[*]-PLUIE :
Personalisable metric in Llm Used for Improved Evaluation.

Anonymous ACL submission

Abstract

Evaluating the quality of text generated by large language models (LLMs) often relies on LLM-as-a-judge methods, where a model is prompted to assess system outputs. While effective, these approaches are computationally expensive and require post-processing of free form text responses. To address these limitations, we build upon PARAPLUIE, a perplexity-based LLM-as-a-judge metric that estimates confidence over "Yes/No" answers without generating text. In this paper, we investigate how this methodology generalises to another language on its original task and beyond paraphrase detection to two other complex semantic tasks : scientific text revision and network intent translation. We also introduce task specific prompting variants of PARAPLUIE and evaluate their alignment with human judgement across multiple datasets. Our experiments show that personalised PARAPLUIE prompts achieve stronger correlations with human ratings and greater interpretability than both traditional similarity metrics (e.g., BLEU, BERTScore) and output based LLM judges, while being six times faster to compute. These findings establish *-PLUIE approaches as efficient, reliable, and adaptable alternatives for automated evaluation across diverse semantic tasks.

1 Introduction

Automatic evaluation is a persistent challenge in NLP tasks. Traditional similarity-based metrics such as BLEU (Papineni et al., 2002) or METEOR (Banerjee and Lavie, 2005) focus on surface-level lexical overlap and often fail to capture meaning-preserving variations or stylistic improvements. Recent advances in LLMs have introduced LLM-as-a-judge methods, which use the reasoning capabilities of LLMs for interpreting user demands (Brown et al., 2020) to evaluate generated text and aptitude. These approaches leverage the semantic understanding and contextual reasoning of LLMs to provide richer, more human-aligned assessments across a variety of NLP tasks (Gu et al., 2024; Doostmohammadi et al., 2024).

However, standard LLM-as-a-judge techniques require generating free-form text responses that must then be parsed into structured judgements. This parsing introduces noise, ambiguity, and additional computational cost, especially when only simple categorical decisions (e.g., Yes/No) are required. Moreover, this text-generation process may not fully exploit the model's internal probabilistic knowledge.

To mitigate these limitations, Lemesle et al. (2025) proposed PARAPLUIE as a perplexity-based alternative to output-based LLM judging. It grants a confidence score for a "Yes"/"No" question by relying on the perplexity of the LLM on these two possible answers. Originally introduced for paraphrase detection, PARAPLUIE achieved strong alignment with human annotations with a minimal computational cost, roughly equivalent to generating a single token.

By definition, PARAPLUIE represents the model's confidence in its answer, offering several appealing interpretability properties. Intuitively, a strong positive score should indicate high confidence in the "Yes" response, while a strong negative score should indicate confidence in "No." However, it has so far been evaluated on a single classification task and only in English. Therefore, its applicability to other tasks and languages, as well as its validity with respect to human judgements, remains to be established.

In this paper, we extend the PARAPLUIE framework to new semantic evaluation tasks beyond paraphrase detection. Specifically, we investigate whether task-specific prompting, i.e. tailoring the question to the evaluation context, can improve the reliability and generalisation of PARAPLUIE. We explore three complementary applications:

(i) **Paraphrase detection in French**, to test cross-lingual generalisation; (ii) **Translation alignment**, to assess its ability to capture correspondence between natural and formal intent representations; and (iii) **Scientific text revision**, a rewriting task that requires capturing instruction following and improvement in text quality.

For each task, we design specialised prompts (PARA-FRENCH-PLUIE, NET-PLUIE, and REV-PLUIE) that adapts the binary question to the evaluation context. We compare our variants against widely used similarity-based metrics (e.g., BLEU, METEOR, SARI, BERTScore) and output-based LLM-as-a-judge baselines for text revision.

Our main contributions are as follows :

(1) We introduce *-PLUIE, a general and personalisable perplexity based method for LLM-as-a-judge evaluation.

(2) We design and evaluate three task-specific variants of *-PLUIE, covering paraphrase detection, intent translation, and text revision, to assess the adaptability of the approach.

(3) Across all tasks, these variants achieve stronger alignment with human judgements and outperform traditional similarity metrics, while remaining competitive with output-based LLM-as-a-judge and up to six times faster to compute.

2 *-PLUIE definition

*-PLUIE is originally defined as the log-likelihood ratio to compare the predominance of an answer "Yes" against "No" to the question raised by the prompt. More formally, given:

- S the source sentence,
- H the hypothetical sentence,
- Prompt(S,H) the prompt filled with the sentences S and H,
- T: length in tokens of the Prompt(S,H),
- $\circ$: the concatenation of 2 sequences of tokens.

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

It can be expanded, using the Bayes formula, to make appear the perplexity of the template:

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

Finally, assuming that the used *LLM* uses log perplexity as a $loss$, we have:

$$\begin{aligned}
&\texttt{*-PLUIE}(S,H) = (T+1)\times \\
&(loss_{LLM}(\text{Prompt}(S,H)\circ\text{No}) - loss_{LLM}(\text{Prompt}(S,H)\circ\text{Yes})) \quad (3)
\end{aligned}$$

Note that, PARAPLUIE is not necessarily symmetrical. The workflow of *-PLUIE is illustrated in Figure 1.

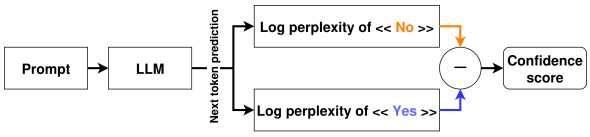


Figure 1: *-PLUIE workflow.

*-PLUIE presents plug-and-play capabilities: both the underlying perplexity model and the prompt template can be modified without altering the fundamental principle of the method. However, a few aspects must be taken into consideration.

First, the perplexity model must be able to perform question-answering tasks, as the metric relies on evaluating the model's confidence between two mutually exclusive answers. Note that a requirement of Equation 2 is that the two answers being opposed each correspond to exactly one token according to the model's tokenizer. Beyond that, the choice of tokens is not restricted (it could be "Dog" vs. "Cat"), even though we only consider "Yes" vs. "No" in this study.

Second, to ensure correct perplexity computation, the tokens being compared must appear as the final tokens of the prompt. Tokenizers of most modern LLM-based chatbots employ a user–assistant dialogue format, where special tokens are inserted to mark the end of each role. These role-ending tokens must be removed prior to perplexity computation, as they would otherwise alter the results. For example, when computing the perplexity of "Yes" in the following dialogue:

`<user>People like "Cats" more than "Dogs".</user> <assistant> Yes </assistant>`

the end-of-turn token `</assistant>` must be removed before calculating perplexity.

In this paper, we rely on the publicly available implementation provided by the original authors[1].

3 Experimental protocol

We investigate the behaviour of the PARAPLUIE metric across three distinct semantic tasks: paraphrase detection, Nile-English translation, and text revision. In the following sections, we define each task, describe the datasets used, and discuss how the PARAPLUIE methodology was adapted through prompt design. To assess the reliability of the metric, we align its outputs to human annotations available in the corresponding datasets and compare it against standard semantic similarity metrics.

3.1 Semantic tasks

This section defines the three studied tasks and introduces the datasets used in our experiments.

3.1.1 Paraphrase detection

As PARAPLUIE demonstrated strong performances in English paraphrase detection (Lemesle et al., 2025), we extend this analysis to French, to examine its generalisation capabilities across languages.

For this purpose, we build on the work of Tytgat et al. (2024) who showed that conventionally used semantic similarity metrics are often more sensitive to surface-level differences than to semantic variations. Their study introduced an expert-annotated French dataset of 355 source sentences, each constructed independently of any specific domain.

For each original sentence, the authors identified the central semantic word, the one that most conveys the meaning of the sentence, and systematically produced three types of modified versions:

Paronym substitution: the word is replaced by a paronym, a word that looks or sounds similar but has a completely different meaning. The resulting sentence becomes semantically different from the original while remaining lexically close.

Synonym substitution: the word is replaced by a synonym, producing a sentence that preserves the original meaning while being lexically close.

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

Synonym of a paronym: the word is replaced by a synonym of the paronym, leading to an additional nuanced variant.

This process results in three alternative versions for each source sentence, an example of these transformations is provided in Table 1.

Transformation	Sentence
None	Les enfants ont boulotté tous les gâteaux.
Paronym	Les enfants ont boulonné tous les gâteaux
Synonym	Les enfants ont mangé tous les gâteaux.
Syn. of Par.	Les enfants ont fixé tous les gâteaux.

Table 1: Characteristic sentence of Tytgat et al. (2024) dataset and its different transformations.

We used this as a paraphrase set. Let's denote the transformation by synonym S and by paronym P. We create a new set composed of couples of sentences formed by the source sentence s and its transformation. Depending on the combination of transformations, we label them as paraphrase or non-paraphrase:

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

In the end, this dataset contains 1914 couples where 67.40% are non-paraphrases.

3.1.2 Network intent translation

The paradigm of Intent Based Networking (IBN) has emerged as a promising approach to simplify network and security management by enabling operators to express high-level intents rather than low-level configurations (Leivadeas and Falkner, 2023). In such systems, user-defined objectives, e.g., performance or security directives, are translated into enforceable policies, which are then deployed across network devices and enforcement points. This shift holds the potential to reduce configuration errors and accelerate deployment times, while offering more natural and flexible interaction with network infrastructures.

Defined in Jacobs et al. (2018), Network Intent LanguagE (Nile) provides a structured yet flexible grammar for expressing access control, Quality of Service, and temporal statements, making it well-suited for bridging the gap between natural language inputs and enforceable network policies. Several studies, such as McNamara et al. (2022), have already relied on Nile for designing translation pipelines or evaluating NLP-driven frameworks, positioning it as the de facto reference language for IBN research. Despite these efforts, the scarcity of robust, large-scale datasets continues to hinder progress, leading to systems that are either too rigid, domain-specific, or error-prone for operational deployment.

To address these challenges, Munson et al. (2025) introduced Nile-English Aligned Translations (NEAT), a methodology used to create a large-scale corpus of aligned English–Nile intents.

In their methodology, the authors formalise the function $align(i, n)$ which assess if a Nile intent (i) and a natural language statement (n) share the same underlying meaning. So far, $align(\cdot, \cdot)$ has only been reliably implemented through manual expert annotation, which limits scalability and reproducibility.

To enable evaluation of an alignment function, Munson et al. (2025) have made available a corpus of over 250 thousand aligned intents. The methodological process to create this corpus includes a human evaluation in which 436 translation triplets of Nile expressions, one from American university's website, and two automatically generated by LLMs, were evaluated using a Mushra test protocol. The Mushra guide to annotation was the following: "Bad" for $[0:20]$, "Poor" for $]20:40]$, "Fair" for $]40:60]$, "Good" for $]60:80]$, and "Excellent" for $]80:100]$.

Here is an example of a typical intent translation: **Nile sentence**: "for group('students') set quota('download', '5000', 'mb/h'), quota('upload', '2000', 'mb/h')" and its translation in **English**: "Quotas for students are 5000 Megabyte per hr download and 2000 Megabyte per hour upload"

3.1.3 Scientific Text revision

Text revision is a writing assistance task that involves substantial modifications of an existing text to improve readability, style, and formality while preserving its original meaning (Du et al., 2022; Li et al., 2022). In the scientific domain, revision is a critical step in the writing process, ensuring clarity, coherence, and adherence to academic standards, as poor writing quality can contribute to paper rejection (Amano et al., 2023).

We focus on the paragraph-level scientific text revision task, where a paragraph and an accompanying instruction specifying the intended modification is provided as input, and the model is expected to produce a revised paragraph that aligns with the given instruction (Jourdan et al., 2025a).

Prior works have relied on human evaluation to assess the performance of text revision systems (Du et al., 2022; Raheja et al., 2023, 2024; Ito et al., 2020; Schick et al., 2023). While effective, such evaluations are resource-intensive and time-consuming, which limits their feasibility for large-scale or iterative use during model development, motivating the need for automatic evaluation metrics that can approximate human judgements.

We use the test split of the ParaReval dataset[2] (Jourdan et al., 2025b), a collection of human pairwise evaluations of automatically generated revisions. Each instance in the dataset consists of two versions of the same paragraph, extracted from computer science papers from OpenReview and authored by the original writers. Pairs of paragraphs are annotated with an instruction describing the underlying revision intention, an example is given in Appendix A.

Based on these annotations, each paragraph was automatically revised by six different models following the corresponding instruction. The resulting generated revisions were evaluated in pairs by human annotators to assess the reliability of automatic evaluation metrics

[2] https://github.com/JourdanL/parareval

for this task. Annotators answered three questions: (1) Did the model follow the instruction? (2) Is the revision correct, *i.e.*, better than the original? and (3) Which revision do they prefer between options A and B? The annotator's responses enable ranking the two options and identifying the preferred one.

The data comprises 258 pairs of revised paragraphs, each annotated with two distinct revision instructions, resulting in a total of 516 evaluation instances. Human pairwise preferences are balanced, with annotators favouring option A in 44% of cases, option B in 41%, and reporting ties in 15% of cases.

3.2 Usual metrics

In this section, we describe the automatic evaluation metrics considered in this study. For consistency, we include the same set of baseline metrics across all tasks, complemented by task-specific metrics where appropriate.

3.2.1 Baseline Metrics

For all tasks, we include a set of standard similarity metrics that have been widely used across text generation.

LEVENSHTEIN: Measures the edit distance between two character strings by counting the minimum number of deletions, insertions, and substitutions required to transform one into the other. A LEVENSHTEIN of 0 indicates an exact match. Because edit distance increases with string length, we normalise scores by the length of the longer string to ensure comparability across examples. To maintain consistency in interpretation with other metrics, where higher scores indicate greater similarity, we report $1-$LEVENSHTEIN.

BLEU (Papineni et al., 2002): Originally developed for evaluating machine translation, BLEU computes the precision of n-gram overlap between a candidate and a reference text, applying a brevity penalty to discourage overly short outputs.

METEOR (Banerjee and Lavie, 2005): A unigram-based metric that combines precision and recall using a harmonic mean. Unlike BLEU, METEOR allows for flexible unigram matching based on surface form, stemming, and synonymy, thus better capturing lexical variation.

BERTSCORE (Zhang et al., 2020): Evaluates semantic similarity between candidate and reference texts by comparing their contextual embeddings from a pretrained BERT model (Devlin et al., 2019). The underlying assumption is that if sentence embeddings are close in representation space, their meanings are semantically aligned.

MODERN BERTSCORE: Since BERTSCORE relies on the 2018 BERT model, its training data and methods are now outdated. To address this, Warner et al. (2025) introduced MODERN BERT, a modernised encoder-only Transformer trained with updated datasets and techniques while preserving the original masked language modelling objective. This model supports longer contexts and demonstrates stronger performance on multiple NLP benchmarks. We therefore use MODERN BERT as the embedding model for BERTSCORE, resulting in an updated variant (MODERN BERTSCORE) that serves as a SOTA baseline.

3.2.2 Metric Specific for Text Revision

For the text revision task, we additionally consider SARI and LLM-as-a-judge approaches that rely on model-generated textual outputs.

SARI (Xu et al., 2016) is an n-gram–based metric originally proposed for text simplification. It compares the system output against both the reference and source texts, rewarding appropriate additions, deletions, and word retention. SARI remains the most widely used automatic metric for evaluating text revision quality.

In addition, we reuse **LLM-as-a-judge** approaches from Jourdan et al. (2025b), where a LLM generates evaluative judgements based on the same questions posed to human annotators. The evaluation decisions are then extracted from the model's generated output We consider two variants: (i) **LLM-choice**, which answers binary Yes/No questions and conducts pairwise comparisons, and (ii) **LLM-Likert**, which assigns scores on a five-point scale. We experiment with LLAMA 3 70B model.

3.3 [*]-PLUIE

The original PARAPLUIE prompting template is structured as a dialogue between a *user* and an *assistant*, which explain the purpose of this task to the assistant. In line with prior works (Rios and Kavuluru, 2018; Brown et al., 2020; Chung et al., 2024), Lemesle et al. (2025) have shown that providing explicit examples of the task (i.e., few-shot prompting) can improve the model's ability to produce accurate and consistent judgements.

While PARAPLUIE has demonstrated strong performance for evaluating text revision using its original paraphrase detection prompt (Jourdan et al., 2025b), we investigate whether adapting the prompt to this specific task can further improve reliability and if it can be generalised to other tasks. We refer to the original paraphrase-oriented template as PARA-PLUIE.

To explore this hypothesis, we design personalised prompts for each of the studied tasks, leveraging the flexibility of PARAPLUIE as a plug-and-play metric that can easily be adapted through prompt modification alone. We construct the following variants:

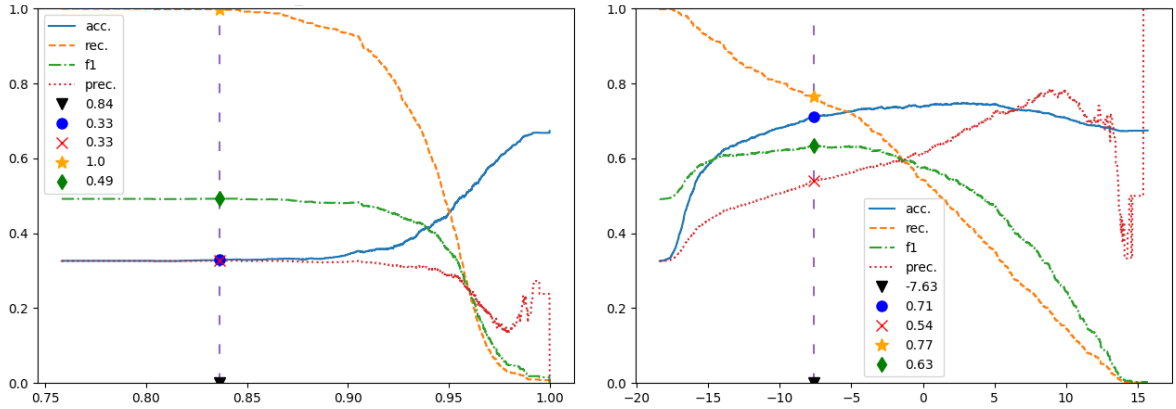
PARA-FRENCH-PLUIE: a French adaptation of the original paraphrase detection prompt (PARA-PLUIE), using translated few-shot examples to test cross-lingual generalisation.

NET-PLUIE: a task-specific prompt for assessing whether two sentences express the same network policy. The few-shot examples used in the prompt are drawn from the Mushra evaluation used in our experiments, with those instances excluded from the evaluation data to prevent bias.

REV-PLUIE: a revision-oriented prompt designed to assess whether a generated revision follows its associated instruction. This variant includes the gold reference as a one-shot example to ground the model's understanding of the task.

All prompts retain the same pairwise response options, "Yes" and "No," to ensure comparability across tasks and are available in Appendix B.

We compare these adapted versions against the original PARAPLUIE template to determine whether



(a) (b)

Figure 2: Score distribution of MODERN BERTSCORE (a) and PARA-FRENCH-PLUIE (b). The blue, orange, red and green curves denote respectively the accuracy, recall, precision and F1-scores according to the decision threshold. Emphasis is placed on the maximum F1-score obtained by the metric.

task-specific personalisation improves alignment with human judgements. Unless explicitly specified otherwise, all *-PLUIE experiments use PHI-4 as the underlying perplexity model (Abdin et al., 2024).

4 Results

We compare the proposed *-PLUIE approaches with standard semantic similarity metrics across our three identified tasks.

Because baseline metric scores are not directly interpretable as categorical decisions, we calibrate each metric by determining an optimal decision threshold. We uniformly sample 10,000 thresholds between the minimum and maximum observed scores, classifying each sentence pair as "Yes" or "No" depending on whether its score exceeds the threshold. Accuracy, precision, recall, and F1 are computed across thresholds, and the optimal threshold is defined as the one that maximises the F1-score.

When human evaluation data are available, we also compute correlations between automatic metric scores and human judgements to further assess reliability.

4.1 Paraphrase detection

Table 2 reports the performance of all evaluated metrics on the task of French paraphrase detection. For each metric, results correspond to the threshold that maximises the F1-score.

We observe that all traditional metrics achieve an accuracy of approximately 33%. Given that the dataset contains slightly more than 67% of non-paraphrase pairs, this implies that these metrics reach their best accuracy when classifying nearly all sentence pairs as non-paraphrases. This finding suggests two things: first, the dataset is particularly challenging; and second, these surface-based metrics struggle to capture large semantic differences when lexical overlap remains high.

By contrast, PARA-PLUIE defines a natural decision threshold at 0 by design. We therefore report results both for this default threshold and for the threshold that maximises F1-score. We find that PARA-PLUIE

Metric	Thr.	Acc.	Rec.	Prec.	F1
LEVENSHTEIN	0.71	0.33	**1.00**	0.33	0.49
BLEU	0.00	0.33	**1.00**	0.33	0.49
METEOR	0.43	0.33	**1.00**	0.33	0.49
BERTSCORE	0.80	0.33	**1.00**	0.33	0.49
MODERN BERTSCORE	0.84	0.33	**1.00**	0.33	0.49
PARA-PLUIE	-7.47	0.73	0.73	0.57	**0.64**
PARA-PLUIE	0	**0.75**	0.49	**0.66**	0.56
PARA-FRENCH-PLUIE	-7.63	0.71	0.77	0.54	0.63
PARA-FRENCH-PLUIE	0	0.74	0.54	0.61	0.58

Table 2: Threshold, accuracy, recall, precision and F1-score according to the best F1-score for each metric. For *-PLUIE one's results with not calibrated threshold (0) are also displayed. Emphasis is placed on maximal results.

achieves substantially higher accuracy and F1-scores than all baselines. Moreover, the difference between the calibrated and default thresholds is minimal, which highlights the interpretability and robustness of PARA-PLUIE's scoring scale.

Interestingly, using a French prompt does not appear to significantly affect classification performance. This observation suggests that large language models are relatively insensitive to the language used in either the input sentences or the evaluation prompt.

To better understand these results, Figure 2 compares the score distributions of MODERN BERTSCORE and PARA-FRENCH-PLUIE. As shown in Figure 2a, MODERN BERTSCORE assigns predominantly high similarity scores, and its accuracy curve doesn't decrease. This indicates that it fails to assign higher similarity scores to sentence pairs that convey the same meaning. In contrast, PARA-FRENCH-PLUIE (Figure 2b) successfully differentiates between paraphrases and non-paraphrases, assigning negative scores to the latter. However, the rapid decrease in recall reveals that the model also produces a number of false negatives pairs labelled as paraphrases but judged as non-paraphrases,

indicating room for improvement.

Overall, this experiment conducted on a particularly difficult dataset reinforces previous findings regarding PARA-PLUIE's strengths. In particular, it confirms that PARA-PLUIE is less influenced by surface lexical similarity and more capable of capturing genuine semantic differences.

4.2 Translation in IBN

For the experiment in this section, we only consider the Llama translations generated from the Mushra human evaluation phase of the NEAT methodology, the other results can be found in Appendix C.

We first examine whether NET-PLUIE can approximate a binary alignment decision and how its natural decision threshold compares with human judgements. Following the NEAT protocol, Mushra scores are binarised using an acceptability cutoff: translations rated as "Good" or "Excellent" (> 60) are considered positive, resulting in 60% positive and 40% negative examples. For each prompting variant (PARA-PLUIE and NET-PLUIE) and model, we vary the decision threshold and source model used and compute precision, recall, and F1, identifying the threshold that maximises the F1-score. This setup allows us to assess how closely the intrinsic decision boundary of NET-PLUIE aligns with the human-defined optimal threshold.

This means that we aim to resolve:

$$align(i,n) \equiv Mushra(i,n) > 60 \equiv Simil(i,n) > X$$

Here, $Simil$ denotes a semantic similarity metric, and X represents the threshold that needs to be optimised to best match the human experts scores.

Figure 3 presents precision, recall, accuracy, and F1-score across different threshold values for MODERN BERTSCORE and NET-PLUIE. Results for the other metrics are provided in Appendix D. For NET-PLUIE (Figure 3b), the precision increases with the threshold, reducing false positives. Accuracy and F1 initially increase for threshold values between approximately -20 and 0, then decline rapidly beyond 0. In contrast, MODERN BERTSCORE (Figure 3a) shows a slower and shorter initial rise in accuracy and F1, followed by a continuous decrease.

Model	Thr.	Acc.	Prec.	Rec.	F1
LEVENSHTEIN	0.17	0.65	0.64	0.95	0.77
BLEU	0.0	0.6	0.6	**1.0**	0.75
METEOR	0.0	0.6	0.6	**1.0**	0.75
BERTSCORE	0.39	0.67	0.64	**1.0**	0.78
MODERN BERTSCORE	0.68	0.63	0.63	0.96	0.76
PARA-PLUIE	-3.43	**0.81**	0.83	0.87	**0.85**
PARA-PLUIE	0	0.73	**0.88**	0.63	0.73
NET-PLUIE	-3.14	**0.81**	0.78	0.95	**0.85**
NET-PLUIE	0	**0.81**	0.84	0.85	**0.85**

Table 3: Threshold per metric according to the best F1 scores for classification of "good" Llama translation in respect to the Mushra evaluation.

Table 3 reports the optimal threshold for each metric and model-prompt variant, along with the corresponding accuracy, precision, recall, and F1-score. The PARAPLUIE variants consistently outperform the other metrics, achieving higher accuracy, precision, and F1 scores. The best performance is observed for NET-PLUIE using the Phi-4 model with an optimal threshold value of -3.14.

Notably, using the natural threshold of 0 yields comparable accuracy, F1, and a higher precision, despite a slightly lower recall. For this task, which involves network configurations, higher precision is preferable to avoid accepting incorrect intents. These results indicate a strong correlation between the NET-PLUIE scores and human judgement, supporting the equivalence $Mushra(i,n) > 60 \equiv \text{PARAPLUIE}(i,n) > 0$, using the default threshold.

4.2.1 Correlation to human evaluation

We assess NET-PLUIE as a proxy for computing $align(i,\ n)$ by calculating its Pearson correlation coefficient with the human evaluation scores. Strong positive correlations indicate that NET-PLUIE reliably reflects human judgements, supporting its use as an automated measure of semantic alignment. Correlations are computed across three translation sources (Llama, Mistral, and Reference) to evaluate consistency.

Model	Llama	Mistral	Reference
LEVENSHTEIN	0.26	0.09	0.17
BLEU	0.26	0.09	0.17
METEOR	0.08	0.03	0.28
BERTSCORE	0.26	0.18	0.36
MODERN BERTSCORE	0.32	0.17	0.37
PARA-PLUIE	0.67	0.32	0.49
NET-PLUIE	**0.69**	**0.35**	**0.52**

Table 4: Pearson's correlation coefficient between the metrics and human judgement on different sources of translation.

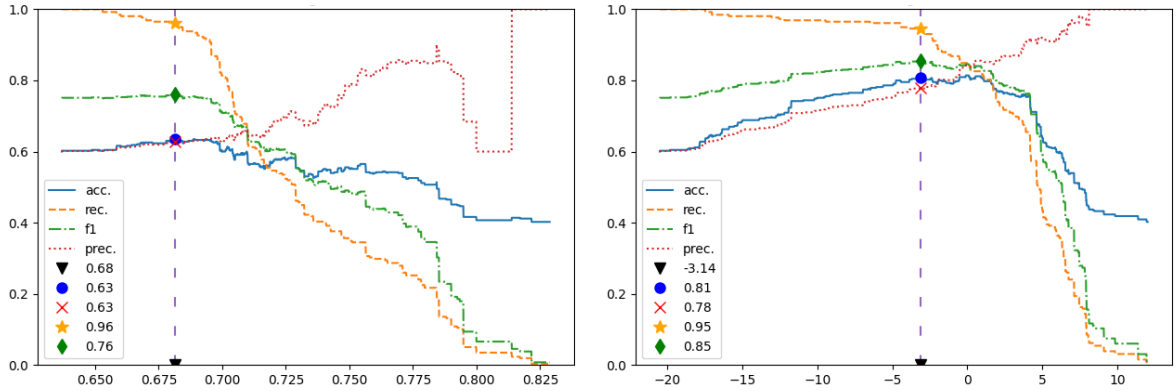
Table 4 shows that all metrics correlate positively with human judgement, with NET-PLUIE achieving substantially higher correlations. The strongest correlation observed on the Llama translations is (0.69), which is close to the threshold generally considered a strong correlation. Correlations are lower for the Reference and Mistral translations (0.52 and 0.35, respectively), which is expected given the difficulty of aligning translation quality with human judgements (Mondshine et al., 2025). Despite these lower values, NET-PLUIE maintains the strongest association with human evaluations, supporting its use as an automated alignment function for evaluating translation quality between Nile and natural language intents.

4.2.2 NET-PLUIE as an alignment function

To validate NET-PLUIE as a semantic alignment function, we require two conditions to be verified presented in Equations 4 and 5.

$$align(i,n_1) \wedge align(i,n_2) \implies \star\text{-PLUIE}(n_1,n_2) > T \wedge \star\text{-PLUIE}(n_2,n_1) > T \tag{4}$$

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



(a) (b)

Figure 3: Accuracy, recall, precision and F1-score score distribution over different threshold values for MODERN BERTSCORE (a) and NET-PLUIE (b). Emphasis is placed on the maximum F1-score obtained by the metric.

Therefore, if two sentences (n_1 and n_2) share the same meaning and one of them is aligned with an intent, the second one should also be align with this intent. This is made possible because the intent language is unambiguous, indeed this relation is not necessarily true with natural language.

We profit from the non-symmetrical property of *-PLUIE and compute the score NET-PLUIE(n_1,n_2) and NET-PLUIE(n_2,n_1). However, when used as the alignment measure between natural language and formal intent language, we only use it once with the formal intent as the reference. This is because natural language expressions usually capture more details than formal languages, these additional details captured in the expression should still be considered as part of a good translation. However, the metric penalises the suppression of information between two expressions, meaning that the loss of information that is not relevant to the actual network configurations when going from natural language to Nile is penalised by the metric.

We consider the Mistral and Llama generated translations as our two sources of natural language expressions and count how many times our two implication are not respected. Overall, of the 431 Nile-NL triplets, 7 do not verify condition 4 and 59 fail condition 5. This corresponds to approximately 15% of the triplets, examples of the translations that failed the tests can be found in Appendix E.

Despite only 15% of the triplets failing our imposed conditions, this makes the use of this metric uncertain for comfortable use as the semantic alignment function without further verification to be carried out on the generated translations before being accepted. However, NET-PLUIE is currently the closest solution we have to an accurate automatic semantic alignment function between natural language and a formal intent language, making it the best available solution.

4.3 Text revision

To assess the alignment between automatic metrics and human judgements, we report pairwise accuracy with tie calibration (Deutsch et al., 2023), Cramér's V (Cramér, 1946), and Cohen's κ (Cohen, 1960). The main results are presented in Table 5, while Table 6 provides a breakdown by model for all LLM-based approaches.

For PARA-PLUIE, we assess whether a generated revision constitutes a paraphrase of the original paragraph (PARA-PLUIE *og-ref*) or the gold reference paragraph (PARA-PLUIE *ref-gen*).

Judge	Pair acc.	V	κ
REV-PLUIE w. gold	**0.61**	**0.32**	**0.33**
REV-PLUIE	0.61	0.31	0.32
PARA-PLUIE og-gen	0.51	0.20	0.15
PARA-PLUIE ref-gen	0.54	0.23	0.20
LLM choice w. gold	0.60	0.30	0.33
LLM choice	0.59	0.28	0.30
LLM likert w. gold	0.54	0.27	0.27
LLM likert	0.54	0.28	0.28
MODERN BERTSCORE	0.36	0.13	-0.07
BERTSCORE	0.45	0.16	0.03
SARI	0.47	0.18	0.07
METEOR	0.42	0.19	0.00
BLEU	0.41	0.17	-0.03
LEVENSHTEIN og-gen	0.49	0.17	-0.08
LEVENSHTEIN ref-gen	0.44	0.13	0.01
Random weighted	*0.36*	*0.03*	*0.00*
Random uniform	*0.33*	*0.03*	*0.00*

Table 5: Alignment of automatic metrics with human judgements across all data. Pairwise accuracy and Cramér's V are defined on [0:1] and Cohen's Kappa on [-1:1]. For REV-PLUIE and PARA-PLUIE we consider PHI-4 scores and for LLM-choice and LLM-likert we consider GPT-4O scores.

4.3.1 Overall Alignment with Human Judgements

Table 6 details the results across metrics. Across all three measures, REV-PLUIE shows the highest alignment with human evaluations, achieving the best or second-best scores on every measure. This performance surpasses both other LLM-as-a-judge methods and traditional similarity metrics. LLM-Choice also performs competitively, while LLM-Likert and PARA-PLUIE achieve moderate yet consistent results.

Judge	Pair acc.	w.g.	V	w.g.	κ	w.g.
REV-PLUIE	**0.61**	**0.62**	**0.31**	**0.32**	**0.32**	**0.34**
PARA-PLUIE	0.52	0.55	0.20	0.23	0.17	0.21
LLM choice	0.59	0.60	0.28	0.30	0.30	0.31
LLM likert	0.44	0.50	0.29	0.27	0.19	0.23

Table 6: Alignment of LLM-based metrics with human judgements. Pairwise accuracy and Cramér's V are defined on [0:1] and Cohen's Kappa on [-1:1]. For PARA-PLUIE, the column without gold (w.g.) correspond to the PARA-PLUIE scores between the original and generated paragraphs and the column with gold correspond to the PARA-PLUIE scores between the reference and generated paragraphs. The perplexity model is LLAMA 3 70B.

Model	Approach	GPUs	Runtime
LLAMA 3 70B	LLM-choice	A100 x2	124 min
	LLM-choice Gold	A100 x2	127 min
	LLM-likert	A100 x2	141 min
	LLM-likert Gold	A100 x2	187 min
	PARA-PLUIE og-gen	A100 x3	**29 min**
	PARA-PLUIE ref-gen	A100 x3	**29 min**
	REV-PLUIE	A100 x3	33 min
	REV-PLUIE Gold	A100 x3	55 min
PHI-4	PARA-PLUIE og-gen	A100 x1	**08 min**
	PARA-PLUIE ref-gen	A100 x1	**08 min**
	REV-PLUIE	A100 x1	09 min
	REV-PLUIE Gold	A100 x1	13 min

Table 7: GPU usage of all LLM based approaches on the paragraph revision task

In contrast, n-gram and embedding-based metrics display considerably weaker correlations with human judgements. This confirms that metrics relying on surface or static embedding similarity fail to capture the nuanced semantic and stylistic improvements targeted in text revision. Lastly, for REV-PLUIE including the reference revision as a one-shot example do slightly improve the alignment judgement, making it the most reliable option to evaluate the task with or without reference.

4.3.2 Model-Equivalent Comparison

Table 6 details the results using the same model: LLAMA 3 70B. We observe that, REV-PLUIE consistently outperforms PARA-PLUIE, demonstrating that the method can be effectively tailored to the task through prompt design. This aligns with expectations, as the paraphrase formulation focuses more on semantic equivalence than on stylistic fidelity to the revision instruction.

Among the output-based evaluators, LLM-Choice remains the most reliable, while LLM-Likert shows greater variability across models. These results confirm that discrete decision formats (e.g. Yes/No or pairwise preference) are generally easier for models to handle than open-ended scoring. Moreover, REV-PLUIE surpasses LLM-Choice in all experimental configurations, suggesting its potential as a lightweight alternative. More results are available in Appendix F.

4.3.3 Computational Efficiency

We also compare the computational cost of all LLM-based metrics in Table 7. On average, PLUIE-based methods are nearly six times faster than output-based alternatives when using the same model, since they compute probabilities over limited answer tokens rather than generating long textual outputs. However, they may require more GPU memory for longer prompts or larger models. Overall, PLUIE methods offer a favourable trade-off between efficiency and alignment, making them an attractive option for scalable LLM-as-a-judge evaluation.

5 Conclusion

We introduced a generalised version of PARAPLUIE, extending the original perplexity-based *LLM-as-a-judge* approach to a broader range of semantic evaluation tasks, including paraphrase detection, translation alignment, and text revision. By adapting task-specific prompts, we showed that *-PLUIE consistently achieves stronger correlations with human judgements than both classical similarity metrics and generation-based LLM judges. Across all experiments, personalised *-PLUIE prompts yield the most reliable results while remaining up to six times faster than output-based approaches. Their interpretable confidence scores and stable decision thresholds make PLUIE-based methods a practical, transparent, and scalable alternative for automatic LLM evaluation. Overall, these results position *-PLUIE as an efficient and adaptable foundation for automatic evaluation in the LLM era.

In this paper, we only considered prompts that can be formulated as "Yes"/"No" questions. However, it would be interesting to extend this work to several tasks that require a different output, so far, as long as it fit on one token. Furthermore, *-PLUIE formula could be extended and evaluated without the one token limitation or more than two exclusives classes. Such tasks could include, sentiment analysis, topic classification or question answering.

Limitations

Our comparison with state-of-the-art *LLM-as-a-judge* methods has so far been conducted only for the text revision task. Although *-PLUIE is structurally less time-consuming, comparing PARA-PLUIE and NET-PLUIE with other *LLM-as-a-judge* methods would further strengthen confidence in the observed performance gains.

Moreover, the perplexity models used with *-PLUIE were not fine-tuned for these experiments. Fine-tuning could potentially enhance both accuracy and task-specific sensitivity.

Finally, while most experiments were carried out in English, with additional tests in French showing consistent results, the behaviour of the method in languages with richer morphology or markedly different syntactic structures remains to be investigated.

Ethical considerations

Data availability For all considered tasks, we use datasets that are openly available. The dataset from Tytgat et al. (2024) and Munson et al. (2025) on available on demand to the original authors. The paragraphs in ParaReval are extracted from scientific articles collected on OpenReview where they fall under different "non-exclusive, perpetual, and royalty-free license".

Computational resources Using LLMs remain resource intensive, however using an LLM of medium size like PHI-4 seems to approach results of larger LLMs in respect with Table 6 and is obviously less computing intensive. Lastly, *-PLUIE computational cost is way lower than other LLM-as-judge methods, as highlighted by Table 7.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Tatsuya Amano, Valeria Ramírez-Castañeda, Violeta Berdejo-Espinola, Israel Borokini, Shawan Chowdhury, Marina Golivets, Juan David González-Trujillo, Flavia Montaño-Centellas, Kumar Paudel, Rachel Louise White, and 1 others. 2023. The manifold costs of being a non-native english speaker in science. *PLoS Biology*, 21(7):e3002184.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

H. Cramér. 1946. *Mathematical Methods of Statistics*. Goldstine Printed Materials. Princeton University Press.

Daniel Deutsch, George Foster, and Markus Freitag. 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*, pages 12914–12929.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ehsan Doostmohammadi, Oskar Holmström, and Marco Kuhlmann. 2024. How reliable are automatic evaluation methods for instruction-tuned LLMs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6321–6336, Miami, Florida, USA. Association for Computational Linguistics.

Wanyu Du, Zae Myung Kim, Vipul Runderstandaheja, Dhruv Kumar, and Dongyeop Kang. 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)*, pages 96–108, Dublin, Ireland. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Takumi Ito, Tatsuki Kuribayashi, Masatoshi Hidaka, Jun Suzuki, and Kentaro Inui. 2020. Langsmith: An interactive academic text revision system. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 216–226, Online. Association for Computational Linguistics.

Arthur Selle Jacobs, Ricardo José Pfitscher, Ronaldo Alves Ferreira, and Lisandro Zambenedetti Granville. 2018. Refining network intents for self-driving networks. In *Proceedings of the Afternoon Workshop on Self-Driving Networks*, SelfDN 2018, page 15–21, New York, NY, USA. ACM.

Léane Jourdan, Florian Boudin, Richard Dufour, Nicolas Hernandez, and Akiko Aizawa. 2025a. ParaRev : Building a dataset for scientific paragraph revision annotated with revision instruction. In *Proceedings of the First Workshop on Writing Aids at the Crossroads of AI, Cognitive Science and NLP (WRAICOGS 2025)*, pages 35–44, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Leane Jourdan, Nicolas Hernandez, Florian Boudin, and Richard Dufour. 2025b. Identifying reliable evaluation metrics for scientific text revision. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6731–6756, Vienna, Austria. Association for Computational Linguistics.

Aris Leivadeas and Matthias Falkner. 2023. A survey on intent-based networking. *Commun. Surveys Tuts.*, 25(1):625–655.

Quentin Lemesle, Jonathan Chevelu, Philippe Martin, Damien Lolive, Arnaud Delhay, and Nelly Barbot. 2025. Paraphrase generation evaluation powered by an LLM: A semantic metric, not a lexical one. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8057–8087, Abu Dhabi, UAE. Association for Computational Linguistics.

Jingjing Li, Zichao Li, Tao Ge, Irwin King, and Michael Lyu. 2022. Text revision by on-the-fly representation optimization. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pages 58–59, Dublin, Ireland. Association for Computational Linguistics.

Joseph McNamara, Erik Aumayr, Liam Fallon, and Enda Fallon. 2022. A flexible interpreter for intent realisation. In *NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium*, pages 1–6.

Itai Mondshine, Tzuf Paz-Argaman, and Reut Tsarfaty. 2025. Beyond n-grams: Rethinking evaluation metrics and strategies for multilingual abstractive summarization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19019–19035, Vienna, Austria. Association for Computational Linguistics.

Daisy Munson, Pierre Alain, and Guillaume Doyen. 2025. Neat: A nile-english aligned translation corpus based on a robust methodology for intent based networking. *Computer Networks*, 271:111519.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Vipul Raheja, Dimitris Alikaniotis, Vivek Kulkarni, Bashar Alhafni, and Dhruv Kumar. 2024. mEdIT: Multilingual text editing via instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 979–1001, Mexico City, Mexico. Association for Computational Linguistics.

Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. 2023. CoEdIT: Text editing by task-specific instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5274–5291, Singapore. Association for Computational Linguistics.

Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3132–3142, Brussels, Belgium. Association for Computational Linguistics.

Timo Schick, Jane A. Yu, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. 2023. PEER: A collaborative language model. In *The Eleventh International Conference on Learning Representations*.

Julie Tytgat, Guillaume Wisniewski, and Adrien Betrancourt. 2024. Évaluation de la similarité textuelle : Entre sémantique et surface dans les représentations neuronales. In *Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position*, pages 85–96, Toulouse, France. ATALA and AFPC.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Example of data from ParaReval dataset

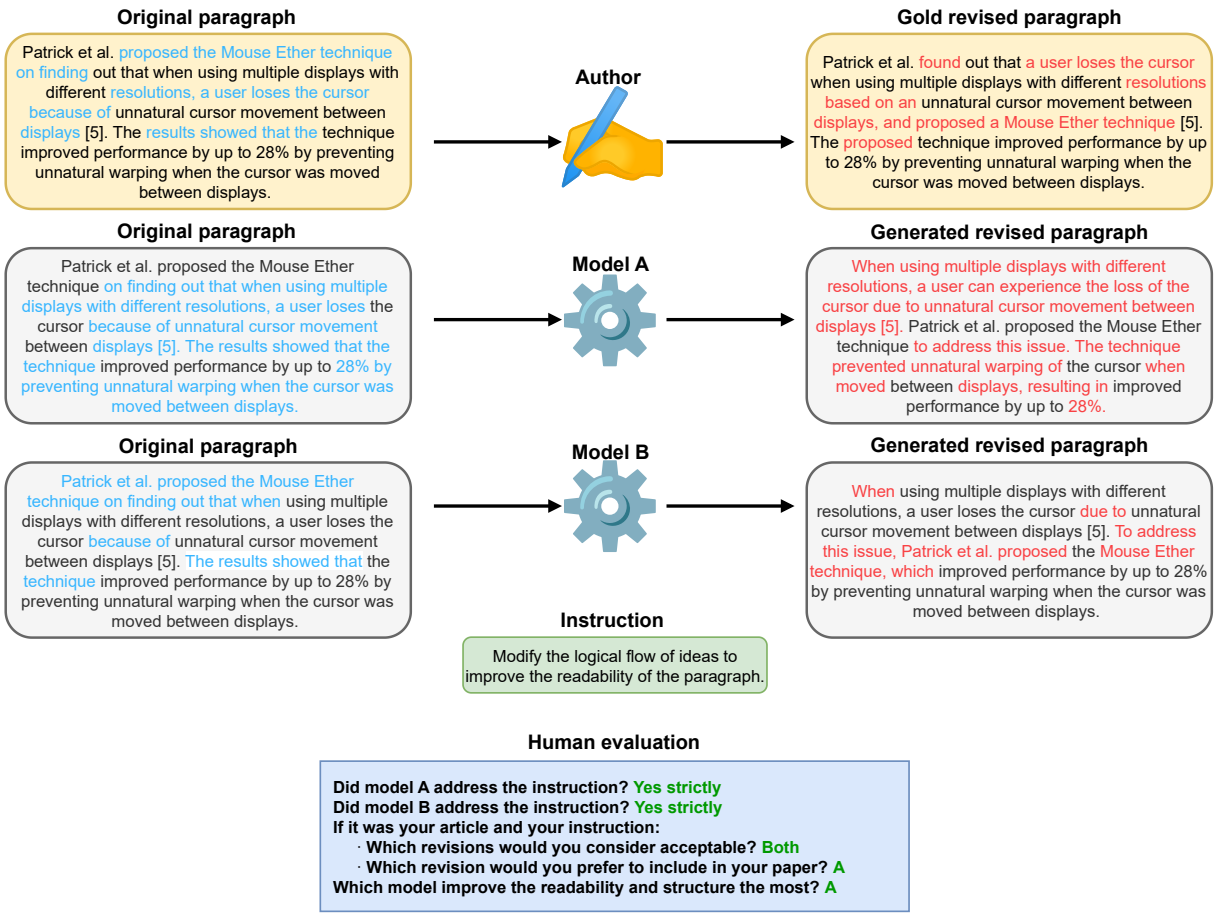


Figure 4: Example of data in the ParaReval dataset

B [*]-PLUIE new task specific templates

Prompt segment 1: PARA-FRENCH-PLUIE new template for french paraphrase detection

```
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}\"""
```

Prompt segment 2: NET-PLUIE new template for network policy

```
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}\""""
```

Prompt segment 3: REV-PLUIE new template for instruction following

```
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
 addressed the instruction I? Answer with only one word \"Yes\" or \"No\"."""

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

<Gold reference can be added here as one-shot>
user="""P1: \"{reference}\"; I: \"{instruction}\"; P2: \"{gold}\""""

assistant= """Yes"""
<end of gold reference inclusion>


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

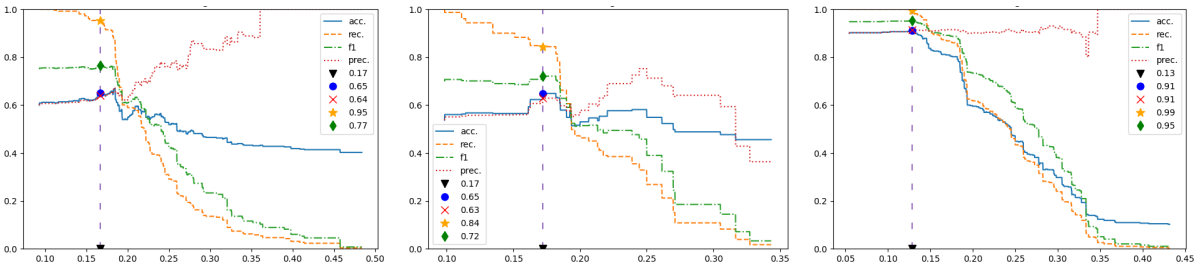
C Mistral and Reference test results

model	syst	thr.	f1	prec.	rec.
levenshtein	reference	0.17	0.72	0.63	0.84
bleu	reference	0.0	0.7	0.54	**1.0**
meteor	reference	0.02	0.74	0.65	0.85
bertscore	reference	0.43	0.73	0.66	0.82
modern_bertscore	reference	0.65	0.7	0.54	**1.0**
NET-PLUIE	reference	-1.91	**0.78**	0.67	0.93
NET-PLUIE	reference	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 8: Metrics evaluation for the Reference translations and the Mistral generated translations

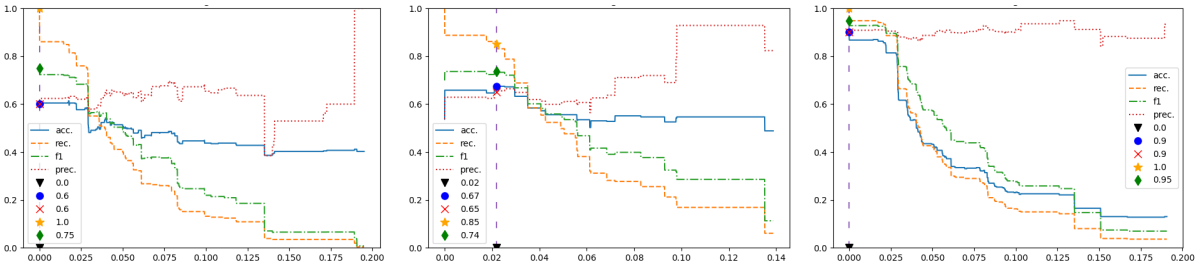
D Classification with PARA-PLUIE using different models

Figures 5, 6, 7, 8, 9, 10, 11 show the different classification results with all the combination of metrics.



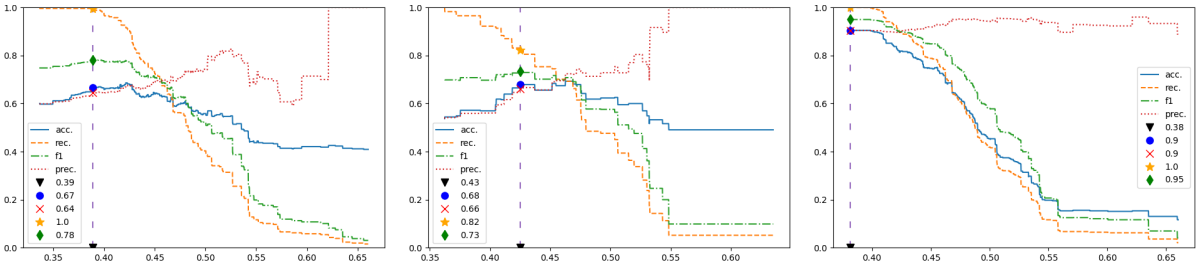
(a) Dataset llama (b) Dataset reference (c) Dataset mistral

Figure 5: Classification with metric levenshtein.



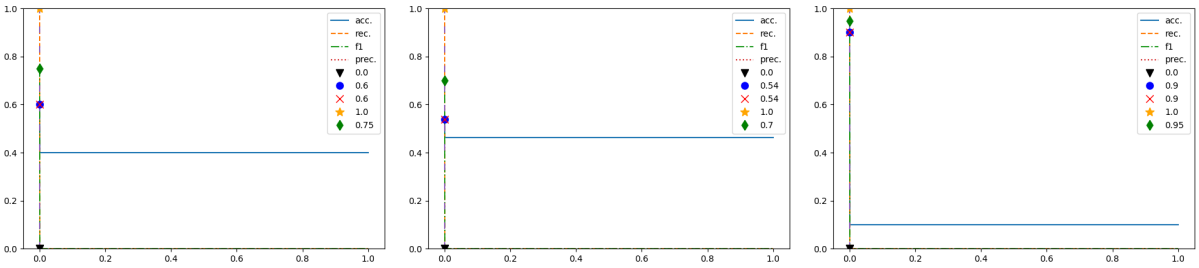
(a) Dataset llama (b) Dataset reference (c) Dataset mistral

Figure 6: Classification with metric meteor.



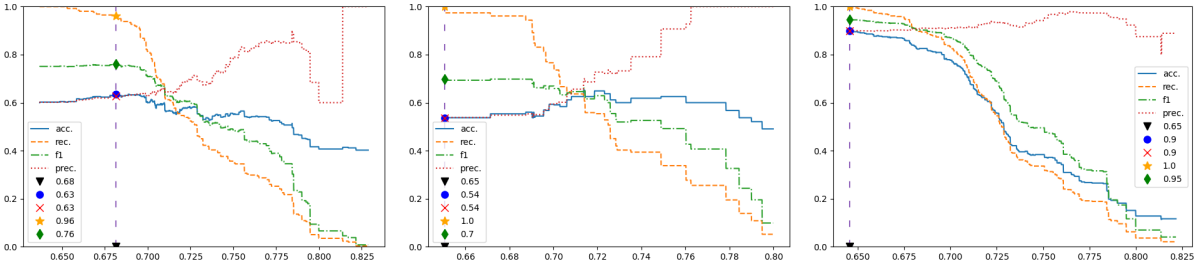
(a) Dataset llama (b) Dataset reference (c) Dataset mistral

Figure 7: PARAPLUIE with model bertscore.



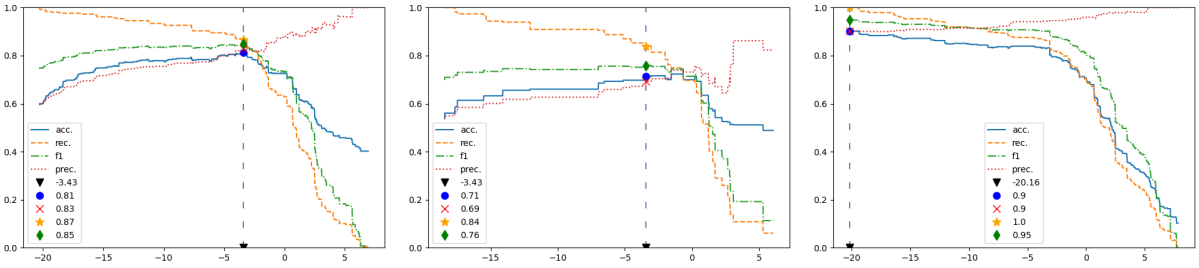
(a) Dataset llama (b) Dataset reference (c) Dataset mistral

Figure 8: PARAPLUIE with model bleu.



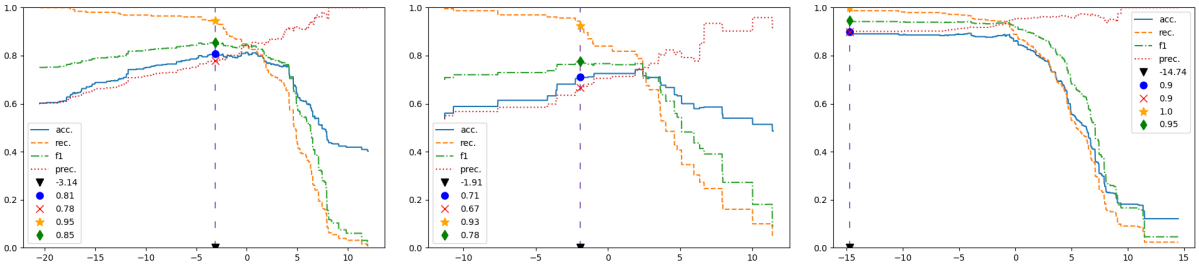
(a) Dataset llama (b) Dataset reference (c) Dataset mistral

Figure 9: Classification with metric modern_bertscore.



(a) Dataset llama (b) Dataset reference (c) Dataset mistral

Figure 10: Classification with metric PARAPLUIE with model phi-4_FS-DIRECT.



(a) Dataset llama (b) Dataset reference (c) Dataset mistral

Figure 11: Classification with metric PARAPLUIE with model phi-4_NETWORK.

E Non-respected implications triplets

Table 9, resp. Table 10, illustrates triplets the does not respect implication 4, resp. implication 5.

Nile sentence	Translation with Llama	Translation with Mistral
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 9: A few example of bad translations that does not respect implication 4, both translation are aligned to the same intent expression, but PARAPLUIE does not consider the translation to be paraphrases.

Nile sentence	Translation with Llama	Translation with Mistral
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 10: Bad translations that does not respect implication 5: One translation is aligned to the Nile expression, both translation are paraphrases, but, on lines 1 to 4 Parapluie does not consider Llama's translation to be aligned to the Nile expression, and on lines 5 to 7 Parapluie does not consider Mistral's translation to be aligned to the Nile expression.

F Additional results for alignment of LLM-based metrics on text revision.

	Judge	Pair acc.		V		κ	
			w.g.		w.g.		w.g.
REV-PLUIE	LLAMA 3 70B	**0.61**	**0.62**	**0.31**	**0.32**	**0.32**	**0.34**
	PHI-4	**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	0.55	0.20	0.23	0.17	0.21
	PHI-4	0.51	0.54	0.20	0.23	0.15	0.20
	LLAMA 3 8B	0.52	0.53	0.21	0.22	0.17	0.19
	MISTRAL 7B	0.54	0.53	0.22	0.22	0.20	0.19
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
	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
	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 11: Alignment of LLM-based metrics with human judgements. Pairwise accuracy and Cramér's V are defined on [0:1] and Cohen's Kappa on [-1:1]. For PARA-PLUIE, the column without gold (w.g.) correspond to the PARA-PLUIE scores between the original and generated paragraphs and the column with gold correspond to the PARA-PLUIE scores between the reference and generated paragraphs.