Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

*-PLUIE : une mesure personnalisable utilisant un LLM pour une meilleure évaluation

Quentin Lemesle[1], Léane Jourdan[2], Daisy Munson[3], Pierre Alain[3], Jonathan Chevelu[1], Arnaud Delhay[1], Damien Lolive[4]

[1]Univ Rennes, CNRS, IRISA, EXPRESSION, 22300 Lannion, France
[2]Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
[3]Univ Rennes, CNRS, IRISA, SOTERN, 22300 Lannion, France
[4]Univ Bretagne Sud, CNRS, IRISA, ARCHIMEDIA, 56000 Vannes, France
Correspondence: quentin.lemesle@irisa.fr

RÉSUMÉ

L'évaluation de la qualité d'un texte généré automatiquement repose souvent sur des méthodes *LLM-as-a-judge* (*LLM*-juge). Bien qu'efficaces, ces approches sont coûteuses en termes de temps de calcul et nécessitent un post-traitement. Pour pallier ces limites, nous étendons ParaPLUIE, une mesure *LLM*-juge basée sur la perplexité qui estime la confiance dans les réponses « *Yes*/*No* » sans générer de texte. Nous introduisons *-PLUIE, des variantes de ParaPLUIE spécifiques à la tâche, et évaluons leur alignement avec le jugement humain. Nos expériences montrent que *-PLUIE atteint des corrélations plus fortes que les autres mesures considérées avec les évaluations humaines tout en conservant un faible coût de calcul.

ABSTRACT

***-PLUIE : Personalisable metric with Llm Used for Improved Evaluation**

Evaluating the quality of automatically generated text often relies on *LLM*-juge methods. While effective, these approaches are computationally expensive and require post-processing. To address these limitations, we build upon ParaPLUIE, a perplexity-based *LLM*-juge metric that estimates confidence over "Yes/No" answers without generating text. We introduce *-PLUIE, task-specific prompting variants of ParaPLUIE and evaluate their alignment with human judgement. Our experiments show that personalised *-PLUIE achieves stronger correlations with human ratings while maintaining low computational cost.

MOTS-CLÉS : mesure, évaluation, sémantique, *LLM* en tant que juge.

KEYWORDS: metric, evaluation, semantic, LLM-as-a-judge.

1 Introduction

L'évaluation automatique reste un défi dans la génération de texte libre. Les mesures traditionnelles basées sur la similarité se concentrent sur les chevauchements lexicaux superficiels et ne parviennent souvent pas à capturer les variations préservant le sens ou les améliorations stylistiques. Les progrès récents dans le domaine des grands modèles de langue (*LLMs*) ont introduit les méthodes *LLM-as-a-*

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 1–32.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

judge (*LLM*-juge), qui utilisent les capacités de raisonnement des *LLMs* à interpréter les demandes des utilisateurs afin d'évaluer le texte généré (Brown *et al.*, 2020). Elles exploitent la compréhension sémantique et le raisonnement contextuel des *LLMs* pour fournir des évaluations plus riches et plus proches de celles des humains dans diverses tâches de traitement automatique des langues (Gu *et al.*, 2024; Doostmohammadi *et al.*, 2024).

Cependant, les méthodes classiques de *LLM*-juge génèrent des réponses sous forme de texte libre qui doit être analysé pour obtenir des jugements structurés. Cette analyse introduit du bruit et de l'ambiguïté, en particulier lorsque des décisions de catégorisation simples (par exemple « *Yes*/*No* ») sont requises. Toutefois, ce processus de génération de texte peut ne pas exploiter pleinement les connaissances internes du modèle. Pour pallier ces limites, Lemesle *et al.* (2025) a proposé ParaPLUIE comme alternative basée sur la perplexité par opposition à l'évaluation *LLM* basée sur la sortie. Cette méthode est formellement définie dans l'annexe A. Elle attribue un score de confiance à une question « *Yes*/*No* » en s'appuyant sur la perplexité du *LLM*. Initialement introduite pour la classification des paraphrases, ParaPLUIE obtient un alignement élevé avec les annotations binaires humaines tout en ayant un coût de calcul faible, équivalent environ à la génération d'un *token*. Par définition, ParaPLUIE représente la confiance du modèle dans sa réponse, offrant plusieurs propriétés d'interprétabilité intéressantes. Intuitivement, un score positif ou négatif élevé devrait indiquer une confiance élevée dans les réponses « *Yes* » ou « *No* » respectivement. Cependant, cette propriété n'a pas été évaluée et ParaPLUIE n'a pas été comparée aux *LLM*-juges basés sur la sortie.

Dans cet article, nous évaluons donc sa généralisation à d'autres tâches, à d'autres langues, son alignement avec les jugements humains dans le cadre de l'évaluation par préférence et nous la comparons aux *LLM*-juges basés sur la sortie. Plus précisément, nous cherchons à savoir si des amorces (*prompts*) spécifiques à la tâche (*-PLUIE), c'est-à-dire des questions adaptées au contexte d'évaluation, peuvent améliorer la fiabilité et la généralisation de ParaPLUIE. Nous menons cette étude sur trois tâches sémantiques : la **classification de paraphrases** en français, l'évaluation de la qualité de la traduction **Network Intent LanguagE (Nile)** et la **révision de textes scientifiques**. Pour chaque tâche, nous comparons *-PLUIE à des mesures largement utilisées basées sur la similarité, aux *LLM*-juges basés sur la sortie et à une approche aléatoire de contrôle.

Nos principales contributions sont les suivantes :

1. Nous introduisons *-PLUIE, une méthode *LLM*-juge générale et personnalisable basée sur la perplexité.
2. Nous concevons et évaluons trois variantes de *-PLUIE spécifiques à chaque tâche, couvrant trois tâches sémantiques, afin d'évaluer l'adaptabilité de l'approche.
3. Nous montrons que *-PLUIE permet un alignement plus fort avec le jugement humain tout en étant jusqu'à près de 8 fois plus rapide à calculer que les autres mesures *LLM*-juge.

2 Protocole expérimental

Dans cette section, nous décrivons les tâches et les données étudiées, les mesures à comparer et la manière dont la méthodologie ParaPLUIE est adaptée grâce au changement de l'amorce.

2.1 Tâches sémantiques

Classification de paraphrases : ParaPLUIE ayant démontré de bonnes performances dans la classification des paraphrases en anglais (Lemesle *et al.*, 2025), nous examinons sa capacité de généralisation au français. Pour cela, nous utilisons le jeu de données proposé par Tytgat *et al.* (2024) ; qui contient des phrases françaises transformées manuellement par substitution de synonymes et de paronymes. Sur cette base, nous créons un jeu de données de paraphrases en français qui contient 33,60% de paires positives. Nous décrivons la méthodologie utilisée en annexe B.

Traduction Nile : Nile (Jacobs *et al.*, 2018) fournit une grammaire structurée mais flexible pour exprimer le contrôle d'accès, la qualité de service et les déclarations temporelles, ce qui le rend particulièrement adapté pour combler le vide entre le langage naturel et les politiques réseau applicables. Munson *et al.* (2025) a introduit *Nile-English Aligned Translations* (NEAT), une méthodologie utilisée pour créer un corpus à grande échelle d'intentions alignées anglais-Nile. Ici, nous utilisons leur évaluation humaine de 436 triplets de traduction d'expressions Nile. Conformément au protocole NEAT, les scores Mushra des traductions Nile sont binarisés à l'aide d'un seuil d'acceptabilité : les traductions notées « *Good* » ou « *Excellent* » sont considérées comme positives, ce qui donne 60% d'exemples positifs.

Révision de textes scientifiques : La révision de textes est une tâche d'aide à l'écriture qui consiste à modifier en profondeur un texte existant afin de l'améliorer tout en préservant son sens original (Du *et al.*, 2022; Li *et al.*, 2022). Dans le domaine scientifique, la révision est une étape cruciale du processus de rédaction qui garantit la clarté, la cohérence et le respect des normes académiques, car une rédaction de mauvaise qualité peut contribuer au rejet d'un article (Amano *et al.*, 2023). Ici, nous nous concentrons sur la tâche de révision de textes scientifiques au niveau des paragraphes (Jourdan *et al.*, 2025a). Nous utilisons l'ensemble de test du jeu de données ParaReval [1] (Jourdan *et al.*, 2025b), un jeu de données de préférences humaines sur des paires de révisions générées. De plus amples détails sur ce jeu de données sont fournis en annexe C.

2.2 Mesures de référence

Pour toutes les tâches, nous incluons un ensemble de mesures de similarité largement utilisées pour la génération de texte : **Levenshtein** (normalisé), **BLEU** (Papineni *et al.*, 2002), **METEOR** (Banerjee & Lavie, 2005) et **BERTScore** (Zhang *et al.*, 2020). Nous utilisons en outre MODERN BERT (Warner *et al.*, 2025), un *Transformer* modernisé, uniquement basé sur un encodeur, entraîné avec des données et des techniques mises à jour. Cela donne une variante mise à jour, **Modern BertScore**, que nous utilisons comme référence état de l'art.

De plus, nous considérons trois variantes *LLM*-juge basées sur la sortie : (a) **LLM-Yes/No**, qui répond à des questions binaires « *Yes*/*No* », (b) ***LLM-choice***, qui émet une préférence sur une paire de propositions, et (c) ***LLM*-Likert**, qui attribue des notes sur une échelle à cinq points. Toutes les amorces sont fournies en annexe D.

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

2.3 Mesures *-PLUIE

Des travaux antérieurs (Rios & Kavuluru, 2018; Brown *et al.*, 2020; Chung *et al.*, 2024; Lemesle *et al.*, 2025) montrent que fournir des exemples explicites de la tâche peut améliorer la capacité du modèle à produire un jugement précis et cohérent. Bien que ParaPLUIE ait démontré de solides performances pour évaluer la révision de texte à l'aide de son amorce de classification de paraphrase originale (Jourdan *et al.*, 2025b), nous cherchons à déterminer si l'adaptation de l'amorce à la tâche peut améliorer la fiabilité et si elle peut être généralisée à d'autres tâches.

Pour explorer cette hypothèse, nous proposons des amorces personnalisées pour chaque tâche étudiée, en considérant ParaPLUIE comme une mesure flexible et prête à l'emploi qui peut facilement être adaptée grâce à la modification de l'amorce et du modèle de perplexité sous-jacent. Nous désignons le modèle original de classification des paraphrases par Para-PLUIE et construisons les variantes suivantes :

- **Fr-PLUIE** : une adaptation française de l'amorce originale Para-PLUIE, utilisant des exemples traduits pour tester la généralisation au français.
- **Net-PLUIE** : une amorce permettant d'évaluer si deux phrases expriment la même intention réseau. Les exemples utilisés dans l'amorce sont tirés et supprimés des données d'évaluation afin d'éviter tout biais.
- **Rev-PLUIE** : une amorce conçue pour évaluer si une révision générée suit l'instruction de l'auteur. Cette variante utilise la révision humaine de référence comme exemple unique pour ancrer la compréhension de la tâche par le modèle.

Toutes les amorces sont disponibles en annexe E. Nous comparons *-PLUIE au modèle original ParaPLUIE afin de déterminer si les amorces spécifiques à une tâche améliorent l'alignement avec les jugements humains. Pour toutes les méthodes basées sur des *LLMs*, nous utilisons Phi-4 14B (Abdin *et al.*, 2024) et Llama 3 70B (AI@Meta, 2024) comme modèles de perplexité.

3 Résultats

Afin d'étudier l'interprétabilité naturelle de *-PLUIE, nous comparons cette approche avec des mesures de similarité standard et des méthodes basées sur les résultats *LLM*-juge. Comme *-PLUIE donne un score continu avec un seuil interprétable, nous l'utilisons pour la **classification** et pour l'**évaluation par préférence**. Pour la classification, l'échantillon est considéré comme positif si le score renvoyé est supérieur à 0, et négatif dans le cas contraire. Pour les préférences, nous classons les options en fonction de leurs scores afin d'identifier celle qui est préférée et de l'aligner sur les préférences humaines.

3.1 Classification

Nous utilisons *-PLUIE pour classer les paires paraphrase/non-paraphrase et les traductions Nile bonnes/mauvaises. Comme les scores des mesures de référence ne sont pas directement interprétables comme des décisions catégorielles, nous calibrons chaque mesure en déterminant un seuil de décision optimal qui maximise le score F1. La table 1 présente les performances de toutes les mesures considérées. Nous présentons également les résultats pour le seuil naturel avec *-PLUIE.

Nous constatons que */Para-PLUIE obtient des performances compétitives ou légèrement supérieures

Tâche	*Classification de paraphrases*					*Traduction Nile*				
Mesure	**Seuil**	**Exc.**	**R.**	**Prec.**	**F1**	**Seuil**	**Exc.**	**R.**	**P.**	**F1**
*-PLUIE Phi	-7.63	0.71	0.77	0.54	0.63	-3.14	0.81	0.95	0.78	0.85
*-PLUIE Phi	0	0.74	0.54	0.61	0.58	0	0.81	0.85	0.84	0.85
*-PLUIE Llama	-4.07	0.67	0.71	0.50	0.59	3.97	0.80	0.92	0.79	0.85
*-PLUIE Llama	0	0.70	0.64	0.53	0.58	0	0.79	0.94	0.77	0.84
Para-PLUIE Phi	-7.47	0.73	0.73	0.57	**0.64**	-3.43	0.81	0.87	0.83	0.85
Para-PLUIE Phi	0	**0.75**	0.49	**0.66**	0.56	0	0.73	0.63	0.88	0.73
Para-PLUIE Llama	-11.91	0.70	0.66	0.54	0.59	-13.73	0.81	0.92	0.80	0.85
Para-PLUIE Llama	0	0.72	0.48	0.58	0.53	0	**0.82**	0.79	**0.89**	0.84
LLM-Yes/No Phi		0.74	0.50	0.64	0.56		0.81	0.88	0.86	0.87
LLM-Yes/No Llama		0.71	0.48	0.56	0.52		0.81	0.90	0.86	**0.88**
Modern BertScore	0.84	0.33	**1.00**	0.33	0.49	0.68	0.63	0.96	0.63	0.76
BERTScore	0.80	0.33	**1.00**	0.33	0.49	0.39	0.67	**1.0**	0.64	0.78
METEOR	0.43	0.33	**1.00**	0.33	0.49	0.0	0.6	**1.0**	0.6	0.75
BLEU	0.00	0.33	**1.00**	0.33	0.49	0.0	0.6	**1.0**	0.6	0.75
Levenshtein	0.71	0.33	**1.00**	0.33	0.49	0.17	0.65	0.95	0.64	0.77
Aléatoire pondéré		*0.56*	*0.32*	*0.32*	*0.32*		0.52	0.60	0.61	0.61
Aléatoire uniforme		*0.50*	*0.50*	*0.32*	*0.40*		0.49	0.48	0.60	0.53

TABLE 1 – Seuils naturel et optimal *a posteriori* pour chaque mesure et variante de modèle. Sont fournis l'exactitude (*Exc.*), le rappel (*R.*), la précision *(P.)* et le score F1 correspondants pour la classification. *-PLUIE correspond respectivement à Fr-PLUIE et Net-PLUIE.

à celles des mesures *LLM*-juge. La différence entre les seuils calibrés et les seuils par défaut est minime, ce qui souligne l'interprétabilité et la robustesse de l'échelle de notation de */Para-PLUIE. Les mesures traditionnelles atteignent une précision d'environ 33% sur la classification des paraphrases en français, ce qui implique que toutes les paires sont classées comme des paraphrases. Ce résultat suggère que le jeu de données est particulièrement difficile et que ces mesures basées sur la surface ont du mal à saisir les grandes différences sémantiques lorsque le chevauchement lexical est élevé.

Ici, nous prenons seulement en compte les traductions Nile générées avec Llama à partir de la méthodologie NEAT ; les autres résultats sont disponibles dans l'annexe F. Nous observons que sur cette tâche, */Para-PLUIE surpasse systématiquement les mesures de similarité et est comparable aux *LLM*-juges basés sur les sorties. Pour la traduction Nile, une précision plus élevée est préférable afin d'éviter une configuration réseau incorrecte. Les meilleures performances sont observées avec Para-PLUIE Llama et le seuil non calibré (0). Une analyse plus approfondie de l'impact du choix du seuil est présentée dans l'annexe G.

3.2 Préférence

Pour évaluer l'alignement entre les mesures automatiques et le jugement humain, nous utilisons l'exactitude par paire avec calibrage des ex æquo (Deutsch *et al.*, 2023), le V de Cramér (Cramér, 1946) et le κ de Cohen (Cohen, 1960). Les résultats sont présentés dans la table 2.

Pour la traduction Nile, toutes les mesures sont positivement corrélées avec le jugement humain, selon

Tâche / **Mesure**	*Traduction Nile* **Exc Pair.**	V	κ	*Révision de textes scientifiques* **Exc Pair.**		V		κ	
*-PLUIE Phi	0.69	0.42	0.40	**0.61**	0.61 réf	**0.31**	**0.32** réf	**0.32**	0.33 réf
*-PLUIE Llama	0.70	0.43	0.42	**0.61**	**0.62** réf	**0.31**	**0.32** réf	**0.32**	**0.34** réf
Para-PLUIE Phi	0.70	0.44	0.42	0.52		0.21		0.15	
Para-PLUIE Llama	**0.72**	**0.46**	0.43	0.52		0.20		0.17	
LLM-choice Phi	0.47	0.42	0.20	0.53	0.55 réf	0.25	0.27 réf	0.24	0.27 réf
LLM-choice Llama	0.47	0.43	0.20	0.59	0.60 réf	0.28	0.30 réf	0.30	0.31 réf
LLM-Likert Phi	0.65	0.37	0.41	0.45	0.52 réf	0.30	0.29 réf	0.21	0.26 réf
LLM-Likert Llama	0.69	0.40	**0.44**	0.44	0.50 réf	0.29	0.27 réf	0.19	0.23 réf
Modern BertScore	0.63	0.41	0.36	0.36		0.13		-0.07	
BERTScore	0.68	0.43	0.40	0.45		0.16		0.03	
METEOR	0.68	0.40	0.37	0.42		0.19		0.00	
BLEU	0.43	N/A	0.00	0.41		0.17		-0.03	
Levenshtein	0.64	0.34	0.29	0.44		0.13		0.01	
Aléatoire pondéré	*0.43*	*0.06*	*0.00*	*0.36*		*0.03*		*0.00*	
Aléatoire uniforme	*0.33*	*0.06*	*0.00*	*0.33*		*0.03*		*0.00*	

TABLE 2 – Alignement des mesures automatiques avec les préférences humaines. L'exactitude par paire et V sont définis sur [0 :1] et κ sur [-1 :1]. « réf » indique que la révision de référence est fournie.

κ. Para-PLUIE obtient la corrélation la plus élevée, suivi de Net-PLUIE. Les approches *LLM*-juge dépendent davantage du choix du modèle, Llama s'imposant comme la meilleure option pour cette tâche. BERTScore affiche le meilleur alignement parmi les mesures de similarité traditionnelles, avec des scores comparables à l'approche *LLM*-Likert.

Pour la révision, sur les trois mesures, Rev-PLUIE affiche l'alignement le plus élevé avec les évaluations humaines, obtenant les meilleurs ou deuxièmes meilleurs scores. Cela surpasse les autres méthodes *LLM*-juge et les mesures de similarité traditionnelles. *LLM-choice* affiche également des performances compétitives, tandis que *LLM*-Likert et Para-PLUIE obtiennent des résultats modérés mais cohérents. En revanche, les mesures basées sur les n-grammes et les plongements obtiennent des corrélations nettement plus faibles avec le jugement humain. Enfin, pour Rev-PLUIE, l'ajout de la révision de référence dans l'amorce en tant qu'exemple unique améliore légèrement l'alignement du jugement, ce qui en fait l'option la plus fiable pour évaluer la tâche avec ou sans référence. Les résultats obtenus avec différents *LLM* sont disponibles dans l'annexe H et montrent que Rev-PLUIE surpasse *LLM-choice* dans toutes les configurations expérimentales.

3.3 Coût calculatoire

Nous comparons le coût calculatoire de toutes les mesures basées sur des *LLMs* dans le Tableau 3. Les approches *-PLUIE sont systématiquement plus rapides que les alternatives à base de génération en utilisant le même modèle, car elles calculent des probabilités sur un ensemble restreint de sous-mots plutôt que de générer de longues sorties textuelles. Dans l'ensemble, les méthodes PLUIE offrent un compromis favorable entre efficacité et alignement, ce qui en fait une option intéressante pour une évaluation par *LLM*-juge à grande échelle.

Modèle	*Tâche* / Approche	*Classification de paraphrases* GPUs	Temps d'exéc.	*Traduction Nile* GPUs	Temps d'exéc.	*Révision de textes scientifiques* GPUs	Temps d'exéc.	
Phi	LLM-Yes/No / *LLM-choice*	MI300 x1	23 min	MI300 x1	11 min	A100 x1	40 min	41 min réf
	LLM-Likert			MI300 x1	10 min	A100 x1	51 min	65 min réf
	Para-PLUIE	MI300 x1	**3.5 min**	MI300 x1	1.8 min	A100 x1	**08 min**	
	*-PLUIE	MI300 x1	3.6 min	MI300 x1	**1.4 min**	A100 x1	09 min	13 min réf
Llama	LLM-Yes/No / *LLM-choice*	MI300 x2	48 min	MI300 x2	22 min	A100 x2	124 min	127 min réf
	LLM-Likert			MI300 x2	21 min	A100 x2	141 min	187 min réf
	Para-PLUIE	MI300 x2	**14 min**	MI300 x2	6.5 min	A100 x3	**29 min**	
	*-PLUIE	MI300 x2	17 min	MI300 x2	**3.4 min**	A100 x3	33 min	55 min réf

TABLE 3 – Temps de calculs sur *GPU* des approches *LLM*-juge, sur les trois tâches sémantiques évaluées, pour les modèles Phi et Llama.

Comme souligné par Nayab *et al.* (2025), le temps d'inférence d'un *LLM* dépend du nombre de sous-mots générés en sortie. En raison de la nature autorégressive des décodeurs *Transformer* (Vaswani *et al.*, 2017), chaque sous-mot en sortie requiert une itération dédiée du décodeur, ce qui signifie que le temps de génération croît linéairement avec la longueur de la réponse et est directement lié à la longueur de l'amorce d'entrée.

Les méthodes *LLM*-juge à base de génération sont directement soumises à cette contrainte : générer un jugement libre de $\mathcal{N}(\hat{y})$ sous-mots nécessite $\mathcal{N}(\hat{y})$ itérations successives du décodeur. De plus, la longueur de l'amorce $\mathcal{N}(x)$ constitue elle-même une source de surcoût souvent négligée. En pratique, les amorces des *LLMs*-juge tendent à s'allonger (Wang *et al.*, 2025), car elles doivent inclure des instructions explicites décrivant le format de sortie attendu, ainsi que des exemples qui conditionnent le modèle à produire une réponse structurée (par exemple une réponse *Oui*/*Non* ou un score de Likert) plutôt qu'un texte libre. De surcroît, les méthodes de raisonnement (Wei *et al.*, 2022) allongent encore les générations des modèles. Sans ce conditionnement, les méthodes à base de génération tendent à produire des réponses difficiles à analyser ou ne respectant pas le schéma attendu.

Les méthodes PLUIE ne subissent pas cette contrainte : puisque le score est dérivé directement des scores bruts du modèle (comme souligné en Annexe A), un score de confiance est obtenu indépendamment de la façon dont le modèle aurait formulé sa réponse. Il n'est donc pas strictement nécessaire d'expliquer le format de réponse ou d'imposer une structure de sortie via de longues instructions. Cela dit, fournir quelques exemples reste bénéfique : même si le respect du format n'est pas nécessaire, les exemples aident à réaligner la distribution du prochain sous-mot du modèle vers les sous-mots d'intérêt (*Yes* et *No*), réduisant la probabilité accordée aux termes non pertinents du vocabulaire et améliorant la discriminabilité du score résultant.

Les méthodes PLUIE nécessitent **exactement une itération du décodeur**, indépendamment du nombre de sous-mots candidats évalués. Cela découle du fait qu'une seule itération du décodeur produit une distribution de probabilité sur l'ensemble du vocabulaire ; les probabilités de *Yes* et *No* sont donc toutes deux directement lisibles dans le vecteur de sortie du modèle, sans coût supplémentaire. En revanche, dans le meilleur des cas, l'utilisation d'une méthode *LLM*-juge basée sur la sortie avec la même amorce coûte au moins deux fois plus cher qu'une méthode PLUIE. En effet, même dans le cas le plus favorable, un LLM-juge basé sur la sortie doit produire au moins deux sous-mots : un pour la réponse et un sous-mot de fin de séquence, nécessitant ainsi **au moins deux itérations du décodeur**. Cela explique l'accélération allant jusqu'à $7.9\times$ rapportée dans le Tableau 3.

4 Reproductibilité

Pour utiliser *-PLUIE, vous pouvez consulter la documentation disponible sur un espace HuggingFace (https://huggingface.co/spaces/qlemesle/parapluie) et utiliser le code source publiquement disponible sur GitLab (https://gitlab.inria.fr/expression/paraphrase-generation-evaluation-powered-by-an-llm-a-semantic-metric-not-a-lexical-one-coling-2025).

Le *LLM* et l'amorce peuvent être remplacés indépendamment sans modifier la logique de calcul du score. Le code suivant illustre comment instancier et exécuter une variante *-PLUIE :

```
from PPLUIE.wrapper import ParaPLUIE

scorer = ParaPLUIE()
scorer.init("microsoft/phi-4", device="auto")
# Utiliser Fr-PLUIE en changeant simplement l'amorce
scorer.setTemplate("FS-DIRECT_FR")

S = ["Les enfants ont boulonné tous les gâteaux."]
H = ["Les enfants ont mangé tous les gâteaux."]

score = scorer.compute(S, H)
print("Result score : ", score) # -4.85
# score > 0 : paraphrase, score < 0 : non-paraphrase
```

La liste complète des modèles et amorces supportés est accessible via `scorer.show_available_models()` et `scorer.show_templates()`. De nouvelles amorces spécifiques à une tâche peuvent être facilement ajoutées en éditant le fichier `template.py`.

5 Conclusion

Nous avons présenté une version généralisée de ParaPLUIE, élargissant l'approche originale à un éventail plus large de tâches sémantiques à évaluer. En adaptant spécifiquement les amorces à chaque tâche, nous avons montré que *-PLUIE obtient systématiquement des corrélations plus fortes avec le jugement humain. Dans toutes les expériences, les amorces personnalisées *-PLUIE sont jusqu'à 7,9 fois plus rapides que les approches basées sur les sorties. Son score interprétable et ses seuils de décision stables la rendent pratique, évitent le post-traitement des sorties *LLM* et permettent une substitution de modèle simple, évolutive et transparente pour l'évaluation automatique par *LLM*.

De plus, le modèle Net-PLUIE proposé peut être utilisé comme fonction d'alignement état de l'art entre le langage naturel et les expressions d'intention Nile, comme décrit dans l'annexe I. Dans l'ensemble, ces résultats positionnent *-PLUIE comme une base efficace et adaptable pour l'évaluation automatique à l'ère des *LLM*.

Limites

Les modèles de perplexité utilisés avec *-PLUIE n'ont pas été affinés pour ces expériences. Un affinage pourrait potentiellement améliorer à la fois l'exactitude et la sensibilité spécifique à la tâche. Les plus grands *LLM* utilisés dans cet article ont 70 milliards de paramètres, et l'utilisation de modèles encore plus grands pourrait améliorer davantage les résultats.

La plupart des expériences ont été menées en anglais, avec des tests supplémentaires en français montrant des résultats cohérents. Le comportement de la méthode dans des langues à la morphologie plus riche ou aux structures syntaxiques nettement différentes reste à étudier.

Dans cet article, nous n'avons pris en compte que les amorces pouvant être formulées sous forme de questions « *Yes*/*No* ». Il serait intéressant d'étendre ce travail à plusieurs tâches nécessitant des sorties différentes. Ces tâches pourraient inclure l'analyse des sentiments, la classification des sujets ou la réponse à des questions à choix multiples. En outre, la formule de *-PLUIE pourrait être modifiée afin de supprimer la limitation à un seul *token*, comme indiqué dans l'annexe A.

Considérations éthiques

Disponibilité des données : Pour toutes les tâches considérées, nous utilisons des jeux de données librement accessibles. Les jeux de données provenant de Tytgat *et al.* (2024) et Munson *et al.* (2025) sont disponibles sur demande auprès des auteurs originaux. Les paragraphes de ParaReval sont extraits d'articles scientifiques collectés sur OpenReview, où ils sont soumis à différentes « licences non exclusives, perpétuelles et libres de droits » (*"non-exclusive, perpetual, and royalty-free licence"*).

Ressources de calcul : L'utilisation des *LLM* reste gourmande en ressources, mais l'utilisation d'un *LLM* de taille moyenne comme Phi-4 14B semble être compétitive par rapport à des *LLM* plus grands comme Llama 3 70B (Tables 1 et 2), et elle est évidemment moins gourmande en ressources de calcul. Enfin, le coût de calcul de *-PLUIE est bien inférieur à celui des autres méthodes *LLM*-juge, comme le montre le tableau 3.

Remerciements

Cette étude a bénéficié de l'accès aux ressources MI300A du CINES dans le cadre de l'allocation 2025–AD011015262R1 accordée par le GENCI, et est soutenue par le Ministère des Armées - Agence de l'Innovation de la Défense.

Cette étude a bénéficié de l'accès aux ressources HPC de l'IDRIS dans le cadre des allocations 2023-AD011013901R1, 2024-AD011013901R2 et 2024-AD011014882R1 accordées par le GENCI.

Cette étude est partiellement financée par l'ANR dans le cadre du projet PIA EUR CyberSchool (ANR-18-EURE-0004).

Références

ABDIN M. *et al.* (2024). Phi-4 technical report.

ABOUELENIN A. *et al.* (2025). Phi-4-mini technical report : Compact yet powerful multimodal language models via mixture-of-loras.

AI@META (2024). Llama 3 model card. (accessed December 2025).

ALLAL L. B., LOZHKOV A., BAKOUCH E., BLÁZQUEZ G. M., PENEDO G., TUNSTALL L., MARAFIOTI A., KYDLÍČEK H., LAJARÍN A. P., SRIVASTAV V., LOCHNER J., FAHLGREN C., NGUYEN X.-S., FOURRIER C., BURTENSHAW B., LARCHER H., ZHAO H., ZAKKA C., MORLON M., RAFFEL C., VON WERRA L. & WOLF T. (2025). Smollm2 : When smol goes big – data-centric training of a small language model.

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, 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.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A., AGARWAL S., HERBERT-VOSS A., KRUEGER G., HENIGHAN T., CHILD R., RAMESH A., ZIEGLER D., WU J., WINTER C., HESSE C., CHEN M., SIGLER E., LITWIN M., GRAY S., CHESS B., CLARK J., BERNER C., MCCANDLISH S., RADFORD A., SUTSKEVER I. & AMODEI D. (2020). Language models are few-shot learners. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems*, volume 33, p. 1877–1901 : Curran Associates, Inc.

CAI Z. *et al.* (2024). Internlm2 technical report.

CHUNG H. W., HOU L., LONGPRE S., ZOPH B., TAY Y., FEDUS W., LI Y., WANG X., DEHGHANI M., BRAHMA S., WEBSON A., GU S. S., DAI Z., SUZGUN M., CHEN X., CHOWDHERY A., CASTRO-ROS A., PELLAT M., ROBINSON K., VALTER D., NARANG S., MISHRA G., YU A., ZHAO V., HUANG Y., DAI A., YU H., PETROV S., CHI E. H., DEAN J., DEVLIN J., ROBERTS A., ZHOU D., LE Q. V. & WEI J. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, **25**(70), 1–53.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46. DOI : 10.1177/001316446002000104.

CRAMÉR H. (1946). *Mathematical Methods of Statistics*. Goldstine Printed Materials. Princeton University Press.

DANG J. *et al.* (2024). Aya expanse : Combining research breakthroughs for a new multilingual frontier.

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.

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.

GRATTAFIORI A. *et al.* (2024). The llama 3 herd of models.

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*.

HURST A. *et al.* (2024). Gpt-4o system card.

JACOBS A. S., PFITSCHER R. J., FERREIRA R. A. & GRANVILLE L. Z. (2018). Refining network intents for self-driving networks. In *Proceedings of the Afternoon Workshop on Self-Driving Networks*, SelfDN 2018, p. 15–21, New York, NY, USA : ACM. DOI : 10.1145/3229584.3229590.

JIANG A. Q. *et al.* (2023). Mistral 7b.

JOURDAN L., BOUDIN F., DUFOUR R., HERNANDEZ N. & AIZAWA A. (2025a). 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., HERNANDEZ N., BOUDIN F. & DUFOUR R. (2025b). Identifying reliable evaluation metrics for scientific text revision. In W. CHE, J. NABENDE, E. SHUTOVA & M. T. PILEHVAR, Éds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 6731–6756, Vienna, Austria : Association for Computational Linguistics. DOI : 10.18653/v1/2025.acl-long.335.

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.

MUNSON D., ALAIN P. & DOYEN G. (2025). Neat : A nile-english aligned translation corpus based on a robust methodology for intent based networking. *Computer Networks*, **271**, 111519. DOI : https ://doi.org/10.1016/j.comnet.2025.111519.

NAYAB S., ROSSOLINI G., SIMONI M., SARACINO A., BUTTAZZO G., MANES N. & GIACOMELLI F. (2025). Concise thoughts : Impact of output length on llm reasoning and cost.

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.

RIOS A. & KAVULURU R. (2018). Few-shot and zero-shot multi-label learning for structured label spaces. In E. RILOFF, D. CHIANG, J. HOCKENMAIER & J. TSUJII, Éds., *Proceedings of the*

2018 Conference on Empirical Methods in Natural Language Processing, p. 3132–3142, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1352.

TEAM G. *et al.* (2024). Gemma 2 : Improving open language models at a practical size.

TEAM Q. (2025). Qwq-32b : Embracing the power of reinforcement learning.

TYTGAT J., WISNIEWSKI G. & BETRANCOURT A. (2024). Évaluation de la similarité textuelle : Entre sémantique et surface dans les représentations neuronales. In M. BALAGUER, N. BENDAHMAN, L.-M. HO-DAC, J. MAUCLAIR, J. G MORENO & J. PINQUIER, Éds., *Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position*, p. 85–96, Toulouse, France : ATALA and AFPC.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems*, volume 30 : Curran Associates, Inc.

VERGA P., HOFSTATTER S., ALTHAMMER S., SU Y., PIKTUS A., ARKHANGORODSKY A., XU M., WHITE N. & LEWIS P. (2024). Replacing judges with juries : Evaluating llm generations with a panel of diverse models.

WANG C., FENG Y., CHEN D., CHU Z., KRISHNA R. & ZHOU T. (2025). Wait, we don't need to "wait" ! removing thinking tokens improves reasoning efficiency. In C. CHRISTODOULOPOULOS, T. CHAKRABORTY, C. ROSE & V. PENG, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2025*, p. 7459–7482, Suzhou, China : Association for Computational Linguistics. DOI : 10.18653/v1/2025.findings-emnlp.394.

WARNER B., CHAFFIN A., CLAVIÉ B., WELLER O., HALLSTRÖM O., TAGHADOUINI S., GALLAGHER A., BISWAS R., LADHAK F., AARSEN T., ADAMS G. T., HOWARD J. & POLI I. (2025). Smarter, better, faster, longer : A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In W. CHE, J. NABENDE, E. SHUTOVA & M. T. PILEHVAR, Éds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2526–2547, Vienna, Austria : Association for Computational Linguistics. DOI : 10.18653/v1/2025.acl-long.127.

WEI J., WANG X., SCHUURMANS D., BOSMA M., XIA F., CHI E., LE Q. V., ZHOU D. *et al.* (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, **35**, 24824–24837.

YANG A. *et al.* (2025). Qwen2.5 technical report.

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

A Définition de ParaPLUIE

ParaPLUIE est initialement défini comme le rapport de vraisemblance logarithmique permettant de comparer la prédominance d'une réponse « *Yes* » par rapport à « *No* » à la question posée par l'amorce (Lemesle *et al.*, 2025), comme illustré dans la figure 1.

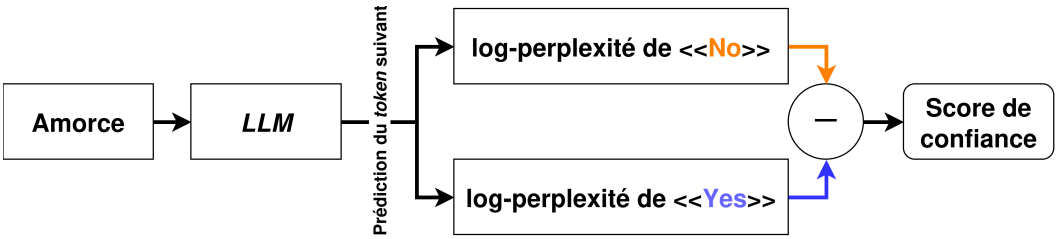


FIGURE 1 – Fonctionnement de *-PLUIE

Plus formellement, soit :

— S la phrase source,

— H la phrase hypothétique,

— Prompt(S, H) l'amorce remplie avec les phrases S et H,

— T : longueur en *token* de Prompt(S, H),

— $\circ$: l'opérateur de concaténation de 2 séquences de *tokens*.

$$\texttt{ParaPLUIE}(S, H) \quad = \quad \log\left(\frac{p\left(\text{Yes}|\text{Prompt}(S,H)\right)}{p\left(\text{No}|\text{Prompt}(S,H)\right)}\right) \tag{1}$$

Il peut être étendu, à l'aide de la formule de Bayes, pour faire apparaître la perplexité du modèle :

$$\begin{aligned}
&= \log\left(\frac{p\left(\text{Prompt}(S,H)\circ\text{Yes}\right)}{p\left(\text{Prompt}(S,H)\right)} \times \frac{p\left(\text{Prompt}(S,H)\right)}{p\left(\text{Prompt}(S,H)\circ\text{No}\right)}\right)\\
&= -\frac{-1}{T+1}\log\left(p\left(\text{Prompt}(S,H)\circ\text{Yes}\right)\right)\times(T+1)\\
&\quad +\frac{-1}{T+1}\log\left(p\left(\text{Prompt}(S,H)\circ\text{No}\right)\right)\times(T+1)\\
&= \log\left(\exp\left(\frac{-1}{T+1}\log\left(p\left(\text{Prompt}(S,H)\circ\text{No}\right)\right)\right)^{T+1}\right)\\
&\quad -\log\left(\exp\left(\frac{-1}{T+1}\log\left(p\left(\text{Prompt}(S,H)\circ\text{Yes}\right)\right)\right)^{T+1}\right)\\
&= \log\left(ppl\left(\text{Prompt}(S,H)\circ\text{No}\right)^{T+1}\right)\\
&\quad -\log\left(ppl\left(\text{Prompt}(S,H)\circ\text{Yes}\right)^{T+1}\right)\\
&= (T+1)\times\log\left(ppl\left(\text{Prompt}(S,H)\circ\text{No}\right)\right)\\
&\quad -(T+1)\times\log\left(ppl\left(\text{Prompt}(S,H)\circ\text{Yes}\right)\right)
\end{aligned} \tag{2}$$

Enfin, en supposant que le *LLM* utilisé utilise la perplexité logarithmique comme $loss$, nous avons :

$$\texttt{ParaPLUIE}(S,H) = (T+1)\times\left[loss_{LLM}(\text{Prompt}(S,H)\circ\text{No}) - loss_{LLM}(\text{Prompt}(S,H)\circ\text{Yes})\right] \tag{3}$$

Notez que ParaPLUIE n'est pas nécessairement symétrique. ParaPLUIE offre des capacités *plug-and-play* : le modèle de perplexité sous-jacent et le patron d'amorce peuvent tous deux être modifiés sans altérer le principe fondamental de la méthode.

Cependant, certains aspects doivent être pris en considération.

Tout d'abord, le modèle de perplexité doit être capable d'effectuer des tâches de question-réponse, car la métrique repose sur l'évaluation de la confiance du modèle entre deux réponses mutuellement exclusives. Notez qu'une exigence de l'équation 2 est que les deux réponses opposées correspondent chacune à un seul *token* selon le *tokenizer* du modèle. Au-delà de cela, le choix des *tokens* n'est pas limité (il pourrait s'agir de « *Dog* » vs « *Cat* »), même si nous ne considérons que « *Yes* » vs « *No* » dans cette étude.

Ensuite, pour garantir un calcul correct de la perplexité, les *tokens* comparés doivent apparaître comme les *tokens* finaux de l'amorce. Les *tokeniseurs* de la plupart des *chatbots* modernes basés sur des *LLM* utilisent un format de dialogue utilisateur-assistant, dans lequel des *tokens* spéciaux sont insérés pour marquer la fin de chaque rôle. Ces *tokens* de fin de rôle doivent être supprimés avant le calcul de la perplexité, car ils altéreraient sinon les résultats. Par exemple, lors du calcul de la perplexité de « *Yes* » dans le dialogue suivant :

`<user>People like “Cats” more than “Dogs”.</user> <assistant> Yes </assistant>` le *token* de fin de tour `</assistant>` doit être supprimé avant de calculer la perplexité.

Nous nous appuyons sur l'implémentation publique fournie par les auteurs originaux [2].

Cette formule peut être généralisée à des réponses de différentes longueurs. Ainsi, si une réponse est $(Y_1 \dots Y_i)$ et la seconde $(N_1 \dots N_j)$, nous pouvons écrire :

$$\texttt{ParaPLUIE}(S,H) = (T+j) \times (loss_{LLM}(\text{Prompt}(S,H) \circ N_1 \dots N_j) - (T+i) \times loss_{LLM}(\text{Prompt}(S,H) \circ Y_1 \dots Y_i)) \quad (4)$$

Un inconvénient ici est que la réponse la plus courte serait favorisée : en effet, il est souvent plus probable de générer une séquence courte plutôt qu'une longue. Considérer la perplexité moyenne de la réponse pourrait être une alternative, mais nous observerions probablement une tendance à centrer chaque réponse autour d'une moyenne similaire, ce qui serait regrettable.

B Jeu français de paraphrases

Ce jeu de données s'appuie sur les travaux de Tytgat *et al.* (2024) qui ont montré que les mesures de similarité sémantique couramment utilisées sont souvent plus sensibles aux différences superficielles qu'aux variations sémantiques. Leur étude a créé un jeu de données en français annoté par des experts, composé de 355 phrases sources, chacune construite indépendamment de tout domaine spécifique.

Pour chaque phrase, les auteurs ont identifié le mot sémantique central, celui qui traduit le mieux le sens de la phrase, et ont systématiquement produit trois types de versions modifiées :

2. https://gitlab.inria.fr/expression/paraphrase-generation-evaluation-powered-by-an-llm-a-semantic-metric-not-a-lexical-one-coling-2025

Transformation	Phrase
Aucune	Les enfants ont boulotté tous les gâteaux.
Paronyme	Les enfants ont boulonné tous les gâteaux
Synonyme	Les enfants ont mangé tous les gâteaux.
Synonyme de Paronyme	Les enfants ont fixé tous les gâteaux.

TABLE 4 – Phrase caractéristique du jeu de données Tytgat *et al.* (2024) et ses différentes transformations.

— **Substitution par un paronyme** : le mot est remplacé par un paronyme, un mot qui ressemble ou qui sonne de manière similaire, mais qui a un sens complètement différent. La phrase obtenue devient sémantiquement différente de l'originale tout en restant proche sur le plan lexical.
— **Substitution par un synonyme** : le mot est remplacé par un synonyme, ce qui donne une phrase qui conserve le sens original tout en étant proche sur le plan lexical.
— **Synonyme d'un paronyme** : le mot est remplacé par un synonyme du paronyme, ce qui donne une variante supplémentaire nuancée. Ce processus donne lieu à trois versions alternatives pour chaque phrase source.

Un exemple de ces transformations est fourni dans le tableau 4.

Soit **S** la transformation d'une phrase **s** par synonyme et **P** par un paronyme. Nous créons un nouvel ensemble composé de couples de phrases formés par la phrase source s et sa transformation. En fonction de la combinaison des transformations, nous les classons comme paraphrases ou non-paraphrases :

(s, S(s)) $\implies$ paraphrase

(P(s), S(P(s))) $\implies$ paraphrase

(s, P(s)) $\implies$ $\overline{\text{paraphrase}}$

(s, S(P(s))) $\implies$ $\overline{\text{paraphrase}}$

(P(s), S(s)) $\implies$ $\overline{\text{paraphrase}}$

(S(s), S(P(s))) $\implies$ $\overline{\text{paraphrase}}$

On obtient 1 914 paires, dont 33,60% sont positives.

C Exemple de données issues du jeu de données ParaReval

Ce jeu de données se concentre sur la tâche de révision de textes scientifiques au niveau du paragraphe, où chaque paragraphe et l'instruction spécifiant la modification souhaitée sont fournis en entrée, et où le modèle doit produire un paragraphe révisé qui correspond à l'instruction donnée (Jourdan *et al.*, 2025a).

Nous utilisons l'ensemble de test du jeu de données ParaReval [3] (Jourdan *et al.*, 2025b), une collection

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

de préférences humaines sur des paires de révisions générées automatiquement. Chaque instance du jeu de données se compose de deux versions du même paragraphe, extraites d'articles en informatique provenant d'OpenReview et rédigées par les auteurs originaux. Les paires de paragraphes sont annotées avec une instruction décrivant l'intention de révision sous-jacente. Un exemple est donné dans l'annexe C.

Sur la base de ces annotations, chaque paragraphe a été automatiquement révisé par six modèles différents suivant l'instruction correspondante. Les révisions générées ont été évaluées par paires par des annotateurs humains afin d'évaluer la fiabilité des mesures d'évaluation automatiques pour cette tâche. Les annotateurs ont répondu à trois questions : (1) Le modèle a-t-il suivi l'instruction ? (2) La révision est-elle correcte, c'est-à-dire meilleure que l'original ? et (3) Quelle révision préfèrent-ils entre les options A et B ? Les réponses des annotateurs permettent de classer les deux options et d'identifier celle qui est préférée.

Les données comprennent 258 paires de paragraphes révisés, chacune annotée avec deux instructions de révision distinctes, ce qui donne un total de 516 instances d'évaluation. Les préférences humaines par paires sont équilibrées, les annotateurs favorisant l'option A dans 44% des cas, l'option B dans 41% des cas et signalent des égalités dans 15% des cas.

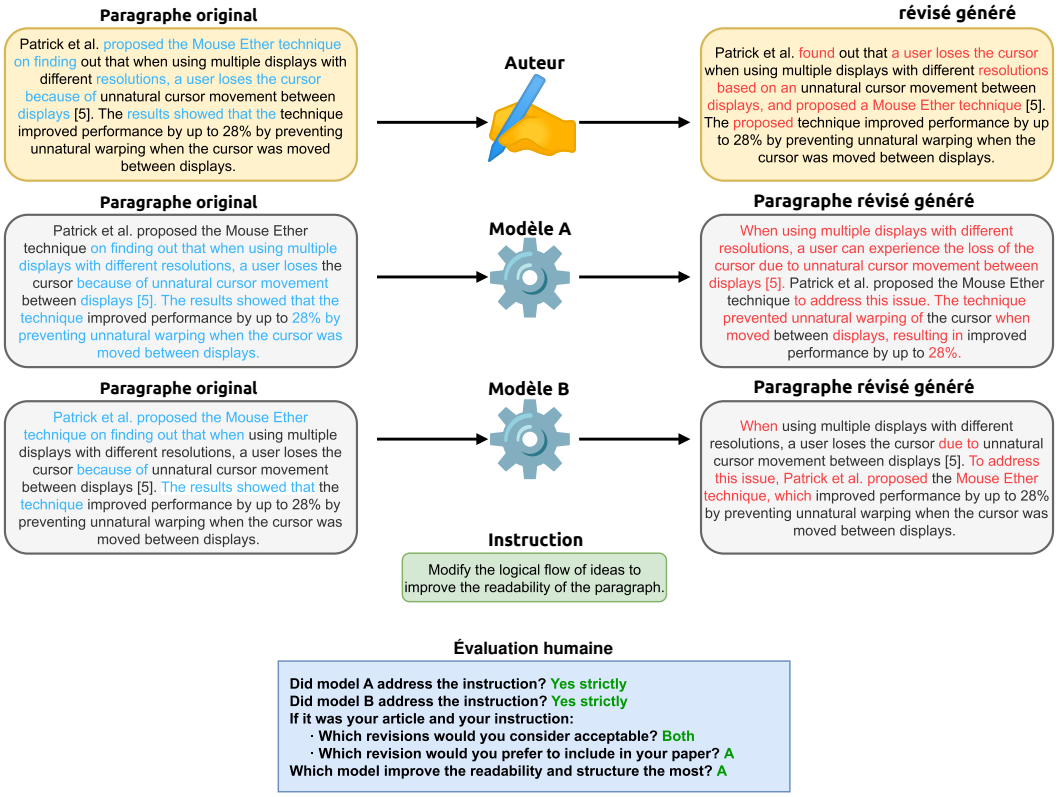


FIGURE 2 – Exemple de données dans le jeu de données ParaReval

D Amorces *LLM*-juge

Amorce 1: LLM-Yes/No pour la détection de paraphrase

```
system="""You will receive two sentences A and B, you will have to identify if they
mean the same thing.In your answer please only provide the answers to the question."""

user="""[BEGIN EXAMPLES]
***
[Sentence A]: Amrozi accused his brother , whom he called  the witness  , of
deliberately distorting his evidence .
[Sentence B]: Amrozi accused his brother , whom he disparagingly referred to as 'the
liar witness', of intentionally twisting his testimony .
No
***
[Sentence A]: Pennmakkal is an Indian Malayalam film from 1966 , produced by J.
Sasikumar and directed by KP Kottarakkara .
[Sentence B]: The Indian Malayalam film 'Pennmakkal', released in 1966, was produced by
 J. Sasikumar and directed by KP Kottarakkara .
Yes
***
[Sentence A]: Sorkin , who faces charges of conspiracy to obstruct justice and lying to
 a grand jury , was to have been tried separately .
[Sentence B]: Despite being accused of conspiring to obstruct justice and perjury ,
Sorkin was supposed to stand trial on his own .
No
***
[Sentence A]: Gilroy police and FBI agents described Gehring as cooperative , but said
Saturday that he had revealed nothing about what had happened to the children .
[Sentence B]: Although Gilroy police and FBI agents reported that Gehring was
cooperative , he hadn't disclosed any information about the children's whereabouts or
what had happened to them as of Saturday .
No
***
[Sentence A]: Whereas '' e `` the electric charge of the particle and A is the magnetic
 vector potential of the electromagnetic field .
[Sentence B]: The electric charge of the particle is denoted by ''e'', and the magnetic
 vector potential of the electromagnetic field is denoted by 'A' .
Yes
***
[Sentence A]: The Jidanul River is a tributary of the Jiul de Vest River in Romania .
[Sentence B]: The Jidanul River is a mere insignificant stream that flows into the
grand Jiul de Vest River in Romania .
No
***
[END EXAMPLES]

[BEGIN DATA]
[Sentence A]: "{source}"
***
[Sentence B]: "{paraphrase}"
***
[END DATA]

Do these two sentences express the same meaning? Answer "Yes" or "No".

"""+"""
You do not need to explain the reason.

Your response must be RFC8259 compliant JSON following this schema:
{{"answer": str }}"""
```

Afin d'évaluer dans des conditions similaires les méthodes Fr-PLUIE et *LLM-Yes/No*, nous avons conçu une amorce en français. Les résultats obtenus avec cette amorce sont disponibles dans l'annexe J.

Amorce 2: Variante française de l'amorce de *LLM-Yes/No* pour la détection de paraphrase

```
system="""Tu vas recevoir deux phrases, A et B, tu vas devoir identifier si elles
signifient la même chose. Dans ta réponse fournis uniquement la réponse à la
question."""

user="""[DEBUT EXEMPLES]
***
[Phrase A]: Amrozi a accusé son frère, qu'il appelait le témoin, d'avoir délibérément
déformé ses preuves .
[Phrase B]: Amrozi a accusé son frère, qu'il désignait de manière péjorative comme le
témoin menteur, d'avoir intentionnellement falsifié son témoignage.
Non
***
[Phrase A]: Pennmakkal est un film indien en malayalam de 1966, produit par J.
Sasikumar et réalisé par KP Kottarakkara .
[Phrase B]: Le film indien en malayalam 'Pennmakkal', sorti en 1966, a été produit par
J. Sasikumar et réalisé par KP Kottarakkara .
Oui
***
[Phrase A]: Sorkin, qui fait face à des accusations de complot pour entraver la justice
et de faux témoignage devant un grand jury, devait être jugé séparément .
[Phrase B]: Malgré les accusations de complot pour entraver la justice et de parjure,
Sorkin devait être jugé seul .
Non
***
[Phrase A]: La police de Gilroy et les agents du FBI ont décrit Gehring comme
coopératif, mais ont déclaré samedi qu'il n'avait révélé aucune information sur ce qui
était arrivé aux enfants.
[Phrase B]: Bien que la police de Gilroy et les agents du FBI aient rapporté que
Gehring était coopératif, il n'avait pas divulgué d'informations sur le lieu où se
trouvaient les enfants ou sur ce qui leur était arrivé samedi .
Non
***
[Phrase A]: Dans lequel "e" représente la charge électrique de la particule et A est le
vecteur du potentiel magnétique du champ électromagnétique .
[Phrase B]: La charge électrique de la particule est désignée par "e", et le vecteur du
potentiel magnétique du champ électromagnétique est désigné par 'A' .
Oui
***
[Phrase A]: La rivière Jidanul est un affluent de la rivière Jiul de Vest en Roumanie .
[Phrase B]: La rivière Jidanul est un simple ruisseau insignifiant qui se jette dans la
grande rivière Jiul de Vest en Roumanie .
Non
***
[FIN EXEMPLES]

[DEBUT DONNEES]
[Phrase A]: "{source}"
***
[Phrase B]: "{paraphrase}"
***
[FIN DONNEES]

Est-ce que ces deux phrases veulent dire la même chose ? Réponds par "Oui" ou "Non".

"""+"""
Tu n'as pas besoin d'expliquer la raison.

Ta réponse doit être compatible RFC8259 JSON et suivre le schéma suivant :
{{"réponse": str }}"""
```

Amorce 3: LLM-Yes/No pour la traduction Nile

```
system="""You are an evaluator of nile network policies to english translations. In
this task, a translation model has been provided with the original nile sentence and
translated it to english. You will be given the original nile sentence and the
proposition from this model and will have to report if the two sentences express the
same network policy. In your answer please only provide the answers to the
questions."""

user="""[BEGIN EXAMPLES]
***
[Sentence A]: Everquest is blocked by the University firewall
[Sentence B]: Everquest is not allowed by the University firewall
Yes
***
[Sentence A]: Quotas for students are 5000 Megabyte per hr download and 2000 Megabyte
per hour upload
[Sentence B]: Students have a download quota of 5000 MB per hour and an upload quota of
 2000 MB per hour
Yes
***
[Sentence A]: from endpoint('guests') to endpoint('servers') for group('host') block
traffic('any')
[Sentence B]: RHIT-OPEN will work in labs, classrooms and residence halls but does not
allow the user to access the internal Rose-Hulman network
No
***
[Sentence A]: Housing does not normally limit the amount of bandwidth
[Sentence B]: There is no bandwidth limit for the dorms
No
***
[Sentence A]: for endpoint('university') add middlebox('firewall') allow traffic('H323
video conferencing')
[Sentence B]: H323 video conferencing is allowed by the University firewall
Yes
***
[Sentence A]: H323 video conferencing is allowed by the University firewall
[Sentence B]: The network firewall does not block H.323 video conferencing
No
***
[END EXAMPLES]

[BEGIN DATA]
[Sentence A]: "{nile}"
***
[Sentence B]: "{translation}"
***
[END DATA]

Do these two sentences express the same network policy? Answer "Yes" or "No".

"""+"""
You do not need to explain the reason.

Your response must be RFC8259 compliant JSON following this schema:
{{"answer": str }}"""
```

Amorce 4: LLM-choice pour la traduction Nile

```
system="""You are an evaluator of nile network policies to english translations. In
this task, two translation models have been provided with the original nile sentence
and translated it to english. You will be given the proposition from the two different
models and a question to identify the best one. In your answer please only provide the
answers to the questions."""

user="""[BEGIN DATA]
***
[Nile network policy]: "{nile}"
***
[Sentence A]: "{translation 1}"
***
[Sentence B]: "{translation 2}"
***
[END DATA]


Which sentence is the best translation of the nile network policy? Answer "A", "B" or "
Tie".

"""+"""
You do not need to explain the reason.

Your response must be RFC8259 compliant JSON following this schema:
{{"best_translation": str }}"""
```

Amorce 5: LLM-Likert pour la traduction Nile

```
system="""You are an evaluator of nile network policies to english translations.
In this task, a translation model has been provided with the original nile sentence and
 translated it to english.
You will be given the proposition from the translation model and an affirmation to
determine the quality of this proposition.
You will answer this affirmation with a grade (int) from 1 to 5 as follows: 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 reference which is a correct
translation to english.
In your answer please only provide the answer to the affirmation."""

user="""[BEGIN DATA]
***
[Nile network policy]: "{nile}"
***
[Model proposed translation]: "{translation}"
***
[END DATA]


The model proposed a correct translation of the nile network policy?

"""+"""
You do not need to explain the reason.

Your response must be RFC8259 compliant JSON following this schema:
{{"translation_quality": int }}"""
```

E *-PLUIE nouvelles amorces spécifiques à la tâche

Dans cette section, nous fournissons les amorces utilisées pour la classification des paraphrases en français et la traduction Nile. Pour la révision de texte, pour *LLM-choice* et *LLM*-Likert, nous avons réutilisé les amorces et les évaluations générées par Jourdan *et al.* (2025b).

Amorce 6: Fr-PLUIE destiné à la détection de paraphrase en français

```
user= """Tu vas recevoir deux phrases, A et B. Est-ce que ces deux phrases veulent dire
la même chose ? Réponds avec un seul mot : "oui" ou "non"."""

assistant="""S'il te plaît, fournis-moi les deux phrases que je dois évaluer."""


user="""A: "Amrozi a accusé son frère, qu'il appelait "le témoin", d'avoir délibérément
déformé ses preuves ."; B: "Amrozi a accusé son frère, qu'il désignait de manière
péjorative comme "le témoin menteur", d'avoir intentionnellement falsifié son
témoignage."""

assistant="""non"""

user="""A: "Pennmakkal est un film indien en malayalam de 1966, produit par J.
Sasikumar et réalisé par KP Kottarakkara ."; B: "Le film indien en malayalam
'Pennmakkal', sorti en 1966, a été produit par J. Sasikumar et réalisé par KP
Kottarakkara ."""

assistant="""oui"""

user="""A: "Sorkin, qui fait face à des accusations de complot pour entraver la justice
et de faux témoignage devant un grand jury, devait être jugé séparément ."; B: "Malgré
les accusations de complot pour entraver la justice et de parjure, Sorkin devait être
jugé seul ."""

assistant="""non"""

user="""A: "La police de Gilroy et les agents du FBI ont décrit Gehring comme
coopératif, mais ont déclaré samedi qu'il n'avait révélé aucune information sur ce qui
était arrivé aux enfants ."; B: "Bien que la police de Gilroy et les agents du FBI
aient rapporté que Gehring était coopératif, il n'avait pas divulgué d'informations sur
le lieu où se trouvaient les enfants ou sur ce qui leur était arrivé samedi ."""

assistant="""non"""

user="""A: "Dans lequel "e" représente la charge électrique de la particule et A est le
vecteur du potentiel magnétique du champ électromagnétique ."; B: "La charge électrique
de la particule est désignée par "e", et le vecteur du potentiel magnétique du champ
électromagnétique est désigné par 'A' ."""

assistant="""oui"""

user="""A: "La rivière Jidanul est un affluent de la rivière Jiul de Vest en Roumanie
."; B: "La rivière Jidanul est un simple ruisseau insignifiant qui se jette dans la
grande rivière Jiul de Vest en Roumanie ."""

assistant="""non"""


user="""A: "{source}"; B: "{paraphrase}"""
```

Amorce 7: Net-PLUIE destiné à la politique réseau

```
user= """You will receive two sentences A and B. Do these two sentences express the
same network policy? Answer with only one word "Yes" or "No"."""

assistant= """Please provide the data for me to evaluate."""

user="""A: "Everquest is blocked by the University firewall"; B: "Everquest is not
allowed by the University firewall""""

assistant="""Yes"""

user="""A: "Quotas for students are 5000 Megabyte per hr download and 2000 Megabyte per
 hour upload"; B: "Students have a download quota of 5000 MB per hour and an upload
quota of 2000 MB per hour""""

assistant="""Yes"""

user="""A: "from endpoint('guests') to endpoint('servers') for group('host') block
traffic('any')"; B: "RHIT-OPEN will work in labs, classrooms and residence halls but
does not allow the user to access the internal Rose-Hulman network""""

assistant="""No"""

user="""A: "Housing does not normally limit the amount of bandwidth"; B: "There is no
bandwidth limit for the dorms""""

assistant="""No"""

user="""A: "for endpoint('university') add middlebox('firewall') allow traffic('H323
video conferencing')"; B: "H323 video conferencing is allowed by the University
firewall""""

assistant="""Yes"""

user="""A: "H323 video conferencing is allowed by the University firewall"; B: "The
network firewall does not block H.323 video conferencing""""

assistant="""No"""

user="""A: "{source}"; B: "{translation}""""
```

Amorce 8: Rev-PLUIE destiné au suivi d'instruction

```
user= """You will receive two paragraphs P1 and P2 and an instruction I. P1 is a
paragraph written for a scientific article, I is an instruction on how to revise P1 and
 P2 is a revised version of P1. Did P2 address the instruction I? Answer with only one
word "Yes" or "No"."""

assistant= """Please provide the data for me to evaluate."""

<La référence peut être ajoutée ici comme exemple>
user="""P1: "{original}"; I: "{instruction}"; P2: "{référence}""""

assistant= """Yes"""
<fin de l'inclusion de l'exemple>


user="""P1: "{original}"; I: "{instruction}"; P2: "{hypothese}""""
```

F Résultats pour les traductions Nile avec Mistral et les intentions sources

Modèle	Système	Seuil	f1	P.	R.
levenshtein	source	0.17	0.72	0.63	0.84
bleu	source	0.0	0.7	0.54	**1.0**
meteor	source	0.02	0.74	0.65	0.85
bertscore	source	0.43	0.73	0.66	0.82
modern_bertscore	source	0.65	0.7	0.54	**1.0**
Net-PLUIE	source	-1.91	**0.78**	0.67	0.93
Net-PLUIE	source	0	0.77	**0.71**	0.84
levenshtein	Mistral	0.13	0.95	0.91	0.99
bleu	Mistral	0.0	0.95	0.9	1.0
meteor	Mistral	0.0	0.95	0.9	1.0
bertscore	Mistral	0.38	0.95	0.9	1.0
modern_bertscore	Mistral	0.65	0.95	0.9	1.0
Net-PLUIE	Mistral	-14.74	0.95	0.9	1.0
Net-PLUIE	Mistral	0	0.92	**0.95**	0.89

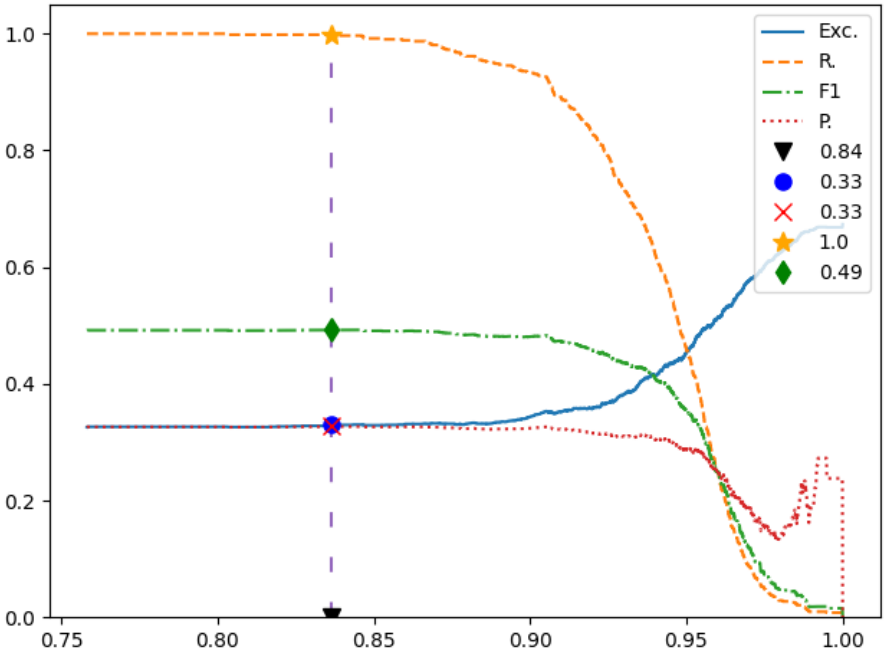
TABLE 5 – Évaluation des mesures pour les traductions générées par Mistral (Jiang *et al.*, 2023) et les intentions sources en anglais. Phi-4 14B est le modèle de perplexité utilisé dans Net-PLUIE.

G Distribution des scores en fonction du seuil de décision

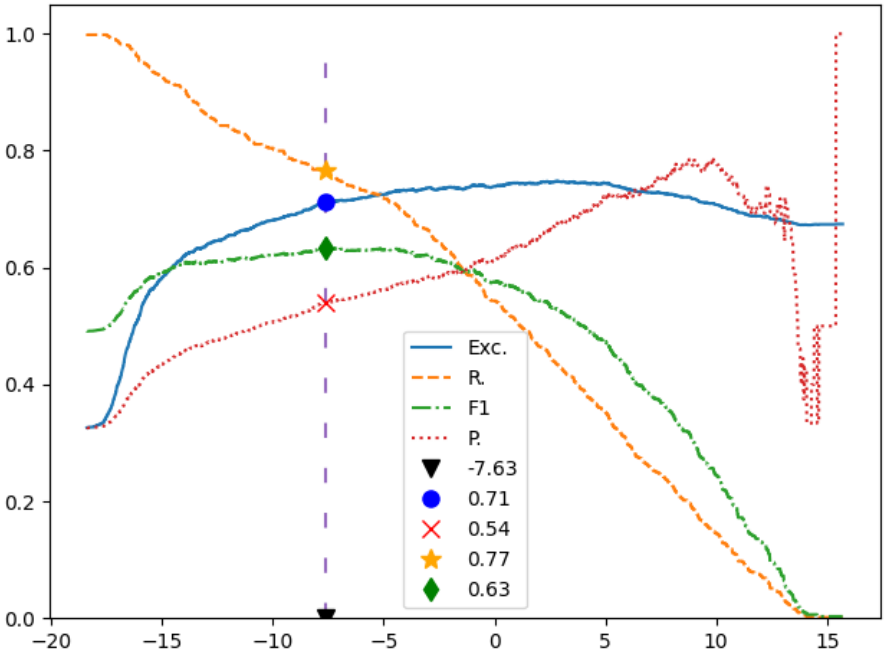
Pour mieux comprendre les résultats de la section 3.1, la figure 3 compare les distributions des scores de Modern BertScore et Fr-PLUIE sur la tâche de détection de paraphrases en français.

Comme le montre la figure 3a, Modern BertScore attribue principalement des scores de similarité élevés, et sa courbe d'exactitude ne diminue pas. La valeur 0,75 sur l'axe des x correspond au score Modern BertScore le plus bas parmi toutes les paires de phrases du jeu de données de paraphrases en français. La version complète de 0 à 1 serait plate de 0 à 0,85. Ce graphique peut sembler un peu déroutant car il ne montre pas la distribution des scores. Le Modern BertScore moyen des paires est de 0,95 (avec un niveau de confiance de 95% et un écart-type de 0,00). La plupart des paires non paraphrasées ont un score supérieur à celui des paires paraphrasées, et la paire ayant le score le plus élevé est étiquetée comme non paraphrasée. Cela indique qu'il ne parvient pas à attribuer des scores de similarité plus élevés aux paires de phrases qui véhiculent le même sens.

En revanche, Fr-PLUIE (figure 3b) différencie avec succès les paraphrases et les non-paraphrases, en attribuant des scores négatifs à ces dernières. Cependant, la diminution rapide du rappel révèle que le modèle produit également un certain nombre de paires de faux négatifs étiquetées comme paraphrases mais jugées comme non-paraphrases, ce qui indique qu'il y a une marge de progression.



(a)

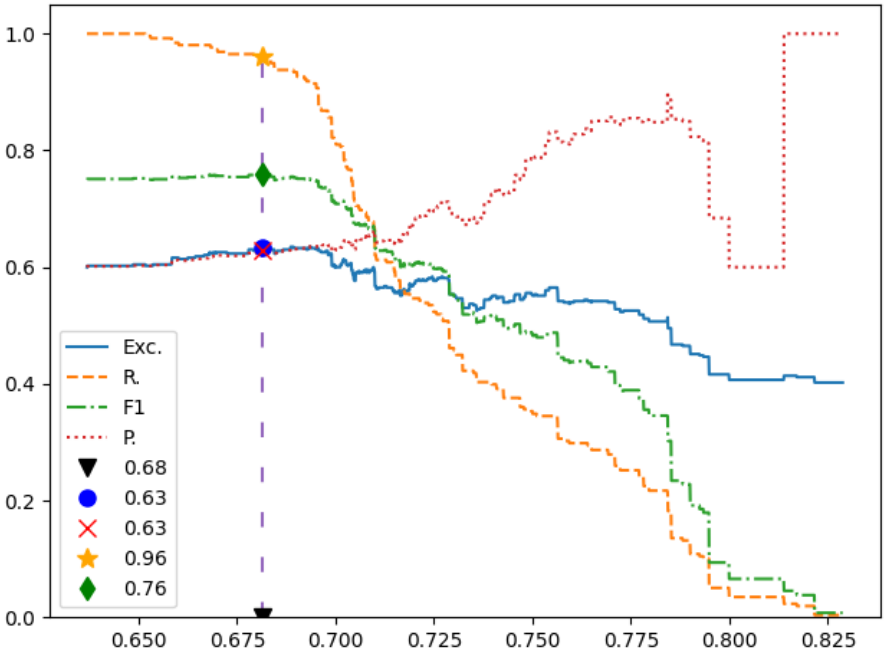


(b)

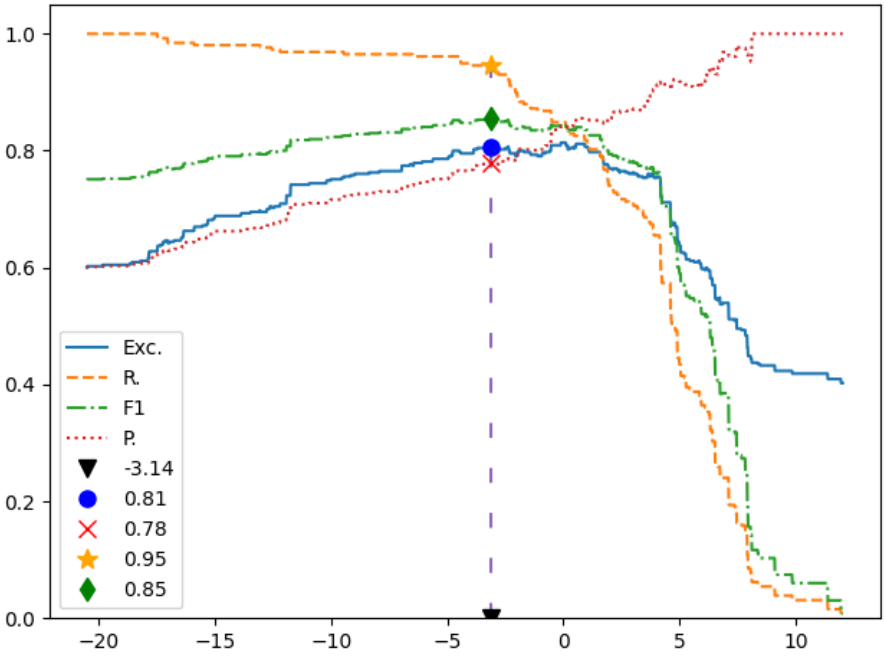
FIGURE 3 – Distribution des scores de Modern BertScore (a) et Fr-PLUIE (b). Les courbes bleu, orange, rouge et verte représentent respectivement les scores exactitude, rappel, précision et F1 en fonction du seuil de décision. L'accent est mis sur le score F1 maximal obtenu par la métrique.

La figure 4 présente la précision, le rappel, l'exactitude et le score F1 pour différentes valeurs de seuils pour Modern BertScore et Net-PLUIE. Pour Net-PLUIE (figure 4b), la précision augmente avec le seuil, réduisant les faux positifs. L'exactitude et le score F1 augmentent initialement pour des valeurs de seuil comprises entre environ -20 et 0, puis diminuent rapidement au-delà de 0. En

revanche, Modern BertScore (figure 4a) montre une augmentation initiale plus lente et plus courte de l'exactitude et du score F1, suivie d'une diminution continue.



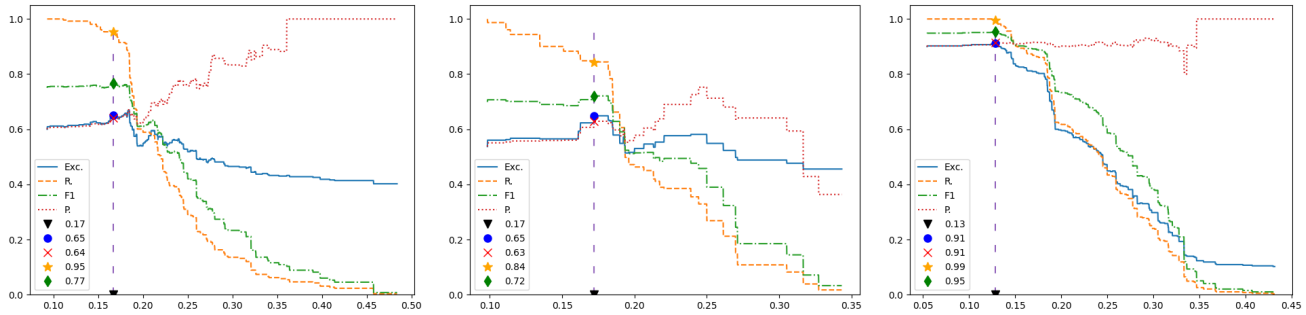
(a)



(b)

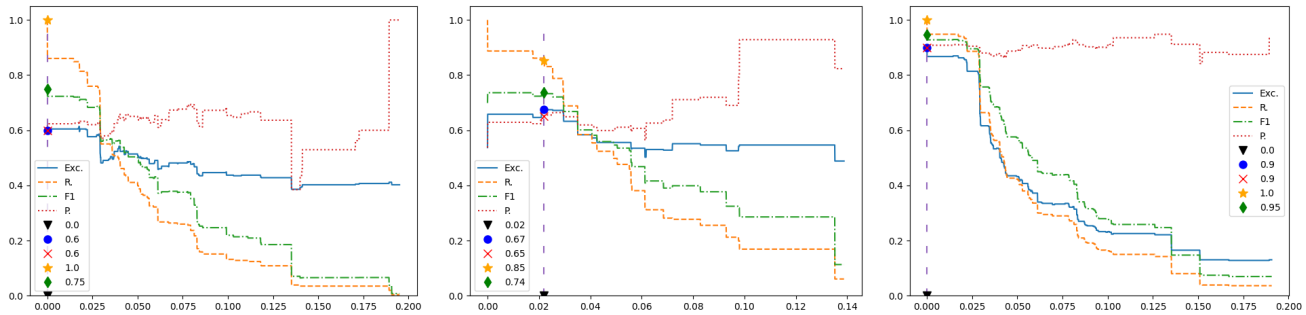
FIGURE 4 – Distribution des scores de l'Exactitude, le rappel, la précision et le score F1 pour différentes valeurs de seuils pour Modern BertScore (a) et Net-PLUIE (b). L'accent est mis sur le score F1 maximal obtenu par la mesure.

Les figures 5, 6, 7, 8, 9, 10, 11 montrent les différents résultats de classification avec toutes les combinaisons de mesures.



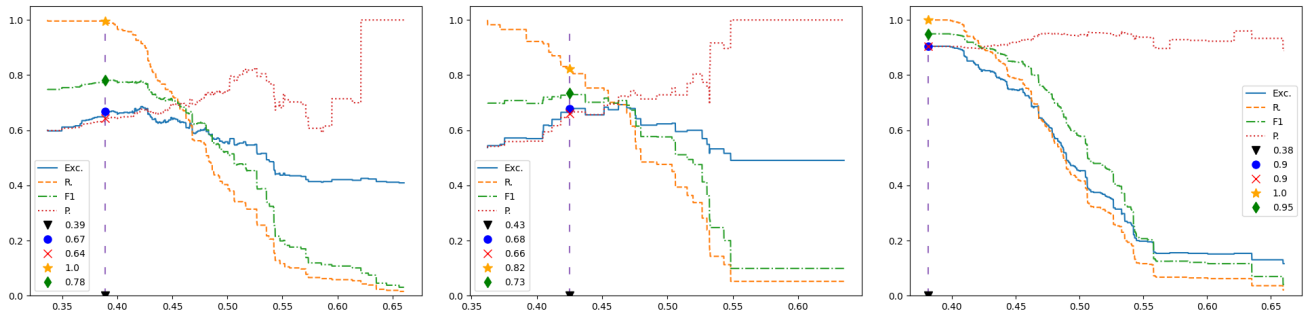
(a) Traductions Llama (b) Intentions source en anglais (c) Traductions Mistral

FIGURE 5 – Classification avec Levenshtein.



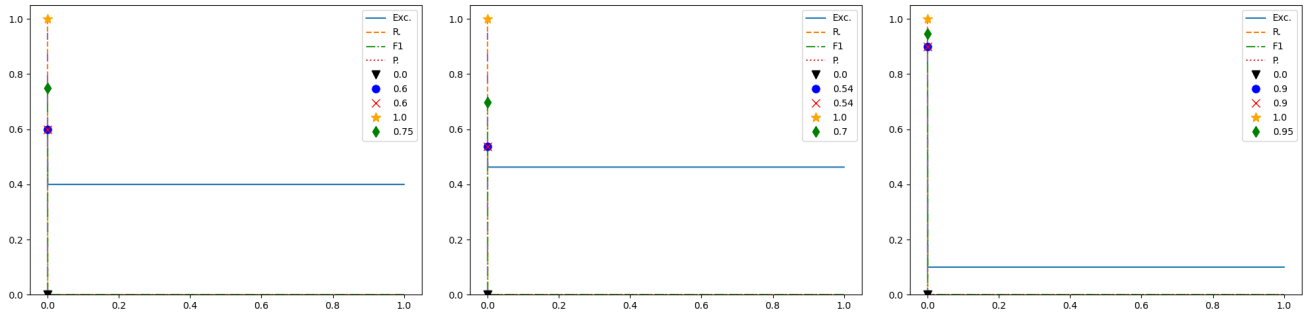
(a) Traductions Llama (b) Intentions source en anglais (c) Traductions Mistral

FIGURE 6 – Classification avec METEOR.



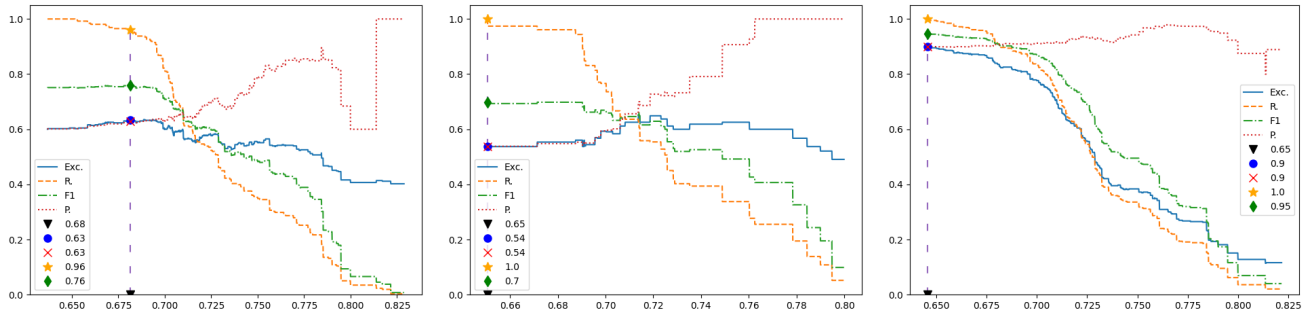
(a) Traductions Llama (b) Intentions source en anglais (c) Traductions Mistral

FIGURE 7 – Classification avec BERTScore.



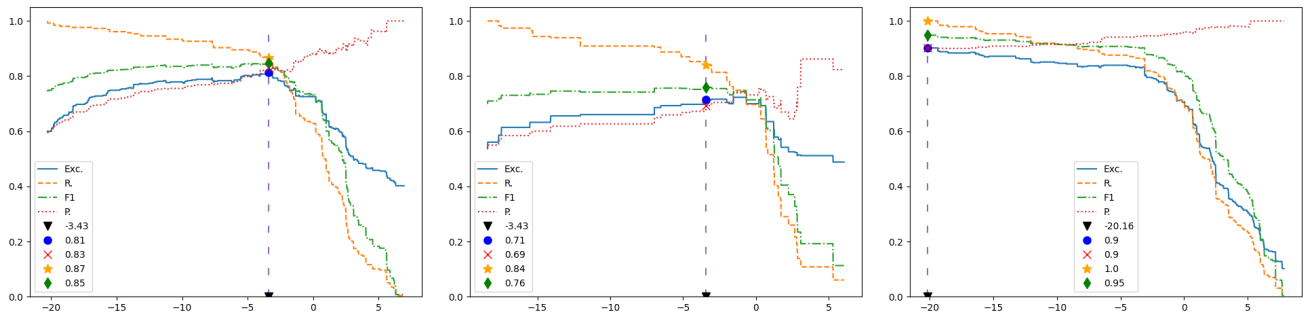
(a) Traductions Llama (b) Intentions source en anglais (c) Traductions Mistral

FIGURE 8 – Classification avec BLEU.



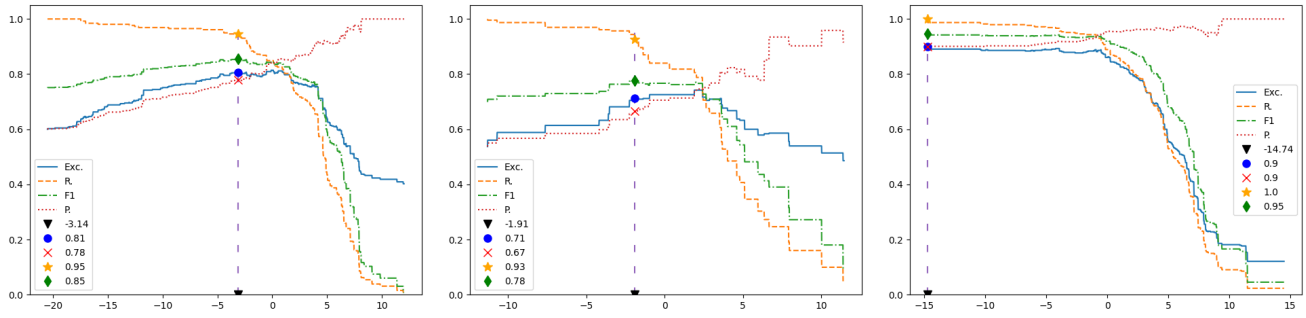
(a) Traductions Llama (b) Intentions source en anglais (c) Traductions Mistral

FIGURE 9 – Classification avec Modern BertScore.



(a) Traductions Llama (b) Intentions source en anglais (c) Traductions Mistral

FIGURE 10 – Classification avec Para-PLUIE Phi-4 14B.



(a) Traductions Llama (b) Intentions source en anglais (c) Traductions Mistral

FIGURE 11 – Classification avec Net-PLUIE Phi-4 14B.

H Résultats supplémentaires pour l'alignement des mesures basées sur les *LLMs* pour la révision de texte

	Juge	Exc. Paire		V		κ	
			réf		réf		réf
Rev-PLUIE	Llama 3 70B	**0.61**	**0.62**	**0.31**	**0.32**	**0.32**	**0.34**
	Phi-4 14B	**0.61**	0.61	**0.31**	**0.32**	**0.32**	0.33
	Llama 3 8B	0.58	0.59	0.27	0.29	0.27	0.29
	Mistral 7B	0.55	0.59	0.24	0.28	0.22	0.28
Para-PLUIE	Llama 3 70B	0.52		0.20		0.17	
	Phi-4 14B	0.52		0.21		0.15	
	Llama 3 8B	0.52		0.21		0.17	
	Mistral 7B	0.54		0.22		0.20	
LLM choice	GPT-4o	0.59	0.60	0.28	0.30	0.30	0.33
	GPT-4o mini	0.57	0.58	0.25	0.26	0.27	0.29
	Llama 3 70B	0.59	0.60	0.28	0.30	0.30	0.31
	Phi-4 14B	0.53	0.55	0.25	0.27	0.24	0.27
	Llama 3 8B	0.54	0.51	0.21	0.18	0.20	0.17
	Mistral 7B	0.53	0.53	0.20	0.17	0.17	0.16
LLM likert	GPT-4o	0.54	0.54	0.28	0.27	0.28	0.27
	GPT-4o mini	0.45	0.51	0.28	0.27	0.21	0.23
	Llama 3 70B	0.44	0.50	0.29	0.27	0.19	0.23
	Phi-4 14B	0.45	0.52	0.30	0.29	0.21	0.26
	Llama 3 8B	0.43	0.45	0.18	0.18	0.15	0.15
	Mistral 7B	0.33	0.28	0.16	0.10	0.09	0.05

TABLE 6 – L'exactitude par paires et le V de Cramér sont définis sur [0 :1] et le Kappa de Cohen sur [-1 :1] "réf" indique que la révision de référence est fournie. Pour Para-PLUIE, la colonne sans référence correspond aux scores Para-PLUIE entre les paragraphes originaux et générés. Les modèles de perplexité utilisés sont Llama-3, Phi-4, Mistral et GPT-4o (Grattafiori *et al.*, 2024; Abdin *et al.*, 2024; Jiang *et al.*, 2023; Hurst *et al.*, 2024).

I Net-PLUIE en tant que fonction d'alignement

Munson *et al.* (2025) a introduit NEAT, une méthodologie utilisée pour créer un corpus à grande échelle d'intentions alignées anglais-Nile. Dans leur méthodologie, les auteurs formalisent la fonction $align(i, n)$ qui évalue si une intention Nile (i) et une déclaration en langage naturel (n) partagent la même signification sous-jacente. Jusqu'à présent, $align(\cdot, \cdot)$ n'a été mis en œuvre de manière fiable que grâce à l'annotation manuelle d'experts, ce qui limite son passage à l'échelle et sa reproductibilité. Pour valider Net-PLUIE en tant que fonction d'alignement sémantique, nous exigeons que deux conditions soient vérifiées, présentées dans les équations 5 et 6.

$$align(i, n_1) \wedge align(i, n_2) \implies \star\text{-PLUIE}(n_1, n_2) > T \wedge \star\text{-PLUIE}(n_2, n_1) > T \quad (5)$$

$$align(i, n_1) \wedge \star\text{-PLUIE}(n_1, n_2) > T \wedge \star\text{-PLUIE}(n_2, n_1) > T \implies align(i, n_2) \quad (6)$$

Par conséquent, si deux phrases (n_1 et n_2) ont le même sens et que l'une d'elles est alignée avec une intention, la seconde devrait également être alignée avec cette intention. Cela est possible car le langage des intentions est sans ambiguïté, ce qui n'est pas nécessairement le cas avec le langage naturel. Nous tirons parti de la propriété non symétrique de *-PLUIE et calculons le score Net-PLUIE(n_1,n_2) et Net-PLUIE(n_2,n_1).

Cependant, lorsqu'il est utilisé comme mesure d'alignement entre le langage naturel et le langage formel des intentions, nous ne l'utilisons qu'une seule fois, en prenant l'intention formelle comme référence. En effet, les expressions en langage naturel capturent généralement plus de détails que les langages formels, et ces détails supplémentaires capturés dans l'expression doivent toujours être considérés comme faisant partie d'une bonne traduction. Cependant, la métrique pénalise la suppression d'informations entre deux expressions, ce qui signifie que la perte d'informations qui ne sont pas pertinentes pour les configurations réseau réelles lors du passage du langage naturel à Nile est pénalisée par la métrique.

Nous considérons les traductions générées par Mistral et Llama comme nos deux sources d'expressions en langage naturel et comptons le nombre de fois où nos deux implications ne sont pas respectées. Au total, sur les 430 triplets Nile-NL, 56 ne vérifient pas la condition 5 et 7 ne vérifient pas la condition 6. Des exemples de traductions qui ont échoué aux tests sont présentés dans le tableau 7 et le tableau 8.

Bien que seulement 15% des triplets ne répondent pas à nos conditions imposées, cela rend l'utilisation de cette métrique incertaine pour une utilisation confortable en tant que fonction d'alignement sémantique sans vérification supplémentaire des traductions générées avant leur acceptation. Heureusement, la condition la plus critique est la condition 6, qui permet de déduire un alignement à partir d'un alignement existant et d'une phrase paraphrasée, et cette condition est satisfaite dans 98% des cas. Nous pouvons donc considérer Net-PLUIE comme la solution la plus proche dont nous disposons pour obtenir une fonction d'alignement sémantique automatique précise entre le langage naturel et un langage formel d'intention, ce qui en fait la meilleure solution disponible.

Phrase Nile	Traduction avec Llama (Grattafiori *et al.*, 2024)	Traduction avec Mistral (Jiang *et al.*, 2023)
for endpoint('dorms'), endpoint('wireless') set quota('any', '6', 'gb/d') start hour('00 :00') end hour('23 :59')	The ResNet and NYURoam wireless networks have a 6 GB per day quota	The ResNet and wireless network have a daily cap of 6 GB per day, from midnight to midnight
for endpoint('dorms'), endpoint('wireless') set quota('any', '6', 'gb/d') start hour('00 :00') end hour('23 :59')	The ResNet cap is currently 6 GB per day	ResNet and UIC wireless have a daily cap of 6 GB for all incoming and outgoing traffic
for endpoint('dorms'), endpoint('wireless') set quota('any', '6', 'gb/d') start hour('00 :00') end hour('23 :59')	The ResNet cap is currently 6 GB per day	ResNet and wireless users have a daily quota of 6 GB, with no time restrictions
for endpoint('dorms'), endpoint('wireless') set quota('any', '6', 'gb/d') start hour('00 :00') end hour('23 :59')	The ResNet cap is currently 6 gigabytes per day	The ResNet cap is currently 6 gigabytes per day for each 24 hour period of combined incoming and outgoing traffic
for endpoint('guests') block service('file transfer')	Guests are not allowed to run file transfer services	Guests are not allowed to transfer files
for endpoint('guests') block service('file transfer')	Guests are not allowed to use file transfer protocols	Guests are not allowed to transfer files
for endpoint('host') set bandwidth('max', '3', 'mbps')	The individual will have a 3 Mbps bandwidth rate limit	The host will have a 3 Mbps bandwidth rate limit
for endpoint('host') set bandwidth('max', '4', 'mbps')	Hosts' network is rate-limited to 4 Mb/s per user	Each host is limited to 4 Mb/s
for endpoint('host') set bandwidth('max', '4', 'mbps')	Hosts' network is rate-limited to 4 Mb/s per user	Hosts are rate-limited to 4 Mb/s
for endpoint('host') set bandwidth('max', '4', 'mbps')	Internet is limited to 4 Mbps per host	Each host is limited to 4 Mbps
for endpoint('network') add middlebox('network border system'), middlebox('ips'), middlebox('firewall'), middlebox('unit firewall')	The network is protected by a firewall, a network border system, and a unit firewall	The network has multiple layers of security, including a network border system, intrusion prevention system, firewall, and unit firewall
for endpoint('network') add middlebox('overuse notification') set quota('any', '10', 'gb/wk')	The network will send an email to the user when they have used 10 GB of bandwidth in a week	If you exceed 10 GB of network usage in a week, you will receive a notification
for endpoint('network') block service('NETBIOS')	The UVA More Secure Network does not allow NetBIOS	NETBIOS is blocked by the University of Virginia's firewall
for endpoint('network') block service('NETBIOS')	The network firewall does prevent some applications from working in their normal manner, for example, NetBIOS	NETBIOS is blocked by the network firewall
for endpoint('network') block service('NETBIOS')	The network firewall does prevent some applications from working in their normal manner, for example, NetBIOS	The network firewall blocks the use of the NETBIOS protocol
for endpoint('network') block traffic('H323 video conferencing')	H323 video conferencing is blocked by the University network	H323 video conferencing is blocked by the network
for endpoint('network') block traffic('H323 video conferencing')	H323 video conferencing is blocked	H323 video conferencing is blocked on the network
for endpoint('network') set quota('any', '10', 'gb/d')	The University of Illinois at Urbana-Champaign has a 10 GB per day limit on all combined traffic	The network has a 10 GB per day data transfer limit

TABLE 7 – Quelques exemples de mauvaises traductions qui ne respectent pas l'implication 5, les deux traductions sont alignées sur la même expression d'intention, mais ParaPLUIE ne considère pas la traduction comme une paraphrase.

Phrase Nile	Traduction avec Llama (Grattafiori *et al.*, 2024)	Traduction avec Mistral (Jiang *et al.*, 2023)
for endpoint('dorms') block protocol('nat')	NAT is not allowed on the ResNet	NAT is blocked on the ResNet
for endpoint('dorms'), endpoint('wireless') block protocol('ftp'), protocol('telnet'), traffic('peer2peer'), protocol('http'), service('irc'), protocol('dns'), service('email')	You may not run any services (e.g., FTP services, telnet, P2P file sharing servers, publishing files via generic file services, etc.) via NYU's ResNet or NYURoam wireless networks	ResNet and wireless networks are blocked from using FTP, Telnet, P2P, HTTP, IRC, DNS, and email services
for endpoint('dorms'), endpoint('wireless') block protocol('ftp'), service('listserver'), traffic('peer2peer')	The University of Washington does not allow FTP, Listserv, or Peer-to-Peer traffic on the ResNet or UW Wireless	ResNet and UIC wireless do not allow FTP, listserv, or peer-to-peer traffic
for endpoint('dorms'), endpoint('wireless') block protocol('ftp'), service('listserver'), traffic('peer2peer')	You may not run FTP, listserver, peer-to-peer, or file sharing/downloading program	ResNet and wireless users cannot use FTP, listserv, or peer-to-peer file sharing
for endpoint('dorms'), endpoint('wireless') set quota('any', '6', 'gb/d') start hour('00 :00') end hour('23 :59')	Dorms and wireless have a quota of 6 GB per day	The quota for dorms and wireless is 6 GB per day, with no time restrictions
for endpoint('dorms'), endpoint('wireless') set quota('any', '6', 'gb/d') start hour('00 :00') end hour('23 :59')	Students have a quota of 6 GB per day The quota is reset at midnight	ResNet and wireless users have a daily quota of 6 GB, with no time restrictions
for endpoint('dorms'), endpoint('wireless') set quota('any', '6', 'gb/d') start hour('00 :00') end hour('23 :59')	The ResNet and NYURoam wireless networks have a 6 GB per day quota	ResNet and wireless network have a daily quota of 6 GB, available 24/7

TABLE 8 – Mauvaises traductions qui ne respectent pas l'implication 6 : une traduction est alignée sur l'expression Nile, les deux traductions sont des paraphrases, mais, aux lignes 1 à 4, Parapluie ne considère pas que la traduction de Llama soit alignée sur l'expression Nile, et aux lignes 5 à 7, Parapluie ne considère pas que la traduction de Mistral soit alignée sur l'expression Nile.

J Résultats supplémentaires pour la classification de paraphrases en français

	Détection de paraphrase en français					
Mesure	**Exc.**	**R.**	**P.**	**F1**	**GPUs**	**Temps**
LLM-Yes/No Phi-4 14B (Français)	0.74	0.53	0.61	0.57	MI300 x1	27 min
LLM-Yes/No Llama 3 70B (Français)	0.71	0.57	0.55	0.56	MI300 x2	57 min

TABLE 9 – Résultats des approches LLM-juge avec une amorce en français.

Mesure	Détection de paraphrase en français Seuil	Exc.	R.	P.	F1
Para-PLUIE SmolLM2-135M-Instruct	0	0.67	N/A	0.00	0.00
Para-PLUIE SmolLM2-135M-Instruct	-2.31	0.33	0.33	1.00	0.49
Para-PLUIE SmolLM2-360M-Instruct	0	0.67	0.35	0.03	0.06
Para-PLUIE SmolLM2-360M-Instruct	-1.26	0.33	0.33	1.00	0.49
Para-PLUIE SmolLM2-1.7B-Instruct	0	0.61	0.41	0.44	0.43
Para-PLUIE SmolLM2-1.7B-Instruct	-0.75	0.41	0.35	0.92	0.51
Para-PLUIE internlm2-chat-1_8b	0	0.39	0.34	0.93	0.50
Para-PLUIE internlm2-chat-1_8b	0.23	0.43	0.35	0.89	0.50
Para-PLUIE gemma-2-2b-it	0	0.61	0.44	0.71	0.54
Para-PLUIE gemma-2-2b-it	-0.67	0.58	0.42	0.77	0.55
Para-PLUIE Phi-4-mini-instruct	0	0.65	0.48	0.68	0.56
Para-PLUIE Phi-4-mini-instruct	-2.52	0.58	0.43	0.85	0.57
Para-PLUIE Mistral-7B-Instruct-v0.2	0	0.67	0.49	0.48	0.49
Para-PLUIE Mistral-7B-Instruct-v0.2	-10.71	0.59	0.43	0.76	0.55
Para-PLUIE Qwen2.5-7B-Instruct	0	0.71	0.55	0.59	0.57
Para-PLUIE Qwen2.5-7B-Instruct	-13.04	0.67	0.50	0.73	0.59
Para-PLUIE aya-expanse-8b	0	0.60	0.43	0.65	0.52
Para-PLUIE aya-expanse-8b	-5.7	0.53	0.40	0.87	0.54
Para-PLUIE Llama-3.1-8B-Instruct	0	0.72	0.58	0.48	0.52
Para-PLUIE Llama-3.1-8B-Instruct	-6.52	0.64	0.47	0.79	0.59
Para-PLUIE gemma-2-9b-it	0	0.70	0.54	0.55	0.54
Para-PLUIE gemma-2-9b-it	-5.83	0.63	0.46	0.78	0.58
Para-PLUIE aya-expanse-32b	0	0.65	0.48	0.63	0.54
Para-PLUIE aya-expanse-32b	-2.65	0.62	0.45	0.74	0.56
Para-PLUIE QwQ-32B	0	0.70	0.53	0.64	0.58
Para-PLUIE QwQ-32B	-0.13	0.70	0.53	0.67	0.59
Para-PLUIE c4ai-command-r-08-2024	0	0.70	0.59	0.29	0.39
Para-PLUIE c4ai-command-r-08-2024	-7.26	0.62	0.45	0.79	0.57

TABLE 10 – Résultats des approches Para-PLUIE avec classification de paraphrases en français à l'aide de différents modèles de perplexité. Dans l'ordre suivant, SmolLM2-Instruct (Allal *et al.*, 2025), internlm2-chat (Cai *et al.*, 2024), gemma-2 (Team *et al.*, 2024), Phi-4-mini-instruct (Abouelenin *et al.*, 2025), Mistral-7B-Instruct (Jiang *et al.*, 2023), Qwen2.5-Instruct (Yang *et al.*, 2025), aya-expanse (Dang *et al.*, 2024), Llama-3.1-Instruct (Grattafiori *et al.*, 2024), QwQ (Team, 2025) et c4ai-command-r (Verga *et al.*, 2024).