

warning "Backend update (October 2026)" GitHub Models inference was retired on July 30, 2026. Any references below are retained only for historical context while this lab is migrated. Use Lab 013 ? Free LLM backends to configure Foundry Local or Ollama before continuing.

Lab 006: What is RAG?

Level: L50 Path: ? RAG Time: ~20 min ? Cost: Free ? No account needed

What You'll Learn

Why LLMs need external knowledge (and why training data alone isn't enough) The complete RAG pipeline: ingest ? chunk ? embed ? store ? retrieve ? generate The difference between keyword search, semantic search, and hybrid search When to use RAG vs. fine-tuning vs. just a bigger context window Real-world RAG architectures

Introduction

Imagine you've built an AI agent for your company. It answers questions beautifully ? until a user asks about a product launched last month, or a policy updated last week.

The LLM doesn't know. Its training data has a cutoff date. And even if the information existed in training data, the model may not have memorized it accurately.

RAG (Retrieval-Augmented Generation) solves this by connecting the LLM to your own, up-to-date knowledge at query time ? without retraining the model.

RAG Pipeline

? Supporting Files

note "Download these files before starting the lab" Save all files to a lab-006/ folder in your working directory.

File | Description | Download | faq_backpacks.txt | FAQ knowledge base file | ? Download | faq_clothing.txt | FAQ knowledge base file | ? Download | faq_footwear.txt | FAQ knowledge base file | ? Download | faq_sleeping_bags.txt | FAQ knowledge base file | ? Download | faq_tents.txt | FAQ knowledge base file | ? Download |

Part 1: The Core Problem

LLMs have two knowledge limitations:

Limitation | Description | Example | Training cutoff | Knowledge stops at a date | "What happened last week?" | Private data | Never saw your documents | "What's our refund policy?" | Hallucination risk | May confabulate when uncertain | Invents plausible-sounding but wrong answer |

The naive solution ? "just put all your documents in the prompt" ? doesn't scale. A 500-page manual is ~375,000 tokens. Most LLMs cap at 128,000 tokens, and even if they didn't, you'd pay for all those tokens on every single query.

RAG's answer: Only retrieve the *relevant* pieces, right when you need them.

Part 2: The RAG Pipeline

RAG has two distinct phases:

Phase 1 ? Ingestion (runs once, or on schedule)

RAG Phase 1 ? Ingestion

Phase 2 ? Retrieval + Generation (runs on every query)

RAG Phase 2 ? Retrieval + Generation

question "? Check Your Understanding" In the RAG pipeline, what is the difference between the ingestion phase and the retrieval phase, and how often does each run?

??? success "Answer" The ingestion phase (Load ? Chunk ? Embed ? Store) runs once (or on a schedule) to prepare your documents. The retrieval phase (Embed query ? Search ? Augment ? Generate) runs on every user query to find relevant chunks and generate an answer.

Ingestion is a batch process; retrieval is real-time.

Part 3: Chunking Strategies

How you split documents matters enormously.

Strategy | How | Best for | Fixed-size | Split every N tokens | Simple, fast, works for most cases | Sentence/paragraph | Split at natural boundaries | Better context preservation | Semantic | Split when topic changes | Best quality, more complex | Recursive | Try paragraph ? sentence ? word | Good default for mixed content |

Overlap is important. If you split at exactly 512 tokens, relevant information that spans a boundary gets lost. Add 50?100 token overlap between chunks.

[diagram or code sample]

question "? Check Your Understanding" Why is it important to add overlap between chunks when splitting documents?

??? success "Answer" Without overlap, relevant information that spans a chunk boundary gets split across two chunks and may be lost during retrieval. Adding 50?100 token overlap ensures that context at the edges is preserved in both adjacent chunks, improving retrieval quality.

Part 4: Search Types

Keyword Search (BM25) Traditional search ? matches exact words. Fast, interpretable, but misses synonyms and intent.

[diagram or code sample]

Semantic Search (Vector) Compares meaning, not words. Finds conceptually similar content.

[diagram or code sample]

Hybrid Search (BM25 + Vector) Best of both worlds ? combines keyword and semantic scores.

[diagram or code sample]

Most production RAG systems use hybrid search because it handles both exact lookups ("SKU-12345") and semantic queries ("something for camping in the rain").

question "? Check Your Understanding" A user searches for "rain-resistant coat" but the document only contains the phrase "waterproof jacket." Will keyword search find it? Will semantic search? Why?

??? success "Answer" Keyword search will miss it ? there are no matching words between "rain-resistant coat" and "waterproof jacket." Semantic search will find it because it compares meaning via vector similarity, not exact words. Both phrases have very similar meanings and will have similar vector representations. This is why hybrid search (combining both) is preferred in production.

Part 5: RAG vs. Fine-tuning vs. Big Context

A common question: "Why not just fine-tune the model on my data?"

Approach | Cost | Freshness | Best for | RAG | Low | ? Real-time | Dynamic data, documents, Q&A | Fine-tuning | High | ? Static | Tone, style, domain vocabulary | Big context | Medium | ? Per-request | Small datasets that fit in context | RAG + Fine-tuning | High | ?

Real-time | Production systems needing both |

Rule of thumb: Use RAG for *knowledge* (facts, documents). Use fine-tuning for *behavior* (tone, style, format). They're complementary, not competing.

Part 6: RAG Quality Metrics

A RAG system can fail in two places:

Failure | Symptom | Fix | Bad retrieval | Retrieved chunks aren't relevant | Better chunking, hybrid search, re-ranking | Bad generation | LLM ignores or misuses retrieved content | Stronger system prompt, citation enforcement |

Key metrics used in Lab 035 and the planned Lab 042:

Groundedness: Is the answer supported by retrieved documents? Relevance: Are the retrieved

chunks actually relevant to the question? Coherence: Is the answer well-structured and readable? Faithfulness: Does the answer stay true to the source material?

Real-World RAG Architectures

Basic RAG [diagram or code sample]

Agentic RAG (covered in Lab 026) [diagram or code sample]

Corrective RAG [diagram or code sample]

? Knowledge Check

question "Q1 (Multiple Choice): What does RAG stand for, and what problem does it primarily solve?"

- A) Recursive Augmented Graph ? it solves multi-step reasoning problems - B)

Retrieval-Augmented Generation ? it grounds LLM answers in private or up-to-date data

without retraining - C) Randomized Agent Generation ? it makes agent responses less

deterministic - D) Ranked Answer Generation ? it improves ranking of search results

??? success "? Reveal Answer" Correct: B ? Retrieval-Augmented Generation

RAG connects the LLM to your own knowledge at query time. Instead of the model relying on training data (which has a cutoff date and doesn't include your private documents), RAG retrieves the most relevant chunks from your data store and includes them in the prompt. No retraining needed.

question "Q2 (Multiple Choice): In the RAG ingestion pipeline, which step comes immediately BEFORE storing vectors in the database?"

- A) Chunking - B) Loading documents - C) Generating embeddings - D) Semantic reranking

??? success "? Reveal Answer" Correct: C ? Generating embeddings

The ingestion order is: Load ? Chunk ? Embed ? Store. You first load raw documents, split them into smaller chunks (~512 tokens with overlap), *then* convert each chunk to a vector embedding using the embedding model, and finally store those vectors in the vector database. Semantic reranking is a retrieval step, not an ingestion step.

question "Q3 (Run the Lab): Open the file lab-006/faq_tents.txt. How many Q&A pairs does it contain, and what is the topic of the LAST question?"

Open lab-006/faq_tents.txt and count the Q&A pairs. The last question starts with "Q:".

??? success "? Reveal Answer" 5 Q&A pairs. The last question is: "Can I use a 2-person tent solo?"

? faq_tents.txt contains exactly 5 Q&A pairs covering: solo backpacking tent choice, 3-season vs 4-season, waterproofing, pole materials, and using a 2P tent solo. This is the kind of knowledge base a RAG system would ingest ? each Q&A pair is a natural chunk for embedding.

Summary

Concept | Key takeaway | Why RAG | LLMs don't know your data or recent events ? RAG fixes

this | Ingestion | Load ? Chunk ? Embed ? Store (runs once) | Retrieval | Embed query ?

Vector search ? Top-k chunks | Chunking | Size + overlap matter; ~512 tokens with 50-token

overlap | Search | Hybrid (keyword + semantic) beats either alone | Evaluation | Measure

groundedness and relevance ? both retrieval and generation |

Next Steps

Understand the embedding vectors: ? Lab 007 ? What are Embeddings? Build a RAG app for

free: ? Lab 022 ? RAG with GitHub Models + pgvector Production RAG on Azure: ? Lab 031 ?

pgvector Semantic Search on Azure

Supporting kit

The site build generates stable PDF and ZIP downloads. The ZIP contains the lab guide and every supporting file declared for this lab.

Supporting file: [faq_backpacks.txt](#)

OutdoorGear Inc. - Frequently Asked Questions

BACKPACKS & DAYPACKS =====

Q: What backpack size do I need for a weekend trip? A: For a 2-3 night trip with lightweight gear: 40-55 liters. For 2-3 nights with standard gear: 55-70 liters. For 4+ nights: 65-80+ liters. The Osprey Atmos 65L (P006) is our most popular 3-season backpacking pack. For day hikes, the DayHiker 22L (P007) fits all essentials. Rule of thumb: a full pack should weigh no more than 20% of your body weight.

Q: How do I fit a backpack correctly? A: Fitting has three steps. Step 1: Measure your torso length (nape of neck to iliac crest on your hip). Most packs come in sizes S/M/L based on torso length ? not body height. Step 2: Load the pack and put it on. The hipbelt padded sections should wrap around your hip bones, not your waist. Tighten the hipbelt first ? 70% of pack weight should sit on your hips. Step 3: Tighten shoulder straps until they curve over your shoulders. Adjust load lifters so the pack top tilts slightly toward your head.

Q: What features should I look for in a backpacking pack? A: Essential features: proper fit system (torso sizing), padded hip belt, top lid with pockets, front panel or side access to main compartment, hydration reservoir compatibility, attachment points for trekking poles and ice axe. Nice-to-have: rain cover, stretchy exterior pockets, removable top lid that converts to a daypack, aluminum frame stays.

Q: How do I pack a backpack efficiently? A: Three layers: Bottom layer (sleeping bag, camp shoes ? light and bulky items rarely needed during the day). Middle layer (clothing, food ? heaviest items packed close to your back and as high as possible). Top layer (items needed during the day: first aid, rain gear, snacks, map). External pockets: water bottles, snacks, trekking poles.

Q: What is the warranty on OutdoorGear backpacks? A: All OutdoorGear packs come with a lifetime guarantee against manufacturing defects. Normal wear and tear (worn straps, zipper pulls) is covered for 5 years from purchase with proof of receipt. To claim warranty, contact support@outdoorgear.example.com with your order number and photos of the defect.

Supporting file: [faq_clothing.txt](#)

OutdoorGear Inc. - Frequently Asked Questions

CLOTHING & LAYERING SYSTEM =====

Q: What is the three-layer system? A: The layering system optimizes warmth, moisture management, and weather protection. Layer 1 ? Base layer (next to skin): wicks moisture away from your body. Never use cotton ? it stays wet and causes hypothermia. Use merino wool (warm, odor-resistant, soft) or synthetic (dries faster, cheaper). Layer 2 ? Mid layer (insulation): traps warm air. Options: fleece (breathable, fast-drying, affordable), down jacket (warmest per gram, compresses small), synthetic insulated jacket (insulates when wet, cheaper than down). Layer 3 ? Shell (outer): blocks wind and rain. Hardshell: fully waterproof, best for heavy rain/snow. Softshell: more breathable, water-resistant only, better for active use in light rain.

Q: How do I choose between a hardshell and softshell jacket? A: Hardshell: choose when you need guaranteed waterproofing in heavy, sustained rain or snow. Look for Gore-Tex or equivalent membrane, sealed seams, and a waterproof zipper. Softshell: choose for high-output activities (ski touring, hiking in light rain) where breathability matters more

than waterproofing. A wet softshell still provides wind protection. Many experienced outdoor athletes carry both: softshell for active use, hardshell packed for emergencies.

Q: What does DWR mean, and why does it wear off? A: DWR (Durable Water Repellent) is a finish applied to the outer fabric of waterproof jackets that causes water to bead and roll off instead of soaking in. When DWR wears off from use and washing, the outer fabric gets "wet out" ? it absorbs water and becomes heavy and cold, even though the waterproof membrane underneath still works. Restore DWR by washing the garment with a DWR wash (like Nikwax Tech Wash) and tumble drying on low heat.

Q: How should I care for waterproof clothing? A: Wash regularly ? dirty waterproof garments lose breathability as oils and dirt block the membrane pores. Use a technical wash (Nikwax Tech Wash or Grangers Performance Wash). Avoid fabric softeners ? they coat the DWR. Tumble dry on low heat after washing to reactivate DWR. Iron on very low heat if needed (protects the waterproofing).

Q: What base layer material is best for warm-weather hiking? A: For warm weather, lightweight synthetic (polyester) is optimal ? it's the fastest-drying and most affordable. Merino wool is excellent for multi-day trips where you need odor resistance (you can wear it 2-3 days without washing). Avoid cotton base layers entirely ? wet cotton against skin is a leading cause of hypothermia even in mild temperatures.

Supporting file: faq_footwear.txt

OutdoorGear Inc. - Frequently Asked Questions

FOOTWEAR & ACCESSORIES =====

Q: What type of hiking boots should I buy? A: Choose based on terrain and load. Trail running shoes / low-cut hiking shoes: best for well-maintained trails, light loads (under 10 kg), warm weather. More agile and faster-drying. Mid-cut hiking boots: most versatile. Ankle support without full rigidity. Good for off-trail, moderate loads, varied terrain. High-cut mountaineering boots: for scrambling, snow, heavy loads (15+ kg), multi-day backpacking with heavy pack. More waterproof and durable.

Q: Should my hiking footwear be waterproof? A: Waterproof boots (Gore-Tex membrane) are great for wet conditions, stream crossings, and cold weather ? feet stay dry longer. Trade-off: waterproof membranes breathe less ? feet get sweaty on warm days, leading to blisters. In dry conditions, non-waterproof boots are often preferred for breathability. A practical approach: waterproof boots for spring/fall, non-waterproof mesh boots for summer.

Q: How do I break in new hiking boots? A: Wear them around the house for a few hours. Then short walks (1-2 km). Then day hikes. Finally, multi-day trips. Never wear new boots on a long trail without breaking them in ? you will get severe blisters. The break-in period is typically 2-4 weeks of occasional use. Hiking socks matter: wear merino wool or synthetic hiking socks (not cotton), with proper cushioning. Using liner socks reduces friction.

Q: What trekking pole features are most important? A: Key features: adjustable length (fits different terrain: shorter uphill, longer downhill), cork or foam grip (reduces vibration, comfortable in cold/wet), anti-shock system (reduces joint impact on downhill), weight (ultralight carbon fiber vs. affordable aluminum), and basket type (small baskets for summer, large snow baskets for winter/mud).

Q: Do I need trekking poles? A: Research shows trekking poles reduce knee stress by up to 25% on downhill sections ? significant on long descents. Benefits: knee protection, better balance on uneven terrain, help when crossing streams, useful as tent poles for ultralight shelters. Most hikers over 40, those with knee problems, or anyone doing long descents benefit greatly. For flat trail running, poles add unnecessary weight and restrict arm

swing.

Supporting file: faq_sleeping_bags.txt

OutdoorGear Inc. - Frequently Asked Questions

SLEEPING BAGS & PADS =====

Q: What temperature rating sleeping bag do I need? A: Choose a bag rated 5-10°C colder than the coldest conditions you expect. This accounts for natural variation in how warm different people sleep. For 3-season camping (down to 0°C nights): use a 0°C or -5°C bag. For winter camping at -10°C: use our ArcticDown -20°C bag (P004). When in doubt, go colder ? you can always vent the zipper if you're too warm.

Q: Down vs. synthetic insulation ? which should I choose? A: Down insulation: warmer per gram, compresses smaller, lasts longer. Weakness: loses insulation when wet. Best for: cold, dry conditions, long trips where weight matters. Synthetic insulation: insulates when wet, cheaper, easier to wash, dries faster. Weakness: heavier and bulkier than equivalent down. Best for: wet environments, casual camping, budget-conscious buyers. Our ArcticDown bags use 800-fill power responsibly sourced down.

Q: What is a sleeping bag's "comfort rating" vs "lower limit"? A: The EN/ISO 13537 standard defines three ratings. Comfort rating: temperature at which a standard woman sleeps comfortably. Lower limit: temperature at which a standard man sleeps without waking. Extreme rating: survival limit (not comfortable, hypothermia risk). Always buy based on the Comfort rating if you tend to sleep cold, or Lower limit if you sleep warm.

Q: How do I store a sleeping bag? A: Never store a sleeping bag compressed in its stuff sack for long periods. The loft (insulation's ability to trap air) degrades permanently if compressed for months. Store it loosely in a large cotton storage sack, hanging, or flat under a bed. Only use the stuff sack during trips.

Q: Can I wash my sleeping bag? A: Yes. Use a front-loading washing machine (never top-loading ? the agitator damages down). Use a down-specific cleaner (like Nikwax Down Wash Direct) on cold, gentle cycle. Dry on low heat with three clean tennis balls to break up down clumps. This takes 2-3 cycles ? patience is key. The bag is dry only when there are zero damp clumps.

Supporting file: faq_tents.txt

OutdoorGear Inc. - Frequently Asked Questions

TENTS & SHELTERS =====

Q: What tent should I buy for solo backpacking? A: For solo backpacking, we recommend the TrailBlazer Solo (P003) for ultra-light trips under 900g, or the TrailBlazer Tent 2P (P001) if you want extra space or sometimes camp with a partner. Both are 3-season tents rated for temperatures down to 0°C.

Q: What is the difference between a 3-season and 4-season tent? A: A 3-season tent (spring, summer, fall) is designed for moderate weather with good ventilation and lighter poles. A 4-season tent handles winter conditions: it has stronger poles for snow load, less mesh for warmth, and a steeper geometry to shed snow. The Summit Dome 4P is our 4-season option.

Q: How do I waterproof a tent? A: Apply a tent seam sealer to all stitched seams on the rain fly and floor. Use a DWR (Durable Water Repellent) spray on the rain fly fabric when water stops beading. Re-apply DWR after every 10-15 uses or when you notice water soaking into the fabric instead of beading off.

Q: What is the pole material in OutdoorGear tents? A: Our entry-level tents use aluminum 7075 poles. The Summit Dome 4P uses DAC Featherlite poles, which are 15% lighter and

stronger than standard aluminum. Carbon fiber poles are available as an upgrade for the TrailBlazer Solo.

Q: Can I use a 2-person tent solo? A: Yes, many backpackers prefer a 2-person tent solo for the extra gear storage space and improved livability. The trade-off is extra weight. If weight is critical (long thru-hikes), choose the solo version. For car camping or weekend trips, the 2P offers much more comfort.