LLMs as Judges for Paraphrase Evaluation: Impact of Model Size and Combinations

Anonymous ACL submission

Abstract

Automatically evaluating whether two sentences convey the same meaning is a challenging task, as it requires a deep understanding of context beyond surface-level lexical similarity. Lexical overlap can be misleading: two sentences may be lexically close yet semantically distant, or conversely, semantically equivalent despite limited lexical overlap.

Recent advances in Large Language Models (LLMs) have demonstrated their ability to capture semantic information in their latent space, opening the door to novel evaluation metrics. In this work, we study ParaPLUIE, a semantic similarity metric based on LLM perplexity. We systematically analyze its behavior using 15 LLMs of varying sizes across four labeled paraphrase datasets.

Our results show that the metric remains robust across models, though its effectiveness is influenced by model size. In particular, we observe a strong correlation between model size (on a logarithmic scale) and accuracy in paraphrase detection. Surprisingly, we find that combining multiple smaller models through a simple voting mechanism can outperform larger individual models, highlighting the potential of ensemble approaches for semantic evaluation.

1 Introduction

Determining whether two sentences express the same meaning is a challenging problem that requires moving beyond surface similarity to capture nuanced semantics. Early automatic metrics based on lexical overlap often fail on such cases, while embedding-based methods and instruction-tuned LLMs offer improved semantic sensitivity at the expense of robustness, interpretability, or computational cost (Papineni et al., 2002; Zhang et al., 2020; Reimers and Gurevych, 2019; Chiang and Lee, 2023). In this paper we focus on ParaPLUIE, an "LLM-as-a-judge" metric that avoids the need to parse free-text outputs. Since its a perplexity-based LLM metric that can be computed from likelihoods rather than generations, it reduce the computational time while increasing the interpretability (Lemesle et al., 2025).

Given the growing diversity of available models, we ask two practical questions: Which models are reliable judges for paraphrase evaluation, and how does model scale affect accuracy and efficiency? Answering these questions is important not only for metric quality but also for computational and energy considerations when deploying evaluation pipelines.

We empirically study ParaPLUIE across a range of model families and sizes, analyze score distributions and inter-model agreement, and measure the trade-offs between size and accuracy. Our contributions are threefold:

- an empirical analysis of ParaPLUIE under different underlying LLMs,
- a examination of the relationship between parameter scale and paraphrase evaluation performance,
- and an investigation of inter-model agreement in paraphrase classification.

The paper is organized as follows. Section 2 reviews neural metrics for semantic similarity, with a focus on LLM-as-a-judge methods and the ParaPLUIE metric. To study these questions, we first describe the language models and datasets used in our experiments (Section 3). We detail our experimental setup in Section 4 and report results in Sections 4.1–4.2, before concluding in Section 5.

2 DNN-based metrics

To evaluate how semantically close two sentences are, early metrics were lexically based, such as BLEU (Papineni et al., 2002), which assumes that sentences sharing similar words are more likely to share meaning. However, small structural changes can drastically alter semantics (Zhang et al., 2019), while conversely, lexically distant sentences may still express the same meaning.

The semantic representations enabled by the Transformer architecture (Vaswani et al., 2017) have led to a new generation of metrics for evaluating meaning similarity using language models. Traditional metrics for assessing semantic proximity at the sentence level rely on encoder-based models, primarily BERT (Devlin et al., 2019). Among the most widely used are BERTSCORE (Zhang et al., 2020) and SENTENCEBERT (Reimers and Gurevych, 2019), which remain standard choices for measuring semantic similarity (Tytgat et al., 2024). However, such metrics often fail to distinguish between sentences that are lexically similar but semantically different (Zhang et al., 2019).

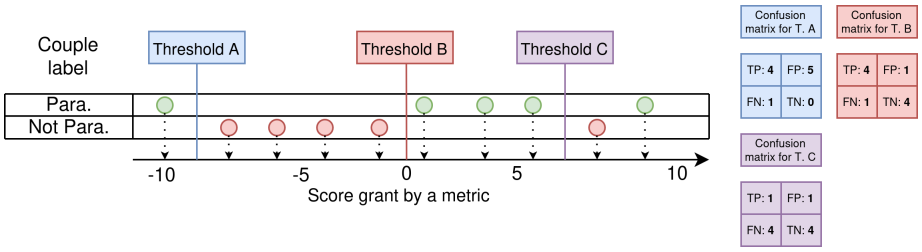


Figure 1: Confusion matrix for dummies data and metric. Green circles denotes a couple labeled paraphrase and red non-paraphrase. By varying the threshold (A,B and C) the results in according confusion matrix change. TP, FP, FN and TN respectively denote True Positives, False Positives, False Negatives and True Negatives.

As an alternative, metrics that rely on LLMs as judges have recently emerged. Previous studies have shown that LLM-as-a-judge approaches can approximate human judgment in several NLP tasks, such as text quality assessment (Chiang and Lee, 2023). Nonetheless, multiple limitations have been identified, including adversarial robustness and inherent biases (Gu et al., 2025; Raina et al., 2024). In the specific case of paraphrase detection, position bias and self-enhancement (or egocentric) bias are particularly critical (Zheng et al., 2023). Position bias has been addressed by varying prompt templates (Lemesle et al., 2025), while mitigating self-enhancement bias requires careful dataset design. Thus, although LLM-as-a-judge metrics show promise, reliability and robustness remain open challenges (Gu et al., 2025).

The ParaPLUIE metric was originally proposed and studied by Lemesle et al. (2025). We highlight here its key features. Instead of interpreting the free-text output of an LLM, ParaPLUIE computes the log-likelihood of the tokens *Yes* and *No* within a templated prompt. Formally, as shown in Equation 1, where Tpl is the prompt filled with the source sentence S and a candidate paraphrase H, and T the length in token of the filled template, the score ParaPLUIE corresponds to the difference in loss assigned by the LLM to the two possible continuations (*Yes* or *No*).

$$\begin{aligned}\text{ParaPLUIE}(S,H) &= \log\left(\frac{p(\text{yes}|\text{Tpl}(S,H))}{p(\text{no}|\text{Tpl}(S,H))}\right)\\ &= (T+1)\times\Big(\text{loss}_{LLM}(\text{Tpl}(S,H)\circ no)\\ &\quad -\text{loss}_{LLM}(\text{Tpl}(S,H)\circ yes)\Big)\end{aligned} \tag{1}$$

The ParaPLUIE metric yields a numerical score in the range $]-\infty,\infty[$. Its main advantage is interpretability: it directly reflects the model's relative confidence in answering *Yes* versus *No*. A score close to 0 indicates uncertainty; increasingly negative values correspond to stronger confidence in *No*, while increasingly positive values correspond to stronger confidence in *Yes*. Interestingly, the initial study reported no clear correlation between this score and simple lexical measures such as edit distance. Moreover, the computational cost is significantly reduced compared to standard LLM-as-a-judge methods, since only a single token needs to be evaluated rather than generating a full response, leading to substantial savings in both time and energy.

3 Experimental Resources

This section describes the core resources used in our experiments: the language models selected to act as LLM-as-a-judge and the datasets employed for paraphrase evaluation. We explain model size groups, specific characteristics, and potential sources of bias in the data.

3.1 Language Models

To study the behavior of the ParaPLUIE metric, we selected a diverse set of publicly available instruction-tuned language models, listed in Table 3. All models were downloaded from the Hugging Face platform[1]. By *instruction-tuned*, we refer to models fine-tuned on dialogue-style tasks, so that they can act as "LLMs-as-a-Judge".

Our selection spans a wide range of parameter sizes, from less than one billion to more than thirty billion. For convenience, we group them into three categories:
Small models ($\leq$2B parameters),
Medium models ($\sim$4–9B parameters),
Large models ($\geq$14B parameters).

Small models. We include three versions of SmolLM2-instruct (0.135B, 0.36B, 1.7B), Gemma2 (2B) and InternLM2-chat (1.8B). The latter is noteworthy for introducing Conditional Online Reinforcement Learning from Human Feedback (COOL-RLHF), a novel online variant of RLHF (Cai et al., 2024).

Medium models. This group contains Phi-4-mini-instruct (4B), Mistral v0.2 (7B), Qwen2.5 (7B), Llama3 (8B), Aya-expanse (8B), and Gemma2 (9B). Mistral adopts a sliding attention window to handle long contexts efficiently, whereas Qwen2.5 benefits from a hybrid post-training combining DPO (offline) and GRPO (online) learning. We retain Mistral v0.2 as a baseline, since it was used in the original ParaPLUIE paper.

Large models. We evaluate Phi-4 (14B), Aya-expanse (32B), Cohere's c4ai (32B), and QwQ-32B. Aya-expanse is trained using data arbitrage (Odumakinde et al., 2025)

[1] https://huggingface.co/

		MRPC	PAWS	MCPG	LLM	HC	MBD
		Baseline					
Bertscore	*Threshold*	0.73	0.96	0.76	*0.61*	0.76	0.79 ±0.01
	Acc.	0.73	0.64	0.84	0.80	0.56	0.62±0.00
	Prec.	0.76	0.68	0.86	0.80	0.97	0.73±0.01
	Rec.	0.88	0.38	0.92	0.99	0.63	0.77±0.01
	F1	0.82	0.48	0.89	0.89	0.59	0.68±0.00
		Small models					
sm-0.13B	*Acc.*	0.33	0.56	0.31	0.21	0.50	0.50±0.00
	Prec.	0.00	0.00	0.00	0.00	0.00	0.00±0.00
	Rec.	0.00	0.00	0.00	0.00	0.00	0.00±0.00
	F1	0.00	0.00	0.00	0.00	0.00	0.00±0.00
sm-0.36B	*Acc.*	0.43	0.53	0.42	0.43	0.54	0.52±0.01
	Prec.	0.72	0.46	0.81	0.79	0.59	0.54±0.01
	Rec.	0.24	0.40	0.22	0.38	0.22	0.30±0.01
	F1	0.36	0.43	0.34	0.51	0.32	0.38±0.01
sm-1.7B	*Acc.*	0.51	0.64	0.64	0.51	0.74	0.64±0.01
	Prec.	0.79	0.63	0.88	0.82	0.80	0.67±0.01
	Rec.	0.37	0.45	0.55	0.49	0.64	0.57±0.02
	F1	0.51	0.52	0.68	0.62	0.71	0.62±0.01
itrn-1.8B	*Acc.*	**0.75**	0.67	0.73	**0.80**	0.75	0.64±0.00
	Prec.	0.76	0.58	0.72	0.80	0.68	0.59±0.01
	Rec.	**0.93**	**0.96**	**0.99**	**0.99**	**0.94**	**0.96**±0.00
	F1	**0.83**	**0.72**	0.84	**0.88**	0.79	0.73±0.01
gem-2B	*Acc.*	0.58	**0.70**	**0.84**	0.66	**0.84**	**0.75**±0.01
	Prec.	**0.89**	**0.65**	**0.91**	**0.87**	**0.88**	**0.74**±0.01
	Rec.	0.43	0.70	0.84	0.67	0.80	0.75±0.01
	F1	0.58	0.67	**0.88**	0.76	**0.84**	**0.74**±0.01
		Medium models					
phi-4B	*Acc.*	0.69	0.76	0.79	0.78	0.82	0.73±0.01
	Prec.	0.89	0.67	0.89	0.85	0.93	0.71±0.01
	Rec.	0.62	0.88	0.79	0.88	0.69	0.79±0.02
	F1	0.73	0.76	0.84	**0.87**	0.79	0.75±0.01
mis-7B	*Acc.*	**0.74**	0.75	0.88	**0.79**	0.87	0.75±0.01
	Prec.	0.84	0.67	0.95	0.84	0.94	0.71±0.01
	Rec.	**0.76**	0.88	0.88	**0.91**	0.79	**0.86**±0.01
	F1	**0.80**	0.76	0.91	**0.87**	0.86	0.77±0.01
qnw-7B	*Acc.*	0.66	0.80	**0.90**	0.76	**0.89**	**0.77**±0.01
	Prec.	0.91	0.78	0.96	0.85	0.98	0.74±0.01
	Rec.	0.55	0.77	**0.89**	0.84	**0.80**	0.82±0.01
	F1	0.68	0.77	**0.92**	0.85	**0.88**	**0.78**±0.01
aya-8B	*Acc.*	0.73	**0.87**	0.89	0.78	0.74	0.69±0.01
	Prec.	0.85	**0.81**	0.96	0.85	0.93	0.69±0.01
	Rec.	0.72	**0.93**	0.88	0.88	0.53	0.70±0.01
	F1	0.78	**0.87**	**0.92**	**0.87**	0.68	0.70±0.01
lla-8B	*Acc.*	0.60	0.78	0.84	0.70	0.77	0.72±0.01
	Prec.	0.94	0.76	0.96	**0.87**	0.95	**0.75**±0.01
	Rec.	0.43	0.72	0.80	0.74	0.56	0.66±0.01
	F1	0.59	0.74	0.88	0.80	0.70	0.70±0.01
gem-9B	*Acc.*	0.55	0.75	0.82	0.71	0.84	0.75±0.01
	Prec.	**0.96**	0.79	**1.00**	0.86	**0.99**	**0.75**±0.01
	Rec.	0.35	0.59	0.74	0.76	0.70	0.74±0.01
	F1	0.51	0.68	0.85	0.81	0.82	0.74±0.01
		Large models					
phi-14B	*Acc.*	0.64	0.78	0.89	0.77	0.81	0.73±0.01
	Prec.	**0.93**	0.75	**0.99**	**0.86**	**0.97**	0.73±0.01
	Rec.	0.50	0.76	0.85	0.84	0.63	0.73±0.01
	F1	0.65	0.76	0.91	0.85	0.76	0.73±0.01
aya-32B	*Acc.*	**0.74**	**0.90**	**0.92**	**0.79**	**0.94**	**0.79**±0.01
	Prec.	0.85	**0.86**	0.92	0.85	**0.97**	0.73±0.01
	Rec.	0.74	0.93	**0.97**	0.90	**0.91**	0.90±0.01
	F1	0.79	**0.89**	**0.94**	0.87	**0.94**	**0.81**±0.00
qwq-32B	*Acc.*	0.72	0.70	0.89	**0.79**	0.90	0.75±0.01
	Prec.	0.74	0.60	0.90	0.83	**0.97**	0.68±0.00
	Rec.	**0.89**	**0.95**	0.95	**0.93**	0.83	**0.96**±0.01
	F1	**0.81**	0.74	0.92	**0.88**	0.89	0.79±0.01
c4ai-32B	*Acc.*	0.65	0.82	0.90	0.76	0.89	0.77±0.01
	Prec.	0.92	0.77	0.93	**0.86**	0.96	**0.74**±0.01
	Rec.	0.51	0.85	0.92	0.84	0.81	0.84±0.01
	F1	0.66	0.81	0.93	0.85	0.88	0.79±0.01

Table 1: Accuracy, precision, recall and F1 of each models on every datasets. The best accuracy for a dataset is in **orange**, precision in **blue**, recall in **black** and F1 in **violet**. Cells colored in **green** denote the best accuracy, precision, recall and F1 for a corpus among all models size.

and model merging techniques (Aakanksha et al., 2024), while QwQ-32B is the only reasoning-focused model in our set, combining supervised fine-tuning and RLHF.

None of the selected models have publicly released training datasets, which raises the possibility of contamination bias (discussed in Section 3.2).

3.2 Datasets

Evaluating semantic similarity at the sentence level requires datasets of labeled sentence pairs. Such corpora should ideally cover a broad range of cases: paraphrases and non-paraphrases that are lexically close but semantically distant, and *vice versa*, along with pairs that are distant in both form and meaning.

We consider five datasets (Table 2):

- **MRPC** (Microsoft Research Paraphrase Corpus) (Dolan and Brockett, 2005),
- **PAWS** (Paraphrase Adversaries from Word Scrambling) (Zhang et al., 2019),
- **MCPG** (Machine-Comprehension Paraphrase Generation) (Fabre et al., 2021),
- **LLM** (Lemesle et al., 2025),
- **HC** (Lemesle et al., 2025).

The first two are long-standing benchmarks, but their presence in pretraining data cannot be ruled out since model training sets are undisclosed. To mitigate this risk, we complement them with three more recent corpora (MCPG, LLM and HC), manually labeled by human judges and not directly available on internet.

A further challenge is dataset imbalance: except for HC, most corpora are skewed toward one label. As MRPC and PAWS are significantly larger, they would dominate aggregated analyses. To counter this, we construct a **Mixed Balanced Dataset (MBD)** by randomly sampling 45 paraphrase and 45 non-paraphrase pairs from each corpus. Reported MBD results are averaged over 100 draws with 95% confidence intervals.

Table 2 summarizes dataset statistics, including size, label distribution, and average edit distance.

4 Experiments & Results

We evaluate the ParaPLUIE metric by applying it with different language models to classify sentence pairs as paraphrases or non-paraphrases. Since the score produced by ParaPLUIE depends on the underlying perplexity model, we examine how the metric behaves across a variety of models. Our analysis focuses on three aspects: the classification performance measured with accuracy, precision, recall, and F1 score; the degree of agreement between models; and the influence of model size on the effectiveness of ParaPLUIE.

To compute these metrics, we must define a decision threshold for classification. This threshold determines the score above which a pair is predicted as a paraphrase and below which it is predicted as a non-paraphrase. Figure 1 illustrates this principle. Since ParaPLUIE provides a natural threshold at 0, we adopt it as the

Dataset	Size	Paraphrase	Non-Paraphrase	Average Edit distance
PAWS	8,000	44%	56%	0.27
MRPC	5,801	67%	33%	0.42
MCPG	146	69%	31%	0.30
LLM	578	79%	21%	0.50
HC	200	50%	50%	0.40
MBD	225	50%	50%	0.38 ± 0.00

Table 2: Size in number of coupled and labeled distribution of each considered dataset. The mean normalized edit distance is also presented, result for MBD is average of 100 draws with 95% confidence interval.

Id	Models	Reference	Size
Small models			
`sm-0.13B`	SmolLM2-135M-Instruct	(Allal et al., 2025)	135M
`sm-0.36B`	SmolLM2-360M-Instruct	(Allal et al., 2025)	360M
`sm-1.7B`	SmolLM2-1.7B-Instruct	(Allal et al., 2025)	1.7B
`itrn-1.8B`	internlm2-chat-1_8b	(Cai et al., 2024)	1.8B
`gem-2B`	gemma-2-2b-it	(Team et al., 2024)	2B
Medium models			
`phi-4B`	Phi-4-mini-instruct	(Microsoft et al., 2025)	4B
`mis-7B`	Mistral-7B-Instruct-v0.2	(Jiang et al., 2023)	7B
`qwn-7B`	Qwen2.5-7B-Instruct	(Qwen et al., 2025)	7B
`aya-8B`	aya-expanse-8b	(Dang et al., 2024)	8B
`lla-8B`	Meta-Llama-3.1-8B-Instruct	(Grattafiori et al., 2024)	8B
`gem-9B`	gemma-2-9b-it	(Team et al., 2024)	9B
Large models			
`phi-14B`	Phi-4	(Abdin et al., 2024)	14B
`aya-32B`	aya-expanse-32b	(Dang et al., 2024)	32B
`qwq-32B`	QwQ-32B	(Team, 2025)	32B
`c4ai-32B`	c4ai-command-r-08-2024	(Verga et al., 2024)	32B

Table 3: Considered language models, categorized by their number of parameters. In the remainder of the paper, we will refer to them using the associated 'Id' identifier.

default: pairs with a score greater than zero are classified as paraphrases, and those with a score less than or equal to zero as non-paraphrases. Predictions are then compared with gold-standard human annotations.

As a historical baseline, we also report classification results obtained using BERTSCORE (Zhang et al., 2020) as a metric. Since BERTSCORE does not provide a natural decision threshold, we adopt the best a-posteriori threshold for classification. Although this approach is not directly practical, it provides an upper bound on the performance that BERTSCORE can achieve.

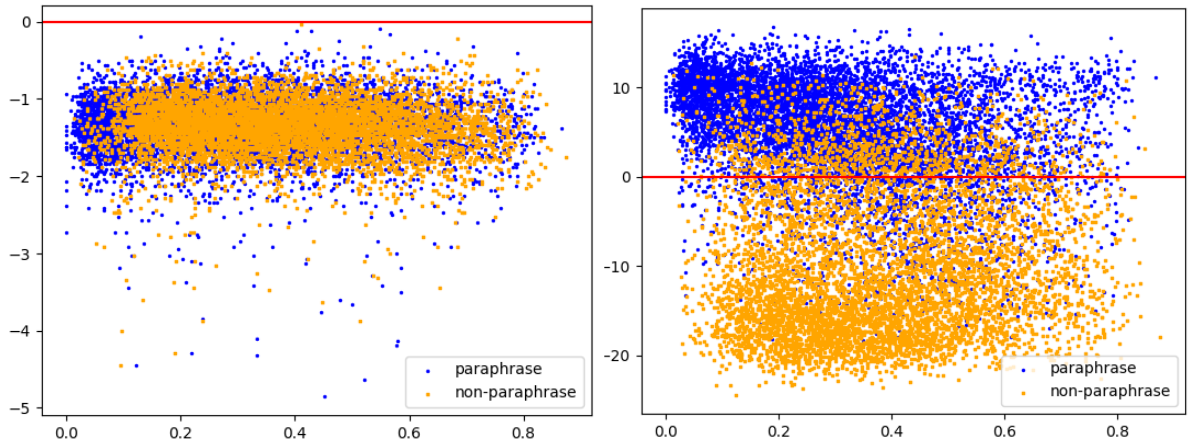
4.1 ParaPLUIE Metric distribution

Table 1 presents the classification results for each model across all datasets, organized by model size.

We observe substantial variation in classification accuracy among the small models. The `sm-0.13B` and `sm-0.36B` models are largely unable to distinguish paraphrases from non-paraphrases. In particular, `sm-0.13B` consistently classifies all pairs as non-paraphrases, as illustrated in Figure 2a, resulting in zero precision and recall for the paraphrase class. Its accuracy of 0.21 on the LLM dataset confirms this behavior. The larger small model, `gem-2B`, achieves the highest accuracy on our control dataset, MBD, suggesting that performance improves with model size.

Medium-sized models show less sensitivity to parameter count on MBD, although `qwn-7B` achieves the highest accuracy in this group. Best overall results are obtained with the largest models. However, `aya-8B` exhibits lower accuracy on MBD compared to other datasets, potentially indicating contamination from its training data. As a reminder, `phi-14B` has 14B parameters, while `aya-32B`, `qwq-32B`, and `c4ai-32B` all have 32B parameters. Interestingly, the improvement in accuracy appears to diminish as model size increases.

By examining the table, we observe some unexpected behaviors. The model `itrn-1.8B` achieves the highest accuracy on both MRPC and LLM corpora. On the LLM dataset, it is only slightly outperformed by BERTSCORE, which should be interpreted with caution since BERTSCORE relies on an a-posteriori threshold. This model also has the highest recall across all corpora, indicating that it rarely misclassifies paraphrases as non-paraphrases. It is worth noting that the LLM corpus is the most unbalanced, with 79% of pairs labeled as



(a) Model sm-0.13B (b) Model aya-32B

Figure 2: Distribution of the ParaPLUIE metric (y-axis) for all sentence pairs according to their edit distance (x-axis). The red line denotes the classification threshold, blue dots are paraphrases, and orange dots are non-paraphrases.

paraphrases.

On the opposite gem-9B show an high precision which mean it classify few non-paraphrases as paraphrases. Notably, aya-32B have the highest accuracy on MBD and on three other corpora (LLM, MCPG and HC). The figure 2 offers a visual representation of the distribution of metric to couple according to the edit distance between the two sentences, for the best and the worst models.

We can note that large model struggles on the LLM dataset which confirms results of previous works that LLM evaluators recognize and favor their own generation (Panickssery et al., 2024; Shimabucoro et al., 2024). Some of lower size models (itrn-1.8B, phi-4B, mis-7B or aya-8B) are not so affected by this bias.

Overall, when comparing the accuracy of all models, we observe a strong correlation with the number of parameters. On a logarithmic scale of model size, the Pearson correlation reaches 0.89, which is quite high. This relationship is illustrated in Figure 3, suggesting that model size indeed impacts performance. However, the logarithmic trend also implies that extremely large models are required to achieve modest gains in accuracy. This may indicate a potential performance ceiling, where exponentially increasing model size provides diminishing returns relative to the improvement in classification performance.

Nevertheless, the top 3 models according to their accuracy on each dataset is pretty volatile. This could indicate that the models do not agree with each other when classifying couples. To verify that, we propose to look forward their mutual agreement in section 4.2.

4.2 Inter-model Agreement and Combinations

An obvious question when working with an LLM-as-a-judge metric is whether the judges agree with one another. Our previous observations suggest that this is often not the case. We therefore first study inter-model agreement at the group level using the Light's Kappa score (Richard and GG., 1977). We then analyze peer-to-peer agreements across all models using the pairwise Kappa score (Cohen, 1960).

Table 4 shows that larger models tend to agree more strongly with others in their group. However, when considering all models together, the percentage agreement is extremely low, even though the Kappa values fall within the "fair" range. This indicates that consensus across models is quite rare. A similar trend is observed across datasets: the HC corpus yields the highest agreement scores, while MRPC produces the lowest. This result is somewhat counterintuitive, as HC was explicitly designed to be challenging, whereas MRPC is likely included in many models' training data.

Pairwise agreements are further illustrated in Figure 4, focusing on our control corpus MBD. Here, the smallest models (sm-0.13B, sm-0.36B, and sm-1.7B) show virtually no agreement with any other models. Among the remaining models, agreements are generally weak, with the exception of the pair (phi-4B, c4ai-32B), which reaches a borderline moderate level of agreement.

As models seem unable to consistently agree on their classifications, yet still perform well in terms of accuracy on the corpora studied, one may wonder whether a majority-vote strategy could further improve performance. In particular, we investigate whether combinations of smaller models can outperform larger ones. To this end, we evaluate all possible triplets and quintuplets of systems whose total parameter count remains below 46B. Mixtures including 32B-parameter models are therefore excluded. Combinations involving sm-0.13B, sm-0.36B, or sm-1.7B are also excluded due to their poor individual performance. Figure 5 presents the accuracy of these combinations on the balanced dataset MBD.

We first observe that certain triplets outperform much larger models. For example, the combination phi-14B+qwn-7B+itrn-1.8B achieves an accuracy of 0.88 with 22.8B parameters. Similarly, phi-4B+mis-7B+itrn-1.8B reaches 0.83 with only 12.8B parameters. However, some combinations perform catastrophically:

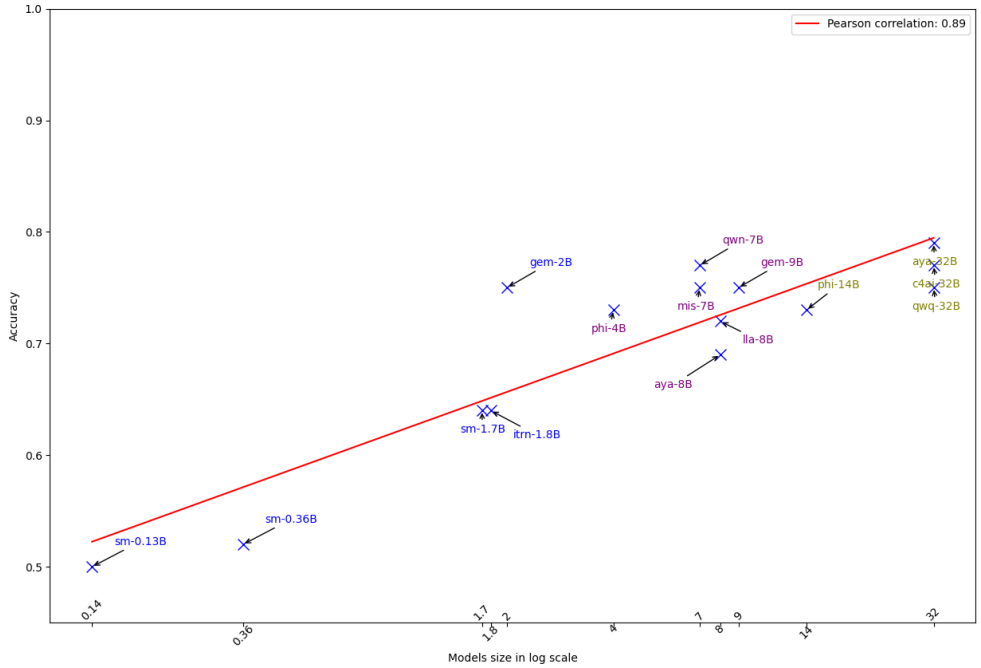


Figure 3: Pearson correlation between accuracy (y-axis) and logarithm scale model size (x-axis) on the MBD dataset.

Dataset	% agreement	Lights Kappa	Agreement	% agreement	Lights Kappa	Agreement
	Smalls models			Mediums models		
MBD	13.06%	0.11	Slight	58.61%	0.63	Substantial
PAWS	14.90%	0.12	Slight	56.20%	0.61	Substantial
MRPC	13.70%	0.08	Slight	43.30%	0.48	Moderate
MCPG	4.11%	0.06	Slight	56.80%	0.63	Substantial
LLM	2.08%	0.08	Slight	69.70%	0.62	Substantial
HC	24.00%	0.15	Slight	63.00%	0.64	Substantial
	Larges models			Every models		
MBD	65.62%	0.63	Substantial	8.71%	0.41	Moderate
PAWS	62.20%	0.60	Substantial	8.60%	0.39	Fair
MRPC	39.60%	0.38	Fair	4.53%	0.28	Fair
MCPG	80.80%	0.77	Substantial	3.42%	0.40	Fair
LLM	76.10%	0.55	Moderate	1.21%	0.35	Fair
HC	78.00%	0.75	Substantial	22.00%	0.46	Moderate

Table 4: Light's Kappa score for each dataset with percentage agreement and it's interpretation. Sub-tables represents the agreement in a group of models.

for instance, `gem-9B+itrn-1.8B+gem-2B` yields 0.56 accuracy, well below the scores of the individual models (0.75, 0.64, and 0.75, respectively).

When combining five models, results are more stable: we did not observe any catastrophic cases, though parameter counts increase substantially. The weakest quintuplet, `phi-4B+lia-8B+gem-9B+itrn-1.8B+gem-2B`, which includes the worst-performing triplet, still achieves 0.85, outperforming the best single model (`aya-32B`, 0.79). The best quintuplet, `phi-4B+qwn-7B+aya-8B+lia-8B+gem-9B`, achieves 0.88 with a total of 46B parameters. Interestingly, some quintuplets below the 32B threshold (e.g., `qwn-7B+lia-8B+gem-9B+itrn-1.8B+gem-2B`, 0.82 with 27.8B parameters) also surpass the larger models considered in this study.

Because ParaPLUIE requires only a single token evaluation per model, combining several smaller models is computationally practical and often favorable, especially when executed in parallel. Memory consumption also tends to be lower compared to using a single large model. Nevertheless, these trends should be interpreted with caution, as model selection was made *a posteriori* based on individual accuracy, and may not generalize universally. But, this suggests that diversity of judges may be as important as size, opening avenues for ensemble approaches.

5 Conclusion

In this paper, we investigated the reliability of large language models as judges for paraphrase evaluation, with a particular focus on the ParaPLUIE metric. Unlike generic LLM-as-a-judge approaches, ParaPLUIE leverages the model's perplexity over a single token to produce a well-calibrated semantic score. Our experiments validate this approach across multiple datasets and model sizes, demonstrating high accuracy

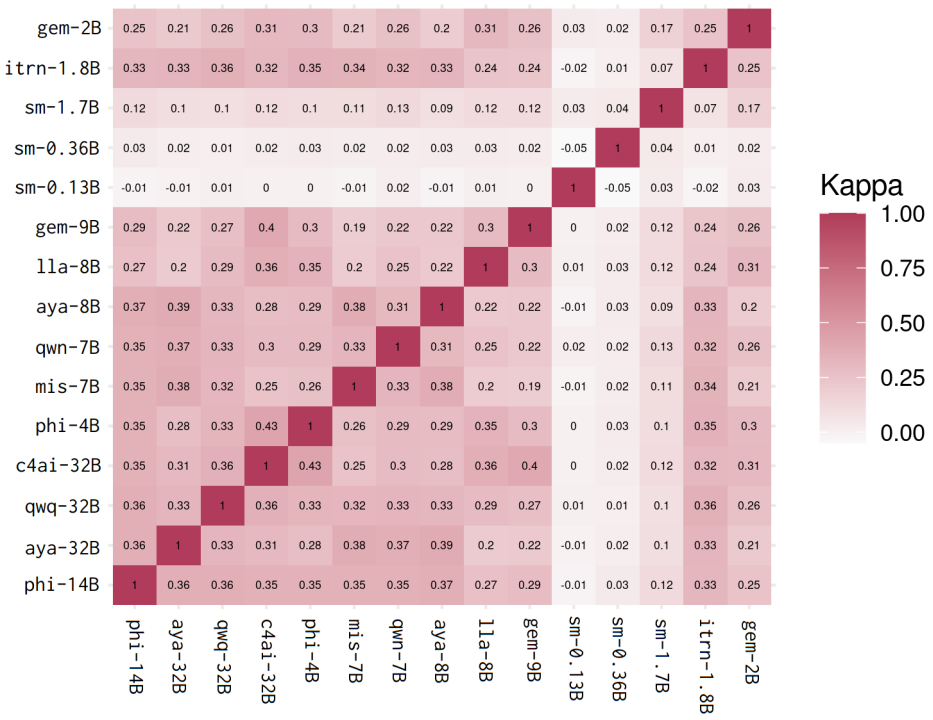


Figure 4: Agreement heatmap between models on MBD corpus

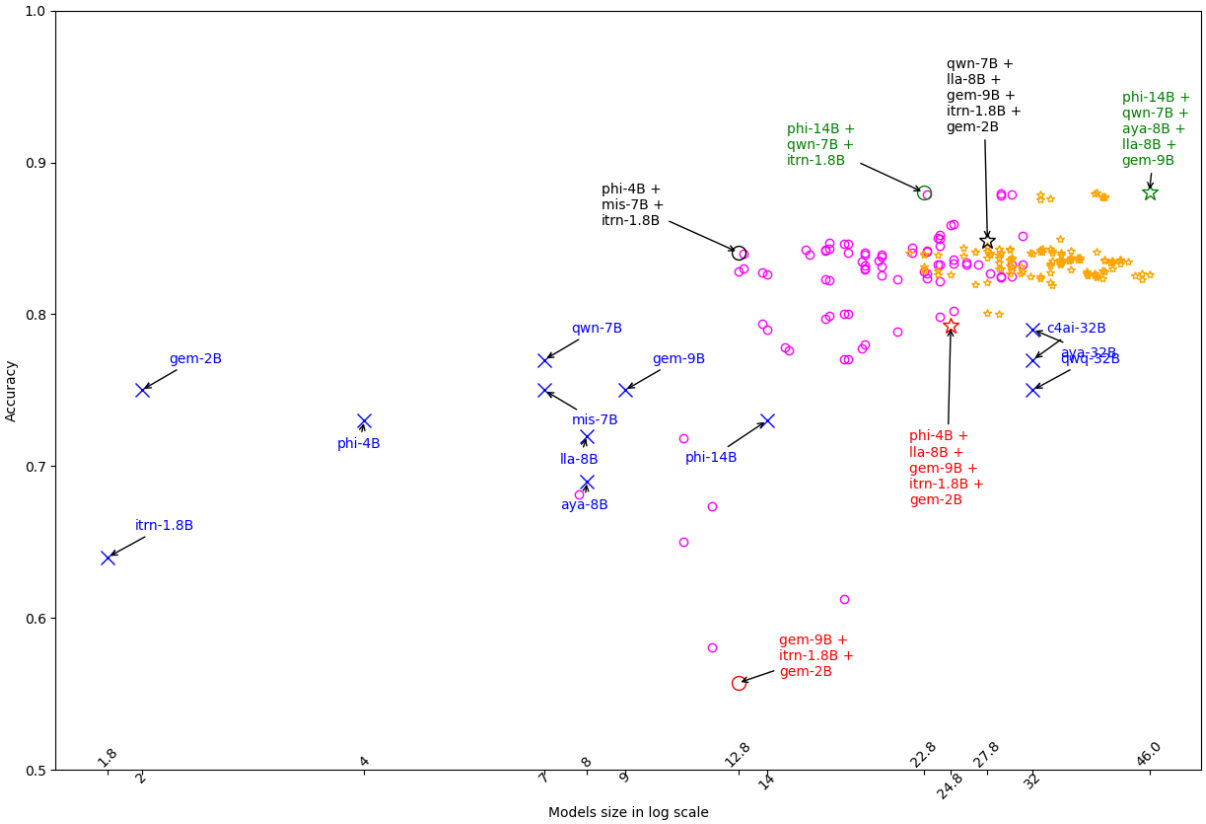


Figure 5: Accuracy (y-axis) of combined models as a function of their total number of parameters on a logarithmic scale (x-axis), evaluated on the MBD dataset. **Blue** crosses denote single models, **fuchsia** circles triplets, and **orange** stars quintuplets. The best and worst combinations are highlighted in **green** and **red**, respectively. The best triplet under 14B parameters and the best quintuplet under 32B parameters are also marked in **black**.

without task-specific fine-tuning.

We observed that model size strongly impacts accuracy, with a logarithmic trend indicating diminishing returns for extremely large models. Despite their performance, models often disagree with each other, highlighting the inherent variability in LLM judgment. Importantly, combining models via majority voting can improve both stability and accuracy, with ensembles of medium-sized models sometimes outperforming single large models while remaining computationally efficient.

Future work could explore more principled ensemble strategies, investigate additional datasets to reduce bias, and extend the approach to other semantic evaluation tasks. It would also be instructive to study extra-large models to determine whether they follow the same scaling law. Overall, our study suggests that carefully designed LLM-based metrics like ParaPLUIE provide a robust alternative for paraphrase evaluation, offering both interpretability and competitive performance.

Limitations

The models used in this study share the same or highly similar Transformer architecture. It would be informative to extend this analysis to alternative architectures, such as *Mamba* (Gu and Dao, 2024). Preliminary results with two Mamba-based models are provided in Appendix B.

All experiments were conducted on English data only, results may not generalize to other languages. Additionally, some datasets used in this work, particularly MRPC and PAWS may be included in the training data of the evaluated models, potentially influencing performances. The datasets also vary considerably in size, which could introduce bias into our analysis. To mitigate this risk, we conducted experiments using a balanced control corpus with equal representation of paraphrase and non-paraphrase pairs across datasets.

Finally, models larger than 32B parameters were not included in our experiments. As a result, conclusions regarding model size and performance correlations may differ when considering a broader range of model scales.

References

Aakanksha, Arash Ahmadian, Seraphina Goldfarb-Tarrant, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. 2024. Mix data or merge models? optimizing for diverse multi-task learning. *Preprint*, arXiv:2410.10801.

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.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, and 3 others. 2025. Smollm2: When smol goes big – data-centric training of a small language model. *Preprint*, arXiv:2502.02737.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, and 81 others. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

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

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark,

and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. *Preprint*, arXiv:2312.00752.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

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.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, and 57 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Ayomide Odumakinde, Daniel D'souza, Pat Verga, Beyza Ermis, and Sara Hooker. 2025. Multilingual arbitration: Optimizing data pools to accelerate multilingual progress. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19142–19164, Vienna, Austria. Association for Computational Linguistics.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. *Association for Computational Linguistics*, pages 311–318.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Vyas Raina, Adian Liusie, and Mark Gales. 2024. Is LLM-as-a-judge robust? investigating universal adversarial attacks on zero-shot LLM assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7499–7517, Miami, Florida, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Landis J. Richard and Koch GG. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159.

Luísa Shimabucoro, Sebastian Ruder, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. 2024. LLM see, LLM do: Leveraging active inheritance to target non-differentiable objectives. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9243–9267, Miami, Florida, USA. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Julie Tytgat, Guillaume Wisniewski, and Adrien Betrancourt. 2024. Évaluation de la similarité textuelle : Entre sémantique et surface dans les représentations neuronales. *ATALA and AFPC*, pages 85–96.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *Preprint*, arXiv:2404.18796.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Jingwei Zuo, Maksim Velikanov, Dhia Eddine Rhaiem, Ilyas Chahed, Younes Belkada, Guillaume Kunsch, and Hakim Hacid. 2024. Falcon mamba: The first competitive attention-free 7b language model. *Preprint*, arXiv:2410.05355.

A Distribution of the ParaPLUIE metric according to their edit distance

Figure 6 provides additional visual representations of the distribution of metric scores across sentence pairs with respect to edit distance, as discussed in Section 4.1.

B Preliminary results with Mamba-based models

Table 5 presents preliminary results obtained with two Mamba-based models. We used the implementation of Mamba 2.8B[2] and Falcon Mamba (Zuo et al., 2024), a 7B model.

		MRPC	PAWS	MCPG	LLM	HC
Mamba	*Acc.*	0.38	0.52	0.48	0.34	0.56
	Prec.	0.68	0.43	0.63	0.81	0.59
	Rec.	0.15	0.25	0.58	0.21	0.41
	F1	0.24	0.31	0.61	0.33	0.48
Falcon-Mamba	*Acc.*	0.59	N/A	0.80	0.68	0.78
	Prec.	0.89	N/A	0.96	0.85	0.98
	Rec.	0.45	N/A	0.74	0.72	0.57
	F1	0.60	N/A	0.84	0.78	0.72

Table 5: Accuracy, precision, recall and F1 of Mamba's models on every datasets.

[2] https://huggingface.co/state-spaces/mamba-2.8b-hf

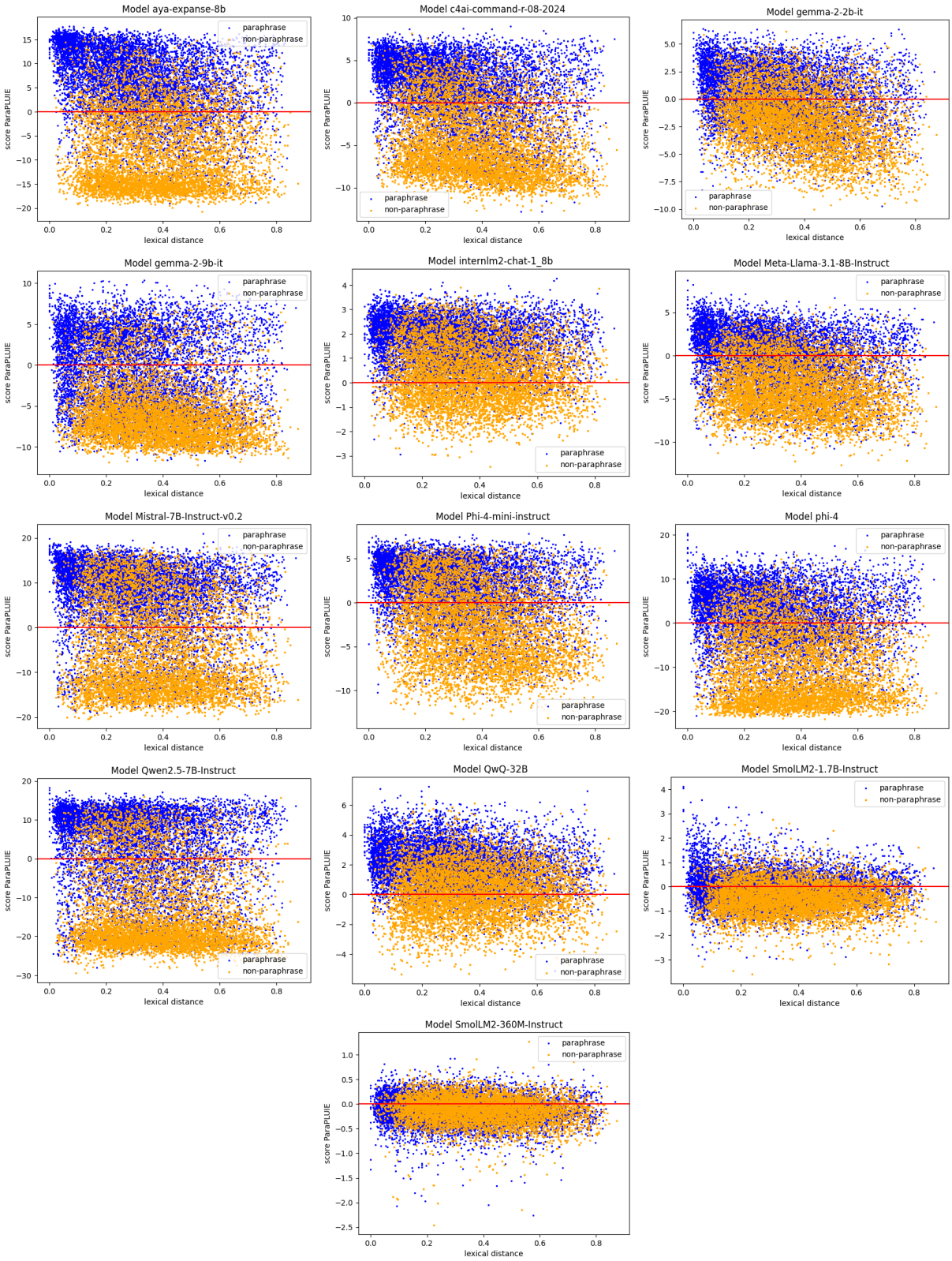


Figure 6: Distribution of the ParaPLUIE metric (y-axis) for all sentence pairs according to their edit distance (x-axis). The red line denotes the classification threshold, blue dots are paraphrases, and orange dots are non-paraphrases.