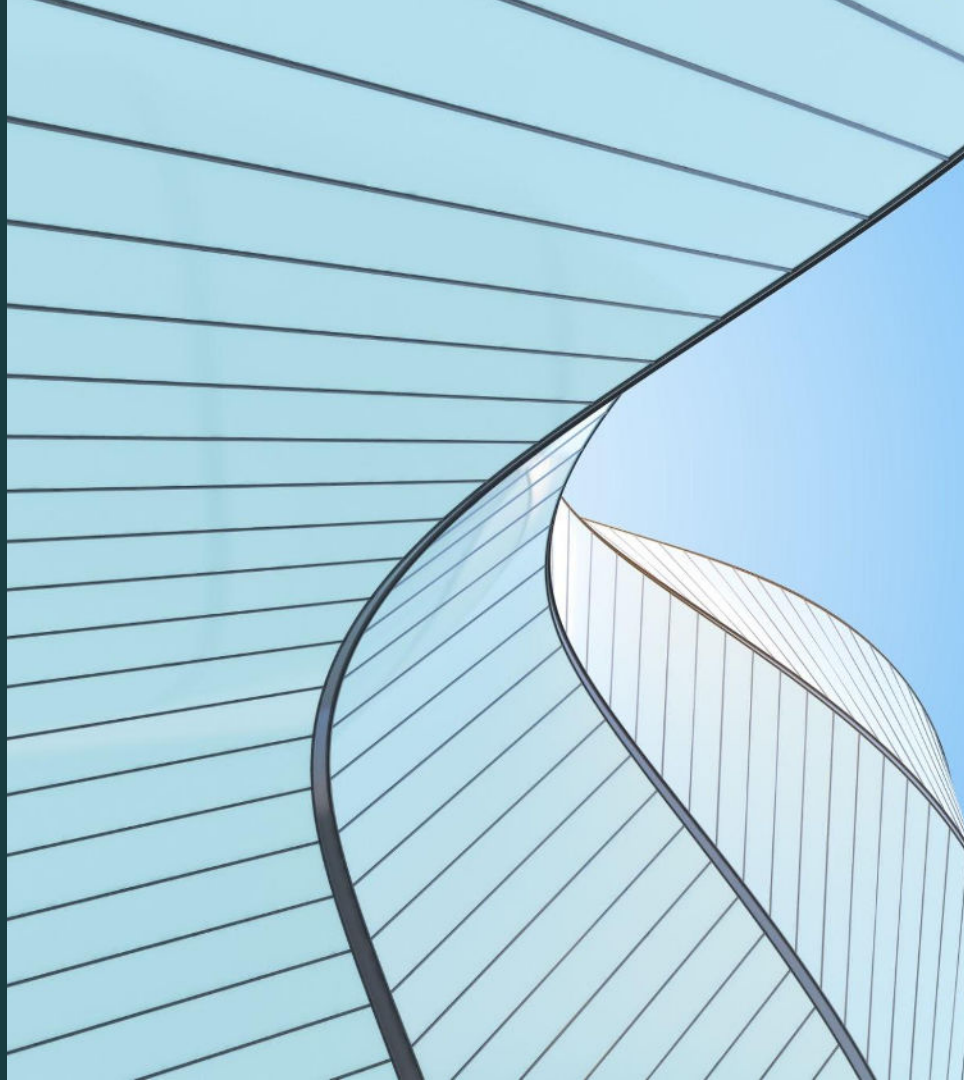
RAG LLM & Tool Calling

Maximilien Andile
Séminaire Aristote
28 Nov 2025

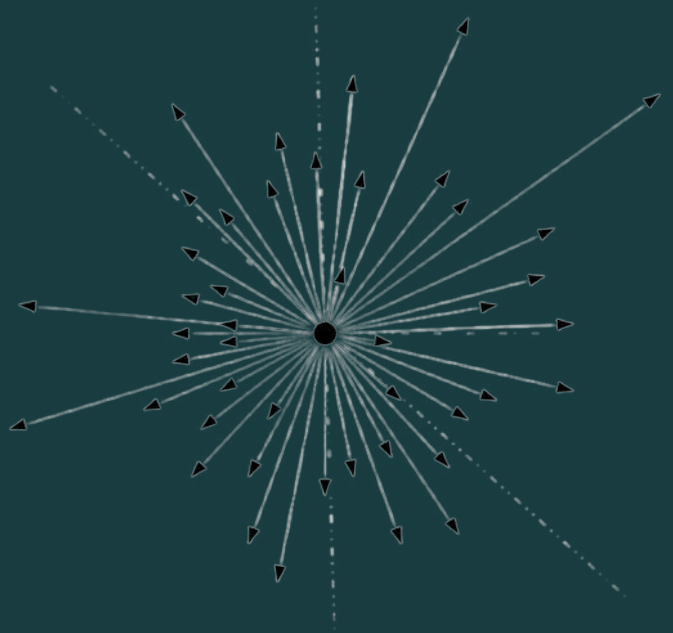


3:26

Florin Pățan

What is it ?

Embeddings - Vectorisation



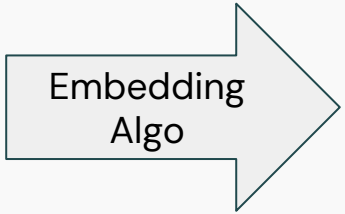
Embedding

Problème : Les ordinateurs ne peuvent pas comprendre directement le texte humain, nous avons besoin d'un format numérique

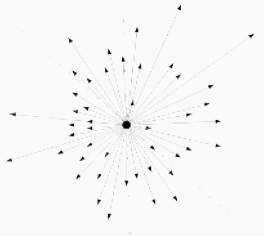
Solution : Transformer le texte en une liste de nombres qui

- Capture la **Syntaxe** (**sparse** vectors)
- Capture **sens et contexte** (Vecteurs denses)

I love golang



[0.016312562 0.010191596
-0.045032285 -0.012354587
-0.01766443 -0.06242634
-0.0018334733 0.048577186
-0.012076703 …
-0.062306173]

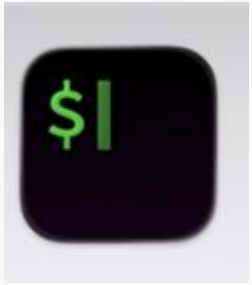


2013 : Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Word2vec => *Limitation:* Only represented single words, not whole sentences.

https://www.pinecone.io/learn/series/nlp/dense-vector-embeddings-nlp/

Wilhelm E J Klein



Démo 1 🦜🔗 LangChainGo

Comment vectoriser un texte en utilisant langchaingo

Modèles :

- Google - gemini-embedding-001
- OpenAI - text-embedding-3-small
- Mistral

Contexte

BOAMP

Détail de l'avis 25-124767

Avis n° 25-124767
Publié le 9 novembre 2025
Date limite de réponse le 22/12/2025 à 17h00

Tierce Maintenance Applicative du Système d'Informations MEDOC et applications associées (lot 1) et des Applications bancaires (lot 2)

DÉPARTEMENT : 93
ACHETEUR : Direction Générale Finances Publiques
TYPE D'AVIS : Avis de marché
PROCÉDURE : Procédure Ouverte

cmd/index-public-offers

Démo 2

Vectorisation + l'indexation des appels d’offres

What is it ?

RAG
=
Retrieval-Augmented Generation
-
Génération à enrichissement contextuel

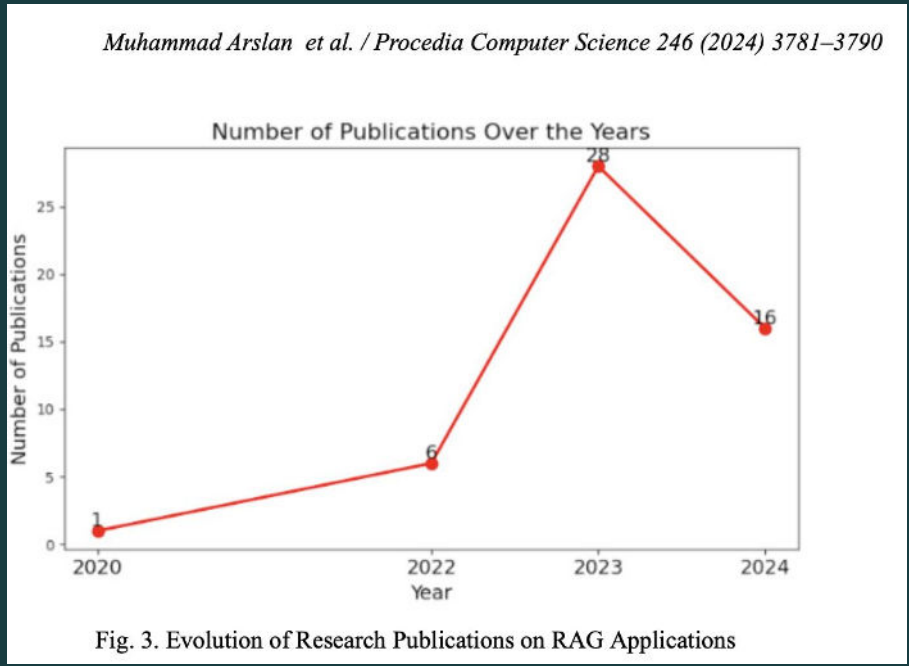


Fig. 3. Evolution of Research Publications on RAG Applications

"Lorsqu'on leur pose des questions spécifiques à un domaine en dehors de leurs données d'entraînement, les LLM peuvent générer des informations incorrectes ou des 'hallucinations'.

Le RAG pallie cette limitation en récupérant des informations à partir d'une source de données externe, qui sont ensuite transmises comme informations contextuelles au modèle LLM pour la génération de la réponse."

Fatehkia, M., Lucas, J. K., & Chawla, S. (2024). T-RAG: lessons from the LLM trenches. *arXiv preprint arXiv:2402.07483*.

30:54

Arslan, M., Ghanem, H., Munawar, S., & Cruz, C. (2024). A Survey on RAG with LLMs. *Procedia computer science*, *246*, 3781-3790.

Où les RAGs sont-ils utilisés et pourquoi ?

Discipline

Discipline: (Count of Publications)

Medical / Biomedical: (9)

Financial: (2)

Educational: (2)

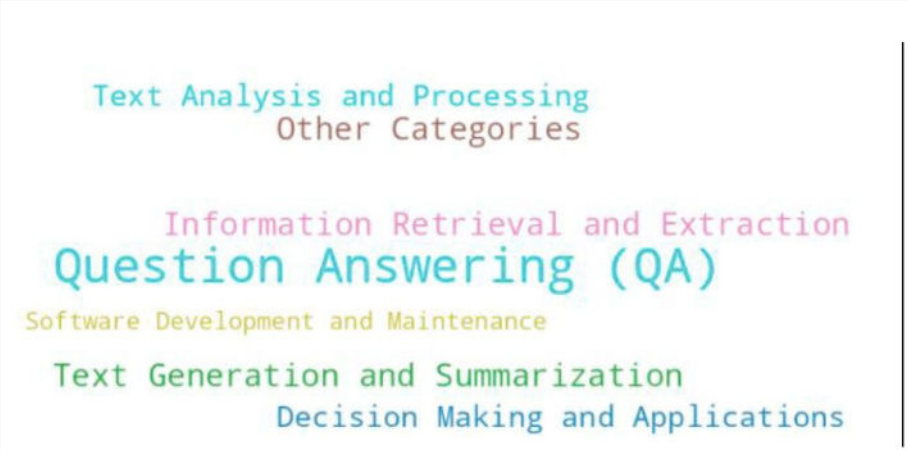
Technology and Software Development: (9)

Social and Communication: (7)

Literature (3)

Other Categories: (8)

Tasks



Task: (Count of Publications)

Question Answering (QA): (20)

Text Generation and Summarization: (6)

Information Retrieval and Extraction: (6)

Text Analysis and Processing: (5)

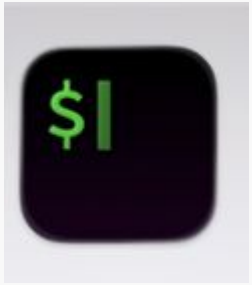
Software Development and Maintenance: (4)

Decision Making and Applications: (5)

Other Categories: (6)

Arslan, M., Ghanem, H., Munawar, S., & Cruz, C. (2024). A Survey on RAG with LLMs. *Procedia computer science*, *246*, 3781-3790.

cmd/chat-with-memory/main.go



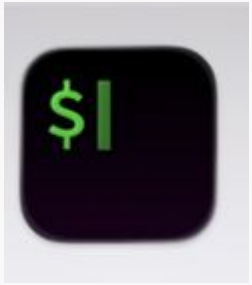
Démo 3

Construire une interface de discussion avec mémoire

https://llm-go-frontend.vercel.app/chat

Test =>

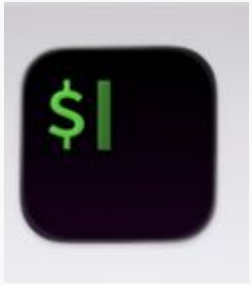
internal/server/chat.go



Démo 4

Construire une interface de chat avec DynamoDB pour la mémoire + déployée sur AWS

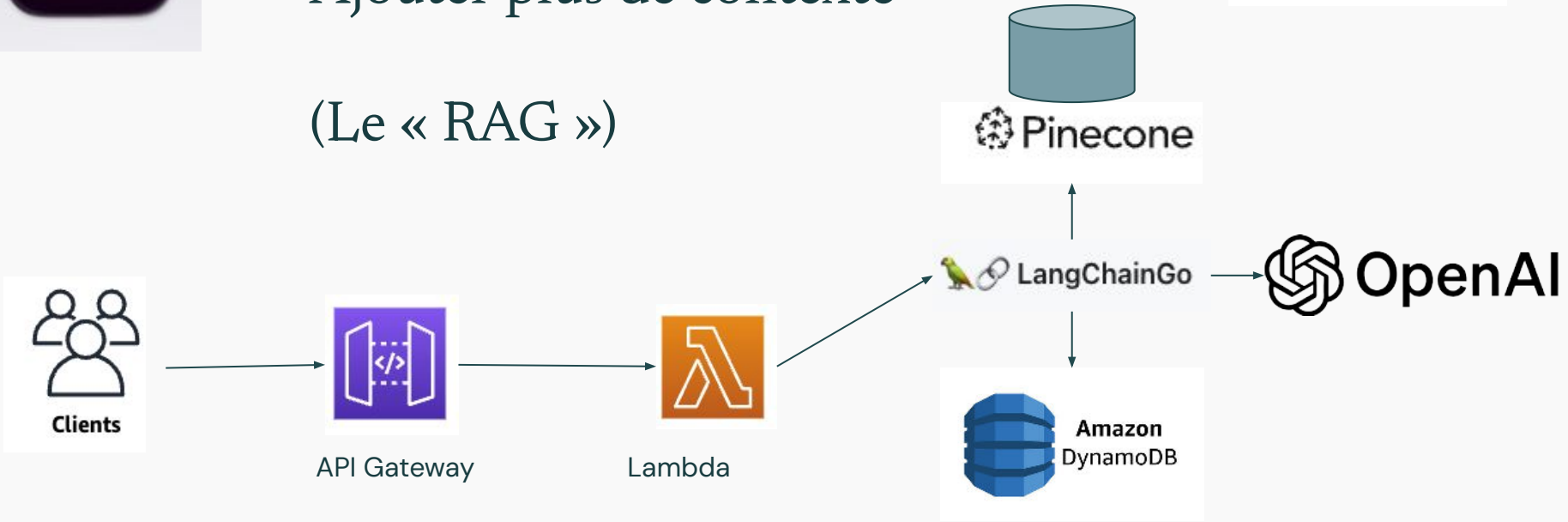
internal/server/chat-with-rag.go



Démo 5

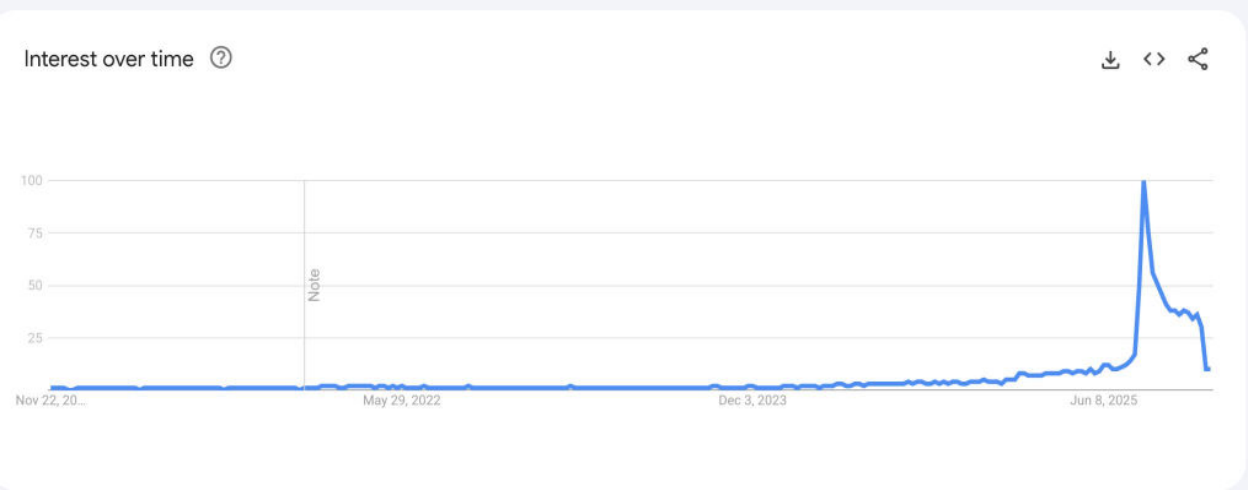
Ajouter plus de contexte

(Le « RAG »)

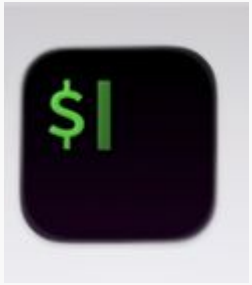


What is it ?

Tool Calling - Outillage des LLM



internal/server/chat-with-rag_with-tools.g
o



Démo 6

Ajouter l'appel d'outil

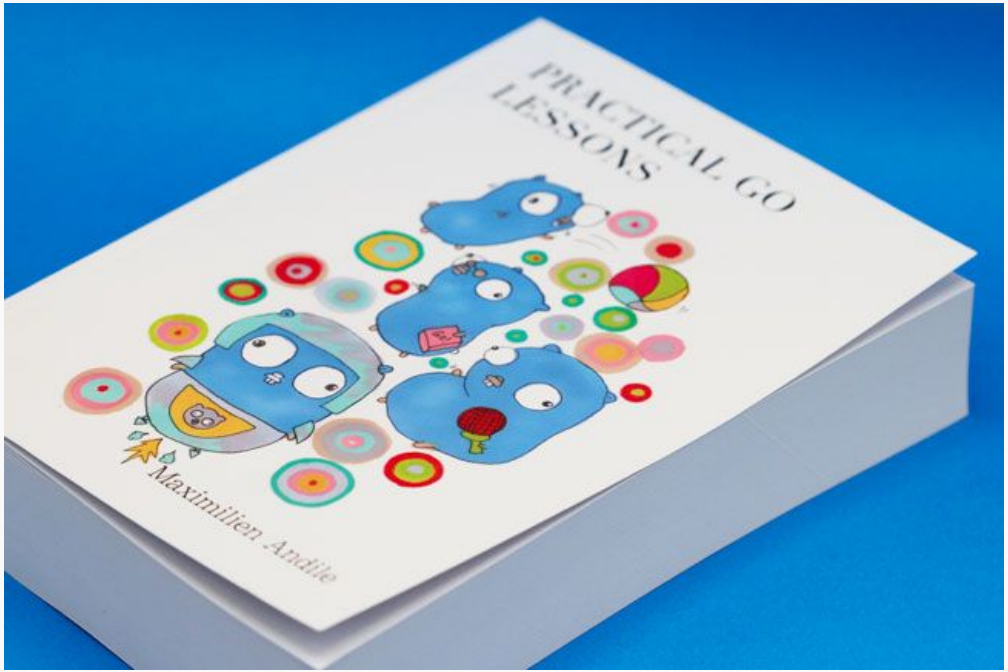
- search_offer
- send_email

API Gateway

Lambda

- Inadéquation de la **Dimension** Vectorielle
  - Problème : Dimension du vecteur (ex. 1536) différente de celle de l'index (ex. 1024).
  - Solution : Réduire la dimension du vecteur de sortie du modèle d'intégration.
- Modèles de **vectorisation Incohérents**
  - Problème : Modèles différents pour l'indexation (A) et la recherche de requêtes (B).
  - Solution : Utiliser le même modèle pour l'intégration du texte et de la requête (cohérence de similarité).
- Sécurité et **Contrôle de l'Appel d'Outil** (Tool Calling)
  - Abus d'Outil/Vecteur d'Attaque
    - Problème : Lister les outils crée un vecteur d'attaque.
    - Solution : Obfusquer les noms d'outils (ex. remboursement => courge24) + interdire l'invite système.
  - Manque de Contrôle
    - Problème : Le modèle accède aux outils sans confirmation.
    - Solution : Mettre en œuvre un contrôle (ex. révision par un LLM séparé) avant l'exécution.
- Difficulté de Découpage des Données (Data **Chunking**)
  - Problème : Découpage précis des données difficile.
  - Note : Pas de solution universelle simple ; dépend du contexte.
- Explosion **Budgétaire**
  - Causes : Coûts élevés des LLM et des bases de données vectorielles.
  - Atténuation : Surveiller les prix et explorer les LLM locaux.

Merci !



Golang Paris Meetup
practical-go-lessons.com