Retrieval-Augmented Generation (RAG)

retrieval-augmented-generation · safety · concept

Source: https://policywindow.org/wiki/retrieval-augmented-generation

Generated 2026-07-09T20:39:00 UTC

Summary

An AI system pattern in which a model's outputs are conditioned on external content retrieved at inference time from a knowledge source — combining the parametric knowledge of the model with the up-to-date or domain-specific knowledge of the retrieval index.

At a glance

Used by

2 instrument(s)

Related concepts

hallucination, prompt-injection, training-data-attribution, ai-supply-chain, in-context-learning

Primary source

Lewis, P., et al. (2020), 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,' NeurIPS — the canonical articulation of RAG.

Details

Retrieval-augmented generation was formalised by Lewis et al. (2020, 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,' NeurIPS) and is now the dominant pattern for deploying LLMs against proprietary, current, or specialised knowledge. The architecture pattern: at inference time, the user query is used to retrieve k documents from an index (vector store, search engine, structured database); those documents are appended to the prompt context; the model generates an answer conditioned on both its parametric memory and the retrieved context. RAG is the substrate for most enterprise LLM deployments — legal assistants citing case law, customer-support agents citing product docs, medical-AI citing clinical guidelines.

Governance relevance opens a distinct surface from pure-LLM outputs. (a) Provenance — retrieved content has its own source attribution that must flow into the output; this is the technical substrate for citation-verifiability requirements (EU AIA Art. 50 transparency for AI-generated content). (b) Hallucination mitigation — RAG reduces but does not eliminate hallucination, because the model may still misquote or compositionally fabricate from retrieved sources. (c) Indirect prompt injection — the retrieval corpus is a primary adversarial-input vector (Greshake et al. 2023); an attacker who can plant content in the retrievable index can hijack the model. (d) Downstream-misinformation risk — RAG systems that surface low-quality sources amplify them with authoritative voice. (e) IP + training-data overlap — RAG creates a deployment-time analogue of training-data attribution questions, since retrieved-and-paraphrased content may infringe copyright at use-time. NIST AI RMF GenAI Profile §2.7 'Value Chain and Component Integration' is the closest binding frame; EU AI Act Art. 53 GPAI obligations apply to the model but the retrieval-index layer is largely unregulated.

How to cite this article

APA

Policy Window. (n.d.). Retrieval-Augmented Generation (RAG) [Wiki article — Concept]. https://policywindow.org/wiki/retrieval-augmented-generation

CHICAGO

Policy Window. n.d.. "Retrieval-Augmented Generation (RAG)." Wiki article (Concept). https://policywindow.org/wiki/retrieval-augmented-generation.

HARVARD

Policy Window (n.d.) 'Retrieval-Augmented Generation (RAG)', Wiki article — Concept, available at: https://policywindow.org/wiki/retrieval-augmented-generation.

OSCOLA

Policy Window, 'Retrieval-Augmented Generation (RAG)' (Wiki article — Concept, n.d.) <https://policywindow.org/wiki/retrieval-augmented-generation> accessed [date].

BIBTEX

```
@misc{policywindow-retrieval-augmented-generation,
title  = {Retrieval-Augmented Generation (RAG)},
author = {Policy Window},
year   = {n.d.},
howpublished = {retrieval-augmented-generation — safety},
url    = {https://policywindow.org/wiki/retrieval-augmented-generation},
note   = {Primary source: https://arxiv.org/abs/2005.11401}
}
```