Amélioration du principe de RAG (Retrieval-Augmented Generation)

Contexte et limites du RAG classique

- **Principe RAG** : combinaison d'un module de récupération d'information (retriever) et d'un générateur (LLM).
- **Objectif** : fournir des réponses plus factuelles en consultant une base externe de documents.
- **Limites du RAG classique** :
  - Récupération parfois hors-sujet ou redondante.
  - Latence élevée (due à la recherche et au traitement).
  - Difficulté à s'adapter à des domaines spécifiques.

Améliorations du module de récupération

- **Hybrid Search et reranking** : combinaison de recherche dense (embedding) et sparse (BM25).
- **Pré-entrainement du retriever** : amélioration des embeddings pour mieux refléter l'intention de la requête.
- **Reformulation automatique de la requête** : meilleure contextualisation grâce au feedback des résultats.
- **Chunking sémantique** : découpage optimisé des documents pour une meilleure récupération.

Optimisation du générateur

- **Prompt engineering** : prompts mieux structurés, incluant des balises ou instructions explicites.
- **Boucles itératives** : alternance entre récupération et génération pour améliorer la précision.
- **Self-RAG** : le modèle évalue lui-même la qualité des documents récupérés et ajuste ses réponses.
- **Mémoire contextuelle** : intégration de mémoire ou contexte long pour améliorer la continuité.

Innovations récentes et cas d'usage

- **GraphRAG** : exploitation de graphes de connaissance pour enrichir les réponses.
- **Parametric RAG** : intégration directe des connaissances dans les paramètres du modèle.
- **Multimodalité** : extension à des données non textuelles (images, vidéos, radiologie...).
- **Domaines d'application** :
  - Santé (diagnostic assisté)
  - Droit (réduction des hallucinations)
  - Industrie (assistance sécurité)