

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

Marine DELVALLEZ

23 juillet 2026

Remerciements

Table des matières

1	Introduction	4
1.1	Environnement du stage	4
1.2	Outils utilisés durant le stage	4
1.3	Contexte et problématiques traitées dans ce mémoire	4
1.3.1	Contexte	4
1.3.2	Problématiques qui ont guidé ce travail	5
2	État de l’art	6
2.1	Systèmes de Génération et Grands Modèles de Langue Augmentés par Extraction (RAG et RALLM)	6
2.1.1	Besoin et Définition	6
2.1.2	L’architecture RAG	6
2.1.3	Architectures aperçues au cours du stage	6
2.1.4	Architecture et Modèles utilisées au cours du stage	6
2.2	Interprétabilité pour les systèmes de RAG	7
2.2.1	Trustworthy AI	7
2.2.2	Explicabilité pour le RAG	7
2.2.3	Interprétabilité Mécaniste	7
2.2.4	Activation Patching	7
3	Travail préalable et préparation des expériences	8
3.1	Challenge RAG@EvalLLM	8
3.2	Construction d’un dataset adapté	8
3.3	Adaptation du modèle RAG4HN	9
3.4	Objectifs	9
3.4.1	Localisation de la langue de spécialité dans un modèle	9
3.4.2	Intégration des travaux dans l’explicabilité post-hoc des modèles d’extraction d’information	9
4	Contribution méthodologique	10
4.1	Construction des perturbations	10
4.1.1	Identification des mots cibles	10
4.1.2	Perturbation replace par niveau	10
4.1.3	Perturbations de type full_replace	10
4.2	Estimation de la pertinence d’une perturbation	10
4.3	Interprétation des résultats de l’Activation Patching	11
4.3.1	Visualisation des résultats proposée par MechIR	11
4.3.2	Autres visualisations de la sensibilité du modèle à la perturbation	11
4.3.3	Critère d’identification des noeuds sensibles	11
4.3.4	Matrice des noeuds sensibles	11
5	Résultats expérimentaux	12
5.1	Résultats et Analyse	12
5.1.1	Application à la cartographie du vocabulaire de la défense français dans mE5	12
5.1.2	Préconisations d’utilisation	12
5.2	Perspectives d’évolution de la méthode proposée	12

6	Conclusions et Perspectives	13
A	Résultats bruts d'expériences	15
A.1	Question Comme Document	15
A.1.1	Représentation et similarité de la question paddées/pertubée	15
A.1.2	Différence des scores de pertinence pour question paddée et question perturbée	16
A.2	Séparer les données positives et négatives	17
A.2.1	Distribution des Scores	17
A.2.2	Scores pour les perturbations full_replace	18
A.3	Niveaux de perturbation	19
A.3.1	Distribution des scores en fonction du niveau de perturbation	19

Chapitre 1

Introduction

1.1 Environnement du stage

- Stage à lieu au LIFO, dans l'équipe CA, encadré par ...
- Stage de fin de master d'informatique Minerve GPEx
- Challenge@EvalLLM comme cas d'usage, participation envisagée mais timing pas OK
- différents événements et sorties organisées pendant le stage
 - TransIA : Tours, Présentations pluridisciplinaires sur l'intégration de l'IA (sous toutes ses formes) dans la société
 - TAL-I : journée du GDR TaL-I à Paris (Jussieu) sur les travaux portant sur l'interprétabilité et l'explicabilité dans le contexte des tâches de traitement de la langue naturelle
 - 2nd Symposium on Mental Health and AI : deux jours de conférence (suivi en pointillé en distanciel) présentation des travaux à la croisée entre santé mentale et intelligence artificielle (impact, outils de détection, de prise en charge)
 - Place du Numérique : événement public de sensibilisation et de communication sur le Numérique organisé place du Martoi à Orléans (Des événements similaires organisés dans chaque département de la région CVL) J'ai participé à la tenue du stand de l'université, été ambassadrice de l'événement et ai joué le rôle d'un témoin dans le tribunal des générations futures portant sur "L'IA nous contrôle-t-elle?"
 - Présentation aux journées d'équipe CA

1.2 Outils utilisés durant le stage

- Développement de l'architecture et réalisation des expérimentations en python dans 3 formats (script, implémentation d'un CLI, notebook)
- Principales bibliothèques python utilisées : PyTorch, TransformerLens, HuggingFace, PyMuPDF4LLM,...
- Utilisation de plusieurs git pour assurer le suivi des versions des outils développés lors du stage
- Pour l'exécution de tâches demandant en charge de calcul (Génération de la base de représentation des données, exécution des différentes architectures explorées et/ou implémentées) : utilisation des grappes de calcul accessibles aux chercheurs du LIFO CaSciModOT et Mirev
- utilisation d'outils à base d'IA très modérée et pour les tâches de l'ordre de reformulation (ChatGPT), traduction (DeepL), exploration de la littérature (LitMaps, Perplexity), rédaction de petites fonctions et outils python (ChatGPT)

1.3 Contexte et problématiques traitées dans ce mémoire

1.3.1 Contexte

- Intro RAG :
 - les modèles ont des soucis pour gérer la connaissance (hallucination, préemption des modèles)

- on a couplé les modèles génératifs avec un base de connaissance entretenue et maitrisée par concept
- Intro XAI
 - modèles opaques
 - besoin de confiance
 - méthode d'explications pour rendre comportement accessible et anticipable
- methodes qui ouvrent la boîte, observent et donnent des explications
- On souhaite cartographier ces modèles
- On choisi un outil
- On découvre les difficultés et problématiques à l'usage
- On propose une méthodologie pour exploiter cet outil dans l'optique de cartographier la langue de sp

à reprendre

- Objectif naif : Identifier les composants du modèle impliqués dans l'identification et la gestion de la langue de spécialité d'un dataset
- Objectif pratique : Identifier et proposer des préconisations d'usage de la librairie MechIR pour la localisation des composant impliqués dans la langue de spécialité
- Définition de langue de spécialité, réduction de l'objectif à l'identification du vocabulaire de spécialité

1.3.2 Problématiques qui ont guidé ce travail

TODO

Chapitre 2

État de l'art

2.1 Systèmes de Génération et Grands Modèles de Langue Augmentés par Extraction (RAG et RALLM)

2.1.1 Besoin et Définition

- Les modèles profonds apprennent sur les données [ref à récupérer de Lewis+20]
- Le problème des modèles génératifs : hallucination et connaissance bornée dans le temps [ref à récupérer de Lewis+20]
- Il y a quelques années, on a cherché à associer une base de connaissances à ces modèles [Lewis+20] (= définition de RAG)
- Ces approches peuvent être appliquées aux tâches et modèles du TALN : RALLM

2.1.2 L'architecture RAG

Description de l'archi de façon plus avancée

- Ces modèles peuvent être utilisés pour différents types de tâches (voir tableau csv sur RAGRights)
- Avec le temps, l'idée de [Lewis+20] a été adaptée et augmentée de composants supplémentaires
- [Fan+23] propose la structure générique ... (exploiter slides Séminaire CA)

2.1.3 Architectures aperçues au cours du stage

Au delà de l'innovation de l'architecture, on retrouve des innovations sur plusieurs caractéristiques de ces modèles : [R24/04] [R16/02]

- Multimodal [Liu+25]
- Chain of Thought [Xu+24]
- Gestion BDD dynamique [Zhu25] (<https://arxiv.org/abs/2508.05662>)
- BDD sparse et interprétable [Prouteau]???

2.1.4 Architecture et Modèles utilisées au cours du stage

Dans le cadre du stage,

- travailler sur une adaptation du modèle RAG4HN. [Tran+25]
- _description de l'architecture RAG4HN_ + parallèle avec l'architecture générique de [Fan+23]
- L'extracteur de RAG4HN est mE5 [Wang+25b] (version multilingue de E5 [Whang+25a])

2.2 Interprétabilité pour les systèmes de RAG

2.2.1 Trustworthy AI

La définition d'IA de confiance (Trustworthy AI) tend à varier en fonction des situation et des utilisations de cette expression. Dans le contexte du Machine Learning, on désigne l'IA de confiance comme l'ensembles des systèmes d'IA dans lesquels il est légitime (dans le sens humainement understandable) d'avoir confiance. Le NIST (National Institute of Standards and Technology (US)) propose différents axes ou composantes de l'IA de confiance : ... [Ni+25] + [Ref citée]

2.2.2 Explicabilité pour le RAG

Dans la suite, nous nous concentrons sur l'explicabilité dans le contexte du RAG. [Ni+25] distingue deux temps pour l'explicabilité : Extraction et génération ... (prévoir la question de l'association/intégration des deux)

- Métrique ??

2.2.3 Interprétabilité Mécaniste

[Somvanshi+25]

- Interprétabilité vs Explicabilité
- Définition d'Interprétabilité Mécaniste et approches (Manual Circuit Tracing, *Intervention-based technique*, Representation analysis, Toy Model and synthetic tasks)

2.2.4 Activation Patching

- Définition de activation patching [Chen+24] [R30/04] [Séminaire CA]
- MechIR [Parry+25]
- Perturbation
- Données produites, visualisation et interprétation des graphiques

Chapitre 3

Travail préalable et préparation des expériences

3.1 Challenge RAG@EvalLLM

- Le sujet de l'extraction d'informatin et lu RAG est d'actualité
- Présentation du Challenge RAG@EvalLLM
 - La tâche
 - Identification des documents pertinents pour une question (phrase parfois complexe ou suite de mots clés)
 - questions fournies au dernier moment
 - pas de gold
 - Les données
 - pdf à traiter sans structure normée du contenu
 - majoritairement en français sauf quelques uns en anglais
 - peuvent contenir des images

3.2 Construction d'un dataset adapté

À REPRENDRE

Base de données:

- Paires
- liste des champs et justification

Choix pour datasets

- Dataset avec des paires question-document (questions différentes et différents pdf, taille des documents, répartition positif/négatif)
- Dataset avec une seule question (construction à partir de 3 pdf)

[description du dataset dans les exps et les specs]

- motivation (reprendre l'expé présentée par MechIR, faciliter la création)
- limites (biais, diversité ~> généralisabilité)

=> Placer le nom des outils utilisés (PyMuPDF4LLM et langchain.text_splitter.MarkdownSplitter)

- Le format de dataset visé
 - tous les dérouler V0, V1 et format V2 (à chaque fois: format, ce qui manque, les modifications proposés)

3.3 Adaptation du modèle RAG4HN

- Choix de ce modèle : traite des données en FR
- Suppression de la partie WebSearch (hors sujet)
- Suppression de la première extraction sur la base des titres (incompatible avec le format des données du challenge)
- Montée en version des librairies pour la compatibilité avec les outils
- Reprise du reranking (mettre au propre la photo du schema sur l'ardoise)

3.4 Objectifs

3.4.1 Localisation de la langue de spécialité dans un modèle

- Définition de langue de spécialité
- Formalisation de l'objectif

3.4.2 Intégration des travaux dans l'explicabilité post-hoc des modèles d'extraction d'information

- Cartographie de modèle d'extraction d'information
- Analyse des éléments identifiés par le modèle lors du passage d'un document

Chapitre 4

Contribution méthodologique

À REMETTRE EN FORME

4.1 Construction des perturbations

- on s'intéresse aux perturbations de type replace pour travailler sur l'impact des mots importants (car liés au vocabulaire de spécialité) en les remplaçant
- La conception de perturbations nécessite de bien choisir les mots à remplacer (traité dans la sous partie 3)
- La sous partie 3 propose un nouveau type de perturbations

4.1.1 Identification des mots cibles

- Le choix des remplaçant est fait par jugement humain
 - Le choix des mots à remplacer se fait à l'aide de
 - fréquence dans l'échantillon de données sélectionné
 - IDF
 - appartenance au vocabulaire de spécialité (jugement humain)
 - présence du mot dans la question étudiée (cas du dataset à 1! question)
- la bonne couverture des documents est importante (commentaire sur NaN) pour avoir des résultats lisibles

4.1.2 Perturbation replace par niveau

- On propose plusieurs candidats au remplacement classé en 4 niveaux ...
- Rq : il faudrait fonder l'estimation du niveau des perturbations

4.1.3 Perturbations de type full_replace

- pour augmenter le nombre de document perturbé et limiter l'impact d'un mot seulement dans la cartographie (avoir une apperçu plus global), on propose un nouveau type de perturbation : full_replace
- on étend les niveaux à ce nouveau type de perturbations

4.2 Estimation de la pertinence d'une perturbation

SI PAS DÉJÀ PRÉSENTÉ DANS PARTIE DE L'ÉTAT DE L'ART SUR MECHIR

- On visualise des scores de similarité entre question et document
- graphique d'exemple + texte d'accompagnement à la lecture
- Une perturbation peut être considérée comme pertinente lorsqu'elle modifie significativement (jugement humain) la distribution des cores pour l'échantillon du dataset étudié

- on propose d'autres visualisations et confrontation de la distribution des scores entre perturbation :
 - observation de la distribution des scores pour les documents pertinents et non-pertinents (sont-ils tr
 - observation de la distribution des scores d'un niveau de perturbation à l'autre pour un mot cible (que

4.3 Interprétation des résultats de l'Activation Patching

Un fois la ou les perturbations choisies, on peut appliquer l'activation patching à l'aide de l'outil MechIR

4.3.1 Visualistaion des résultts proposée par MechIR

SI PAS DÉJÀ FAITE DANS SOTA sur MechIR

- graphique + accompagnement à la lecture (on obtient une sensibilité de chaque noeuds pour chaque paire (Question, document), on visualise la moyenne)

4.3.2 Autres visualisations de la sensibilité du modèle à la perturbation

à placer : confronter les matrices d'une perturbation à l'autre

- étant donné la construction de ces valeurs, on s'intéresse aussi aux écarts-types
- pour faciliter la visualisation et la comparaison en cas de sensibilité variable d'une perturbation à l'autre on propose de rescaler les poids des matrices

4.3.3 Critère d'identification des noeuds sensibles

- L'identification des noeuds sensibles doit reposer sur un critère commun. On propose un critère de sensibilité ...

4.3.4 Matrice des noeuds sensibles

- Pour faciliter le besoin de mise en perspective des matrice de snesibilité d'un perturbation à l'autres
- Visualisation des noeuds sensibles ****commun à plusieurs perturbation**** (matrice des noeuds sensibles)
 - > croiser les différentes caractéristiques (mot remplacé, type, niveau)
 - > critique des noeuds sensibles de la dernière couche

Chapitre 5

Résultats expérimentaux

À REPRENDRE

5.1 Résultats et Analyse

5.1.1 Application à la cartographie du vocabulaire de la défense français dans mE5

- Grand nombre de noeuds de la dernière couche du modèle
- pas de noeuds clairement associé au vocabulaire
- ...

5.1.2 Préconisations d'utilisation

- des perturbations qui couvrent tous ou très grande partie des documents étudiés
- utiliser la question comme document témoin

5.2 Perspectives d'évolution de la méthode proposée

- échelle pour situer le niveau d'un mot candidat remplaçant en fonction du mot remplacé
- critère d'identification du vocabulaire de spécialité
- intégration d'autres informations au critère de sensibilité

Chapitre 6

Conclusions et Perspectives

TODO

Bibliographie

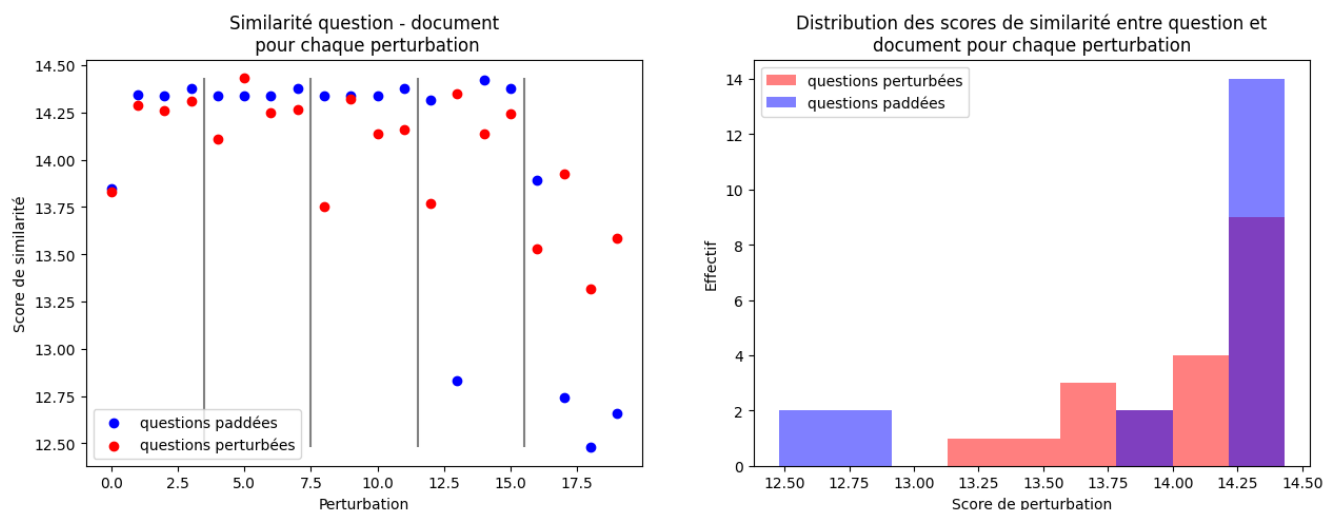
[Parry et al., 2025] Parry, A., Chen, C., Eickhoff, C., and MacAvaney, S. (2025). MechIR : A Mechanistic Interpretability Framework for Information Retrieval. In *Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part V*, volume 15576 of *Lecture Notes in Computer Science*, pages 89–95. Springer.

Annexe A

Résultats bruts d'expériences

A.1 Question Comme Document

A.1.1 Représentation et similarité de la question paddées/pertubée



Dataset manipulé Une paire (question, question) où la question est exploitée comme question mais aussi comme document

Perturbations appliquées

drone -> engin autonome
limitation -> contrainte
système -> dispositif
technique -> spécialisé
drone -> vecteur
limitation -> bornage
système -> ensemble
technique -> difficile
drone->gazon
limitation->animation
système->panneau
technique->magique
drone->ver de verre
limitation->plastique anarchique
système->esperluette en vent

```

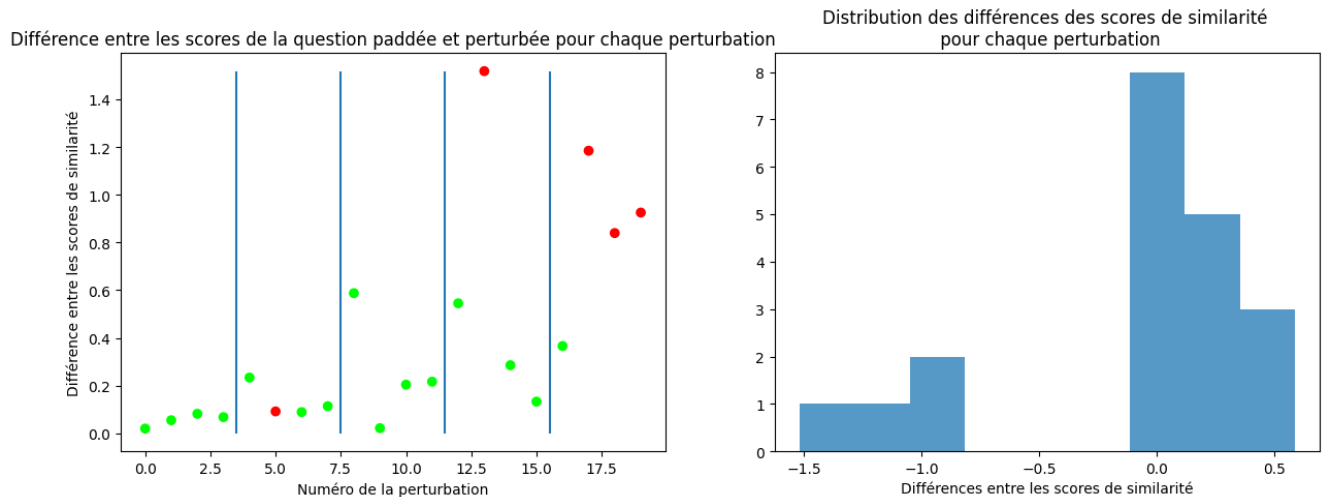
technique->feuille entubée
Full-replace Synonymes 1
    drone -> engin autonome
    limitation -> contrainte
    système -> dispositif
    technique -> spécialisé
Full-replace Polysémie 1
    drone -> vecteur
    limitation -> bornage
    système -> ensemble
    technique -> difficile
Full-replace Elephants 1
    drone -> gazon
    limitation -> animation
    système -> panneau
    technique -> magique
Full-replace Elephants Roses 1
    drone -> ver de verre
    limitation -> plastique anarchique
    système -> esperluette en vent
    technique -> feuille entubée

```

Données représentées Le graphique de gauche représente les scores de pertinence de la question comme document pour chaque perturbation. Les questions non modifiées (représentées en bleu) sont juste paddées pour faire correspondre la taille du document non modifié avec celui perturbé (ici de la question perturbée, représentées en rouge). Les scores de pertinence pour chaque perturbation sont dans l'ordre des perturbations citées. Les lignes verticales séparent les perturbations par niveau (de 1 à 4 en allant de gauche à droite) puis regroupent les perturbation de type full_replace (le plus à droite, de niveau 1 à 4 en allant de gauche à droite) Le graphique de droite est un histogramme des scores (en bleu pour la question paddée et en rouge pour la question perturbée)

Commentaires On observe bien la différence de représentation entre les questions et les question perturbées. On constate aussi que le système de padding is en place par [Parry et al., 2025] provoque une perturbation des documents. Certaines perturbations améliorent la similarité de la question avec elle même. Cela est surprenant

A.1.2 Différence des scores de pertinence pour question paddée et question perturbée



Données et perturbation Question comme document
Perturbation présentées précédemment

Lecture des graphiques Le graphique de gauche associe à chaque perturbation (dans l'ordre cité) la différence entre les scores de similarité à la question de la question perturbée et la question paddée pour la perturbation. Les points en rouge correspondent à aux différences négatives et les points en vert aux différences positives. On attend des scores positifs car perturber un objet (en l'occurrence une question) réduit sa similarité à lui même. Les points sont séparés pour identifier les perturbations de type replace de différent niveau (1 à 4 de gauche à droite). Le cinquième (dernier) groupe correspond aux perturbations de type full_replace (niveau 1 à 4 de gauche à droite) Le graphique de droite représente la distribution de ces différences.

Analyse des graphiques Pour le premier graphique, on constate une augmentation de la différence pour les niveaux les plus élevés concernant les perturbations de type replace. Les perturbations de type full_replace ont un effet de perturbation moins important que le padding induit par le modèle. On peut lier ce phénomène au fait que les modèles récents sont capables d'associer les expressions de même sens et de faire abstraction du contenu incohérent au sein d'un texte.

A.2 Séparer les données positives et négatives

Données Question : 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 ?

Documents : 3 pdf portant sur 3 sujets distincts (limitations matérielles du drone, code de la défense - matériel de guerre, fausses infos pdt les JO/JP) dont sont extraits 28 textes (50% de documents pertinents)

Sujet du PDF	Documents Positifs	Documents Négatifs
Limitations techniques de drones	14 (100%)	0
Code de la Défense (Chapitre sur le matériel de guerre)	0	6 (43%)
Desinformation pendant les Jeux Olympiques	0	8 (57%)

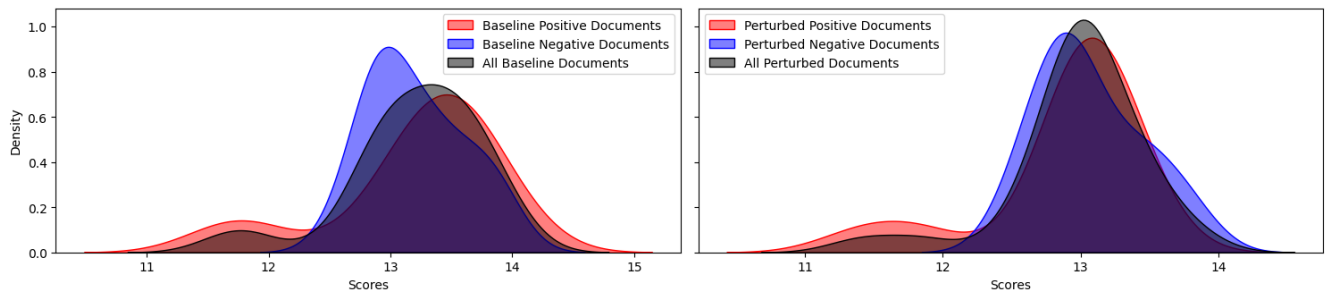
Perturbations

TODO

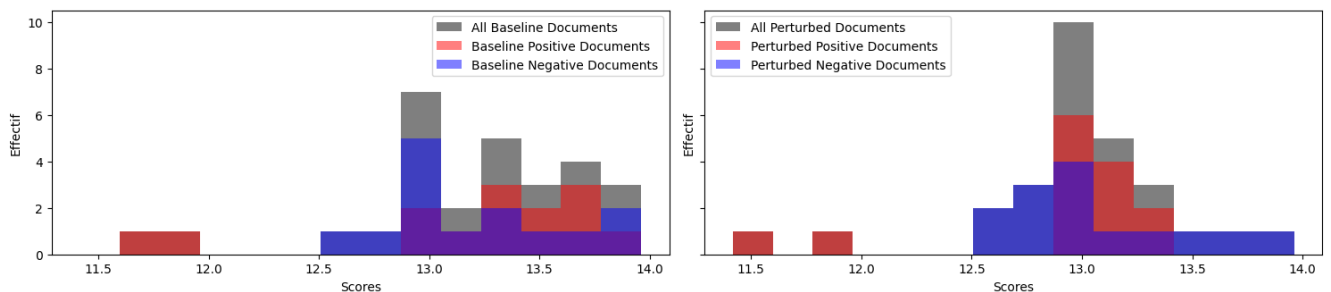
Full-replace

A.2.1 Distribution des Scores

Distribution des scores de similarité des documents par rapport à la question



Histogramme des scores de similarité entre le document et la question

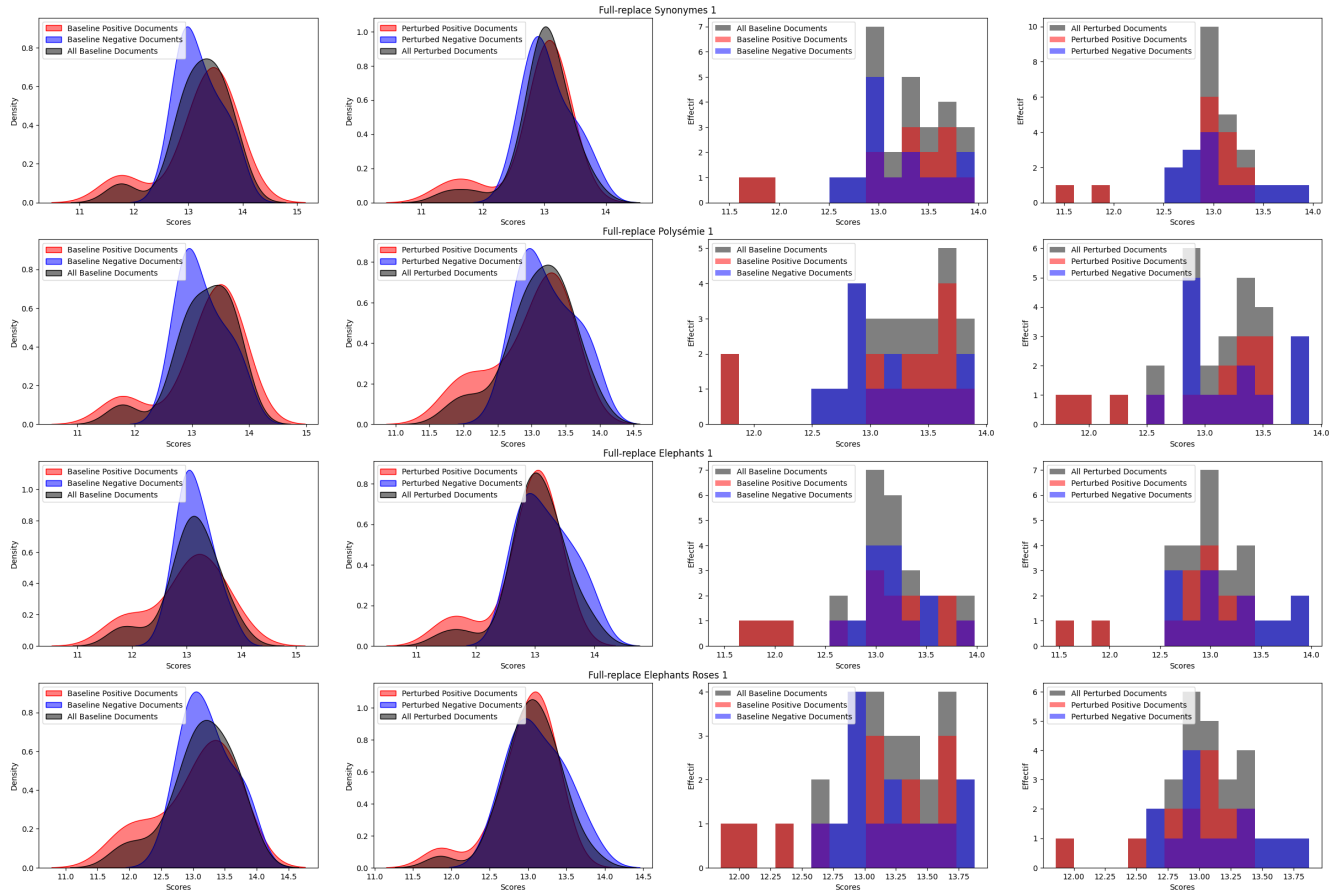


Données visualisées Ces graphiques représentent la répartition des scores de similarité entre la question et chaque document pour la perturbation **Full-replace Synonyme 1**. En rouge sont représentées les documents désignés pertinents vis à vis de la question, en bleu les documents non pertinents et en noire l'ensemble des documents. Le graphique de gauche correspond au cas des documents paddés et les graphiques de droite correspond aux documents perturbés.

Analyse

A.2.2 Scores pour les perturbations full_replace

Distribution et Histogramme des scores de similarité des documents par rapport à la question



Données Visualisées Chaque ligne de graphique traite une perturbation. Les graphiques représentent la distribution/ l'histogramme des scores de similarité à la question de chaque document. En rouge, on considère les documents désignés comme pertinents, en bleu les documents considérés comme non-pertinents et en noir l'ensemble des documents. Pour chaque paire de graphique, celui de gauche concerne les documents paddés (ramenés à la même longueur que les documents perturbés) et celui de droite concerne les documents perturbés par la perturbation indiquée dans le titre de la ligne de graphique.

Analyse Dans un premier temps, concernant les documents paddés, on retrouve une répartition des documents pertinents et non-pertinents cohérente : les documents pertinents ont un score de similarité à la question plus élevé que les documents non-pertinents. On notera que cette variation de distribution reste assez faible.

Finalement, on ne fait que vérifier le comportement des scores vis-à-vis de l'étiquette des documents. La visualisation des scores des documents perturbés n'est pas exploitée. J'ai l'impression que les documents positifs voient leur score diminuer et que les documents négatifs voient leur score augmenter. Ce comportement (1) est

suffisamment faible pour n'être qu'une impression, (2) n'a pas de sens étant donné la nature de la perturbation (en tout cas je ne l'explique pas). Peut-être que la visualisation des scores pour les documents non perturbés suffit.

A.3 Niveaux de perturbation

Intégrer dans l'analyse de la proposition de perturbation par niveau

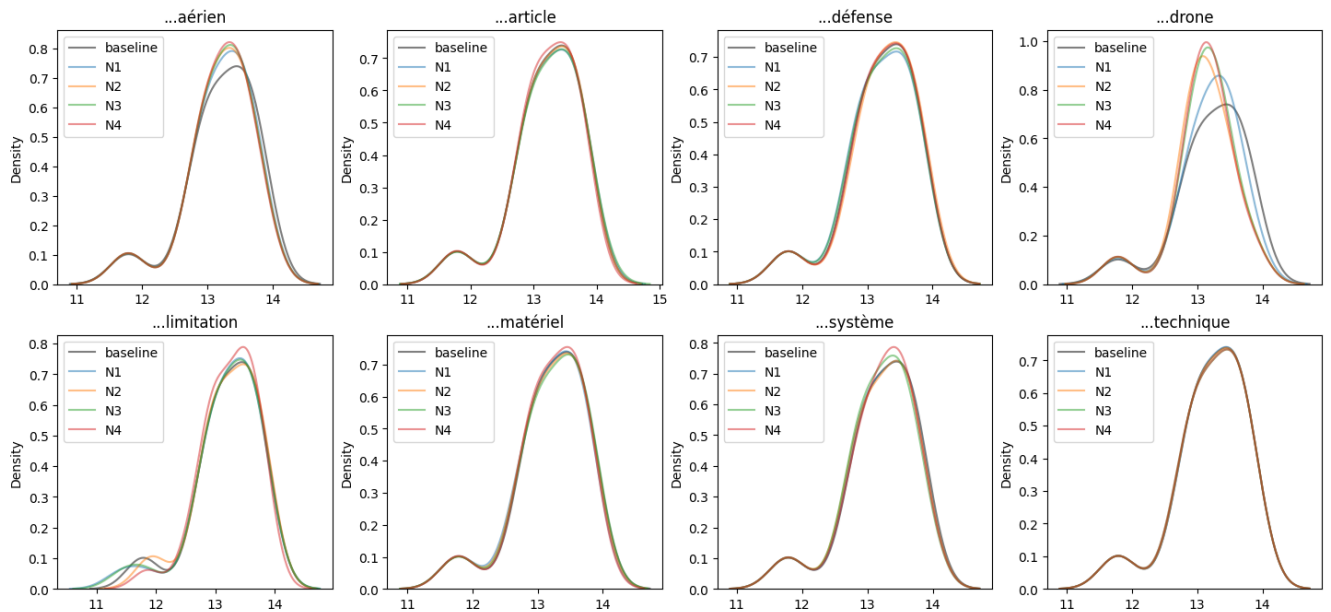
Dataset Idem que précédemment

Perturbations

TODO

A.3.1 Distribution des scores en fonction du niveau de perturbation

Distribution des scores pour chaque type de perturbation appliquée au mot ...



Données visualisées On visualise la distribution des scores de similarité à la question des documents. Les distributions sont regroupées par mot remplacé (appelé mot cible dans la suite du paragraphe). Pour chaque mot on observe la distribution des scores de perturbation pour les documents non modifiés (paddés pour la perturbation de niveau 1) en noir, avec le mot cible remplacé par un synonyme (perturbation de niveau 1, en bleu), avec le mot cible remplacé par le synonyme d'un autre de ses sens que celui utilisé dans le document (perturbation de niveau 2, en orange), avec le mot cible remplacé par un mot ne partageant aucun sens avec le mot cible mais respectant sa nature (perturbation de niveau 3, en vert), et avec le mot cible remplacé par un groupe de mot absurde (perturbation de niveau 4, en rouge).