# Evaluation de Génération de Texte en Domaine Spécifique, Cas d'étude

Ulysse Oliveri[1,2] Guillaume Gadek[1], Alexandre Dey[3], Arnaud Delhay-Lorrain[2], Damien Lolive[2], Benjamin Costé[3], Bruno Carron[1], Claude Fendzi [1], Bruno Grilheres[1]

(1) Airbus Defence & Space, 1 Bd Jean Moulin, 78990 Élancourt, France

(2) Univ Rennes/IRISA/CNRS, 6 Rue de Kerampont, CS 80518, 22305 Lannion Cedex, France

(3) Airbus Cybersecurity, 3 Rue Louis Braille, 35000 Rennes, France

ulysse.oliveri@irisa.fr

RÉSUMÉ

Un des problèmes ouverts de la Génération de Texte est l'évaluation en domaine spécifique. Ce *vision paper* étudie l'apport des techniques d'évaluation existantes appliquées au domaine particulier de la Défense et plus spécifiquement au cas d'usage des Demandes d'Informations. Ces dernières sont cependant confidentielles rendant impossible la constitution de jeux de données d'évaluation. Pour compléter l'indispensable évaluation humaine, nous proposons une méthodologie d'évaluation des Demandes d'Information synthétiques selon trois axes distincts et leur prise en compte par les métriques usuelles.

## 1 Introduction

La Génération de Langage Naturel, sous-domaine du Traitement de Langage Naturel, est un domaine de l'Intelligence Artificielle qui consiste à générer du texte servant à différents cas d'usage telles que la réponse à des questions [2], du résumé de texte [28] ou encore la traduction de données structurées en langage naturel grâce à des modèles d'Apprentissage Profond. Ce dernier a évolué de manière drastique grâce à de nouvelles architectures de modèles telles que les Réseaux Récurrents, les Transformers [33] ou encore les State Space Models [12] ainsi qu'à des nouvelles méthodes d'entraînement telles que l'Apprentissage par Renforcement avec Retour Humain [25], la spécification par Instructions [37] ou encore les méthodes de Spécialisation par Efficience des Paramètres (Parameter-Efficient Fine-Tuning - PEFT) [14]. Un besoin dans ce domaine est de pouvoir comparer les modèles sur des tâches spécifiques, aussi bien sur la qualité que sur la diversité des sorties d'un modèle sur une tâche donnée. Bien qu'atteignant les résultats que des experts humains produiraient dans certaines tâches ou domaines, l'évaluation de ces modèles reste une tâche ardue [11]. En effet, la génération de texte est dite ouverte, ce qui signifie que pour une entrée (instruction) donnée, il existe une multitude de sorties possibles pour un modèle, rendant compliqué le fait de se référer à une vérité terrain unique. Malgré cela, plusieurs jeux de données ainsi que divers benchmarks, souvent peuplés par des humains et majoritairement en anglais, sont mis à disposition du public permettant de comparer et évaluer les différents modèles entre eux sur des tâches génériques (Raisonnement, Code, Question/Réponse, ...) [3].

Dans ce but d'évaluation, de nombreuses métriques existent, mais demeurent imparfaites [13] et la

meilleure évaluation reste à ce jour l'humain. Cependant, l'évaluation humaine reste très coûteuse, longue et peine à juger de la diversité de la génération.

L'évaluation de modèles génératifs spécifiés au monde de la défense, et plus précisément au monde du renseignement, rencontre plusieurs contraintes supplémentaires. La première est l'impossibilité de comparer les générations à une vérité terrain issue d'un jeu d'évaluation. En effet, les jeux de données ne sont pas publics voire classifiés donc non-accessibles.

La solution adoptée est la production de données vraisemblables par un expert opérationnel, ce qui reste très coûteux dû au niveau de formation et au niveau d'expérience de ce dernier. De plus, pour obtenir un jeu de données annoté de façon objective, il est souhaitable d'avoir un nombre conséquent d'annotateurs ce qui rend très compliqué cette production de données dans un cas réel. Les jeux de données sont donc bien souvent biaisés et peuvent ne pas être réellement représentatifs et exhaustifs des situations possibles.

Nous étudions ici le cas des Demandes d'Informations (RFI - Request For Information) [1], qui sont des questions posées en français ou en anglais par des opérationnels en mission sur tous les sujets propres au renseignement, telles que la position de troupes, d'infrastructures ou plus généralement de ressources.

La production en masse de RFI synthétiques sert notamment de base pour l'entraînement de moteurs de recommandations [9] utilisés pour synthétiser les RFIs similaires, réduisant la redondance dans les demandes à remonter au renseignement. Ces jeux de données produits sont exploités pour améliorer les modèles. Cependant, une démarche d'évaluation de ces jeux de données produits n'est à ce jour pas effectuée et demeure un problème ouvert.

Nous parlerons dans ce vision paper de trois axes d'évaluation des RFIs générées. Nous proposerons aussi une combinaison de métriques afin d'améliorer l'existant dans le domaine de la défense et dans cette tâche. Le premier axe concerne l'évaluation de la linguistique de la génération.Le deuxième axe porte sur une discussion sur les critères opérationnels dans le domaine des RFIs pour une génération à l'aide de modèles larges de langue. Le dernier axe est dédié aux connaissances et à la vraisemblance de la question à vérifier sur une génération.

## 2 Cas d'étude : Évaluation de la Génération de Demandes d'Informations

Les Demandes d'Informations, ou RFI en anglais, sont un cas d'utilisation particulier du renseignement qui consiste à diffuser de l'information au sein de ses effectifs. Un opérationnel sur le terrain pose une question à sa hiérarchie opérationnelle sur un sujet en rapport avec les opérations en cours sur sa localisation d'affectation. Il demande des précisions sur un élément spécifique lié aux événements en cours.

Plus particulièrement, nous nous intéressons aux champs « sujet »(Subject) et demande de « détails »(Needs Details) de la Demande d'Informations. Le premier champ porte sur le sujet général de la demande (ex : Système Anti-Missile à Bakhmut) et le champ needs details porte sur la question posée (ex : Les Systèmes Anti-Missiles russes à Bakhmut sont-ils fonctionnels ?). Une particularité à noter est que le champ needs details dépend forcément du champ sujet et se doit d'être cohérent.

1. https://www.intelligence101.com/an-introduction-to-the-intelligence-cycle/

Ces questions doivent respecter un formalisme et quelques critères, détaillés dans la section suivante. Ces critères rendent nécessaire l'analyse des aspects métier à une évaluation de la pertinence d'une RFI. A l'heure actuelle, plusieurs Demandes d'Informations peuvent porter sur les mêmes sujets et on observe alors une redondance de la question. Chacune de ces questions similaires entraîne alors une séquence d'actions humaines entraînant un travail non négligeable.

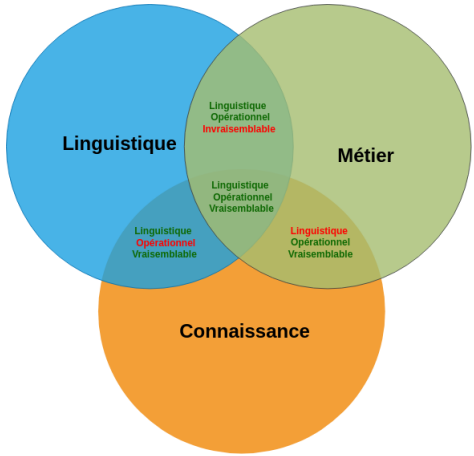


FIGURE 1 – Différents axes à évaluer pour constituer une RFI valide

Dans ce contexte, Fendzi et al [9] propose de grouper les RFIs similaires à l'aide d'un moteur de similarité entraîné sur des jeux de données synthétiques générés à l'aide du modèle GPT2 [27], entraîné sur une base de données d'environ 300 RFIs manuellement écrites par un opérationnel. Cependant, l'évaluation dans ce papier se focalise sur la robustesse du moteur de recommandation et éclipse la mesure de la qualité des contenus générés. Nous proposons ici d'apporter une discussion critique sur les méthodes d'évaluation possibles pour cette génération à travers trois axes : linguistique, métier et connaissances.

Le premier axe vise à évaluer si la génération est bien formée au niveau de la langue au niveau lexical, grammatical, de la diversité et du point de vue de la cohérence entre le Subject et le Needs Details. Le deuxième axe vise à vérifier si les pré-requis d'une RFI sont validés d'un point de vue opérationnel à travers une série de règles issues du monde opérationnel. Le but du dernier axe est de s'assurer que les connaissances évoquées dans une RFI sont vraisemblables à travers la construction d'une base de connaissances et l'utilisation de grands modèles de langues.

Ces trois axes sont illustrés dans la Figure 1, laquelle montre les différentes caractéristiques représentées par les trois axes ainsi que leur nécessité.

## 2.1 Axe Linguistique

L'axe linguistique vise à évaluer la qualité intrinsèque du texte. Cette qualité est liée à la grammaticalité, l'intelligibilité du texte ou encore sa diversité. Dans cet exercice d'évaluation, deux axes principaux existent : l'évaluation automatique et l'évaluation humaine.

### 2.1.1 Évaluation Humaine

Le plus souvent, les évaluations humaines sont effectuées à l'aide d'échelles de Likert ou d'échelles RankMe [35]. Le choix de l'échelle est déterminant, notamment dans le nombre de choix - pair ou impair - qui laisse la possibilité du choix "Neutre", forçant (ou non) une prise de décision de la part de l'évaluateur. Une échelle de Likert posant la question "La génération est-elle bien formée ?" en 5 choix (Pas du tout d'accord - Tout à fait d'accord) est une première tentative d'évaluation sur cette thématique.

L'évaluation humaine est en général considérée comme la meilleure évaluation même si l'humain a du mal à évaluer la diversité [13]. De plus, l'accord inter-annotateur est reconnu comme étant assez faible en génération de texte [32], il est donc important d'avoir un certain nombre d'évaluateurs de tout milieu, âge et genre, ce qui permet de minimiser les biais d'évaluation pour effectuer ce travail. Enfin, cette méthode d'évaluation a pour défaut d'être particulièrement coûteuse et longue à effectuer, limitant son utilisation à l'échelle.

### 2.1.2 Évaluation Automatique

Pour résoudre ce problème, de nombreuses métriques automatiques sont développées. La cible pour une métrique automatique est d'obtenir un score de corrélation élevé avec une expertise humaine sur la même tâche, mais de manière automatique, afin de réduire les coûts humains et en temps.

Les métriques automatiques les plus courantes sont basées sur le recouvrement de n-grammes comme BLEU [26], ROUGE [22] ou encore METEOR [1]. Cependant, ces métriques échouent à évaluer la paraphrase, qui est particulièrement présente en génération de texte appliquée aux RFIs.

La qualité peut aussi être évaluée en utilisant des modèles externes. La Perplexité [15] est notamment utilisée pour mesurer l'incertitude d'un modèle sur sa génération, c'est-à-dire la mesure d'à quel point un modèle est « surpris » quand présenté face aux nouvelles étapes de génération. Malheureusement, la perplexité possède ses propres faiblesses et n'est pas assez représentative de la qualité de génération [29]. Cependant, celle-ci est intéressante lorsque comparée à la perplexité générée par un humain lors d'écriture manuelle de Demandes d'Informations.

Dans la même idée, des modèles externes de régression sont utilisés et ont pour but de scorer la génération notamment sur sa grammaticalité [5], vérifiant que la partie générée est bien construite grammaticalement. Néanmoins, les RFIs sont souvent rédigées par des personnels non-natifs de la langue du pays de provenance de la RFI - par exemple un personnel français qui rédige en anglais - ou tout simplement qui ont des niveaux de langage différent (hétérogénéité des niveaux d'expression écrite) pour un texte écrit en français par un natif. La grammaticalité stricte n'est alors pas représentative de la qualité de la RFI, les usages grammaticaux et orthographiques réels ne respectant pas systématiquement la règle. Néanmoins, la grammaticalité permet de discriminer et de filtrer les générations complètement insensés. Dernièrement, de nouvelles méthodes utilisent des grands modèles de langue comme évaluateurs [19, 4], montrant une corrélation assez élévée avec ce qu'un évaluatueur humain peut produire sur certaines métriques. Dans le cas de Prometheus-2 [19], le processus est disponible en source ouverte, important dans un contexte de défense et les métriques peuvent être personnalisées. Cependant, ces modèles doivent être adaptés au français et au vocabulaire spécifique des RFIs pour être utilisés dans notre cas d'usage.

Dans notre besoin d'évaluer la cohérence entre le « subject »et la partie « needs details », des méthodes

comme BertScore [36] ou encore YiSi [23] utilisent des modèles de plongement tels que Bert [6] ou Word2Vec [24] et permettent de mesurer la distance entre les plongements de la partie Subject et les plongements de la partie Needs Details en vérifiant si les deux parties traitent du même sujet.

Pour générer un jeu de données synthétiques, il est nécessaire d'avoir des contenus avec une haute diversité permettant de refléter les différentes façons d'écrire, les différents champs lexicaux et sémantiques. Pour évaluer de manière automatique la diversité, principalement deux métriques sont utilisées dans la littérature. Distinct-n [21] évalue la diversité intra-génération en comptant le nombre moyen de répétitions des n-grammes et SELF-BLEU [38] permet de mesurer la diversité inter-génération en évaluant le score BLEU entre chaque paire de contenus générées. De manière similaire, il est possible, en utilisant les métriques BertScore ou YiSi, d'obtenir une métrique analogue à SELF-BLEU pour la diversité sémantique, que l'on appellera SELF-BertScore.

Illustré dans la figure 2, nous recommandons l'utilisation d'une combinaison de métriques automatiques dont le BertScore, SELF-BLEU, Distinct-n et l'utilisation de modèles de langue dédiés à l'évaluation de la langue. Cette combinaison reflète la diversité et la justesse linguistique d'une génération. De plus, deux types de métriques automatiques apparaissent. Le premier type rassemble les métriques intra-corpus (SELF-BLEU, SELF-BertScore ..) et le deuxième type adresse les évaluations unitaires de chaque RFI.

Évaluer les caractéristiques intrinsèques de qualité et de diversité est nécessaire mais ne reflète pas l'ensemble des attendus d'un domaine spécifique tel que la Défense. Ce domaine implique certaines connaissances opérationnelles ou connaissances métiers lors d'applications à certaines tâches. Pour des tâches spécifiques telles que la génération de RFI, nous avons besoin de spécifier quels sont les critères opérationnels devant être mesurés et potentiellement automatisés.

## 2.2 Axe Métier

Pour s'assurer de la vraisemblance d'un point de vue opérationnel d'une Demande d'Informations, il est nécessaire d'établir un certain nombre de critères opérationnels permettant de discriminer les générations.

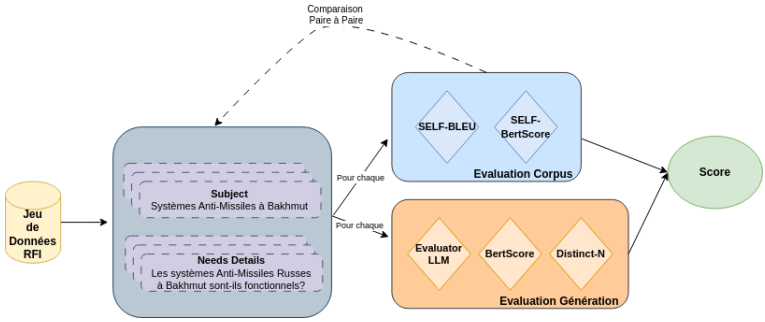


FIGURE 2 – Évaluation Linguistique Proposée de la base de données RFI

### 2.2.1 Critères d'évaluation

Ces règles, utilisées dans le monde du Renseignement, spécifient qu'une RFI est une question atomique (unique) et doît être précise. De manière intuitive, la précision d'une question est proportionnelle à sa difficulté théorique. Par exemple, la question "La vie a-t'elle 5 millions d'années ?" est très vague, très complexe à répondre. Une question plus précise, "Le président de la République a-t-il mangé au restaurant La Fayette la semaine dernière ?" est théoriquement plus simple à répondre, sous réserve d'avoir accès aux ressources. Nous commençons par nous interroger sur la manière de mesurer ce critère de précision par le biais d'une évaluation humaine ou automatique.

### 2.2.2 Evaluation Opérationnelle

Tout d'abord, une évaluation sous la forme d'une échelle de Likert semble pertinente pour cette évaluation. Cette échelle contient des affirmations et attend pour réponse une gradation de l'accord de l'évaluateur. Par exemple, de « Pas du tout d'accord »à « Tout à fait d'accord ». Cependant, le nombre d'opérationnels disponibles étant très restreint et leur évaluation de jeux de données étant très coûteuse, le nombre potentiel d'évaluateurs est de manière conséquente réduit. L'accord inter-évaluateur sur de la génération de texte étant connu pour être très faible [32], il faut s'attendre à ce que cela soit le cas pour une évaluation de notre cas d'usage. Ainsi, les résultats sont parfois non-représentatifs du cas général car subjectifs au vu de la non-représentativité en nombre et diversité de compétences du panel évaluateur.

L'évaluation de la précision est automatisable avec une démarche analogue à [16], en entraînant un classifieur dont l'objectif est d'attribuer un score de précision de la question. Ainsi, se baser sur des jeux d'entraînement tels que fait par Jeong et al [16] serait un progrès vers la production d'une métrique efficace évaluant la précision et le niveau de détail d'une question.

De manière supplémentaire, la question posée ne doit pas être rhétorique, c'est-à-dire qu'elle ne doit pas traiter un sujet dont la réponse est évidente (ex : « Est-il dangereux de se rendre dans une zone du front en Ukraine le 10 juillet 2024 ? »), ce qui nous interroge sur la *pertinence opérationnelle* d'une question. Cette pertinence est définie comme la probabilité d'attribution de la génération d'une RFI à un opérationnel. En d'autres mots, une question opérationnellement pertinente est une demande apportant un gain d'information à un opérationnel.

| **La question...** | **... est précise** | **... n'est pas rhétorique** | **... est pertinente** |
|---|---|---|---|
| Le président de la République a-t'il mangé au restaurant ... | Tout à fait d'accord | Tout à fait d'accord | Plutôt d'accord |
| Est-il dangereux de se rendre dans une zone du front en Ukraine le 10 juillet 2024 ? | Plutôt d'accord | Pas du tout d'accord | Plutôt pas d'accord |

TABLE 1 – Exemple d'évaluation humaine de l'axe métier des RFIs pour le champs Needs Details.

Ainsi, une affirmation dans une échelle de Likert donne « La RFI est pertinente ». En l'absence de jeux de données de questions annotées s'approchant de cette tâche, il est compliqué d'automatiser cette démarche, l'entraînement de modèles étant rendu complexe. Nous recommandons donc l'utilisation d'un opérationnel expert humain pour évaluer de façon la plus juste cette partie, comme présenté dans la table 1.

Dans l'exemple de question précédemment donné, la connaissance de la situation en Ukraine permet

de juger de la nature rhétorique de la question. Ainsi, cette connaissance opérationnelle est un élément immédiatement discriminant de la qualité d'une RFI. De plus, une question du type « Le porte-avions Charles De Gaulle était-il présent lors du sommet des ministres le 24 mai 2023 à Milan ? » permet de soulever deux interrogations sur la vraisemblance de la question. La première porte sur la nature même du bâtiment Charles De Gaulle qui est un porte-avions, qui ne peut être présent à Milan, au coeur des terres italiennes. La deuxième concerne la présence d'un sommet des ministres à cette date précise. Cette problématique multi-niveaux permet d'argumenter la nécessité d'un troisième axe d'évaluation qui est donc l'axe concernant les connaissances évoquées lors de la génération d'une RFI.

## 2.3 Axe Connaissance

La vérification de la vraisemblance des connaissances évoquée est une tâche en ébullition ces dernières années au sein de l'IA générative. La vraisemblance fait référence à la notion de vérité, difficile à définir et source de nombreux débats [31, 34]. Dans une logique de simplicité, nous ferons ici référence à la vraisemblance comme une information possible compte tenu de références sources.

### 2.3.1 Hallucinations

Une problématique liée à l'IA générative est la génération de contenus inventés ou faux. Cette problématique, aussi nommée hallucination [17] en regroupe deux types. Le premier type, les **hallucinations intrinsèques**, concerne les générations qui contredisent leur donnée source et sont factuellement fausses par rapport à une référence. Cette catégorie d'hallucination n'est pas forcément problématique pour la génération de RFI, l'opérateur pouvant se tromper. Cependant, il est important de mesurer grâce à des métriques la divergence par rapport à cette référence et statuer sur *à quel point la génération est fausse*. En effet, si cette divergence est trop grande, la question perdra toute crédibilité pour l'opérateur et ne sera donc pas vraisemblables pour un analyste manipulant les demandes d'informations. Ainsi, des métriques inspirées de la démarche de Li et al [20] basées sur de la vérification de connaissance grâce à une base de données externe doivent être envisagées.Il serait alors possible de mesurer cette divergence grâce à ces bases et l'entraînement de modèles de plongement et de langue spécialisés, entraînant une mesure de la **vraisemblance** de la génération.

L'autre type d'hallucination, les **hallucinations extrinsèques** (ou confabulation), est défini comme un contenu qui ne peut pas être statué vrai ou faux par rapport à la base de référence. Ce type d'hallucination pose particulièrement problème ici, car il dépend du jeu de données d'entraînement du modèle et du raisonnement de la génération. En reprenant notre exemple « Le porte-avions Charles De Gaulle était-il présent lors du sommet des ministres le 24 mai 2023 à Milan ? », en partant du principe que les informations relatives au porte-avions, au sommet des ministres ainsi qu'à la ville de Milan soient présents dans le jeu de données. Si cet événement est absent de l'entraînement du modèle, alors l'information reliant nos informations est absente. Il est donc compliqué de statuer immédiatement sur la **vraisemblance** de la question. Il est alors nécessaire de profiter soit d'un expert humain, soit d'un modèle de langue montrant des capacités sur des tâches liées au raisonnement tels que LLama [30] ou encore Mistral [18].

Une technique reconnue pour réduire la quantité d'hallucinations est le RAG (Retrieval-Augmented Generation ou Génération augmentée par la récupération - de connaissances -) [10]. Cette technique permet de renforcer la capacité d'un modèle à générer des éléments factuels. Une base de connaissance

est ajoutée au modèle de génération, qui permet à celui-ci d'augmenter la vraisemblance ou la factualité de la génération. La première étape pour cela consiste à peupler cette base de connaissances.

### 2.3.2 Base de Connaissances

Dans le domaine des RFIs, un expert évaluateur spécialisé dans un domaine spécifique n'a pas connaissance de tous les tenants et aboutissants de chaque région du globe, ce qui rend compliqué de juger de la vraisemblance de chaque RFI présentée. Ainsi, la construction de bases de connaissances regroupant les divers évènements et faits prouvés autour de zones d'intérêts pour la création de RFIs, permet d'avoir un point de comparaison pour juger de la vraisemblance d'une génération. Cependant, plusieurs interrogations concernant la création de cette base peuvent émerger. La première repose sur la possibilité technique d'effectuer cette création due à des opérations passées ou en cours relevant du secret de la défense nationale. La deuxième interrogation porte sur l'assurance que la véracité des informations, portant donc sur la crédibilité de la source.

En se basant sur l'hypothèse de l'obtention d'une base de connaissance adéquate, il est alors possible de se référer à celle-ci pour vérifier la génération. Cependant, dans des bases de données à l'échelle, il est nécessaire d'utiliser des outils de recommandation récupérant les bonnes connaissances à comparer à la génération. Ainsi, des problématiques liées aux domaines de la Recherche d'Information ou encore du RAG [10] permettent de faire émerger des méthodes permettant d'implémenter des moteurs de similarité récupérant l'information correspondante la plus probable dans la base de connaissances. Comparant une requête utilisateur (ici la génération de question) encodée grâce à des modèles de plongement au contenu encodé lui aussi de la base de connaissance, il est possible de scorer la vraisemblance de la génération.

### 2.3.3 Calcul de Score de Vraisemblance

La base de connaissance est encodée grâce à un modèle de plongement et ses éléments sont comparés à la requête (ici la Demande d'Information) grâce à une similarité cosinus, ce qui permet de retrouver les éléments de la base de connaissance qui sont similaires à la génération initiale. Une fois les informations récupérées, elles sont passées au modèle pour estimer si le modèle arrive à trouver une logique entre la question et les éléments récupérés. Utilisant des modèles de langue montrant la capacité à raisonner sur certaines tâches, différentes méthodes [7, 20] mettent en place une chaîne de traitement permettant de vérifier ou assurer la factualité de la génération. Dhuliawala et al [7] réutilise un modèle de langue afin de vérifier les éléments d'une génération en la décomposant en différentes étapes et ainsi vérifier leur factualité de manière itérative. Illustré dans la figure 3, une méthodologie analogue permettrait de vérifier la vraisemblance d'une génération en récupérant les différents éléments dans la base de connaissance grâce à du Retrieval-Augmented Generation, vérifiant la vraisemblance du lien effectué lors de la génération. Néanmoins, ces méthodes sont destinées à évaluer une génération en langue anglaise et leur jeu de données d'entraînement contient une quantité minoritaire de documents français, biaisant le modèle sur celle-ci. Il est alors important de développer ces solutions sur des jeux de données en français avant de les réutiliser pour l'évaluation des RFIs. Ainsi, des pistes utilisant des modèles tels que CroissantLLM [8] sont envisageables.

Cependant, il est important de rappeler que la mise en place de bases de connaissances octroyant une exhaustivité des situations opérationnelles est une tâche non-triviale voire infaisable et que selon le

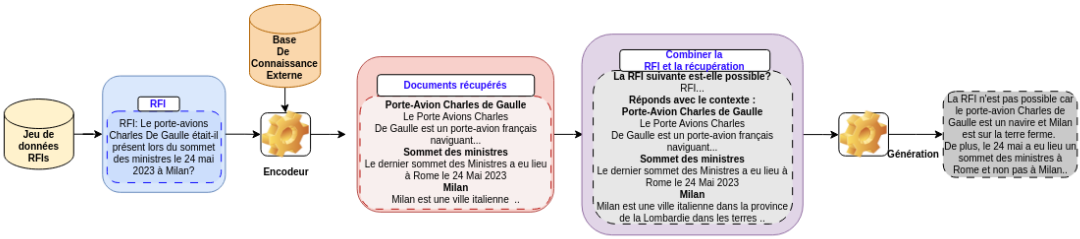


FIGURE 3 – Vérification de la vraisemblance des RFIs à l'aide de modèles externes

cas d'usage et la classification des opérations, la vraisemblance opérationnelle sera donnée par un évaluateur humain.

# 3 Synthèse

L'évaluation des Demandes d'Informations requiert l'étude d'au moins trois axes distincts qui ont été étudiés dans ce vision paper. Le premier axe, l'aspect linguistique, mobilise les connaissances actuelles concernant l'Etat de l'Art en génération de texte. Ainsi, ressortent des métriques telles que le BertScore permettant de vérifier la cohérence entre un Subject et un Needs détails générés. De plus, d'autres métriques telles que le SELF-BLEU ou une version analogue avec le BertScore permettent d'évaluer la diversité sémantique et lexicale au sein du corpus. Au sein d'une génération, Distinct-n permet d'évaluer la répétition d'un modèle selon plusieurs n-grams. Enfin, des métriques issues d'évaluations utilisant de grands modèles de langues peuvent être utilisées afin d'évaluer la bonne construction de la génération.

Le deuxième axe concerne l'aspect métier et opérationnel qui doit être présent au sein d'une génération. Cet axe, permet de spécifier le niveau de détail requis dans la question générée et sa pertinence opérationnelle. Cette pertinence fait aussi référence à la mesure de rhétoricité d'une question.

Enfin, le troisième axe s'intéresse à la présence de connaissances spécifiques dans les RFIs générées et s'assure de la vraisemblance de la génération. Cet aspect s'adresse principalement à la réduction et la détection des hallucinations.

Ainsi, la table 2 résume les différentes métriques recommandées dans ce vision paper afin d'optimiser l'évaluation des Demandes d'Informations générées. Nous remarquons l'absence de métriques automatiques permettant de mesurer la rhétoricité, la vraisemblance ainsi que la pertinence. Nous encourageons la production de recherches allant dans ce sens permettant d'automatiser cette tâche pour l'évaluation des Demandes d'Informations.

# 4 Conclusion

En conclusion, nous avons présenté lors de ce *vision paper* l'évaluation de Demandes d'Informations générées liées au domaine de la Défense. L'évaluation de ces données est une tâche compliquée par la confidentialité des données utilisées et le manque d'experts opérationnels capables d'évaluer les

| Champ d'évaluation | Connaissance | Linguistique | Métier | Références |
|---|---|---|---|---|
| **Linguistique** | - | Evaluation via LLM, Grammaticalité, Perplexité | - | [19, 4, 15, 5] |
| **Précision** | - | - | Classifieur | [16] |
| **Rhétoricité** | - | - | Opérationnel humain | |
| **Vraisemblance** | Evaluation par RAG, Opérationnel humain | - | - | [16, 10, 20] |
| **Pertinence** | - | - | Opérationnel humain | |
| **Cohérence** | - | BertScore | - | [36] |
| **Diversité** | - | Distinct-n, Self-BLEU, SELF-BertScore | - | [21, 38, 36] |

TABLE 2 – Synthèse des différentes métriques recommandées

données générées par un modèle de langue.

Dans ce contexte, nous avons cherché à évaluer un jeu de données de Demandes d'Informations préalablement généré et nous avons proposé une méthodologie d'évaluation afin de juger de la qualité d'une génération. Pour cela, nous avons proposé trois axes distincts nécessaires à l'établissement d'une méthodologie d'évaluation de cette problématique. Ces axes que nous proposons couvrent la qualité linguistique de la génération, le respect des règles métier pour la rédaction de ces demandes ainsi que la vraisemblance de ces dernières par rapport à une base de connaissance de référence. Le premier axe s'intéresse à la qualité intrinsèque du texte généré, que ce soit en terme de diversité inter-générations et intra-génération, de cohérence lors de la génération de différentes parties de la requête ou encore de la qualité grammaticale du texte.

Le deuxième axe porte sur le respect des règles métier observées par les opérationnels pour écrire des Demandes d'Informations. Ces règles (questions précises, non rhétoriques..) permettent de spécifier différents critères d'évaluation de la requête sur l'axe métier.

Le troisième et dernier axe couvre la vraisemblance de la demande. Nous avons notamment proposé différentes pistes s'inspirant du RAG (Retrieval-Augmented Generation - Génération augmentée à l'aide de connaissance récupérée) permettant de vérifier automatiquement cette vraisemblance.

Cette réflexion vise à structurer et rationaliser la mesure de la qualité des contenus générés, en identifiant et outillant chacune des dimensions d'intérêt. Cet outillage, spécifique au cas d'emploi, fournit des caractéristiques des jeux de données générées et rend possible la comparaison de générations entre différentes instances ou modèles de génération. Des travaux futurs permettront de dérouler cette évaluation pour augmenter, de manière itérative, la qualité d'un ensemble de Demandes d'Informations synthétiques, à exploiter par la suite pour la mise au point d'algorithmes et de systèmes.

# Remerciements

Ce projet a été financé par le Programme Européen pour l'Industrie de la Défense (EDIDP) sous l'accord de subvention No : EDIDP-AI-2020-066-AI4DEF.

# Références


[1] Satanjeev Banerjee and Alon Lavie. METEOR : An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[2] Alessandra Cervone, Chandra Khatri, Rahul Goel, Behnam Hedayatnia, Anu Venkatesh, Dilek Hakkani-Tur, and Raefer Gabriel. Natural Language Generation at Scale : A Case Study for Open Domain Question Answering, September 2019. arXiv :1903.08097 [cs].

[3] Yiran Chen, Zhenqiao Song, Xianze Wu, Danqing Wang, Jingjing Xu, Jiaze Chen, Hao Zhou, and Lei Li. MTG : A Benchmark Suite for Multilingual Text Generation. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics : NAACL 2022*, pages 2508–2527, Seattle, United States, July 2022. Association for Computational Linguistics.

[4] Cheng-Han Chiang and Hung-yi Lee. Can Large Language Models Be an Alternative to Human Evaluation ? 2023.

[5] Davide Colla, Matteo Delsanto, and Elisa Di Nuovo. EliCoDe at MultiGED2023 : fine-tuning XLM-RoBERTa for multilingual grammatical error detection. *Swedish Language Technology Conference and NLP4CALL*, pages 24–34, May 2023.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. arXiv :1810.04805 [cs].

[7] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-Verification Reduces Hallucination in Large Language Models, September 2023. Issue : arXiv :2309.11495 arXiv :2309.11495 [cs].

[8] Manuel Faysse, Patrick Fernandes, Nuno M. Guerreiro, António Loison, Duarte M. Alves, Caio Corro, Nicolas Boizard, João Alves, Ricardo Rei, Pedro H. Martins, Antoni Bigata Casademunt, François Yvon, André F. T. Martins, Gautier Viaud, Céline Hudelot, and Pierre Colombo. CroissantLLM : A Truly Bilingual French-English Language Model, March 2024. arXiv :2402.00786 [cs].

[9] Claude Fendzi, Bruno Carron, and Guillaume Gadek. AI-based Text Generation for Semantic Search Robustness : Application to Defence. 2023.

[10] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-Augmented Generation for Large Language Models : A Survey, March 2024. arXiv :2312.10997 [cs].

[11] Cristina Garbacea and Qiaozhu Mei. Why is constrained neural language generation particularly challenging ?, June 2022. Issue : arXiv :2206.05395 Issue : arXiv :2206.05395 arXiv :2206.05395 [cs].

[12] Albert Gu and Tri Dao. Mamba : Linear-Time Sequence Modeling with Selective State Spaces, May 2024. arXiv :2312.00752 [cs].

[13] Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. Unifying Human and Statistical Evaluation for Natural Language Generation, April 2019. Issue : arXiv :1904.02792 Issue : arXiv :1904.02792 arXiv :1904.02792 [cs, stat].

[14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA : Low-Rank Adaptation of Large Language Models, October 2021. Issue : arXiv :2106.09685 Issue : arXiv :2106.09685 arXiv :2106.09685 [cs].

[15] Frederick Jelinek, Robert L. Mercer, Lalit R. Bahl, and Janet M. Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *Journal of the Acoustical Society of America*, 62, 1977.

[16] Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. Adaptive-RAG : Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity, March 2024. arXiv :2403.14403 [cs].

[17] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Wenliang Dai, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12) :1–38, December 2023. arXiv :2202.03629 [cs].

[18] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, October 2023. arXiv :2310.06825 [cs].

[19] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2 : An Open Source Language Model Specialized in Evaluating Other Language Models, May 2024. arXiv :2405.01535 [cs].

[20] Jiarui Li, Ye Yuan, and Zehua Zhang. Enhancing LLM Factual Accuracy with RAG to Counter Hallucinations : A Case Study on Domain-Specific Queries in Private Knowledge-Bases, March 2024. arXiv :2403.10446 [cs].

[21] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A Diversity-Promoting Objective Function for Neural Conversation Models, June 2016. Issue : arXiv :1510.03055 arXiv :1510.03055 [cs].

[22] Chin-Yew Lin. ROUGE : A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[23] Chi-kiu Lo. YiSi - a Unified Semantic MT Quality Evaluation and Estimation Metric for Languages with Different Levels of Available Resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2 : Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy, 2019. Association for Computational Linguistics.

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space, September 2013. arXiv :1301.3781 [cs].

[25] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, March 2022. Issue : arXiv :2203.02155 Issue : arXiv :2203.02155 arXiv :2203.02155 [cs].

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for*

*Computational Linguistics - ACL '02*, page 311, Philadelphia, Pennsylvania, 2001. Association for Computational Linguistics.

[27] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. 2019.

[28] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, July 2020. Issue : arXiv :1910.10683 Issue : arXiv :1910.10683 arXiv :1910.10683 [cs, stat].

[29] Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models, April 2016. Issue : arXiv :1511.01844 arXiv :1511.01844 [cs, stat].

[30] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA : Open and Efficient Foundation Language Models, February 2023. Issue : arXiv :2302.13971 Issue : arXiv :2302.13971 arXiv :2302.13971 [cs].

[31] Werner Ulrich. A Philosophical Staircase for Information Systems Definition, Design and Development. The Journal of Information Technology Theory and Application (JITTA) :55–84, 2001.

[32] Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. Human evaluation of automatically generated text : Current trends and best practice guidelines. *Computer Speech & Language*, 67 :101151, May 2021.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, December 2017. Issue : arXiv :1706.03762 Issue : arXiv :1706.03762 arXiv :1706.03762 [cs].

[34] Arjan Vreeken. Notions of Information : A Review of Literature. 2002.

[35] Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. A Survey of Controllable Text Generation using Transformer-based Pre-trained Language Models. January 2022. arXiv : 2201.05337 Publisher : arXiv.

[36] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore : Evaluating Text Generation with BERT, February 2020. Issue : arXiv :1904.09675 Issue : arXiv :1904.09675 arXiv :1904.09675 [cs].

[37] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-Following Evaluation for Large Language Models, November 2023. Issue : arXiv :2311.07911 arXiv :2311.07911 [cs].

[38] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen : A Benchmarking Platform for Text Generation Models, February 2018. Issue : arXiv :1802.01886 arXiv :1802.01886 [cs].