

warning "Mise à jour du backend (octobre 2026)" L'inférence GitHub Models a été retirée le 30 juillet 2026. Les références ci-dessous sont conservées uniquement comme contexte historique pendant la migration de ce lab. Utilisez le Lab 013 ? Backends LLM gratuits pour configurer Foundry Local ou Ollama avant de continuer.

Lab 006 : Qu'est-ce que le RAG ?

Niveau : L50 Parcours : ? RAG Durée : ~20 min ? Coût : Gratuit ? Aucun compte requis

Ce que vous apprendrez

Pourquoi les LLM ont besoin de connaissances externes (et pourquoi les données d'entraînement seules ne suffisent pas) Le pipeline RAG complet : ingestion ? découpage ? embedding ? stockage ? récupération ? génération La différence entre la recherche par mots-clés, la recherche sémantique et la recherche hybride Quand utiliser le RAG vs l'affinage vs simplement une plus grande fenêtre de contexte Des architectures RAG du monde réel

Introduction

Imaginez que vous avez construit un agent IA pour votre entreprise. Il répond aux questions magnifiquement ? jusqu'à ce qu'un utilisateur demande un produit lancé le mois dernier, ou une politique mise à jour la semaine dernière.

Le LLM ne sait pas. Ses données d'entraînement ont une date de coupure. Et même si l'information existait dans les données d'entraînement, le modèle peut ne pas l'avoir mémorisée avec précision.

RAG (Génération Augmentée par la Récupération) résout ce problème en connectant le LLM à vos propres connaissances à jour au moment de la requête ? sans ré-entraîner le modèle.

Pipeline RAG

? Fichiers de support

note "Téléchargez ces fichiers avant de commencer le lab" Enregistrez tous les fichiers dans un dossier lab-006/ de votre répertoire de travail.

Fichier | Description | Téléchargement | faq_backpacks.txt | Fichier de base de connaissances FAQ | ? Télécharger | faq_clothing.txt | Fichier de base de connaissances FAQ | ? Télécharger | faq_footwear.txt | Fichier de base de connaissances FAQ | ? Télécharger | faq_sleeping_bags.txt | Fichier de base de connaissances FAQ | ? Télécharger | faq_tents.txt | Fichier de base de connaissances FAQ | ? Télécharger |

Partie 1 : Le problème fondamental

Les LLM ont deux limitations de connaissances :

Limitation | Description | Exemple | Date de coupure | Les connaissances s'arrêtent à une date | « Que s'est-il passé la semaine dernière ? » | Données privées | N'a jamais vu vos documents | « Quelle est notre politique de remboursement ? » | Risque d'hallucination | Peut confabuler en cas d'incertitude | Invente une réponse plausible mais fausse |

La solution naïve ? « mettre tous vos documents dans le prompt » ? ne passe pas à l'échelle. Un manuel de 500 pages représente environ 375 000 tokens. La plupart des LLM plafonnent à 128 000 tokens, et même s'ils ne le faisaient pas, vous paieriez pour tous ces tokens à chaque requête.

La réponse du RAG : Ne récupérer que les morceaux *pertinents*, exactement quand vous en avez besoin.

Partie 2 : Le pipeline RAG

Le RAG comporte deux phases distinctes :

Phase 1 ? Ingestion (exécutée une fois, ou de façon programmée)

RAG Phase 1 ? Ingestion

Phase 2 ? Récupération + Génération (exécutée à chaque requête)

RAG Phase 2 ? Récupération + Génération

question "? Vérifiez votre compréhension" Dans le pipeline RAG, quelle est la différence entre la phase d'ingestion et la phase de récupération, et à quelle fréquence chacune s'exécute-t-elle ?

??? success "Réponse" La phase d'ingestion (Charger ? Découper ? Vectoriser ? Stocker) s'exécute une fois (ou de façon programmée) pour préparer vos documents. La phase de récupération (Vectoriser la requête ? Rechercher ? Augmenter ? Générer) s'exécute à chaque requête utilisateur pour trouver les morceaux pertinents et générer une réponse.

L'ingestion est un processus par lots ; la récupération est en temps réel.

Partie 3 : Stratégies de découpage

La façon dont vous découpez les documents est extrêmement importante.

Stratégie | Comment | Idéale pour | Taille fixe | Découper tous les N tokens | Simple, rapide, fonctionne pour la plupart des cas | Phrase/paragraphe | Découper aux limites naturelles | Meilleure préservation du contexte | Sémantique | Découper quand le sujet change | Meilleure qualité, plus complexe | Récursive | Essayer paragraphe ? phrase ? mot | Bon défaut pour le contenu mixte |

Le chevauchement est important. Si vous découpez exactement à 512 tokens, les informations pertinentes qui chevauchent une limite sont perdues. Ajoutez un chevauchement de 50 à 100 tokens entre les morceaux.

[diagram or code sample]

question "? Vérifiez votre compréhension" Pourquoi est-il important d'ajouter un chevauchement entre les morceaux lors du découpage des documents ?

??? success "Réponse" Sans chevauchement, les informations pertinentes qui chevauchent une limite de morceau sont réparties entre deux morceaux et peuvent être perdues lors de la récupération. Ajouter un chevauchement de 50 à 100 tokens garantit que le contexte aux bords est préservé dans les deux morceaux adjacents, améliorant la qualité de la récupération.

Partie 4 : Types de recherche

Recherche par mots-clés (BM25) Recherche traditionnelle ? correspond aux mots exacts.

Rapide, interprétable, mais manque les synonymes et l'intention.

[diagram or code sample]

Recherche sémantique (vectorielle) Compare le sens, pas les mots. Trouve du contenu conceptuellement similaire.

[diagram or code sample]

Recherche hybride (BM25 + vectorielle) Le meilleur des deux mondes ? combine les scores par mots-clés et sémantiques.

[diagram or code sample]

La plupart des systèmes RAG en production utilisent la recherche hybride car elle gère à la fois les recherches exactes (« SKU-12345 ») et les requêtes sémantiques (« quelque chose pour camper sous la pluie »).

question "? Vérifiez votre compréhension" Un utilisateur cherche « rain-resistant coat » mais le document ne contient que l'expression « waterproof jacket ». La recherche par mots-clés le trouvera-t-elle ? La recherche sémantique ? Pourquoi ?

??? success "Réponse" La recherche par mots-clés ne le trouvera pas ? il n'y a pas de mots correspondants entre « rain-resistant coat » et « waterproof jacket ». La recherche sémantique le trouvera car elle compare le sens via la similarité vectorielle, pas les mots exacts. Les deux expressions ont des significations très similaires et auront des représentations vectorielles similaires. C'est pourquoi la recherche hybride (combinant les

deux) est préférée en production.

Partie 5 : RAG vs Affinage vs Grand contexte

Une question courante : « Pourquoi ne pas simplement affiner le modèle sur mes données ? »

Approche | Coût | Fraîcheur | Idéale pour | RAG | Faible | ? Temps réel | Données dynamiques, documents, Q&R | Affinage | Élevé | ? Statique | Ton, style, vocabulaire métier | Grand contexte | Moyen | ? Par requête | Petits jeux de données qui tiennent dans le contexte | RAG + Affinage | Élevé | ? Temps réel | Systèmes de production nécessitant les deux |

Règle de base : Utilisez le RAG pour les *connaissances* (faits, documents). Utilisez l'affinage pour le *comportement* (ton, style, format). Ils sont complémentaires, pas en concurrence.

Partie 6 : Métriques de qualité RAG

Un système RAG peut échouer à deux endroits :

Échec | Symptôme | Correction | Mauvaise récupération | Les morceaux récupérés ne sont pas pertinents | Meilleur découpage, recherche hybride, re-classement | Mauvaise génération | Le LLM ignore ou mal utilise le contenu récupéré | Prompt système plus fort, obligation de citation |

Métriques clés utilisées dans le Lab 035 et le Lab 042 prévu :

Ancrage : La réponse est-elle étayée par les documents récupérés ? Pertinence : Les morceaux récupérés sont-ils réellement pertinents par rapport à la question ? Cohérence : La réponse est-elle bien structurée et lisible ? Fidélité : La réponse reste-t-elle fidèle au matériel source ?

Architectures RAG du monde réel

RAG basique [diagram or code sample]

RAG agentique (couvert dans le Lab 026) [diagram or code sample]

RAG correctif [diagram or code sample]

? Vérification des connaissances

question "Q1 (Choix multiple) : Que signifie RAG, et quel problème résout-il principalement ?"

- A) Recursive Augmented Graph ? il résout les problèmes de raisonnement multi-étapes - B) Retrieval-Augmented Generation ? il ancre les réponses LLM dans des données privées ou à jour sans ré-entraînement - C) Randomized Agent Generation ? il rend les réponses des agents moins déterministes - D) Ranked Answer Generation ? il améliore le classement des résultats de recherche

??? success "? Voir la réponse" Correct : B ? Retrieval-Augmented Generation

Le RAG connecte le LLM à vos propres connaissances au moment de la requête. Au lieu que le modèle se fie aux données d'entraînement (qui ont une date de coupure et n'incluent pas vos documents privés), le RAG récupère les morceaux les plus pertinents de votre magasin de données et les inclut dans le prompt. Aucun ré-entraînement nécessaire.

question "Q2 (Choix multiple) : Dans le pipeline d'ingestion RAG, quelle étape vient immédiatement AVANT le stockage des vecteurs dans la base de données ?"

- A) Le découpage - B) Le chargement des documents - C) La génération des embeddings - D) Le re-classement sémantique

??? success "? Voir la réponse" Correct : C ? La génération des embeddings

L'ordre d'ingestion est : Charger ? Découper ? Vectoriser ? Stocker. Vous chargez d'abord les documents bruts, les découpez en morceaux plus petits (~512 tokens avec chevauchement), *puis* convertissez chaque morceau en un vecteur d'embedding à l'aide du modèle d'embedding, et enfin stockez ces vecteurs dans la base de données vectorielle. Le

re-classement sémantique est une étape de récupération, pas d'ingestion.

question "Q3 (Exécutez le lab) : Ouvrez le fichier lab-006/faq_tents.txt. Combien de paires Q&R contient-il, et quel est le sujet de la DERNIÈRE question ?"

Ouvrez lab-006/faq_tents.txt et comptez les paires Q&R. La dernière question commence par « Q: ».

??? success "?" Voir la réponse" 5 paires Q&R. La dernière question est : « Can I use a 2-person tent solo? »

? faq_tents.txt contient exactement 5 paires Q&R couvrant : le choix d'une tente de randonnée solo, 3 saisons vs 4 saisons, l'imperméabilisation, les matériaux des arceaux, et l'utilisation d'une tente 2P en solo. C'est le type de base de connaissances qu'un système RAG ingérerait ? chaque paire Q&R est un morceau naturel pour l'embedding.

Résumé

Concept | Point clé | Pourquoi le RAG | Les LLM ne connaissent pas vos données ni les événements récents ? le RAG corrige cela | Ingestion | Charger ? Découper ? Vectoriser ? Stocker (exécuté une fois) | Récupération | Vectoriser la requête ? Recherche vectorielle ? Top-k morceaux | Découpage | La taille + le chevauchement comptent ; ~512 tokens avec 50 tokens de chevauchement | Recherche | L'hybride (mots-clés + sémantique) surpasse chacune seule | Évaluation | Mesurer l'ancrage et la pertinence ? tant pour la récupération que pour la génération |

Prochaines étapes

Comprendre les vecteurs d'embedding : ? Lab 007 ? Que sont les embeddings ? Construire une application RAG gratuitement : ? Lab 022 ? RAG avec GitHub Models + pgvector RAG en production sur Azure : ? Lab 031 ? Recherche sémantique pgvector sur Azure

Kit d'accompagnement

Le build du site génère des téléchargements PDF et ZIP stables. Le ZIP contient le guide du lab et tous les fichiers d'accompagnement déclarés.

Télécharger le PDF généré Télécharger le ZIP généré

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.