



Interprétabilité Mécaniste pour l'Extraction d'Information dans les systèmes de RAG

Soutenance de stage de Master 2

Marine DELVALLEZ

Master 2 Informatique GPEx Minerve

Jury : Thi-Bich-Hanh DAO
Anaïs HALFTERMEYER
Marie HÉNAULT

Tutrice de stage : Anaïs HALFTERMEYER

2 Septembre 2026

Sommaire

- 1 État de l'art
 - Retrieval Augmented Generation - RAG
 - Explicabilité pour le RAG
- 2 Objectifs de ce stage
- 3 Préliminaires
- 4 Expérimentations
 - Construction des perturbations
 - Évaluation d'une perturbation
 - Lecture des résultats de l'activation patching
 - Résultats
- 5 Conclusion et perspectives

Sommaire

1 État de l'art

- Retrieval Augmented Generation - RAG
- Explicabilité pour le RAG

2 Objectifs de ce stage

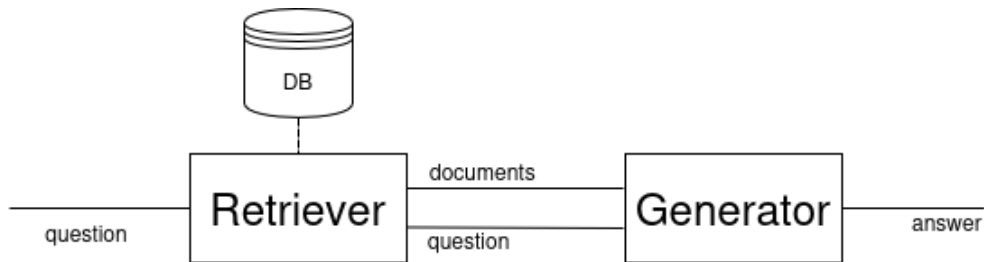
3 Préliminaires

4 Expérimentations

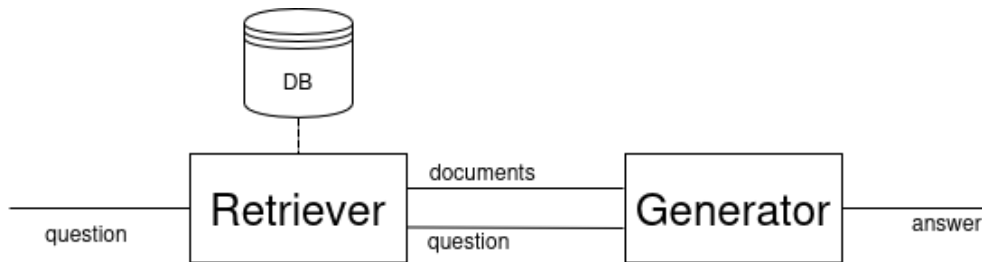
- Construction des perturbations
- Évaluation d'une perturbation
- Lecture des résultats de l'activation patching
- Résultats

5 Conclusion et perspectives

Retrieval Augmented Generation - RAG [LPP⁺20]



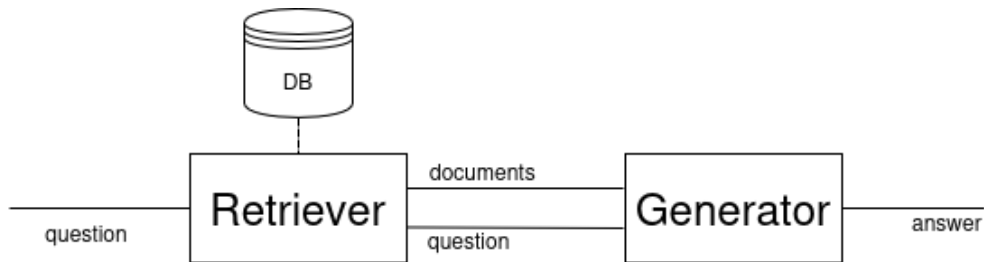
Retrieval Augmented Generation - RAG [LPP⁺20]



Tâches du RAG

- Question Answering
- Fact Checking
- Dialogue

Retrieval Augmented Generation - RAG [LPP⁺20]



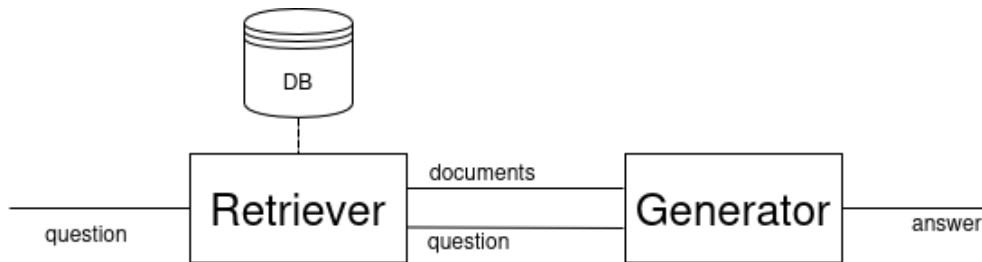
Tâches du RAG

- Question Answering
- Fact Checking
- Dialogue

Sous-tâches

- Extraction de document/
d'information
- Synthèse
- Génération de texte

Retrieval Augmented Generation - RAG [LPP⁺20]



Tâches du RAG

- Question Answering
- Fact Checking
- Dialogue

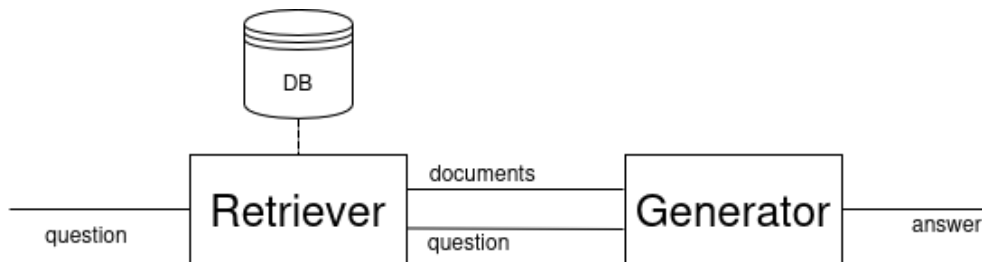
Sous-tâches

- **Extraction de document/**
d'information
- Synthèse
- Génération de texte

Explicabilité de l'IA et RAG

On a besoin d'expliquer les systèmes RAG !

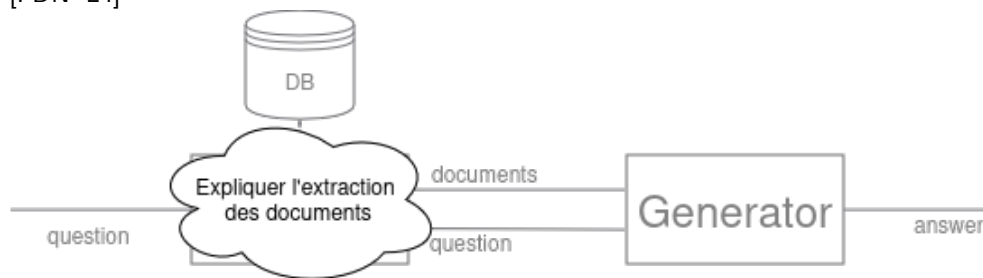
[FDN⁺24]



Explicabilité de l'IA et RAG

On a besoin d'expliquer les systèmes RAG !

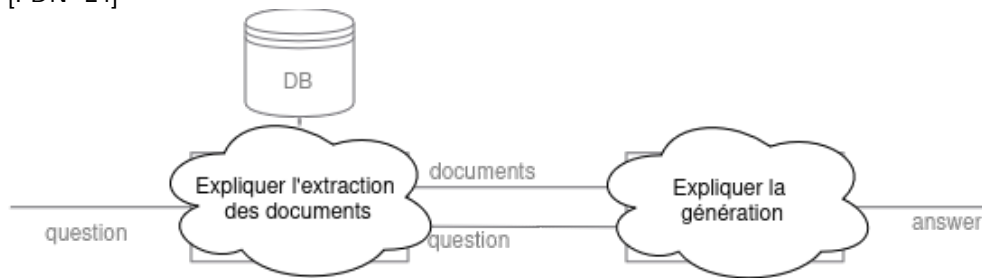
[FDN⁺24]



Explicabilité de l'IA et RAG

On a besoin d'expliquer les systèmes RAG !

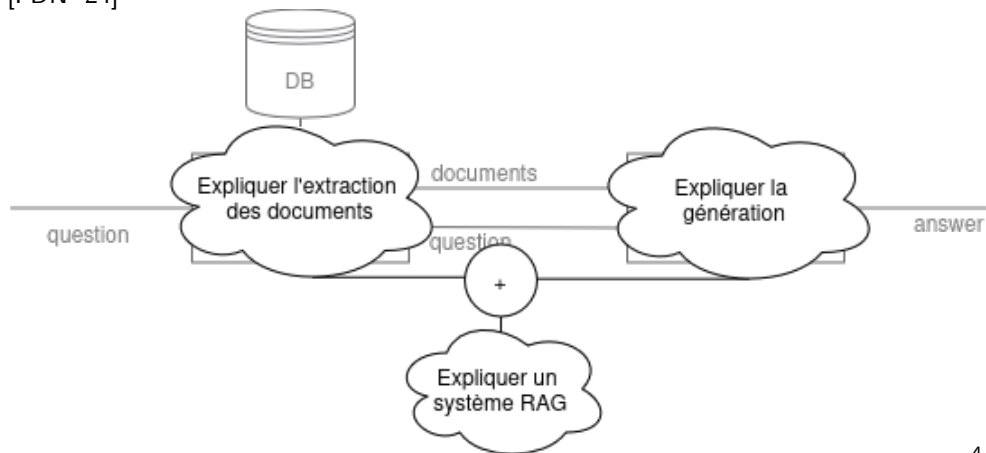
[FDN⁺24]



Explicabilité de l'IA et RAG

On a besoin d'expliquer les systèmes RAG !

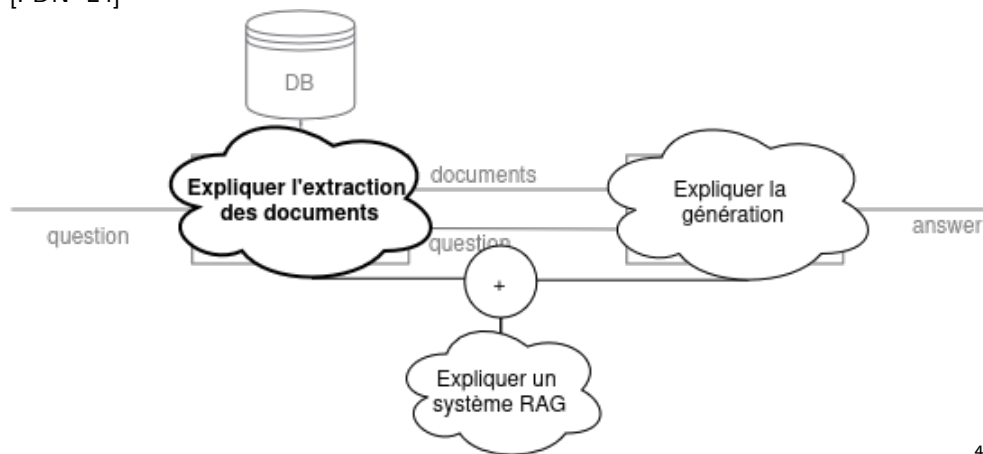
[FDN⁺24]



Explicabilité de l'IA et RAG

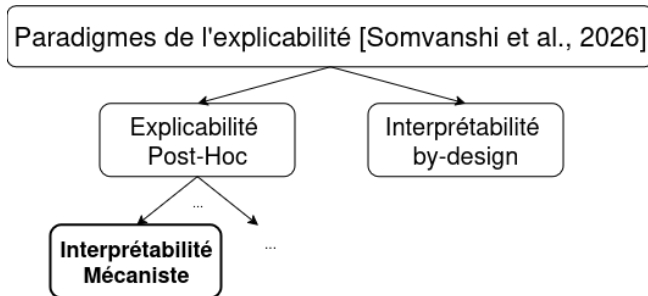
On a besoin d'expliquer les systèmes RAG !

[FDN⁺24]



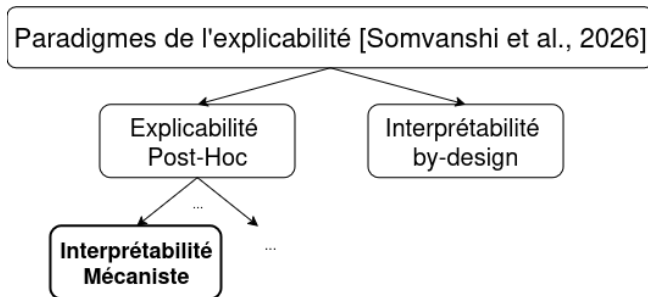
Explicabilité de l'IA et Interprétabilité Mécaniste

[SIR⁺26]



Explicabilité de l'IA et Interprétabilité Mécaniste

[SIR⁺26]



Interprétabilité Mécaniste

Explorer les **activations des composants** du modèle pour **identifier ceux qui provoquent ou influencent** un comportement cible

Interprétabilité Mécaniste [SIR⁺26]

Interprétabilité Mécaniste

Explorer les **activations des composants** du modèle pour **identifier ceux qui provoquent ou influencent** un comportement cible

Plusieurs approches :

- Manual Circuit Tracing
- Representation Analysis
- Toy Models and Synthetic Tasks
- Intervention-based Techniques

Interprétabilité Mécaniste [SIR⁺26]

Interprétabilité Mécaniste

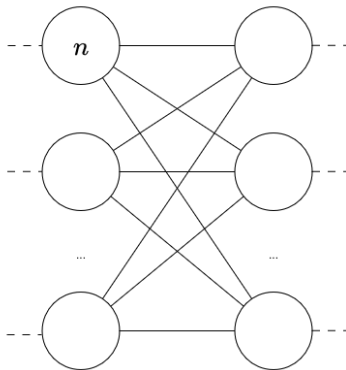
Explorer les **activations des composants** du modèle pour **identifier ceux qui provoquent ou influencent** un comportement cible

Plusieurs approches :

- Manual Circuit Tracing
- Representation Analysis
- Toy Models and Synthetic Tasks
- **Intervention-based Techniques**
 - ▶ Activation Patching
 - ▶ ...

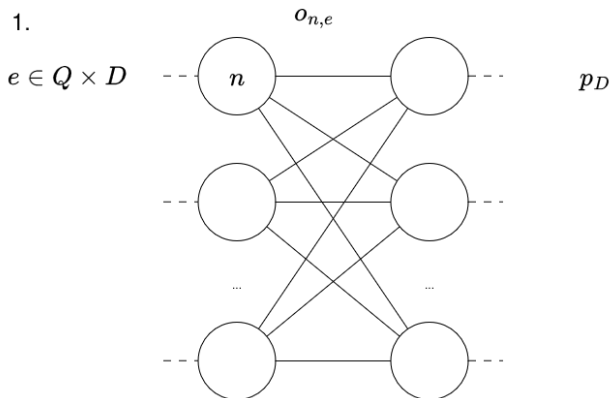
Activation Patching pour l'IR [CME24] : intuition

0.



Activation Patching pour l'IR [CME24] : intuition

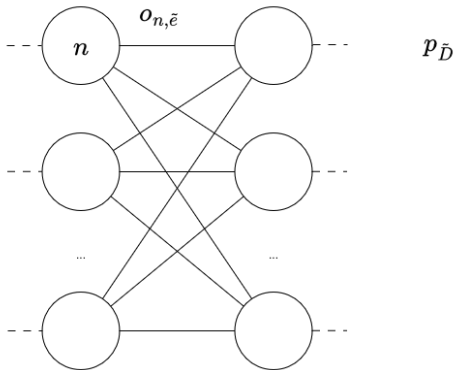
1.



Activation Patching pour l'IR [CME24] : intuition

2.

$$\tilde{e} \in Q \times \tilde{D}$$

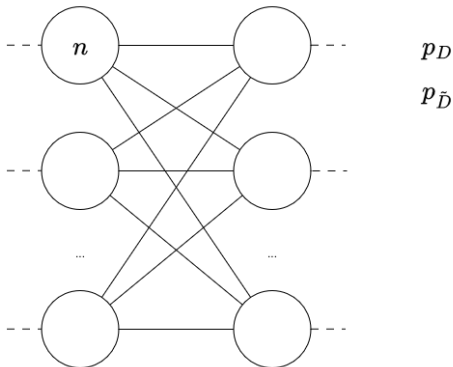


Activation Patching pour l'IR [CME24] : intuition

3.

$$e \in Q \times D$$

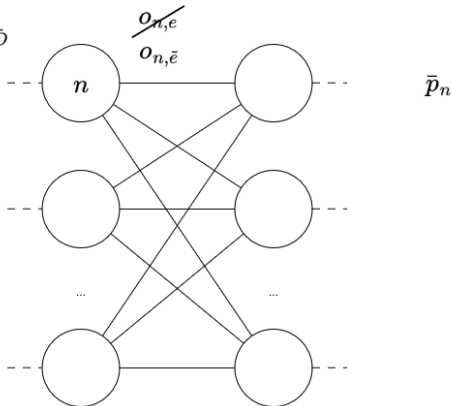
$$\tilde{e} \in Q \times \tilde{D}$$



Activation Patching pour l'IR [CME24] : intuition

4. if $p_D < p_{\bar{D}}$

$e \in Q \times D$



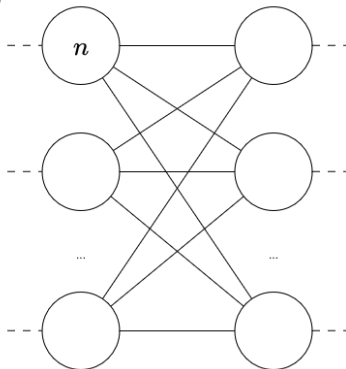
Activation Patching pour l'IR [CME24] : intuition

5. if $p_D < p_{\tilde{D}}$

$e \in Q \times D$

$\tilde{e} \in Q \times \tilde{D}$

$e \in Q \times D$



p_D

$p_{\tilde{D}}$

$\bar{p}_n$

$$P_n = \frac{\bar{p}_n - p_D}{p_{\tilde{D}} - p_D}$$

Sommaire

- 1 État de l'art
 - Retrieval Augmented Generation - RAG
 - Explicabilité pour le RAG
- 2 Objectifs de ce stage
- 3 Préliminaires
- 4 Expérimentations
 - Construction des perturbations
 - Évaluation d'une perturbation
 - Lecture des résultats de l'activation patching
 - Résultats
- 5 Conclusion et perspectives

Hypothèses et Objectifs

Langue de spécialité [Cab98]

Ensemble de sous-codes linguistiques spécifiques à un domaine de spécialité (lexique, syntaxe,...)

Hypothèses

- H1 : Le vocabulaire de spécialité (lexique) repose sur certains composants du modèle
- H2 : L'activation patching permet de les localiser

Objectifs

- Appliquer l'activation patching à la localisation du vocabulaire de spécialité de la *défense*
- Proposer une méthodologie d'utilisation de MechIR

Sommaire

- 1 État de l'art
 - Retrieval Augmented Generation - RAG
 - Explicabilité pour le RAG
- 2 Objectifs de ce stage
- 3 Préliminaires
- 4 Expérimentations
 - Construction des perturbations
 - Évaluation d'une perturbation
 - Lecture des résultats de l'activation patching
 - Résultats
- 5 Conclusion et perspectives

Cas d'application

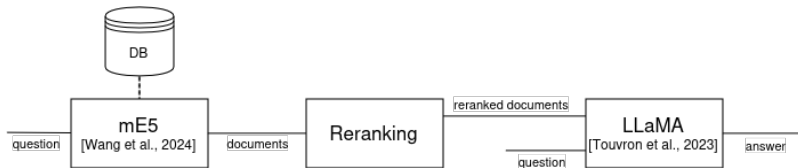


Figure – Version simplifiée de RAG4HistoricalNewspapers [TGGD24]

mE5-small [WYH⁺24]

- $\approx 17M$ de paramètres
- encodeur-only
- bi-encodeur

Cas d'application

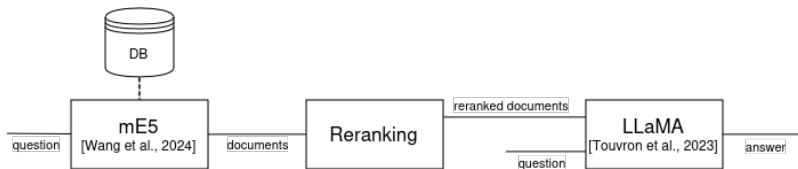


Figure – Version simplifiée de RAG4HistoricalNewspapers [TGGD24]

mE5-small [WYH⁺24]

- $\approx 17M$ de paramètres
- encodeur-only
- bi-encodeur

MechIR [PCEM25]

- Alternative à TransformerLens [NB22] pour les encodeur-only
- Implante des outils pour l'Activation Patching

Dataset

Contenu thématique

- Extraits de documents issus des données du challenge RAG@EvalLLM
- Question sur les *limitations techniques des systèmes de drones*
- Trois documents
 - ▶ Limitations techniques du drone
 - ▶ Matériel de guerre - Code de la défense
 - ▶ Fausses informations pendant les JO/JP

Structure

- 1 question
- 28 extraits
- 50% de documents pertinents

Sommaire

- 1 État de l'art
 - Retrieval Augmented Generation - RAG
 - Explicabilité pour le RAG
- 2 Objectifs de ce stage
- 3 Préliminaires
- 4 Expérimentations**
 - Construction des perturbations
 - Évaluation d'une perturbation
 - Lecture des résultats de l'activation patching
 - Résultats
- 5 Conclusion et perspectives

Construire des perturbations

Qu'est-ce qu'une bonne perturbation ?

- Modification minimale du document [CME24]
- Modification significative de la sortie [CME24]
- Grande couverture du dataset

Perturbations proposées par [PCEM25]

- `prepend` : ajouter un mot au début du document
- `append` : ajouter un mot à la fin du document
- `replace` : remplacer un mot par un autre

Construire des perturbations

Qu'est-ce qu'une bonne perturbation ?

- Modification minimale du document [CME24]
- Modification significative de la sortie [CME24]
- Grande couverture du dataset

Perturbations proposées par [PCEM25]

`prepend` : ajouter un mot au début du document

`append` : ajouter un mot à la fin du document

`replace` : remplacer un mot par un autre

Exemple : remplacer *aérien* par *poétique*

Certaines limitations sont plus particulièrement liées aux caractéristiques du vecteur aérien

Construire des perturbations

Qu'est-ce qu'une bonne perturbation ?

- Modification minimale du document [CME24]
- Modification significative de la sortie [CME24]
- Grande couverture du dataset

Perturbations proposées par [PCEM25]

`prepend` : ajouter un mot au début du document

`append` : ajouter un mot à la fin du document

`replace` : remplacer un mot par un autre

Exemple : remplacer *aérien* par *poétique*

*Certaines limitations sont plus particulièrement liées aux caractéristiques du vecteur **poétique***

Choisir le mot à remplacer

Critères de choix des mots

- Fréquence dans le dataset
- TF-IDF du mot dans le dataset
- Présence du mot dans la question
- Porteur d'un sens marqué
- **Appartenance à au moins un autre sous-code**

Choisir le mot à remplacer

Critères de choix des mots

- Fréquence dans le dataset
- TF-IDF du mot dans le dataset
- Présence du mot dans la question
- Porteur d'un sens marqué
- **Appartenance à au moins un autre sous-code**

Mot	Fréquence	TF-IDF
matériel	16	0.063
système	11	0.054
limitation	6	0.050
drone	10	0.044
défense	10	0.044
prendre	6	0.043
compte	5	0.040
action	8	0.040
article	9	0.040
ministre	8	0.039
aérien	7	0.037
acteur	6	0.034
guerre	8	0.034
informationnel	5	0.031
information	6	0.030

Choisir le mot remplaçant

Étant donné le mot cible (à remplacer)

Niv	Mot remplaçant	Phénomène cible	Exemple
1	Synonyme dans la langue de spécialité	Localisation du mot remplacé	aérien → de vol
2	Synonyme dans le langage courant	Localisation du sens	aérien → poétique

Choisir le mot remplaçant

Étant donné le mot cible (à remplacer)

Niv	Mot remplaçant	Phénomène cible	Exemple
1	Synonyme dans la langue de spécialité	Localisation du mot remplacé	aérien → de vol
2	Synonyme dans le langage courant	Localisation du sens	aérien → poétique
3	Mot sémantiquement incompatible	Localisation du sens (sans réparation)	aérien → épicurien
4	Groupe de mots <i>absurde</i> sémantiquement incompatibles	Localisation du sens (avec bruit)	aérien → acide tracté

Perturbation de type `full_replace`

Perturbation `full_replace`

Appliquer une liste de perturbations `replace` **simultanément**

Exemple :

- remplacer *aérien* par *poétique*
- remplacer *limitation* par *bornage*

Certaines limitations sont plus particulièrement liées aux caractéristiques du vecteur aérien

Buts

- Augmenter la couverture du dataset
- Effet d'ensemble

Perturbation de type `full_replace`

Perturbation `full_replace`

Appliquer une liste de perturbations `replace` **simultanément**

Exemple :

- remplacer *aérien* par *poétique*
- remplacer *limitation* par *bornage*

*Certaines **bornages** sont plus particulièrement liées aux caractéristiques du vecteur **poétique***

Buts

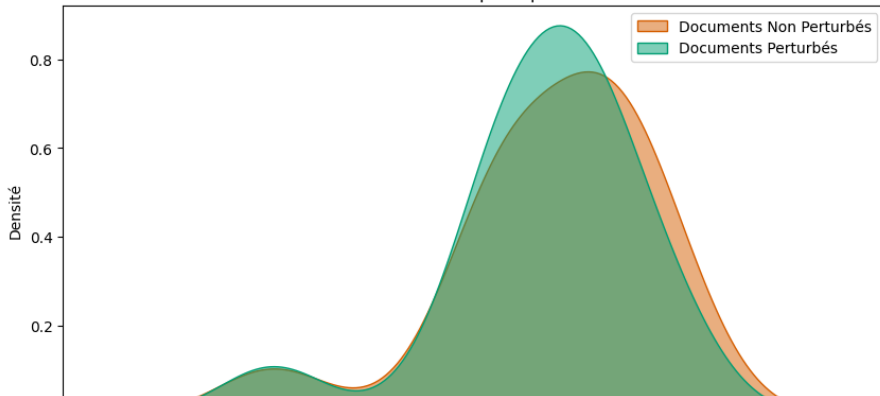
- Augmenter la couverture du dataset
- Effet d'ensemble

Distribution des scores de pertinence des documents

Objectif : Distributions de scores différentes pour les documents et documents perturbés [PCEM25]

Distribution des Scores de pertinence des documents

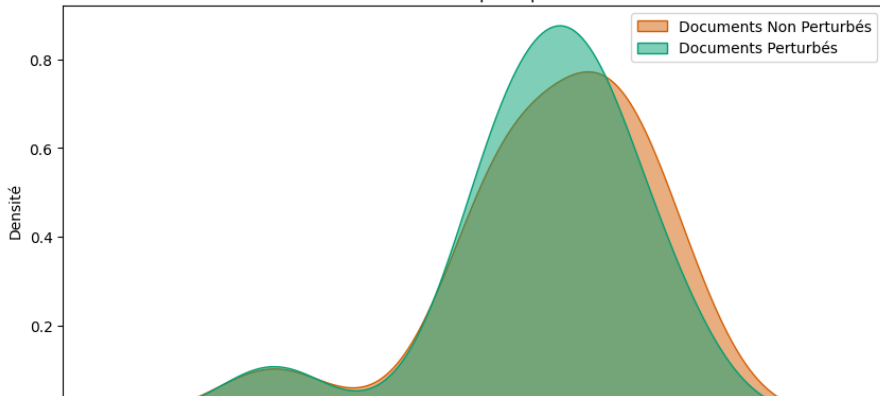
aérien -> poétique



Distribution des scores de pertinence des documents

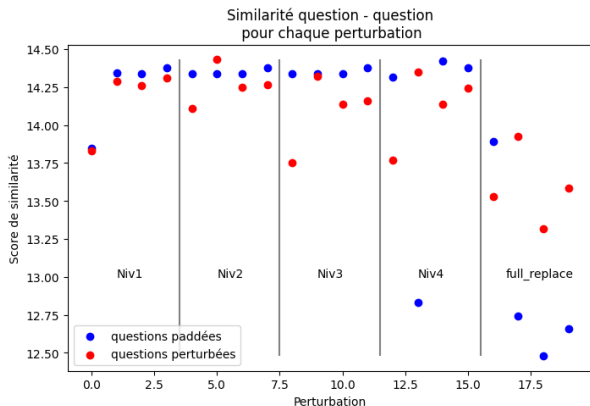
Objectif : Distributions de scores différentes pour les documents et documents perturbés [PCEM25]

Distribution des Scores de pertinence des documents
aérien -> poétique



En observant l'impact de la perturbation sur la question

Objectif : Estimer l'impact d'une perturbation sur la question dans l'espace de représentation.



Quelles sont les trois principales catégories de limitations techniques affectant les systèmes de drones, et comment chacune influence-t-elle l'efficacité opérationnelle de ces vecteurs ?

Mots ciblés

● drone

● limitation

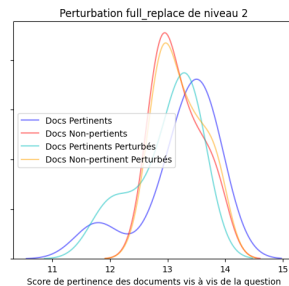
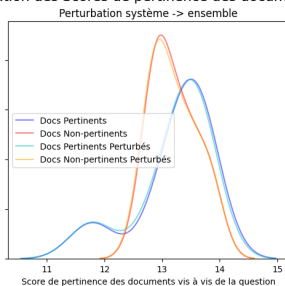
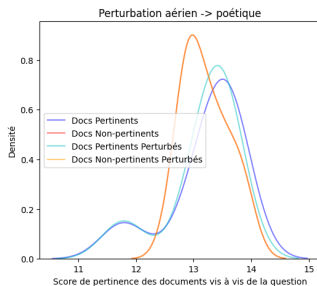
● système

● technique

En séparant les documents pertinents et non-pertinents

Objectif : Identifier les documents perturbés et leur distribution en fonction de leur pertinence.

Distribution des Scores de pertinence des documents



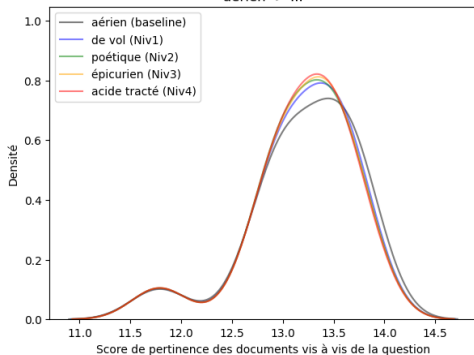
⇒ Répartition des mots ciblés biaisée

Confronter les perturbations entre elles

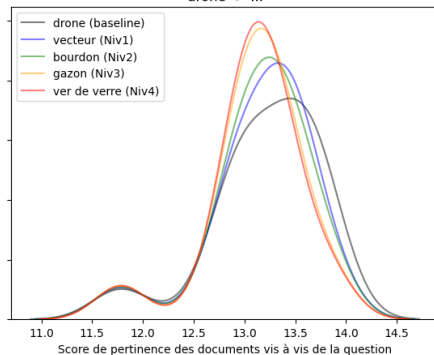
Objectif : Comparer l'impact des perturbations entre elles. Confirmer une progressivité. Choisir parmi plusieurs perturbations.

Distribution des Scores de pertinence des documents par niveau de perturbation

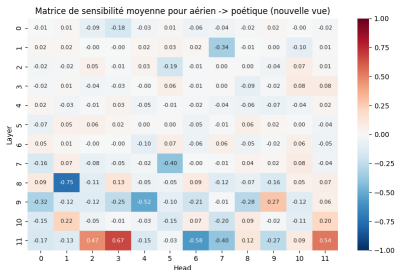
aérien -> ...



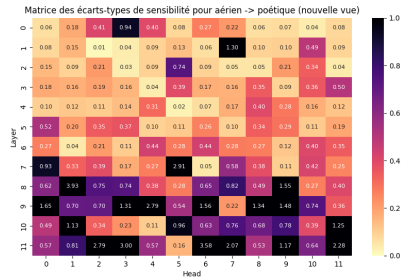
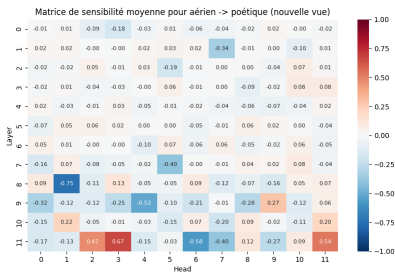
drone -> ...



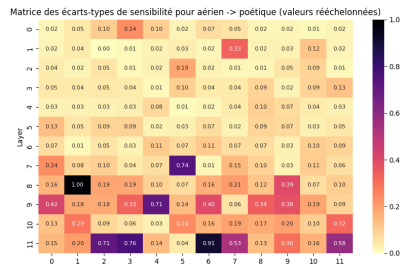
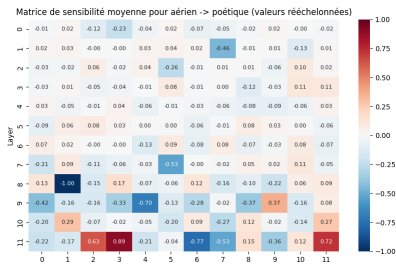
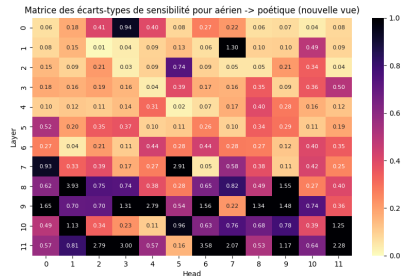
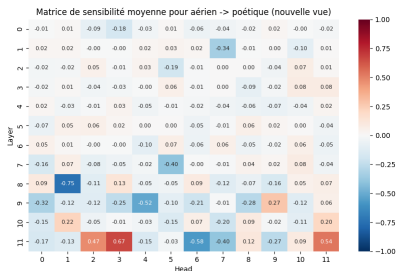
Matrice de sensibilité



Matrice de sensibilité



Matrice de sensibilité



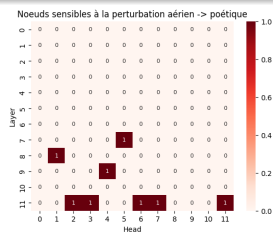
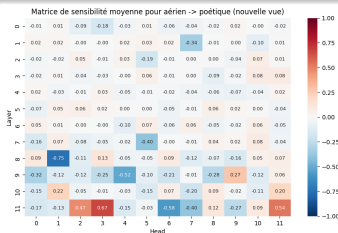
Identification des noeuds sensibles

Qu'est-ce qu'un noeud sensible ?

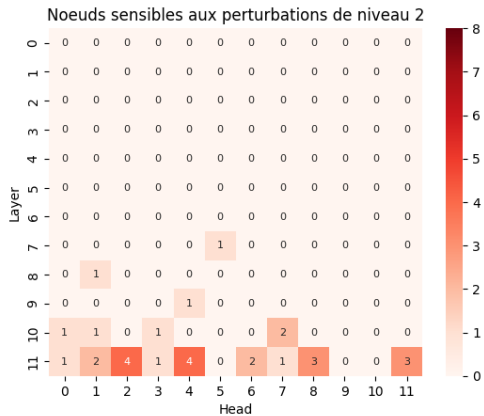
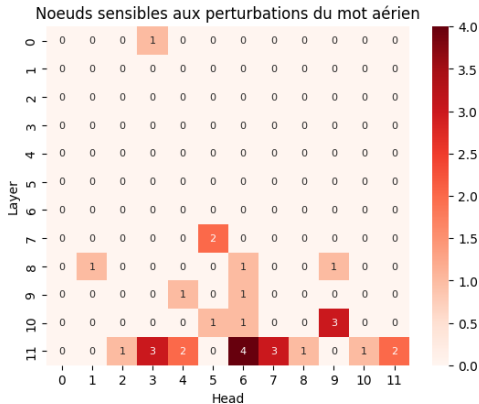
- Sensibilité au patching forte
- ... plus forte que les autres noeuds
- ... pour certains documents

Critère de sensibilité

$$((|mean| > 0.1) \wedge (|resc_mean| > 0.5)) \vee ((std > 0.1) \wedge (resc_std > 0.5))$$



Matrice des noeuds sensibles



Résultats

Quelques noeuds identifiés

Mot / Niveau	Noeuds identifiés
article	(11,4)
aérien	(11,6), (11,3), (10,9)
drone	(11,4)
défense	(11,2)
matériel	(11,4), (11,9)

système	(11,1)
Niveau 3	(11,3)
Niveau 4	(1,7)
Full replace	(11,1), (11,5), (11,6), (11,3)

Résultats

Quelques noeuds identifiés

Mot / Niveau	Noeuds identifiés
article	(11,4)
aérien	(11,6), (11,3), (10,9)
drone	(11,4)
défense	(11,2)
matériel	(11,4), (11,9)

système	(11,1)
Niveau 3	(11,3)
Niveau 4	(1,7)
Full replace	(11,1), (11,5), (11,6), (11,3)

La méthodologie actuelle ...

- ... met en évidence des noeuds sensibles *pour une perturbation*
- ... met en évidence des noeuds sensibles *pour un mot perturbé* (parfois)
- ... ne met pas en évidence de tendance à *l'échelle du vocabulaire*

Analyse

H1 : Le vocabulaire de spécialité (lexique) repose sur certains composants du modèle

- On peut identifier une spécialisation de certains composants à l'échelle du mot
- Sans savoir si elle relève du vocabulaire de spécialité, du mot ou du bruit

⇒ On ne peut pas conclure pour l'instant

H2 : L'activation patching permet de les localiser

- L'activation patching permet de localiser les composants sensibles d'un modèle

⇒ On ne peut pas conclure

Sommaire

- 1 État de l'art
 - Retrieval Augmented Generation - RAG
 - Explicabilité pour le RAG
- 2 Objectifs de ce stage
- 3 Préliminaires
- 4 Expérimentations
 - Construction des perturbations
 - Évaluation d'une perturbation
 - Lecture des résultats de l'activation patching
 - Résultats
- 5 Conclusion et perspectives

À retenir

Bonnes pratiques

- Perturbations couvrantes
- Perturbations pertinentes (cas de *article*)
- Vigilance et prise en compte des biais
- Vigilance pour les noeuds de la dernière couche

Résultats

- L'activation patching permet de localiser les composants sensibles au sens de certains mots

Perspectives d'amélioration

- Fonder et outiller l'échelle de niveau des perturbations [GC79]

Perspectives d'amélioration

- Fonder et outiller l'échelle de niveau des perturbations [GC79]
- Méthodologie pour mettre en perspective différents datasets

Perspectives d'amélioration

- Fonder et outiller l'échelle de niveau des perturbations [GC79]
- Méthodologie pour mettre en perspective différents datasets
- Diversification des perturbations (prepend , append , ...)

Perspectives d'amélioration

- Fonder et outiller l'échelle de niveau des perturbations [GC79]
- Méthodologie pour mettre en perspective différents datasets
- Diversification des perturbations (prepend , append , ...)
- Étendre à la langue de spécialité

Intégration dans l'XRAG

- Amélioration de la méthodologie

Intégration dans l'XRAG

- Amélioration de la méthodologie
- Transformer les noeuds identifiés en cartographie

Intégration dans l'XRAG

- Amélioration de la méthodologie
- Transformer les noeuds identifiés en cartographie
- Protocole d'extraction d'explication

Intégration dans l'XRAG

- Amélioration de la méthodologie
- Transformer les noeuds identifiés en cartographie
- Protocole d'extraction d'explication
- Rendre les explications accessibles à l'utilisateur non-informaticien

Intégration dans l'XRAG

- Amélioration de la méthodologie
- Transformer les noeuds identifiés en cartographie
- Protocole d'extraction d'explication
- Rendre les explications accessibles à l'utilisateur non-informaticien
- Protocole d'extraction d'améliorations

Conclusion sur les travaux du stage

- Défrichage bibliographique
 - ▶ RAG et RA-LLM
 - ▶ Explicabilité pour le RAG et le RA-LLM
 - ▶ Interprétabilité mécaniste
 - ▶ Activation Patching
- Challenge RAG@EvalLLM
 - ▶ Reprise d'un système RAG
 - ▶ Structuration itérative d'un dataset à la main
- Proposition d'une méthode d'utilisation de MechIR
- Proposition d'une méthode d'interprétation de l'activation patching

Références I



Maria Teresa Cabré.

La terminologie : théorie, méthode et application / Maria Teresa Cabré.
U. Linguistique. Armand Colin. Paris, 1998.



Catherine Chen, Jack Merullo, and Carsten Eickhoff.

Axiomatic Causal Interventions for Reverse Engineering Relevance
Computation in Neural Retrieval Models.

*In Proceedings of the 47th International ACM SIGIR Conference on
Research and Development in Information Retrieval*, pages 1401–1410,
Washington DC USA, July 2024. ACM.

Références II



Wenqi Fan, Yajuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li.

A Survey on RAG Meeting LLMs : Towards Retrieval-Augmented Large Language Models, June 2024.

[arXiv :2405.06211.](#)



Algirdas Julien Greimas and Joseph Courtès.

Sémiotique. Dictionnaire raisonné de la théorie du langage.

Hachette, Paris, 1979.

Références III



Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela.

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.
arXiv : Computation and Language, 2020.



Neel Nanda and Joseph Bloom.

TransformerLens, 2022.



Andrew Parry, Catherine Chen, Carsten Eickhoff, and Sean MacAvaney.
MechIR : A Mechanistic Interpretability Framework for Information Retrieval.

In Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025,

Références IV

Proceedings, Part V, volume 15576 of *Lecture Notes in Computer Science*, pages 89–95. Springer, 2025.



Shriyank Somvanshi, Md Monzurul Islam, Amir Rafe, Anannya Ghosh Tusti, Arka Chakraborty, Anika Baitullah, Tausif Islam Chowdhury, Nawaf Alnawmasi, Anandi Dutta, and Subasish Das.

Bridging the Black Box : A Survey on Mechanistic Interpretability in AI.
ACM Computing Surveys, 58(8) :1–35, June 2026.



The Trung Tran, Carlos-Emiliano González-Gallardo, and Antoine Doucet.

Retrieval Augmented Generation for Historical Newspapers.

In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, pages 1–5, Hong Kong China, December 2024. ACM.

Références V



Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei.

Multilingual E5 Text Embeddings : A Technical Report, February 2024.

Sommaire

6 Annexes

Activation Patching pour l'IR [CME24]

Soit $Q \times D \subset \mathcal{Q} \times \mathcal{D}$ l'ensemble des paires question, document. Soit $Q \times \tilde{D}$ le même ensemble mais avec les documents perturbés à la place

- ① Faire une passe pour toutes les paires $Q \times D$
 - ▶ noter o_n^e la sortie de chaque nœud $n, \forall e \in Q \times D$
 - ▶ noter p_D la performance du modèle sur les données non perturbées
- ② Faire une passe pour toutes les paires $Q \times \tilde{D}$
 - ▶ noter $o_n^{\tilde{e}}$ la sortie de chaque nœud $n, \forall \tilde{e} \in Q \times \tilde{D}$
 - ▶ noter $p_{\tilde{D}}$ la performance du modèle sur les données perturbées
- ③ On renomme $D, e, \tilde{D}$ et $\tilde{e}$ respectivement
 - ▶ $\hat{D}, \hat{e}, \check{D}$ et $\check{e}$ si $p_D > p_{\tilde{D}}$
 - ▶ $\check{D}, \check{e}, \hat{D}$ et $\hat{e}$ sinon

Pour chaque nœud n

- ④ Faire une passe de $Q \times \check{D}$ en remplaçant $o_{i,j}^{\check{e}}$ par $o_n^{\hat{e}}$ pour chaque $\check{e}$. On note la performance $\bar{p}_n$
- ⑤ $P_n = \frac{\bar{p}_n - p_{\hat{D}}}{p_{\check{D}} - p_{\hat{D}}}$ exprime l'impact de la perturbation sur la performance du modèle pour le nœud n

Liste des perturbations pratiquées

Mot remplacé	Niveau 1	Niveau 2	Niveau 3	Niveau 4
aérien	de vol	poétique	épicurien	acide tracté
article	sujet	déterminant	table	luge pontificale
défense	protection	corne	fromage	poire retournée
drone	vecteur	bourdon	gazon	ver de verre
limitation	contrainte	bornage	animation	plastique anarchique
matériel	outil	concret	post-it	porte sans poivre
système	dispositif	ensemble	panneau	esperluette en vent
technique	spécialisé	difficile	magique	feuille entubée

Matrice des noeuds sensibles pour le groupe de perturbation ...

