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

AllSummedUp : un framework open-source pour comparer les métriques d'évaluation de résumé

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

RÉSUMÉ

Cet article examine les défis de reproductibilité dans l'évaluation automatique des résumés de textes. À partir d'expériences menées sur six métriques représentatives allant de méthodes classiques comme ROUGE à des approches récentes basées sur les LLM (G-Eval, SEval-Ex), nous mettons en évidence des écarts notables entre les performances rapportées dans la littérature et celles observées dans notre cadre expérimental. Nous proposons un framework unifié et open-source, appliqué au jeu de données SummEval et ouvert à de futurs jeux de données, facilitant une comparaison équitable et transparente des métriques. Nos résultats révèlent un compromis structurel : les métriques les mieux alignées avec les jugements humains sont aussi les plus coûteuses en calculs et les moins stables. Au-delà de cette analyse comparative, notre étude met en garde contre l'utilisation croissante des LLM dans l'évaluation, en soulignant leur nature stochastique, leur dépendance technique et leur faible reproductibilité.

ABSTRACT

AllSummedUp : an open-source framework for comparing summary evaluation metrics

This paper investigates reproducibility challenges in automatic text summarization evaluation. Based on experiments conducted across six representative metrics ranging from classical approaches like ROUGE to recent LLM-based methods (G-Eval, SEval-Ex), we highlight significant discrepancies between reported performances in the literature and those observed in our experimental setting. We introduce a unified, open-source framework, applied to the SummEval dataset and designed to support fair and transparent comparison of evaluation metrics. Our results reveal a structural trade-off : metrics with the highest alignment with human judgments tend to be computationally intensive and less stable across runs. Beyond comparative analysis, this study highlights key concerns about relying on LLMs for evaluation, stressing their randomness, technical dependencies, and limited reproducibility. We advocate for more robust evaluation protocols including exhaustive documentation and methodological standardization to ensure greater reliability in automatic summarization assessment.

MOTS-CLÉS : métriques d'évaluation, résumé automatique, reproductibilité, grands modèles de langue.

KEYWORDS: evaluation metrics, automatic summarization, reproducibility, large language models.

1 Introduction

La reproductibilité des résultats expérimentaux constitue un pilier fondamental de la recherche scientifique, en particulier dans le domaine du traitement automatique du langage naturel (TALN).

Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM), pages 11–21.

Pourtant, un nombre croissant d'études souligne la fragilité de cette reproductibilité, notamment face à l'essor des modèles de langue de grande taille (LLM). Ces modèles, utilisés pour évaluer ou produire des textes, manifestent une variabilité inquiétante, même dans des conditions que l'on pourrait croire déterministes (Atil *et al.*, 2025; Biderman *et al.*, 2024).

Cette instabilité remet en cause la fiabilité des métriques d'évaluation, essentielles en résumé automatique où les jugements humains sont coûteux à mobiliser. En effet, comme le montrent Salinas & Morstatter ou He *et al.*, la reformulation du prompt — parfois réduite à l'ajout d'un simple espace — peut affecter significativement les sorties des modèles, révélant leur extrême sensibilité aux prompts. De plus, la forte dépendance à des modèles spécifiques pose également un risque de non-reproductibilité à moyen terme, dans la mesure où certains modèles, comme GPT-3, peuvent être rendus inaccessibles ou supprimés par leurs fournisseurs (Biderman *et al.*, 2024), rendant impossible toute reproduction a posteriori si les conditions initiales ne sont pas rigoureusement conservées.

Dans ce contexte, la reproductibilité ne se limite pas à la réexécution d'un code identique, mais implique aussi la robustesse des résultats face à de légères variations de conditions expérimentales. La distinction conceptuelle entre répétabilité et reproductibilité, bien établies en métrologie (Belz *et al.*, 2021), est rarement prise en compte dans les études sur les métriques d'évaluation.

Or, de telles différences peuvent engendrer des biais dans l'interprétation des performances, comme le montrent nos résultats empiriques où certaines métriques — pourtant largement utilisées, comme BERTScore (Zhang *et al.*, 2020) ou BARTScore (Yuan *et al.*, 2021) — affichent une corrélation faible et non reproductible avec les jugements humains dans différents contextes d'exécution.

Notre travail s'inscrit dans cet élan critique en proposant un framework unifié et open-source pour l'évaluation des résumés, incluant six métriques issues de la littérature. Ce cadre vise à tester rigoureusement la reproductibilité des résultats sur un jeu de données standardisé, SummEval (Fabbri *et al.*, 2021), et a été développé pour accueillir de futurs jeux de données. En contribuant à la standardisation et à la transparence méthodologique, nous espérons non seulement faciliter la comparaison équitable entre métriques, mais aussi réduire la variabilité technique introduite par l'utilisation de modèles stochastiques, tels que les LLM.

Pour ce faire, cet article est structuré comme suit : la section 2 présente l'état de l'art portant sur la reproductibilité en TALN et les différentes métriques d'évaluation de résumés. La section 3 détaille notre approche méthodologique, en décrivant le framework que nous avons développé, les métriques implémentées, les données d'évaluation utilisées, ainsi que le choix de notre mesure de corrélation. Les expérimentations menées, leurs résultats en termes de corrélation et de temps d'exécution, ainsi qu'une comparaison des métriques sont ensuite présentés dans la section 4. Enfin, la section 5 analyse les implications et les limites de notre étude, avant que la section 6 ne vienne clore cet article en proposant des pistes pour les travaux futurs.

2 Travaux Connexes

2.1 Reproductibilité

La reproductibilité représente un enjeu majeur en apprentissage automatique et notamment en traitement automatique du langage naturel (TALN) (Belz *et al.*, 2021). Ces derniers identifient la complexité des pipelines expérimentaux, le non-déterminisme (lié aux générateurs aléatoires et

à l'hétérogénéité matérielle) et la documentation incomplète comme principaux obstacles. Selon cette étude, seulement $14\,\%$ des reproductions exactes sont réussies malgré l'accès aux ressources originales.

En réponse, diverses initiatives communautaires ont vu le jour, telles que les listes de contrôle proposées lors des grandes conférences en TALN (Whitaker, 2017) et des sessions dédiées à la reproductibilité expérimentale dans certains événements (Branco *et al.*, 2020). Malgré ces efforts, (Xue *et al.*, 2023) soulignent que même des pratiques standardisées, comme l'utilisation d'une partition fixe des données, peuvent introduire des biais significatifs. Cela indique clairement un besoin de protocoles expérimentaux plus robustes pour garantir des comparaisons fiables entre modèles.

2.2 Métriques d'évaluation de résumés

L'évaluation automatique des résumés s'est historiquement appuyée sur des métriques de chevauchement lexical comme ROUGE (Lin, 2004), largement adoptée grâce à sa simplicité et son rôle central dans les compétitions DUC/TAC (Bhandari *et al.*, 2020). Toutefois, sa capacité limitée à capturer la paraphrase, l'équivalence sémantique et la cohérence factuelle a motivé le développement de nouvelles approches.

Des métriques basées sur les plongements contextuels (embeddings), telles que BERTScore (Zhang *et al.*, 2020) et MoverScore (Zhao *et al.*, 2019), évaluent la similarité sémantique entre textes au-delà du simple chevauchement lexical. D'autres, comme QuestEval (Scialom *et al.*, 2021), qui repose sur la génération et la réponse à des questions, et FactCC (Kryściński *et al.*, 2019), basé sur l'inférence de relations d'entaillement, visent à mesurer la factualité. Des approches probabilistes, telles que BARTScore (Yuan *et al.*, 2021), estiment la qualité des résumés en utilisant la probabilité de génération conditionnelle. Enfin, des cadres d'évaluation multidimensionnelle, comme UniEval (Zhong *et al.*, 2022) et G-Eval (Liu *et al.*, 2023), cherchent à unifier l'évaluation sur plusieurs axes qualitatifs, tels que la cohérence, la factualité et l'adéquation. Pour finir, SEval-Ex (Herserant & Guigue, 2025), propose une métrique basée sur un pipeline LLM explicable spécialisée dans la factualité.

Cependant, les nouvelles métriques, souvent coûteuses en calculs et peu standardisées, posent des défis en termes de reproductibilité et de comparabilité entre les travaux.

2.3 Variabilité et reproductibilité des LLM dans l'évaluation automatique

Les grands modèles de langue (LLM) jouent un rôle croissant dans l'évaluation automatique de la génération de texte, notamment via des métriques comme G-Eval (Liu *et al.*, 2023), UniEval (Zhang, 2024), ou SEval-Ex (Herserant & Guigue, 2025). Toutefois, leur utilisation soulève des préoccupations croissantes en matière de stabilité et reproductibilité des résultats.

Plusieurs travaux récents révèlent une variabilité importante des sorties des LLM, même dans des conditions censément déterministes. (Atil *et al.*, 2025) montrent que des exécutions répétées avec les mêmes paramètres peuvent produire des réponses différentes, remettant en cause l'idée d'un comportement stable. D'autres études s'intéressent à la sensibilité des modèles à la formulation des prompts. (Salinas & Morstatter, 2024) montrent que de simples modifications superficielles — comme l'ajout d'un espace ou d'un saut de ligne — peuvent entraîner des changements radicaux dans la sortie du modèle, phénomène surnommé l'effet papillon (*Butterfly Effect*). (He *et al.*, 2024) confirment cette

sensibilité en comparant les performances de LLM selon différents formats de prompt (*plain text*, JSON, Markdown, etc.), révélant que cette instabilité ne dépend pas uniquement de la tâche mais constitue une propriété générale du comportement des LLM.

Enfin, ces problèmes de variabilité sont exacerbés par des difficultés pratiques de reproductibilité. (Biderman *et al.*, 2024) insistent sur les nombreux facteurs pouvant altérer l'évaluation : version des modèles, méthode de décodage, stratégie de normalisation, ou encore dépendance aux bibliothèques. (Xue *et al.*, 2023) montrent que même les pratiques réputées standard, comme l'utilisation d'un split fixe des données, peuvent introduire des biais significatifs. Ils proposent des méthodes statistiques basées sur le ratio signal-bruit pour quantifier la stabilité des comparaisons entre modèles.

3 Méthodologie

3.1 Framework

Nous présentons un cadre modulaire complet pour l'évaluation des résumés de texte qui implémente six métriques diverses au sein d'une interface cohérente. Le cadre est structuré autour de trois composants principaux :

Interface de Métrique : Une classe abstraite standardisée (`TextMetric`) qui définit l'interface commune que toutes les métriques doivent implémenter, incluant des méthodes pour le prétraitement, le calcul et le formatage des résultats.
Évaluateur : Un orchestrateur central (`TextEvaluator`) qui gère les instances de métriques, les applique aux textes d'entrée et agrège les résultats dans un format cohérent.
Générateur de Rapports : Un composant (`Report`) qui transforme les résultats d'évaluation en divers formats (JSON, YAML, CSV) et génère des visualisations pour faciliter l'analyse.

Cette architecture garantit que toutes les métriques suivent le même flux de travail et produisent des sorties comparables, facilitant les comparaisons directes et réduisant la variabilité liée à l'implémentation.

Par ailleurs, le framework est conçu pour être extensible et permet l'ajout de nouveaux jeux de données d'évaluation. [1]

3.2 Métriques implémentées

Notre cadre implémente les métriques suivantes, sélectionnées pour représenter diverses approches d'évaluation de résumés, allant des comparaisons basées sur les n-grammes aux méthodes plus récentes basées sur les plongements et les modèles de langue :

Métriques de Chevauchement Lexical : Ces métriques évaluent le chevauchement lexical entre le résumé généré et une référence.

— ROUGE (Lin, 2004) : Orientée rappel, elle calcule le chevauchement de n-grammes (ROUGE-1, ROUGE-2) ou la plus longue sous-séquence commune (ROUGE-L). C'est la métrique la plus populaire.

1. Le github sera partagé après acceptation de l'article

Métriques Basées sur des Modèles (*Reference-Free*) : Ces métriques n'ont pas besoin de résumé de référence humain.

— QuestEval (Scialom *et al.*, 2021) : Évalue la fidélité et la pertinence en générant des questions à partir de la source (pour le rappel) et du résumé (pour la précision), puis en y répondant avec l'autre texte. Montre une bonne corrélation avec le jugement humain, même sans résumé de référence. Un modèle est accessible sur `HuggingFace`.
— BARTScore (Yuan *et al.*, 2021) : Évalue le texte généré en calculant sa probabilité selon un modèle pré-entraîné (BART), permettant d'évaluer sous différents angles (information, fluidité, factualité).
— UniEval (Zhong *et al.*, 2022) : Propose une évaluation multidimensionnelle via un modèle entraîné à répondre à des questions oui/non spécifiques guidant l'évaluation. Un modèle est accessible sur `HuggingFace`.
— G-Eval (Liu *et al.*, 2023) : Utilise un grand modèle de langue (LLM) comme GPT-4 avec des prompts dédiés pour évaluer la qualité selon plusieurs critères.
— SEval-Ex (Herserant & Guigue, 2025) : Basée sur des énoncés atomiques — des unités d'information autonomes — la métrique permet une analyse entièrement interprétable. Grâce à une architecture en deux temps, elle permet de bien capter les informations présentes à la fois dans le texte et le résumé.

Pour chaque métrique, nous implémentons soigneusement les pipelines de prétraitement, les paramètres et le formatage des résultats conformément aux publications originales. Nous reprenons systématiquement les codes GitHub ou les packages déjà codés et les adaptons si besoin.

3.3 Données d'évaluation

Afin d'évaluer la reproductibilité des différentes métriques dans des contextes variés, nous nous appuyons sur le jeu de données SummEval (Fabbri *et al.*, 2021). Celui-ci comprend 1600 résumés générés par 16 systèmes de résumé automatiques distincts, accompagnés d'annotations humaines selon quatre dimensions d'évaluation : la cohérence textuelle, la cohérence factuelle, la fluidité, et la pertinence.

3.4 Choix de la mesure de corrélation

La littérature ne présente pas de consensus clair quant à la mesure de corrélation à privilégier pour évaluer l'alignement entre métriques automatiques et jugements humains. Les auteurs du jeu de données SummEval utilisent la corrélation de *Kendall*, QuestEval rapporte des coefficients de *Pearson*, tandis que des approches plus récentes comme G-Eval et UniEval optent pour *Spearman*. Cette hétérogénéité complique les comparaisons et rend difficile l'interprétation systématique des résultats.

Dans notre étude, nous avons choisi d'utiliser la corrélation de *Spearman* pour plusieurs raisons :
— Elle est de plus en plus adoptée dans les travaux récents d'évaluation de résumés (G-Eval), ce qui renforce la comparabilité de nos résultats avec l'état de l'art.
— Elle ne suppose pas de relation linéaire entre les variables en se basant sur les rangs.
— Elle est moins sensible aux valeurs extrêmes que la corrélation de Pearson.

Ce choix vise à assurer une robustesse méthodologique et une meilleure cohérence avec les pratiques actuelles dans la communauté.

4 Expériences

4.1 Protocole d'évaluation

L'ensemble des expériences a été réalisé en local sur une machine équipée d'un GPU NVIDIA A6000 (48 Go de VRAM), garantissant une capacité suffisante pour exécuter les LLM utilisés (`Gemma-3-27b` et `Qwen-2.5-72b`).

Les modèles nécessaires à l'utilisation de certaines métriques ont été obtenus via deux sources principales : Les modèles utilisés pour QuestEval, UniEval, BERTScore et BARTScore ont été téléchargés depuis la plateforme `HuggingFace`, en utilisant les versions publiques les plus récentes disponibles au moment de l'expérimentation. Les variantes basées sur LLM pour G-Eval et SEval-Ex ont été exécutées via `Ollama`, une infrastructure locale permettant d'héberger et d'interroger des LLM tout en évitant la dépendance à une API externe, souvent commerciale. Afin d'assurer la comparabilité entre les différentes métriques, tous les traitements ont été effectués dans le même environnement. Pour les LLM, nous utilisons les valeurs par défaut, par exemple pour `Gemma-3-27b` : `temperature` = 1, `top_k` = 64 et `top_p` = 0,95.

4.2 Résultats

4.2.1 Corrélation Spearman

La Table 1 présente les corrélations de Spearman entre les scores fournis par les métriques automatiques et les jugements humains sur quatre dimensions (Cohérence, Consistance, Fluidité, Pertinence), ainsi que pour deux modèles de résumé distincts pour les métriques G-Eval et SEval-Ex : **`Gemma-3-27b`** et **`Qwen-2.5-72b`**. Les résultats issus de notre propre expérimentation sont comparés avec ceux rapportés dans la littérature. L'objectif est d'évaluer dans quelle mesure ces corrélations, souvent citées comme indicateurs de validité des métriques, peuvent être reproduites dans un cadre expérimental distinct. L'un des obstacles majeurs à cette évaluation réside dans l'indisponibilité partielle ou totale des corrélations de Spearman dans les travaux originaux. Lorsqu'elles n'étaient pas explicitement fournies, nous avons pris comme références des articles ayant reproduit les expériences.

Les résultats révèlent une variabilité notable entre les valeurs attendues et celles effectivement mesurées. Si certaines métriques montrent une relative stabilité (ex. UniEval sauf Fluidité et SEval-Ex avec `Qwen-2.5-72b`), d'autres présentent des écarts importants voire des inversions de tendance (G-Eval sur la fluidité : -0.45 ; SEval-Ex-`Gemma-3-27b` sur la Pertinence : $+0.12$). Ces différences suggèrent que les performances rapportées dans la littérature ne sont pas systématiquement généralisables, même en conservant un jeu de données identique.

Nous n'avons pas pu répliquer exactement les résultats de ROUGE en raison d'un manque de détails (les résultats sont présentés dans la Table 2). En effet, dans l'article BARTScore (Yuan *et al.*, 2021), le premier article présentant les résultats pour SummEval, il est spécifié que ROUGE est calculé entre les candidats et les références de chaque document de SummEval, mais il ne précise pas si

toutes les références sont utilisées ni comment (par ex., une moyenne des scores entre un résumé et toutes les références). Ici, un manque de clarté empêche la reproduction. En ce qui concerne les autres métriques, plusieurs facteurs peuvent expliquer une dérive des résultats : l'utilisation d'un LLM différent, un prompt modifié ou encore des versions divergentes des bibliothèques logicielles. Ces résultats soulignent la nécessité d'un encadrement plus rigoureux des protocoles expérimentaux, incluant des éléments tels que les poids des LLM utilisés, la publication des configurations exactes (prompt, température, méthode de génération).

TABLE 1 – Corrélations Spearman entre les dimensions évaluatives et les métriques automatiques pour les modèles **Gemma** et **Qwen**, avec écart par rapport aux valeurs de référence.

Source	Dimension	BartScore	BertScore	QuestEval	UniEval	GEval	SEval-Ex
Gemma3 :27b	Coh	0.45 (+0.05)	0.16 (-0.12)	0.26 (+0.08)	0.53 (-0.04)	0.55 (-0.02)	0.21 (-0.05)
	Con	0.24 (-0.14)	0.27 (+0.16)	0.29 (-0.01)	0.43 (-0.01)	0.62 (+0.18)	0.43 (-0.15)
	Flu	0.22 (-0.13)	0.19 (+0.00)	0.18 (-0.10)	0.34 (-0.10)	-0.01 (-0.45)	0.34 (-0.01)
	Rel	0.37 (+0.01)	0.30 (+0.01)	0.37 (+0.10)	0.42 (+0.00)	0.52 (+0.10)	0.42 (+0.12)
Qwen2.5 :72b	Coh	—	—	—	—	0.38 (-0.19)	0.27 (+0.01)
	Con	—	—	—	—	0.47 (+0.03)	0.57 (-0.01)
	Flu	—	—	—	—	-0.07 (-0.51)	0.33 (-0.02)
	Rel	—	—	—	—	0.48 (+0.06)	0.31 (+0.01)

TABLE 2 – Corrélations de Spearman de la métrique **Rouge** obtenus sur SummEval.

Dimension	Rouge -nous (1/2/L)	Rouge - réel (1/2/L)
Coh	0.18 / 0.15 / 0.17	0.167 / 0.184 / 0.128
Con	0.14 / 0.13 / 0.12	0.160 / 0.187 / 0.115
Flu	0.08 / 0.06 / 0.08	0.115 / 0.159 / 0.105
Rel	0.3 / 0.25 / 0.24	0.326 / 0.290 / 0.311

4.2.2 Temps d'exécution

TABLE 3 – **Temps total d'exécution des métriques pour Gemma 3 27B et Qwen 2.5 72B**. Seul GEval et SEval-Ex sont impactées.

Métrique	**Gemma 3 27B (s)**	**Qwen 2.5 72B (s)**
SEval_Ex	13449.63	38302.47
Questeval	8185.54	8232.21
GEval_coherence	4827.80	18660.82
GEval_consistency	3844.54	18787.04
GEval_relevance	3744.85	13354.72
GEval_fluency	3527.36	12828.27
Unieval	297.42	300.46
Bartscore	32.43	30.16
Rouge	11.47	11.36
Total	**37944.20**	**202299.97**

Toutes les métriques ont été testées sur plusieurs exécutions distinctes. Les résultats, présents dans la Table 3 montrent une distinction claire entre les métriques légères (ex. ROUGE, BARTScore) qui s'exécutent en quelques secondes, et les métriques lourdes impliquant des modèles LLM (G-Eval,

SEval-Ex, QuestEval) dont les temps de traitement peuvent atteindre plusieurs heures (10h 38m 22s pour SEval-Ex avec `Qwen-2.5-72b`). L'impact de la taille du modèle évalué est particulièrement marquant : pour des métriques identiques, les temps d'exécution avec `Qwen-2.5-72b` sont systématiquement beaucoup plus élevés qu'avec `Gemma-3-27b`, en général trois à quatre fois supérieurs. Cela souligne l'importance de prendre en compte la complexité du modèle de référence dans l'analyse des performances et des coûts d'évaluation. Nos temps d'exécution sont à relativiser par rapport à notre environnement d'exécution.

4.3 Comparaison

La comparaison croisée des résultats de performance (Section 4.2.1) et des temps d'exécution (Section 4.2.2) met en lumière des compromis marqués entre précision d'évaluation, reproductibilité et coût calculatoire des différentes métriques.

Les métriques légères, comme ROUGE ou BERTScore, s'exécutent en quelques secondes à peine (entre 11 et 32 secondes) sur tout un jeu de données, ce qui les rend particulièrement adaptées à des évaluations rapides ou à grande échelle. Toutefois, leurs corrélations avec les jugements humains restent globalement modestes ou inconstantes, notamment pour des dimensions complexes comme la fluidité ou la cohérence globale.

À l'autre extrême, les métriques basées sur LLM, comme G-Eval, SEval-Ex et QuestEval, affichent des corrélations parfois plus élevées, mais au prix de temps de calcul très importants — jusqu'à plus de 8 heures cumulées pour SEval-Ex (`Qwen-2.5-72b`) sur une seule exécution. Les métriques basées sur un LLM, pouvant avoir plusieurs miliards de paramètres (`Qwen-2.5-72b` a plus de deux fois plus de paramètres que `Gemma-3-27b`, sont difficilement utilisables sans infrastructure dédiée en fonction de la taille du jeu de données à essayer.

Entre ces deux extrêmes, des approches intermédiaires comme UniEval offrent un bon compromis : elles nécessitent des temps modérés (300 s), tout en fournissant des résultats relativement stables sur plusieurs dimensions, notamment la pertinence et la cohérence. UniEval se distingue notamment par sa constance, avec des performances élevées et peu de dérive par rapport aux valeurs de référence, sauf sur la fluidité.

Ce panorama met en évidence plusieurs points :

- *Performance vs. Frugalité* : les métriques les plus performantes en termes de corrélation sont aussi les plus coûteuses en calculs, et donc les moins frugales.
- *Performance vs. Reproductibilité* : certaines métriques à haute performance, comme G-Eval, montrent une grande variabilité selon le modèle ou les conditions d'exécution.
- *Simplicité vs. Pertinence* : les métriques classiques, bien que simples à exécuter, ont leur pouvoir discriminant limité dans des contextes modernes (LLM, paraphrases, factualité).

Ces observations plaident pour une approche différenciée selon les contraintes d'usage : les métriques frugales conviennent à des évaluations exploratoires ou en production à grande échelle, tandis que les métriques basées sur LLM sont utiles pour des évaluations approfondies mais ponctuelles. L'existence de telles tensions souligne l'importance de proposer des outils capables d'intégrer plusieurs métriques et de guider leur usage selon les contextes, comme le permet notre framework.

5 Discussion

La création de notre framework et la comparaison avec les articles originaux révèle plusieurs limitations qui doivent être soulignées :

Reproductibilité partielle des métriques originales. Certaines métriques, notamment celles basées sur des embeddings comme BERTScore ou des modèles spécifiques, montrent des divergences notables entre les corrélations rapportées dans la littérature et celles obtenues via notre framework. Ces écarts peuvent être dus, par exemple, à des différences de version des modèles ou de tokenisation.

Accès restreint aux modèles propriétaires. Certaines métriques récentes, telles que *G-Eval* reposant sur GPT-4, dépendent de modèles ou d'API fermées, ce qui soulève plusieurs enjeux critiques pour la reproductibilité :

- un **biais d'accessibilité**, lié aux ressources financières ou matérielles nécessaires pour interroger ces modèles ;
- des **limitations techniques**, telles que les quotas d'utilisation, les coûts d'exécution ou la latence ;
- une **fragilité de la reproductibilité** à moyen terme, certains modèles pouvant être modifiés ou rendus inaccessibles (comme GPT-3), ce qui remet en question la pérennité des expériences.

Ces contraintes illustrent des points de tension : entre *performance* et *reproductibilité*, entre *stabilité* et *accessibilité*, ou encore entre *qualité des résultats* et *frugalité calculatoire*. Dans notre étude, nous avons privilégié l'utilisation d'alternatives open-source, quitte à perdre en fidélité par rapport aux conditions d'évaluation originales.

Données d'évaluation limitées à un seul benchmark. Notre étude repose exclusivement sur le jeu de données *SummEval*, bien que riche. Des expérimentations sur d'autres corpus seraient nécessaires pour généraliser nos conclusions. Notre framework a cependant été conçu pour faciliter l'ajout de nouveaux jeux de données.

Temps de calcul prohibitif pour certaines métriques. Certaines métriques, comme *SEval-Ex*, *QuestEval* ou *G-Eval*, présentent un temps de traitement très élevé (jusqu'à plusieurs heures par exécution), les rendant peu adaptées aux environnements à contrainte de ressources ou aux évaluations à grande échelle.

6 Conclusion

Notre étude met en lumière les défis majeurs que pose la reproductibilité dans l'évaluation automatique des résumés. Les écarts parfois significatifs entre les performances rapportées dans la littérature et celles mesurées au sein de notre cadre expérimental soulignent l'urgence d'adopter une approche plus rigoureuse, explicite et systématique dans ce domaine.

Trois tensions structurantes émergent de nos analyses. Premièrement, la performance s'oppose souvent à la frugalité calculatoire : les métriques les plus corrélées aux jugements humains (comme G-Eval ou

SEval-Ex) sont également les plus coûteuses en ressources. Deuxièmement, la performance s'oppose parfois à la reproductibilité : certaines métriques, bien que prometteuses, montrent une variabilité préoccupante selon le modèle ou l'environnement d'exécution. Troisièmement, la simplicité d'usage entre en conflit avec la pertinence évaluative : les métriques classiques telles que ROUGE, faciles à implémenter, peinent à capturer la richesse des résumés produits par des modèles modernes.

Notre framework open-source constitue une réponse partielle à ces défis en offrant une plateforme unifiée facilitant la comparaison transparente et équitable des métriques. Pour aller plus loin, nous recommandons à la communauté d'adopter des protocoles standardisés, de publier systématiquement les paramètres expérimentaux complets (modèles, prompts, températures, graines aléatoires (seeds), etc.), et d'intégrer des mesures statistiques de variabilité dans les rapports de performance.

Ces efforts de standardisation sont d'autant plus essentiels que l'évaluation automatique s'oriente vers des approches fondées sur les LLM, dont le comportement intrinsèquement stochastique accentue les risques de non-reproductibilité. À long terme, de telles pratiques favoriseront non seulement une meilleure compréhension des outils existants mais aussi la conception de nouvelles métriques, capables de concilier exigence méthodologique, robustesse empirique et efficacité calculatoire.

Références

ATIL B., AYKENT S., CHITTAMS A., FU L., PASSONNEAU R. J., RADCLIFFE E., RAJAGOPAL G. R., SLOAN A., TUDREJ T., TURE F., WU Z., XU L. & BALDWIN B. (2025). Non-Determinism of "Deterministic" LLM Settings. arXiv :2408.04667 [cs], DOI : 10.48550/arXiv.2408.04667.

BELZ A., AGARWAL S., SHIMORINA A. & REITER E. (2021). A Systematic Review of Reproducibility Research in Natural Language Processing. In P. MERLO, J. TIEDEMANN & R. TSARFATY, Éds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, p. 381–393, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-main.29.

BHANDARI M., GOUR P., ASHFAQ A., LIU P. & NEUBIG G. (2020). Re-evaluating evaluation in text summarization. *arXiv preprint arXiv :2010.07100*.

BIDERMAN S., SCHOELKOPF H., SUTAWIKA L., GAO L., TOW J., ABBASI B., AJI A. F., AMMANAMANCHI P. S., BLACK S., CLIVE J., DIPOFI A., ETXANIZ J., FATTORI B., FORDE J. Z., FOSTER C., HSU J., JAISWAL M., LEE W. Y., LI H., LOVERING C., MUENNIGHOFF N., PAVLICK E., PHANG J., SKOWRON A., TAN S., TANG X., WANG K. A., WINATA G. I., YVON F. & ZOU A. (2024). Lessons from the Trenches on Reproducible Evaluation of Language Models. arXiv :2405.14782 [cs], DOI : 10.48550/arXiv.2405.14782.

BRANCO A., CALZOLARI N., VOSSEN P., VAN NOORD G., VAN UYTVANCK D., SILVA J., GOMES L., MOREIRA A. & ELBERS W. (2020). A shared task of a new, collaborative type to foster reproducibility : A first exercise in the area of language science and technology with reprolang2020. In *Proceedings of The 12th Language Resources and Evaluation Conference*, p. 5539–5545 : European Language Resources Association (ELRA).

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.

HE J., RUNGTA M., KOLECZEK D., SEKHON A., WANG F. X. & HASAN S. (2024). Does Prompt Formatting Have Any Impact on LLM Performance ? arXiv :2411.10541 [cs], DOI : 10.48550/arXiv.2411.10541.

HERSERANT T. & GUIGUE V. (2025). Seval-ex : A statement-level framework for explainable summarization evaluation. Accepté pour publication, prépublication en cours sur arXiv.

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

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.

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], DOI : 10.48550/arXiv.2303.16634.

SALINAS A. & MORSTATTER F. (2024). The Butterfly Effect of Altering Prompts : How Small Changes and Jailbreaks Affect Large Language Model Performance. arXiv :2401.03729 [cs], DOI : 10.48550/arXiv.2401.03729.

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*.

WHITAKER K. (2017). The mt reproducibility checklist. https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf. Accessed : 2025-04-29.

XUE Y., CAO X., YANG X., WANG Y., WANG R. & LI J. (2023). We Need to Talk About Reproducibility in NLP Model Comparison. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 9424–9434, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.586.

YUAN W., NEUBIG G. & LIU P. (2021). BARTScore : Evaluating Generated Text as Text Generation. arXiv :2106.11520 [cs], DOI : 10.48550/arXiv.2106.11520.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2020). BERTScore : Evaluating Text Generation with BERT. arXiv :1904.09675 [cs], DOI : 10.48550/arXiv.1904.09675.

ZHANG Y. (2024). Unraveling Text Generation in LLMs : A Stochastic Differential Equation Approach. arXiv :2408.11863 [cs], DOI : 10.48550/arXiv.2408.11863.

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 :1909.02622 [cs], DOI : 10.48550/arXiv.1909.02622.

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], DOI : 10.48550/arXiv.2210.07197.