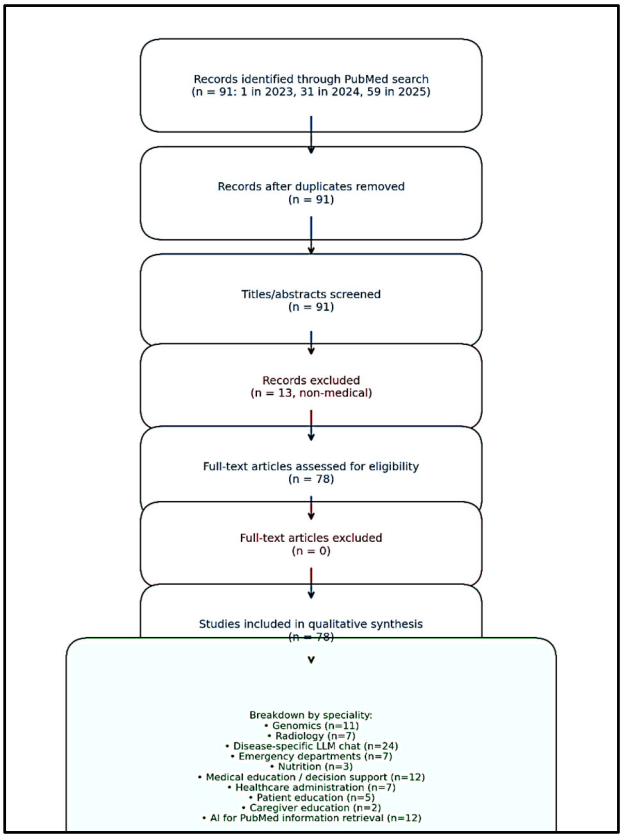


**Fig. 4.** PRISMA-adapted flow diagram of the study selection process

Table I summarizes the comparative performance of LLMs used alone versus RAG-enhanced models across different healthcare domains. In general, LLMs alone demonstrated moderate to inconsistent reliability, often producing misinformation or low accuracy (e.g., in radiology, diabetes, or nutrition). By contrast, RAG-enhanced models consistently provided higher accuracy, structured information, and better user satisfaction, particularly in clinical decision support, radiology, and patient education. This demonstrates the potential of RAG integration to improve both precision and applicability in healthcare contexts significantly. Discussion will support the synthetic affirmations.

**Table I. Comparative analysis of potential applications of LLMs alone versus RAG-enhanced models in healthcare**

| **Domains** | **Use Cases** | **LLM Alone** | **RAG-Enhanced Model** |
|---|---|---|---|
| **Medical Education** (**for clinicians**, **patients**, and **caregivers**) | Virtual tutors, quizzes, and content creation | Moderate | High |
| **Clinical Support (guidelines)** | Diagnosis, treatment recommendations | Unreliable | More accurate |
| **Healthcare Administration** | Report writing, documentation | Good | Excellent |
| **Radiology** | Lung cancer staging | Poor–Moderate | 86% accuracy (NotebookLM) |
| **Diabetes Management** | Multilingual guideline support | Inconsistent | High reliability |
| **Nutrition / Supplement interaction info retrieval** | Supplement interaction info retrieval | Misinformation | 99% accurate (iDISK2.0) |
| **Nursing** | Breast cancer care answers | Lacks precision | Higher satisfaction & accuracy |
| **Surgery (plastic)** | Evidence synthesis, multilingual support | Limited | Structured, grounded info |

**5. Discussion** A retrieval-augmented information system (RAG) for an enhancement framework significantly increased the percentage of correct answers, by 7% for GPT-4, 19% for Claude 2, and 9% for Google Bard. [12]
Comparing RAG implementation to baseline LLMs, performance increased by an overall odds ratio of 1.35 [13,14], exhibiting an accuracy improvement from 44.46% to 48.54% in medical question-answering. [15]

The classic LLMs (mentioned above) only respond based on information accumulated from prior training, which includes data up to 2023 or later, depending on the version. It does not browse the internet in real time. RAG-equipped LLMs (RAG-LLM) provide reliability in clinical applications. However, ChatGPT Plus (especially GPT-4 Turbo) sometimes uses similar RAG functions.

Advanced RAG techniques considerably enhance theranostics chatbots (conversational artificial intelligence systems developed to integrate two essential functions in medicine, per their name, a combination of "Therapeutics" and "Diagnostics") [16]

There are some similarities between these AI-based technologies: both LLM and RAG are based on automatic machine learning methods and natural language processing technologies, using AI models to answer users' questions. Both technologies are capable of natural interaction, generating responses in an easy-to-understand language, which facilitates interaction with the user.

Strategies such as RAG use collections of verified external information and documents, rather than relying on the pre-trained model, to represent the key differences.

**Table II** presents a comparative synthesis highlighting the differences between LLMs and RAG systems. It focuses on information sources, verification capabilities, risks of hallucination, and practical benefits.

**Table II. Main differences between the performances of LLMs** and **RAG systems**

| Feature | LLM (Large Language Model) | RAG (Retrieval-Augmented Generation) |
|---|---|---|
| **1. Source of information** | Trained on massive datasets (books, websites, forums), static, and pre-defined | Access to reliable external knowledge; real-time data access to verified external sources |
| **2. Real-time access** | **No** real-time access; relies on pre-training knowledge | Real-time access to external verified sources (e.g., scientific databases, clinical guides) to generate accurate answers |
| **3. Verification of information** | Can**not** verify in real time; may generate answers that are plausible but sometimes incorrect ("hallucinated"), or incomplete content. | Can cross-check and dynamically perform real-time verification of information during the generation of the answer |
| **4. Update frequency** | Static knowledge (cutoff date) requires retraining for updates | Dynamically retrieves updated content without retraining |
| **5. Risk of hallucinations** | **High**—may mix facts or invent details ("dream-like" creativity) | **Greatly reduced**—pulls from validated sources to ground answers |
| **6. Response accuracy** | **Variable** — depends on the prompt and model limitations. Responses can be fluent and coherent, but they can also be wrong or based on outdated information. | **High**—uses evidence-based, verifiable sources. Responses are precise and based on validated data, free from "hallucinations." |
| **7. Integration of external sources** | **Limited** — knowledge must be part of the previous training data | **High** — can include current clinical guidelines, webinars, recent articles, structured databases |
| **8. Creativity and Flexibility** | Creativity and flexibility are its strengths. | More rigid in terms of creativity<br>Much more reliable when it comes to the accuracy of responses |

| | | |
|---|---|---|
| **9. Users' trust and satisfaction** | May decrease when hallucinations occur or answers seem uncertain („agony and ecstasy", „Prometheus and Pandora") | Improved due to transparency, accuracy, and source attribution |
| **10. Use case in science, healthcare, and academia** | Limited to up-to-date medical/scientific queries | Ideal for clinical decision support, scientific literature summarization, and real-time updates |
| **11. Dependency on OpenAI's training** | Fully based on internal training. | Independent—can be fine-tuned and customized with its retrieval tools and content. |

**a. Medical Education**

a.1. In **medical education** and didactic domains, LLMs provide junior trainees in various medical specialties with individualized tutoring and the viewpoint of virtual patients.[17]

Semantic clinical AI can also be applied to the medical licensing exam, focused on current clinical knowledge. [18]

People living with chronic diseases are increasingly seeking health information online. **Patient (self)-education** is a healthcare concept that involves raising the public`s awareness with evidence-based medical information. This biomedical information promotes a healthier life and better management of their conditions. Traditional educational materials often lack reliability and fail to engage or motivate effectively. A RAG-based technology integrated with an LLM chatbot can answer real-world clinical questions, deliver contextually appropriate, empathetic, individualised, and clinically credible responses to queries. [12; 19-21]

In healthcare applications, AI techniques increase accuracy and decrease misinformation. Innovative approaches such as LLMs powered by RAG enhance health literacy by delivering interactive, medically accurate, and user-focused resources based on trusted sources. These platforms are capable of producing human-sounding text and, by extension, patient education materials.[21]

RAG-LLMs' convergent methods mitigate content issues by integrating AI-generated content with curated knowledge sources, ensuring more accurate and contextually relevant responses.

A RAG-enhanced LLM's capacity to raise journal club participation and learning, and revolutionize public education. [22]

a.2. In the fields of **diabetes nutritional management** and **dietary supplements** (DS), the risk of misinformation is high due to the prevalence of unverified online sources. To address this challenge, providing accurate and reliable medical information can empower the public to make informed decisions for self-education and self-management, particularly in conditions like diabetes.

Over 8091 DS ingredients are found in 163,806 DS products worldwide, and they are linked to 786 diseases. Intelligent iDISK 2.0 database provides scientific information with 99% accuracy regarding DS interactions, supplements, and medications, significantly reducing the risk of errors and supporting safe decision-making in healthcare. [12, 19, 23, 24]

**a.3 Basic medical education for the carers**

One in five adults in the US is a family carer for someone who has a severe illness or disability. Family caregivers, in contrast to professional ones, frequently take on this role without any official training or specialized preparation.

Enhancing family carers' ability to deliver high-quality assistance is therefore critically needed. [11]

Modern AI technology, as a useful educational tool or an adjunct to care, has the potential to enhance the learning and caregiving capabilities of family members.

Reliability of information is crucial when using LLMs as caregivers' primary assistance.

**b**. **AI-assisted decision-making**

Regarding AI-assisted clinical decision-making, a synergistic RAG-LLM could potentially aid in diagnostics, treatment recommendations, and retrieving medical knowledge. [17]

The RAG system supports the development of **clinical guidelines** by providing rapid access to up-to-date scientific evidence and automatically synthesizing it.

RAG technology enables the development of dynamic, focused, and adapted „living" relevant **guidelines** by ensuring accuracy, source transparency, reduced errors, time efficiency, and continuous updates. [14, 25]

The **emergency departments** are characterised by a high level of stress and a short amount of time available for information collection. A performant LLM could be a useful tool for identifying patients presenting critical illnesses at admission, requiring an alert extraction of information from the patient's complaints and anamnesis. [26]

**c. Healthcare Administration**

In **healthcare administration**, RAG-LLM technology reduces the administrative burden on healthcare professionals, automatically performing some „boring" tasks such as clinical note summarization, routine data extraction, and report generation. [17]

d. **Radiology and nuclear medicine imaging.**

A **radiology**-specific approach with RAG-based LLMs offers a promising streamlining of workflows by integrating reliable, verifiable, and customizable information. [27-29]

The RAG systems bring significant benefits to radiology by providing real-time access to validated medical sources, reducing AI „hallucinations," and supporting automated medical report generation. By integrating databases and archives, it enables the retrieval of similar cases and offers evidence-based clinical justifications. It also facilitates continuous medical learning and contributes to the automatic triage of imaging studies, supporting faster and more reliable clinical decisions.[27, 30-32]

RAG may also improve LLM performance and clinical applicability in nuclear medicine imaging. [33]

RAG-LLMs' specific platforms may enhance guideline adherence by dynamically integrating external information and could support **clinical decision-making.** [25, 27, 29]

The RAG system offered significantly better performance than pure LLMs. The synergistic cooperation with LLMs improved GPT-4 performance to 81.2% vs 75.5% (P = .04) and Command R+ (Cohere) to 70.3% vs 62.0% (P = .02), but has no significant benefits for Claude Opus (Anthropic), Mixtral (Mistral AI), or Gemini 1.5 Pro (Google DeepMind). [34]

To illustrate the incredible capacity of medical imaging analysis, I will give as an example the radiograph of a patient admitted to our NeuroRehabilitation Clinic after a SCI with a cauda equina lesion. I asked ChatGPT-4.0 to describe this postsurgical radiograph (Fig. 4a,b), without giving it additional information regarding the etiology and neurosurgical type of intervention.

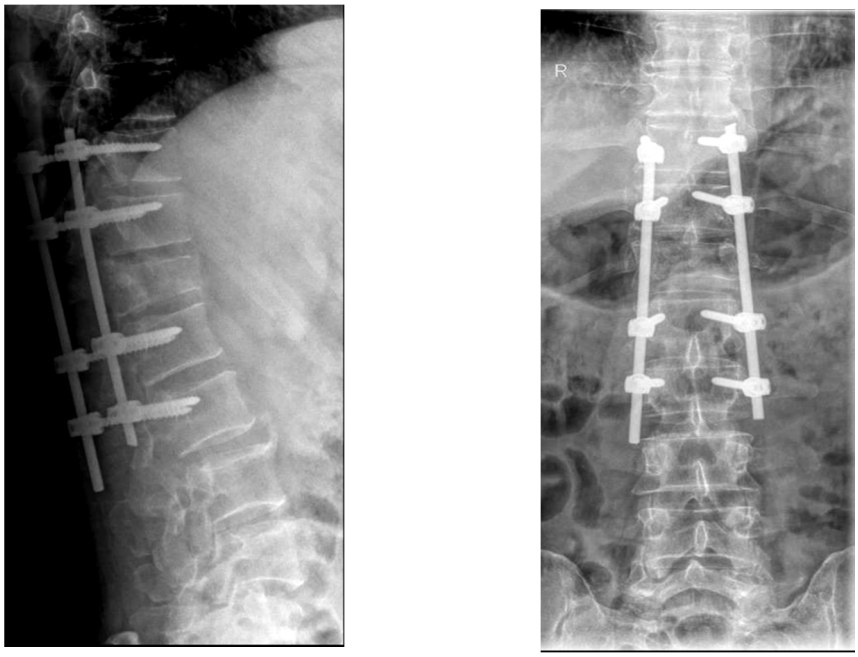


**Fig. 5 (a,b)**: Radiologic imagery, clinical case: FFO, male, 50 years old, SCI (car accident, L3 AIS-C paraplegia), T12 somatic comminuted fracture with myelo-radicular compressive fragments and L1, 2, and L1,2,3 transverse process fractures, **operated**: metallic osteosynthesis T10-T11-L1-L2 with 8 transpedicular screws and 2 interconnecting bars (Medtronic MRI compatible system)osteosynthesis

The AI`s interpretation was excellent. ChatGPT said: "This is a lateral x-ray of the lumbar spine, showing a spinal fixation system with pedicle screws and rods. Key observations include:

- Spinal fixation with metal screws and rods: These are used to **stabilize the spine**, possibly following surgery to correct a fracture, scoliosis, spondylolisthesis, or other spinal pathology.
- **Vertebral alignment**: There is no major spinal deformity in the visible area.
- Bone structure: The vertebrae in the fixed area appear to be fused, suggesting a **spinal fusion procedure**."

ChatGPT said: "This is an anteroposterior radiograph of the lumbar spine, similar to the previous one, but from a different angle. Key observations include:

- Pedicle screw and rod spinal fixation system: These implants are used to stabilize the spine, indicating surgery, most likely to treat a pathology such as vertebral fractures, spondylolisthesis, scoliosis, or spinal instability.
- Correct implant placement: The screws are symmetrically positioned along the vertebrae, and the rods are parallel, suggesting proper placement.
- Vertebral alignment: No significant deviation is seen, indicating effective fixation.
- **Osteophytes and degenerative changes**: Signs of wear and tear on the spine may be seen, possibly as part of the initial pathological process.
- This **radiograph confirms spinal stabilization surgery**."

I must emphasize the correct and balanced answer of the AI: "If you need a detailed interpretation or a correlation with clinical symptoms, **I recommend consulting a**

**specialist** in orthopedics or neurosurgery. Do you have information about the reason for the surgery?"

The rapid advancement of AI has raised significant professional concerns among radiologists:

a) Economic pressure – AI could be used to cut costs, reducing radiologist positions or outsourcing image interpretation.
b) Falling behind technology – The fast pace of AI development creates anxiety about constant retraining and the risk of becoming obsolete.
c) Job replacement – AI may take over repetitive tasks (like nodule detection, fracture identification, tumor classification), threatening radiologists' roles.
d) Loss of expertise status – There is concern that radiologists may become mere algorithm supervisors, losing autonomy and professional prestige.
e) Legal and ethical responsibility – It's unclear who is liable for AI-generated errors; radiologists fear bearing responsibility for systems they don't control.

Despite these concerns, many experts now see a balanced reality — AI as a partner, not a replacement—enhancing accuracy, speed, and allowing more time for clinical and multidisciplinary tasks**.**

Clinical decision-making support in plastic surgery and nursing for breast cancer can benefit from collaboration between RAG-LLM and medical teams. [35,36]

Despite advancements in AI, the issue of hallucinations in LLMs continues to hinder the adoption of AI in high-stakes applications.

Even the most advanced RAG systems fall short of 80% accuracy in producing relevant and faithful outputs, highlighting the need for further improvement.[37]

ACURAI is a novel framework designed to enhance the reliability and factual accuracy of LLMs by significantly reducing or eliminating hallucinations.

The name **ACURAI** combines **ACU** (from *accuracy*) and **RAI** (referring to *Responsible AI*).

The ACURAI methodology overcame the limitations of current systems in preventing erroneous responses.

The ACURAI system improves output fidelity by restructuring noun phrases and disambiguating meaning within context. The underlying theoretical (mathematical) models are based on the ***Dominance of Noun Phrases*** (understanding which concepts dominate a sentence), ***Discrete Functional Units*** (breaking text into modular chunks for analysis and correction), and ***Reference Remapping*** (by restoring the original terminology after output generation).

This technique involves reforming queries and contextual data before model input, ensuring strict alignment between the prompt and the generated output. Unlike reactive strategies that block hallucinations post-generation, ACURAI prevents them at the source.

The systematic reformulation techniques of ACURAI transform LLM responses from ”hallucinatory” to verifiable ones.

ACURAI is not directly affiliated with OpenAI, but ACURAI collaborated in validating its methodology, offering a pioneering solution for hallucination-free outputs.

A step forward is the ACURAI associated with the RAGTruth Corpus. This synergistic approach demonstrated nearly total hallucination-free responses for GPT-3.5 and GPT-4.0 (estimated 91% and 100%). [8]

Its ability to filter out inaccurate information and provide consistent and coherent outputs makes it a candidate for domains that demand reliability and trustworthiness, such as medical diagnostics, law, and education.

Compared to direct queries to the LLM, the primary challenges and limitations of ACURAI-RAG systems are implementation, scalability, supplementary latency, and higher computational consumption (mean time processing per query with Acurai vs. standard LLMs).

This methodological association of synergistic systems combining RAG-based evidence (ACURAI and RAG-Truth Dataset Caveats) can be compared to the Secksperean character, Theseus, representing order, logic, and reason-controlled dreams grounded in reality, with constant access to verified and valid answers.

The present study has obvious limitations—such as the short timeframe of this new subject and its dedicated literature—and has some methodological issues, but it could constitute an opportunity for future work (e.g., broader systematic reviews or empirical testing of ACURAI in clinical settings). Future advancements in the performance and dependability of LLMs utilizing RAG techniques, fine-tuning, and reinforcement learning will be crucial in ensuring patient safety and improving healthcare delivery.

## 5. Conclusions

Large language models (LLMs) have brought significant potential to revolutionize and transform various aspects of healthcare.

Because they lack specialized medical knowledge, commercially available LLMs are still prone to producing inaccurate information.

Despite AI's remarkable advancements, hallucinations can still occur in LLMs and RAG (although less frequently) due to generalization errors, misunderstandings of semantic proximity, and limitations in training data. Therefore, eliminating hallucinations remains a challenging problem.

As mentioned in one of our recent papers, our global research is not intended to "crush the miracle corolla of the world" – as a well-known Romanian poet says – but to strengthen the confidence (at least for now), in the classic academic way of carrying out systematic literature reviews, using the standardized PRISMA methodology. [6]

Nonetheless, ACURAI's synergistic approach significantly reduces error rates when combined with validated corpora as RAGTruth. Its ability to filter out inaccurate information and provide coherent outputs makes it an ideal candidate for domains that demand reliability and trustworthiness, as medical diagnostics and education.

The implementation of Retrieval-Augmented Generation (RAG) could represent a promising avenue for enhancing the factual consistency of LLMs, reducing hallucinations, and supporting ACURAI's framework for reliable clinical and scientific applications.

Human intelligence and reasoning will ultimately determine the outcome. The need for human verification and critical oversight is mandatory, especially in high-stakes domains such as medicine, science, and law.

**Author Contributions:** All authors had equal contributions.
All authors have read and agreed to the published version of the manuscript.
**Funding:** This research received no external funding.
**Institutional Review Board Statement:** Not applicable
**Informed Consent Statement:** Not applicable.
**Data Availability Statement:** Not applicable here.
**Conflicts of Interest:** The authors declare no conflict of interest

**References**


1. Beutel G, Geerits E, Kielstein JT. Artificial hallucination: GPT on LSD? Crit Care. 2023 Apr 18;27(1):148. doi: 10.1186/s13054-023-04425-6
2. https://amsconsulting.com/articles/ai-as-myth-from-prometheus-to-pandora/ (accessed on the 30th June 2025)
3. https://www.economist.com/culture/2024/01/30/are-the-myths-of-pandora-and-prometheus-a-parable-for-ai (accessed on the 30th June 2025)
4. Aurelian Anghelescu, Ilinca Ciobanu, Constantin Munteanu, Lucia Ana Maria Anghelescu, Gelu Onose. ChatGPT: "To be or not to be" ... in academic research. The human mind's analytical rigor and capacity to discriminate between AI bots' truths and hallucinations. Balneo and PRM Research Journal. 2023;14(4):614 Full Text DOI 10.12680/balneo.2023.614
5. Anghelescu A, Firan FC, Onose G, Munteanu C, Trandafir A-I, Ciobanu I, Gheorghița Ș, Ciobanu V. PRISMA Systematic Literature Review, including Meta-Analysis vs. Chatbot/GPT (AI) regarding Current Scientific Data on the Main Effects of the Calf Blood Deproteinized Hemoderivative Medicine (Actovegin) in Ischemic Stroke. Biomedicines. 2023; 11(6):1623. https://doi.org/10.3390/biomedicines11061623
6. Anghelescu A, Munteanu C, Spinu A, Ciobanu V, Popescu C, Cioca IE, Andone I, Stoica SI, Mandu M, Rebedea A, Giuvara S, Malaelea AD, Vladulescu-Trandafir AI, Morcov MV, Onose G. Standardized clinical assessments and advanced AI-driven instruments used to evaluate neurofunctional deficits, including within biomarker based framework, in Parkinson's disease - human intelligence made vs. AI models - systematic review. Front Med (Lausanne). 2025 Jun 13;12:1565275. doi: 10.3389/fmed.2025.1565275
7. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE. High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. Cureus. 2023 May 19;15(5):e39238. doi: 10.7759/cureus.39238.
8. Wood MC & Forbes AA. 100% Elimination of Hallucinations on RAGTruth for GPT-4 and GPT-3.5 Turbo. (2024). ArXiv. https://arxiv.org/abs/2412.05223
9. Jim Waldo and Soline Boussard. Gpts and hallucination: Why do large language models hallucinate? Queue, 22(4):19–33, September 2024. ISSN 1542-7730. doi:10.1145/3688007. URL https://doi.org/10.1145/3688007
10. Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. Factuality challenges in the era of large language models, 2023.URL https://arxiv.org/abs/2310.05189
11. Parmanto B, Aryoyudanta B, Soekinto TW, Setiawan IMA, Wang Y, Hu H, Saptono A, Choi YK. A Reliable and Accessible Caregiving Language Model (CaLM) to Support Tools for Caregivers: Development and Evaluation Study. JMIR Form Res. 2024 Jul 31;8:e54633. doi: 10.2196/54633.
12. Wang D, Liu Z, Shi W, Shi D, Li F, Qu B, Zheng Y. Enhancement of the Performance of Large Language Models in Diabetes Education through Retrieval-Augmented Generation: Comparative Study. J Med Internet Res. 2024 Nov 8;26:e58041. doi: 10.2196/58041.
13. Li M, Kilicoglu H, Xu H, Zhang R. BiomedRAG: A retrieval augmented large language model for biomedicine. J Biomed Inform. 2025 Feb;162:104769. doi: 10.1016/j.jbi.2024.104769. Epub 2025 Jan 13. PMID: 39814274
14. Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. J Am Med Inform Assoc. 2025 Apr 1;32(4):605-615. doi: 10.1093/jamia/ocaf008
15. Shi Y, Xu S, Yang T, Liu Z, Liu T, Li X, Liu N. MKRAG: Medical Knowledge Retrieval Augmented Generation for Medical Question Answering. AMIA Annu Symp Proc. 2025 May 22;2024:1011-1020.
16. Koller P, Clement C, van Eijk A, Seifert R, Zhang J, Prenosil G, Sathekge MM, Herrmann K, Baum R, Weber WA, Rominger A, Shi K. Optimizing theranostics chatbots with context-augmented large language models. Theranostics. 2025 Apr 21;15(12):5693-5704. doi: 10.7150/thno.107757
17. Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A Review of Large Language Models in Medical Education, Clinical Decision Support, and Healthcare Administration. Healthcare (Basel). 2025 Mar 10;13(6):603. doi: 10.3390/healthcare13060603
18. Elkin PL, Mehta G, LeHouillier F, Resnick M, Mullin S, Tomlin C, Resendez S, Liu J, Nebeker JR, Brown SH. Semantic Clinical Artificial Intelligence vs Native Large Language Model Performance on the USMLE. JAMA Netw Open. 2025 Apr 1;8(4):e256359. doi: 10.1001/jamanetworkopen.2025.6359.
19. Kelly A, Noctor E, Ryan L, van de Ven P. The Effectiveness of a Custom AI Chatbot for Type 2 Diabetes Mellitus Health Literacy: Development and Evaluation Study. J Med Internet Res. 2025 May 5;27:e70131. doi: 10.2196/70131
20. Low YS, Jackson ML, Hyde RJ, Brown RE, Sanghavi NM, Baldwin JD, Pike CW, Muralidharan J, Hui G, Alexander N, Hassan H, Nene RV, Pike M, Pokrzywa CJ, Vedak S, Yan AP, Yao DH, Zipursky AR, Dinh C, Ballentine P, Derieg DC, Polony V, Chawdry RN, Davies J, Hyde BB, Shah NH, Gombar S. Answering real-world clinical questions using large language model, retrieval-augmented generation, and agentic systems. Digit Health. 2025 Jun 9;11:20552076251348850. doi: 10.1177/20552076251348850

21. AlSammarraie A, Househ M. The Use of Large Language Models in Generating Patient Education Materials: A Scoping Review. Acta Inform Med. 2025;33(1):4-10. doi: 10.5455/aim.2024.33.4-10
22. Umer F, Naved N, Naseem A, Mansoor A, Kazmi SMR. Transforming education: tackling the two sigma problem with AI in journal clubs - a proof of concept. BDJ Open. 2025 May 8;11(1):46. doi: 10.1038/s41405-025-00338-4.
23. Hou Y, Bishop JR, Liu H, Zhang R. Improving Dietary Supplement Information Retrieval: Development of a Retrieval-Augmented Generation System With Large Language Models. J Med Internet Res. 2025 Mar 19;27:e67677. doi: 10.2196/67677
24. Mashatian S, Armstrong DG, Ritter A, Robbins J, Aziz S, Alenabi I, Huo M, Anand T, Tavakolian K. Building Trustworthy Generative Artificial Intelligence for Diabetes Care and Limb Preservation: A Medical Knowledge Extraction Case. J Diabetes Sci Technol. 2024 May 20:19322968241253568. doi: 10.1177/19322968241253568.
25. Su AY, Knebel A, Xu AY, Kaper M, Schmitt P, Nassar JE, Singh M, Farias MJ, Kim J, Diebo BG, Daniels AH. Evaluation of retrieval-augmented generation and large language models in clinical guidelines for degenerative spine conditions. Eur Spine J. 2025 Jul 7. doi: 10.1007/s00586-025-08994-8
26. Gao H, Wang K, Yuan Y, Wang Y, Liu Q, Wang Y, Sun J, Wang W, Wang H, Zhou S, Jin K, Zhang M, Lai Y. A large language model-based pipeline for extracting information from patient complaints and anamnesis in clinical notes for severity assessment. Sci Rep. 2025 Jul 14;15(1):25345. doi: 10.1038/s41598-025-07649-4.
27. Mugu VK, Carr BM, Olson MC, Schupbach JC, Eguia FA, Schmitz JJ, Khandelwal A. Increasing Adherence to Societal Recommendations in Radiology Reporting: A Feasibility Study Using Society of Radiologists in Ultrasound Guidelines for Incidentally Detected Gallbladder Polyps. Ultrasound Q. 2024 Dec 17;41(1):e00699. doi: 10.1097/RUQ.0000000000000699.
28. Li R, Mao S, Zhu C, Yang Y, Tan C, Li L, Mu X, Liu H, Yang Y. Enhancing Pulmonary Disease Prediction Using Large Language Models With Feature Summarization and Hybrid Retrieval-Augmented Generation: Multicenter Methodological Study Based on Radiology Report. J Med Internet Res. 2025 Jun 11;27:e72638. doi: 10.2196/72638
29. Dietrich N, Stubbert B. Evaluating Adherence to Canadian Radiology Guidelines for Incidental Hepatobiliary Findings Using RAG-Enabled LLMs. Can Assoc Radiol J. 2025 Feb 27:8465371251323124. doi: 10.1177/08465371251323124
30. Fink A, Rau A, Reisert M, Bamberg F, Russe MF. Retrieval-Augmented Generation with Large Language Models in Radiology: From Theory to Practice. Radiol Artif Intell. 2025 Jul;7(4):e240790. doi: 10.1148/ryai.240790.
31. Tayebi Arasteh S, Lotfinia M, Bressem K, Siepmann R, Adams L, Ferber D, Kuhl C, Kather JN, Nebelung S, Truhn D. RadioRAG: Online Retrieval-augmented Generation for Radiology Question Answering. Radiol Artif Intell. 2025 Jun 18:e240476. doi: 10.1148/ryai.240476.
32. Zhu Z, Liu J, Hong CW, Houshmand S, Wang K, Yang Y. Multimodal Large Language Model With Knowledge Retrieval Using Flowchart Embedding for Forming Follow-Up Recommendations for Pancreatic Cystic Lesions. AJR Am J Roentgenol. 2025 Jul 16:1-12. doi: 10.2214/AJR.25.32729
33. Choi H, Lee D, Kang YK, Suh M. Empowering PET imaging reporting with retrieval-augmented large language models and reading reports database: a pilot single-center study. Eur J Nucl Med Mol Imaging. 2025 Jun;52(7):2452-2462. doi:10.1007/s00259-025-07101-9.
34. Weinert DA, Rauschecker AM. Enhancing Large Language Models with Retrieval-Augmented Generation: A Radiology-Specific Approach. Radiol Artif Intell. 2025 May;7(3):e240313. doi: 10.1148/ryai.240313
35. Xu R, Hong Y, Zhang F, Xu H. Evaluation of the integration of retrieval-augmented generation in a large language model for breast cancer nursing care responses. Sci Rep. 2024 Dec 28;14(1):30794. doi: 10.1038/s41598-024-81052-3
36. Ozmen BB, Mathur P. Evidence-based artificial intelligence: Implementing retrieval-augmented generation models to enhance clinical decision support in plastic surgery. J Plast Reconstr Aesthet Surg. 2025 May;104:414-416. doi: 10.1016/j.bjps.2025.03.053