Lecture 5: RAG and knowledge pipelines

From source documents to grounded, cited answers

Patrick Hall

Department of Decision Sciences

November 2, 2026

© 2026–2027 Patrick Hall (jphall@gwu.edu). Some rights reserved. Text and copy are licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0).

Agenda

What RAG changes

Build the knowledge pipeline

Retrieve and rank

Generate, cite, and operate

Synthesis

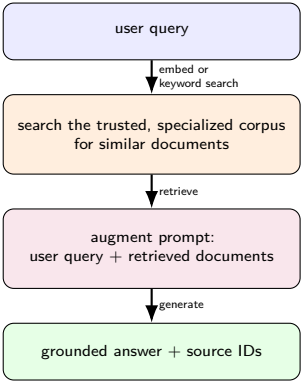
References

RAG adds an explicit, inspectable knowledge path

Retrieval-augmented generation

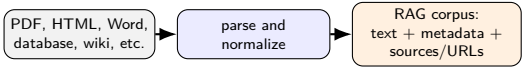
At answer time, retrieve relevant passages from a specialized document corpus, add them to the model context, and generate an answer constrained by those passages (Lewis et al. 2020).

- **Specialize answers** around trusted documents such as policies, manuals, research, or product records.
- **Reduce hallucination risk** by grounding answers in retrieved evidence; relevance and citation support must still be checked.
- **Enable citation** through passage-level source IDs and links.



A query and retrieved documents form an augmented prompt that produces a grounded answer with source IDs.

Ingestion makes a source usable and accountable



Source files are parsed into a RAG corpus that preserves text, metadata, and links to the originating source.

Keep a stable source ID, version or retrieval date, access-control label, title, section, and canonical URL. These fields make later citations and audits possible.

Ingestion checks

- Is the source authorized and current?
- Did extraction retain headings, tables, and links?
- Can each chunk lead a reader back to the original source?

Fixed-length chunking trades context, recall, and attribution

Fixed-length chunks from AI RMF GOVERN 1.1–1.6 (National Institute of Standards and Technology 2023)

Chunk 1 | begins at GOVERN 1.1; 128-token cutoff
GOVERN 1.1 Legal and regulatory requirements involving AI are understood, managed, and documented. **GOVERN 1.2** The characteristics of trustworthy AI are integrated into organizational policies, processes, procedures, and practices. **GOVERN 1.3** Processes, procedures, and practices are in place to determine the needed level of risk management activities based on the organization's risk tolerance. **GOVERN 1.4** The risk management process and its outcomes are established through transparent policies, procedures, and other controls based on organizational risk priorities. **GOVERN 1.5** Ongoing monitoring and periodic review of the risk management process and its outcomes are planned and organizational roles and responsibilities **[cutoff]**

Chunk 2 | 16 repeated tokens + new tokens
[repeat 16] *review of the risk management process and its outcomes are planned and organizational roles and responsibilities.* **[new] GOVERN 1.5** clearly defined, including determining the frequency of periodic review. **GOVERN 1.6** Mechanisms are in place to inventory AI systems and are resourced according to organizational risk priorities. . .

The boundary is tokenizer-specific; here it visibly divides GOVERN 1.5 rather than respecting its subcategory boundary.

Sizing and overlap heuristic

- Start from a meaningful record; do not split a record that is already shorter than the target size.
- Use shorter chunks (for example, 128 tokens) for detail and citation locality; use larger chunks for themes and summaries.
- Use 10–20% overlap for ordinary prose. Here, $16/128 = 12.5\%$; raise it for boundary-crossing clauses, tables, or lists.
- Test size, overlap, retrieval depth k, and answer faithfulness (Lewis et al. 2020).

Semantic chunking respects structure and meaning

AI RMF GOVERN records as structure-aligned chunks (National Institute of Standards and Technology 2023)

Chunk 1 | **GOVERN 1.1** Legal and regulatory requirements involving AI are understood, managed, and documented.

Chunk 2 | **GOVERN 1.2** The characteristics of trustworthy AI are integrated into organizational policies, processes, procedures, and practices.

Chunk 3 | **GOVERN 1.3** Processes, procedures, and practices are in place to determine the needed level of risk management activities based on the organization's risk tolerance.

Chunk 4 | **GOVERN 1.4** The risk management process and its outcomes are established through transparent policies, procedures, and other controls based on organizational risk priorities.

Chunk 5 | **GOVERN 1.5** Ongoing monitoring and periodic review of the risk management process and its outcomes are planned and organizational roles and responsibilities clearly defined, including determining the frequency of periodic review.

Chunk 6 | **GOVERN 1.6** Mechanisms are in place to inventory AI systems and are resourced according to organizational risk priorities.

...

- Use clear document structure when available: headings, sections, table rows, enumerated controls, or other authored boundaries.
- When structure is weak, embed adjacent sentences or paragraphs and start a new chunk when similarity falls below a calibrated threshold (Karpukhin et al. 2020).
- Retain the heading and source ID, and enforce practical minimum and maximum chunk lengths.
- An embedding boundary is not automatically a meaningful policy, legal, or table boundary. Inspect failures.

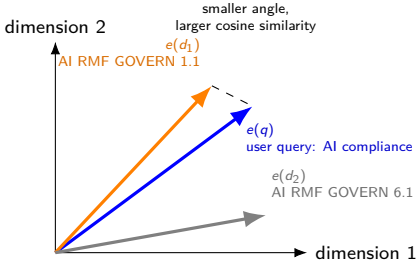
Embeddings represent query–passage similarity

For a query q and chunks d_j, an embedding model maps text to vectors:

$$e(q), e(d_j) \in \mathbb{R}^D,$$

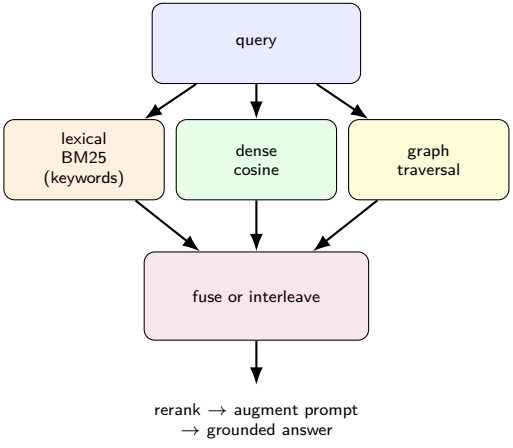
$$s_{\text{dense}}(q, d_j) = \frac{e(q)^\top e(d_j)}{\|e(q)\|_2 \|e(d_j)\|_2}.$$

Rank chunks by s_{dense} and preserve each chunk's source metadata. Dense retrieval can match related language even when query and passage do not share exact terms (Karpukhin et al. 2020; OpenAI 2026).



An AI-compliance query embedding lies closer in direction to the GOVERN 1.1 embedding than to the GOVERN 6.1 embedding; cosine similarity compares those directions.

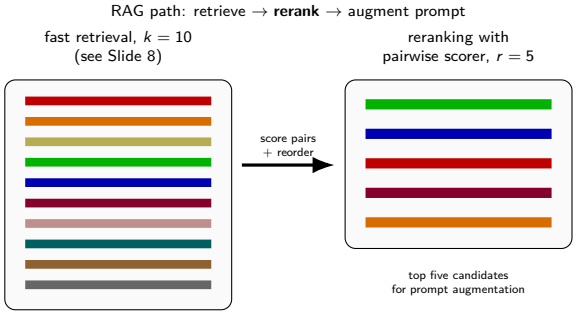
Hybrid search combines complementary signals



Lexical, dense, and graph retrieval candidates are fused.

- Lexical search is strong for exact control IDs, names, and rare terms.
- Dense search is strong for paraphrases and related meanings.
- GraphRAG can surface connected entities, relationships, or community-level context when an answer spans passages (Edge et al. 2024; Microsoft Research 2026).
- Normalize or combine ranks before comparing scores from different systems (Robertson and Zaragoza 2009; Karpukhin et al. 2020).

Reranking spends more computation on fewer candidates



Ten fast-retrieval candidates are scored against the query and reordered; the five highest-ranked candidates continue to prompt augmentation.

- A cross-encoder is a Transformer encoder (such as BERT) that reads the query and one candidate document together, then outputs a relevance score; an LLM judge can play a similar role.
- Reranking is more expensive than first-stage retrieval, so apply it only to a small candidate set.
- Compare recall@k, ranking quality, latency, and support for final claims on held-out questions (Nogueira and Cho 2019).

Grounded generation makes use of clear contracts

System prompt instruction

```
# Grounded-answer policy
- Use only the **Retrieved context** below.
- Add [source ID] after every factual claim.
- If unsupported, say:  "I don't have enough
evidence."
```

Example retrieved context

`[MS-2.1]` Test sets, metrics, and details about the tools used during testing, evaluation, verification, and validation are documented.

`[MG-4.1]` Post-deployment AI-system monitoring plans include mechanisms for user and stakeholder input, appeal and override, decommissioning, incident response, recovery, and change management.

Example cited answer

Question: What should an organization document and include in its monitoring plan?

Document the test sets, metrics, and tools used during testing, evaluation, verification, and validation (TEVV) `[MS-2.1]`. Include user input, appeals, incident response, recovery, and change management in the monitoring plan `[MG-4.1]`.

References:
`[MS-2.1]` NIST AI RMF Playbook, Measure 2.1.
`[MG-4.1]` NIST AI RMF Playbook, Manage 4.1.

Citations need coverage and correctness checks

Check	Question
Coverage	Does each material claim have at least one citation?
Entailment	Does the cited passage actually support that claim?
Completeness	Were relevant passages missed before generation?
Traceability	Can a reader open the source and find the passage?

Citations do not establish faithfulness.

Test retrieval independently (Lewis et al. 2020).

A claim is linked through a source ID to its passage and URL, enabling support checks and failure triage.

Claim

Document the test sets, metrics, and tools used during testing, evaluation, verification, and validation (TEVV) [MS-2.1].

Source ID

[MS-2.1] resolves in the NIST AI RMF Playbook to: *"Test sets, metrics, and details about the tools used during testing, evaluation, verification, and validation are documented."*

flag unsupported or over-broad claims

Tool chains are replaceable components, not the architecture

SDKs

Call services directly; choose and connect every component yourself.

- OpenAI SDK
- Anthropic SDK
- Google GenAI SDK

Frameworks

Compose integrations and chains; inspect defaults and generated requests.

- LangChain
- LlamaIndex
- Haystack

Products

Package workflows; still verify sources, permissions, and citations.

- Gemini Notebook
- AnythingLLM
- Bedrock Knowledge Bases

Every option should make RAG stages, citations, and cross-cutting controls inspectable.

A RAG evaluation starts before the demo

- Stratify the authorized corpus, then generate diverse questions.
- Score recall, relevance, groundedness, completeness, and safety.
- Calibrate automated scores with human labels; retain uncertainty.
- Stress test ambiguous, adversarial, and out-of-distribution inputs.
- Slice failures by topic, query type, and risk; revise and repeat.

Lecture 7 covers AI evaluation in depth.

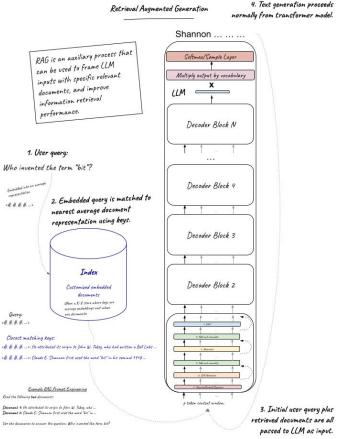


Source: Sudjianto et al. (2024), *Human-Calibrated Automated Testing and Validation of Generative Language Models*.

Colored clusters illustrate topical strata in a document collection, supporting diverse automated test generation for a RAG system.

Putting the pieces together

- A user makes a query.
- The retrieval system searches for document chunks similar to the query.
- The LLM prompt is augmented with the retrieved documents; system instructions provide citation guidelines.
- The prompt is submitted to the LLM or generation system.
- A grounded response is generated with citations.



A query retrieves relevant documents, which augment the LLM prompt before a cited answer is generated.

References

Edge, Darren, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson (2024). *From Local to Global: A Graph RAG Approach to Query-Focused Summarization*. arXiv:2404.16130. URL: https://arxiv.org/abs/2404.16130.

Karpukhin, Vladimir, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih (2020). "Dense Passage Retrieval for Open-Domain Question Answering". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 6769–6781. URL: https://aclanthology.org/2020.emnlp-main.550/.

Lewis, Patrick et al. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in Neural Information Processing Systems*. Vol. 33. URL: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Nogueira, Rodrigo and Kyunghyun Cho (2019). *Passage Re-ranking with BERT*. arXiv:1901.04085. URL: https://arxiv.org/abs/1901.04085.

Robertson, Stephen and Hugo Zaragoza (2009). "The Probabilistic Relevance Framework: BM25 and Beyond". In: *Foundations and Trends in Information Retrieval* 3.4, pp. 333–389. URL: https://doi.org/10.1561/1500000019.

References

Microsoft Research (2026). *GraphRAG Documentation*. URL: https://microsoft.github.io/graphrag/.

National Institute of Standards and Technology (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. Tech. rep. NIST AI 100-1. U.S. Department of Commerce. URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

OpenAI (2026). *Embeddings Guide*. URL: https://platform.openai.com/docs/guides/embeddings.