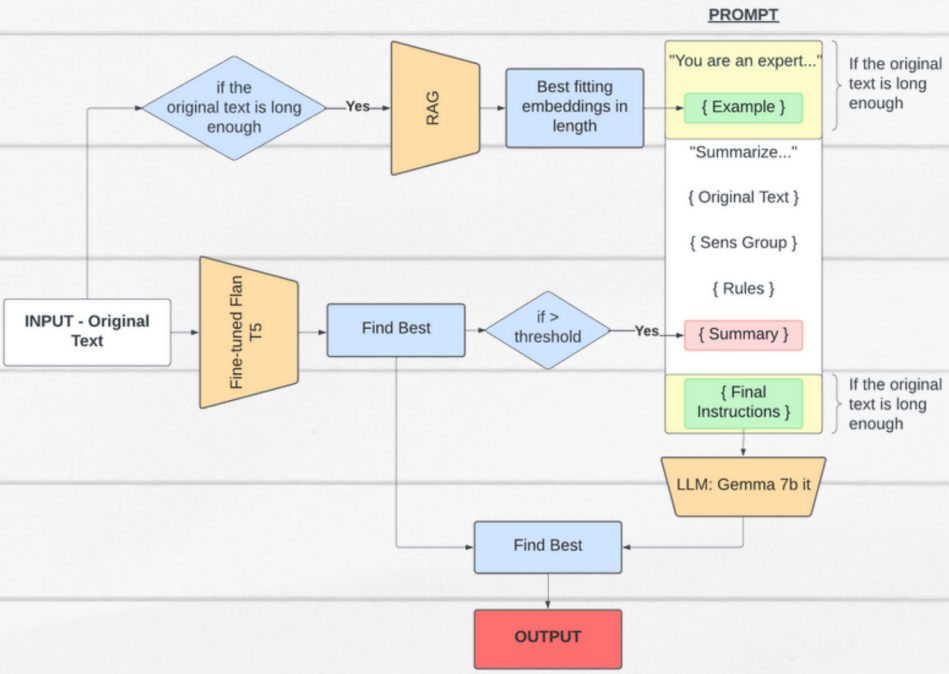
HACKATSUKI

CRANCÉE ELIOTT, HAN CLÉA, LABEYRIE YANIS, TRIQUET LÉA, ZABBAN ADRIEN



2 modèles en 1

Transformer: Tout les textes sont résumés par ce modèle léger. Dans deux tiers des cas, ce résultat suffit.
Ce modèle produit 10 résumés différents et on sélectionne le meilleur.
LLM avec RAG: Dans un tiers des cas, un LLM est utilisé pour tenter d'améliorer les résultats du transformer.
On compare ensuite son résultat avec celui du transformer, et on prend le meilleur.

Le transformer: Flan T5 Fine-tuned

Fine tunning sur 10 epochs, génération de 10 résumés, sélection du meilleur. Si le score du résumé est supérieur à 0.82, on renvoie directement ce résultat.

Le processus de sélection: Find Best

Une moyenne entre la métrique rouge, la similarité avec le texte original et la prédiction d'un xgboost entraîné pour donner la métrique d'évaluation.

Le RAG: ChromaDB

Si le texte n'est pas dans la catégorie "très court": Comparaison avec les exemple ingérés, sélection des 3 plus proches embedings, sélection du texte le plus proche en taille.

Le LLM: Gemma instruction tuned

L'instruction dépend de la taille du texte. 5 catégories de textes: Très court (ne passe pas par le LLM), court, standard, long et très long.

Le LLM: Le prompt

- L'exemple sélectionné par le RAG
- La consigne
- Les groupes de sens identifiés auxquels faire attention
- Les règles de résumés déduites de la base d'entrainement
- Le résumé produit par le transformer
- Les dernières instructions