Paraphrase Generation Evaluation Powered by LLM: A Semantic Metric, Not a Lexical One

Anonymous ACL submission

Abstract

Evaluating automatic paraphrase production systems is a difficult task as it involves, among other things, assessing the semantic proximity between two sentences. Usual measures are based on lexical distances, or at least on semantic embedding alignments. The rise of large language models has provided tools to model relationships within a text thanks to the attention mechanism. In this article, we introduce ParaPLUIE, a new measure based on a log likelihood ratio from a LLM, to assess the quality of a potential paraphrase. This measure is compared with usual measures on three datasets of manually labeled paraphrases and non-paraphrases pairs. Two datasets, posterior to this study, are known for their quality or difficulty on this task. The third one, build for this occasion, is composed of LLM outputs. According to evaluations, the proposed measure is better for sorting pairs of sentences by semantic proximity. In particular, it is much more independent to lexical distance and offer a easy classification threshold between paraphrases and non-paraphrases.

1 Introduction

In the field of automatic generation of paraphrases, plenty of definitions of paraphrases have been proposed (Mel'čuk, 1997; Barzilay and McKeown, 2001; Sekine, 2005; Zhao et al., 2009; Fabre et al., 2021). All those definitions point the importance of meaning conservation, that is inherently an ambiguous concept.

Despite this, paraphrase generation systems need semantic measures to be trained or evaluated. Usually, metrics work either with lexical matching (Papineni et al., 2002) or embedding matching (Zhang et al., 2020). By design, lexical matching approach struggles to reconcile simple transformations like synonym replacement (Banerjee and Lavie, 2005). Moreover, they have difficulties to reject sentences with an opposed meaning if they are lexically close. On the other hand, metrics that use semantic embedding matching, are laid on sub-phrasal alignments without taking into account a global view of sentences. These two points have been highlighted by Zhang et al. (2019) and lead to the construction of PAWS dataset.

The TRANSFORMER architecture (Vaswani et al., 2017) and the emergence of Large Language Models have led to many advances in the area of natural language processing. Specifically, the self-attention mechanism, can capture semantic relations in a large context. Chen et al. (2023) have demonstrated that a LLM is capable of scoring the quality of reference-free sentences. We offer to explore the development of a new semantic metric for paraphrase classification, ParaPLUIE, based on a LLM and its output perplexity.

The paper is organized as follows. First, we sum up metrics usually used to classify paraphrases in section 2. The reference evaluation datasets are then described in section 3, including a new dataset of human labeled paraphrases. In section 4, we propose a novel automatic metric dedicated to semantic proximity, ParaPLUIE. State of the art metrics are evaluated together with ParaPLUIE in section 5. It should be noted that, despite their variety, automatic metric scores seem correlated to the edit distance, which is not the case for ParaPLUIE.

2 Automatic metrics

Metrics usually used to evaluate meaning conservation between two sentences, can be split into two main groups. One that involves metrics measuring how much the lexical structure is similar between two sentences, thanks to a lexical distance. Another that involves metrics estimating the semantic proximity between two sentences, thanks to embedding matching.

In the first group, we can include the Levenshtein distance (LEV.) (Levenshtein, 1965), METEOR

(Banerjee and Lavie, 2005) and BLEU (Papineni et al., 2002).

LEV. gives a measurement of differences between two character strings. This metric is counting the number of minimal deletions, insertions and replacements of characters, needed to transform one string into another. As the considered strings are getting longer, the LEV. distance increases, then it is usually normalised by the longest string length.

BLEU has been designed to measure translation quality. It consists in computing the n-gram overlap between a candidate and reference sentence as well as a brevity penalty. Usually, n-grams up to 4 words long are considered. In this paper, we use the *Torchtext*[1] implementation of BLEU with the default settings.

METEOR echoes the design of BLEU, by computing an harmonic mean of the uni-gram precision and recall between the hypothesis and the source. Moreover, METEOR considers a synonym matching to compute its score. METEOR has shown a better correlation with human judgement than BLEU.

One might argue that, if two sentences have a close lexical structure, they are more likely to be paraphrases. This is why, even if it seems not adequate, lexical metrics can be used to assess if two sentences share a common meaning. The weakness of this argument is that, even if two sentences share a common structure, they can convey a different meaning as with these two sentences:
"The cat is alive" and *"The cat was alive"*.

To address this issue, a research effort has been made to create another group of metrics. Those metrics rely on semantic distances and use token embeddings to symbolize words inside a LLM. In this second group, we can include BERT$_{\text{score}}$ (Zhang et al., 2020) and ParaScore (Shen et al., 2022).

BERT$_{\text{score}}$ is a score of similarity between each token embeddings of a hypothesis and of a source. Its definition is based on the following assumption: if a pairing between two sentences exists such as, all embeddings that form them are close, then their meaning is close. In the experiments, we use the BERT base uncased model (Devlin et al., 2019) from *Hugging Face*[2].

Shen et al. (2022) points out that, while lexical distances between two sentences increase, the performance of metrics decreases. To deal with this issue, they propose ParaScore, a metric that extends BERT$_{\text{score}}$ by including the normalized LEV. distance to determine a similarity score.

It is important to note that semantic similarity metrics take into account a word to word matching, without considering higher level semantic relations. This carries a risk concerning the quality of classification of paraphrases.

3 Datasets

Evaluating automatic metrics on sentence to sentence semantic proximity involves the use of datasets of labeled pair sentences as paraphrases/not paraphrases. Optimally, for assessing the relevance of metrics in challenging cases, labeled pairs of non-paraphrases should be lexically or semantically close (without being considered as paraphrases by a human). Our choice thus settled on two English corpora, PAWS (Zhang et al., 2019), designed to fool lexical metrics, and MRPC (Dolan and Brockett, 2005) including examples of semantic inference (but asymmetric).

The MRPC (MS-SSLA licence) dataset used in this paper is available on HuggingFace[3]. It contains 5,801 couples where 3,900 has been labeled as paraphrases, representing 67% of the dataset. This dataset has been generated automatically from a large corpus of newspapers organized by themes. During the labeling, the procedure was the following: for each couple of sentences, two evaluators have been asked if the pair can be considered as semantically equivalent. They were constrained to answer only by yes or no. In case of disagreement between them, a third one answers with the same guideline. This dataset is mostly composed of entailments. Here is a characteristic example of non-paraphrase entailment from MRPC: "*Last year, Bush appointed him to the Homeland Security Advisory Council.*" and "*He has also served on the president's Homeland Security Advisory Council.*".

For PAWS (free use licence), we are using the *dev* subset. This one is composed of 8,000 couples from which 3,539 are paraphrases. They represent 44% of the dataset. The whole dataset counts 108,463 couples and has been generated in a semi-automatic manner by word swapping and reverse

[1] https://pytorch.org/text/stable/data_metrics.html

[2] https://huggingface.co/spaces/evaluate-metric/bertscore

[3] https://huggingface.co/docs/datasets/v1.13.0/about_dataset_features.html?highlight=mrpc

translation. For each generated couple, 5 humans have labeled the couple as paraphrases or non-paraphrases. PAWS has been designed to be a challenge for automatic paraphrase classification systems. Indeed, generating sentences by word swapping often creates non-paraphrases, while maintaining a close lexical distance with the source sentence. Here is a typical example of non-paraphrase couple from PAWS: "*flights from New York to Florida*" and "*flights from Florida to New York*".

3.1 New LLM generated paraphrase dataset

The purpose of this new dataset is to have sentence pairs with a significant lexical distance. To do so, we use MISTRAL (Jiang et al., 2023) and LLAMA2 (Touvron et al., 2023) to generate paraphrases. Models have not been fine-tuned to generate paraphrases. Source sentences are randomly picked up from PAWS and MRPC sets. Two prompt templates are used for MISTRAL and one for LLAMA2. Moreover, to create diversity, and be more likely to generate non-paraphrases, a vulgar template has been designed. As LLAMA2 refuses to generate with this template, it was only used with MISTRAL. Hypothesis paraphrases generated with this template contain a wider range of vocabulary.

We have generated 605 hypothetical paraphrases with LLM from 605 source sentences. Each hypothesis paraphrase has been classified by at least one human judge. The evaluators were volunteers, non-experts in NLP domain. The evaluation protocol was as follows. Each judge was proposed to label up to 55 couples on a web-application (Fayet et al., 2020) in which 5 couples were reserved for the training trial. The training trial was the same for all judges.

Sentence pairs have been shown one by one, one sentence above the other. Presentation order of the sentence pair is chosen randomly. For each pair, judges had 5 possible answers: [Very different, Slightly similar, Mostly similar, Same meaning, Don't know], presented in this order. Evaluation guidelines with examples were also provided.

At the end of the evaluations, 276 couples have been labeled as "Same meaning", 181 as "Mostly similar", 93 "Slightly similar", 28 "Very different" and 22 "Don't know". We consider couples labeled as "Very different" and "Slightly similar" as non-paraphrases. "Mostly similar" and "Same meaning" labeled couples are considered as paraphrases. It is interesting to note that LLMs seem capable to generate paraphrases, and most of the times very good paraphrases.

Here is an example of a non-paraphrase in our dataset: "*Trading volume was incredibly light at 500.22 million shares, below an already thin 611.45 million exchanged at the same point Thursday.*" and "*The trading volume was significantly lower than usual on this day, with only 500.22 million shares exchanged compared to 611.45 million shares traded at the same time the previous day.*"

Overall, 457 couples have been labeled as paraphrases, 121 as non-paraphrases, 22 as indeterminate and 5 have been used for the training trial and were not taken in account. We did not include couples labeled as indeterminate in the dataset. In the end, the dataset is composed of 79% of paraphrases.

Details for reproducibility – prompts used and evaluation guidelines – are provided in appendix A.1 and A.2. The annotated corpus is provided as supplementary material.

4 ParaPLUIE

Usual metrics are focused on lexical proximity, or at best on token embedding alignments. As a result, their capacity to catch complex relations between sentences is limited. Recently, advancement with the TRANSFORMER architecture, thanks to the self-attention mechanism (Vaswani et al., 2017), has demonstrated that, it is possible to more effectively consider the internal relationships within a text.

LLMs are intended to model the probabilities associated to a token, knowing the previous ones. It is thus possible to compare two similar sequences to calculate a class belonging degree while considering intricate and subtle relations inside sentences.

We propose ParaPLUIE, a novel semantic proximity metric, relying on a learnt probabilistic model of a LLM. ParaPLUIE is defined as the log likelihood ratio of "*yes*" versus "*no*" knowing a template (Tpl, see section 4.1) filled with the source (S) sentence to paraphrase and the evaluated hypothesis (H), i.e:

$$\text{ParaPLUIE}(S, H) = \log\left(\frac{p\left(\text{yes}|\text{Tpl}(S, H)\right)}{p\left(\text{no}|\text{Tpl}(S, H)\right)}\right)$$

The intuition beside ParaPLUIE comes from the fact that, if LLM are capable to criticize sentences while generating, their surprise on the appearance of a token can be used as a metric. A positive score is given to a couple of sentences if the system estimates that they are likely to be paraphrases. On the opposite, the system gives a negative score when it estimates that they are not paraphrases. This propriety helps the interpretation of results unlike other scoring metrics because it creates a natural threshold decision at zero. Score is a real value whose range depends on the learnt probabilistic model used.

4.1 Templates

A template is an incomplete prompt that is then filled with sentences to evaluate. In the following, **S** is the source sentence and **H** is a candidate paraphrase of **S**. The template mimics a dialog with a user and an assistant because the model used in these experiments is a fine-tuned LLM, learnt to work as a conversational agent. We considered three different templates in our experiments.

4.1.1 Template: DIRECT

This naive template directly explains the task intended by the model and the expected output format.

$\mathbf{Tpl}_{\text{Direct}}(\mathbf{S}, \mathbf{H})$:

$(user)$: *You will receive two sentences A and B. Do these two sentences mean the same thing? Answer with only one word "yes" or "no".*

$(assistant)$: *Please provide the sentences for me to evaluate.*

$(user)$: *A: "**S**"; B: "**H**"*

4.1.2 Template: INDIRECT

(Qiao et al., 2023) points out that using a chain of thoughts may help the LLM to answer correctly. In other words, letting a LLM generate context or explanations about a question makes it more likely to be right in its answer. Inspired by this, a template involving a generation step, denoted **E**, has been created. First, the model generates its answer. Then it is requested to summarize it using only one word.

$\mathbf{Tpl}_{\text{Indirect}}(\mathbf{S}, \mathbf{H})$:

$(user)$: *You will receive two sentences A and B. Do these two sentences mean the same thing?*

$(assistant)$: *Please provide the sentences for me to evaluate.*

$(user)$: *A: "**S**"; B: "**H**"*

$Generation \longrightarrow E$

$(assistant)$: **E**

$(user)$: *Summarize your answer with only one word "yes" or "no".*

4.1.3 Template: FS-DIRECT

Numerous studies have shown that a few-shots approach helps LLMs to give an accurate answer (Rios and Kavuluru, 2018; Brown et al., 2020; Chung et al., 2024). We have used an improved version of the DIRECT template which contains few examples of the task resolution. The examples were generated using a LLM and labeled by three experts. We have intentionally picked examples where ParaPLUIE with the DIRECT template made scoring errors. More precisely, we have picked examples for whom the associated score was likely to classify them as paraphrase, while they are non-paraphrases and reciprocally. We have also picked some examples where the model was right with it prediction. The complete template is available in A.3.

4.2 Practical computation

To compute the prediction score with ParaPLUIE, we evaluate the ratio between the probability that the template is followed by the token "yes" and the probability that the template is followed by the token "no". As the templates differ by only one token ("yes" or "no"), we can reformulate the equation using perplexities. This is convenient as the perplexity reflects the "surprise" of the model for the prediction of a token.

$$\begin{aligned}\text{ParaPLUIE}(S, H) &= \log\left(\frac{p\,(\text{yes}|\text{Tpl}(S,H))}{p\,(\text{no}|\text{Tpl}(S,H))}\right)\\ &= \log\left(\frac{\mathit{ppl}\,(\text{Tpl}(S,H)\circ\text{no})^{T+1}}{\mathit{ppl}\,(\text{Tpl}(S,H)\circ\text{yes})^{T+1}}\right)\end{aligned}$$

where T is the number of tokens that made up the template and "$\circ$" a text concatenation operator.

Moreover, as LLMs are trained by using the perplexity as a loss function, we can use it directly. Then, the metric equation becomes:

$$\begin{aligned}&\text{ParaPLUIE}(S, H) = (T+1)\times\\ &(loss_{LLM}(\text{Tpl}(S,H)\circ\text{no}) - loss_{LLM}(\text{Tpl}(S,H)\circ\text{yes}))\end{aligned}$$

In our experiments, we use the MISTRAL *7B Instruct v0-2*[4] version of MISTRAL, in half-precision configuration. This model is a medium size language with 7 billion parameters. It is based on the TRANSFORMER architecture and uses a sliding attention window to reduce computing costs. The dataset used for its training is not disclosed. With

[4] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

	MRPC		PAWS		LLM		Global	
	yes	no	yes	no	yes	no	yes	no
Lev. ↓	0.38±0.16	0.51±0.13	0.20±0.15	0.32±0.15	0.49±0.17	0.55±0.16	0.31±0.19	0.38±0.17
Bleu ↑	0.40±0.21	0.28±0.18	0.62±0.18	0.49±0.19	0.24±0.20	0.18±0.18	0.49±0.23	0.42±0.21
Meteor ↑	0.69±0.14	0.56±0.15	0.91±0.06	0.88±0.07	0.64±0.18	0.54±0.20	0.79±0.16	0.78±0.18
BERT$_{\text{score}}$ ↑	0.82±0.07	0.74±0.08	0.94±0.04	0.91±0.04	0.82±0.08	0.76±0.11	0.87±0.08	0.86±0.10
ParaScore ↑	0.83±0.07	0.76±0.09	0.92±0.03	0.92±0.04	0.82±0.08	0.76±0.10	0.87±0.08	0.87±0.09
ParaPLUIE DIRECT ↑	20.02±8.94	4.41±15.43	22.04±6.65	12.80±13.46	23.84±5.27	16.44±13.81	21.15±7.89	10.41±14.60
ParaPLUIE INDIRECT ↑	14.71±12.79	−2.61±14.88	18.33±9.59	6.96±16.09	19.07±8.07	10.91±14.84	16.58±11.37	4.22±16.33
ParaPLUIE FS-DIRECT ↑	5.00±7.92	−4.89±8.30	9.10±7.36	−3.82±10.38	10.16±7.02	4.05±11.19	7.14±7.91	−3.99±9.91

Table 1: Average scores and standard deviation of each measure on MRPC, PAWS and the LLM paraphrases corpus. Datasets have been split according to the hypothesis sentence label: yes, it is a paraphrase or no. The ↑ associated to a metric indicates that, the higher the score, the closer the sentences. The ↓ sign means the opposite.

this configuration, the model needs approximately 15 GB of memory. We have conducted our experiments on a computer equipped with a Nvidia RTX 4090 GPU. The ParaPLUIE code is released with the supplementary material.

5 Results

In our evaluation scenario, we are given a source sentence (S) and the associated candidate paraphrase (H). (H) label as paraphrase or non-paraphrase given by human evaluators is considered as the gold label. As the goal is to evaluate (H) as a paraphrase candidate for (S), our evaluation takes place in a reference-free context.

5.1 Scores range

For each dataset and metric, the score of each sentence is computed and compared to human annotations. Table 1 presents the mean scores obtained for the different metrics on the datasets. Results are divided into subsets such as all paraphrase pairs, denoted as "yes" and non-paraphrase pairs, denoted as "no".

We can observe that the mean edit distance i.e Lev. distance, is the lowest on PAWS dataset, medium on MRPC and the highest on the LLM generated. This offers us a large overview of different paraphrases/non-paraphrases. We can point out that, mean scores of every metrics, excluding ParaPLUIE, strongly overlap. This can be explained by the deliberately misleading nature of the corpora considered in this experiment. We can observe that, ParaPLUIE mean scores on paraphrase pairs overlap less on non-paraphrase scores. Moreover, the mean scores of non-paraphrases is always lower than paraphrases' ones. In addition, we can view that the global mean score of ParaPLUIE FS-DIRECT on non-paraphrases is negative. Overall, this is not true for the other templates. It may confirm that giving examples to the LLM helps it solve complex tasks such as paraphrase detection.

5.2 Metrics accuracy

Let us now look at the best threshold score to classify paraphrases/non-paraphrases with the best accuracy for each metric. By looking at the results, presented in table 2, for metrics other than ParaPLUIE, we can notice that a good threshold for a corpus is not applicable on another one. Moreover, we can notice that, as explained by Zhang et al. (2020), BERT$_{\text{score}}$ is not well performing on PAWS corpus. Despite that, BERT$_{\text{score}}$ seems to be a good metric for this task. We can also observe that choosing a common threshold to all the dataset for a given metric significantly reduces its performance. This indicates two things, metrics struggle to correctly classify paraphrases and they are not resilient. Surprisingly, we can notice that the Lev. distance looks like to be the best metrics. Obviously Lev. is not a good metric for semantic evaluation. Consequently, this may suggest that other metrics are correlated with the edit distance.

By looking at the results of the different ParaPLUIE templates, we can see that their accuracy is significantly higher than others metrics on all datasets, except for the LLM corpus. Their

		MRPC	PAWS	LLM	Global
LEV.	Max. acc.	0.69	0.70	0.79	0.58
	Threshold	0.52	0.12	0.87	0.40
	F1	0.78	0.55	0.88	0.65
	Recall	0.81	0.41	**1.00**	0.71
	Precision	0.75	0.84	0.79	0.60
BLEU	Max. acc.	0.67	0.67	0.79	0.57
	Threshold	0.00	0.67	0.00	0.48
	F1	0.80	0.54	0.88	0.59
	Recall	**1.00**	0.47	**1.00**	0.57
	Precision	0.67	0.69	0.79	0.61
METEOR	Max. acc.	0.73	0.59	0.80	0.57
	Threshold	0.52	0.92	0.26	0.52
	F1	0.81	0.53	0.88	0.70
	Recall	0.87	0.51	0.98	0.92
	Precision	0.76	0.54	0.81	0.56
BERT$_{\text{score}}$	Max. acc.	0.73	0.64	0.80	0.57
	Threshold	0.73	0.96	0.61	0.73
	F1	0.82	0.48	0.89	0.71
	Recall	0.88	0.38	0.99	**0.93**
	Precision	0.76	0.68	0.80	0.57
ParaScore	Max. acc.	0.72	0.56	0.80	0.57
	Threshold	0.74	1.00	0.61	0.74
	F1	0.80	0.00	0.89	0.70
	Recall	0.86	0.00	0.99	0.92
	Precision	0.75	0.40	0.80	0.56
ParaPLUIE DIRECT	Max. acc.	**0.78**	0.69	**0.83**	0.71
	Threshold	8.76	22.84	12.42	21.09
	F1	**0.85**	0.69	**0.90**	0.76
	Recall	0.89	0.77	0.97	0.81
	Precision	0.81	0.62	0.84	0.71
ParaPLUIE INDIRECT	Max. acc.	**0.78**	0.65	0.81	0.70
	Threshold	−8.26	19.10	−11.90	13.50
	F1	0.84	0.69	0.89	0.75
	Precision	0.81	0.57	0.82	0.69
	Recall	**0.88**	**0.86**	**0.98**	**0.82**
ParaPLUIE FS-DIRECT	Max. acc.	0.76	**0.77**	0.81	**0.75**
	Threshold	−3.25	5.53	−7.53	0.37
	F1	0.82	**0.75**	0.88	**0.78**
	Recall	0.83	**0.79**	0.94	0.82
	Precision	0.82	0.71	0.83	0.75

Table 2: F1, recall, precision and associate score threshold of each metrics on every datasets, according to their best classification accuracy. Best accuracy's is in **black**, F1 in **blue**, precision in **orange** and recall in **violet**.

performance does not sharply decrease if we are working on all datasets. This is highlighted by the figure 1.

This leads us to the assumption that ParaPLUIE is resilient and is able to classify paraphrases without being influenced by the edit distance. ParaPLUIE FS-DIRECT seems to be the best template as it shows the best accuracy and the best F1 overall. The best threshold overall is really close to zero. This is convenient because it follows the inherent natural threshold of ParaPLUIE. It is interesting to see that the INDIRECT ParaPLUIE template always gets the best recall. This means that, with the INDIRECT template, we can be really confident about the couples labeled as paraphrases with a threshold chosen retrospectively.

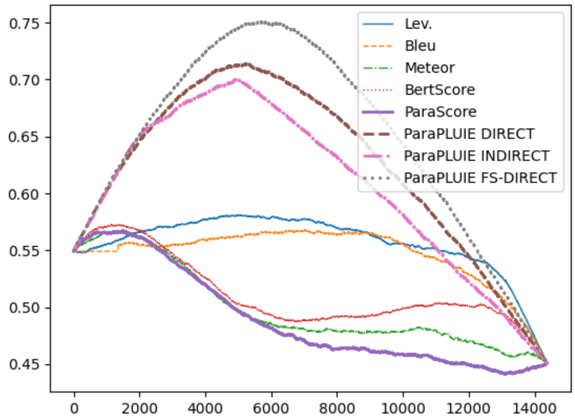


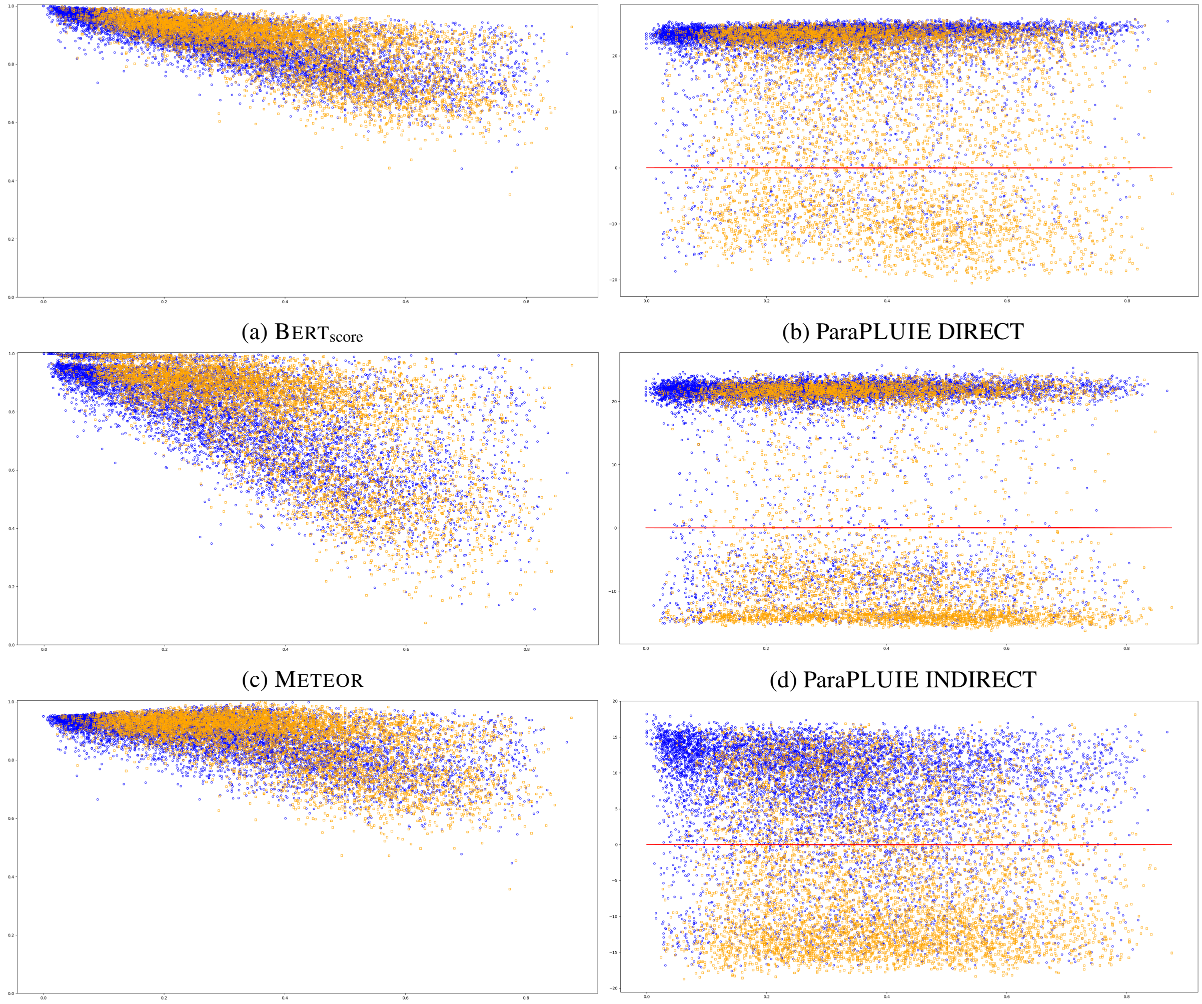
Figure 1: Evolution of the classification accuracy when sentence couples are ranked according to their score. For each rank, couples ranked higher are considered paraphrases and are compared to their gold label.

5.3 Correlation with edit distance

To confirm our previous assumptions, the Pearson correlation between metrics and edit distance is presented in table 3. It is focused on correlation inside each class – paraphrase/non-paraphrase. Indeed, the extent of belonging to a category should be related to semantic distance and not lexical distance. Undoubtedly, other metrics than ParaPLUIE are correlated with the edit distance. This is a concern because the semantic proximity estimation among two sentences should not be guided by the edit distance that separates them. We can observe that, all ParaPLUIE templates are much less correlated with the edit distance. This observation is highlighted by figure 2. More precisely, BERT$_{\text{score}}$, ParaScore and METEOR scores are linked to the edit distance and they are not able to create clusters of paraphrases/non-paraphrases. On the opposite, different ParaPLUIE templates are clearly less correlated and are able to cluster sentence pairs. We can regret that there are many false positives in the paraphrase clusters of the templates DIRECT and INDIRECT. The ParaPLUIE FS-DIRECT made less false positive and false negative errors. It's interesting to point out that, between the two clusters made by ParaPLUIE FS-DIRECT we can notice an area of uncertainty. The closer we are to zero, the less certain the system is to classify the hypothesis sentence. This is a really attractive propriety as it enhances the natural dynamic of the measure. In other words, the higher the score of a hypothesis

	MRPC		PAWS		LLM		Global	
	yes	no	yes	no	yes	no	yes	no
WER	0.88	0.79	0.93	0.91	0.87	0.81	0.93	0.90
BLEU	−0.67	−0.60	−0.66	−0.57	−0.59	−0.64	−0.76	−0.68
METEOR	−0.63	−0.57	−0.45	−0.47	−0.53	−0.59	−0.69	−0.65
BERT$_{\text{score}}$	−0.72	−0.63	−0.61	−0.55	−0.63	−0.63	−0.76	−0.69
ParaScore	−0.56	−0.54	−0.10	−0.23	−0.56	−0.56	−0.60	−0.58
ParaPLUIE DIRECT	**−0.08**	**−0.14**	−0.07	**−0.02**	**−0.00**	**−0.13**	**−0.06**	−0.17
ParaPLUIE EXPLAINS	−0.13	−0.15	**−0.04**	**−0.02**	−0.04	**−0.13**	−0.11	−0.15
ParaPLUIE FS-DIRECT	−0.17	−0.15	−0.05	−0.13	−0.08	−0.22	−0.19	**−0.12**

Table 3: Pearson correlation coefficients between evaluated metrics and the edit distance for each corpus and each class. Emphasis is placed on the weakest correlations.



(a) BERT$_{\text{score}}$

(b) ParaPLUIE DIRECT

(c) METEOR

(d) ParaPLUIE INDIRECT

(e) ParaScore

(f) ParaPLUIE FS-DIRECT

Figure 2: Score distribution of paraphrase couples, represented by blue circles and non-paraphrase, represented in orange squares, in regards of the edit distance. BERT$_{\text{score}}$, METEOR are ParaScore are between zero and one. For ParaPLUIE, the red line denotes the natural threshold at zero.

sentence is, the higher we can be confident in the classification and vice versa.

5.4 Equal Error Rate (EER)

We might view paraphrase classification as a spoofing detection task, where the spoofs are the non-paraphrase sentences. Then, as in spoofing detection, we can compute the EER for each metrics, as in table 4. The EER indicates the point where the false acceptance rate and false rejection rate are equal. It provides an interesting indicator when classes are not balances like in the LLM corpus. The lower the EER, the better the system is.

Thus, all ParaPLUIE templates, compared to

		MRPC	PAWS	LLM	Global
Lev.	EER	0.66	0.66	0.60	0.58
	Threshold	0.44	0.23	0.54	0.33
Bleu	EER	0.37	0.35	0.42	0.48
	Threshold	0.35	0.57	0.20	0.43
Meteor	EER	0.31	0.41	0.40	0.52
	Threshold	0.62	0.91	0.61	0.84
BERT$_{\text{score}}$	EER	0.31	0.37	**0.36**	0.51
	Threshold	0.79	0.93	0.80	0.89
ParaScore	EER	0.33	0.46	0.37	0.53
	Threshold	0.80	0.93	0.81	0.90
ParaPLUIE DIRECT	EER	**0.25**	0.31	0.41	0.31
	Threshold	21.01	23.29	24.72	22.90
ParaPLUIE INDIRECT	EER	**0.25**	0.35	0.39	0.34
	Threshold	16.13	21.18	21.19	20.82
ParaPLUIE FS-DIRECT	EER	0.26	**0.23**	0.40	**0.26**
	Threshold	1.24	6.58	11.29	4.18

Table 4: EER for each metric on each corpus. The lowest error rates are emphasized.

other metrics, have lower or at least similar EER on individual corpora. When all corpora are considered together, then ParaPLUIE templates EER's is much lower than other metrics. Despite of that, for the LLM corpus, BERT$_{\text{score}}$ and ParaScore performs better. This could be explained by the fact that mistakes made by a LLM when generating paraphrases are reproduced while the LLM criticizes them. In other words, it seems more difficult for a LLM to spot LLM errors. That is why ParaPLUIE DIRECT and INDIRECT EER threshold is so high, around 20. They tend to give a high score to many sentences. However, ParaPLUIE FS-DIRECT threshold is much closer than natural classification threshold from ParaPLUIE formula.

5.5 Natural threshold

ParaPLUIE, defined by the logarithm of a ratio, possesses a natural frontier at zero between paraphrases and non-paraphrases. The table 5 offers an overview of different ParaPLUIE templates accuracy and F1 on each datasets, by taking a score of zero as a classification threshold.

We can see that, the ParaPLUIE FS-DIRECT template gets the best accuracy overall. Hence, each ParaPLUIE system, using this a priori threshold perform better – or similar for ParaPLUIE FS-DIRECT on the LLM corpus – than any other metrics using the best threshold a posteriori.

These experiments on complex corpora seem to indicate that ParaPLUIE is a good semantic measure. The FS-DIRECT template looks resilient and able to classify paraphrases and non-paraphrases better than the state of the art. Its natural threshold makes it understandable and usable without having prior knowledge on candidates data.

		MRPC	PAWS	LLM	Global
ParaPLUIE DIRECT	Accuracy	**0.77**	0.56	**0.82**	0.66
	F1	**0.85**	0.66	**0.89**	0.75
ParaPLUIE INDIRECT	Accuracy	0.76	0.64	0.80	0.69
	F1	0.82	0.69	0.88	0.75
ParaPLUIE FS-DIRECT	Accuracy	0.75	**0.75**	0.79	**0.75**
	F1	0.80	**0.76**	0.87	**0.78**

Table 5: Accuracy on each datasets for each ParaPLUIE templates according to a decision threshold fixed at zero. Best accuracy on each datasets is emphasis in **black** and F1 in **blue**.

6 Conclusion

We propose ParaPLUIE, a new metric for evaluating semantic proximity between two sentences. ParaPLUIE is relying on a learned probabilistic model of LLM. It is designed to return scores that can be directly interpreted thanks to the natural threshold of this metric. We have conducted experiments with various templates for ParaPLUIE on three English paraphrase corpora. One of them is created automatically with LLM and delivered with this paper. Our experiments have shown that ParaPLUIE performs better than commonly used measures. Interesting properties of ParaPLUIE are the easy comprehensible threshold between paraphrase and non paraphrases, and that it is marginally correlated to the edit distance.

In future work, we will look into the development of Small Language Models dedicated to the generation of paraphrases, thanks to learning methods like knowledge distillation (Hsieh et al., 2023) and ParaPLUIE as a loss.

7 Ethical considerations

It is important to keep in mind that, ParaPLUIE is not using learning methods. The best template ParaPLUIE FS-DIRECT does not need a generation step and Mistral is a medium size language model. For these reasons, scoring with ParaPLUIE is not much computing intensive. The use of a larger LLM could lead to better results. Nevertheless we appeal to not do that. As we live in a world of limited resources and energy, the research effort should be put into the adaptation and creation of small model dedicated to this task. Therefore, as hint by ParaPLUIE FS-DIRECT good results, distilling and fine-tuning a model for paraphrase evaluation could improve ParaPLUIE.

8 Limitations

This section aims to discuss about other limits than those already discussed.

Experiments in this study were lead on a limited quantity of data. The entire PAWS corpus was not used but only the *dev* subset. This is due to the high computational cost needed to use an LLM, specifically with the ParaPLUIE INDIRECT which needs a generation step. Results on the whole PAWS dataset may vary.

LLAMA2 and MISTRAL, although producing different results, are likely to be trained on very similar data. They both have TRANSFORMER-style architecture and have the same magnitude of weights. Other model architecture may produce different results.

Most sentence pairs inside the LLM dataset made for this experiment were labeled by only one human. Hence, inter-annotator agreement is not available. This corpus is smaller than other corpora, so it has weak impact on the overall evaluation. This corpus turns out to be highly unbalanced since the generation systems produce good paraphrase overall. Moreover it appears that none of the metrics are able to perform very well on it.

The LLM used for ParaPLUIE could have been trained on corpora evaluated during these experiments as they are well known dataset and the training corpus isn't disclosed.

For ParaPLUIE FS-DIRECT, since sources sentences used to build the 6 examples are extracted from MRPC and PAWS, they shared the "style" as the considered corpora. Evaluations on other writing styles could be performed to validate results.

Throughout this paper, three ParaPLUIE versions have been shown. We have tested other templates with different generation step strategies. The prompt plays a critical role here, as small changes in it can involve major differences in scoring. Templates versions proposed may not be optimal.

References

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

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 50–57, Toulouse, France. 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, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374, Nusa Dua, Bali. Association for Computational Linguistics.

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, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Betty Fabre, Tanguy Urvoy, Jonathan Chevelu, and Damien Lolive. 2021. Neural-driven search-based paraphrase generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2100–2111, Online. Association for Computational Linguistics.

Cédric Fayet, Alexis Blond, Grégoire Coulombel, Claude Simon, Damien Lolive, Gwénolé Lecorvé, Jonathan Chevelu, and Sébastien Le Maguer. 2020. FlexEval, création de sites web légers pour des campagnes de tests perceptifs multimédias. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)*, pages 22–25, Nancy, France. ATALA.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: Association for Computational Linguistics 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Vladimir Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Dokl. Akad. Nauk SSSR*, 163:845–848.

Igor M Mel'čuk. 1997. *Vers une linguistique sens-texte: leçon inaugurale faite le vendredi 10 janvier 1997*. Collège de France.

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.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. 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.

Satoshi Sekine. 2005. Automatic paraphrase discovery based on context and keywords between NE pairs. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Lingfeng Shen, Lemao Liu, Haiyun Jiang, and Shuming Shi. 2022. On the evaluation metrics for paraphrase generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3178–3190, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proceedings of North American Chapter of the Association for Computational Linguistics*.

Shiqi Zhao, Xiang Lan, Ting Liu, and Sheng Li. 2009. Application-driven statistical paraphrase generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 834–842, Suntec, Singapore. Association for Computational Linguistics.

A Appendix

A.1 Templates used to generate paraphrases with LLM

We used two models to generate paraphrases, MISTRAL and LLAMA2. Both models are fine-tuned to work as conversational agents. The templates mimic a dialog with a user and the assistant. Here **S** denote the sentence which we want to paraphrase. To create diversity in generated paraphrases, we create a vulgar template. This one shows an example of paraphrase for a vulgar sentence. We empirically observe that, with this kind of example the model is more inventive and tends to use a wider range of vocabulary.

MISTRAL naive(S) :
$(user)$: *You have to transform a sentence A into a paraphrase B. The purpose of A to B transformation is to maintain the original meaning of the sentence A in B.*
$(assistant)$: *I've understood the instructions perfectly.*
$(user)$: *A: "The little cat refreshes himself with water every morning."*
$(assistant)$: *B: "Every morning, the little cat refreshes himself by drinking water."*
$(user)$: *A: "**S**"*

LLAMA2 naive(S) :
$(system)$: *You have to transform a sentence A into a paraphrase B. The purpose of A to B transformation is to maintain the original meaning of the sentence A in B.*
$(user)$: *A: "The little cat refreshes himself with water every morning."*
$(assistant)$: *B: "Every morning, the little cat refreshes himself by drinking water."*
$(user)$: *A: "**S**"*

MISTRAL vulgar(S) :
$(user)$: *You have to transform a sentence A into a paraphrase B. The purpose of A to B transformation is to maintain the original meaning of the sentence A in B. You can use offensive and disrespectful statements.*
$(assistant)$: *I've understood the instructions perfectly.*
$(user)$: *A: "Fuck both of you."*
$(assistant)$: *B: "Go fuck yourselves."*
$(user)$: *A: "**S**"*

A.2 Details about the LLM generated paraphrase dataset

Each judge had to self-assess his English level. They were able to select between, poor, good, advanced and native to self-assess their confidence in their English. To avoid bias in judgement, every button to choose was in grey except for the "Don't know" option which was colored in light-blue. To help judges, word differences between sentences were highlighted in yellow. Additionally, explanations about their task and examples of expected responses were available at any moment. Here are the explanations provided : *"You are going to see two sentences. You are asked to estimate the extent to which the two sentences share a common meaning. To help you, the differences between the sentences are highlighted in yellow of which several examples are shown bellow."* Here are the examples provided:

- The cat drinks water.
- The cat eats kibble.
- Very different
- The associated actions have nothing in common even though the two sentences have the same subject.

- The cat drinks milk.
- The cat drinks water.
- Slightly similar
- The subjects and actions are similar, but water is not milk.

- The cat drinks water.
- The cat quenches its thirst.
- Mostly similar
- The only difference is that the first sentence specifies the type of liquid that is being drunk.

- The cat eats the mouse.
- The mouse is eaten by the cat.
- Same meaning
- Differences linked to context interpretation in these two sentences are too small to say that their meaning is different.

- The cat drinks tomato soup.
- Cat tomato soup.

- Don’t know
- The second sentence doesn’t make any sense. We can’t draw any conclusions from it.

A.3 Complete few shots template

Tpl$_{\text{FS-Direct}}(\mathbf{S}, \mathbf{H})$:

($user$): *You will receive two sentences A and B. Do these two sentences mean the same thing? Answer with only one word "yes" or "no".*

($assistant$): *Please provide the sentences for me to evaluate.*

($user$): *A: "Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence ."; B: "Amrozi accused his brother, whom he disparagingly referred to as ’the liar witness’, of intentionally twisting his testimony."*

($assistant$): *No*

($user$): *A: "Pennmakkal is an Indian Malayalam film from 1966, produced by J. Sasikumar and directed by KP Kottarakkara."; B: "The Indian Malayalam film ’Pennmakkal’, released in 1966, was produced by J. Sasikumar and directed by KP Kottarakkara."*

($assistant$): *Yes*

($user$): *A: "Sorkin , who faces charges of conspiracy to obstruct justice and lying to a grand jury , was to have been tried separately."; B: "Despite being accused of conspiring to obstruct justice and perjury, Sorkin was supposed to stand trial on his own."*

($assistant$): *No*

($user$): *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 ."; 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 ."*

($assistant$): *No*

($user$): *A: "Whereas “e” the electric charge of the particle and A is the magnetic vector potential of the electromagnetic field."; B: "The electric charge of the particle is denoted by “e”, and the magnetic vector potential of the electromagnetic field is denoted by ’A’."*

($assistant$): *Yes*

($user$): *A: "The Jidanul River is a tributary of the Jiul de Vest River in Romania."; B: "The Jidanul River is a mere insignificant stream that flows into the grand Jiul de Vest River in Romania."*

($assistant$): *No*

($user$): *A: "*$\mathbf{S}$*"; B: "*$\mathbf{H}$*"*