Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Explicabilité par Perturbations pour les Systèmes RAG

Yongxin Zhou Philippe Mulhem Didier Schwab
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000, Grenoble, Frances
{yongxin.zhou, philippe.mulhem, didier.schwab}@univ-grenoble-alpes.fr

RÉSUMÉ

Les systèmes de Génération Augmentée par Récupération (RAG) ont pour objectif d'améliorer les Grands Modèles de Langage (LLM) en intégrant des informations provenant de sources externes pour générer des réponses, mais leur manque de transparence en terme d'explicabilité soulève des préoccupations, particulièrement dans des domaines tels que la santé, la finance ou le droit. Les méthodes par perturbations fournissent une explicabilité *post-hoc*, avec des RAG considérés comme des boîtes noires, en modifiant systématiquement les entrées ou documents récupérés pour évaluer la stabilité des réponses et l'attribution des sources. Ce document présente un aperçu de l'explicabilité des systèmes RAG, en se concentrant sur les approches basées sur des exemples et des perturbations. Nous proposons une taxonomie des techniques de perturbation à différents niveaux de granularité, montrant comment elles offrent des indicateurs interprétables sur le comportement des modèles.

ABSTRACT

Explainability through Perturbations for Retrieval-Augmented Generation (RAG) Systems.

Retrieval-Augmented Generation (RAG) systems enhance Large Language Models (LLMs) by incorporating information retrieved from external sources to generate responses, but their lack of transparency in terms of explainability raises concerns, especially in domains such as healthcare, finance and law. Perturbation-based methods provide *post-hoc* explainability, with RAGs considered as black boxes, by systematically modifying inputs or retrieved documents to evaluate response stability and source attribution. This document presents an overview of the explainability of RAG systems, focusing on example-based and perturbation-based approaches. We present a taxonomy of perturbation techniques at different granularities, showing how they offer interpretable insights into model behavior.

MOTS-CLÉS : Génération Augmentée par Récupération (RAG), Explicabilité, Perturbations.

KEYWORDS: Retrieval Augmented Generation (RAG), Explainability, Perturbations.

1 Introduction

Les systèmes de Génération Augmentée par Récupération (RAG) (Lewis *et al.*, 2020) améliorent les Grands Modèles de Langage (LLM) en intégrant une récupération de connaissances externe, effectuée par un système de recherche d'information, réduisant ainsi les hallucinations et améliorant la précision des réponses. Cependant, la force même des systèmes RAG - leur capacité à synthétiser des informations provenant de multiples sources - amène de nouveaux défis en matière de transparence et d'interprétabilité. La complexité de ces systèmes, comme celle de nombreux autres modèles d'IA, rend souvent leurs processus décisionnels opaques, ce qui soulève des inquiétudes quant à leur fiabilité

Actes de l'atelier Accès à l'information basé sur le dialogue et grands modèles de langage 2025 (DIAG-LLM), pages 1–6.

- particulièrement dans des domaines sensibles tels que la santé, la finance ou le droit (Arrieta *et al.*, 2020).

Pour répondre à ce problème, les méthodes d'IA explicable (XAI) se sont imposées comme des outils pertinents pour auditer les systèmes RAG. Ces approches visent à produire des explications compréhensibles par les humains, sous forme de langage naturel ou d'autres formats interprétables. Elles sont donc étroitement liée à la perspective utilisateur. Les humains utilisent des exemples pour comprendre, expliquer et démontrer leurs arguments, et les explications basées sur des exemples sont considérées comme l'une des méthodes d'explication les plus simples à comprendre (Humer *et al.*, 2022).

Dans ce survol, nous discutons de l'explicabilité des systèmes RAG à travers des explications basées sur des exemples, en particulier les perturbations comme outil pour les systèmes RAG, qui offre un moyen de vérifier la robustesse et l'interprétabilité des RAG en modifiant systématiquement les entrées afin d'observer le comportement du modèle.

2 Taxonomie des Approches Explicatives

Le domaine de l'IA explicable (XAI) propose diverses taxonomies pour l'interprétabilité des modèles, comme le décrivent les publications récentes (Poché *et al.*, 2023). On distingue notamment les explications locales des explications globales, les premières clarifiant des prédictions individuelles tandis que les secondes décrivent le comportement global du modèle. Une autre classification importante oppose les méthodes *post-hoc*, appliquées après l'entraînement du modèle, aux approches intrinsèques expliquant soit le processus d'apprentissage soit le modèle appris, ainsi qu'aux méthodes explicables par conception comme les arbres de décision (Zhou *et al.*, 2021), intrinsèquement interprétables par construction. Les formats d'explication varient également considérablement, allant des attributions de caractéristiques et des justifications en langage naturel (Tekkesinoglu & Kunze, 2024) aux exemples adversariaux et modèles substitutifs.

Dans le domaine de la Recherche d'Information (RI), les techniques d'explicabilité adoptent souvent des stratégies *post-hoc* (Anand *et al.*, 2022), telles que l'attribution de caractéristiques ou des explications en texte libre, qui mettent en évidence des termes clés ou bien génèrent des justifications succinctes pour les résultats retrouvés.

Les perturbations, un sous-ensemble des méthodes *post-hoc*, sont particulièrement utiles pour les systèmes RAG. En introduisant des modifications délibérées et minimales aux entrées - comme des substitutions de synonymes ou des erreurs typographiques - ces techniques révèlent les *vulnérabilités* du modèle et fournissent des explications contrastives. Par exemple, les exemples adversariaux démontrent comment de légères modifications des entrées peuvent altérer les prédictions (Anand *et al.*, 2022), exposant ainsi les frontières décisionnelles du modèle et favorisant une compréhension plus approfondie de son comportement.

3 Explicabilité par Perturbations des RI et RAG

Les explications par l'exemple englobent plusieurs formats, notamment : les exemples similaires (factuels), les contrefactuels, les semi-factuels, les instances influentes, les prototypes et les visualisa-

Cibles	Méthodes	Types	Définitions	Granularité	Exemples
Document	Sous-ensemble	Suppression	Identification d'un sous-ensemble minimal de contenu dont la suppression dégrade le classement au-delà du rang k.	Universel	Suppression de phrases : classement (Rorseth *et al.*, 2023), QA (Sudhi *et al.*, 2024)
		Combinaison	Analyse de l'influence des combinaisons de sous-ensembles via échantillonnage aléatoire.	Universel	QA ouvert (Rorseth *et al.*, 2024)
	Permutation	Réordonnancement	Modification de l'ordre des sources pour contrer le biais de position.	Universel	QA ouvert (Rorseth *et al.*, 2024)
		Réorganisation	Modification de l'ordre des mots dans chaque source.	Mots	QA (Sudhi *et al.*, 2024)
	Remplacement	Phrase/Passage	Remplacement ciblé de phrases ou passages.	Phrases/Passages	Classement (Goren *et al.*, 2020)
		Entités	Remplacement des noms et noms propres par des termes aléatoires.	Mots	QA (Sudhi *et al.*, 2024)
		Antonymes	Substitution par des termes de sens opposé.	Mots	QA (Sudhi *et al.*, 2024)
		Synonymes	Remplacement sémantiquement neutre.	Mots	QA (Sudhi *et al.*, 2024), Classement (Wu *et al.*, 2023)
	Bruit	Injection	Insertion de termes aléatoires dans les sources.	Mots	QA (Sudhi *et al.*, 2024), Classement (Raval & Verma, 2020)
Requête	Modification	Préfixe	Insertion de préfixes pour dévier les réponses.	Universel	QA (Hu *et al.*, 2024)
		Termes	Augmentation minimale de la requête.	Universel	Classement (Rorseth *et al.*, 2023)

TABLE 1 – Vue d'ensemble des perturbations possibles en recherche d'information explicable. * : *source* = *token* utilisé dans (Sudhi *et al.*, 2024).

tions de caractéristiques (Poché *et al.*, 2023). Parmi ceux-ci, les contrefactuels - définis comme les échantillons les plus proches entraînant des prédictions différentes du modèle - présentent un intérêt particulier pour les méthodes basées sur les perturbations.

Dans la Recherche d'Information (RI), les perturbations adversariales ont fait l'objet d'études approfondies pour évaluer la robustesse des modèles. Des recherches démontrent que des modifications minimes de tokens (1 à 3 altérations) peuvent générer des documents sémantiquement similaires trompant les algorithmes de classement et altérant les scores de documents (Raval & Verma, 2020).

Les techniques de perturbations adversariales exploitent souvent l'optimisation par gradient pour modifier stratégiquement des tokens dans des documents pertinents ou non pertinents, induisant ainsi des fluctuations dans le classement (Wang *et al.*, 2022). Au-delà des perturbations au niveau documentaire, les perturbations de prompts (comme les injections de préfixes) ont été explorées pour manipuler les sorties RAG. La technique *Gradient Guided Prompt Perturbation (GGPP)*, par exemple, optimise de courts préfixes pour orienter les LLM vers des réponses incorrectes ciblées, révélant ainsi des vulnérabilités liées à la dépendance aux prompts dans les systèmes RAG (Hu *et al.*, 2024).

Ces perturbations sont systématiquement catégorisées dans le Tableau 1, selon :

— La **cible** (document, requête), qui indique quel élément est perturbé.
— La **méthode** qui indique quel **type** de perturbation est effectué (suppression de tokens, modifications adversariales), suivie des **définitions** des différents types de perturbation.
— La **granularité** des perturbations réalisées (des mots aux documents entiers), suivie d'**exemples**.

On constate à la vue de l'état de l'art que, de nombreuses études sur les perturbations dans les systèmes RAG restent préliminaires. Par exemple, la stratégie *leave-one-token-out* - évaluée sur seulement 100 paires de questions-réponses en anglais et allemand - représente l'une des rares approches validées pour des explications approximatives dans les systèmes RAG (Sudhi *et al.*, 2024). D'autres démonstrations, comme celles présentées dans (Rorseth *et al.*, 2023, 2024), bien qu'illustrant des explications obtenues par différentes méthodes de perturbation, n'ont pas fait l'objet d'évaluation ou de validation rigoureuse, soulignant le besoin de cadres de validation évolutifs.

Les perturbations peuvent être catégorisées selon deux types :

— **Universelles** : applicables dans tous les contextes d'utilisation ;
— **Spécifiques** : adaptées à certaines applications particulières de recherche d'information (RI), comme les systèmes question-réponse.

Ces approches présentent différentes granularités, pouvant s'appliquer aussi bien à des mots individuels qu'à des documents entiers.

4 Défis et Perspectives

L'utilisation de techniques de perturbation présente certaines limites. Ces perturbations artificielles peuvent ne pas refléter fidèlement le bruit présent dans des conditions réelles d'utilisation. De plus, leur application s'avère souvent coûteuse en calcul, nécessitant l'exécution de multiples expérimentations perturbées pour obtenir des résultats significatifs comme dans (Sudhi *et al.*, 2024).

Afin d'avancer vers davantage de réalisme dans ces approches à base de perturbations, les directions de recherche ci-dessous sont des directions possibles d'exploration :

— Des approches hybrides combinant méthodes par perturbation avec des techniques d'explicabilité intrinsèque, comme les architectures auto-explicatives, par exemple Self-RAG (Asai *et al.*, 2024).
— Le développement de collections de test standardisés pour évaluer de manière cohérente les explications basées sur les perturbations.
— Développement d'outils légers et extensibles pour l'intégration aux frameworks RAG existants, visant à réduire l'empreinte computationnelle des méthodes par perturbations et à en démocratiser l'usage.

Ces avancées permettraient d'améliorer à la fois la robustesse et l'adoption des méthodes d'explication pour les systèmes RAG, tout en réduisant leur coût computationnel. Une attention particulière devrait être portée sur le développement de perturbations représentatives des conditions réelles d'utilisation.

Remerciements

Ce travail a été partiellement financé par l'axe Systèmes Intelligents pour les Données, les Connaissances et les Humains du Laboratoire d'Informatique de Grenoble. Ce travail est aussi réalisé dans le cadre de la Chaire AugmentIA dirigée par Didier Schwab et portée par la Fondation Grenoble INP grâce au mécénat du Groupe Artelia. Cette chaire bénéficie également d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre de France 2030 portant la référence ANR-23-IACL-0006 (MIAI Cluster).

Références

ANAND A., LYU L., IDAHL M., WANG Y., WALLAT J. & ZHANG Z. (2022). Explainable information retrieval : A survey.

ARRIETA A. B., DÍAZ-RODRÍGUEZ N., DEL SER J., BENNETOT A., TABIK S., BARBADO A., GARCÍA S., GIL-LÓPEZ S., MOLINA D., BENJAMINS R. *et al.* (2020). Explainable artificial intelligence (xai) : Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, **58**, 82–115.

ASAI A., WU Z., WANG Y., SIL A. & HAJISHIRZI H. (2024). Self-RAG : Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

GOREN G., KURLAND O., TENNENHOLTZ M. & RAIBER F. (2020). Ranking-incentivized quality preserving content modification. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, p. 259–268, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3397271.3401058.

HU Z., WANG C., SHU Y., PAIK H.-Y. & ZHU L. (2024). Prompt perturbation in retrieval-augmented generation based large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, p. 1119–1130, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3637528.3671932.

HUMER C., HINTERREITER A., LEICHTMANN B., MARA M. & STREIT M. (2022). Reassuring, misleading, debunking : Comparing effects of xai methods on human decisions. DOI : 10.31219/osf.io/h6dwz.

LEWIS P., PEREZ E., PIKTUS A., PETRONI F., KARPUKHIN V., GOYAL N., KÜTTLER H., LEWIS M., YIH W.-T., ROCKTÄSCHEL T., RIEDEL S. & KIELA D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems*, volume 33, p. 9459–9474 : Curran Associates, Inc.

POCHÉ A., HERVIER L. & BAKKAY M.-C. (2023). Natural example-based explainability : A survey. In L. LONGO, Éd., *Explainable Artificial Intelligence*, p. 24–47, Cham : Springer Nature Switzerland.

RAVAL N. & VERMA M. (2020). One word at a time : adversarial attacks on retrieval models.

RORSETH J., GODFREY P., GOLAB L., KARGAR M., SRIVASTAVA D. & SZLICHTA J. (2023). Credence : Counterfactual explanations for document ranking. *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, p. 3631–3634.

RORSETH J., GODFREY P., GOLAB L., SRIVASTAVA D. & SZLICHTA J. (2024). Rage against the machine : Retrieval-augmented llm explanations.

SUDHI V., BHAT S. R., RUDAT M. & TEUCHER R. (2024). Rag-ex : A generic framework for explaining retrieval augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, p. 2776–2780, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3626772.3657660.

TEKKESINOGLU S. & KUNZE L. (2024). From feature importance to natural language explanations using llms with rag.

WANG Y., LYU L. & ANAND A. (2022). Bert rankers are brittle : A study using adversarial document perturbations. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '22, p. 115–120, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3539813.3545122.

WU C., ZHANG R., GUO J., DE RIJKE M., FAN Y. & CHENG X. (2023). Prada : Practical black-box adversarial attacks against neural ranking models. *ACM Trans. Inf. Syst.*, **41**(4). DOI : 10.1145/3576923.

ZHOU Y., BOUSSARD M. & DELABORDE A. (2021). Towards an xai-assisted third-party evaluation of ai systems : Illustration on decision trees. In D. CALVARESI, A. NAJJAR, M. WINIKOFF & K. FRÄMLING, Éds., *Explainable and Transparent AI and Multi-Agent Systems*, p. 158–172, Cham : Springer International Publishing.