La méthode HyDE (Hypothetical Document Embeddings) et les techniques avancées de RAG (Retrieval-Augmented Generation) font partie des approches modernes combinant la recherche d'information avec la génération de texte par des modèles de langage.
HyDE est une méthode qui améliore la recherche dans les systèmes RAG. Plutôt que d'utiliser directement la requête de l'utilisateur pour rechercher les documents pertinents, HyDE commence par générer un document hypothétique qui répondrait idéalement à la question posée. Ce document n'est pas présenté à l'utilisateur, mais il est converti en vecteur (embedding), et ce vecteur sert ensuite à interroger une base de données vectorielle contenant les documents réels. L'hypothèse est qu'un document généré par un modèle de langage bien entraîné est plus représentatif de l'information recherchée que la requête brute, ce qui améliore la pertinence des documents retrouvés.
RAG (Retrieval-Augmented Generation) est un cadre dans lequel un modèle de langage interagit avec une base de connaissances externe (comme une base de documents ou une base vectorielle) pour produire des réponses plus précises et informées. Le fonctionnement classique de RAG se déroule en deux étapes : (1) une étape de récupération (retrieval) où l'on identifie les documents les plus pertinents pour une requête donnée à l'aide d'un moteur de recherche vectoriel, souvent basé sur des embeddings sémantiques ; (2) une étape de génération où un modèle (comme GPT ou BERT modifié) lit ces documents et génère une réponse en s'appuyant sur leur contenu.

Les techniques avancées de RAG incluent :

- **Multi-hop RAG : qui récupère plusieurs documents en plusieurs étapes pour répondre à des questions complexes nécessitant des chaînes de raisonnement.**
- **Fusion-in-Decoder (FiD) : qui permet au modèle de traiter séparément chaque document récupéré avant de fusionner leur contenu pendant la génération de réponse.**
- **Query rewriting : où la question originale est reformulée automatiquement pour mieux correspondre aux documents de la base.**
- **Re-ranking ou reranking : après la récupération initiale de documents, un second modèle les réévalue pour en sélectionner les plus pertinents.**
- **Memory-Augmented RAG : où le système garde une mémoire à long terme des interactions précédentes pour fournir des réponses plus contextuelles.**

En combinant HyDE avec RAG, on obtient un système plus intelligent et pertinent, capable d'anticiper les réponses possibles et d'enrichir la recherche documentaire, ce qui améliore significativement la qualité de la génération finale.