SEval-EX : Un paradigme basé sur les phrases atomiques pour une évaluation explicable de la qualité des résumés

Tanguy Herserant[1] Vincent Guigue[1]
(1) AgroParisTech - MIA, 22 place de l'Agronomie, 91120 Palaiseau, France
{tanguy.herserant, vincent.guigue}@agroparistech.fr

RÉSUMÉ

L'évaluation de la qualité des résumés de texte demeure un défi critique en Traitement Automatique du Langage Naturel. Les approches actuelles font face à un compromis entre performance et interprétabilité. Nous présentons SEval-Ex, un framework qui comble cette lacune en décomposant l'évaluation des résumés en phrases atomiques, permettant à la fois une haute performance et une explicabilité. SEval-Ex emploie un pipeline en deux étapes : extraction des phrases atomiques à partir du texte source et du résumé via un LLM, puis mise en correspondance de ces phrases. Contrairement aux approches existantes qui ne fournissent que des scores globaux, notre méthode génère un parcours détaillé des décisions grâce à un alignement entre les phrases. Les expériences sur SummEval démontrent que SEval-Ex atteint des performances état de l'art avec une corrélation de 0.580 sur la cohérence avec les jugements humains, surpassant GPT-4 (0.521) tout en maintenant l'interprétabilité et la robustesse contre l'hallucination.

ABSTRACT

SEval-Ex : A Statement-Level Framework for Explainable Summarization Evaluation

Evaluating text summarization quality remains a critical challenge in Natural Language Processing. Current approaches struggle with the trade-off between performance and interpretability. We introduce SEval-Ex, a framework addressing this gap by breaking down summary evaluation into atomic statements, enabling both high performance and explainability. SEval-Ex employs a two-step pipeline : extracting atomic statements from the source text and summary using an LLM, followed by statement alignment. Unlike existing approaches that only provide summary-level scores, our method produces a detailed decision trace via phrase-level alignments. Experiments on the SummEval benchmark demonstrate that SEval-Ex achieves state-of-the-art performance with a 0.580 correlation on coherence with human judgments, outperforming GPT-4 evaluators (0.521) while maintaining interpretability and robustness against hallucination.

MOTS-CLÉS : Évaluation de résumés, Explicabilité, Traitement Automatique du Langage.

KEYWORDS: Summarization evaluation, Explainability, Natural Language Processing.

ARTICLE : **Accepté à** PAKDD 2025.

1 Introduction

L'évaluation de la génération de texte est devenue un défi critique en Traitement Automatique du Langage Naturel (TALN), particulièrement avec les Grands Modèles de Langage (LLMs) qui révolutionnent notre capacité à générer du texte semblable à celui produit par les humains [2, 19, 10]. Bien que ces avancées aient permis une fluidité sans précédent dans la génération de texte, elles ont également mis en évidence un défi fondamental : comment pouvons-nous évaluer de manière fiable la cohérence factuelle du contenu généré tout en maintenant l'interprétabilité de nos méthodes d'évaluation ?

Les approches traditionnelles d'extraction d'information, particulièrement la Reconnaissance d'Entités Nommées (NER), ont montré des limitations significatives dans des contextes spécialisés, échouant souvent à atteindre des performances satisfaisantes à travers divers domaines. Bien que le NER soit performant dans l'identification des faits explicites, sa structure peine à gérer les nuances d'expression propres au langage naturel. Parallèlement, les métriques sémantiques opérant dans l'espace latent, comme BERTScore [21], offrent des possibilités intéressantes mais font face à une ambiguïté cruciale : mesurent-elles vraiment l'exactitude factuelle, ou capturent-elles principalement la fluidité linguistique ?

Exploiter les capacités avancées des LLMs présente une voie interessante pour répondre à ces défis. Les LLMs démontrent un haut niveau de compétence en reformulation de texte et en compréhension [18]. Nous proposons d'utiliser les LLMs pour l'extraction d'information à travers une approche de reformulation textuelle, contournant ainsi les complexités et les erreurs potentielles associées à la transformation du texte en formats de données structurées. Cette méthodologie capitalise sur la compréhension inhérente du langage des LLMs pour identifier et aligner les unités fondamentales de connaissance, que nous appelons phrases atomiques, au sein du texte. Notre approche s'inspire de méthodologies récentes telles que RAGAS [4], mais est spécifiquement adaptée pour relever les défis techniques associés à l'évaluation de textes plus longs nécessitant une analyse détaillée.

Nous présentons SEval-Ex, un cadre novateur qui décompose l'évaluation des résumés en phrases atomiques, permettant à la fois une haute performance et une évaluation interprétable de la cohérence factuelle. Notre approche apporte trois contributions techniques clés :

- Une **méthodologie d'Extraction des phrases atomiques** qui utilise les LLMs pour décomposer les textes source et les résumés en phrases atomiques
- Un **Pipeline de Raisonnement pour le Verdict :** Nous concevons un pipeline qui fournit des décisions d'évaluation interprétables en faisant correspondre et en classifiant les énoncés atomiques en Vrais Positifs (TP), Faux Positifs (FP) et Faux Négatifs (FN)
- Un **protocole d'évaluation complet** qui démontre la robustesse à travers diverses formes d'hallucination (au niveau des entités, des événements et des détails) validé par une expérimentation approfondie

Pour valider l'efficacité de notre approche, nous avons mené des tests rigoureux en utilisant le benchmark SummEval [5]. Nous montrons que SEval-Ex atteint un coefficient de corrélation de Spearman état de l'art de 0.580 avec les jugements humains de cohérence, surpassant les évaluateurs basés sur GPT-4 (0.521). Il est important de noter que notre cadre maintient une interprétabilité complète, une caractéristique critique pour les applications où la compréhension du raisonnement derrière les décisions d'évaluation est essentielle. SEval-Ex démontre une robustesse particulière sur divers types d'incohérences factuelles (hallucinations), soulignant sa valeur pour les applications qui exigent une haute fiabilité et transparence dans l'évaluation du contenu.

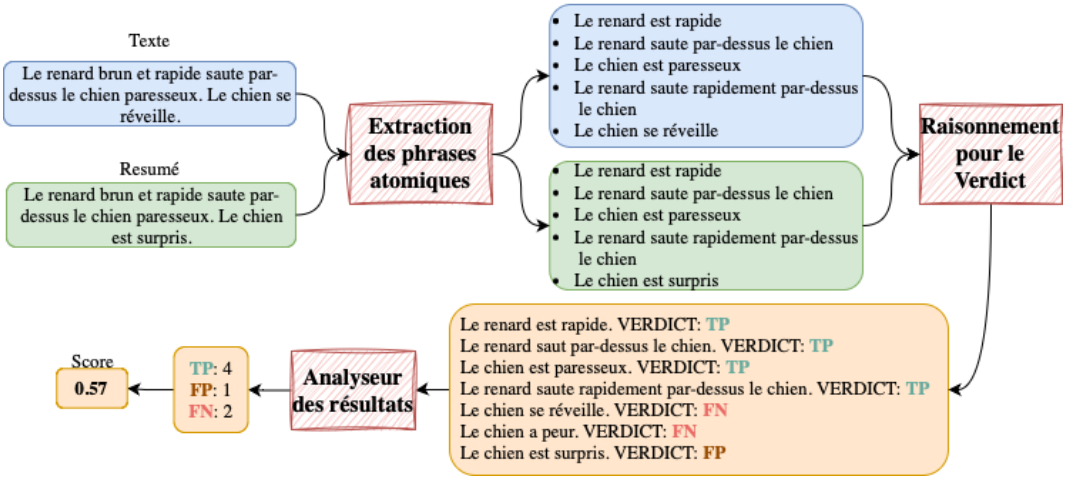

FIGURE 1 – Pipeline d'évaluation SEval-Ex. Lors la première phase un LLM réalise une (1) Extraction des phrases atomiques, puis durant la phase (2) de Raisonnement pour le Verdict, un LLM étiquette les énoncés. Enfin, un (3) analyseur des résultats extrait la matrice de confusion qui constitue le score.

2 Travaux Connexes

L'évaluation des systèmes de résumé a considérablement évolué, passant des techniques rudimentaires de correspondance lexicale aux modèles neuronaux avancés. Dans cette section, nous structurons notre analyse autour de cinq paradigmes clés : *les métriques traditionnelles de chevauchement lexical*, *les approches basées sur les embeddings*, *les cadres d'évaluation spécifiques à la tâche*, *les méthodes fondées sur l'inférence en langage naturel (NLI)* et *les évaluateurs basés sur les modèles de grande taille (LLM)*.

Métriques traditionnelles de chevauchement lexical

Les premières métriques utilisées pour l'évaluation des résumés, telles que ROUGE [13] et BLEU [16], reposaient principalement sur la correspondance textuelle en mesurant le chevauchement des *n*-grammes entre les résumés générés et les textes de référence. Bien que ces métriques soient computationnellement efficaces et offrent une interprétabilité intuitive grâce au chevauchement lexical observable, elles présentent des limites importantes. Notamment, elles ne prennent pas en compte les paraphrases ni la synonymie, peinent à capturer une équivalence sémantique plus profonde et dépendent fortement de la correspondance exacte des mots.

Approches basées sur les embeddings

L'introduction des embeddings contextuels, tels que ceux de BERT [3], a permis le développement de métriques d'évaluation plus avancées, capables de capturer la similarité sémantique au-delà du simple chevauchement lexical. Par exemple, BERTScore [21] évalue la similarité entre résumés en utilisant des représentations de mots dans un espace vectoriel, tandis que MoverScore [22] exploite la distance

de Earth Mover pour comparer les distributions d'embeddings des textes candidats et de référence.

Cependant, ces approches présentent aussi des limites. Comme souligné par Hanna et al. [8], BERTScore peut manquer de sensibilité face aux nuances sémantiques et donner des scores trompeusement élevés à des textes lexicalement proches mais sémantiquement divergents. Cette limitation nuit à la précision de l'évaluation, notamment lorsque des erreurs factuelles sont présentes mais difficiles à détecter via une simple analyse des embeddings.

Cadres d'évaluation spécifiques à la tâche

Pour pallier les limites des méthodes génériques, des cadres spécialisés ont été développés afin d'évaluer des aspects spécifiques des résumés, tels que la cohérence factuelle et la qualité globale. Par exemple, FactCC [11] utilise des modèles d'inférence en langage naturel (NLI) pour identifier les incohérences factuelles entre un résumé et son texte source. De leur côté, QAGS [20] et QuestEval [17] s'appuient sur des techniques de question-réponse afin de mesurer l'exactitude factuelle en générant et en répondant à des questions dérivées des résumés et des documents sources.

Les cadres d'évaluation multi-aspects tels que UniEval [23] et SUPERT [7] offrent une évaluation plus complète en intégrant des critères tels que la fluidité, la cohérence et le caractère informatif. Toutefois, ces approches sont souvent limitées par leur coût computationnel élevé, une spécialisation à des domaines spécifiques réduisant leur généralisabilité, ainsi que des défis en matière d'interprétabilité, ce qui limite leur adoption généralisée.

Approches basées sur l'inférence en langage naturel (NLI)

Les méthodes fondées sur l'inférence en langage naturel appliquent des modèles d'implication logique afin d'évaluer la cohérence factuelle des résumés. Par exemple, SummaC [12] mesure la capacité d'un résumé à être inféré à partir du texte source en calculant des scores de probabilité d'implication entre paires de phrases.

Bien que prometteuses pour la détection des inexactitudes factuelles, ces approches rencontrent des défis majeurs, notamment en termes de passage à l'échelle. Le calcul des relations d'implication entre phrases engendre une complexité combinatoire qui se traduit par un coût computationnel élevé, particulièrement pour les textes longs. Liu et al. [14] ont récemment proposé un raisonnement contextuel pour atténuer ces limitations, mais des obstacles persistent quant à l'adaptabilité de ces modèles à différents domaines et à leur capacité à traiter des textes de grande longueur avec précision.

Évaluateurs basés sur les modèles de grande taille (LLM)

L'essor des modèles de grande taille (LLM) a conduit à l'émergence de nouvelles méthodes d'évaluation, exploitant des modèles comme GPT-3 [2] et GPT-4. Des approches telles que GPTScore [6] et G-Eval [15] utilisent ces modèles pour générer des scores d'évaluation ou effectuer des comparaisons entre différents résumés.

Ces méthodes bénéficient de la compréhension approfondie du langage acquise par les LLM au cours de leur pré-entraînement massif. Toutefois, elles sont confrontées à plusieurs limitations, notamment

leur nature de *boîte noire*, ce qui réduit leur transparence et leur interprétabilité. De plus, les coûts computationnels élevés rendent difficile leur application à grande échelle ou en temps réel.

Enfin, la dépendance à des modèles propriétaires pose des problèmes de reproductibilité et d'accessibilité [1], soulevant des préoccupations quant à leur viabilité pour une évaluation fiable et ouverte des systèmes de résumé.

3 Méthodologie

Nous introduisons **SEval-Ex**, un cadre d'évaluation des résumés basé sur des phrases atomiques — unités d'information autonomes — permettant une analyse entièrement interprétable. En mettant l'accent sur la transparence à chaque étape, SEval-Ex offre aux utilisateurs la possibilité de retracer l'ensemble du processus d'évaluation, depuis l'extraction des phrases atomiques jusqu'à l'attribution du score final. Contrairement aux approches de type *boîte noire*, qui se limitent à fournir un score global au niveau du résumé, notre méthode apporte une granularité d'analyse en s'appuyant sur trois étapes distinctes et explicites :

1. **Extraction des phrases atomiques** : interprétation des textes sources et des résumés sous forme de phrases atomiques ;
2. **Raisonnement pour le verdict** : alignement et classification des phrases atomiques en catégories (Vrais Positifs, Faux Positifs, Faux Négatifs) ;
3. **Analyse des résultats** : calcul de métriques d'évaluation basées sur ces alignements.

Afin d'améliorer l'efficacité et de préserver le contexte, la vérification des phrases atomiques est directement intégrée dans les *prompts* du modèle de langage (LLM), ce qui évite une comparaison exhaustive par paires. Cette approche garantit à la fois l'évolutivité et l'interprétabilité. La Figure 1 illustre l'architecture du pipeline et ses principales fonctionnalités à travers un exemple concret.

3.1 Formalisation du problème

Étant donné un document source D et un résumé généré S, l'objectif est d'évaluer la qualité du résumé via une analyse fondée sur les phrases atomiques. Définissons :

— $\mathcal{S}$: l'ensemble des phrases atomiques possibles ;

— $\mathcal{Y} = \{\text{TP}, \text{FP}, \text{FN}\}$: l'espace des catégories d'évaluation (Vrai Positif, Faux Positif, Faux Négatif).

Les trois fonctions fondamentales du cadre SEval-Ex sont définies comme suit :

Extraction des phrases atomiques : La fonction $E(\cdot)$ segmente un texte en un ensemble des phrases atomiques :

$$E(D) = \{d_1, d_2, \ldots, d_n\} \quad \text{où } d_i \in \mathcal{S}, \text{ énoncé extrait de } D, \tag{1}$$

$$E(S) = \{s_1, s_2, \ldots, s_m\} \quad \text{où } s_j \in \mathcal{S}, \text{ énoncé extrait de } S. \tag{2}$$

Vérification des phrases atomiques : La fonction $V(\cdot,\cdot)$ détermine les relations entre les phrases atomiques extraites :

$$V : E(D) \times E(S) \to \mathcal{Y}. \tag{3}$$

Calcul du score : La fonction $F(\cdot,\cdot)$ génère la métrique d'évaluation :

$$F(D,S) = \mathrm{F}_1(V(E(D),E(S))) \to \mathbb{R}. \tag{4}$$

3.2 Extraction des phrases atomiques

L'extraction des phrases atomiques repose sur des LLM quantifiés et sur des instructions optimisées à travers une série d'expérimentations. L'objectif est d'identifier des unités informationnelles minimales tout en préservant leur autonomie sémantique.

3.3 Raisonnement de verdict

La fonction de raisonnement de verdict V classe chaque énoncé extrait selon les règles suivantes :

— **Vrai Positif (TP)** : $s_j = TP$ si $d_i \equiv s_j$.
— **Faux Positif (FP)** : $s_j = FP$ si $\nexists d_i : d_i \equiv s_j$.
— **Faux Négatif (FN)** : $d_i = FN$ si $\nexists s_j : d_i \equiv s_j$.

où $\equiv$ représente l'équivalence sémantique entre les phrases atomiques. Deux phrases atomiques sont considérés comme équivalents lorsqu'ils véhiculent le même fait, indépendamment de différences linguistiques, cette relation étant évaluée par un LLM.

L'objectif est d'évaluer si les informations contenues dans le résumé sont bien supportées par le texte source. Un résumé est jugé factuellement correct si ses phrases atomiques principaux trouvent des correspondances validées dans le document original. Les scores de précision, rappel et F1 sont ensuite calculés pour mesurer la fidélité du résumé.

3.4 Optimisation du pipeline pour les textes longs

L'extraction et la mise en correspondance des phrases atomiques posent deux défis majeurs :

1. **Perte d'information contextuelle** : la segmentation en phrases atomiques peut altérer les relations entre phrases et entraîner une perte de cohérence globale.
2. **Dérive sémantique** : la mise en correspondance des phrases atomiques isolées peut conduire à une mauvaise interprétation de leur sens initial.

Ces défis sont particulièrement marqués pour les textes longs, où la cohérence et la fidélité doivent être maintenues à travers des segments plus vastes. Pour y remédier, nous avons exploré plusieurs stratégies d'amélioration du pipeline de base (Figure 1) :

1. **Optimisation de l'extraction des phrases atomiques** :
   — **Approche basique** : le texte est traité en une seule unité via un unique appel au LLM. Bien que rapide, cette méthode risque de perdre certaines connexions contextuelles essentielles.

— **Approche segmentée (3-Chunk)** : le texte est divisé en segments de trois phrases, permettant une meilleure préservation du contexte local. Cette méthode améliore la robustesse de l'analyse tout en maintenant une charge computationnelle raisonnable.

2. **Optimisation du raisonnement pour le verdict** :
   — **Approche basique** : toutes les phrases atomiques extraites sont mises en correspondance, mais cette approche peut générer une dérive sémantique en raison d'un manque de contexte.
   — **Méthode StSum_Text** : les phrases atomiques du résumé sont directement alignées avec les phrases du texte source, garantissant ainsi une meilleure préservation du contexte durant la vérification.

Après de nombreuses expériences comparatives, nous avons sélectionné le modèle Qwen2.5 :72B (quantifié en 4 bits). L'implémentation complète, incluant les modèles de prompts et le code d'optimisation, sera rendue publique sur Github [1].

4 Expérimentations

Nous évaluons notre approche sur les benchmarks de référence pour l'évaluation des résumés, notamment :

SummEval [5] : un benchmark complet comprenant 1 600 résumés générés par 16 systèmes de résumé différents à partir d'articles de CNN/DailyMail. Chaque résumé est annoté selon quatre dimensions évaluées par des experts humains :

— **Cohérence** : qualité collective de l'ensemble des phrases du résumé. Cette dimension mesure si le résumé est bien structuré et organisé, formant un ensemble d'informations cohérent plutôt qu'une simple juxtaposition de faits reliés.
— **Exactitude factuelle** (*Consistency*) : alignement des faits entre le résumé et le document source. Un résumé factuellement correct ne doit contenir que des affirmations qui découlent du texte d'origine, sans hallucinations ni distorsions d'information.
— **Fluidité** : qualité linguistique des phrases individuelles. Cette dimension évalue l'absence de problèmes de formatage, d'erreurs de capitalisation et de fautes grammaticales (ex. phrases incomplètes, éléments manquants) susceptibles de nuire à la lisibilité.
— **Pertinence** : sélection du contenu important à partir de la source. Un bon résumé doit inclure uniquement les informations essentielles issues du document d'origine.

Notre travail étant centré sur les phrases atomiques, la dimension d'**Exactitude factuelle** est celle qui s'applique le mieux à notre analyse. En effet, comme nous n'intégrons pas de module spécifique à l'évaluation de l'importance des phrases, nous ne pouvons pas mesurer correctement la pertinence des résumés.

4.1 Corrélation avec le jugement humain

Nos expériences initiales (Tableau 1) aident à mieux comprendre les forces et les faiblesses de notre architecture en comparant les variantes présentées dans la Section 3.4. *Préservation du Contexte*

1. https ://github.com/TanguyHsrt/seval-ex

Local : La variante **3-Chunk** préserve le contexte local en traitant des fragments de trois phrases, améliorant la corrélation de cohérence de 0,30 (Base) à 0,39.
Traiter le texte en fragments plus petits et sémantiquement cohérents peut améliorer la précision d'extraction des phrases atomiques car les dépendances et les références sont généralement résolues dans un contexte local. Lors du traitement de longs documents dans leur ensemble, le LLM peut avoir du mal à maintenir la cohérence à travers des sections distantes, manquant potentiellement des indices contextuels clés qui sont plus faciles à capturer dans un champ plus restreint.

Comparaison Directe à la Source : La version **StSum_Text** implémente une comparaison directe à la source, améliorant davantage la corrélation de cohérence à 0,58. Éviter les représentations intermédiaires du document source lors de la comparaison du contenu du résumé réduit la perte d'information car chaque étape de transformation introduit des distorsions potentielles. La comparaison directe préserve les relations sémantiques originales et les nuances contextuelles présentes dans le texte source.

TABLE 1 – Spearman Correlation (F1) des différentes expériences sur SummEval

Métriques	**Base**	**3-Chunk**	**StSum_text**
Fluidité	0.126	0.207	**0.351**
Exactitude factuelle	0.231	0.306	**0.580**
Cohérence	0.165	0.152	**0.264**
Pertience	0.210	0.209	**0.300**

L'augmentation de performance observée de *StSum_Text* (+0,28 de corrélation) par rapport à *Base* peut s'expliquer par l'hypothèse que le contexte local et la comparaison directe à la source sont tous deux importants. Cependant, cette conception spécialisée montre des compromis attendus : tout en atteignant une corrélation de cohérence état de l'art, la performance sur d'autres dimensions (pertinence : 0,30, cohérence : 0,26, fluidité : 0,35) reste modérée. Puisque les résumés dans le jeu de données SummEval sont courts, la différence dans le nombre des phrases atomiques extraites entre le traitement du texte complet et 3-Chunk est négligeable. Par conséquent, nous utiliserons la version **StSum_Text** pour nos analyses ultérieures.

Comparaison avec les approches existantes

Le Tableau 2 révèle plusieurs observations clés sur les approches d'évaluation de résumés :
Métriques Traditionnelles : Les approches basées sur les N-grammes (famille ROUGE) montrent une corrélation systématiquement faible sur toutes les dimensions (0,11-0,19), mais peuvent être interprétables.
Métriques Sémantiques : Malgré des espaces de représentation vectorielle de mots sophistiqués, BERTScore et MOVERScore, qui sont des boîtes noires, atteignent une corrélation de cohérence limitée (0,11-0,16). QuestEval présente une amélioration significative (0,306) mais, basée sur des questions générées et des réponses générées, l'explicabilité est controversée. *Méthodes basées sur les LLM :* Bien que les évaluateurs LLM à usage général comme G-Eval montrent de fortes performances sur toutes les dimensions (0,52-0,58), ils sont coûteux à exécuter et restent des boîtes noires. Notre approche spécialisée atteint une corrélation de cohérence supérieure (0,58 contre 0,52) avec interprétabilité.

Ces résultats valident notre principe de conception clé : en se concentrant spécifiquement sur l'évalua-

TABLE 2 – Comparaison de la corrélation de Spearman entre notre approche et d'autres métriques de résumé sur le jeu de données SummEval

Architecture	Métriques	Fluidité	Exactitude factuelle	Cohérence	Pertience	Moyenne
GPT4	**G-Eval (Best)**	**0.540**	0.521	**0.582**	**0.547**	0.516
GPT3	**GPTScore**	0.403	0.449	0.434	0.381	0.417
n-gramme	**ROUGE-1**	0.115	0.160	0.167	0.326	0.192
	ROUGE-2	0.159	0.187	0.184	0.290	0.205
	ROUGE-L	0.105	0.115	0.128	0.311	0.165
basé sur la représentation vectorielle	**BERTScore**	0.193	0.110	0.284	0.312	0.225
	MOVERScore	0.129	0.157	0.159	0.318	0.191
	BARTScore	0.356	0.382	0.448	0.356	0.385
T5	**QuestEval**	0.228	0.306	0.182	0.268	0,246
	UniEval	0.449	0.446	0.575	0.426	0.474
qwen2.5 :72b	**SEval-Ex**	0.351	**0.580**	0.264	0.300	0.373

tion de la cohérence et en implémentant une comparaison directe à la source, nous pouvons surpasser même les approches sophistiquées basées sur les LLM sur cette dimension cruciale. La performance inférieure sur d'autres aspects est un compromis attendu de cette conception spécialisée, suggérant qu'une évaluation complète des résumés pourrait nécessiter la combinaison de plusieurs métriques spécialisées.

4.2 Analyse de la détection des Hallucinations

Une hallucination est un contenu génératif d'IA qui semble plausible mais n'est pas soutenu par des faits [9], cela peut se produire avec tous les LLM. Comme les hallucinations représentent un problème critique de qualité dans le texte généré, une métrique d'évaluation de cohérence fiable devrait montrer une sensibilité à leur présence, démontrant des scores de corrélation plus bas lorsque des hallucinations sont présentes dans le résumé. Pour évaluer la robustesse de notre métrique face à différents types d'hallucinations, nous avons mené une analyse systématique en utilisant le jeu de données SummEval. Nous avons développé trois catégories distinctes d'hallucinations synthétiques pour tester la capacité de notre cadre à détecter ces incohérences.

4.2.1 Types d'Hallucinations

Notre objectif est d'établir que les résumés contenant des hallucinations reçoivent un score plus bas selon notre métrique. À cette fin, nous avons conçu et implémenté trois types distincts d'hallucinations, illustrés dans la Figure 2.

1. **Remplacement d'Entités** : Substitution systématique d'entités nommées par des entités incorrectes tout en maintenant la structure globale du résumé.
2. **Événements Incorrects** : Modification de la séquence d'événements en introduisant de fausses relations temporelles ou causales. Ce type d'hallucination préserve les entités, mais déforme le flux narratif et la séquence factuelle des événements.

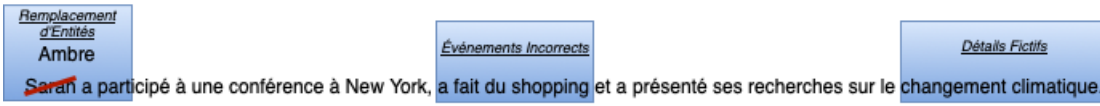

FIGURE 2 – Exemples d'hallucinations divisés en 3 types : Remplacement d'Entités, Événements Incorrects et Détails Fictifs.

3. **Détails Fictifs** : Ajout de détails plausibles mais non soutenus. Cela représente une forme plus subtile d'hallucination dans laquelle l'information principale reste intacte mais est embellie avec des détails non soutenus.

4.2.2 Préparation du Jeu de Données

Nous avons utilisé le jeu de données SummEval comme base, créant un jeu de données équilibré de 1 600 échantillons. Le jeu de données a été divisé uniformément en trois groupes égaux par type d'hallucination. Pour chaque résumé, nous avons généré une version hallucinée, basée sur un type d'hallucination, en utilisant un promptage contrôlé via un LLM.

4.2.3 Analyse des Résultats

Notre analyse a révélé des modèles distincts dans la façon dont notre métrique répond à différents types d'hallucinations (Figure 3) :

Remplacement d'Entités : A montré un impact modéré avec une réduction moyenne de score de 0,435 dans le score d'exactitude (de 0,964 à 0,529). Cette baisse peut s'expliquer car, au sein d'un énoncé atomique, un événement différent change entièrement sa signification et a des répercussions sur le raisonnement du verdict.

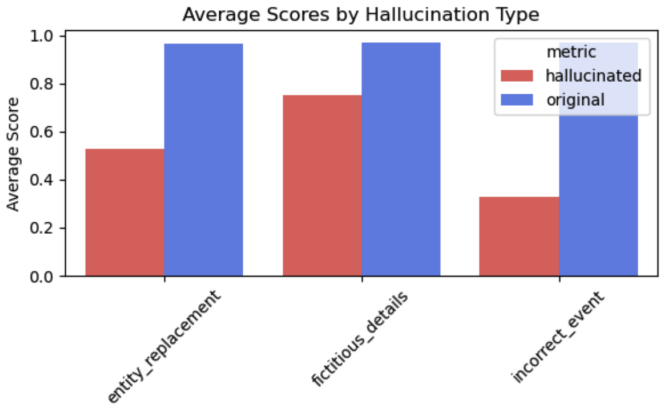

FIGURE 3 – Comparaison des scores moyens des métriques à travers différents types d'hallucination, montrant l'impact sur le score SEval-Ex.

Événements Incorrects : Ont démontré l'impact le plus sévère, avec une réduction de score de 0,65 en exactitude (de 0,970 à 0,328). C'est la même explication que pour le Remplacement d'Entités.

Détails Fictifs : Ont montré un impact plus petit mais toujours significatif, avec une diminution de score de 0,223 en exactitude (de 0,973 à 0,750). Comme nous n'ajoutons qu'un peu de bruit, le score ne baisse qu'un peu car le reste du résumé reste bon. Toutes les différences étaient statistiquement significatives ($p < 0,0001$), confirmant la capacité robuste de notre métrique à détecter diverses formes d'hallucination. Les amplitudes variables de réduction de score à travers différents types d'hallucinations suggèrent que notre métrique est particulièrement sensible aux modifications structurelles des séquences d'événements tout en maintenant une tolérance appropriée pour les élaborations mineures.

5 Discussions et limitations

Nos résultats expérimentaux démontrent que StEval-Ex réussit à combler l'écart entre performance et interprétabilité dans l'évaluation des résumés, particulièrement pour l'évaluation de la cohérence. Cependant, plusieurs limitations et considérations importantes peuvent être discutées. **Efficacité Computationnelle :** Bien que notre utilisation de LLMs légers et quantifiés rende le cadre plus accessible, le processus en deux étapes (extraction de phrases atomiques suivie de vérification) augmente la charge computationnelle par rapport à des métriques plus simples comme ROUGE. Ce compromis entre coût computationnel et qualité d'évaluation doit être pris en compte dans les applications pratiques. **Sensibilité au Prompt :** La performance du cadre dépend significativement de la qualité des prompts d'extraction de phrases atomiques et de vérification. Bien que nous ayons optimisé ceux-ci à travers des tests approfondis, la sensibilité à la conception des prompts suggère que des améliorations supplémentaires pourraient être possibles grâce à des techniques d'ingénierie de prompts plus sophistiquées. Pour assurer la reproductibilité, tous les prompts seront publiés sur le dépôt GitHub associé à l'article. **Portée de l'Évaluation :** Bien que SEval-Ex excelle dans l'évaluation de la cohérence factuelle, sa performance sur d'autres dimensions, particulièrement sur la pertinence, suggère que, selon nos expériences, nous devons incorporer de nouveaux types d'information pour évaluer si les phrases sélectionnées sont les phrases les plus pertinentes du texte original.

6 Conclusion

Notre travail présente SEval-Ex, un nouveau cadre pour évaluer la qualité des résumés avec des capacités interprétables tout en fournissant une corrélation état de l'art avec le jugement humain sur l'évaluation de la cohérence. Nos contributions clés sont :

- Un nouveau cadre d'évaluation basé sur les phrases qui décompose l'évaluation des résumés en une analyse fine des phrases atomiques, offrant à la fois une haute corrélation avec le jugement humain et un retour interprétable. Notre méthode atteint une corrélation état de l'art (0,580) sur l'évaluation de la cohérence.
- Le développement d'une méthodologie complète de détection d'hallucinations.
- Une implémentation pratique utilisant des LLMs légers et quantifiés, rendant notre approche accessible pour la recherche et les applications pratiques tout en maintenant une performance robuste à travers différentes longueurs de texte.

Ces réalisations suggèrent que la décomposition de l'évaluation des résumés en phrases atomiques offre une direction prometteuse pour développer des méthodes d'évaluation plus transparentes et efficaces. À mesure que les systèmes de résumé continuent de progresser, de tels cadres d'évaluation interprétables deviendront de plus en plus cruciaux pour assurer leur fiabilité et faciliter leur amélioration. Les travaux futurs pourraient explorer l'extension de cette approche pour mieux capturer les propriétés au niveau du document, comme l'importance des phrases, et investiguer des moyens d'optimiser l'efficacité computationnelle tout en maintenant la performance.

Références

[1] BOMMASANI R., HUDSON D. A., ADELI E., ALTMAN R., ARORA S., VON ARX S., BERNSTEIN M. S., BOHG J., BOSSELUT A., BRUNSKILL E. *et al.* (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv :2108.07258*.

[2] BROWN T. B. (2020). Language models are few-shot learners. *arXiv preprint arXiv :2005.14165*.

[3] DEVLIN J. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

[4] ES S., JAMES J., ESPINOSA-ANKE L. & SCHOCKAERT S. (2023). Ragas : Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv :2309.15217*.

[5] FABBRI A. R., KRYŚCIŃSKI W., MCCANN B., XIONG C., SOCHER R. & RADEV D. (2021). Summeval : Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, **9**, 391–409.

[6] FU J., NG S.-K., JIANG Z. & LIU P. (2023). GPTScore : Evaluate as You Desire. arXiv :2302.04166 [cs].

[7] GAO Y., ZHAO W. & EGER S. (2020). Supert : Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. *arXiv preprint arXiv :2005.03724*.

[8] HANNA M. & BOJAR O. (2021). A fine-grained analysis of bertscore. In *Proceedings of the Sixth Conference on Machine Translation*, p. 507–517.

[9] HUANG L., YU W., MA W., ZHONG W., FENG Z., WANG H., CHEN Q., PENG W., FENG X., QIN B. *et al.* (2023). A survey on hallucination in large language models : Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

[10] JIANG A. Q., SABLAYROLLES A., MENSCH A., BAMFORD C., CHAPLOT D. S., CASAS D. D. L., BRESSAND F., LENGYEL G., LAMPLE G., SAULNIER L. *et al.* (2023). Mistral 7b. *arXiv preprint arXiv :2310.06825*.

[11] KRYŚCIŃSKI W., MCCANN B., XIONG C. & SOCHER R. (2019). Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv :1910.12840*.

[12] LABAN P., SCHNABEL T., BENNETT P. N. & HEARST M. A. (2022). Summac : Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, **10**, 163–177.

[13] LIN C.-Y. (2004). ROUGE : A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, p. 74–81, Barcelona, Spain : Association for Computational Linguistics.

[14] LIU H., CUI L., LIU J. & ZHANG Y. (2021). Natural language inference in context-investigating contextual reasoning over long texts. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, p. 13388–13396.

[15] LIU Y., ITER D., XU Y., WANG S., XU R. & ZHU C. (2023). G-Eval : NLG Evaluation using GPT-4 with Better Human Alignment. arXiv :2303.16634 [cs].

[16] PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, p. 311–318.

[17] SCIALOM T., DRAY P.-A., GALLINARI P., LAMPRIER S., PIWOWARSKI B., STAIANO J. & WANG A. (2021). Questeval : Summarization asks for fact-based evaluation. *arXiv preprint arXiv :2103.12693*.

[18] SHU L., LUO L., HOSKERE J., ZHU Y., LIU Y., TONG S., CHEN J. & MENG L. (2024). Rewritelm : An instruction-tuned large language model for text rewriting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, p. 18970–18980.

[19] TOUVRON H., MARTIN L., STONE K., ALBERT P., ALMAHAIRI A., BABAEI Y., BASHLYKOV N., BATRA S., BHARGAVA P., BHOSALE S. *et al.* (2023). Llama 2 : Open foundation and fine-tuned chat models. *arXiv preprint arXiv :2307.09288*.

[20] WANG A., CHO K. & LEWIS M. (2020). Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv :2004.04228*.

[21] ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2020). BERTScore : Evaluating Text Generation with BERT. arXiv :1904.09675 [cs].

[22] ZHAO W., PEYRARD M., LIU F., GAO Y., MEYER C. M. & EGER S. (2019). Moverscore : Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv :1909.02622*.

[23] ZHONG M., LIU Y., YIN D., MAO Y., JIAO Y., LIU P., ZHU C., JI H. & HAN J. (2022). Towards a Unified Multi-Dimensional Evaluator for Text Generation. arXiv :2210.07197 [cs].