Rapport d’expérimentation - Optimisation du chunking pour le RAG

Expérimentation :

- Différentes valeurs de chunk_size
- Différentes valeurs de chunk_overlap
- Utilisation du modèle d’embeddings all-MiniLM-L6-v2 afin d’identifier la configuration offrant le meilleur compromis entre performance et coût de traitement

Résultats obtenus :

256 / 0 → MRR = 0.145, nb_chunks = 308
256 / 64 → MRR = 0.214, nb_chunks = 352
384 / 96 → MRR = 0.192, nb_chunks = 246
512 / 0 → MRR = 0.233, nb_chunks = 181
512 / 128 → MRR = 0.282 (meilleur), nb_chunks = 194
512 / 100 → MRR = 0.222, nb_chunks = 188
768 / 0 → MRR = 0.241, nb_chunks = 145
768 / 192 → MRR = 0.271, nb_chunks = 151

Analyse :

Influence de la taille des chunks Les expérimentations montrent une tendance claire : les chunks de petite taille donnent systématiquement un MRR plus faible. Les valeurs 256/0 (MRR = 0.145) et 256/64 (MRR = 0.214) illustrent bien cette fragmentation excessive de l’information. Les chunks plus longs, en particulier entre 512 et 768 tokens, obtiennent de meilleurs résultats. Par exemple, 768/0 atteint 0.241 et 768/192 atteint 0.271. Ces tailles conservent un contexte plus cohérent et permettent au retrieval de mieux capturer les passages pertinents.

Influence du chunk_overlap L’overlap joue un rôle essentiel dans la performance du retrieval. Une augmentation du chevauchement entraîne presque toujours une amélioration du MRR. L’exemple le plus parlant est la transition de 512/0 (MRR = 0.233) à 512/128 (MRR = 0.282), qui constitue le meilleur score obtenu. Le chevauchement permet d'éviter la perte d’information située aux frontières des chunks, réduisant ainsi la fragmentation du contexte sémantique.

Compromis entre coût et précision Un overlap élevé ou une réduction de la taille des chunks augmente mécaniquement le nombre total de segments à indexer. Cela peut alourdir le coût de calcul et la mémoire nécessaire. Par exemple, 256/64 génère 352 chunks, contre seulement 145 pour 768/0. Il faut donc trouver un équilibre entre cohérence contextuelle et efficacité computationnelle.

Interprétation :

Les résultats permettent de dégager plusieurs lignes directrices importantes : Les chunks courts dégradent fortement la qualité du retrieval. La zone optimale se situe autour de 512 à

768 tokens. L’overlap améliore clairement la cohérence et la performance. 4. La meilleure configuration identifiée est : chunk_size = 512 et chunk_overlap = 128, avec un MRR de 0.282. Cette combinaison offre un excellent équilibre entre précision, cohérence et coût.

Ces expérimentations démontrent l’importance du paramétrage du chunking dans un système RAG. Un bon découpage améliore nettement la stabilité et la précision du retrieval. Le couple (512, 128) constitue la configuration la plus efficace parmi celles testées. Ces observations confirment que le chunking doit être traité comme un véritable hyperparamètre à optimiser dans tout pipeline RAG sérieux.