

warning "Atualização de backend (outubro de 2026)" A inferência do GitHub Models foi descontinuada em 30 de julho de 2026. As referências abaixo são mantidas apenas como contexto histórico enquanto este lab é migrado. Use o Lab 013 ? Backends de LLM gratuitos para configurar o Foundry Local ou o Ollama antes de continuar.

Lab 006: O que é RAG?

Nível: L50 Trilha: ? RAG Tempo: ~20 min ? Custo: Gratuito ? Nenhuma conta necessária
O que Você Vai Aprender

Por que LLMs precisam de conhecimento externo (e por que os dados de treinamento sozinhos não são suficientes) O pipeline completo de RAG: ingestão ? fragmentação ? embedding ? armazenamento ? recuperação ? geração A diferença entre busca por palavras-chave, busca semântica e busca híbrida Quando usar RAG vs. ajuste fino vs. apenas uma janela de contexto maior Arquiteturas RAG do mundo real

Introdução

Imagine que você construiu um agente de IA para sua empresa. Ele responde perguntas lindamente ? até que um usuário pergunte sobre um produto lançado no mês passado, ou uma política atualizada na semana passada.

O LLM não sabe. Seus dados de treinamento têm uma data de corte. E mesmo que a informação existisse nos dados de treinamento, o modelo pode não tê-la memorizado com precisão.

RAG (Geração Aumentada por Recuperação) resolve isso conectando o LLM ao seu próprio conhecimento atualizado no momento da consulta ? sem retreinar o modelo.

Pipeline RAG

? Arquivos de Apoio

note "Baixe estes arquivos antes de iniciar o lab" Salve todos os arquivos em uma pasta lab-006/ no seu diretório de trabalho.

Arquivo | Descrição | Download | [faq_backpacks.txt](#) | Arquivo de base de conhecimento FAQ |
? Download | [faq_clothing.txt](#) | Arquivo de base de conhecimento FAQ | ? Download |
[faq_footwear.txt](#) | Arquivo de base de conhecimento FAQ | ? Download | [faq_sleeping_bags.txt](#)
| Arquivo de base de conhecimento FAQ | ? Download | [faq_tents.txt](#) | Arquivo de base de conhecimento FAQ | ? Download |

Parte 1: O Problema Central

LLMs têm duas limitações de conhecimento:

Limitação | Descrição | Exemplo | Corte de treinamento | O conhecimento para em uma data |
"O que aconteceu na semana passada?" | Dados privados | Nunca viu seus documentos | "Qual é nossa política de reembolso?" | Risco de alucinação | Pode confabular quando incerto |
Inventa resposta que soa plausível mas está errada |

A solução ingênua ? "coloque todos os seus documentos no prompt" ? não escala. Um manual de 500 páginas tem ~375.000 tokens. A maioria dos LLMs limita em 128.000 tokens, e mesmo que não limitassem, você pagaria por todos esses tokens em cada consulta individual.

A resposta do RAG: Recupere apenas as partes *relevantes*, exatamente quando você precisa delas.

Parte 2: O Pipeline RAG

RAG tem duas fases distintas:

Fase 1 ? Ingestão (executa uma vez, ou por agendamento)

RAG Fase 1 ? Ingestão

Fase 2 ? Recuperação + Geração (executa em cada consulta)

RAG Fase 2 ? Recuperação + Geração

question "? Verifique Seu Entendimento" No pipeline RAG, qual é a diferença entre a fase de ingestão e a fase de recuperação, e com que frequência cada uma executa?

??? success "Resposta" A fase de ingestão (Carregar ? Fragmentar ? Fazer Embedding ? Armazenar) executa uma vez (ou por agendamento) para preparar seus documentos. A fase de recuperação (Fazer embedding da consulta ? Buscar ? Aumentar ? Gerar) executa a cada consulta do usuário para encontrar fragmentos relevantes e gerar uma resposta. A ingestão é um processo em lote; a recuperação é em tempo real.

Parte 3: Estratégias de Fragmentação

Como você divide documentos importa enormemente.

Estratégia | Como | Melhor para | Tamanho fixo | Divide a cada N tokens | Simples, rápido, funciona para a maioria dos casos | Sentença/parágrafo | Divide em limites naturais | Melhor preservação de contexto | Semântica | Divide quando o tópico muda | Melhor qualidade, mais complexo | Recursiva | Tenta parágrafo ? sentença ? palavra | Bom padrão para conteúdo misto |

Sobreposição é importante. Se você dividir exatamente em 512 tokens, informação relevante que cruza um limite se perde. Adicione 50?100 tokens de sobreposição entre fragmentos.

[diagram or code sample]

question "? Verifique Seu Entendimento" Por que é importante adicionar sobreposição entre fragmentos ao dividir documentos?

??? success "Resposta" Sem sobreposição, informação relevante que cruza um limite de fragmento fica dividida entre dois fragmentos e pode se perder durante a recuperação.

Adicionar 50?100 tokens de sobreposição garante que o contexto nas bordas seja preservado em ambos os fragmentos adjacentes, melhorando a qualidade da recuperação.

Parte 4: Tipos de Busca

Busca por Palavras-chave (BM25) Busca tradicional ? corresponde palavras exatas. Rápida, interpretável, mas perde sinônimos e intenção.

[diagram or code sample]

Busca Semântica (Vetorial) Compara significado, não palavras. Encontra conteúdo conceitualmente similar.

[diagram or code sample]

Busca Híbrida (BM25 + Vetorial) O melhor dos dois mundos ? combina pontuações de palavras-chave e semânticas.

[diagram or code sample]

A maioria dos sistemas RAG em produção usa busca híbrida porque ela lida tanto com buscas exatas ("SKU-12345") quanto com consultas semânticas ("algo para acampar na chuva").

question "? Verifique Seu Entendimento" Um usuário busca por "rain-resistant coat" mas o documento contém apenas a frase "waterproof jacket." A busca por palavras-chave vai encontrá-lo? E a busca semântica? Por quê?

??? success "Resposta" A busca por palavras-chave vai perder ? não há palavras correspondentes entre "rain-resistant coat" e "waterproof jacket." A busca semântica vai encontrar porque ela compara significado via similaridade vetorial, não palavras exatas.

Ambas as frases têm significados muito similares e terão representações vetoriais similares. É por isso que a busca híbrida (combinando ambas) é preferida em produção.

Parte 5: RAG vs. Ajuste Fino vs. Contexto Grande

Uma pergunta comum: "Por que não simplesmente ajustar finamente o modelo nos meus dados?"

Abordagem | Custo | Atualidade | Melhor para | RAG | Baixo | ? Tempo real | Dados dinâmicos, documentos, Q&A | Ajuste fino | Alto | ? Estático | Tom, estilo, vocabulário de domínio | Contexto grande | Médio | ? Por requisição | Conjuntos de dados pequenos que cabem no contexto | RAG + Ajuste fino | Alto | ? Tempo real | Sistemas em produção que precisam de ambos |

Regra prática: Use RAG para *conhecimento* (fatos, documentos). Use ajuste fino para *comportamento* (tom, estilo, formato). Eles são complementares, não concorrentes.

Parte 6: Métricas de Qualidade RAG

Um sistema RAG pode falhar em dois pontos:

Falha | Sintoma | Correção | Recuperação ruim | Fragmentos recuperados não são relevantes | Melhor fragmentação, busca híbrida, re-ranking | Geração ruim | LLM ignora ou usa mal o conteúdo recuperado | Prompt de sistema mais forte, exigência de citação |

Métricas-chave usadas no Lab 035 e no Lab 042 planejado:

Fundamentação: A resposta é suportada pelos documentos recuperados? Relevância: Os fragmentos recuperados são realmente relevantes para a pergunta? Coerência: A resposta é bem estruturada e legível? Fidelidade: A resposta permanece fiel ao material de origem?

Arquiteturas RAG do Mundo Real

RAG Básico [diagram or code sample]

RAG Agêntico (abordado no Lab 026) [diagram or code sample]

RAG Corretivo [diagram or code sample]

? Verificação de Conhecimento

question "Q1 (Múltipla Escolha): O que significa RAG, e qual problema ele resolve principalmente?"

- A) Recursive Augmented Graph ? resolve problemas de raciocínio multi-etapas - B)

Retrieval-Augmented Generation ? fundamenta respostas do LLM em dados privados ou

atualizados sem retreinamento - C) Randomized Agent Generation ? torna respostas de agentes menos determinísticas - D) Ranked Answer Generation ? melhora o ranking de resultados de busca

??? success "? Revelar Resposta" Correta: B ? Retrieval-Augmented Generation

RAG conecta o LLM ao seu próprio conhecimento no momento da consulta. Em vez do modelo depender dos dados de treinamento (que têm uma data de corte e não incluem seus documentos privados), o RAG recupera os fragmentos mais relevantes do seu armazenamento de dados e os inclui no prompt. Sem necessidade de retreinamento.

question "Q2 (Múltipla Escolha): No pipeline de ingestão RAG, qual etapa vem imediatamente ANTES de armazenar vetores no banco de dados?"

- A) Fragmentação - B) Carregamento de documentos - C) Geração de embeddings - D)

Re-ranking semântico

??? success "? Revelar Resposta" Correta: C ? Geração de embeddings

A ordem de ingestão é: Carregar ? Fragmentar ? Fazer Embedding ? Armazenar. Você primeiro carrega documentos brutos, divide-os em fragmentos menores (~512 tokens com sobreposição), *depois* converte cada fragmento em um embedding vetorial usando o modelo de embedding, e finalmente armazena esses vetores no banco de dados vetorial. Re-ranking semântico é uma etapa de recuperação, não de ingestão.

question "Q3 (Execute o Lab): Abra o arquivo lab-006/faq_tents.txt. Quantos pares de P&R ele contém, e qual é o tópico da ÚLTIMA pergunta?"

Abra lab-006/faq_tents.txt e conte os pares de P&R. A última pergunta começa com "Q:".

??? success "? Revelar Resposta" 5 pares de P&R. A última pergunta é: "Can I use a 2-person tent solo?"

? faq_tents.txt contém exatamente 5 pares de P&R cobrindo: escolha de barraca para mochileiro solo, 3-estações vs 4-estações, impermeabilização, materiais de hastes e uso de barraca 2P solo. Este é o tipo de base de conhecimento que um sistema RAG ingeriria ? cada par de P&R é um fragmento natural para embedding.

Resumo

Conceito | Principal aprendizado | Por que RAG | LLMs não conhecem seus dados ou eventos recentes ? RAG corrige isso | Ingestão | Carregar ? Fragmentar ? Fazer Embedding ? Armazenar (executa uma vez) | Recuperação | Fazer embedding da consulta ? Busca vetorial ? Top-k fragmentos | Fragmentação | Tamanho + sobreposição importam; ~512 tokens com 50 tokens de sobreposição | Busca | Híbrida (palavras-chave + semântica) supera qualquer uma isoladamente | Avaliação | Meça fundamentação e relevância ? tanto recuperação quanto geração |

Próximos Passos

Entenda os vetores de embedding: ? Lab 007 ? O que são Embeddings? Construa um app RAG gratuitamente: ? Lab 022 ? RAG com GitHub Models + pgvector RAG em produção no Azure: ? Lab 031 ? Busca Semântica pgvector no Azure

Kit de apoio

O build do site gera downloads estáveis em PDF e ZIP. O ZIP contém o guia do lab e todos os arquivos de apoio declarados para este lab.

Baixar PDF gerado Baixar ZIP gerado

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.