Benchmark des techniques de spécialisation de LLM

Rapport de projet -
CentraleSupélec - Université Paris-Saclay, Gif-Sur-Yvette

Groupe
BARAKAT Kenzy – DAVID Erwan – PUTEGNAT Theo

Résumé du projet

Ce projet vise à améliorer la spécialisation d'un modèle de langage sur des domaines de connaissances spécifiques, en surmontant les limites liées à l'apprentissage sur des données statiques. En exploitant les techniques de RAG et GraphRAG, nous cherchons à enrichir les réponses du modèle avec des informations contextuelles pertinentes. Nous comparerons l'efficacité de ces approches sur différentes tâches, avec des nouveaux sujets non appris initialement par le LLM. L'objectif est de réaliser un benchmark des meilleures pratiques pour adapter et spécialiser un modèle de langage afin d'optimiser ses performances dans des contextes spécifiques.

Table des matières

Contexte et problème

Contexte

Les modèles de langage offrent un compromis intéressant entre performance et coût de calcul, en particulier lorsqu'ils sont accessibles via des API. Leur capacité à comprendre et générer du texte les rend adaptés à une large gamme d'applications. Toutefois, leur formation sur des corpus massifs, mais figés peut limiter leur connaissance des sujets récents ou très spécifiques.
Pour pallier ces limitations, nous nous intéressons à des techniques de spécialisation permettant d'adapter un modèle à des tâches précises ou à des domaines de connaissance particuliers. Parmi ces techniques, la génération augmentée par la récupération (RAG) et sa variante GraphRAG, qui intègre des graphes de connaissances, offrent des bons compromis entre efficacité technique et efficacité énergétique. Elles offrent des perspectives prometteuses. En combinant les capacités de génération de langage d'un modèle avec la recherche d'informations pertinentes dans une base de connaissances, ces approches permettent d'enrichir les réponses et de mieux contextualiser l'information.

Problèmes

Les modèles de langage peuvent rencontrer des difficultés lorsqu'ils sont confrontés à des informations qui ne sont pas présentes ou suffisamment représentées dans leurs données d'entraînement. Ces limitations se manifestent de deux manières principales :

- Connaissances inconnues : Les modèles peuvent être incapables de traiter des sujets émergents ou très spécifiques, car ils n'ont pas été exposés à ces informations lors de leur apprentissage initial.
- Connaissances incomplètes ou erronées : Même si un sujet est présent dans les données d'entraînement, il peut être sous-représenté ou associé à des informations incorrectes, ce qui peut conduire à des hallucinations ou à des réponses imprécises.

Pour remédier à ces problèmes, l'utilisation de techniques telles que RAG ou GraphRAG apparaît comme une solution prometteuse. Cependant, leur mise en œuvre soulève plusieurs défis :

- Pertinence de la récupération : Comment sélectionner les informations les plus pertinentes dans une base de connaissances pour répondre à une requête donnée ?
- Intégration des informations : Comment intégrer de manière cohérente les informations récupérées dans la génération de texte ?
- Complexité des tâches : Comment gérer des tâches qui nécessitent une compréhension fine des relations entre les entités et les concepts ?

Notre projet vise à étudier comment ces techniques peuvent être appliquées pour améliorer la capacité d'un modèle LLM sur des connaissances soit absentes de ses données d'entraînement, soit sous-représentées. Nous identifierons les limites et les défis associés à leur utilisation.

Présentation RAG & GraphRAG

GraphRAG et RAG sont au cœur des recherches actuelles sur les LLM car ils permettent d'enrichir les réponses générées par les modèles en les connectant à des bases de connaissances externes.

Fonctionnement du RAG

Le RAG fonctionne en deux étapes principales :

1. **Représentation des données**
   Cette étape n'est nécessaire qu'un fois lors du passage initial des documents de contexte au LLM.

   - Découpage : Les documents sont d'abord découpés en morceaux de taille adaptée (chunks), par exemple à l'aide d'un RecursiveCharacterTextSplitter (Langchain), afin d'optimiser la granularité des futures recherches.
   - Encodage : Chaque chunk est ensuite transformé en un vecteur numérique (embedding) qui capture son sens et son contexte. Dans notre cas, nous utilisons les OpenAIEmbeddings pour cette opération.
   - Stockage : L'ensemble de ces vecteurs est ensuite stocké dans une base vectorielle (ici ChromaDB), qui permet de retrouver rapidement les morceaux de texte les plus proches d'une requête.

2. **Récupération et génération du prompt**
   - Recherche sémantique : Lorsqu'un utilisateur pose une question, celle-ci est, elle aussi, encodée en vecteur. Une recherche de similarité cosinus est effectuée dans ChromaDB pour identifier les passages les plus pertinents.
     Remarque : il aurait été intéressant de tester une recherche pondérée entre cosine et mot clé. Cependant, la base de données ChromaDB ne le propose pas nativement.
   - Construction du prompt : Les chunks retrouvés sont assemblés avec la question initiale pour former un prompt enrichi. Ce prompt est ensuite envoyé au modèle (ici GPT-4o via API), qui génère une réponse en tenant compte à la fois du contexte trouvé et de l'historique de la conversation (grâce notamment au buffer memory de Langchain).

Nous avons en particulier utilisé les outils suivants : *GPT-4o, ChromaDB, OpenAIEmbeddings, langchain (RecursiveCharacterTextSplitter, ConversationBufferMemory, ChatOpenAI).*

Voici un schéma récapitulatif du pipeline :

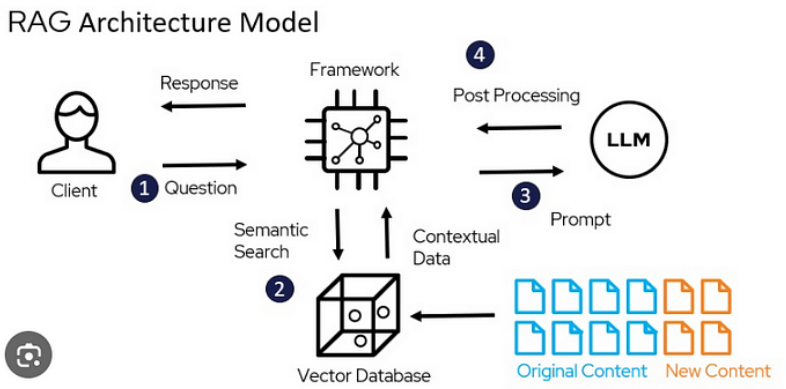


Figure 0 : Schéma explicatif du RAG

Fonctionnement du GraphRAG

Le GrapRAG est une variante du RAG pour laquelle est utilisé un graphe de connaissances pour représenter des relations sémantiques et retrouver des informations plus contextuelles.

Le graphRAG fonctionne de la manière suivante :

1. **Représentation des données**
   Cette étape n'est nécessaire qu'une fois lors du passage initial des documents de contexte au LLM
   - Extraction des entités et des relations : Chaque chunk de document est analysé par un LLM pour extraire automatiquement : Les entités clés (personnes, organisations, concepts…) & Les relations entre ces entités (appartenance, collaboration, dépendance…)
   - Construction du graphe de connaissances : Ces entités et relations sont utilisées pour construire un graphe local pour chaque chunk. Les graphes issus des différents documents sont ensuite fusionnés pour créer un graphe de connaissances global, structurant l'ensemble des données sous forme de nœuds (entités) et d'arêtes (relations).
   - Identification des communautés : L'algorithme de clustering *Leiden* est appliqué au graphe global pour détecter des sous-ensembles fortement connectés, appelés *communautés*. Ces communautés correspondent à des thématiques ou des ensembles d'informations naturellement liées.
   - Résumé des communautés : Un LLM est utilisé pour générer, pour chaque communauté détectée, un résumé concis et représentatif de son contenu. Ce résumé servira de point d'entrée rapide pour la recherche ultérieure.

2. **Récupération et génération du prompt**
   - Recherche locale (*Local Search*) : Une recherche de similarité est réalisée dans le graphe à partir des entités présentes dans la requête. Les nœuds les plus proches sont identifiés, et les chunks de texte associé à ces nœuds sont récupérés.
   - Recherche globale (*Global Search*) : En parallèle, les communautés précédemment identifiées sont classées par un LLM en fonction de leur pertinence vis-à-vis de la requête. La communauté jugée la plus pertinente, ainsi que les chunks liés, sont sélectionnés.
   - Génération du prompt : Les résultats des recherches locales et, ou globales sont ensuite combinés pour construire un prompt qui sera envoyé au LLM. Ce prompt peut inclure les entités clés et leurs relations, les résumés de communautés ou les passages de texte les plus pertinents.

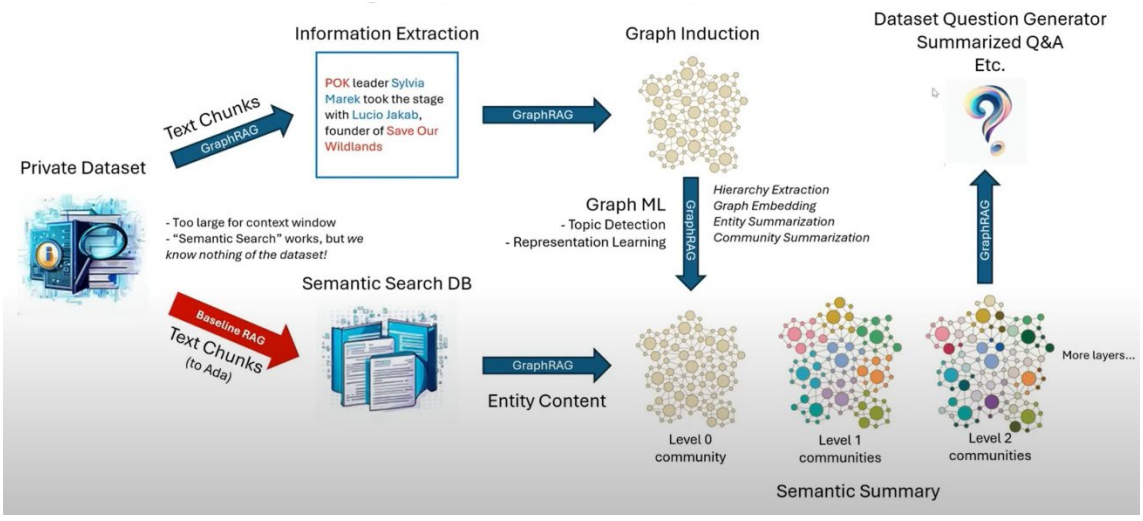


Figure 0 bis : Schéma explicatif du fonctionnement du GraphRAG de Microsoft

Schéma d'ensemble

Fonctions conçues

Notre projet s'articule autour de la recherche de méthode d'actualisation de LLM ainsi que l'évaluation de ces méthodes. Nous avons implémenté des algorithmes, dont la liste est donnée ci-dessous :

- Implémentation RAG
- Implémentation GraphRAG
- Scraping web + création corpus articles (python)

Livrable

Le livrable le plus important de notre projet reste l'analyse des différentes méthodes ainsi que leurs évaluations respectives, regroupées sous le terme de Benchmark. Notre travail a également conduit à des livrables intermédiaires que vous pourrez retrouver dans la liste ci-dessous :

- Benchmark GraphRAG – RAG – LLM online
- Notation humaine via questionnaire
- Notation par LLM juges
- Création d'un corpus de test cohérent
- Outil d'analyse de résultats de l'évaluation de LLMs (Excel)

Corpus de documents

Choix du corpus

Le choix du corpus sur lequel réaliser nos tests a été un choix crucial. Ce corpus nécessitait d'avoir une grande quantité d'articles écrits dans un bon anglais et de bonne qualité. De plus, le corpus doit parler d'un thème suffisamment récent pour garantir que chat GPT ne se soit pas encore entrainé sur des données reliées à cet élément / ce thème. Il a donc fallu trouver un sujet d'actualité et suffisamment couvert par les médias. Un corpus sur les élections américaines et l'accession au pouvoir de Donald Trump s'est rapidement imposé comme un très bon choix. Sachant que le dernier entrainement de GPT date de juin 2024, nous nous sommes également limités dans le temps, en prenant tous les articles sur le thème de Donald Trump entre le 3 juin 2024 et le 26 janvier 2025, ce qui couvre une bonne partie de la campagne électorale ainsi que l'élection du nouveau président et quelques-unes de ses premières déclarations tout en garantissant de traiter des informations qui ne sont pas contenues dans les poids de GPT.

Extraction des articles

Une fois le thème choisi, il a fallu extraire les données depuis internet. Pour cela, de nombreuses API proposent de récupérer des articles en fonction d'un thème donné, dans notre cas le thème recherché étant assez large : "Trump".
Dans un premier temps, l'utilisation de librairies python classiques comme *the-news-api* ou encore *news-api* a très bien fonctionné. Cependant, la qualité des articles était beaucoup trop hétérogène. Nous avons relevé quelques points problématiques :

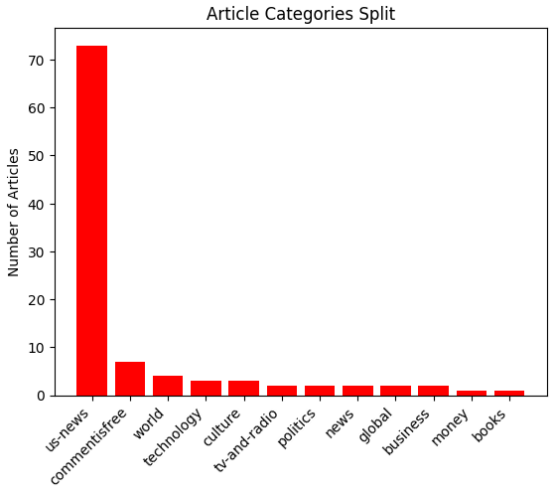
- Mauvaise qualité d'articles liés à un site peu fiable / peu connu (mauvaise couverture du sujet en général, articles "people" avec peu de connaissances)
- Mauvaise qualité du découpage de l'article (article vide, présence de pub, de liens hypertextes, etc.…)

Face à ces problèmes, il a fallu modifier notre approche. L'API du Guardian, journal Anglais propose une API de très bonne qualité : articles complets et fiables, bonne couverture des événements, contenu parfaitement

découpé sans artefacts. Nous avons donc choisi cette API qui possède toutes les qualités recherchées. La présence d'un seul journal dans le corpus pourrait créer quelques biais qui sont en réalité négligeables dans notre étude qui reste qualitative sur le graphe créé.

Analyse du corpus obtenu

Une fois les articles récupérés, nous pouvons essayer rapidement d'analyser la qualité du corpus ainsi créé. Premièrement, comme prévu par l'API, tous les articles comportent au minimum une mention au nouveau président élu dans leur titre, ce qui garantit une cohérence du thème global du corpus.



Ensuite, les articles récupérés sont classés par le Guardian dans différentes catégories présentées ci-contre (Fig.1), qui garantissent une fois encore une relative bonne couverture du sujet et de ses enjeux.

Fig. 1 - Répartition des articles selon les catégories du Guardian

La dernière chose à vérifier était la couverture temporelle des évènements, afin de garantir que la campagne et l'élection du président était couverte. Comme le montre le graphique suivant (Fig.2), les articles couvrent parfaitement toute la période que nous avons voulue scraper initialement.

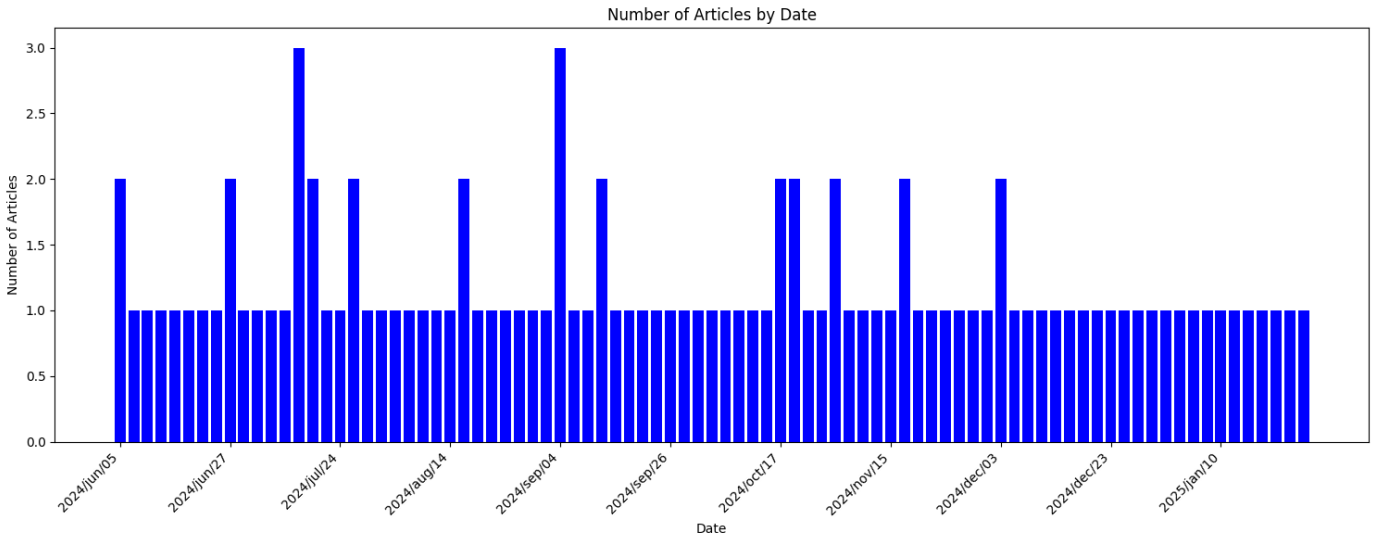


Fig. 2 - Répartition temporelle des articles

Finalement, on obtient le corpus suivant autour de Donald Trump et de l'élection présidentielle : 102 articles, 89 400 mots. Nous appellerons par la suite le corpus présenté ici **Trump24-25**.

Le knowledge Graph – GraphRAG

Définition knowledge graph

Un knowledge graph (ou graphe de connaissances) est une structure de données qui représente des informations sous forme de graphes, reliant des entités (personnes, lieux, objets, concepts) par des relations sémantiques. Il permet de modéliser et d'organiser la connaissance de façon structurée et interconnectée. Chaque nœud représente une entité et chaque lien une relation spécifique entre ces entités. Les knowledge graphs sont largement utilisés pour améliorer la recherche d'informations, l'intelligence artificielle et la compréhension automatique des données.

Exemple de knowledge graph

YAGO est une base de connaissances qui structure des millions de faits sous forme de graphes en reliant des entités (personnes, lieux, événements) via des relations sémantiques précises. Ces graphes sont construits automatiquement en extrayant des informations de sources fiables comme Wikipédia, WordNet et GeoNames, puis en les organisant sous une ontologie hiérarchique cohérente. Nous avons extrait une partie de ce knowledge graph consacré à Donald Trump, afin de comparer ce knowledge graph à ceux que nous allons générer.

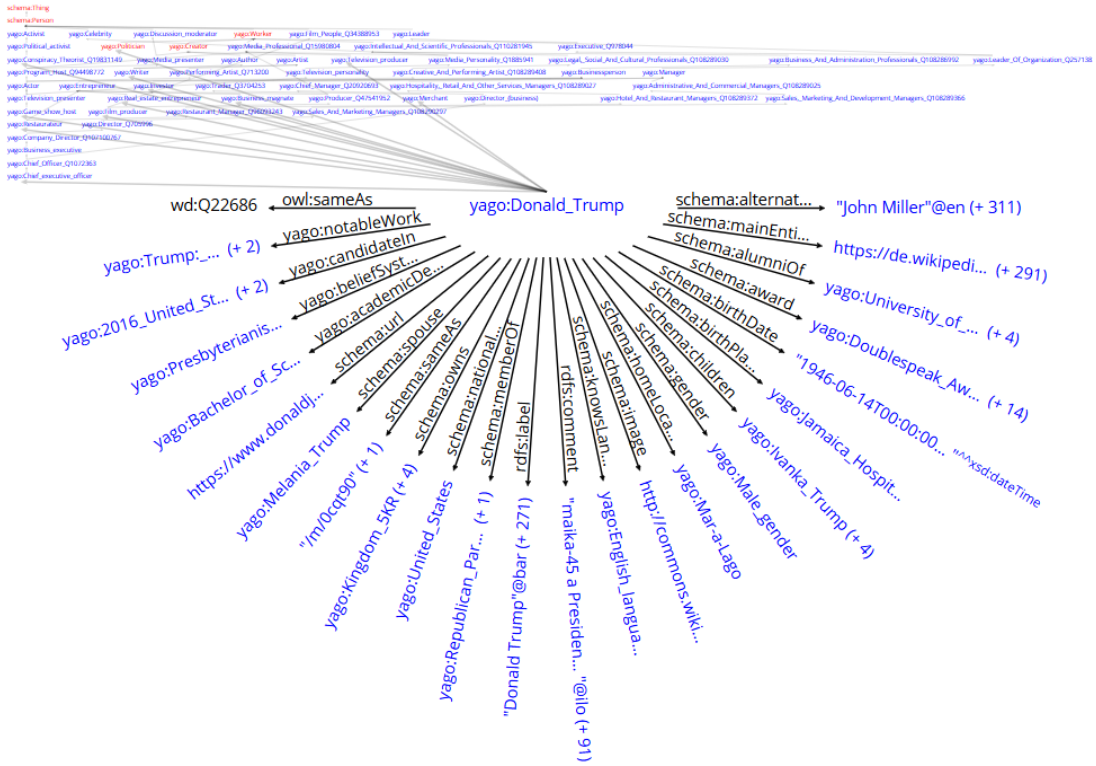


Fig. 3 - Knowledge graph YAGO et nœud "Donald Trump"

On voit que le knowledge graph (Fig.3) est vraiment axé sur une description générale de la personne : âge, enfants, étude, métier ou activité, etc. Nous ne trouvons pas ici d'informations sur des détails ou des évènements récents qui auraient pu arriver à cette personne. C'est un pur graphique informatif et descriptif de D. Trump. On peut s'attendre à générer des graphes plus spécialisés, contenant des informations sur des évènements spécifiques de la campagne par exemple.

Présentation du Graph

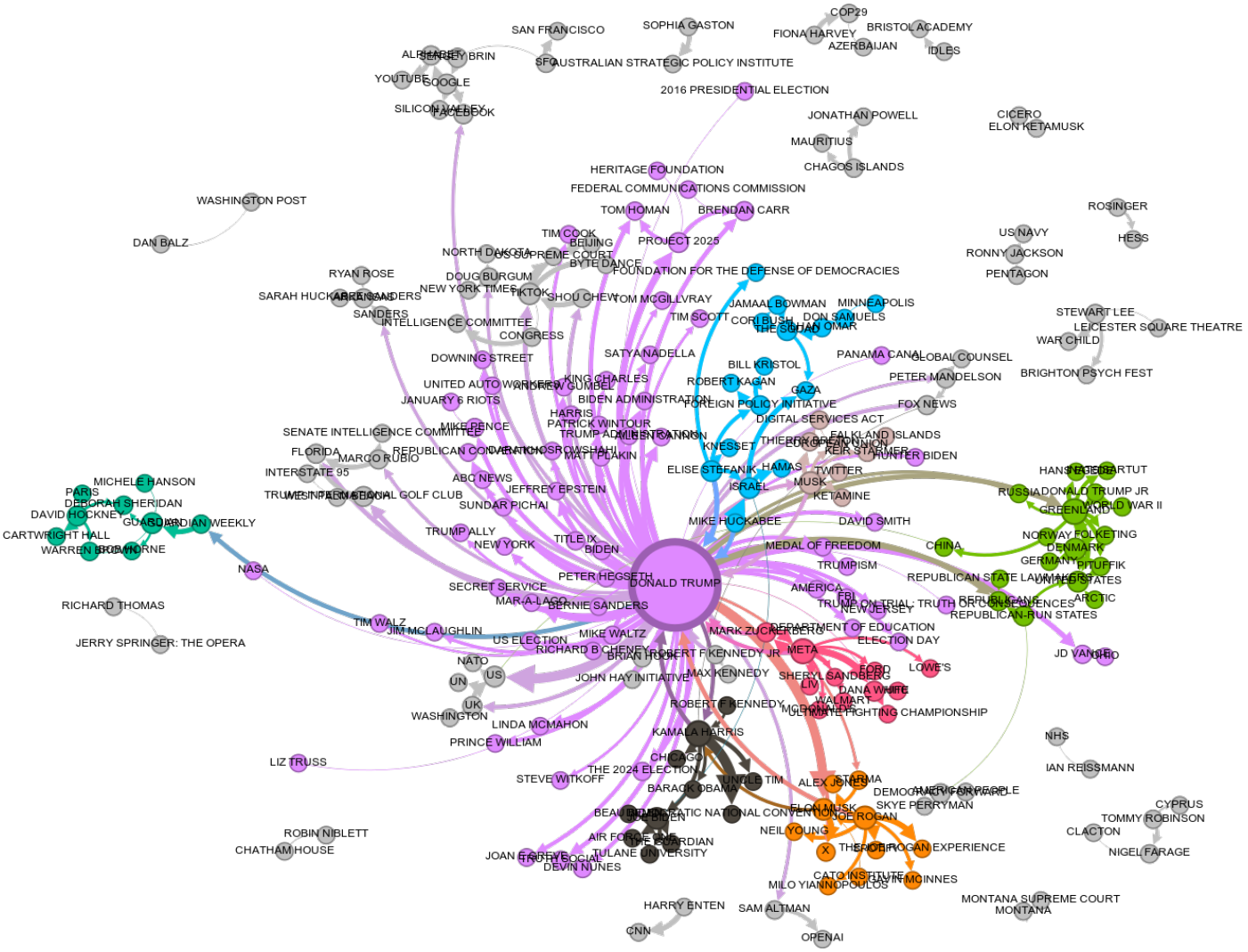


Fig 4 - Graphe de connaissance obtenu via le corpus Trump24-25

Le graphique ci-dessus (Fig. 4) a été obtenu après traitement par GPT-4o de notre corpus Trump24-25, en utilisant le dépôt GraphRAG officiel de Microsoft. La précision de certains hyperparamètres sera plus détaillé dans la section suivante qui y est consacrée.
On observe déjà une relative cohérence du graphe obtenu. A noté que le graphe présenté a été légèrement édité visuellement pour fusionner les nœuds “Trump” et “Donald Trump”. On voit clairement que ce nœud agit comme un nœud central du graphe, ce qui confirme la bonne qualité du corpus d’articles autour de Donald Trump. Plusieurs clusters sont identifiables par les couleurs, correspondant également à des groupes cohérents et en relation avec le nouveau président des États-Unis. Par exemple, le bloc noir pourrait être catégorisé comme le bloc démocrate, faisant référence à Kamala Harris ou encore Barack Obama.

GraphRAG calcule également des poids qui décrivent l’importance des transitions entre nœuds. Ils sont calculés en fonction du nombre de mentions dans le corpus et de leur pertinence contextuelle calculée au sein des communautés.

Le poids moyen d'une connexion est 6,67, avec une majorité des relations entre 6 et 8. Ils vont de 1 à 18.

Connexions les plus fortes (poids > 16)

- Kamala Harris → Democratic National Convention (18) : Logique, car Harris y a probablement joué un rôle majeur.
- Skye Perryman → Democracy Forward (18) : Une connexion forte, logique étant donné que Skye Perryman est le CEO de Democracy Forward
- Trump → US (17) : Une relation attendue, mais son poids élevé montre une forte présence du lien Trump-USA dans le corpus.
- Donald Trump → Elon Musk (16) : Indique une connexion médiatique fréquente entre les deux.
- Trump → Truth Social (16) : Lien fort avec sa propre plateforme de communication.

Connexions les plus faibles (poids = 1)

- Donald Trump → Panama Canal : Une mention unique d'un lien entre Trump et le canal de Panama.
- Donald Trump → Dana White : Bien qu'ils soient amis, le lien est peu mentionné dans le corpus.
- Trump → NASA : Curieux que ce lien soit faible, peut-être en raison d'un désintérêt du corpus sur ce sujet.
- Kamala Harris → Fox News : Surprenant, car on pourrait s'attendre à plus de mentions sur ses interactions avec la chaîne conservatrice.
- Mike Pence → January 6 Riots : Malgré son rôle crucial dans cet évènement, les liens n'ont pas dû être mentionnés dans beaucoup d'articles.

Étude de variabilité et d'influence des hyperparamètres

Une question importante lors de la génération de ce type de graphe reste la répétabilité et la stabilité ou non de ce type de génération. Étant donné que ces graphes de connaissances sont générés par des LLMs et prompts adaptés, le caractère non déterministe des LLMs nous invite à penser que les graphes obtenus n'ont aucune garantie d'être identique d'une génération à l'autre.
Pour vérifier cela, nous avons lancé différentes générations de graphes avec le corpus Trump24-25 en faisant varier quelques hyperparamètres de l'outil G-RAG comme la *chunk_size* et *max_cluster_size*. Nous avons également voulu tester l'influence de la concaténation des articles dans un seul document, ou à l'inverse leur séparation dans des documents distincts. Les résultats sont présents ci-dessous (Fig5.)

Par souci de clarté, nous noterons :
max_cluster_size: MC,
chunk_size: CS,
documents_séparés: S = TRUE / FALSE,
claim_extraction: CE = TRUE/FALSE.

Identifiant	1	2	3	4	5	6	7
MC	10	30	30	10	10	10	10
CS	800	800	800	800	800	1200	800
S	False	True	False	True	True	False	False
CE	True	True	True	True	True	False	True
Nœuds	228	137	108	119	149	127	116
Relations	247	147	105	146	164	131	117

Fig 5 - Caractéristiques des différents knowledge graph générés

Le premier résultat est sans appel : **la génération de G-RAG n'est pas déterministe** comme le montre le tableau (Fig 5.). En plus de cela, la variation est suffisamment importante pour rendre difficile l'interprétation des influences de chaque hyperparamètre. Clairement, il aurait été nécessaire de mettre en place ici une étude statistique sur beaucoup plus d'échantillons de graphes, mais le coup monétaire et temporel de génération de G-RAG via Open-AI nous en a dissuadé pour cette étude préliminaire. Quelques interprétations sont quand même à noter.

Premièrement, nous n'avons jamais réussi à retrouver une valeur de nœuds et de relations similaires à ceux du graphe utilisé pour notre étude. Le graphe utilisé pour notre benchmark est bien plus fourni que ceux générés ici. Pour estimer la variation, le plus rigoureux est ici de regarder les deux colonnes possédant exactement les mêmes paramètres (4 & 5). On voit malgré cela que le nombre de nœuds varie de 25 % passant de 119 à 149 et que les relations varient de 12 % passant de 146 à 164.
Clairement, la variation provoquée par le non-déterminisme du LLM nous fait poser quelques questions quant au caractère reproductible de ce type de technique.

Ensuite, il semblerait que certaines hypothèses soient vérifiées, bien qu'encore ces résultats soient à prendre avec beaucoup de recul étant donné la variabilité présentée ci-dessus. Si on compare les colonnes 5 et 7, il semblerait que séparer les documents aide légèrement le LLM à dissocier les articles et leur contenu, aboutissant à une légère augmentation du nombre de nœuds. De plus, il semblerait que réduire la taille maximale des clusters augmente légèrement le nombre de nœuds, en tout cas, c'est une hypothèse que l'on peut raisonnablement poser. À l'inverse, nous ne pouvons pas avec les résultats présents, conclure un quelconque impact sur la taille des chunks.

Voici les 6 knowledge graphs générés (Fig. 6) correspondant aux identifiants dans le tableau précédant, à noter que quelques-uns sont tous en annexe dans un format plus lisible. Ils possèdent une certaine cohérence structurelle et visuelle, mais aucune garantie d'existence ne peut être donné sur un nœud spécifique.

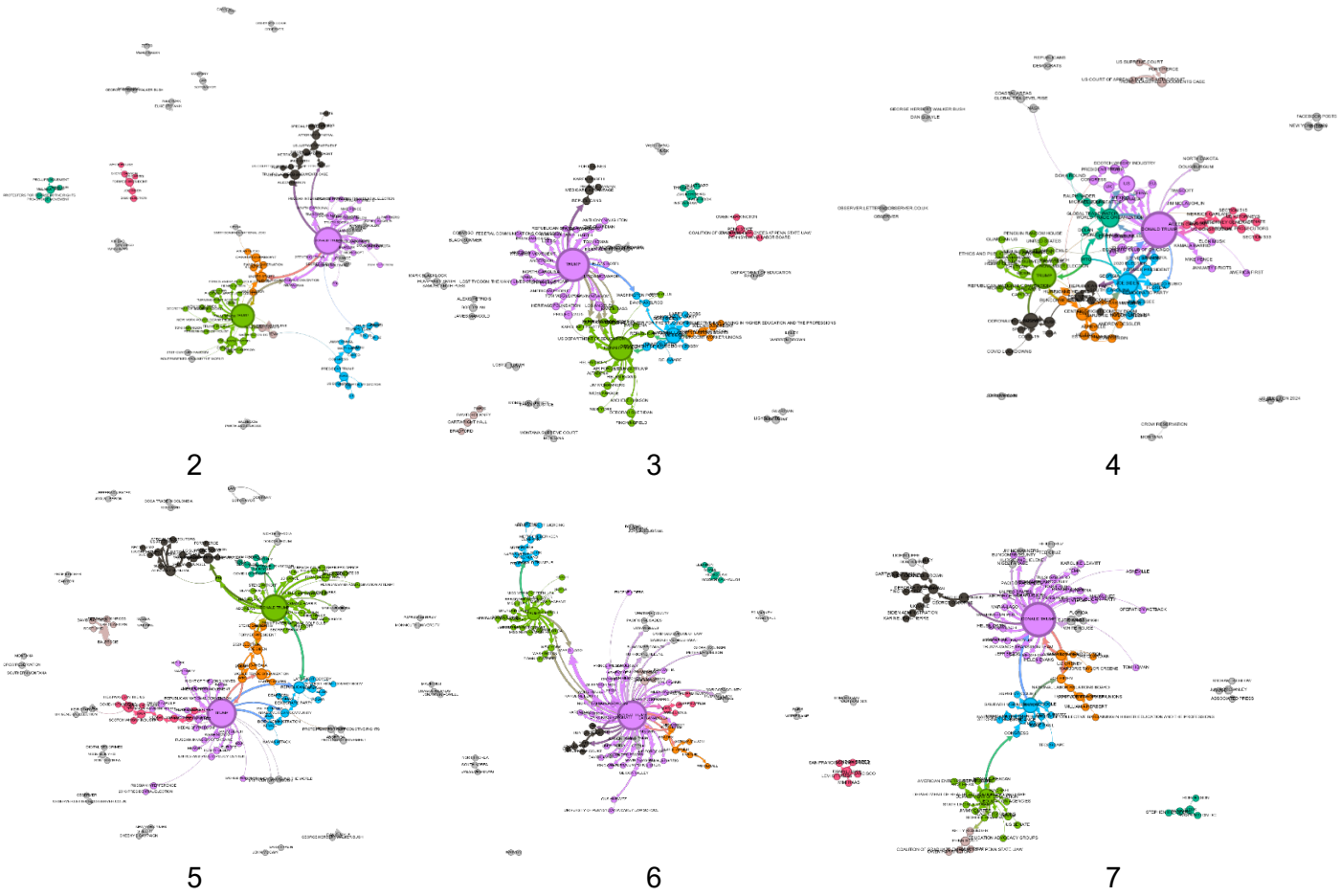


Fig 6 - Knowledges graphs visualisés par Gephi

Complexification du graphe

Nous nous intéressons dans cette partie à l'influence de l'ajout d'un second corpus de texte lors de la génération d'un graphe. L'idée est ici de développer une intuition sur la potentielle scalabilité de cette méthode d'actualisation. Est-il possible de rajouter de nouvelles données, sans détériorer les données précédemment captées par le graphe.

Pour rappel, notre GraphRAG a été généré à partir d'un corpus de 102 articles, c'est-à-dire 89 400 mots.
Nous ajoutons à cette base de connaissance un nouveau corpus sur le conflit en Ukraine. Sa taille est relativement identique au corpus précédant : 99 articles pour 104 956 mots et s'étend sur la même période. Les articles proviennent également du Guardian.

Pour ces 2 graphes, nous utiliserons les paramètres suivants :

- Séparation des articles
- *Max_cluster_size* = 10
- *Chunk_size* = 800
- *Claim_extraction* = True

Voici le knowledge graph généré avec GraphRAG pour le corpus sur le conflit en Ukraine seul (Fig. 7) :

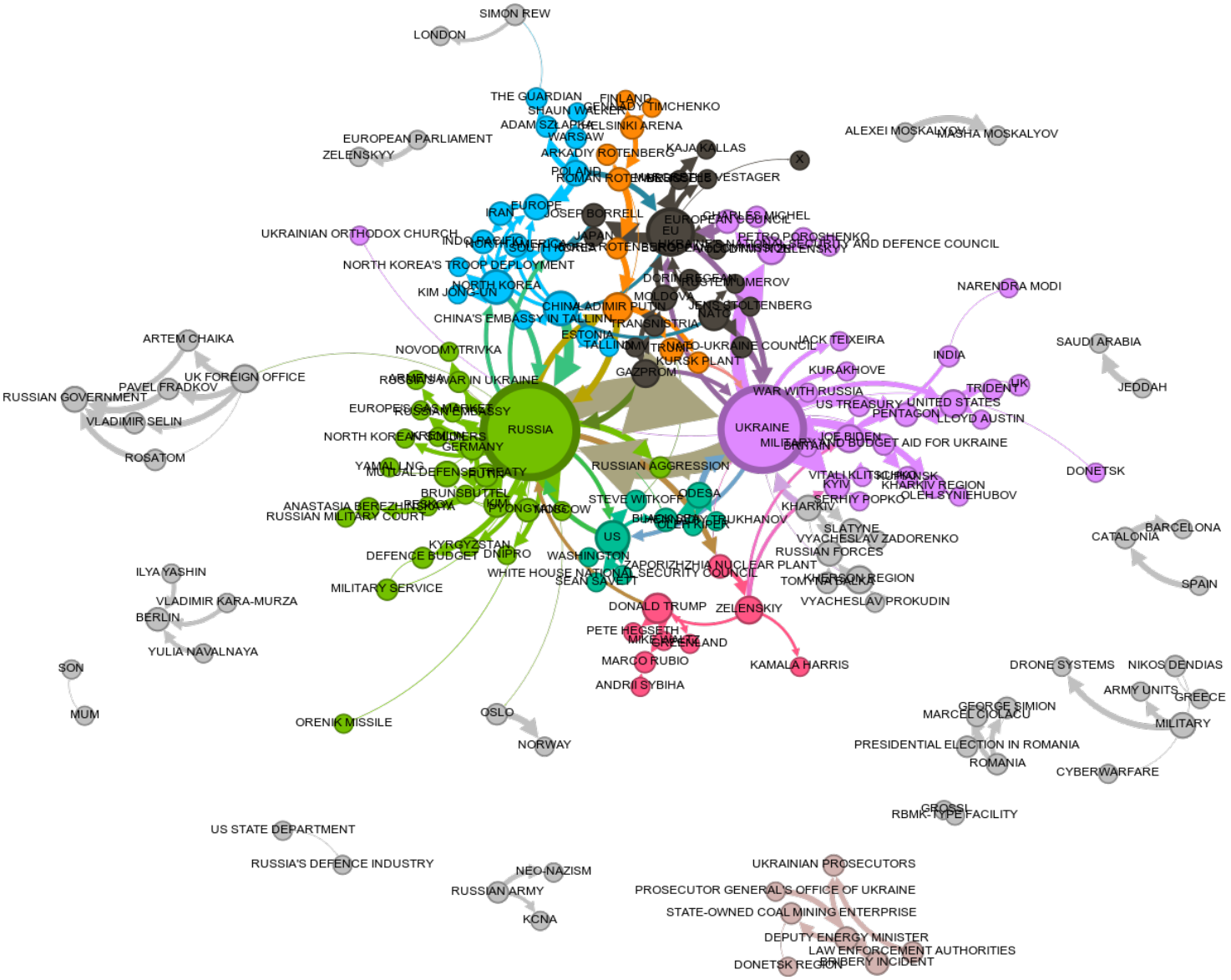


Fig 7 - Graph de connaissance Ukraine, 170 noeuds et 215 relations

Voici le knowledge graph généré avec GraphRAG sur les 2 corpus, soit près de 200 articles :

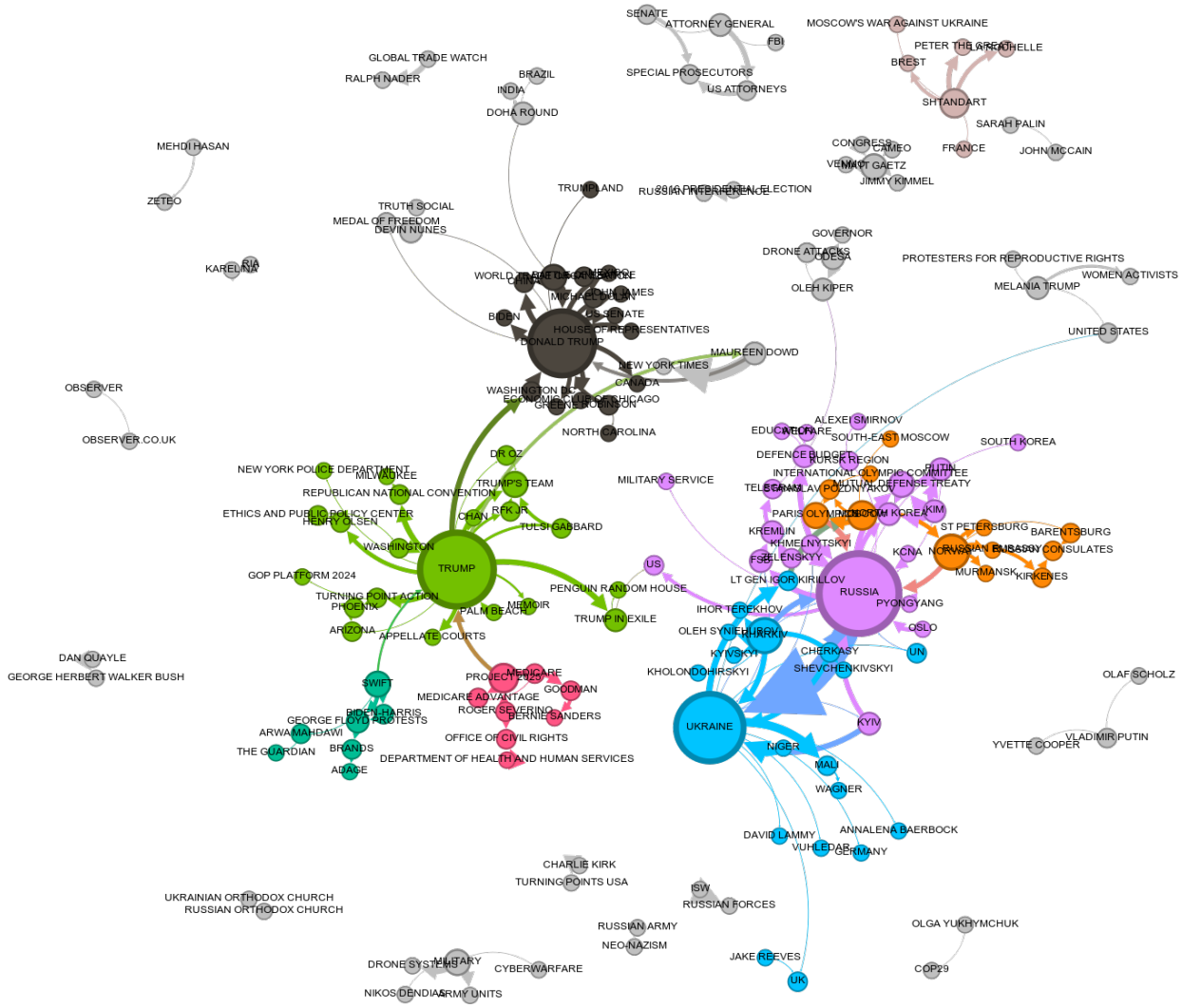


Fig. 8 - Graph de connaissance Ukraine, 170 nœuds et 182 relations

C’est étonnant de constater qu’il ne semble pas avoir de corrélation directe entre la taille du corpus et la taille du knowledge graph. Peut-être que l'ajout du corpus sur l'Ukraine à Trump24-25 à un impact destructif sur certains liens. En mélangeant les sujets, certaines relations fortes dans le corpus sur l’Ukraine peuvent être noyées ou jugées moins pertinentes par le modèle, réduisant ainsi leur pondération et donc leur inclusion dans le graphe final. Il y a donc eu une simplification ou une perte d’information.

Le graphe révèle une connexion unique et surprenante entre les corpus "Trump" et "Ukraine", qui ne se rejoignent qu’à travers le livre *Trump in Exile*, publié aux États-Unis (nœud "US", Fig. 8). Étonnamment, les États-Unis apparaissent aussi comme un acteur surveillant les actions russes, mais avec un poids de seulement 5, bien que les autres relations entre l’Ukraine et la Russie affichent des poids allant jusqu’à 36. Cette sous-pondération est frappante, d'autant plus que Trump est mentionné plus de 370 fois dans le corpus ukrainien, suggérant qu'il devrait être bien plus central dans les connexions. Il est d’ailleurs directement relié à Russie, avec un faible poids, dans le knowledge graph Ukraine (Fig7.). De manière encore plus troublante, Elon Musk est totalement absent du graphe côté Trump, alors qu’il est cité 15 fois dans le corpus Ukraine, malgré son rôle clé dans les communications militaires ukrainiennes via Starlink. Ces déséquilibres suggèrent soit une distorsion dans les sources, soit une absence de liens explicites dans les données, remettant en question la manière dont les relations entre ces entités sont structurées par la méthode GraphRAG.

Benchmark RAG vs GraphRAG

Une fois les graphes de connaissances générés, nous avons souhaité développer une méthode permettant de comparer le plus précisément possible différentes approches d'actualisation d'un LLM sur notre corpus **Trump24-25**.

Notre benchmark

Peu importe la méthode d'évaluation retenue, toutes reposent sur un ensemble de 20 questions que nous avons soigneusement rédigées dans un objectif précis : évaluer la capacité d'actualisation du LLM. Ces questions couvrent différents types de situations. Certaines visent à déterminer si le LLM se limite à ses connaissances générales ou s'il est capable de fournir une information à jour, intégrant les dernières évolutions. Par exemple, à la question « Quel président des États-Unis a subi une tentative d'assassinat ? », quelle sera sa réponse ? Le LLM mentionnera-t-il Donald Trump ou non ? Étant donné que l'événement a eu lieu en juillet 2024, il ne figure pas dans les poids actuels de GPT. D'autres questions se concentrent sur des faits plus spécifiques mentionnés dans les articles du corpus.

Classement humain

Nous commençons par la méthode d'évaluation la plus classique et attendue : une évaluation humaine des réponses générées.

Méthode order rank et méthode de Dowdall

Nous comparons ainsi quatre méthodes :

- **ChatGPT en ligne**, qui s'appuie sur une connexion active à Internet pour actualiser ses réponses ;
- **Un système RAG classique** appliqué à notre jeu de données ;
- **Graph RAG** avec une **stratégie locale** ;
- **Graph RAG** avec une **stratégie globale** ;

Pour chacune des 20 questions, nous générons quatre réponses, une par méthode. Il est alors demandé aux évaluateurs de classer l'ensemble des réponses de la meilleure à la moins pertinente. À partir de ces classements, nous calculons un score selon la méthode de Dowdall.
La méthode de Dowdall attribue à chaque réponse un score proportionnel inverse à son rang : un classement en première position rapporte 100 points, un classement en deuxième position rapporte 50 points, en troisième 33,33 points, etc. Le score final d'une réponse correspond à la moyenne arithmétique des scores obtenus sur l'ensemble des évaluations.

Afin de comparer les différentes approches, nous agrégeons les résultats en calculant la moyenne et l'écart type des scores pour chaque méthode. Ces deux statistiques permettent d'établir un classement global ainsi que d'analyser la variabilité de la qualité des réponses fournies.

Les réponses étaient présentées anonymisées et dans un ordre aléatoire.

Résultats score humain

Nous avons utilisé l'outil OpinionX, qui permet de calculer automatiquement les scores selon la méthode de Dowdall pour chaque question. Au total, nous avons recueilli 12 formulaires complets, chacun portant sur 19 questions, ce qui représente 228 classements au total.

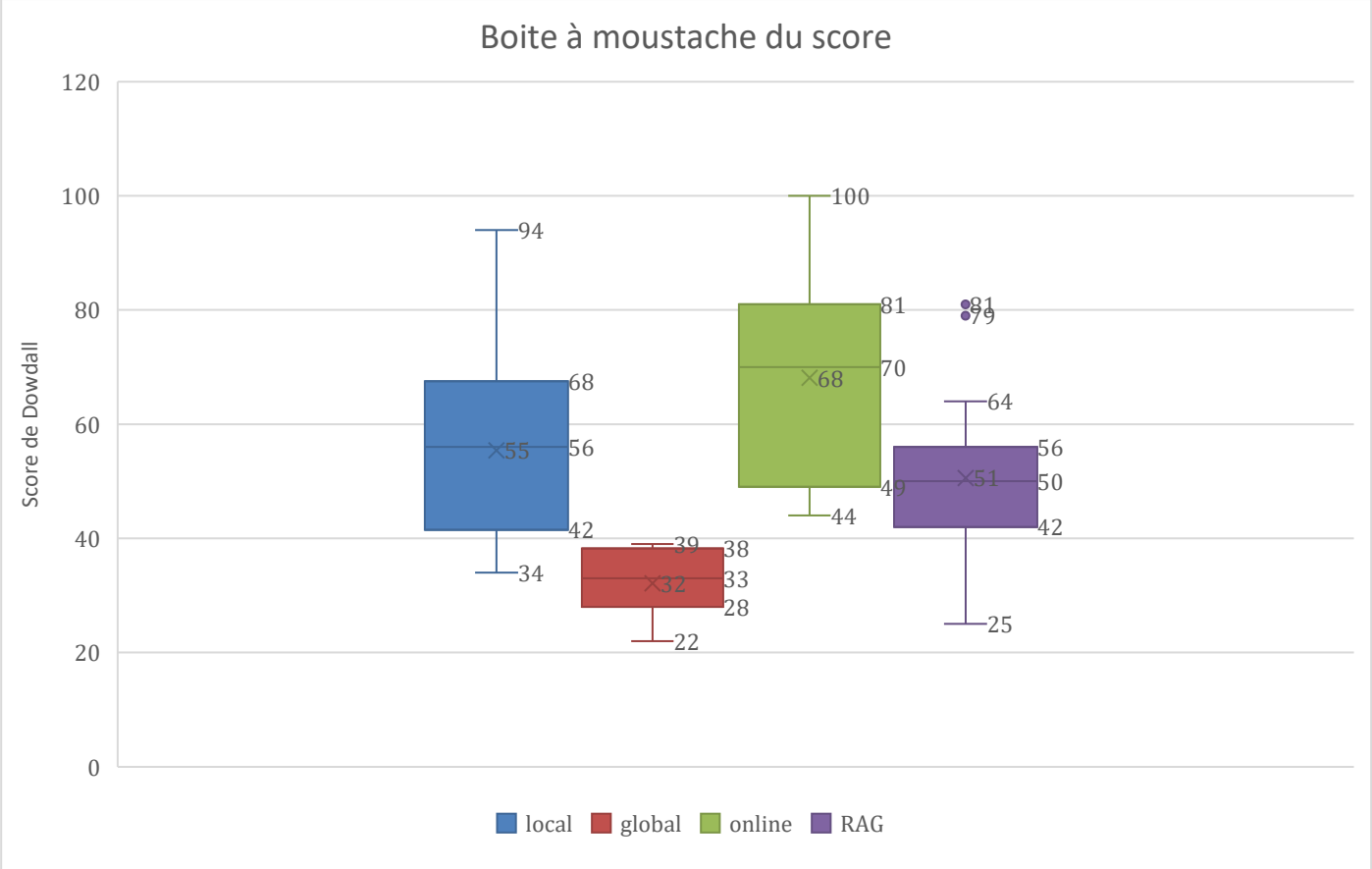


Fig. 9 – Scores de Dowdall pour les différents LLMs, juges humains

Les résultats obtenus révèlent une tendance nette (voir Fig. 9) : *ChatGPT Online* surpasse largement les autres méthodes, confirmant son efficacité telle qu'elle est perçue par les évaluateurs humains. En deuxième position, les approches *Local Graph-RAG* et *RAG classique* affichent des performances comparables et intermédiaires.

En revanche, la méthode *Global Graph-RAG* se distingue négativement, se classant systématiquement en dernière position. Ce résultat s'explique principalement par son incapacité récurrente à fournir des réponses, ou par des réponses très brèves et peu informatives.

Sur le plan de la variabilité des jugements (écart-type), seule la méthode *Global Graph-RAG* présente une grande stabilité, avec un écart-type très faible, autour de 6. Cela s'interprète comme le reflet de performances constamment faibles. À l'inverse, les autres méthodes — *Local Graph-RAG, RAG classique* et *ChatGPT Online* — présentent des écarts-types relativement similaires, compris entre 15 et 17, traduisant des performances plus hétérogènes selon les questions posées.

Enfin, compte tenu du nombre total d'échantillons (calculé comme le produit du nombre de questions par le nombre de réponses évaluées), les intervalles de confiance à 95 % sont suffisamment étroits pour permettre des conclusions robustes quant au classement général observé.

L'ensemble des données et des résultats est fourni en annexe de ce rapport.

LLM as a judge

Dans cette partie, nous nous inspirons de certaines analyses présentées dans le papier "Judging LLM as A Judge" [2]. L'idée est ici de regarder comment d'autres LLMs évaluent la qquestionsualité des réponses de nos 4 modèles. En cas de mise à l'échelle, l'évaluation par des humains deviendrait trop compliquée pour un nombre de question important et le recours à ce type de méthode serait alors quasiment obligatoire. A titre d'information, la comparaison sur notre faible corpus de question prenait déjà environ 30 à 40 min pour un humain déjà informé sur le sujet de Donald Trump.
Comme détaillé dans l'article, nous avons dû nous prémunir de certains biais que peuvent avoir les LLMs (mais aussi les humains) lors de l'évaluation des questions. En voici un rapide résumé, accompagnées des mesures que nous avons mises en place pour s'en protéger :

Biais de "self enhancement":
Un LLM aura tendance à surévaluer une réponse à une tâche si c'est lui-même qui l'a résolu. Nos GraphRAG et le modèle online sont basés sur GPT-4o, c'est pour cette raison que nous n'avons pas utilisé ChatGPT pour analyser nos réponses, mais trois autres LLMs : Mistral, DeepSeek et Gemini.

Biais de positionnement :
Les LLMs (et les humains également) peuvent avoir des évaluations biaisées dans le cas où les réponses des modèles se trouvent toujours dans le même ordre. Nous avons donc appliqué des permutations aléatoires sur l'ordre des propositions, que cela soit pour les LLMs, mais aussi pour l'évaluation humaine.

Biais de verbosité :
Ce biais n'a pas été traité, ce d'autant qu'il a semblé que DeepSeek ne l'applique pas systématiquement. Ce biais consiste à privilégier une réponse plus longue à une réponse plus courte, même si les deux réponses possèdent la même qualité d'informations à l'intérieur.

Biais de limitation mathématique :
Le dernier biais ne nous concerne pas. Dans l'article, ils évaluent parfois les réponses à des questions mathématiques, et les LLMs semblent avoir plus de difficultés à évaluer la qualité d'une réponse, alors même qu'ils sont capables de trouver la réponse au problème. Ce biais a été ignoré pour nous.
Une fois ces biais traités, nous avons pu réaliser les évaluations.

Résultats LLM as a judge

Dans cette partie, nous n'avons pas utilisé le score de Dowdall mais un score plus simple et également très répandu : le score de Borda.
Le scoring de Borda est une méthode d'agrégation de préférences qui attribue des points en fonction du classement des différentes options. Plus une option est bien classée, plus elle obtient de points.
Ci-dessous, les scores de Borda moyens et les écarts-types :

Moyenne	**Local**	**Global**	**Online**	**Rag**
Gemini	1,47	0,64	1,84	1,26
Deepseek	1,29	0,86	2,05	0,89
Mistral	1,35	0,50	2,21	1,26

Écart-type	**Local**	**Global**	**Online**	**Rag**
Gemini	0,80	0,96	1,21	0,99
Deepseek	0,92	0,95	1,22	0,66
Mistral	1,00	0,65	0,98	0,87

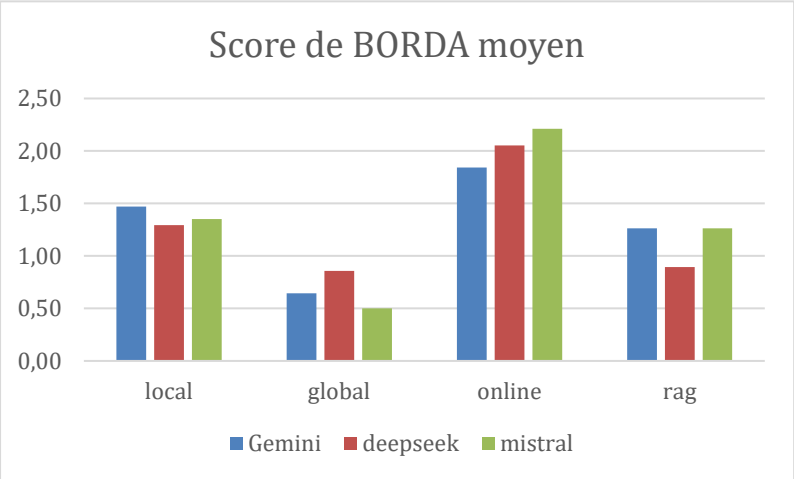


Fig. 10 – Scores de Bordas, LLMs as Judges
Fig. 11 – Tableau résumé des statistiques liées aux scores de Borda, LLMs as Judges

La première observation est que les LLMs semblent globalement confirmer l'évaluation humaine, en tout cas dans l'aspect général.

Globalement, Online GPT obtient les meilleurs scores de Borda moyen (entre 1,84 et 2,21), ce qui montre qu'il est en général préféré par les juges LLM. En revanche, ses écarts-types sont les plus élevés (autour de 1,2), ce qui traduit une performance moins stable selon les cas. Local GraphRAG arrive en seconde position, avec de bons scores moyens (entre 1,29 et 1,47) et des écarts-types modérés. À l'opposé, RAG présente les scores moyens les plus bas (entre 0,9 et 1.26), traduisant une efficacité perçue plus faible par les 3 juges LLM, mais avec des écarts-types faibles, signe de stabilité. Enfin, Global GraphRAG se situe entre Local et Global, mais sans se démarquer clairement, avec des scores moyens plutôt faibles.

Ci-dessous, le graphique qui présente le Score de Borda obtenu par les modèles juge Gemini, DeepSeek et Mistral sur 12 questions générales et cinq questions précises ayant pour but d'éprouver la recherche Local du GraphRAG. Ces graphiques comparent, pour chaque juge, les quatre approches de récupération d'information : Local, Global, Online et RAG. Voici les graphiques :

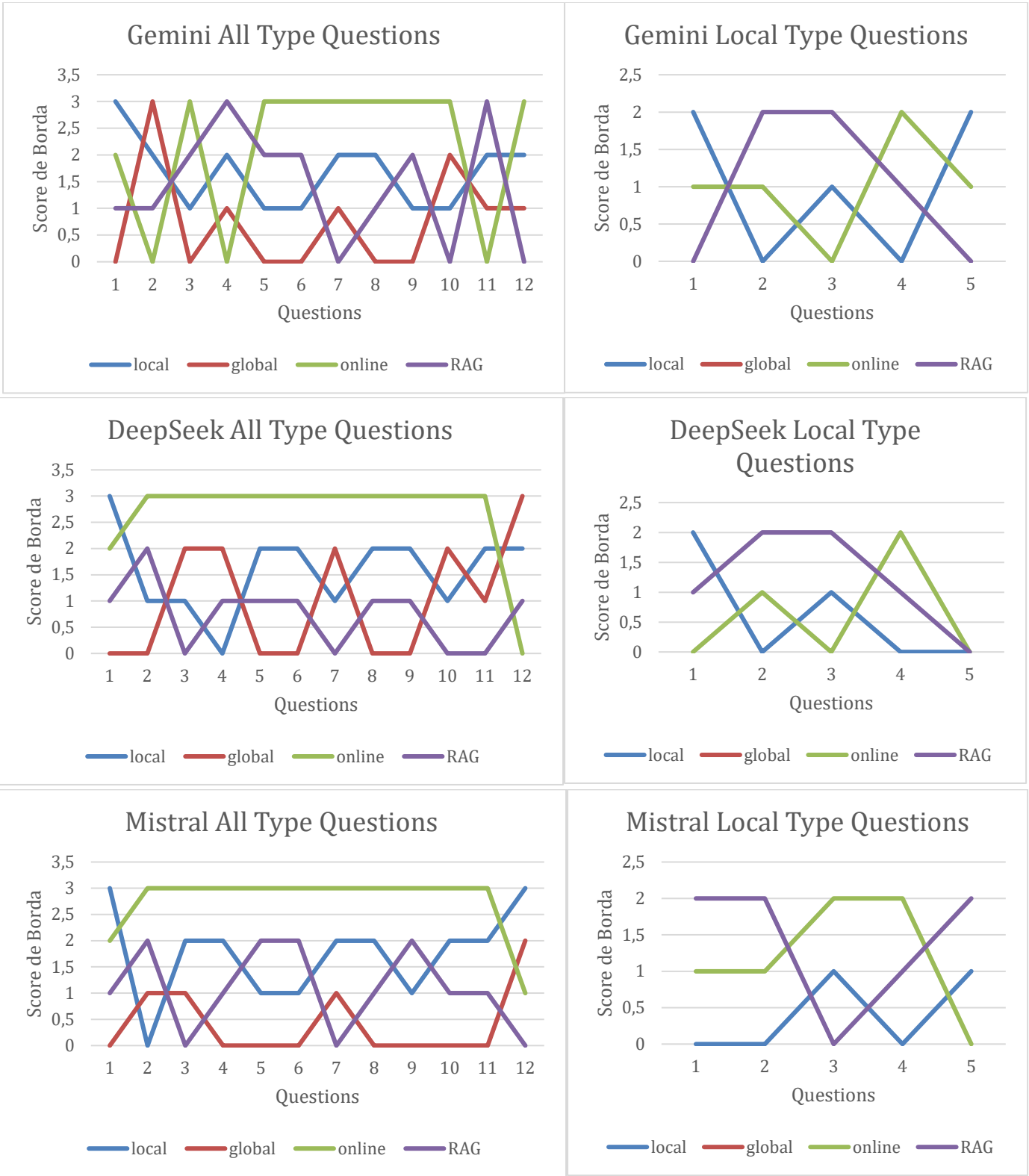


Fig. 12 – Graphiques de comparaison des scores de Borda des 4 technologies de spécialisation de LLM pour 3 juges différents

Sur l'ensemble des questions variées, une tendance claire se dégage : la technologie *online* obtient systématiquement les meilleurs scores de Borda, notamment avec DeepSeek et Mistral où elle domine très largement ses concurrents. Le GraphRAG local se positionne juste derrière, montrant une grande stabilité et une capacité d'adaptation satisfaisante. La technologie RAG arrive juste après étant très instable. À l'inverse, la technologie GraphRAG global peine à rivaliser, avec des scores de Borda trop faibles.
En revanche, lorsque l'on se concentre uniquement sur les questions très précises (Local Type Questions), conçues pour mettre en difficulté les approches globales et favoriser les méthodes locales, la hiérarchie évolue sensiblement. Ici, les technologies *local* et *RAG* prennent clairement l'avantage, avec des scores proches et souvent meilleurs que ceux des autres méthodes. Les performances d'*online*, si brillantes sur les questions variées, chutent considérablement, confirmant sa difficulté à traiter des requêtes très spécifiques et contextuelles. Enfin, la technologie GraphRAG global reste la moins performante, quelle que soit la nature des questions posées.

Remarque sur la victoire de Chat GPT Online :
GPT Online semble être la méthode d'actualisation la plus efficace. Cependant, au fur et à mesure des questions, nous avons remarqué un comportement étonnant. GPT Online est très performant car il a accès à Internet, et permet ainsi de s'actualiser seul, en allant chercher des informations sur des sites sans avoir besoin de mettre à jour ses poids. Cette option est très puissante, et probablement imbattable par une quelconque autre méthode si elle est utilisée parfaitement. Cependant, lorsque la question est tournée d'une certaine manière, Chat GPT ne va pas systématiquement aller chercher sur Internet une vérification de ce qu'il dit. Par exemple : à la question "Qui est Kamala Harris", le chatbot sait parfaitement qui est Kamala Harris. Il ne va pas aller chercher sur Internet une quelconque actualisation, et donc, ne pas indiquer qu'elle a perdu les dernières élections américaines. Il va plutôt la présenter comme la Vice-Présidente Américaine. On voit donc clairement que cette méthode n'est pas infaillible. En effet, dans certaines situations où des affirmations éridiques au moment de l'entrainement se retrouve finalement fausse après un travail d'investigation récent, Chat GPT Online pourrait donc propager une information totalement fausse au moment où l'utilisateur s'informe. Pour éviter cela, il faut bien vérifier si la réponse de Chat GPT s'est appuyée sur une source Internet. A l'inverse, des méthodes comme RAG et GraphRAG offrent un peu plus de garanti sur l'utilisation des données actualisation, puisqu'elles ont été ajoutées spécifiquement.

Méthodologie du projet

Le projet s'est articulé en différentes phases :

- Octobre 2024 : Définition des objectifs
- Novembre 2024 : Prise en main des outils RAG / GRAPHRAG
- Décembre 2024 : Définition du futur corpus d'étude
- Janvier 2025 : Scrapping des documents, nettoyage du corpus et analyse de qualité
- Février 2025 : Création des questions d'évaluation et spécialisation des LLMs sur le corpus
- Mars 2025 : Analyse des performances sur le benchmark et études complémentaires
- Avril 2025 : Finalisation du projet et rédaction du rapport

Pour la répartition des tâches, ils nous ont semblé important de répartir équitablement la charge de travail entre les 3 membres du groupe. Cependant, nous avons fait en sorte que chaque membre manipule chaque étape du processus ci-dessus afin que chacun d'entre nous garde une maitrise globale du projet. Il y a cependant eux certaines parties qui ont été plus profondément traitées par certaines d'entre nous (Théo : Scraping des documents, Erwan : étude de RAG, Kenzy : Mise en place d'une partie de l'évaluation).

Conclusion

En définitive, cette étude aura fait ressortir quelques résultats importants. Premièrement, la version Online de ChatGPT avec accès à Internet reste probablement la méthode actuelle la plus efficace pour utiliser des LLMs tout en manipulant des informations inconnues à l'entrainement. Cependant, les méthodes de GraphRAG et RAG possèdent toutes 2 des performances correctes. On notera tout de même qu'il était très compliqué de proposer ici une méthode d'évaluation absolue des différents LLMs, et que tous nos résultats sont comparatifs (entre les solutions) et non pas absolues.
Malgré les meilleures évaluations de GPT Online, cette méthode a des limites. D'abord, nous avons pris ici l'exemple de données publiques, accessibles via des articles de journaux. Cependant, dans le cas de données internes à des entreprises par exemple, la solution chat GPT Online devient non pertinente et le recours à des méthodes alternatives comme GraphRAG ou RAG devient intéressant. Également, l'évaluation a permis de porter l'attention sur un point important : le contrôle du LLM. En effet, si on ne vérifie pas que Chat GPT va effectivement bien chercher l'information sur Internet, il pourrait passer à côté d'informations majeures, et nous induire en erreur dans notre raisonnement.

Bibliographie

[1] Han, H., Wang, Y., Shomer, H., Guo, K., Ding, J., Lei, Y., ...Tang, J. (2024). Retrieval-Augmented Generation with Graphs (GraphRAG). arXiv, 2501.00309. Retrieved from https://arxiv.org/abs/2501.00309v2

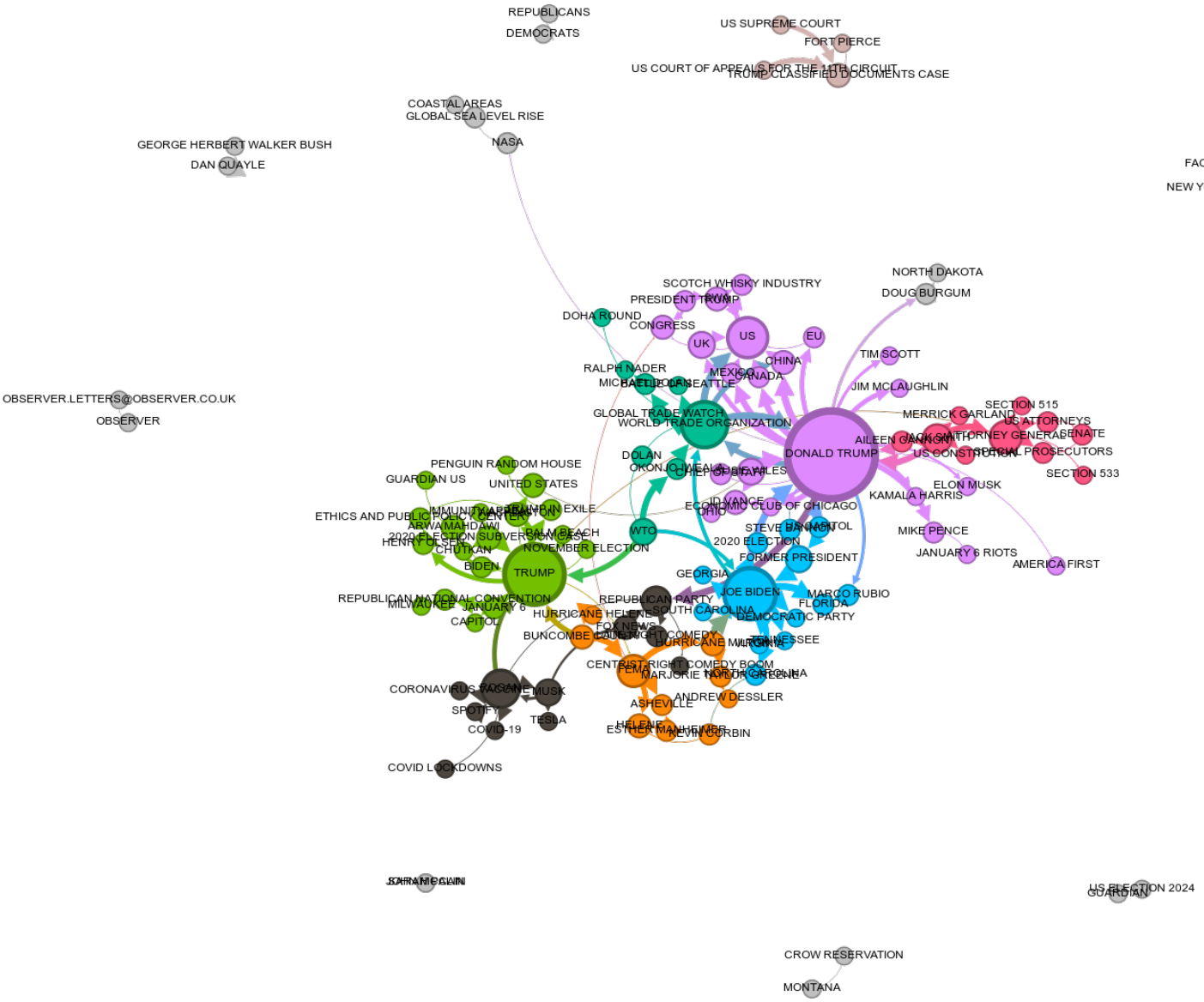
[2] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ...Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv, 2306.05685. Retrieved from https://arxiv.org/abs/2306.05685v4

[3] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ...Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv, 2005.11401. Retrieved from https://arxiv.org/abs/2005.11401v4

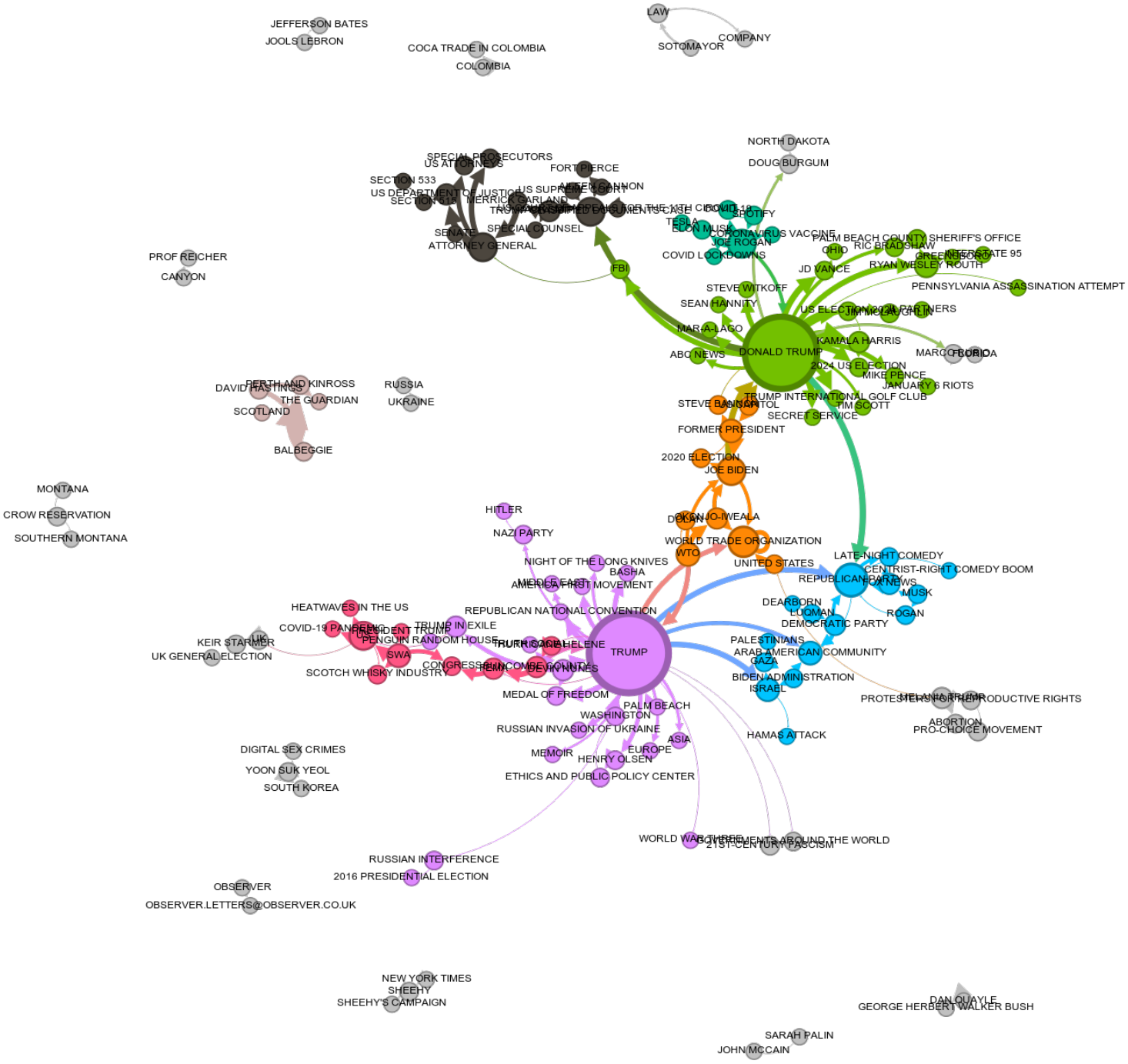
Annexes

Voici, pour illustrer la variance de la technologie de GraphRAG, 2 graphes de connaissance générés sur le même corpus **Trump24-25** avec exactement les mêmes paramètres :

- *Séparation des .txt*
- *Max_cluster_size = 10*
- *Chunk_size = 800*
- *Claim_extraction=true*



119 nœuds, 146 relations



149 nœuds, 164 relations

Score de Dowdall						Correspondance Numéro de réponse			
Question	Numéro de réponse					Technologie utilisée			
	1	2	3	4		local	global	online	RAG
Is Musk supporting Democrates or Republicans?	56	58	58	22		2	4	3	1
Who is the current president of the USA, and the next one?	53	85	34	33		3	4	2	1
Who is Kamala Harris?	38	64	63	25		2	1	3	4
When will be the next USA's elections?	52	31	45	80		3	2	4	1
Give me the list of the American president on which there has been an assassination attempt ?	25	70	28	72		2	1	4	3
What are the links between Musk and Trump?	38	39	79	46		1	2	4	3
Give me examples of territories that trump want to add to the USA	72	42	49	39		1	4	3	2
What is the vision of Trump and Kamala Harris on Isreal and the conflict in Gaza?	46	69	65	22		3	4	2	1
Analyse the link between medias, fake news and Donald Trump.	41	56	39	70		2	3	4	1
Summarize the debate arround immigration during this political Campaign	78	50	41	36		3	4	1	2
What is the current status on Trump and justice?	32	29	100	40		4	2	3	1
What are the key points of the political campaign?	45	47	33	73		4	3	1	2
What is the "lock up her" affair? Who is her and who pronounced it?	48	75	46			3		2	1
Who is suing the Texas' department of criminal justice. Why? And in what context.	56	81	42			3		1	2
Why were advisers concerned about the debate between Trump and Haris but still are willing to take the risk?	94	44	42			1		2	3
Why Joe Biden thought about pardoning members of congress? To what extent would it be unprecedented? What are the risks?	81	42	56			2		1	3
What is the opinion of Trump on the Federal Emergency Management Agency? /!\ erreur	62	48	64			1		2	3
Give all the opportunity and risks over Trump debate with Kamala Harris, whether from Harris' perspective or Trump's. Then, expose the outcomes.	81	63	35				3	1	2
What can you say between Trump and fascism?	29	56	94				1	3	2

Score de Dowdall				
Question	**local**	**global**	**online**	**RAG**
Is Musk supporting Democrates or Republicans?	58	22	58	56
Who is the current president of the USA, and the next one?	34	33	85	53
Who is Kamala Harris?	64	38	63	25
When will be the next USA's elections?	45	31	80	52
Give me the list of the American president on which there has been an assassination attempt ?	70	25	72	28
What are the links between Musk and Trump?	38	39	46	79
Give me examples of territories that trump want to add to the USA	72	39	49	42
What is the vision of Trump and Kamala Harris on Isreal and the conflict in Gaza?	65	22	69	46
Analyse the link between medias, fake news and Donald Trump.	56	39	70	41
Summarize the debate arround immigration during this political Campaign	41	36	78	50
What is the current status on Trump and justice?	40	29	100	32
What are the key points of the political campaign?	73	33	45	47
What is the "lock up her" affair? Who is her and who pronounced it?	46	#N/A	75	48
Who is suing the Texas' department of criminal justice. Why? And in what context.	42	#N/A	56	81
Why were advisers concerned about the debate between Trump and Haris but still are willing to take the risk?	94	#N/A	44	42
Why Joe Biden thought about pardoning members of congress? To what extent would it be unprecedented? What are the risks?	42	#N/A	81	56
What is the opinion of Trump on the Federal Emergency Management Agency? /!\ erreur	62	#N/A	48	64
Give all the opportunity and risks over Trump debate with Kamala Harris, whether from Harris' perspective or Trump's. Then, expose the outcomes.	#N/A	35	81	63
What can you say between Trump and fascism?	#N/A	29	94	56
MOYENNE	55,4117647	32,1428571	68,1052632	50,5789474
Ecart-type	16,3288809	6,30425172	17,0941664	14,8185911
Intervalle de confiance x12 (12 réponses)	2,24072949	0,9532943	2,21885333	1,92347959
Min	53,1710352	31,1895628	65,8864098	48,6554678
Max	57,6524942	33,0961514	70,3241165	52,502427