



# From RAG to Riches? Corporate Control in the AI Research Era



Wilhelmina Randtke, Nikki Canon-Rech, Kevin Reagan

Charleston Conference, Charleston, South Carolina, November 15, 2024

# Abstract

Search engines tout retrieval augmented generation (RAG) as the future of search. There are popular and academic interests in generative artificial intelligence (AI), and how it can enable new methods of research. RAG provides a way to mitigate AI misinformation and improve results. RAG also raises ethical, commercial, and legal issues that information professionals must address. AI-generated misinformation is subject to consumer protection regulations, though Section 230 protections provide immunity from externally generated, website-hosted misinformation. For instance, Meta is not liable for harmful medical information that users produce on Facebook and Instagram, though Meta would be responsible for harmful medical information it produces. Similarly, copyright law affects RAG differently than traditional search, with implications for library collections and scholarly communication. This presentation explores how RAG impacts academic libraries and the research process:

- Challenges in evaluating AI-generated information
- Implications for information literacy instruction
- Effects on collection development and resource licensing
- Potential changes to scholarly publishing models
- Ethical considerations for library-provided AI tools

This presentation examines how centralized corporate control of AI technology influences academic research and library services. By understanding these issues, librarians can better advocate for responsible AI use in academic settings and guide researchers in navigating this evolving landscape.

# Retrieval Augmented Generation (RAG)

= incorporate search results into a generative AI response

RAG solves common  
problems with generative AI

Misinformation:  
Generative AI  
makes things  
up

RAG can  
compare  
against real  
world info



# Background on high training costs

GPT = Generative **Pre-trained** Transformer

Training is expensive

- GPT 3 training cost = over \$12 million
- Hugging Face's BLOOM training cost = around \$10 million

Training requires specialized hardware:

- Training GPT4 took 10,000 NVIDIA chips. (fancy, specialized, more than consumer grade chips)

Experts oversee the training process

# Problem: Outdated Training Data

Huge cost for one training round for a large language model (LLM)

→ If info is out of date, welp, you can't retrain it every day.

RAG can compare to up-to-date sources



## Problem: Outdated Training Data

Application development requires some amount of stability.

→ Predictability may outweigh out of date info in a software process

**RAG updates info without changing software behavior**



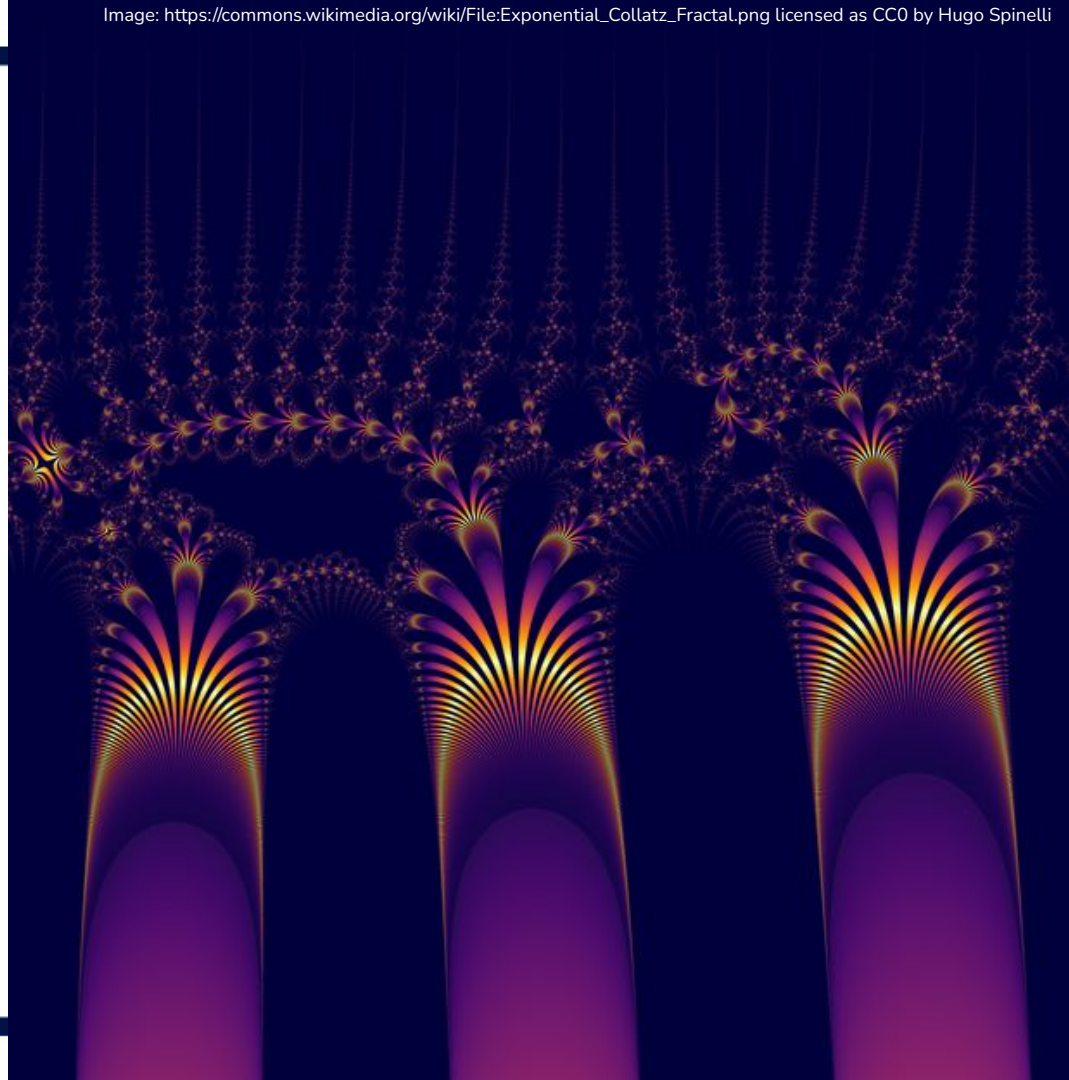


# Problem: Most AI tools are generalist

High training costs mean only a handful of very large tech companies are training models, and training infrequently.

→ Practically everything is built on an API for the same few tools

## RAG brings in domain specific info



# Corporate Benefits of RAG

The image shows a search engine results page for the query "retrieval augmented generation". The page is split into two main sections: a left sidebar and a main content area. The left sidebar contains navigation options like "SEARCH", "COPILOT", "SCHOOL", "VIDEOS", "IMAGES", "MAPS", "NEWS", "SHOPPING", and "MORE". The main content area displays search results, including a featured snippet and a list of links. The featured snippet defines RAG as an advanced AI technique that combines information retrieval with text generation. The list of links includes "What is RAG? - Retrieval-Augmented Generation ...", "What Is Retrieval-Augmented Generation...", and "Retrieval-Augmented Generation (RAG) Guid".

Microsoft Bing

About 373,000 results

**Retrieval-Augmented Generation (RAG)** is an advanced AI technique that combines information retrieval with text generation. It allows AI models to retrieve relevant information from a knowledge source and incorporate it into generated text [1](#) [2](#) [3](#) [4](#) [5](#).

Learn more:

- 1 What is RAG? - Retrieval-Augmented Generation ... [aws.amazon.com](#)
- 2 What Is Retrieval-Augmented Generation... [blogs.nvidia.com](#)
- 3 Retrieval-Augmented Generation (RAG) Guid [datastax.com](#)

AI Overview

Retrieval-augmented generation (RAG) is a framework that combines information retrieval systems with large language models (LLMs) to improve the accuracy and relevance of generated text. RAG works by:

1. **Retrieving and pre-processing:** Using search algorithms to find relevant data from external sources, such as knowledge bases, web pages, and databases
2. **Integrating:** Incorporating the pre-processed information into the LLM

RAG can help LLMs provide more relevant and up-to-date answers to prompts

Show more

# RAG can maintain the status quo for dominant search engines

- Chatbots compete with search.
- → Converging chatbot and search engine may favor already-dominant tech companies.

The screenshot shows the Statista website interface. At the top, there is a navigation bar with the Statista logo, a search bar containing 'Search Statistics', and a menu with items like 'Prices & Access', 'Statistics', 'Reports', 'Insights', 'Research AI' (marked as 'NEW'), 'Daily Data', and 'Services'. Below the navigation bar, the breadcrumb trail reads 'Internet > Online Search'. The main heading is 'AI-powered online search - Statistics & Facts'. The article text discusses the impact of ChatGPT on the search market, mentioning Google and Microsoft. A 'KEY INSIGHTS' sidebar on the right lists: 'Main generative AI usage by U.S. adults Answer a question', 'U.S. adults using AI-powered search tools as first option 13m', and 'Investments in AI for broad use cases (search engines/LLMs) 12.9bn USD'. A 'DIGITAL & TRENDS' section is partially visible at the bottom right.

statista

Search Statistics

Prices & Access Statistics Reports Insights Research AI NEW Daily Data Services

Internet > Online Search

## AI-powered online search - Statistics & Facts

The release of ChatGPT sent shockwaves through various digital sectors, none more so than the [online search market](#). The disruptive modality of artificial intelligence-powered search, led by chatbots engaging users in conversational interactions to deliver results, has thrown giants like [Google](#) and [Microsoft](#) into a frenzied competition with start-up companies and international challengers. Risking companies' investments while casting doubt on the reliability of online information due to inherent inaccuracies of the generated search results, the AI-powered search presents alluring prospects but emerges as a risky gamble for the companies willing to adopt it.

### What is AI-powered online search?

Largely based on large language processing models (LLMs), AI-powered search engines are conversational [generative AIs](#) that combine [deep learning](#) techniques, users' inputs, and growing datasets, to provide search results in a dialogue-style output. Despite slightly different ways of presenting their answers, the revolutionary aspect of turning online search into a chat makes it more interactive and dynamic than the classic search result experience, setting a new paradigm for the industry and opportunities for search advertising.

### The main AI-powered search names

Despite the solid and long-running monopoly of Google's search engine and Microsoft's [Bing](#) competition in the global online search market. San Francisco-based start-up [OpenAI](#) became the

Subscribe

KEY INSIGHTS

Main generative AI usage by U.S. adults  
[Answer a question](#)

U.S. adults using AI-powered search tools as first option  
**13m**

Investments in AI for broad use cases (search engines/LLMs)  
**12.9bn USD**

[Get more insights](#)

DIGITAL & TRENDS

# RAG limits copyright infringement liability

Microsoft Bing

retrieval augmented generation

SEARCH COPILOT SCHOOL VIDEOS IMAGES MAPS NEWS SHOPPING MORE

About 373,000 results

K2view  
<https://www.k2view.com/generative-ai/rag>

## Retrieval Augmented Generation | Make Your GenAI Apps Better

Sponsored Leverage LLMs, RAG, and enterprise data for GenAI apps while ensuring data privacy. Get exclusive Gartner research to get started!  
Realtime Data Integration · Secure and Cost-Effective · Micro-Database Technology  
Service catalog: Anonymization, Pseudonymization, Referential Integrity, PII Discovery

<h3>What is RAG?</h3> <p>Understand the meaning of RAG and how it can help your GenAI apps</p>	<h3>Real-time Data Fusion</h3> <p>K2view RAG Tools Grow sales &amp; customer intimacy</p>
--	---

**Retrieval-Augmented Generation (RAG)** is an advanced AI technique that combines information retrieval with text generation. It allows AI models to retrieve relevant information from a knowledge source and incorporate it into generated text [1](#) [2](#) [3](#) [4](#) [5](#).

Learn more:

- 1 What is RAG? - Retrieval-Augmented Generation ... [aws.amazon.com](https://aws.amazon.com)
- 2 What Is Retrieval-Augmented Generation... [blogs.nvidia.com](https://blogs.nvidia.com)
- 3 Retrieval-Augmented Generation (RAG) Guid [datastax.com](https://datastax.com)

Wikipedia  
<https://en.wikipedia.org/wiki/Retrieval-augmented...>

## Retrieval-augmented generation - Wikipedia

Google

retrieval augmented generation

AI Overview

Retrieval-augmented generation (RAG) is a framework that combines information retrieval systems with large language models (LLMs) to improve the accuracy and relevance of generated text. RAG works by:

1. **Retrieving and pre-processing:** Using search algorithms to find relevant data from external sources, such as knowledge bases, web pages, and databases
2. **Integrating:** Incorporating the pre-processed information into the LLM

RAG can help LLMs provide more relevant and up-to-date answers to prompts

Show more

NVIDIA Blog  
<https://blogs.nvidia.com> · blog · what-is-retrieval-augme...

## What Is Retrieval-Augmented Generation aka RAG

Nov 15, 2023 – Retrieval-augmented generation (RAG) is a technique for enhancing the accuracy and reliability of generative AI models with facts fetched ...

Amazon Web Services  
<https://aws.amazon.com> · ... · Generative AI

## What is RAG (Retrieval-Augmented Generation)?

Retrieval-Augmented Generation (RAG) is the process of optimizing the output of a large language model, so it references an authoritative knowledge base ...

Google Cloud  
<https://cloud.google.com> · use-cases · retrieval-augmen...

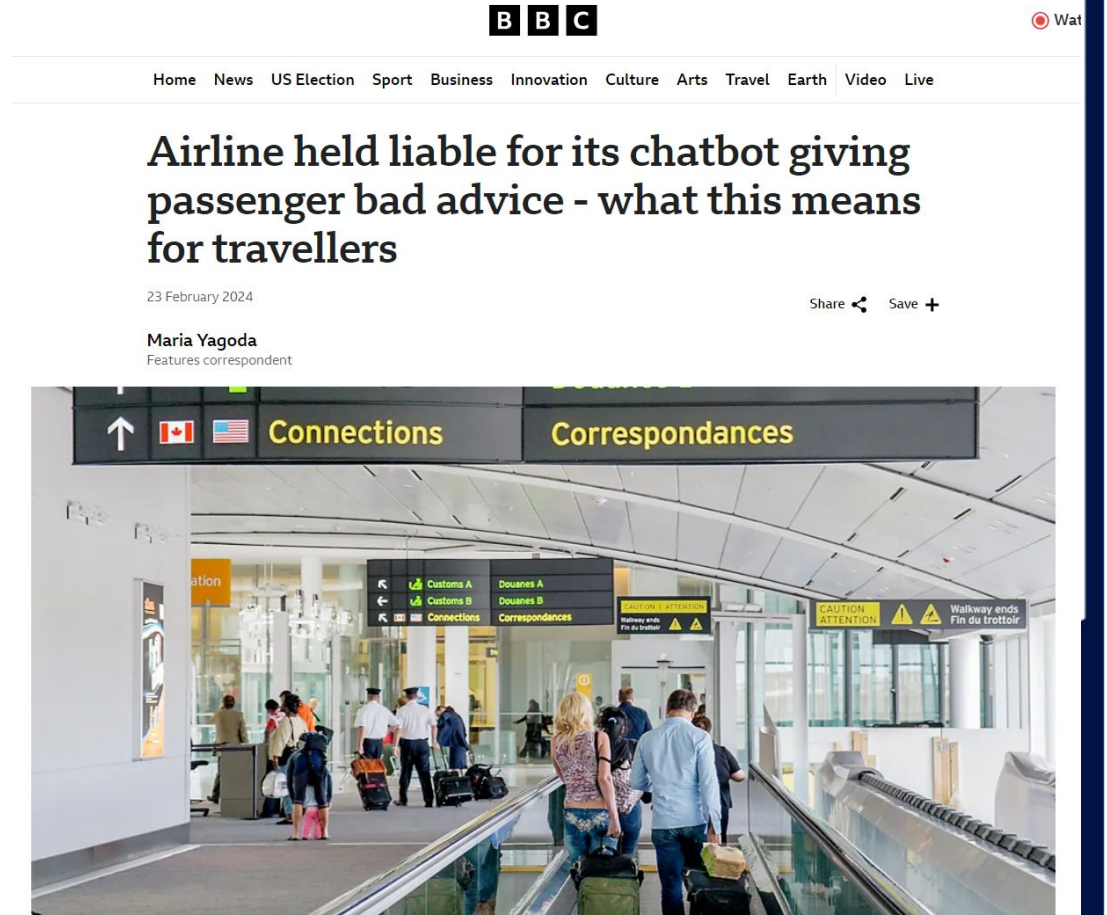
## What is Retrieval-Augmented Generation (RAG)?

Retrieval-augmented generation (RAG) combines LLMs with external knowledge bases to improve their outputs. Learn more with Google Cloud.

# RAG limits misinformation liability

Section 230 protects a web platform hosting content made by someone else.

Chatbot text is not made by someone else, so no section 230 immunity. But... search results are someone else's content.



The image is a screenshot of a BBC news article. At the top, the BBC logo is visible on the left, and a 'Wat' logo is on the right. Below the logo is a navigation bar with links for Home, News, US Election, Sport, Business, Innovation, Culture, Arts, Travel, Earth, Video, and Live. The main headline reads 'Airline held liable for its chatbot giving passenger bad advice - what this means for travellers'. Below the headline, the date '23 February 2024' is shown on the left, and 'Share' and 'Save' icons are on the right. The author's name, 'Maria Yagoda', and her title, 'Features correspondent', are listed below the date. The bottom half of the screenshot shows a photograph of an airport terminal. In the foreground, there is a moving walkway with several people walking. Above the walkway, there are large directional signs. The signs include an upward arrow, the Canadian and US flags, and the words 'Connections' and 'Correspondances'. Below these signs, there are smaller signs for 'Customs A', 'Customs B', 'Douanes A', and 'Douanes B'. To the right, there are 'CAUTION ATTENTION' signs with a yellow triangle and the text 'Walkway ends Fin du trottoir'.



# From RAG to Riches? Corporate Control in the AI Research Era



Wilhelmina Randtke, Nikki Canon-Rech, Kevin Reagan

Charleston Conference, Charleston, South Carolina, November 15, 2024