ÉVALUATION AUTOMATIQUE DU RETOUR À LA SOURCE DANS UN CONTEXTE HISTORIQUE

Les débats parlementaires de la Troisième République française

Aurélien Pellet[1 2], Julien Perez[2], Marie Puren[2 3]

4 juillet 2025 - Journée HNIA à PFIA

[1]Epitech [2]LRE, Epita [3]Centre Jean-Mabillon, Ecole nationale des chartes

LES DÉBATS PARLEMENTAIRES DURANT LA IIIE RÉPUBLIQUE

Journal officiel du 5 décembre 1897 — CHAMBRE — SÉANCE DU 4 DÉCEMBRE — Session extraordinaire de 1897 — 2723

CHAMBRE DES DÉPUTÉS

6e législ. — Session extraordinaire de 1897.

COMPTE RENDU IN EXTENSO. — 31e SÉANCE

1re séance du samedi 4 décembre.

SOMMAIRE

Procès-verbal : M. Bourrillon.
Excuses et demande de congé.
Discussion de l'interpellation de M. Chiché au sujet de l'attitude de M. le ministre des travaux publics en présence de combinaisons ayant pour but d'accaparer le trafic du port de Bordeaux au profit de la société des appontements de Paulilac et de la compagnie des chemins de fer du Midi : MM. Chiché, le comte du Périer de Larsan, le ministre des travaux publics, Grujet.
Discussion de l'interpellation de M. Dejeante sur les causes du nouvel accident du puits Sainte-Eugénie, à Montceau-les-Mines : MM. Dejeante, le ministre des travaux publics, Basly. — Ordre du jour motivé de M. Dejeante et autres : M. Dejeante. — Adoption de l'ordre du jour pur et simple.
Communication, par M. le président : 1° d'une demande d'interpellation adressée par M. Marcel Sembat au Gouvernement sur son attitude dans l'affaire Dreyfus ; 2° d'une demande d'interpellation adressée par M. le comte Albert de Mun au ministre de la guerre au sujet de l'affaire Dreyfus. — MM. Marcel Sembat, le président du conseil, ministre de l'agriculture.
Question adressée par M. André Castelin à M. le président du conseil, ministre de l'agriculture, et réponse de M. le président du conseil, ministre de l'agriculture. — Demande, par M. Marcel Sembat, de transformation de la question en interpellation. Adoption. — Sur la discussion immédiate : MM. Marcel Sembat, Castelin, le président du conseil, le comte Albert de Mun. — Suspension de la séance. — Reprise de la séance. — Discussion des interpellations de M. Marcel Sembat et de M. le comte Albert de Mun : MM. le ministre de la guerre, Marcel Sembat, le président du conseil, Millerand. — Ordres du jour motivés : 1° de M. Jumel ; 2° de M. le comte Albert de Mun ; 3° de MM. Lavertujon et Thoulouse ; 4° de M. Bourgeois-du-Jura ; 5° de MM. Pierre Richard et Marcel Habert ; 6° de M. Marcel Sembat ; 7° de MM. Millerand, Viviani, Rouanet et Humbert. — Demande de priorité pour l'ordre du jour de M. Jumel. Rejet, au scrutin. — Demande de priorité pour l'ordre du jour de M. le comte Albert de Mun. — Scrutin. — Pointage. — Annonce du résultat du dépouillement du scrutin vérifié. Rejet. — Additions à l'ordre du jour de MM. Lavertujon et Thoulouse : 1° de MM. le comte du Périer de Larsan et autres ; 2° de MM. Marcel Habert et Pierre Richard. — Nouvelle rédaction de l'ordre du jour de MM. Millerand, Viviani, Rouanet et Humbert. — Demande de priorité pour l'ordre du jour de MM. Lavertujon et Thoulouse. Adoption, au scrutin. — Retrait, par M. du Périer de Larsan, de son addition à l'ordre du jour de MM. Lavertujon et Thoulouse. — Adoption, au scrutin, de la 1re partie de l'ordre du jour de MM. Lavertujon et Thoulouse. — Reprise, par M. Alphonse Humbert, de l'addition de M. du Périer de Larsan. — Adoption, au scrutin, de la 2e partie de l'ordre du jour de MM. Lavertujon et Thoulouse. — Sur l'addition de M. du Périer de Larsan reprise par M. Alphonse Humbert : MM. le président du conseil, Alphonse Humbert. Adoption, au scrutin. — Scrutin sur l'addition de MM. Marcel Habert et Pierre Richard. Nul, faute du quorum.
Levée de la séance.

PRÉSIDENCE DE M. HENRI BRISSON

La séance est ouverte à deux heures.

M. Henry Cochin, *l'un des secrétaires,* donne lecture du procès-verbal de la séance d'hier.

CHAMBRE. — IN EXTENSO.

M. le président. La parole est à M. Bourrillon, sur le procès-verbal.

M. Bourrillon. Je suis porté par erreur au *Journal officiel* comme ayant voté « contre » l'ensemble de la proposition de loi relative à la suppression des taxes d'octroi sur les boissons hygiéniques. Je déclare avoir voté « pour ».

Ayant dans tous les scrutins qui ont eu lieu émis un vote conforme aux vues de la commission, il n'est en effet pas admissible que j'aie voté « contre » l'ensemble.

M. le président. Il n'y a pas d'autre observation sur le procès-verbal ?...

Le procès-verbal est adopté.

EXCUSES ET DEMANDE DE CONGÉ

M. le président. MM. Ermant et Couyba s'excusent de ne pouvoir assister aux séances de ce jour.

M. Theulier s'excuse de ne pouvoir assister aux séances de ce jour et demande un congé.

La demande sera renvoyée à la commission des congés.

DISCUSSION D'UNE INTERPELLATION DE M. CHICHÉ, CONCERNANT LA SOCIÉTÉ DES APPONTEMENTS DE PAULILAC ET LA COMPAGNIE DU MIDI

M. le président. L'ordre du jour appelle la discussion de l'interpellation de M. Chiché au sujet de l'attitude de M. le ministre des travaux publics en présence de combinaisons ayant pour but d'accaparer le trafic du port de Bordeaux au profit de la société des appontements de Paulilac et de la compagnie des chemins de fer du Midi.

La parole est à M. Chiché, pour développer son interpellation.

M. Albert Chiché. Messieurs, l'interpellation que je vieus développer devant vous n'intéresse pas seulement le port de Bordeaux. Sa portée est plus haute, car les faits sur lesquels je me propose d'appeler votre attention vous montreront à quel point la rapacité des grandes compagnies de chemins de fer et des grandes sociétés capitalistes peut devenir dangereuse pour notre commerce, pour notre industrie, pour notre marine marchande, en un mot, pour toutes les branches de notre activité nationale.

Deux sociétés extrêmement riches et puissantes cherchent à accaparer à leur profit exclusif le trafic du port de Bordeaux, et elles y sont déjà parvenues en partie ; l'une, la compagnie du chemin de fer du Midi, a sa tête de ligne au sud de Bordeaux ; l'autre, la société des appontements de Paulilac, est installée à quelques kilomètres de l'embouchure de la Gironde, et, par la ligne du Médoc, pénètre jusqu'au cœur de Bordeaux. Elles ne sont donc séparées que par les 6 kilomètres de quais qui s'étendent depuis la gare Saint-Jean jusqu'aux docks. Leur ambition est de se réunir directement l'une à l'autre, afin que les marchandises puissent transiter directement de Paulilac à Cette et de Cette à Paulilac sans emprunter le cours du fleuve. Ce serait, vous le comprenez, la ruine du port de Bordeaux au profit des deux grandes sociétés, qui ont, l'une et l'autre, l'heureuse fortune d'être dirigées par un membre de la famille Pereire. (*Très bien ! très bien ! à l'extrême gauche et sur plusieurs bancs à gauche.*)

Ces projets n'ont été, malheureusement, jusqu'ici que trop favorisés par plusieurs des ministres des travaux publics qui se sont succédé au pouvoir dans ces dernières années. Après avoir accordé à la compagnie des chemins de fer du Midi la concession d'une ligne ferrée longeant les quais de Bordeaux depuis la gare du Midi jusqu'à la gare du Médoc, on a donné à M. Eugène Pereire l'autorisation de construire à Paulilac de vastes appontements ; les navires qui jadis remontaient le fleuve, s'y arrêtent ; on y opère leur débarquement, puis on transporte les marchandises, par la ligne du Médoc, au nord de Bordeaux. Cette ligne n'a été concédée que comme ouillage du port, pour faciliter les relations entre la navigation et les gares, et non pour unir les deux gares entre elles.

Mais les Pereire se mettent au-dessus des lois, des décrets et des conventions. Ils les violent ; ils emploient cette ligne des quais à tous les usages, et ils parviennent ainsi à ce double résultat : de faire à la navigation une concurrence désastreuse et de rendre les quais du port de Bordeaux impraticables en y faisant circuler d'interminables convois qui interceptent les communications entre le fleuve et les maisons, chais, magasins ou bureaux qui lui font face.

Voilà le premier empiètement sur lequel je crois devoir appeler l'attention de M. le ministre des travaux publics, car il cause un préjudice au port de Bordeaux. Mais la compagnie du Midi ne veut pas s'en tenir là ; les grandes compagnies, vous le savez, sont insatiables. La compagnie du Midi a jeté son dévolu sur une nouvelle proie : elle veut absorber le transport des marchandises lourdes qui ont encore intérêt à emprunter le cours du fleuve.

Actuellement, 400,000 tonnes de houilla arrivent d'Angleterre, par le fleuve, sur les quais de Bordeaux, et 40,000 tonnes de poteaux de mines, tirés de nos forêts de pins, arrivent de Langon par le haut de la rivière et servent de fret de retour. Comment faire pour accaparer ce transit ? Il faudrait tout simplement que le prix des transports, qui était autrefois de 2 fr. 25 par tonne et qui a été réduit à 1 fr. 25, soit abaissé à 50 centimes par tonne. La compagnie du Midi, qui ne doute de rien, a soumis ce nouveau tarif réduit à l'homologation de M. le ministre des travaux publics.

Cette demande d'homologation a produit dans la ville de Bordeaux une émotion considérable. Les gabariers, les batelliers, les cultivateurs, tous les ouvriers qui vivent en travaillant sur le fleuve ou sur les quais, ont immédiatement signé des protestations. Le conseil général de la Gironde, le conseil municipal et la chambre de commerce de Bordeaux s'y sont associés, et ont pris des délibérations hostiles à la réduction de tarif sollicitée.

J'ai ces protestations dans mon dossier, et il n'est peut-être pas inutile que je vous en donne lecture ; car, en pareille matière, il est toujours bon de donner la parole aux intéressés. Ces documents ont été envoyés à M. le ministre des travaux publics. Mais vous savez que les ministres n'ont pas toujours le loisir de lire les pétitions qu'on leur adresse. Par conséquent, M. le ministre des travaux publics me permettra de lui en imposer la lecture :

Voici d'abord la pétition des bateliers et gabariers :

« Bordeaux, le 15 février 1897.

« Monsieur le ministre,

« La compagnie des chemins de fer du Midi faisait payer autrefois 2 fr. 25 pour la traction des marchandises de Bordeaux-Saint-Jean aux docks ; elle a réduit ce prix à 1 fr. 75, et il est actuellement question de l'abaisser à 50 centimes par 1,000 kilogr. Cette réduction de tarif, qui va supprimer la batellerie, plongerait dans la plus pro-

55

Source gallica.bnf.fr / Bibliothèque et Archives de l'Assemblée nationale

Séance parlementaire du 4 décembre 1897

Compréhension des motivations des députés, des oppositions aux projets de loi, et des sujets traités

- Analyse thématique du corpus et exploitation des données sérielles
- Interrogation des documents pour répondre à des questions spécifiques
- Annotation automatique pour faciliter l'exploration du corpus
- Validation des affirmations et données extraites en revenant aux sources

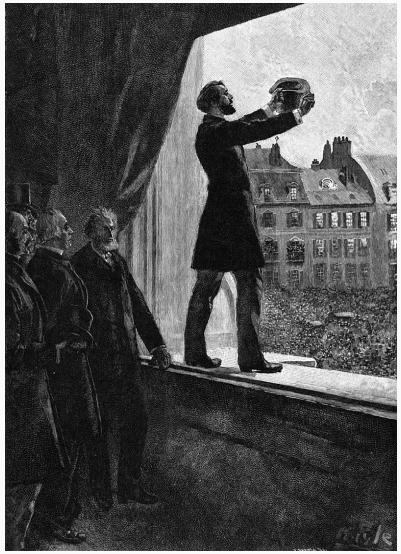
LES LLMS PEUVENT-ILS NOUS AIDER ?

Contexte

- Troisième République : archives parlementaires riches mais volumineuses
- Test des LLM pour l'exploration de corpus historiques

Questions de recherche

- Comment les LLM aident-ils l'exploration d'archives ?
- Comment évaluer la pertinence et la frugalité ?
- Mesurer le retour à la source



Gambetta proclamant la République (Frontispice de Mary Platt Parmele, A Short History of France, 1907)

Séance du 4 Décembre 1897 :
« Il n'y a pas d'affaire Dreyfus [...] il n'y a pas en ce moment et il ne peut pas y avoir d'affaire Dreyfus. »

UN EXEMPLE D'OPINION TRANCHÉE

« l'anarchisme. [...] a ses deux versants : sur l'un, la doctrine fleurit dans toute sa pureté mystique ; sur l'autre, l'action apparaît avec toute la brutalité révolutionnaire. [...] C'est aussi sur les théoriciens que vous avez déchaîné votre police, et votre erreur est d'autant plus condamnable, monsieur le ministre, que le régime sous lequel vous la pratiquez est issu lui-même de la double gestation de la violence et de la pensée.

Il y a du sang sur toutes les doctrines, même sur celles que vous invoquez contre le sang ! »

— M. Clovis Hugues

Extrait d'un discours sur les répressions contre les Anarchistes à la Chambre des députés (27 janvier 1894).

UN EXEMPLE D'OPINION TRANCHÉE

« Aujourd'hui je demande à un paysan :

— Qu'est-ce que la patrie ? Est-ce ton champ ?

— Il répondra : Oui.

Cependant, il sait bien qu'il fait partie d'une nation qui s'appelle la France ; mais comment la connaît-il ?

Un jour, on publie à son de trompe l'appel des conscrits. Son fils va prendre un numéro dans une boîte. Soldat pendant cinq ans, c'est la patrie qui le demande.

De temps en temps, il voit apparaître un gendarme qui lui dresse un procès-verbal : c'est une nouvelle incarnation de la patrie.

Les charges, il les conçoit très bien. Quant aux avantages, il les aperçoit moins distinctement.

L'idée de la patrie est fort obscurcie chez lui, et quand, au moment d'une guerre, on vient lui parler en phrases pompeuses de ses obligations, il répond :

« Mes obligations, je les connais ; mais quelles sont les obligations de la patrie envers moi ? »

— **Yves Guyot**, Les Droits de l'Homme, 27 mars 1876.
Cité par **Clovis Hugues**, discours à la Chambre des députés, 27 janvier 1894.

Un corpus « massif » : 15 législatures entre 1881 et 1940 à la Chambre des députés, avec 10 000 à 12 000 images par législature

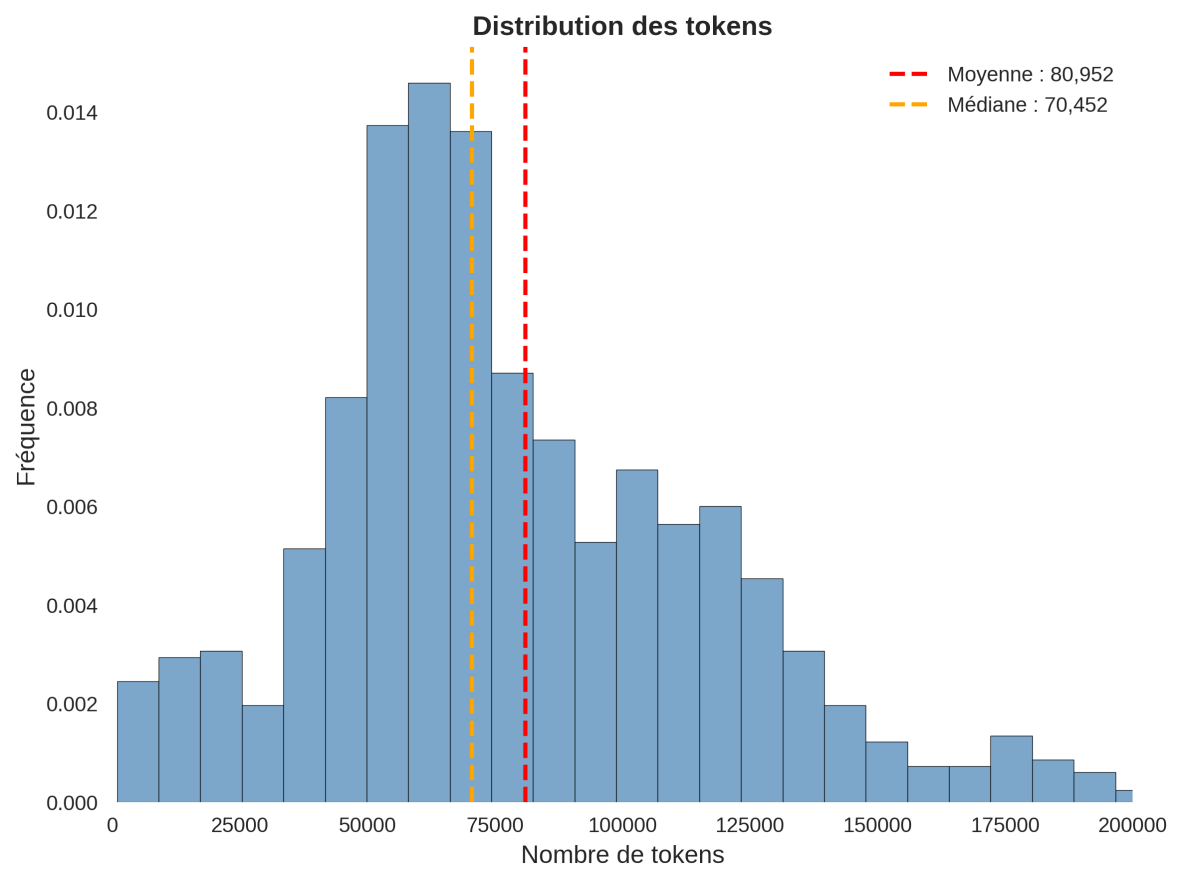
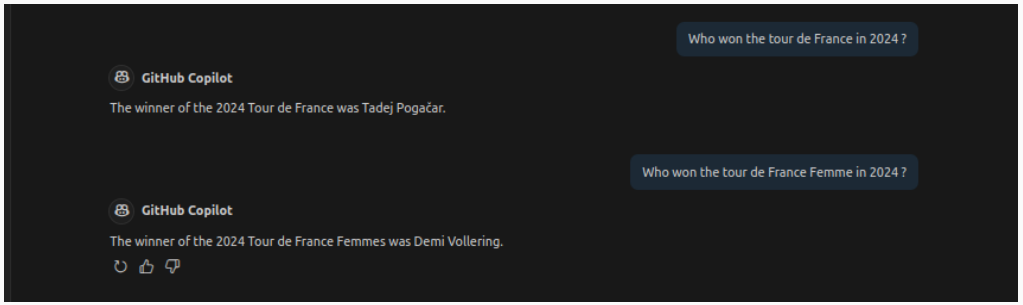


Figure – Distribution du nombre de tokens

LE RAG POUR LA RECHERCHE HISTORIQUE

LLM ET AGENTIFICATION



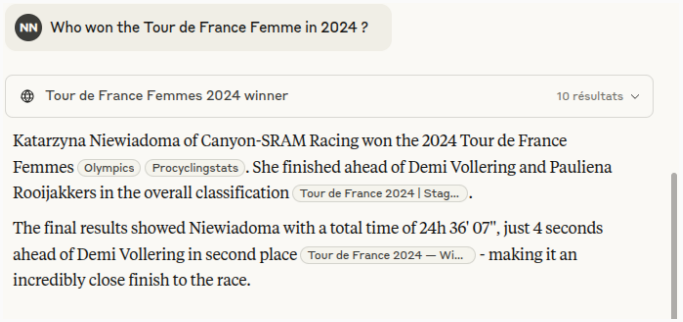
Réponses erronées de GitHub Copilot

Limites

- Modèles fermés
- Limite dans le temps
- Encodage du savoir

Pour pallier ces limites : agentification des LLM

- Recherche sur Internet
- Interrogation d'API
- RAG (Retrieval-Augmented Generation)



Résultat correct via recherche web

LE PROBLÈME DE LA SPÉCIFICITÉ ET DE LA TAILLE

- Les architectures Transformer sont limitées par la taille maximale de la séquence qu'elles peuvent traiter.
- Le risque d'hallucination apparaît bien avant d'atteindre cette limite.

Modèle	Contexte max.	Date de sortie
BERT	512 tokens	Octobre 2018
GPT-3.5	4 096 tokens	Mars 2023
LLaMA 2	4 096 tokens	Juillet 2023
LLaMA 3.1 (70B)	128 000 tokens	Juillet 2024
GPT-4o	128 000 tokens	Mai 2024
Qwen2.5	128 000 tokens	Septembre 2024
Gemma-3 :27b	128 000 tokens	Mars 2025

Capacités contextuelles des modèles de langue

LE PROBLÈME DE LA SPÉCIFICITÉ ET DE LA TAILLE

Query	Model	Response
Quelles sont les deux propositions de résolutions déposées par M. Louis Marin ?	GPT-3.5 Turbo	Code d'erreur : 400 - {'error' : {'message' : "La longueur maximale du contexte de ce modèle est de 16385 tokens. Cependant, vos messages ont généré 49791 tokens.
	LLaMA 3.1 70B	Les deux propositions de résolution déposées par M. Louis-Marin concernent la pétition n° 1615 et portent sur le renvoi de cette affaire à la commission des réclamations, avec un vote différent pour chacune : l'une est favorable à un renvoi simple (proposition n° 1) et l'autre préconise un renvoi avec une déclaration formelle (proposition n° 2).
Qui est le président de la séance du jour ?	LLaMA 3.1 70B (full)	M. André
	LLaMA 3.1 70B (5k chars)	M. Jules Méline

Limites des LLM sur les séances des débats parlementaires

Points clés

- **Contexte limité :** Les modèles comme GPT-3.5 ne disposent pas d'une taille de contexte suffisante pour traiter un débat complet.
- **Risque d'hallucination :** Même avec un modèle très récent et une taille de contexte plus grande, le risque d'hallucination reste élevé.

RETRIEVAL-AUGMENTED GENERATION (RAG)

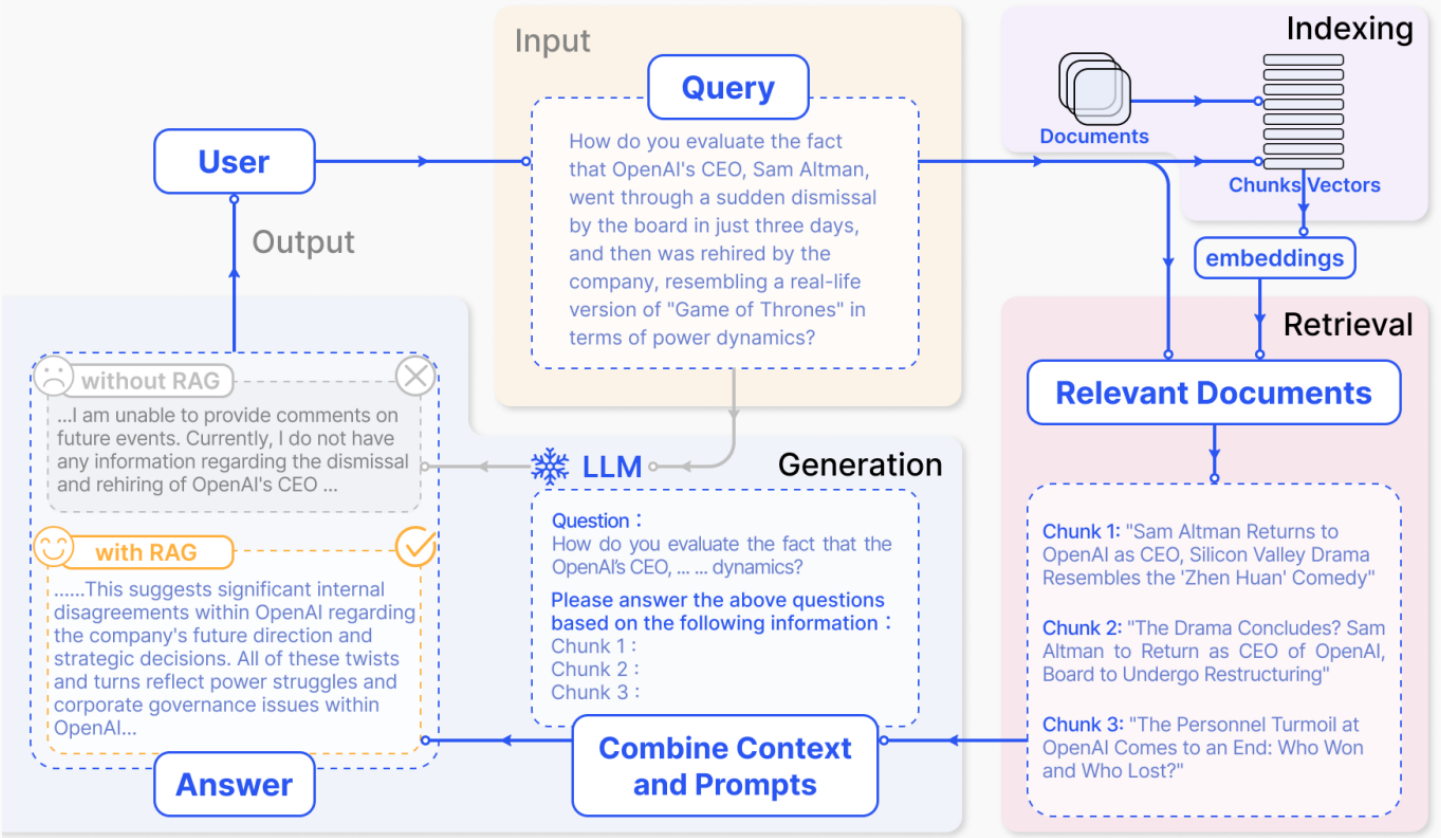


Schéma d'un RAG simple, Yufan Gao, Retrieval-Augmented Generation for Large Language Models : A Survey, 2023

EXEMPLE DE RAG NAÏF

1. Question

Quel est le montant total des dégâts subis par les routes locales du département des Côtes-du-Nord ?

2. Phase de Récupération

Modèle d'embedding : `mGTE`

Documents récupérés :

[Chunk n°1] Les dégâts sont beaucoup plus considérables; d'après un document officiel — le rapport du préfet à la commission départementale — la dépense pour les routes locales à la charge du département s'élève à 275 060 francs.

[...]

[Chunk n°k] ...

3. Template de Prompt

SYSTÈME : Je vais vous fournir des extraits de débats parlementaires de la Troisième République. Analysez-les et utilisez-les pour répondre à la question donnée.

UTILISATEUR :

- Extrait 1 : [Premier chunk]
- ...
- Extrait K : [K-ème chunk]

Question : [Question]

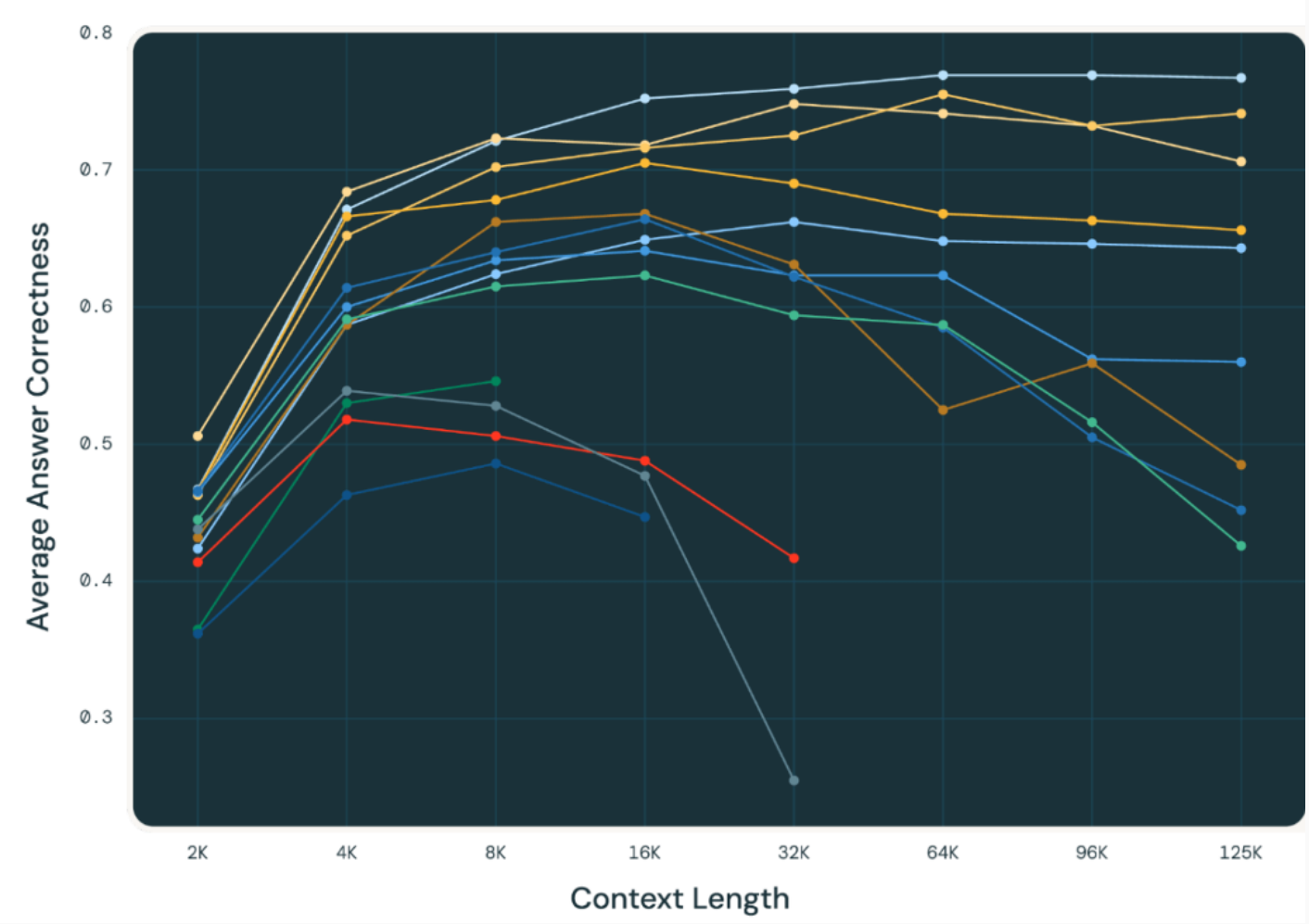
4. Réponse Générée

Le montant total des dégâts subis par les routes locales du département des Côtes-du-Nord est de **275 060 francs**.

EXEMPLE DE QUESTIONS

Source	Query	Chunk retrieved (top 1 chunk)	Answer
1881-01-22	Quel est le montant total des dégâts subis par les chemins vicinaux dans le département des Côtes-du-Nord ?	Les dégâts sont bien plus considérables ; ils s'élèvent, — et j'ai un document officiel, c'est le rapport du préfet à la commission départementale, — rien que pour les chemins vicinaux qui sont à la charge du département, ils s'élèvent à la somme de 275,060 fr.	Le montant total des dégâts subis par les chemins vicinaux dans le département des Côtes-du-Nord s'élève à 275 000 francs.
1881-01-22	"Quel est le montant de l'emprunt autorisé pour le département de l'Allier et à quoi sera-t-il utilisé ?"	t'rPROJET«Art. 1er. — Le département de l'Allier est autorisé, conformément la demande que le conseil général en a faite, à emprunterala caisse des lycées, collèges et écoles, aux conditions de cet établissement, une somme de 310,000 fr. applicable au payement de subvenventions destinées aux travaux des lycées de Moulins et de Montluçon.	Le montant de l'emprunt autorisé pour le département de l'Allier est de 310,000 fr. Il sera utilisé pour le payement de subventions destinées aux travaux des lycées de Moulins et de Montluçon.
1882-03-23	"Quel est le montant de l'emprunt autorisé pour le département des Bouches-du-Rhône ?"	1882-03-23	Le montant de l'emprunt autorisé pour le département des Bouches-du-Rhône n'est pas mentionné dans les extraits de débat parlementaire fournis.
	Qui est le président de la séance du jour ?	PRÉSIDENCE DE M. JULES MÉLINE La séance est ouverte à deux heures un quart. M. Charles Chevalier (Manche) , l'un des secrétaires, donne lecture du procès-verbal de la séance d'hier.	M. Jules Méline

LLM, CONTEXTE ET HALLUCINATION



Long Context RAG Performance of LLMs, databricks, https://www.databricks.com/blog/long-context-rag-performance-llms

Inférence et frugalité

- Limites des LLM généralistes : inefficaces sur des corpus historiques longs et spécialisés.
- Objectif du RAG : maximiser la qualité des réponses tout en limitant la taille des entrées.

Objectifs de l'étude

- **Chunking** : le découpage du texte influence la qualité du retrieval et des réponses.
- **Évaluation** : comparaison de plusieurs stratégies de découpage dans un pipeline RAG.
- **Reranking** : étude de l'impact du reranker sur le retour à la source.

Évaluation des capacités de retrieving

GÉNÉRATION AUTOMATIQUE DE QUESTIONS

Pour explorer les capacités de retrieving sur de longs documents et constituer un corpus conséquent de questions, nous utilisons un LLM pour générer des questions.

- À partir de débats prédécoupés (pour réduire leur taille), un LLM est utilisé pour **générer une liste de questions** et **fournir le passage du débat utilisé** pour chaque question.
- Llama 3.1 :70B est utilisé pour la génération de questions.
- Vérification systématique que la source renvoyée par le LLM est bien présente dans le texte d'origine.
- Le prompt inclut une phase de réflexion et de vérification pour améliorer la qualité des résultats.
- On cherchera à réduire les biais potentiels dans la génération en **retravaillant le prompt et en validant les questions générées**.

REFORMULATION ET ENRICHISSEMENT DES QUESTIONS

Question initiale	Reformulation	Nouvelles questions	Reformulation
Selon le comte de Douville-Maillefeu, quel facteur est primordial pour assurer la victoire d'une nation dans un conflit, même face à un adversaire de force égale ?	En 1881, lors des discussions sur le budget de la marine, quel est le facteur primordial, selon le comte de Douville-Maillefeu, pour assurer la victoire d'une nation dans un conflit, même face à un adversaire de force égale ?	– Quel arsenal le comte de Douville-Maillefeu propose t-il de supprimer lors des discussions sur le budget de la marine en 1881 ? - Quels sont les investissements que le comte de Douville-Maillefeu propose de revoir lors des discussions sur le budget de la marine en 1881 ?	Ajout de la date et du sujet de la discussion.

Exemple de reformulation et génération manuelle de sous-questions

PREMIÈRES OBSERVATIONS

- Pour le moment, l'évaluation est réalisée par une seule experte du domaine. Il paraît nécessaire de **réaliser l'évaluation à plusieurs mains** et de croiser les résultats.
- Les questions générées se concentrent sur des é**léments factuels, faciles à vérifier et non controversés**.
- Cette méthode ne remplace pas la génération manuelle de questions, qui demeure **le moyen le plus fiable** de produire un corpus de qualité.

La question reste : comment évaluer l**es stratégies de chunking dans leur capacité à fournir des informations suffisantes** (retrieving) pour répondre correctement à une question ?

EXPÉRIENCE (1/2)

Questions : 1400 questions « single-hop » issues de débats parlementaires, avec traçabilité de leur source d'origine.

Base vectorielle ChromaDB (L2) – **LLM** : Gemma 3 - 27B – **Reranker** : BGE-v2-M3 – **Embeddings** : mGTE – **GPU** : NVIDIA A40 (46 Go)

Métriques :

$$\text{Recall@k} = \frac{|\text{Rel} \cap \text{Ret@k}|}{|\text{Rel}|} \quad | \quad \text{nDCG@k} = \frac{\text{DCG@k}}{\text{IDCG@k}}, \text{ où } \text{DCG@k} = \sum_{i=1}^{k} \frac{\text{rel}_i}{\log_2(i+1)}$$

EXPÉRIENCE (2/2)

Méthode	Description	Implémentation	Taille moyenne des chunks (nombre de tokens)
S	Découpage par sections	Une Regex identifie les débuts de sections par séquences de majuscules	$10\,989 \pm 14\,302$
S&PP	Sections + Prises de parole	Prises de parole identifiées par le préfixe « M. »	808 ± 1267
S&PP&MaxN	Sections + Prises de parole + Limite de taille	Découpage interrompu si le chunk contient moins de N caractères	204 ± 66 – 1412 ± 940 ($N = 1000$ – $N = 10\,000$)

Description des stratégies de chunking étudiées. Pour chaque méthode, un chunking récursif est appliqué, et les frontières entre les différentes sections sont identifiées à l'aide d'expressions régulières.

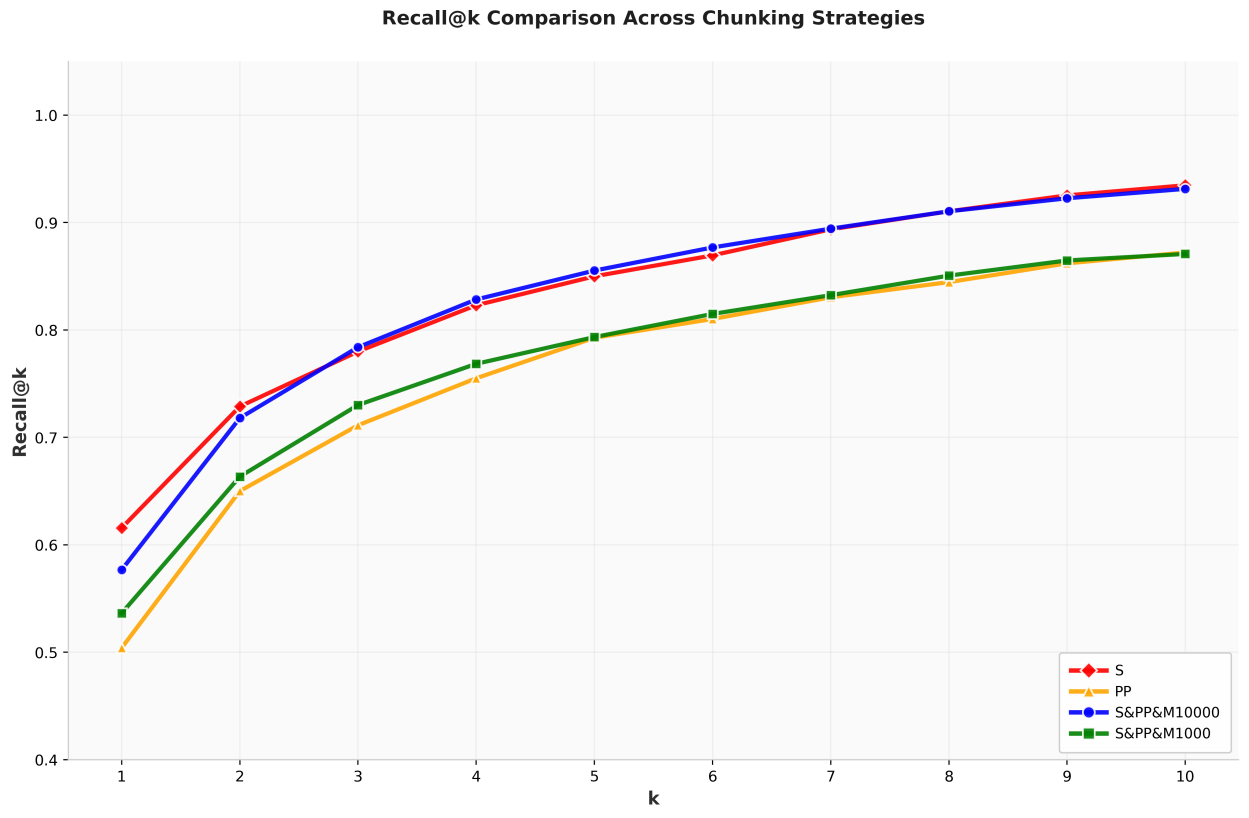
DÉTAIL DU CHUNKING

Extrait d'un chunk en fonction de différentes stratégies	S	S&PP	S&PP&Max10000
	PRÉSIDENCE DE M. HENRI BRISSON La séance est ouverte à deux heures. M. Henry Cochin, l'un des secrétaires, donne lecture du procès-verbal de la séance d'hier. M. le président. La parole est à M. Bourrillon, sur le procès-verbal. M. Bourrillon. Je suis porté par erreur au Journal officiel comme ayant voté « contre » l'ensemble de la proposition de loi relative à la suppression des taxes d'octroi sur les boissons hygiéniques. Je déclare avoir voté « pour ». Ayant dans tous les scrutins qui ont eu lieu émis un vote conforme aux vues de la commission, il n'est en effet pas admissible que j'aie voté « contre » l'ensemble. M. le président. Il n'y a pas d'autre observation sur le procès-verbal ? Le procès-verbal est adopté.	Idem colonne S.	PRÉSIDENCE DE M. HENRI BRISSON La séance est ouverte à deux heures. M. Henry Cochin, l'un des secrétaires, donne lecture du procès-verbal de la séance d'hier.
Nombre de chunks	59	79	255

Impact des stratégies sur la taille des chunks

RÉSULTATS

STRATÉGIES DE CHUNKING VS RECALL@K



Chunking vs Recall@k

Points clés

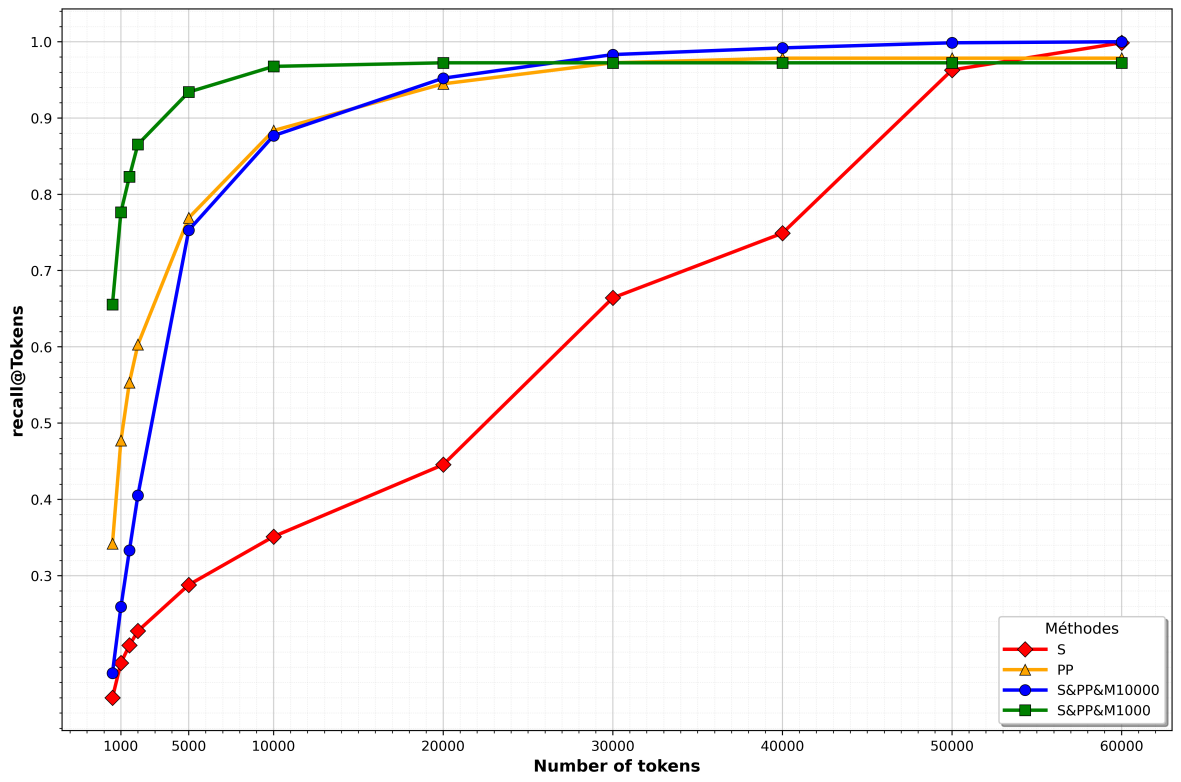
- **S** offre les meilleurs résultats grâce à des chunks très longs (11k tokens), mais limite le retour à la source.
- **S&PP&Max10000** constitue un bon compromis : deuxième meilleure performance avec bien moins de tokens, et un recall proche de S dès 3 documents.
- **S&PP&Max1000** surpasse **PP** malgré des chunks 4× plus petits, probablement grâce à une taille plus homogène favorisant des embeddings de meilleure qualité.
- Pour les petits modèles, **S&PP&Max1000** semble la stratégie la plus adaptée.

DÉTAIL DES RÉSULTATS

Stratégie de Chunking		Taille moyenne des chunks (nombre de tokens)	Recall@k					nDCG@k	
Granularité	**Taille limite** (caractères)		1	2	3	5	10	5	10
S&PP	1000	204 ± 66	0.54	0.66	0.73	0.79	0.87	0.68	0.70
	10000	1412 ± 940	0.58	0.72	0.78	0.86	0.93	0.73	0.74
S	None	10989 ± 14302	**0.62**	0.73	0.78	0.85	0.93	0.74	0.77
PP		808 ± 1267	0.50	0.65	0.71	0.79	0.87	0.66	0.68

Évaluation du retriever pour différentes stratégies de chunking. Pour chaque configuration, nous calculons le nombre moyen de tokens par chunk récupéré sur l'ensemble des questions. La stratégie de segmentation par section donne presque toujours les meilleurs résultats, mais elle est très coûteuse en nombre de tokens. La stratégie **S&PP&Max10000** offre un bon compromis entre coût à l'entrée et performance du retriever.

STRATÉGIES DE CHUNKING VS RECALL@TOKENS

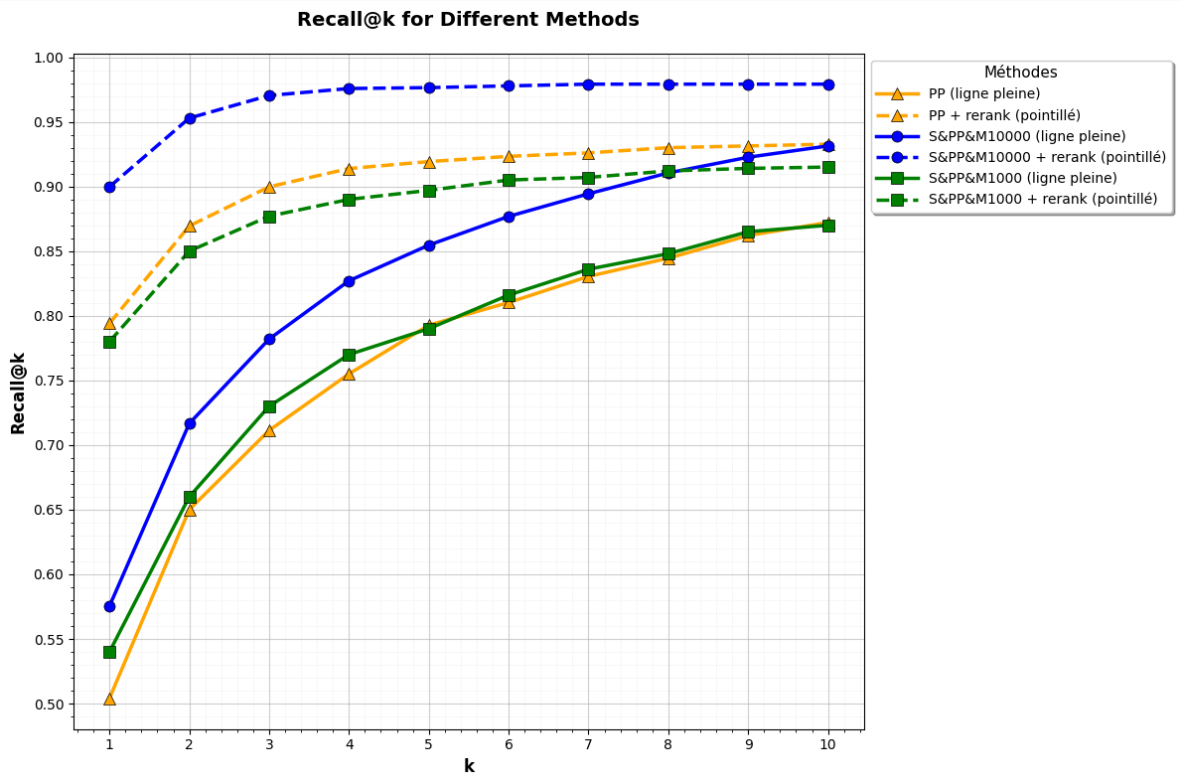


Évaluation des stratégies de chunking,
incluant les questions sans documents récupérés.

Points clés

- Évaluation à nombre de tokens constant : pour chaque stratégie, on récupère un maximum de documents sans dépasser un seuil de tokens fixé. Une question pour laquelle aucun chunk n'est récupéré est considérée comme un échec du retriever.
- S est désavantagée : ses chunks très longs diluent l'information et ne s'adaptent pas aux contextes de petite taille.
- S&PP&Max1000 apparaît comme la stratégie la plus adaptée pour un fabile nombre de tokens en entréee. À partir de 5000 tokens, S&PP&Max10000 et PP offrent des performances similaires, mais la plus faible variance des chunks générés par la première peut en faire une option plus robuste.

L'APPORT DU RERANKER



Comparaison des résultats sur le retriever avec et sans reranking. En ligne pleine les résultats avec un RAG classique et en pointillé avec un RAG augmenté par reranking

Points clés

- Pour chaque question, 20 documents sont récupérés, puis reclassés avec le modèle `bge-reranker-v2-m3` selon un score de similarité (question, document). Le recall est recalculé après reclassement.
- Le reranking améliore nettement les performances : pour $k = 1$, le recall de **PP** passe de 0.5 à 0.79. À partir de $k = 3$, cette stratégie gagne en moyenne 2 points sur **S&PP&Max1000**.
- Le reranker corrige les imprécisions des embeddings, rendant **PP** plus performante. **S&PP&Max10000** reste compétitive, avec un recall supérieur d'au moins 5 points sur l'ensemble des évaluations.

DÉTAIL DES RÉSULTATS

Stratégie de Chunking		Taille moyenne des chunks (nombre de tokens)	Coût de calcul (temps)	Recall@k			
Granularité	**Taille limite** (caractères)			1	2	5	10
S	—	10989 ± 14302	<1min	0.62	0.73	0.85	0.93
S&PP	1000	204 ± 66	<1min	**0.54**	0.66	0.79	0.87
	10000	1412 ± 940	<1min	0.58	0.72	0.86	0.93
PP	—	808 ± 1267	<1min	0.50	0.65	0.79	0.87
S&PP+reranking	1000	204 ± 66	9min	0.78	0.85	0.89	0.92
	10000	1412 ± 940	1h47min	**0.90**	0.95	0.98	0.98
PP+reranking	—	808 ± 1267	2h25min	0.79	0.87	0.92	0.93

Évaluation du retriever pour différentes stratégies de chunking. Pour chaque configuration, nous calculons le nombre moyen de tokens par chunk récupéré sur l'ensemble des questions. La stratégie de segmentation par section donne presque toujours les meilleurs résultats, mais elle est très coûteuse en nombre de tokens. La stratégie S&PP&Max10000 offre un bon compromis entre coût à l'entrée et performance du retriever.

Points clés

- Le coût du reranking est élevé : environ 2h d'inférence pour **S&PP&Max10000** et **PP**. **S&PP&Max1000** constitue un bon compromis : pas de reranking, temps d'inférence < 10 min, et résultats proches des meilleures stratégies.

CONCLUSION

- Améliorer la phase de génération de questions
- Produire des questions portant sur plusieurs sections d'un même débat ou sur plusieurs débats
- Après la phase de retrieval, évaluer la qualité de la réponse générée dans le cadre du RAG
- Poursuivre l'annotation automatique et introduire de nouvelles tâches pour comparer les stratégies de chunking

MERCI POUR VOTRE ATTENTION !

Aurélien Pellet : aurelien.pellet@epita.fr

Julien Perez : julien.perez@epita.fr

Marie Puren : marie.puren@epita.fr