Identification de mesures d'évaluation fiables pour la révision de textes scientifiques

Léane Jourdan[1] Florian Boudin[1, 2] Nicolas Hernandez[1] Richard Dufour[1]
(1) Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
(2) JFLI, CNRS, Nantes University, France
{prénom}.{nom}@univ-nantes.fr

RÉSUMÉ

L'évaluation de la révision des textes scientifiques reste un défi, car les métriques traditionnelles telles que ROUGE et BERTScore se concentrent sur la similarité à une référence plutôt que sur les améliorations réalisées. Nous analysons et identifions les limites de ces métriques et explorons des méthodes d'évaluation alternatives qui s'alignent mieux sur le jugement humain. Nous évaluons d'abord manuellement différentes révisions pour estimer leur qualité. Ensuite, nous examinons la possibilité d'utiliser des métriques d'évaluation sans référence provenant de domaines connexes du traitement automatique des langues (TAL) ainsi que des approches GML en tant que juge. Nos résultats montrent que GMLs évaluent efficacement le suivi des instructions mais peinent à évaluer l'acceptabilité, alors que les métriques spécifiques au domaine fournissent des informations complémentaires. Nous recommandons une approche hybride combinant l'évaluation GML en tant que juge et les mesures spécifiques à la tâche offrant l'évaluation la plus fiable de la qualité de la révision.

ABSTRACT

Identifying Reliable Evaluation Metrics for Scientific Text Revision

Evaluating text revision in scientific writing remains a challenge, as traditional metrics such as ROUGE and BERTScore primarily focus on similarity rather than capturing meaningful improvements. In this work, we analyse and identify the limitations of these metrics and explore alternative evaluation methods that better align with human judgments. We first conduct a manual annotation study to assess the quality of different revisions. Then, we investigate reference-free evaluation metrics from related NLP domains and LLM-as-a-judge approaches. Our results show that LLMs effectively assess instruction-following but struggle with correctness, while domain-specific metrics provide complementary insights. We find that a hybrid approach combining LLM-as-a-judge evaluation and task-specific metrics offers the most reliable assessment of revision quality.

MOTS-CLÉS : révision de texte, article scientifique, évaluation, métriques.

KEYWORDS: text revision, scientific article, evaluation, metrics.

ARTICLE : **Accepté à** ACL 2025.

1 Introduction

La révision est une étape essentielle de la rédaction scientifique, car elle garantit la clarté, la cohérence et le respect des normes académiques. Le processus d'écriture se compose généralement de quatre étapes : 1) la pré-écriture, 2) le brouillon, 3) la révision et 4) l'édi-

tion (Jourdan *et al.*, 2023). L'étape de la révision implique des modifications substantielles pour améliorer la lisibilité, le style et la formalité (Du *et al.*, 2022; Li *et al.*, 2022). Cette étape est cruciale, car une rédaction de mauvaise qualité brouille la compréhension des résultats de recherche et augmente le risque de rejet d'un article (Amano *et al.*, 2023).

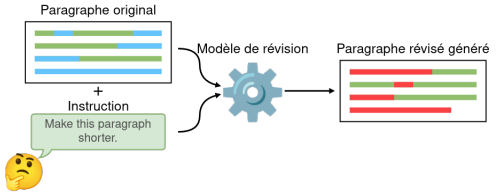

FIGURE 1 – Aperçu de la tâche de révision

Comme l'illustre la Figure 1, la tâche de révision prend en entrée un paragraphe original et une instruction spécifiant la modification requise. Le résultat attendu est un paragraphe révisé s'alignant sur l'instruction donnée. Compte tenu de l'importance de cette tâche, il est crucial de pouvoir l'évaluer de façon fiable.

Comme pour les autres tâches de génération de texte, la révision de texte est évaluée grâce à des métriques bien établies telles que ROUGE (Lin, 2004) ou BERTScore (Zhang *et al.*, 2020). Bien que les métriques basées sur les plongements (e.g. BERTScore) capturent en partie la similarité sémantique, elles restent centrées sur le chevauchement lexical et les caractéristiques de surface, plutôt que sur des aspects plus profonds de la qualité du texte.

Dans le cadre de la révision de texte, les métriques de similarité seules échouent à pleinement capturer la qualité de la révision. Au-delà de la similarité à un texte de référence, l'évaluation de la révision requiert de considérer les améliorations par rapport au texte original, la préservation du sens et le respect de l'instruction. Plusieurs études se sont reposées sur l'évaluation humaine pour évaluer les systèmes de révision (Du *et al.*, 2022; Raheja *et al.*, 2023, 2024; Ito *et al.*, 2020; Schick *et al.*, 2023), mais celle-ci apparaît coûteuse et chronophage, la rendant impraticable pour l'évaluation à grande échelle. Face à cette limitation, nous explorons différentes approches d'évaluation automatiques capables d'estimer la qualité des révisions tout en passant à l'échelle.

La révision de texte englobant diverses sous-tâches (e.g. paraphrase, résumé, simplification, transfert de style, correction grammaticale) (Li *et al.*, 2022; Raheja *et al.*, 2024; Ito *et al.*, 2019; Kim *et al.*, 2022), nous explorons tout d'abord les métriques sans référence communément utilisées pour évaluer ces tâches. Ainsi, ces métriques comparent directement les textes originaux et révisés au lieu de se reposer sur la référence. De plus, nous explorons différentes approches par grands modèles de langue (*Large Language Models* ; GMLs) en tant que juge (LLM-as-a-judge en anglais, et abbrégé en GML-juge dans cet article), celles-ci pouvant considérer l'instruction de révision. Avec la croissance rapide des GMLs, ces approches connaissent une utilisation croissante pour l'évaluation de diverses tâches (Gu *et al.*, 2024). Toutefois, des études ont montré que les GMLs rencontrent des pertes de performances quand la référence n'est pas fournie et sont parfois surpassés par des méthodes plus simples et moins coûteuses, les rendant moins attrayants (Doostmohammadi *et al.*, 2024; Mita *et al.*, 2024). Dans cet article, nous cherchons à savoir si ces résultats se généralisent à notre tâche et à connaître l'impact de fournir ou non la référence. Nos contributions sont les suivantes :

— Nous introduisons ParaReval, un jeu de données de préférences humaines sur des paires de révisions générées. [1]
— Nous montrons que les métriques traditionnelles échouent à évaluer la révision de texte avec exactitude.
— Nous montrons que GML-juge peut efficacement estimer le suivi de l'instruction sans avoir besoin de la référence.
— Nous montrons que les métriques de similarité sont complémentaires des approches GML-juge

1. https://github.com/JourdanL/parareval

pour faire face aux cas difficiles à juger.
— Bien que GML-juge offre les meilleures performances, nous montrons que la métrique ParaPLUIE (Lemesle *et al.*, 2025) est une alternative moins coûteuse pour mesurer la préservation du sens.

2 État de l'art

Dans cette section, nous classons les approches d'évaluation en trois catégories : les métriques de similarité des n-grammes, celles de similarité basées sur les plongements et les méthodes GML-juge.

Métriques de similarité des n-grammes Les métriques basées sur la similarité au niveau des n-grammes sont couramment utilisées pour évaluer les tâches de génération de texte dont la révision fait partie. Ces métriques mesurent le chevauchement lexical entre le texte généré et la référence. Cependant, elles ne peuvent pas capturer l'équivalence sémantique ou l'amélioration par rapport au texte original. Les plus fréquemment utilisées sont :

— **BLEU** (Papineni *et al.*, 2002) : Métrique de traduction automatique, utilisée par (Du *et al.*, 2022; Raheja *et al.*, 2024; Jourdan *et al.*, 2024; Dwivedi-Yu *et al.*, 2024; Mücke *et al.*, 2023).
— **ROUGE-L** (Lin, 2004) : Métrique de résumé automatique, utilisée par (Du *et al.*, 2022; Jourdan *et al.*, 2024, 2025; Dwivedi-Yu *et al.*, 2024).
— **METEOR** (Banerjee & Lavie, 2005) : Métrique de correspondance des unigrammes pour la traduction automatique, moins sensible à la paraphrase que BLEU, utilisée par Mücke *et al.* (2023).
— **GLEU** (Napoles *et al.*, 2015) : Variante de BLEU adaptée à la correction grammaticale, utilisée par (Dwivedi-Yu *et al.*, 2024; Raheja *et al.*, 2023). Elle récompense les modifications et zones inchangées correctes et pénalise les modifications grammaticalement incorrectes.
— **SARI** (Xu *et al.*, 2016) : Métrique de simplification de texte, utilisée par (Du *et al.*, 2022; Raheja *et al.*, 2023, 2024; Jourdan *et al.*, 2024, 2025; Dwivedi-Yu *et al.*, 2024). Elle récompense les ajouts, suppressions et conservations de n-grammes corrects.

SARI et GLEU sont les seules métriques qui considèrent la source, ce qui est essentiel pour observer les améliorations apportées par la révision. Bien que les métriques n-grammes offrent l'avantage de l'interprétabilité, elles sont en difficulté pour les tâches nécessitant une compréhension sémantique plus profonde, comme mesurer le suivi de l'instruction ou l'amélioration par rapport au texte original.

Métriques de similarité des plongements Les métriques basées sur les plongements sont conçues pour capturer la similarité sémantique au-delà du chevauchement lexical. Ces méthodes comparent les plongements du texte généré et de référence pour estimer leur alignement. Pour la révision, seul **BERTScore** a été utilisé (Mücke *et al.*, 2023; Jourdan *et al.*, 2024).

Approches GML-Juge Des travaux récents ont exploré l'utilisation des GML-juges pour évaluer les tâches de génération au-delà de la similarité de surface. Ces approches traitent l'évaluation comme une tâche de jugement, où un GML évalue le texte généré sur la base de multiples critères. Plusieurs taxonomies ont été proposées : Gu *et al.* (2024) propose quatre catégories : *Scores*, *Oui ou Non*, *Paires* et *Choix Multiples*. Zheng *et al.* (2023) propose trois catégories : *comparaison par paires*, *notation d'une seule réponse* et *notation guidée par une référence*.

De plus, Doostmohammadi *et al.* (2024) proposent d'évaluer le texte généré selon trois dimensions : *naturalité* (Est-ce que la génération sonne naturelle et fluide ?), *suivi de l'instruction* (Est-ce que

la génération est liée à l'amorce (*prompt*) et est au format requis ?), et *acceptabilité* (Est-ce que la génération est correcte ?, dont la signification varie suivant la tâche). Pour l'évaluation de la révision de texte, Mita *et al.* (2024) proposent une évaluation de comparaison par paires et montrent que cette approche sous-performe comparativement à un classifieur BERT affiné.

Pour nos approches GML-juges, nous nous basons sur ces travaux et proposons des combinaisons des trois approches de (Zheng *et al.*, 2023).

3 Protocole expérimental

Afin d'examiner les limites des métriques de similarité traditionnelles, nous générons d'abord de multiples versions révisées en utilisant différents GMLs et nous les évaluons manuellement.

Jeu de données Nous utilisons la partie d'évaluation du jeu de données `ParaRev` (Jourdan *et al.*, 2025) qui contient 258 paires de paragraphes révisées extraites d'articles scientifiques. Chaque paire est annotée avec deux instructions de révisions différentes, ce qui donne un total de 516 révisions à effectuer par modèle. De plus, chaque paragraphe est étiqueté avec son type d'intention de révision, ce qui sera utilisé dans nos analyses (i.e., `Rewriting+{Light, Medium, Heavy}`, `Concision`, `Content Deletion`). La taxonomie complète est disponible en Annexe A.

Modèles de révision Afin de s'assurer de la diversité en termes de qualité de l'ensemble des révisions, des paragraphes révisés sont générés pour chaque paire de *paragraphe original* + *instruction* en utilisant 6 modèles différents. Les modèles utilisés sont : **CoEdIT-XL**, un modèle T5 affiné pour la tâche de révision de phrases (Raheja *et al.*, 2023), les GMLs semi-ouverts **Llama 3 8B Instruct**, **Llama 3 70B Instruct** et **Mistral 7B Instruct v0.2** et enfin les GMLs fermés **GPT 4o mini** et **GPT 4o**. Les amorces utilisées sont disponibles en Annexe B.

Tâche d'annotation Dans l'objectif d'identifier les métriques qui reflètent au mieux la qualité réelle des révisions, nous les annotons manuellement et comparons ce jugement humain aux scores automatiques. Pour cela, nous créons une tâche d'annotation où l'on compare une paire de révisions candidates et on émet une préférence.

L'annotation a été menée avec l'aide de 10 annotateurs·rices : 3 professeurs et 7 doctorants·tes, non-natifs·ves de l'anglais mais ayant de l'expérience dans la lecture et l'écriture d'articles scientifiques. Chaque instance d'annotation consiste en une paire de suggestions de révision pour un paragraphe donné, accompagnée de l'instruction de révision correspondante. Les annotateurs·rices répondent à une série de questions pour estimer la qualité des révisions :

- Q1A et Q1B **Suivi de l'instruction** : *Did model A/B address the instruction ? {Yes strictly, Yes with additional modifications, No}*
- If it was your article and your instruction :
  - Q2 **Acceptabilité** : *Which revisions would you consider acceptable ? {Both, A, B, None}*
  - Q3 **Préférence** : *Which revision would you prefer to include in your paper ? {Both, A, B, None}*

Les questions spécifiques au type de révision et une capture d'écran de l'environnement d'annotation sont disponibles en Annexe C. Pour garantir une évaluation équitable, nous avons équilibré les comparaisons par paire entre les modèles, en veillant à ce que chaque modèle soit comparé aux autres un nombre similaire de fois.

Résultats de la phase d'annotation

À partir du jeu d'évaluation de ParaRev, 1 548 paires de paragraphes révisés ont été générées pour l'annotation. Parmi elles, 129 paires (8.33%) sont annotées en double pour mesurer l'accord inter-annotateurs·rices. Ces scores d'accord (Cohen's Kappa κ) sont rapportés dans l'Annexe 1.
Dans le cadre de notre analyse, notre objectif étant d'étudier la capacité des métriques à identifier la meilleure révision parmi deux propositions, nous introduisons la notion de Préférence Étendue. Même si *None* est sélectionné lors de l'annotation pour la question de Préférence, une proposition est tout de même considérée préférable si elle est la seule à être *Acceptable* ou à *suivre l'instruction*. Nous considérons les *Both* et *None* restants comme des *ex æquo*.
La Figure 2 présente la distribution des préférences humaines à travers les modèles. Sur la base de l'annotation humaine, GPT-4o émerge comme étant le meilleur modèle de révision, étant strictement favorisé dans 58.33% des comparaisons. Llama 3 70B obtient aussi de bonnes performances, avec un taux de préférence à 53.68%. Lorsque l'on compare ces modèles deux à deux, Llama 3 70B ressort comme étant à égalité avec GPT-4o. Pour plus de détails, on rapporte la préférence des modèles deux à deux en Annexe D.

Question	κ	Accord
Suivi de l'instruction	0.54	Modéré
Acceptabilité	0.55	Modéré
Préférence	0.33	Moyen
Concision	0.22	Moyen
Rewriting light	0.41	Modéré
Rewriting medium	0.48	Modéré

TABLE 1 – Kappa de Cohen (κ) pour chaque question.

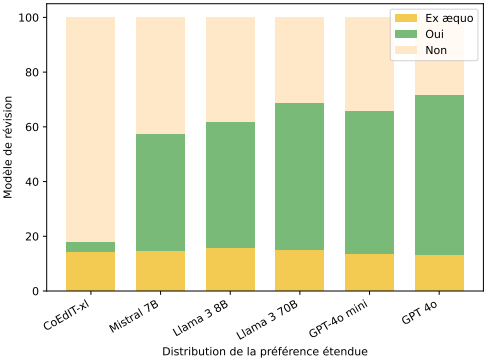

FIGURE 2 – Distribution de la préférence étendue humaine pour chaque modèle de révision. La zone verte correspond aux cas où le modèle est préféré.

4 Limites des métriques basées sur la similarité

Dans cette section, nous évaluons les révisions générées et étudions les faiblesses des métriques basées sur la similarité.

Performances des modèles de révision avec les métriques basées sur la similarité

Afin de déterminer le meilleur modèle de révision, nous évaluons chaque révision générée en utilisant les métriques de similarité traditionnelles en les comparant à une référence. De plus, nous comparons les scores de ces modèles avec CopieEntrée, une approche de contrôle qui recopie l'entrée sans y ajouter de modifications. Les résultats sont présentés dans la Table 2. Toutes les métriques, excepté GLEU, considèrent CopieEntrée comme la meilleure approche, et CoEdIT-XL comme étant un concurrent sérieux. Cependant, après avoir manuellement analysé les résultats, nous remarquons que CoEdIT-XL tend à effectuer des révisions minimales telles que corriger la grammaire et l'orthographe, ou, à l'opposé, supprime de manière excessive une partie du paragraphe. Cela suggère que ces métriques favorisent le fait de ne pas effectuer de changement au lieu de récompenser des révisions plus profondes, et impactantes.

Modèle de révision	BLEU	ROUGE-L	METEOR	GLEU	SARI	Bertscore	BLANC	BETS	ParaPLUIE
CopieEntrée | réf.	**66.00**	**78.30**	**83.80**	25.78	**60.63**	**95.95**	*55.21*	*2.461*	*20.93*
CoEdIT-XL	50.24	67.46	66.66	23.84	39.60	93.90	**58.96**	1.554	19.35
Mistral-7B	27.77	50.79	54.02	15.38	31.63	92.14	41.59	2.491	**23.02**
Llama-3-8B	41.66	62.07	62.00	25.78	39.33	93.53	49.09	2.364	22.67
Llama-3-70B	46.78	65.61	67.20	30.31	42.74	93.90	52.27	2.386	22.58
GPT4o-mini	51.68	69.54	72.70	**32.67**	45.06	94.80	54.89	**2.497**	22.74
GPT4o	49.34	68.20	69.88	31.35	43.54	94.45	53.62	2.454	22.86

TABLE 2 – Résultats initiaux sur la tâche de révision de paragraphes. Le score en **gras** est le plus haut et le score souligné est le deuxième plus haut.

Redondance et corrélation entre métriques

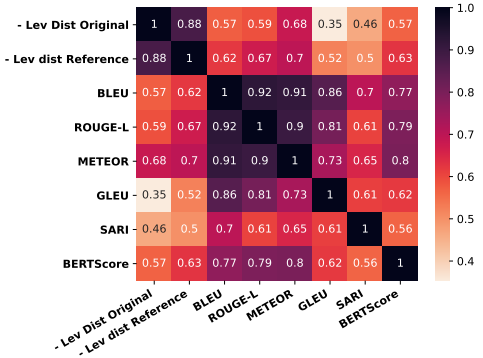

FIGURE 3 – Corrélation des métriques de similarité

Afin d'approfondir ce problème, nous analysons la corrélation entre différentes métriques ainsi que leur relation à la distance d'édition (distance de Levenshtein). Nous calculons la corrélation de Pearson entre toutes les métriques (voir Figure 3) et à la distance de Levenshtein négative. Nous observons que la majorité des métriques sont fortement corrélées, suggérant qu'elles fournissent des informations redondantes. La seule exception est SARI, qui diffère de la plupart des autres métriques, car elle considère le texte original, la révision générée et la référence. Ces résultats suggèrent que, bien qu'on utilise différentes métriques dans l'espoir de pouvoir étudier les révisions sous plusieurs angles, la majorité d'entre elles transmet en réalité la même information.

Sensibilité des métriques à la distance d'édition

Les deux premières colonnes de la Figure 3 montrent une forte corrélation entre les métriques de similarité et la distance d'édition, à la fois en relation aux paragraphes originaux et de référence. Cette relation est davantage illustrée dans l'Annexe E. Nous en tirons deux observations majeures :

— **Tout d'abord, les métriques capturent seulement la similarité de surface**. La forte corrélation avec la distance entre la référence et la révision générée suggère que les métriques traditionnelles reflètent principalement à quel point le modèle parvient à répliquer la révision de référence plutôt qu'évaluer la qualité de la révision elle-même. Même BERTScore, bien que basée sur des plongements et plus coûteuse à calculer, fournit en dernier lieu des informations similaires à des métriques plus simples basées sur la distance.

— **De plus, les révisions substantielles sont pénalisées**. La forte corrélation entre les métriques et la distance entre le texte original et généré indique que plus une révision dévie du paragraphe original, plus son score va être faible. Cela suggère que les métriques traditionnelles ne récompensent pas les changements substantiels et qualitatifs tels que la restructuration de phrases ou l'amélioration de la clarté. Au lieu de cela, elles encouragent des modifications plus conservatrices qui correspondent davantage à la référence.

Ce phénomène crée un biais d'évaluation majeur : les modèles qui produisent des modifications minimales vont recevoir un score haut tandis que les changements valides, mais différents, seront pénalisés. Dans la plupart des cas, ne faire aucune modification permet ainsi d'obtenir un score plus haut qu'en proposant des modifications plus conséquentes, comme le montre l'exemple dans

l'Annexe F. Parmi toutes les métriques, SARI et GLEU se démarquent en ayant une corrélation plus faible à la distance d'édition (≤ 0.52), puisqu'elles pénalisent de façon explicite le texte inchangé, encourageant ainsi la révision.

5 Exploration d'approches alternatives pour l'évaluation

Cette section vise à identifier les métriques d'évaluation les plus corrélées avec l'évaluation humaine de la qualité des textes révisés.

Métriques de domaines connexes du TAL

Nous émettons l'hypothèse que la comparaison au texte original est un facteur essentiel pour la révision de texte, les métriques qui calculent la similarité à une révision de référence tendant à négliger l'amélioration par rapport au texte original. SARI et GLEU sont largement utilisées dans la révision de texte, car elles considèrent à la fois les textes originaux et de référence.

De plus, la révision de texte comprend plusieurs sous-tâches en fonction du type de modifications menées. Raheja *et al.* (2024) a classifié les révisions en trois catégories principales : correction grammaticale, simplification, et paraphrase, en les évaluant avec des ensembles de métriques distincts. Cela suggère qu'une unique métrique peut ne pas être suffisante, étant donné qu'on n'essaie pas de capturer le même phénomène suivant le type de révision.

Nous explorons ainsi des métriques de domaines du TAL proches, en sélectionnant celles qui considèrent le texte original et s'alignent avec différents types de révision (e.g., résumé de texte pour la concision ou paraphrase pour les catégories de re-écriture). Nous identifions trois métriques candidates considérant le paragraphe original et la révision générée en entrée :

— **BETS** (Zhao *et al.*, 2023) : Conçue pour la simplification de texte, estime la préservation du sens et la variation de la simplicité des paires de mots modifiées avec les plongements BERT.
— **BLANC** (Vasilyev *et al.*, 2020) (variante BLANC-help) : Conçue pour le résumé automatique pour remplacer ROUGE. Elle mesure à quel point le résumé aide à la compréhension du texte à l'aide d'un BERT affiné.
— **ParaPLUIE** (Lemesle *et al.*, 2025) : Conçue pour la détection de paraphrase, elle évalue si un texte B est une paraphrase d'un texte A. Elle prompt un Mistral 7B et utilise les scores de perplexité lorsque les réponses *Yes* et *No* sont suggérées au lieu du texte généré.

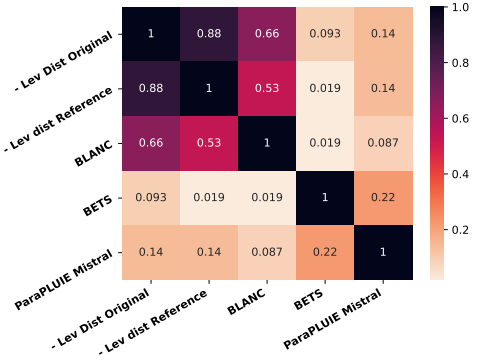

FIGURE 4 – Corrélation des métriques connexes

La Table 2 présente les résultats de l'évaluation à l'aide des trois métriques candidates, nous ajoutons également les résultats de la comparaison entre le paragraphe original et le paragraphe de référence. BETS et ParaPLUIE présentent un classement similaire des modèles et classent CoEdIT-XL en dernière position comme l'annotation humaine. Au contraire, BLANC suit celui des métriques de similarité.
La figure 4 présente les corrélations entre ces nouvelles métriques et la distance de Levenshtein négative. À l'exception de BLANC, ces métriques présentent une faible corrélation avec la distance d'édition. En outre, l'Annexe E permet de visualiser ces relations de corrélation.

BETS et ParaPLUIE apparaissent comme des candidats prometteurs pour l'évaluation de la révision de texte, tandis que BLANC semble moins approprié. Nous étudions davantage leur alignement avec les annotations humaines pour confirmer leur efficacité dans les sections 5 et 6.

GML-Juges

Une hypothèse supplémentaire est que la révision du texte ne doit pas seulement tenir compte du texte original mais aussi évaluer la capacité du modèle à suivre les instructions de manière efficace. Nous explorons les approches GML-Juges pour évaluer ces deux aspects.

Nous expérimentons différentes approches pour le GML-Juge, basées sur le travail de Doostmohammadi *et al.* (2024), qui a utilisé GPT-4 en tant que juge pour évaluer trois critères clés dans le texte généré : 1) *naturalité*, 2) *suivi de l'instruction*, et 3) *acceptabilité*. Puisque notre tâche consiste à modifier un texte existant plutôt qu'à le générer à partir de zéro, *naturalité* n'est pas pertinent pour notre évaluation. Cependant, *suivi de l'instruction*, s'aligne sur Q1A et Q1B de notre annotation humaine et *acceptabilité* s'aligne directement sur Q2. Nous structurons notre amorce de manière similaire à la tâche d'annotation humaine afin d'évaluer ces deux aspects.

Nous explorons deux approches de Gu *et al.* (2024) pour les GML-Juges : *Scores* (*GML-Likert*) où on présente une seule révision à noter, et *Oui ou Non + Paires* (*GML-Choix*) où on présente une paire de révisions et le modèle choisit celle qu'il préfère ou déclare un ex æquo. Comme Doostmohammadi *et al.* (2024) montre une baisse des performances lorsque la référence n'est pas fournie, nous expérimentons ces deux approches avec et sans référence. Les amorces sont fournies en Annexe G.

Nos révisions candidates étant générées par différents GMLs, nous utilisons également différents GMLs comme juges afin de réduire les biais potentiels, où un modèle favoriserait ses propres révisions. En Annexe H, nous analysons les préférences de chaque juge et discutons de ce biais potentiel.

Rev. Model	GML-Choix		GML-Likert	
	base	**+ réf**	**base**	**+ réf**
CoEdIT-XL	6.15	8.82	3.385	3.487
Mistral-7B	**64.95**	56.73	**4.816**	**4.497**
Llama-3-8B	58.88	<u>57.08</u>	<u>4.789</u>	4.472
Llama-3-70B	<u>59.34</u>	**60.20**	4.784	<u>4.479</u>
GPT4o-mini	50.52	52.33	4.750	4.447
GPT4o	50.85	51.41	4.743	4.443

TABLE 3 – Résultats sur la tâche de révision de paragraphe avec l'évaluation par les GML-Juges. Pour GML-choix, on rapporte la préférence étendue et pour GML-likert, la moyenne de tout les critères.

Pour chaque modèle excepté GPT-4o nous présentons des accords moyennés sur trois exécutions en raison des contraintes de coût. De plus, dans le corps principal de l'article, afin de présenter les résultats de manière plus concise, nous calculons la moyenne des modèles par approche.

Les résultats obtenus avec les approches GML-Juges sont présentés en Table 3. Toutes les approches s'accordent avec l'annotation humaine sur le fait que CoEdIT-XL est le modèle de révision le moins performant et que Llama 3 70B est une bonne option pour cette tâche. Cependant, les scores des GML-likert sont très proches les uns des autres.

Résultats

Après avoir identifié les métriques candidates, nous évaluons leur alignement avec le jugement humain en utilisant trois mesures distinctes pour exprimer cet accord : le Kappa de Cohen (κ), le V de Cramér (V), et l'exactitude par paire pour tenir compte des ex æquo (Deutsch *et al.*, 2023).

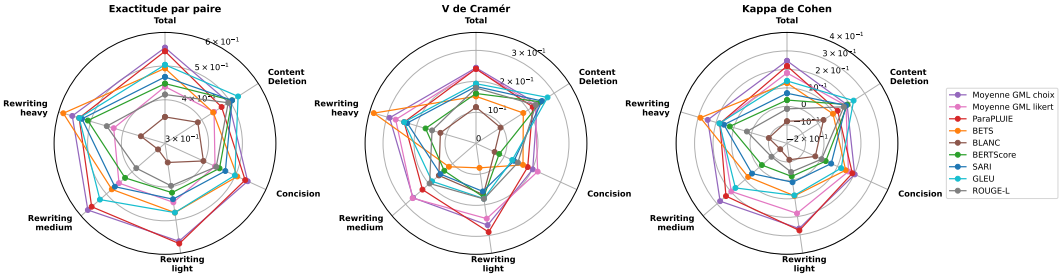

FIGURE 5 – Alignement des métriques automatiques avec les annotations humaines par catégorie de révision

La Table 4 présente l'alignement des méthodes d'évaluation automatique aux jugements humains, on ajoute également une approche aléatoire à titre de comparaison. BLEU, ROUGE et METEOR étant fortement corrélés, on garde uniquement les résultats de ROUGE-L dans le corps principal de l'article.
GML-Choix apparaît comme l'option la plus fiable, suivie par ParaPLUIE. GML-Likert et GLEU présentent également un fort alignement. Cependant, alors que GML-Likert atteint une plus grande précision lorsqu'une décision est prise, elle a tendance à surclasser les cas comme des ex æquo (voir l'Annexe I).
Nous conservons ici uniquement les approches GML-Juges sans référence mais nous étudions l'impact de la disponibilité de la référence sur l'alignement de ces approches en Annexe J.

Juge	Exact.	V	κ
Moy. GML Choix	**0.564**	**0.244**	**0.247**
Moy. GML likert	0.436	0.240	0.181
ParaPLUIE	<u>0.551</u>	<u>0.241</u>	<u>0.218</u>
BETS	0.492	0.152	0.127
BLANC	0.357	0.117	-0.080
BERTScore	0.445	0.161	0.034
SARI	0.465	0.183	0.071
GLEU	0.504	0.193	0.138
ROUGE-L	0.414	0.179	-0.013
Aléatoire	*0.334*	*0.027*	*0.003*

TABLE 4 – Alignement des mesures automatiques sur les jugements humains pour l'ensemble des données. L'exactitude par paire et le V de Cramér sont définis sur [0 :1] et le Kappa de Cohen sur [-1 :1].

6 Performances par aspects

Dans cette section, on analyse les performances à une granularité plus fine sur deux aspects afin de voir si elles varient en fonction du type de révision ou de la difficulté à distinguer la meilleure révision.

Performance par catégorie de révision

Pour tester notre hypothèse de la section 5, nous analysons l'alignement des jugements humains et automatiques sur différents types de révision en utilisant les étiquettes annotées dans `ParaRev` (voir Figure 5). Pour la majorité des catégories, GML-Choix est l'approche la plus fiable. Cependant, dans les cas où le contenu du paragraphe est peu modifié (`Rewriting Light` ou `Medium` et `Concision`), ParaPLUIE semble être une bonne alternative pour capturer la préservation du sens, car elle est moins coûteuse que les approches GML-Juges. En effet, elle a traité notre jeu de données en 11 minutes seulement, contre 1 heure et 22 minutes pour Mistral 7B-Choix sur une GPU V100.

Pour `Content Deletion`, les métriques basées sur la similarité n-grammes telles que GLEU et

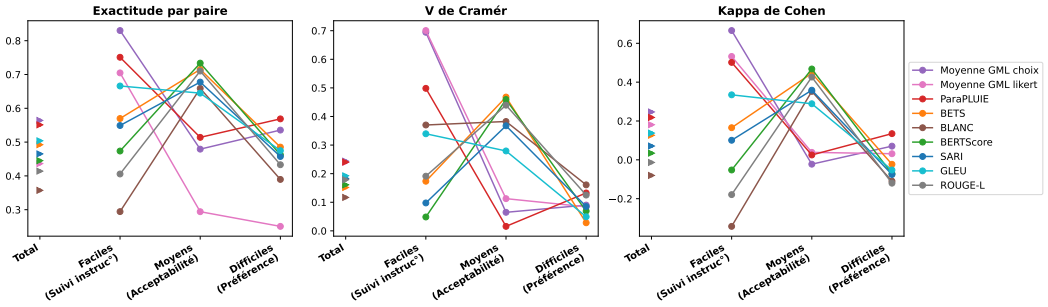

FIGURE 6 – Alignement des métriques automatiques avec les annotations humaines, par difficulté. Les triangles de la première colonne représentent l'accord sur l'ensemble des données de la Table 4

SARI offrent une alternative plus économique, s'alignant aussi bien voire mieux que GML-Choix avec les préférences humaines en tirant parti des informations de suppression basées sur les références.

Enfin, pour `Rewriting Heavy`, BETS surpasse les autres métriques et s'aligne mieux sur l'annotation humaine que dans les autres catégories de révisions. Dans cette catégorie, le sens du paragraphe doit rester le même, tout en faisant l'objet d'une restructuration en profondeur et d'une reformulation de la majeure partie du contenu. Dans le jeu de données, de nombreuses instructions associées à ces paragraphes visent à les rendre plus clairs, plus lisibles ou plus faciles à comprendre, ce qui peut être lié à la tâche de simplification du texte. BETS est un score équilibré entre la préservation du sens et la simplification du texte, ce qui explique probablement ses bonnes performances dans cette catégorie.

Performance par difficulté

Pour mieux évaluer l'efficacité des métriques, nous analysons les performances en fonction du niveau de difficulté de la paire, déterminé par l'annotation humaine :

— Cas Faciles (530 paires) : définis par Q1A et Q1B, un modèle a suivi l'instruction tandis que l'autre ne l'a pas fait.
— Cas Moyens (214 paires) : définis par Q2, les deux modèles ont suivi les instructions, mais un seul a produit une révision acceptable.
— Cas Difficiles (575 paires) : définis par Q3, les deux révisions sont acceptables, mais l'une d'entre elles est préférée.

Nous présentons l'alignement par difficulté dans la Figure 6. Nous observons que GML-Choix est le plus performant dans les cas Faciles, avec une exactitude de 0.821. Cela suggère que les GMLs sont particulièrement efficaces pour reconnaître si une révision suit l'instruction donnée, une capacité qu'aucune des autres métriques ne possède. Cependant, pour les cas Moyens, où les deux révisions sont conformes à l'instruction, les métriques basées sur la similarité surpassent les GMLs dans l'identification de la meilleure option. Nous supposons que ces métriques peuvent exploiter les informations de la référence pour évaluer la révision de manière plus efficace.

Pour les cas Difficiles, aucune des métriques ne donne de bons résultats, toutes les méthodes montrant une faible concordance avec les jugements humains, la tâche devenant encore plus subjective. Dans de telles situations, la préférence pour une révision plutôt qu'une autre peut dépendre largement du style d'écriture et de l'intention de l'auteur, ce qui rend l'évaluation automatique difficile. Cependant, ParaPLUIE semble être la meilleure option pour évaluer ces cas, en garantissant que le sens original du paragraphe est préservé pendant le processus de révision et en détectant d'éventuelles hallucinations.

Puisque les GMLs ont des difficultés avec l'évaluation de l'acceptabilité, nous analysons cet aspect dans l'Annexe K, en corrélant les résultats avec les questions préliminaires d'annotation humaine. Une analyse complète de l'alignement est disponible dans l'annexe L.

7 Discussion et conclusion

Dans cet article, nous avons identifié les métriques les plus fiables pour évaluer la révision de texte scientifique. L'identification de ces métriques pourra permettre meilleure connaissance des stratégies de révision de textes pour par évaluer les solutions existantes et mettre au point des modèles pour cette tâche.

Nos résultats suggèrent que les méthodes de GML-Juges évaluent efficacement si une révision suit l'instruction mais peinent à départager deux bons candidats et ont besoin d'être complétées par d'autres métriques. Les métriques de similarité traditionnelles, bien qu'elles ne soient pas conçues pour évaluer le respect de l'instruction, fournissent, par leur capacité à comparer les révisions à une référence, un mécanisme pour départager lorsque les GMLs ne parviennent pas à faire une distinction claire.

Cependant, les méthodes GML-Juges restent coûteuses en termes de calcul. Nous recommandons ainsi l'utilisation d'un ensemble complémentaire de métriques qui établit un équilibre entre le coût, l'interprétabilité et l'alignement avec le jugement humain. Ce sous-ensemble pourrait inclure un petit GML pour évaluer le suivi des instructions, ParaPLUIE pour la préservation du sens, et des mesures basées sur la similarité comme SARI et GLEU, qui exploitent les informations de la référence pour aider à départager les cas les plus difficiles. [2]

Remerciements

Nous remercions Anas Belfathi, Maël Houbre, Trung Hieu Ngo, Xavier Pillet, Mohamed Reda Marzouk and Thomas Sebbag pour leur participation à l'évaluation humaine.

Nous remercions Thomas Sebbag pour sa relecture du papier avant la soumission.

Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers des allocations de ressources 2023-AD011013901R1, 2024-AD011013901R2 et 2024-AD011014882R1 attribuées par GENCI.

Références

AMANO T., RAMÍREZ-CASTAÑEDA V., BERDEJO-ESPINOLA V., BOROKINI I., CHOWDHURY S., GOLIVETS M., GONZÁLEZ-TRUJILLO J. D., MONTAÑO-CENTELLAS F., PAUDEL K., WHITE R. L. *et al.* (2023). The manifold costs of being a non-native english speaker in science. *PLoS Biology*, **21**(7), e3002184.

BANERJEE S. & LAVIE A. (2005). METEOR : An automatic metric for MT evaluation with improved correlation with human judgments. In J. GOLDSTEIN, A. LAVIE, C.-Y. LIN & C. VOSS,

2. Les sections limites et considérations éthiques présentes dans la version originale sont disponibles en Annexes M et N.

Éds., *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, p. 65–72, Ann Arbor, Michigan : Association for Computational Linguistics.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

DEUTSCH D., FOSTER G. & FREITAG M. (2023). Ties matter : Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 12914–12929.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

DOOSTMOHAMMADI E., HOLMSTRÖM O. & KUHLMANN M. (2024). How reliable are automatic evaluation methods for instruction-tuned LLMs ? In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2024*, p. 6321–6336, Miami, Florida, USA : Association for Computational Linguistics.

DU W., KIM Z. M., RUNDERSTANDAHEJA V., KUMAR D. & KANG D. (2022). Read, revise, repeat : A system demonstration for human-in-the-loop iterative text revision. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, p. 96–108, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.in2writing-1.14.

DWIVEDI-YU J., SCHICK T., JIANG Z., LOMELI M., LEWIS P., IZACARD G., GRAVE E., RIEDEL S. & PETRONI F. (2024). EditEval : An instruction-based benchmark for text improvements. In L. BARAK & M. ALIKHANI, Éds., *Proceedings of the 28th Conference on Computational Natural Language Learning*, p. 69–83, Miami, FL, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.conll-1.7.

GU J., JIANG X., SHI Z., TAN H., ZHAI X., XU C., LI W., SHEN Y., MA S., LIU H. *et al.* (2024). A survey on llm-as-a-judge. *arXiv preprint arXiv :2411.15594*.

ITO T., KURIBAYASHI T., HIDAKA M., SUZUKI J. & INUI K. (2020). Langsmith : An interactive academic text revision system. In Q. LIU & D. SCHLANGEN, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 216–226, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-demos.28.

ITO T., KURIBAYASHI T., KOBAYASHI H., BRASSARD A., HAGIWARA M., SUZUKI J. & INUI K. (2019). Diamonds in the rough : Generating fluent sentences from early-stage drafts for academic writing assistance. In K. VAN DEEMTER, C. LIN & H. TAKAMURA, Éds., *Proceedings of the 12th International Conference on Natural Language Generation*, p. 40–53, Tokyo, Japan : Association for Computational Linguistics. DOI : 10.18653/v1/W19-8606.

JOURDAN L., BOUDIN F., DUFOUR R. & HERNANDEZ N. (2023). Text revision in scientific writing assistance : An overview. In I. FROMMHOLZ, P. MAYR, G. CABANAC, S. VERBERNE & J. BRENNAN, Éds., *13th International Workshop on Bibliometric-enhanced Information Retrieval (BIR)*, volume 3617 de CEUR Workshop Proceedings, p. 22–36, Aachen.

JOURDAN L., BOUDIN F., DUFOUR R., HERNANDEZ N. & AIZAWA A. (2025). ParaRev : Building a dataset for scientific paragraph revision annotated with revision instruction. In M. ZOCK, K. INUI & Z. YUAN, Éds., *Proceedings of the First Workshop on Writing Aids at the Crossroads of AI, Cognitive Science and NLP (WRAICOGS 2025)*, p. 35–44, Abu Dhabi, UAE : International Committee on Computational Linguistics.

JOURDAN L., BOUDIN F., HERNANDEZ N. & DUFOUR R. (2024). CASIMIR : A corpus of scientific articles enhanced with multiple author-integrated revisions. In N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI & N. XUE, Éds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 2883–2892, Torino, Italia : ELRA and ICCL.

KIM Z. M., DU W., RAHEJA V., KUMAR D. & KANG D. (2022). Improving iterative text revision by learning where to edit from other revision tasks. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 9986–9999, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.678.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

LEMESLE Q., CHEVELU J., MARTIN P., LOLIVE D., DELHAY A. & BARBOT N. (2025). Paraphrase generation evaluation powered by an LLM : A semantic metric, not a lexical one. In O. RAMBOW, L. WANNER, M. APIDIANAKI, H. AL-KHALIFA, B. D. EUGENIO & S. SCHOCKAERT, Éds., *Proceedings of the 31st International Conference on Computational Linguistics*, p. 8057–8087, Abu Dhabi, UAE : Association for Computational Linguistics.

LI J., LI Z., GE T., KING I. & LYU M. (2022). Text revision by on-the-fly representation optimization. In T.-H. K. HUANG, V. RAHEJA, D. KANG, J. J. Y. CHUNG, D. GISSIN, M. LEE & K. I. GERO, Éds., *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, p. 58–59, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.in2writing-1.7.

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.

MITA M., SAKAGUCHI K., HAGIWARA M., MIZUMOTO T., SUZUKI J. & INUI K. (2024). Towards automated document revision : Grammatical error correction, fluency edits, and beyond. In E. KOCHMAR, M. BEXTE, J. BURSTEIN, A. HORBACH, R. LAARMANN-QUANTE, A. TACK, V. YANEVA & Z. YUAN, Éds., *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, p. 251–265, Mexico City, Mexico : Association for Computational Linguistics.

MIZRAHI M., KAPLAN G., MALKIN D., DROR R., SHAHAF D. & STANOVSKY G. (2024). State of what art ? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, **12**, 933–949. DOI : 10.1162/tacl_a_00681.

MÜCKE J., WALDOW D., METZGER L., SCHAUZ P., HOFFMAN M., LELL N. & SCHERP A. (2023). Fine-tuning language models for scientific writing support. In A. HOLZINGER, P. KIESEBERG, F. CABITZA, A. CAMPAGNER, A. M. TJOA & E. WEIPPL, Éds., *Machine Learning and Knowledge Extraction*, p. 301–318, Cham : Springer Nature Switzerland.

NAPOLES C., SAKAGUCHI K., POST M. & TETREAULT J. (2015). Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2 : Short Papers)*, p. 588–593, Beijing, China : Association for Computational Linguistics.

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, Philadelphia, Pennsylvania, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

RAHEJA V., ALIKANIOTIS D., KULKARNI V., ALHAFNI B. & KUMAR D. (2024). mEdIT : Multilingual text editing via instruction tuning. In K. DUH, H. GOMEZ & S. BETHARD, Éds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, p. 979–1001, Mexico City, Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2024.naacl-long.56.

RAHEJA V., KUMAR D., KOO R. & KANG D. (2023). CoEdIT : Text editing by task-specific instruction tuning. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 5274–5291, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.350.

SCHICK T., YU J. A., JIANG Z., PETRONI F., LEWIS P., IZACARD G., YOU Q., NALMPANTIS C., GRAVE E. & RIEDEL S. (2023). PEER : A collaborative language model. In *The Eleventh International Conference on Learning Representations*.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

VASILYEV O., DHARNIDHARKA V. & BOHANNON J. (2020). Fill in the BLANC : Human-free quality estimation of document summaries. In S. EGER, Y. GAO, M. PEYRARD, W. ZHAO & E. HOVY, Éds., *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, p. 11–20, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.eval4nlp-1.2.

XU W., NAPOLES C., PAVLICK E., CHEN Q. & CALLISON-BURCH C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, **4**, 401–415. DOI : 10.1162/tacl_a_00107.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2020). Bertscore : Evaluating text generation with bert. In *International Conference on Learning Representations*.

ZHAO X., DURMUS E. & YEUNG D.-Y. (2023). Towards reference-free text simplification evaluation with a BERT Siamese network architecture. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Findings of the Association for Computational Linguistics : ACL 2023*, p. 13250–13264, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-acl.838.

ZHENG L., CHIANG W.-L., SHENG Y., ZHUANG S., WU Z., ZHUANG Y., LIN Z., LI Z., LI D., XING E. P., ZHANG H., GONZALEZ J. E. & STOICA I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA : Curran Associates Inc.

A Taxonomie de ParaRev

Voir Table 5, seules les catégories présentes dans le jeu d'évaluation sont listées.

Type		**Description**
Rewriting	Light	Changements mineurs dans le choix des mots ou de la formulation.
	Medium	Reformulation complète de phrases.
	Heavy	Reformulation significative affectant au moins la moitié du paragraphe.
Concision		Même idée, formulée de façon plus concise en supprimant des détails non nécessaires.
Content	Deletion	Modification du contenu par la suppression d'une idée.

TABLE 5 – Taxonomie des catégories de au niveau du paragraphe.

B Amorce pour la révision de texte

Segment d'amorce 1 : Messages de l'amorce pour la révision de texte

```
system_message
= """You are a writing assistant specialised in academic writing.
Your task is to revise the
 original paragraph from a research paper draft that will be given
 according to the author's instruction. The input will follow the
 pattern ' <author instruction> : "<Original paragraph>" '.
Please answer only by "Revised paragraph: <
revised version of the paragraph>". Please limit your modifications
 only to what is requested in the author's instruction. Do
 not make any other modifications to the rest of the paragraph."""

user_message= """{instruct} : \"{parag}\" """
```

C Environnement d'annotation et questions supplémentaires

Voir Figure 7 pour l'environnement d'annotation.

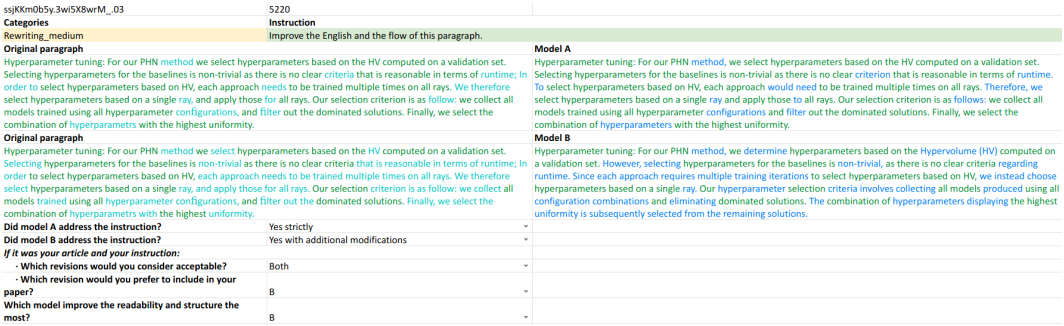

FIGURE 7 – Capture d'écran de l'environnement d'annotation

Les questions supplémentaires, spécifiques à chaque type de révision sont les suivantes, les réponses possibles pour chaque question sont *{Both, A, B, None}* :

- Rewriting light : *Which model improves the academic style and English the most ?* (Quel modèle améliore le plus le style académique et l'anglais ?)
- Rewriting medium : *Which model improves the readability and structure the most ?* (Quel modèle améliore le plus la lisibilité et la structure ?)
- Rewriting heavy : *Which model improves the readability and clarity the most ?* (Quel modèle améliore le plus la lisibilité et la clarté ?)
- Concision : *Which model manages the most to give a shorter version while keeping all the important ideas ?* (Quel modèle parvient au mieux à fournir une version plus courte tout en conservant les idées importantes ?)

D Préférences humaines des modèles de révision par paires

Voir Figure 8.

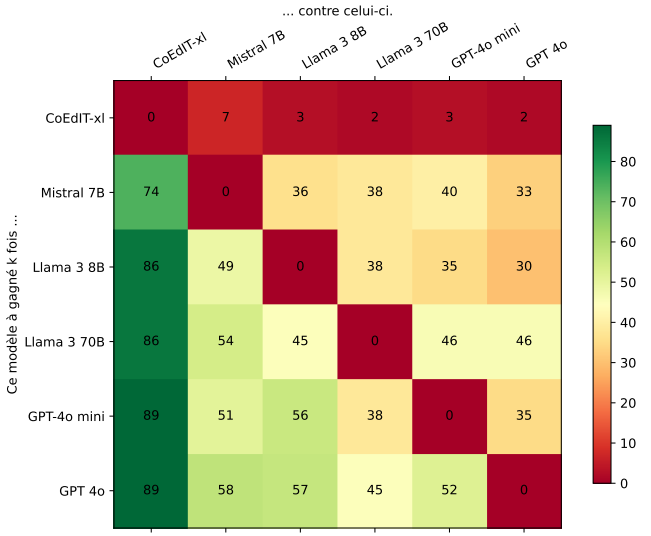

FIGURE 8 – Comparaison par paire des préférences humaines sur les modèles de révision.

E Relation entre les métriques et la distance de Levenshtein

Voir les Figures 9, 10, 11 and 12

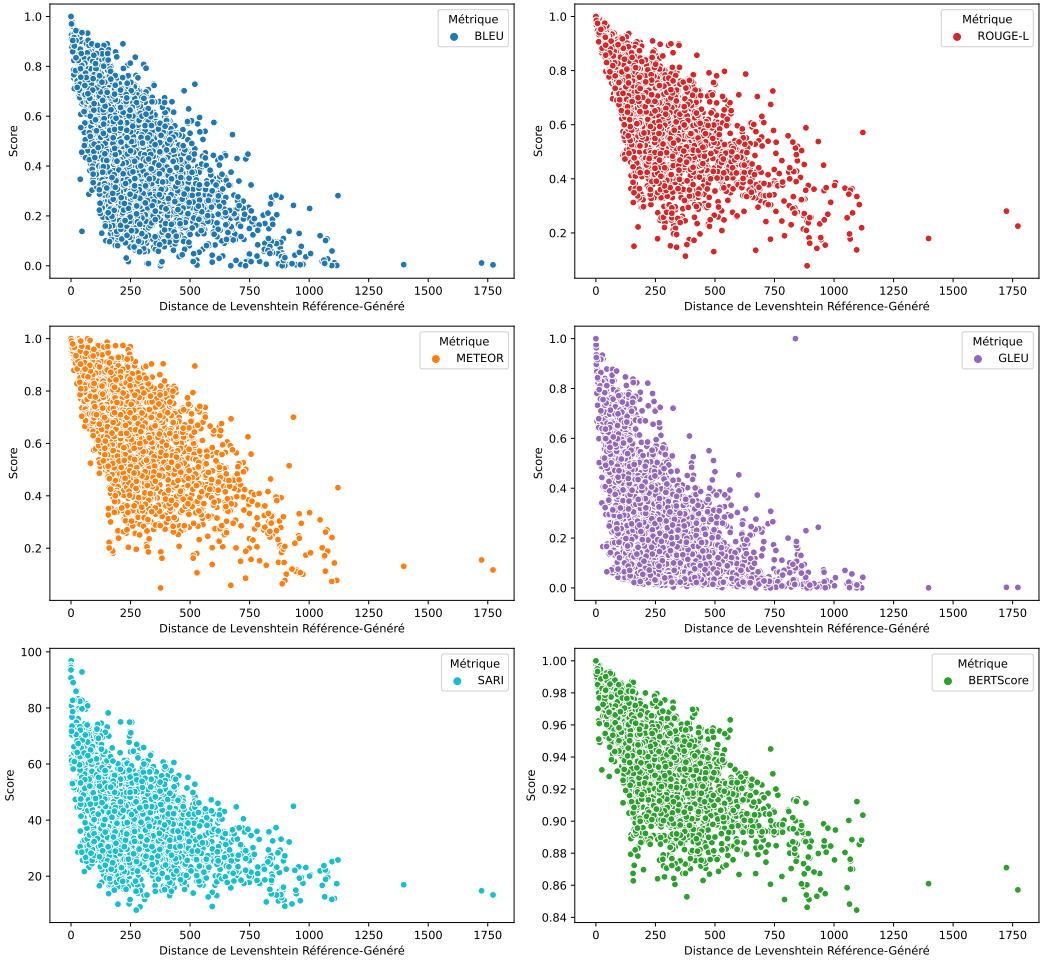

FIGURE 9 – Distribution des scores des métriques de similarité selon la distance de Levenshtein entre le paragraphe généré et de référence.

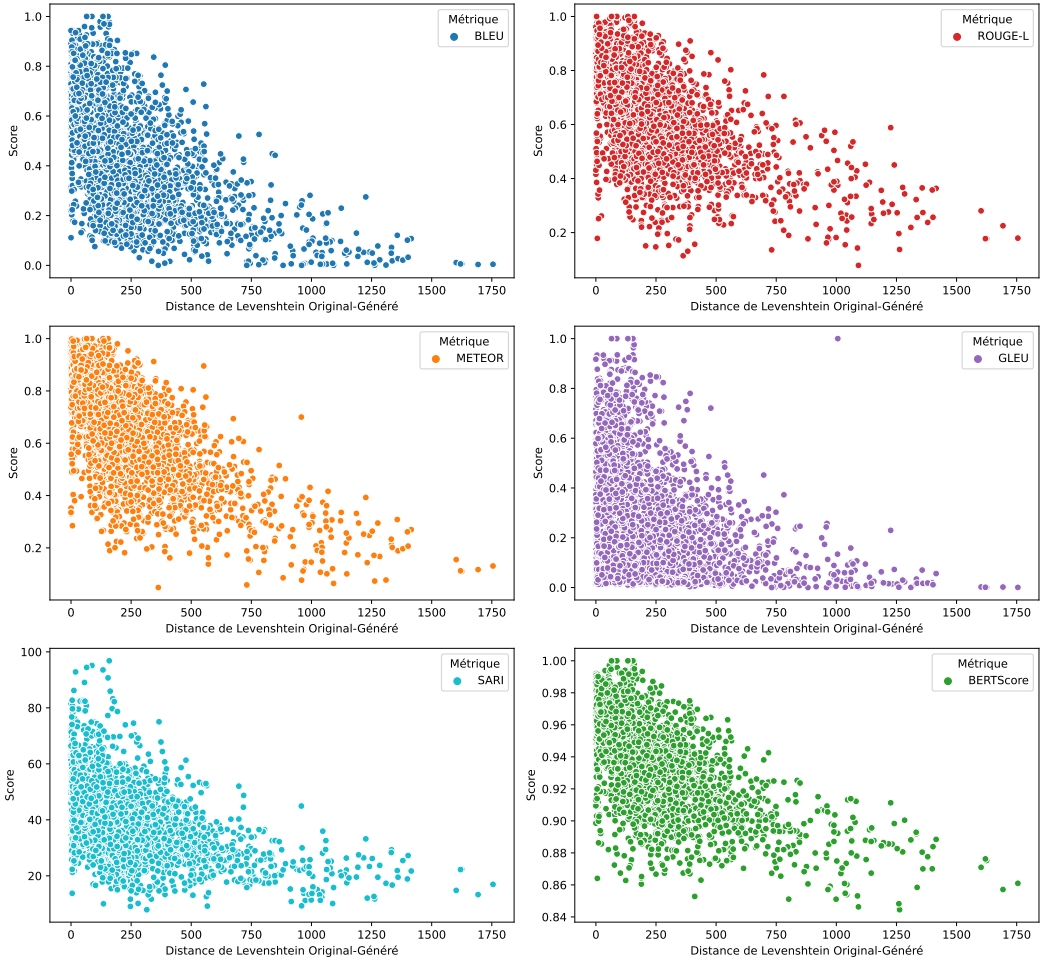

FIGURE 10 – Distribution des scores des métriques de similarité selon la distance de Levenshtein entre le paragraphe original et généré.

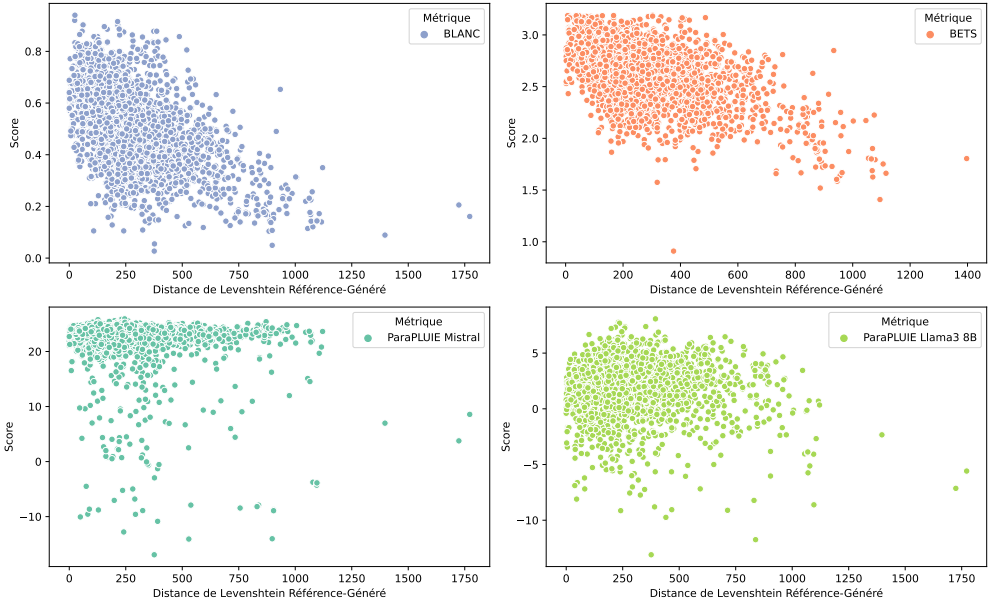

FIGURE 11 – Distribution des scores des métriques alternatives selon la distance de Levenshtein entre le paragraphe généré et de référence.

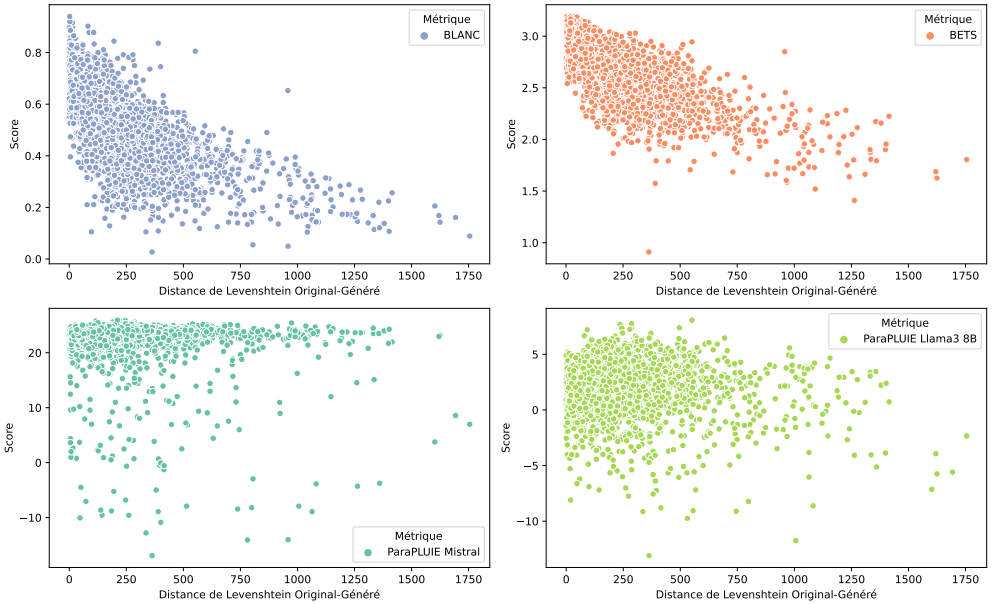

FIGURE 12 – Distribution des scores des métriques alternatives selon la distance de Levenshtein entre le paragraphe original et généré.

F Exemple d'évaluation automatique ne s'alignant pas avec le jugement humain

Voir Figure 13

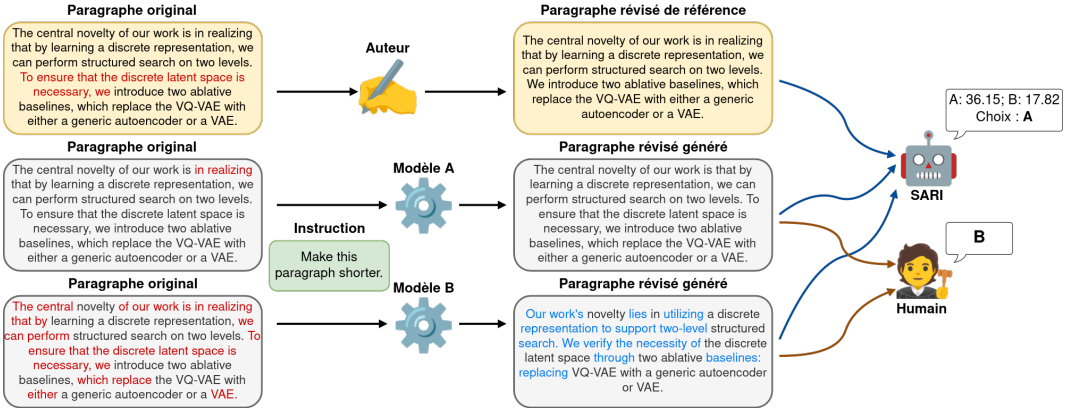

FIGURE 13 – Aperçu de l'évaluation dans un cas où l'évaluation automatique (ici SARI) et le jugement humain ne s'alignent pas.

G Amorces utilisées pour les GML-Juges

Nos amorces sont inspirées par celles utilisées dans Doostmohammadi *et al.* (2024).

Segment d'amorce 2 : Message système GML-Choix sans référence

```
You
 are an evaluator of academic writing on the task of text revision.
In this task, two revision models have
 been provided with the original paragraph written for a scientific
 article and a revision instruction on how to revise the paragraph.
You will be given the proposition
 from the two different models and several questions to determine
 the quality of those propositions and identify the best one
. In your answer please only provide the answers to the questions.
```

Segment d'amorce 3 : Message système GML-Choix avec référence

You are an evaluator of academic writing on the task of text revision. In this task, two revisions models have been provided with the original paragraph written for a scientific article and a revision instruction on how to revise the paragraph. You will be given the proposition from the two different models and several questions to determine the quality of those proposition and identify the best one. **To help you in this task you will also be given the gold paragraph which is the version revised by the author themselves.** In your answer please only provide the answers to the questions.

Segment d'amorce 4 : Questions spécifiques aux catégories de révisions pour GML-Choix avec et sans référence

```
Rewriting_light:"""
Which model improves the academic style and English the most?"""

Rewriting_medium
:"""Which model improves the readability and structure the most?"""

Rewriting_heavy
:"""Which model improves the readability and clarity the most?"""

Concision:"""Which model manages the most
to give a shorter version while keeping all the important ideas?"""
```

Segment d'amorce 5 : Message utilisateur GML-Choix avec et sans référence

```
[BEGIN DATA]
***
[Original paragraph]: \"{original_paragraph}\"
***
[Revision instruction]: \"{instruction}\"
***
[Model A]: \"{modelA_generated_revised_paragraph}\"
***
[Model B]: \"{modelB_generated_revised_paragraph}\"
***
[END DATA]

1. Did model A address the instruction?
Answer "Yes strictly", "Yes with additional modifications" or "No":
 -
 Yes strictly : The model proposition matches what is required in
the instruction. Here, the quality of the revision does not matter.
 -
 Yes with additional modifications : The model proposed additional
 modifications to the one required in the instruction. But some
 of the modification address the needs stated in the instruction.
- No : The model proposition does not match the instruction.

2. Did model B address the instruction? (
Answer "Yes strictly", "Yes with additional modifications" or "No")

3. Is model A revision acceptable? Answer "Yes" or "No". Answer
 "Yes" if the model made a good quality revision proposition that
 should replace the original paragraph in the scientific article.

4. Is model B revision acceptable? (Answer "Yes" or "No")

5. Which model proposed the best revision? (
Answer preferably "A" or "B", you can answer "Both" if it is really
 a tie. Answer "None" if you answered "No" to question 3 and 4.)"""
<Additional category questions
 depending on the revision intention labels of the instance>

"""For all questions, you do not need to explain the reason.

Your response must be RFC8259 compliant JSON following this schema:
{{"1": str, "2": str , "3": str , "4": str , "5": str """
< """ "6": str """ and """, "7": str """
 can be added depending on the number of labels of the instance.>
"""}}
```

Segment d'amorce 6 : Questions spécifiques aux catégories de révisions pour GML-Likert avec et sans référence

```
Rewriting_light
: """The academic style and english has been improved."""

Rewriting_medium
: """The readability and structure has been improved."""

Rewriting_heavy: """The paragraph has been rewritten in a more
 well organized and clear version, fitting the academic style."""

Concision: """The generated
revision is a shorter version that kept all the important ideas."""
```

Segment d'amorce 7 : Message système GML-Likert sans référence

You are an evaluator of academic writing on the task of text revision. In this task, **a revision model** have been provided with the original paragraph written for a scientific article and a revision instruction on how to revise the paragraph.
You will be given the proposition from the **revision model** and several **affirmations** to determine the quality of **this** proposition.
You will answer each affirmation with a grade (int) from 1 to 5 as following: 1 = Strongly disagree , 2 = Disagree , 3 = Neutral ,4 = Agree , 5 = Strongly agree
In your answer please only provide the answers to the **affirmations**.

Segment d'amorce 8 : Message Utilisateur GML-Likert sans référence

```
[BEGIN DATA]
***
[Original paragraph]: \"{original_paragraph}\"
***
[Revision instruction]: \"{instruction}\"
***
[Model proposed revision]: \"{model_generated_revised_paragraph}\"
***
[END DATA]

1. Relatedness: The generated revision correctly addressed the instruction.

2. Correctness: The generated revision is better than original version in my
opinion."""
<Additional category questions
 depending on the revision intention labels of the instance>

"""For all questions, you do not need to explain the reason.

Your response must be RFC8259 compliant JSON following this schema:
{{"1": str, "2": str , "3": str , "4": str , "5": str """
< """ "6": str """ and """, "7": str """
 can be added depending on the number of labels of the instance.>
"""}}
```

Segment d'amorce 9 : Message système GML-Likert avec référence

```
You
 are an evaluator of academic writing on the task of text revision.
In this task, a revision model have
 been provided with the original paragraph written for a scientific
 article and a revision instruction on how to revise the paragraph.
You will be given the proposition from the revision model and
 several affirmations to determine the quality of this proposition.
You will answer each affirmation
 with a grade (int) from 1 to 5 as following: 1 = Strongly disagree
 , 2 = Disagree , 3 = Neutral ,4 = Agree , 5 = Strongly agree
To help you in this task you will also be given the gold paragraph which is the
version revised by the author themselves.
In your answer please only provide the answers to the affirmations.
```

Segment d'amorce 10 : Message utilisateur GML-Likert avec référence

```
[BEGIN DATA]
***
[Original paragraph]: \"{original_paragraph}\"
***
[Revision instruction]: \"{instruction}\"
***
[Model proposed revision]: \"{model_generated_revised_paragraph}\"
***
[Gold revised paragraph]: \"{gold}\"
***
[END DATA]

1. Gold similarity: The generated revision is similar to gold revision.

2. Relatedness: The generated revision correctly addressed the
instruction.

3. Correctness: The generated revision is better than original
version in my opinion."""
<Additional category questions
 depending on the revision intention labels of the instance>

"""For all questions, you do not need to explain the reason.

Your response must be RFC8259 compliant JSON following this schema:
{{"1": str, "2": str , "3": str , "4": str , "5": str """
< """ "6": str """ and """, "7": str """
 can be added depending on the number of labels of the instance.>
"""}}
```

H Biais des modèles GMLs sur leurs propres révisions

Comme nous avons utilisé plusieurs GML comme juges que nous avions déjà utilisés pour la révision, nous vérifions dans la Table 6 s'ils sont biaisés en faveur de leur propre proposition. Nous n'observons pas un tel biais, et nous remarquons même que les résultats ont tendance à être cohérents entre les modèles de juges. Cependant, comme ils ont tous tendance à favoriser Mistral 7B nous avons également calculé la distance d'édition moyenne entre le texte original et les textes révisés générés pour tous les modèles de révision. Comme Mistral a la moyenne la plus élevée, cela pourrait indiquer un biais opposé à celui véhiculé par les mesures de similarité : Les approches GML en tant que juge ont tendance à favoriser les propositions avec des révisions plus importantes.

Juge↓/Modèle de révision→	CoEdIT	Mistral 7B	Llama3 8B	Llama3 70B	GPT-4o mini	GPT 4o
Humain	3.29	42.83	46.12	53.68	52.13	**58.33**
Mistral 7B Choix	21.58	**60.92**	56.20	55.17	50.65	52.65
Llama 3 8B Choix	3.94	**64.66**	58.98	62.02	58.08	52.07
Llama 3 70B Choix	1.49	**73.00**	61.76	59.23	46.06	48.13
GPT-4o mini Choix	1.81	**66.86**	58.53	57.30	47.61	48.90
GPT-4o Choix	1.94	59.30	58.91	**62.98**	50.19	52.52
Mistral 7B Gold Choix	23.64	54.65	56.33	**56.52**	51.16	54.07
Llama 3 8B Gold Choix	14.67	51.62	59.24	**60.06**	53.68	50.74
Llama 3 70B Gold Choix	1.49	**64.47**	58.85	61.95	51.10	48.19
GPT-4o mini Gold Choix	2.00	**62.73**	58.08	59.50	51.61	51.16
GPT-4o Gold Choix	2.33	50.19	52.91	**62.98**	54.07	52.91
Mistral 7B Likert	5.04	25.65	24.55	23.84	**25.97**	23.25
Llama 3 8B Likert	6.98	38.11	**39.99**	36.24	38.56	36.05
Llama 3 70B Likert	2.26	**40.38**	37.40	32.69	29.20	30.10
GPT-4o mini Likert	1.87	**36.95**	33.85	31.98	31.52	29.39
GPT-4o Likert	1.36	41.28	47.87	**49.61**	47.67	46.90
Mistral 7B Likert Gold	11.24	18.41	17.89	19.64	19.05	**19.77**
Llama 3 8B Likert Gold	10.92	**46.25**	41.15	42.96	42.18	39.60
Llama 3 70B Likert Gold	2.45	46.19	**47.22**	44.51	39.28	38.76
GPT-4o mini Likert Gold	2.58	44.06	43.99	45.41	**46.51**	44.19
GPT-4o Likert Gold	3.88	42.25	49.22	49.22	48.45	**51.16**
ParaPLUIE Mistral	17.83	**72.48**	56.20	52.91	47.67	50.39
ParaPLUIE Llama3 8B	20.35	**70.74**	52.33	54.84	52.33	46.90
Distance d'édition (Original-Generated)	190.82	**342.69**	270.47	234.95	175.02	197.36

TABLE 6 – Distribution de la préférence étendue stricte (pas d'ex æquo) de chaque GML-juge par modèle de révision.

I Distribution des préférences étendues de chaque GML-Juge

Voir Table 7.

Juge↓/Choix→	Ex æquo	A	B
Humain	14.53	44.25	41.21
Mistral 7B Choix	0.95	82.43	16.63
Llama 3 8B Choix	0.08	53.92	46.00
Llama 3 70B Choix	03.45	52.97	43.58
GPT-4o mini Choix	06.33	39.28	54.39
GPT-4o Choix	04.72	53.62	41.67
Mistral 7B Gold Choix	01.21	77.76	21.04
Llama 3 8B Gold Choix	03.33	23.79	72.85
Llama 3 70B Gold Choix	04.65	51.14	44.21
GPT-4o mini Gold Choix	4.98	39.27	55.75
GPT-4o Gold Choix	08.21	49.94	41.86
Mistral 7B Likert	57.24	22.37	20.39
Llama 3 8B Likert	34.69	34.28	31.03
Llama 3 70B Likert	42.66	28.71	28.64
GPT-4o mini Likert	44.81	27.99	27.2
GPT-4o Likert	21.77	39.02	39.21
Mistral 7B Likert Gold	64.66	17.29	18.05
Llama 3 8B Likert Gold	25.65	38.20	36.16
Llama 3 70B Likert Gold	27.20	35.44	37.36
GPT-4o mini Likert Gold	24.42	37.96	37.62
GPT-4o Likert Gold	18.60	41.54	39.86

TABLE 7 – Distribution de la préférence étendue pour chaque GML-juge

J Impact de la disponibilité de la référence pour les approches GML-Juges

Nous cherchons à savoir si la disponibilité de la référence influence la performance des méthodes GML-Juge et nous trouvons un impact minime, voir Figures 14 et 15. L'exactitude de GML-Choix a légèrement variée de 0,564 à 0,563 lorsque la référence est fournie et de 0,436 à 0,457 pour GML-Likert. Nos résultats contrastent avec Doostmohammadi *et al.* (2024), qui indique qu'en l'absence d'une référence, GPT-4o présente un alignement plus faible avec les jugements humains. Ceci suggère que les GMLs s'appuient fortement sur leur propre raisonnement interne plutôt que sur des comparaisons directes avec une révision de référence.

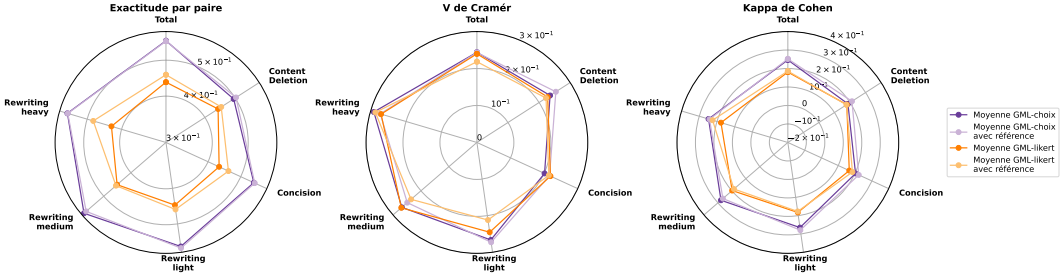

FIGURE 14 – Alignement des approches GML-juge avec les annotations humaines par catégories de révision.

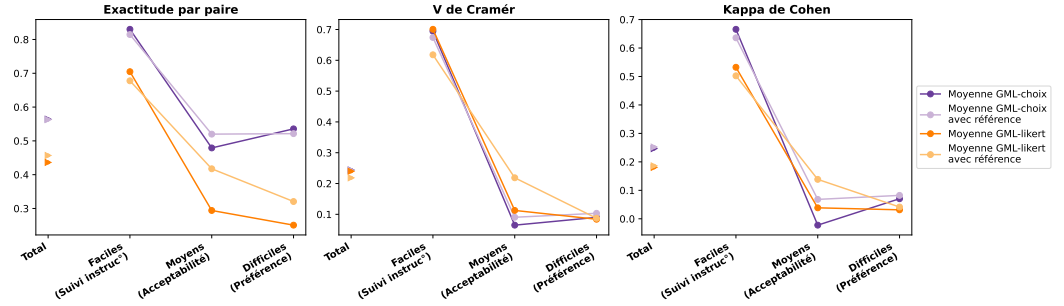

FIGURE 15 – Alignement des approches GML-juge approches avec les annotations humaines par difficulté. Les triangles dans la première colonne représentent l'accord sur le jeu de données complet.

K Distribution du *suivi de l'instruction* et de l'*accéptabilité* pour les approches GML-Juges

Voir Table 8

Modèle	Suivi de l'instruction (Exactitude Stricte)	Suivi de l'instruction (Exactitude souple)	Acceptabilité (Exactitude)
gpt-4o	**67.03**	84.69	76.78
gpt-4o-gold	62.99	84.66	**76.94**
gpt-4o-mini	62.26	**85.30**	76.04
gpt-4o-mini-gold	58.95	85.14	75.91
llama3-70b	66.66	82.11	75.96
llama3-70b-gold	66.48	82.37	76.10
llama3-8b	45.41	80.16	72.30
llama3-8b-gold	40.55	77.70	61.93
mistral-gold	58.78	75.19	62.37
mistral	58.98	75.27	62.58

TABLE 8 – Exactitude de GML-Choix sur les questions préliminaires : *suivi de l'instruction (Relatedness)* et de l'*accéptabilité (Correctness)*. Pour le *suivi de l'instruction*, en exactitude souple, on fusionne les catégories "Yes stricly" et "Yes with additional modifications".

L Alignement de toutes les métriques

Voir les Figures 16 et 17

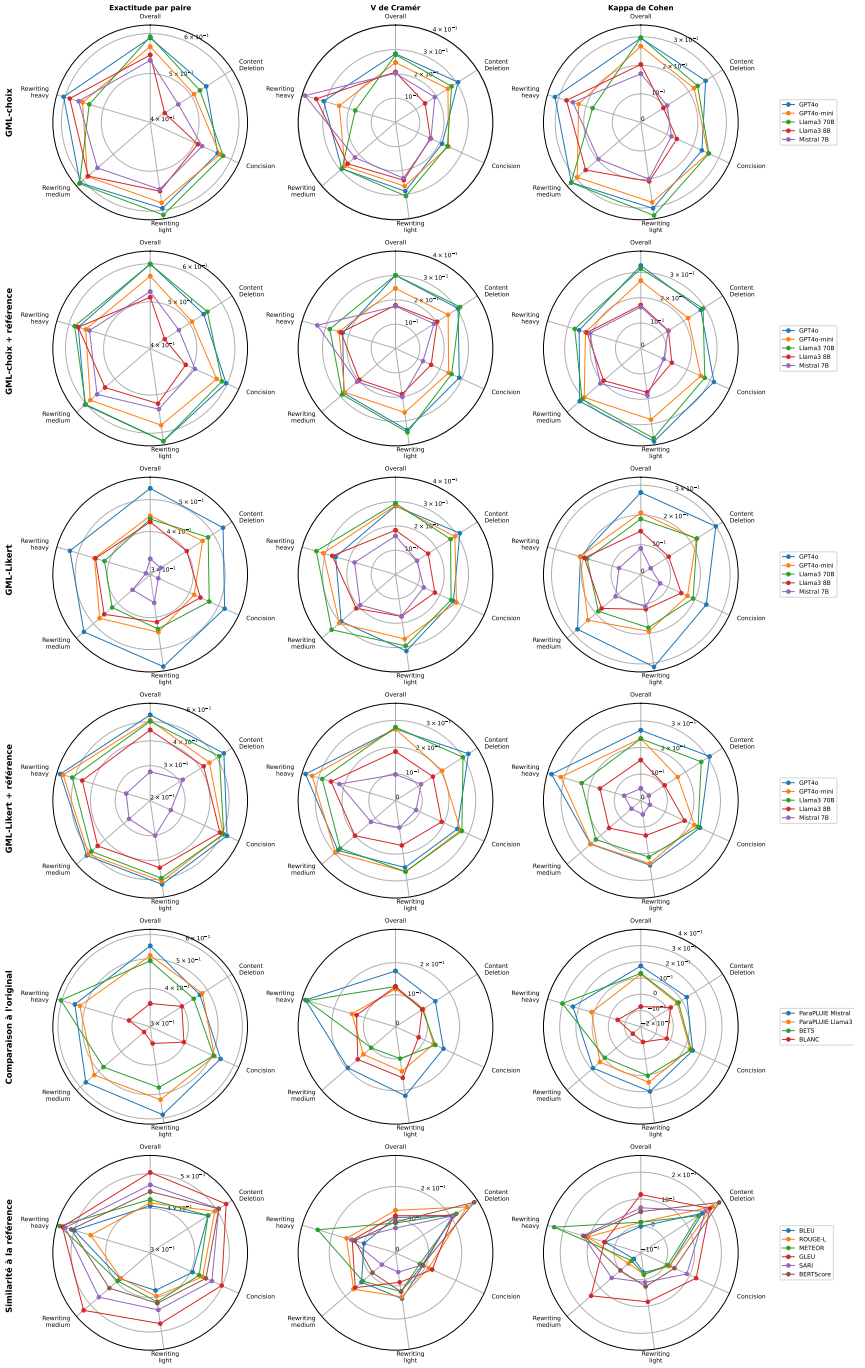

FIGURE 16 – Alignement des métriques aux annotations humaines par type de métrique et de révision.

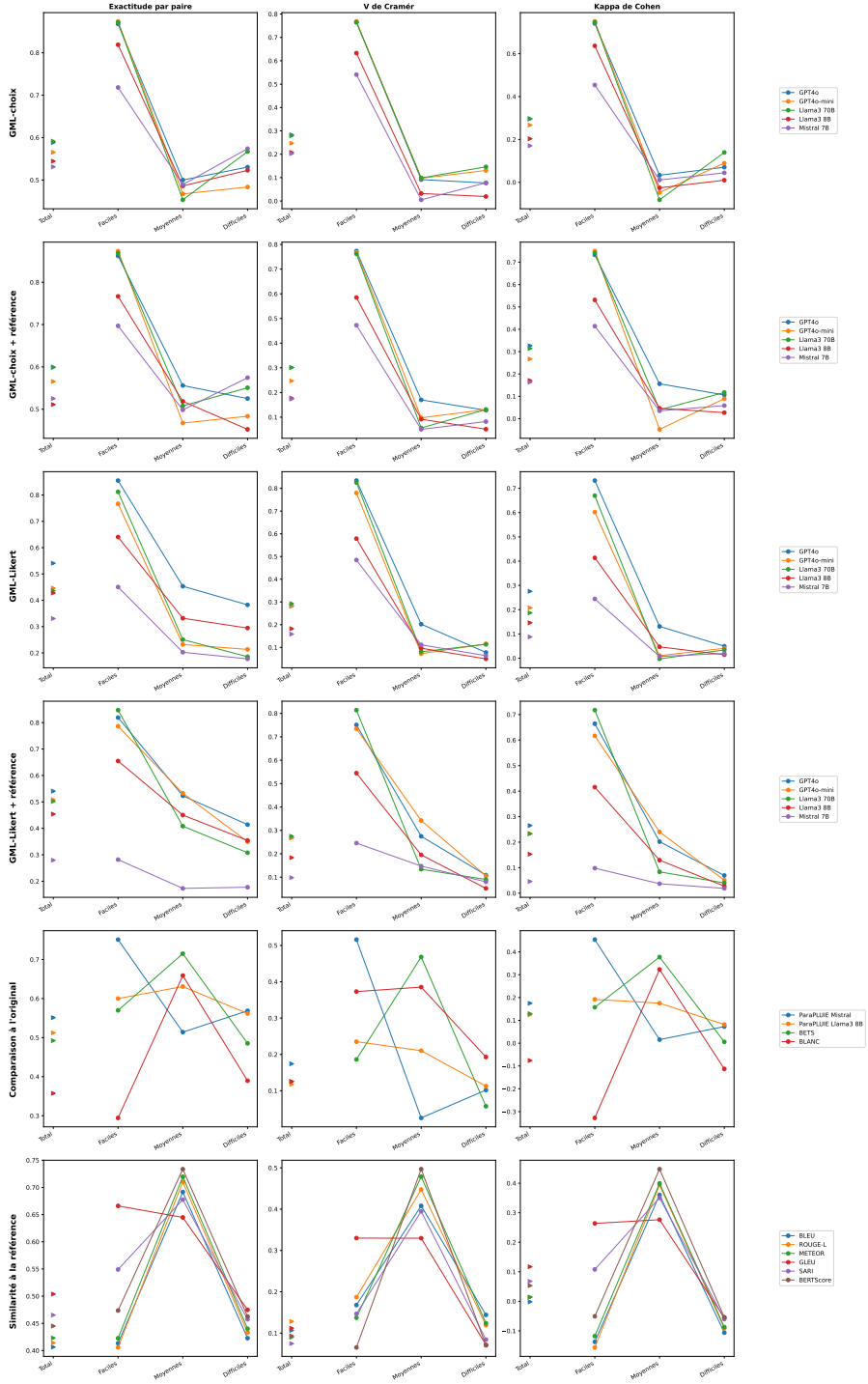

FIGURE 17 – Alignement des métriques aux annotations humaines par type de métrique et par difficulté.

M Limites

La principale limite de ce travail est la taille du jeu de données, car nous étions limités par la taille du jeu l'évaluation de ParaRev et nous ne disposions que d'un nombre limité de chercheurs volontaires pour l'annotation manuelle. Cette annotation devant être effectuée par des annotateurs·rices qualifiés·ées (des chercheurs·euses), nos ressources humaines sont très limitées conduisant à un jeu double annoté plus petit. Une plus grande quantité de données annotées augmenterait la fiabilité de notre analyse, renforçant ainsi les observations que nous avons faites dans ce papier.

En outre, l'annotation basée sur les préférences est intrinsèquement subjective, comme le montrent les scores Kappa de Cohen dans la Table 1. Pour le jeu de données ParaReval, nous avons d'abord annoté un sous-ensemble avec des doubles annotations et conservé les annotations des chercheurs dont l'accord était le plus élevé. Ceux qui ont obtenu les scores d'accord les plus élevés ont poursuivi le processus d'annotation afin d'améliorer la fiabilité. Le choix des annotateurs (domaine du TAL uniquement) peut également introduire un biais.

Nous avons également été limités par les ressources de calcul, as GMLs utilisant beaucoup d'énergie. Notre priorité a été de s'assurer que nos résultats sont indépendants du modèle utilisé, nous avons donc effectué la même expérience avec différents modèles mais en utilisant un seul prompt. Cependant,il aurait été préférable de s'assurer également de l'aspect indépendent du prompt des résultats (Mizrahi *et al.*, 2024).

Enfin, de nombreuses méthodes et métriques ont été proposées pour évaluer les tâches de génération de texte au fil des années. Pour garder notre analyse claire, nous en avons considéré un nombre limité. Pour les approches basées sur le GML, un plus grand nombre d'exécutions aurait été préférable pour GPT-4o mais pour des raisons de coût, nous avons dû le limiter à une exécution par approche.

N Considérations éthiques

Disponibilité des données Toutes les données proviennent du corpus ParaRev les paragraphes sont extraits d'articles scientifiques collectés sur OpenReview où ils sont soumis à différentes "non-exclusive, perpetual, and royalty-free license" [3].

Ressources de calcul

— Pour générer des révisions avec Co-edit et expérimenter avec les métriques basées sur BERT, nous avons utilisé un GPU local GeForce RTX 2080 11Go pendant environ 12 heures.
— Pour utiliser ParaPLUIE et les différents GMLs ouverts pour générer des révisions et des évaluations, nous avons utilisé des GPU V100 et A100 pendant un total de 249 heures sur un supercalculateur, équivalant à 0,009 tonne de CO_2.
— Pour utiliser GPT-4o mini et GPT-4o nous avons dépensé 29,01$ en crédits API GPT.

3. https://openreview.net/legal/terms