L'intelligence artificielle au service de la découvrabilité des contenus scientifiques en français : potentiel, limites et pistes de réflexion

Marie-Jean Meurs

92ème Congrès de l'ACFAS

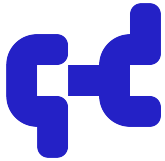
05 mai 2025

Chaire de recherche du Québec
Découvrabilité des contenus scientifiques en français

Fonds de recherche Québec

CIRST
Centre interuniversitaire de recherche sur la science et la technologie

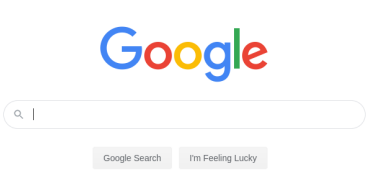
Recherche d'information (RI)

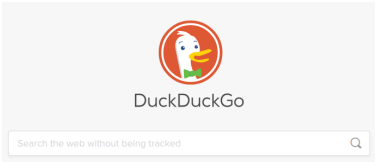


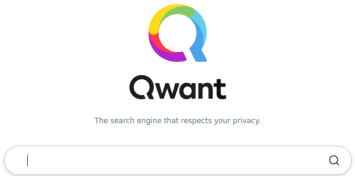
- Une personne a besoin d'une information.
- Elle rédige une **requête** décrivant l'information cherchée.
- Le système de recherche d'information retrouve un ensemble de documents pertinents.
- Les documents sont **ordonnés** par ordre de pertinence.

En entrée : une requête et une **collection de documents**.
En sortie : une liste de documents ordonnés contenant les informations demandées dans la requête.

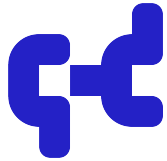
Faire correspondre la meilleure liste de documents à une requête







Recherche d'information (RI)



Des modèles basés sur :

- **la théorie des ensembles**. Les documents sont représentés comme des ensembles de mots et les similarités entre entre documents et requêtes sont obtenues par des opérations ensemblistes.
- **la représentation algébrique des documents**. Requêtes et documents sont modélisés vectoriellement et les similarités sont des scalaires.
- **les probabilités**. Les similarités sont mesurées comme la probabilité qu'un document soit pertinent au regard d'une requête donnée.

⚠ Biais de "source" : les modèles neuronaux privilégient les documents générés par LLM au détriment de ceux sémantiquement équivalents écrits par des humains [1].

Systèmes génératifs

- Basés sur des **algorithmes d'apprentissage**
  → renforcement à partir de rétroaction humaine

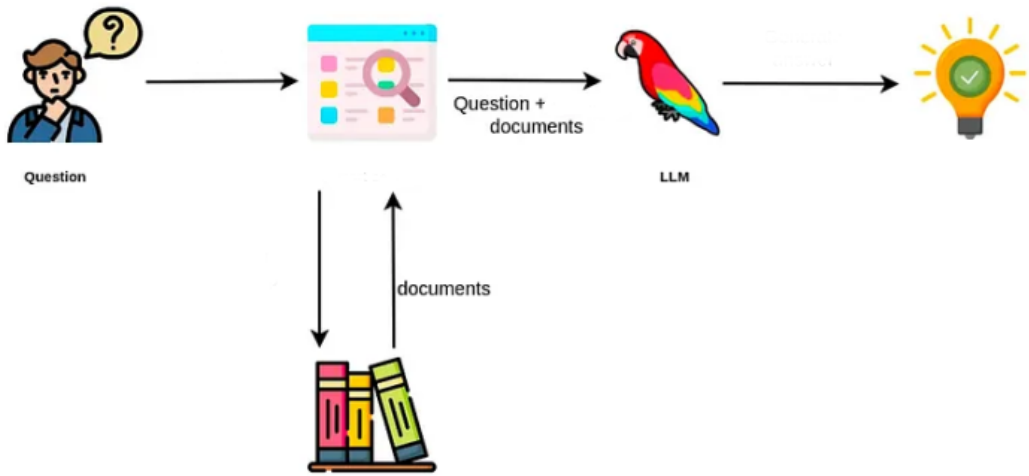
- **Génèrent** du texte, des images, en réponse à une invite de commande

- **Modèles probabilistes** construits à partir de grandes quantités de documents existants

  Fabriquer un contenu élégant en réponse à contenu fourni.

⇒ **pas** de raisonnement, **pas** de compréhension, **pas** d'analyse
pas non plus d'"hallucinations", mais des contenus non valides.

Génération augmentée de récupération (RAG)



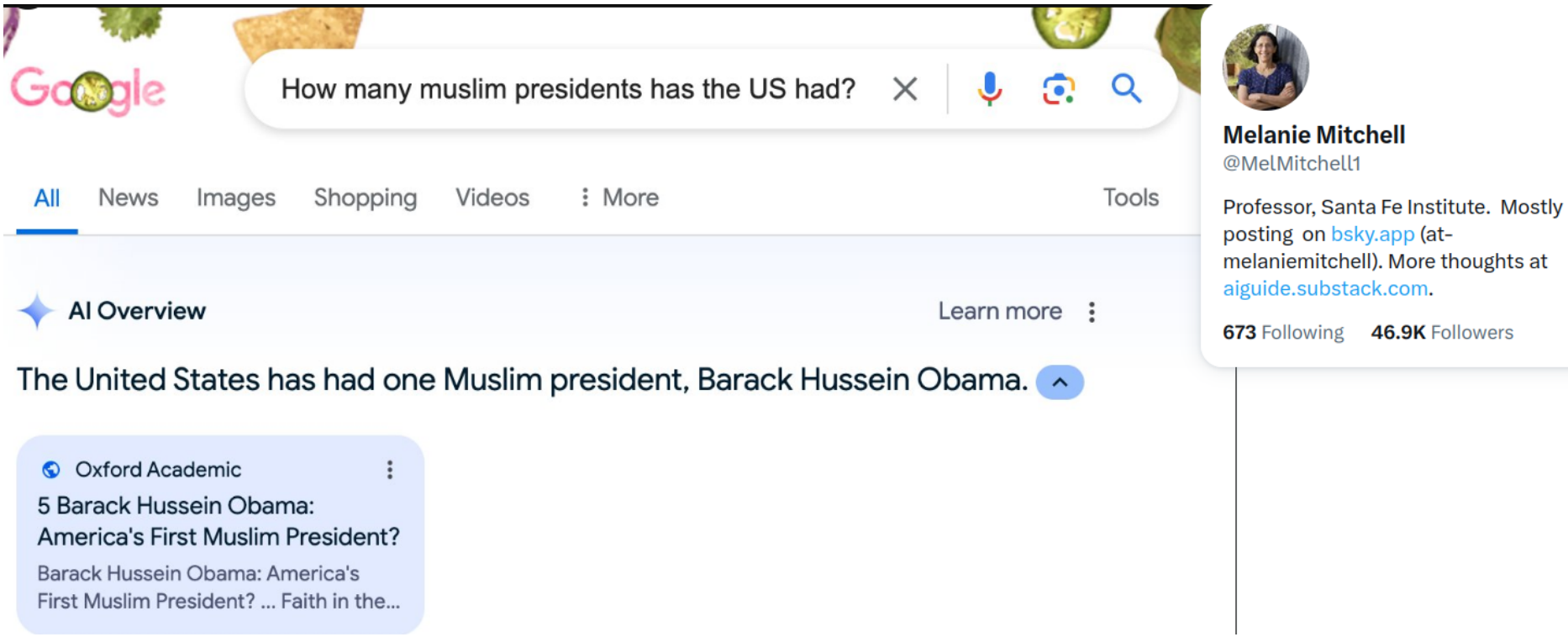
source : neo4j.com

- RI pour intégrer de **nouvelles informations** dans les modèles génératifs, sans ré-entraînement.
- Modèle de langue contraint à se référer à un **ensemble de documents choisi**, en complément de son entraînement.
- Adaptation par domaine et limitation de l'**obsolescence** des LLM

Toujours aucune garantie de validité des contenus générés.

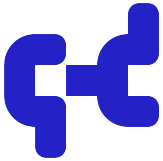
(créer du faux avec du vrai [2])

Génération augmentée de récupération (RAG)



source : Melanie Mitchell

Traduction automatique de contenus scientifiques



- Que traduire ? *résumé, métadonnées, texte complet, mots clés...*
- Quel objectif ? *meilleure indexation, accessibilité, découvrabilité...*
- Qualité de la traduction ? *utilisable mais imprécise* [3]
- Impact sur les contenus disponibles ? *création de contenus imparfaits*
- Impact sur les professionnel.les ? *perte de compétences, d'intérêt*

Traduction professionnelle au coeur de la **réflexion sur les usages**

Usages et enjeux

- Revue de la **littérature** ⛔
  → Outils opaques, biais, limitations, perte de compétences
- Extraction d'**information** ⛔
  → Erreurs factuelles
- **Résumé** automatique ❓
  → Utile pour évaluer le thème d'un document ou pour résumer un document que l'on connaît et dont on pourra valider le résumé
- Aide à la **réflexion** ❓
  → Pas de contrôle sur l'orientation des questionnements
- Aide à la **rédaction** ❓
  → Utile pour préparer un brouillon de document

Et aussi : Enjeux environnementaux - Contenus non valides
© Propriété intellectuelle

Bibliographie

Sunhao Dai, *et al.* (2024), *Cocktail: A Comprehensive Information Retrieval Benchmark with LLM-Generated Documents Integration.*

Rhiannon Williams (2024), *Why Google's AI Overviews gets things wrong*, MIT Technology Review.

Paul Lerner et François Yvon (2024), *Vers la traduction automatique des néologismes scientifiques*, JEP-TALN-RECITAL 2024.