Mini-PLUIE: Efficient Paraphrase Evaluation by Distilling an LLM-As-A-Judge

Anonymous ACL submission

Abstract

Evaluating automatic paraphrase generation systems requires measuring the semantic proximity between many sentence pairs, a task for which recent Large Language Model (LLM)-based metrics such as ParaPLUIE have shown strong performance. However, the computational cost of such metrics becomes prohibitive at scale. We propose Mini-PLUIE, a lightweight siamese encoder trained by distillation from Aya-expanse-32b on six paraphrase datasets, which is two orders of magnitude faster than its teacher while remaining competitive. A systematic ablation over 36 configurations (decoder architecture, encoder, fine-tuning regime) reveals that the bottleneck lies in the *encoder*: when fine-tuned, all decoder variants are statistically equivalent and an MLP head suffices; when frozen, a single cross-attention layer is the practical optimum while deeper stacks degrade. Our distillation protocol and framework are released to enable the creation of task-specific lightweight evaluators for efficient semantic evaluation at scale.

1 Introduction

Following the definitions proposed by Mel'čuk (1997); Barzilay and McKeown (2001); Sekine (2005); Zhao et al. (2009); Fabre et al. (2021), the defining criterion of a paraphrase transformation is the preservation of meaning, which is inherently a fuzzy concept. Assessing whether two sentences share the same meaning therefore amounts to estimating their semantic distance, a difficult problem requiring deep linguistic knowledge and fine-grained understanding of context. While the lexicon carries meaning, relying solely on it is risky: two sentences can be lexically close while semantically distant, and vice versa (Zhang et al., 2019).

Automatic paraphrase generation systems rely on semantic proximity measures both to evaluate outputs and to compare systems at scale, often requiring the scoring of hundreds of thousands of pairs. Recent LLM-based metrics have achieved state-of-the-art performance on classification tasks (Gu et al., 2025). ParaPLUIE (Lemesle et al., 2026), for instance, directly correlates with human ranking judgements and is interpretable by design. However, their computational cost grows linearly with the size of the underlying LLM used and quickly becomes prohibitive. This raises a natural question: can a much smaller model learn to imitate such a judge closely enough to serve as a practical substitute?

In this paper, we propose Mini-PLUIE, a lightweight model trained by distillation to reproduce the ParaPLUIE scores of a LLM. Our main contributions are: **(1)** A systematic ablation of 36 configurations across decoder architecture, sentence encoder, and fine-tuning regime, identifying a simple Multi-Layer Perceptron (MLP) head over a fine-tuned BERT-base as optimal. **(2)** An extensive evaluation of Mini-PLUIE characteristics and performance, including a carbon cost estimation that further motivates the usage of such a distilled system for large-scale evaluation. **(3)** A new self-contained distillation framework (annotation, training, evaluation)[1] to create a ready-to-use task-specific version of Mini-PLUIE. We show that Mini-PLUIE is competitive and correlated with its teacher while being $\times\mathbf{175}$ faster and $\times\mathbf{280}$ smaller.

2 Related Work

To assess the semantic proximity of two sentences, several strategies are available. Historical metrics rely on lexical comparison such as Levenshtein distance (Levenshtein, 1965), BLEU (Papineni et al., 2002), or METEOR (Banerjee and Lavie, 2005), which struggle by construction with simple transformations such as negation. The Transformer architecture (Vaswani

[1] https://anonymous.url

et al., 2017) enabled embedding-based metrics such as BERTScore (Zhang et al., 2020) and Sentence-BERT (Reimers and Gurevych, 2019). They produce continuous scores but rely on subsentential alignment without a global view of the sentence, and fail on examples that are trivial for humans, as shown by Zhang et al. (2019) and Tytgat et al. (2024). These scores are hard to interpret and tend to be uniformly high regardless of actual semantic relation, due to anisotropy (Ethayarajh, 2019; Li et al., 2020). Proposed remedies include post-hoc whitening (Su et al., 2021) and contrastive fine-tuning (Gao et al., 2021) or multiple encoders (Humeau et al., 2019), but neither have proven effective in real-world scenarios on adversarial benchmarks such as PAWS (Zhang et al., 2019). Embedding-based metrics also typically require threshold calibration to separate paraphrases from non-paraphrases. Feature-matching approaches such as TinyBERT (Jiao et al., 2020) or DistilBERT (Sanh et al., 2019) are designed to compress teachers by aligning hidden states but therefore the student model must share the teacher architecture.

Instruction-tuned LLMs have profoundly transformed evaluation paradigms. Using an LLM-as-a-judge, leveraging its reasoning abilities to interpret an instruction and evaluate a text (Brown et al., 2020), has become a standard in many NLP tasks (Gu et al., 2025; Doostmohammadi et al., 2024). However, extracting a decision from free-form generation introduces noise and ambiguity, typically yielding a binary score at non-negligible computational cost. Several works distil LLM judges into smaller generative models. Kim et al. (2023) train Prometheus, a 13B evaluator fine-tuned on GPT-4 feedback with custom score rubrics. Zhu et al. (2023) fine-tune JudgeLM at 7B–33B to mimic GPT-4 open-ended judgements. Wang et al. (2023) follow a similar strategy for instruction-tuning comparison. All three keep the generative paradigm and therefore remain slow at inference.

To overcome the limitations of LLM-as-a-judge, Lemesle et al. (2025) used the perplexity of an LLM as a confidence score making it directly interpretable. Their metric ParaPLUIE is up to 8 times faster than a standard LLM-as-a-judge setup on the same model while being at least as accurate and easily personalisable (Lemesle et al., 2026). However, the quality of ParaPLUIE scales logarithmically with the size of the underlying LLM, so it remains too costly for large-scale use (or as a training signal). We propose to overcome this limitation by distilling the confidence score of a large LLM into a much smaller siamese encoder.

3 Mini-PLUIE

Mini-PLUIE is a siamese sentence encoder paired with a lightweight regression decoder. Both sentences s_1 and s_2 are independently embedded by the same pretrained encoder E, yielding two sequences of token embeddings $E(s_1)$ and $E(s_2)$. The decoder then maps these representations to a scalar score approximating a ParaPLUIE score.

We explore three decoder variants of increasing capacity: (i) **MLP**, which pools each sequence into a fixed-size vector, concatenates them, and applies a regression head; (ii) **TF Enc**, which applies a single Transformer Encoder layer with cross-attention over the concatenated token-sequence representations before pooling; and (iii) **TF Enc ZIP**, which interlaces the two token sequences before the same cross-attention step, enabling finer-grained alignment. We also vary the backbone encoder (BERT-base, ModernBERT, all-mpnet-base-v2, all-MiniLM-L6-v2) and the fine-tuning regime (frozen or fully fine-tuned), for a total of 36 configurations evaluated. We report results for the BERT-base encoder variant in the main paper; details of all configurations are available in Appendix A.

We train Mini-PLUIE by distillation against the ParaPLUIE scores produced by Aya-expanse-32b (Dang et al., 2024), used as the teacher model. Training minimises the mean squared error between teacher and student scores:

$$\mathcal{L} = \frac{1}{b} \sum_{i=1}^{b} \left(S_i^{\text{teacher}} - \hat{S}_i^{\text{student}} \right)^2$$

We use batches of 512 pairs and the AdamW optimiser with a fixed learning rate of 10^{-7}. Because the loss is applied directly on the continuous ParaPLUIE scores, the student preserves the interpretable threshold of the teacher and does not require any calibration at inference. This training objective encourages the student to reproduce the sentence-level semantic sensitivity of its ParaPLUIE teacher.

4 Experiments and Results

We use six English binary annotated paraphrase datasets (Table 1), splitting each into train, dev and

test while preserving its original paraphrase/non-paraphrase ratio. PAWS-Wiki, its automatically annotated part (PAWS-Wiki-Noisy), and PAWS (Zhang et al., 2019) dominate in volume and target adversarial cases in which surface similarity is decoupled from meaning. HC, LLM, MCPG (Lemesle et al., 2025) and MRPC (Dolan and Brockett, 2005) provide more diverse and challenging examples, at the cost of a much smaller size.

Dataset	Train	Dev	Test	Para.
PAWS-Wiki	70 238	8 779	8 781	32%
PAWS-Wiki-Noisy	524 520	65 566	65 566	50%
PAWS	6 399	800	801	56%
HC	160	20	20	50%
LLM	461	58	59	78%
MCPG	116	14	16	69%
MRPC	4 640	580	581	67%
Total	606 534	75 817	75 824	48%

Table 1: Number of sentence pairs per dataset and proportion of positive (paraphrase) pairs. Overall, the combined dataset is near-balanced with 48% positive pairs.

For each test pair, a model emits a score; we binarise it using either the sign (for ParaPLUIE and Mini-PLUIE) or a threshold tuned on the training set to maximise F1, and report accuracy (**Acc**), precision (**P**), recall (**R**), and **F1**. We also measure agreement with the teacher (Aya-expanse-32b) via ranking and classification metrics, and assess inference speed and carbon cost.

4.1 Architecture ablation

We compare three decoder architectures (MLP, TF Enc, TF Enc ZIP) on top of frozen and unfrozen encoders across 36 configurations (full results and details in Appendix A). The key factor is whether the encoder is fine-tuned: freezing it costs up to 30 accuracy points (Table 2). When the encoder is fine-tuned, all three decoders are statistically indistinguishable. When the encoder is frozen, a single cross-attention layer recovers up to $+9$ F1 points over MLP, but deeper stacks degrade, acting as a compensation mechanism for missing encoder adaptation. These findings translate into two concrete design rules. When encoder fine-tuning is feasible, use the simplest and fastest decoder: an MLP head is statistically equivalent to the cross-attention variants and cheaper at inference. On the other hand, if fine-tuning is infeasible, use a single cross-attention layer as a decoder. In the remainder of this paper, we use the MLP unfrozen variant with 1 layer as the Mini-PLUIE configuration.

Decoder	Freeze	Acc	F1	P	R	r
MLP_{114M}	no	.84	.85	.78	.94	.89
TF Enc_{125M}	no	.84	.85	.78	.93	.89
TF Enc ZIP_{125M}	no	.84	.85	.78	.93	.90
MLP_{114M}	yes	.62	.58	.62	.54	.50
TF Enc_{125M}	yes	.68	.67	.66	.68	.61
TF Enc ZIP_{125M}	yes	.68	.67	.66	.68	.61
MLP (5L)_{131M}	yes	.63	.59	.63	.55	.51
$\text{TF Enc (5L)}_{188M}$	yes	.56	.44	.57	.36	.35
$\text{TF Enc ZIP (5L)}_{188M}$	yes	.60	.53	.60	.48	.40

Table 2: Aggregated test performance for 9 Mini-PLUIE configurations. All bootstrap 95% confidence intervals (accuracy, F1, P, R) are $< 10^{-2}$. r is the Pearson correlation with the teacher (Aya-expanse-32b). "5L" denotes the 5-layer decoder variant; all others use 1 layer. Subscripts indicate the number of parameters.

4.2 Classification performance

Mini-PLUIE reaches .84 accuracy and .85 F1 on the aggregated test set (Table 3), outperforming all baselines including Mistral ($+.11$ points) and closing most of the gap to the teacher ($-.02$ points). All differences are significant ($p < 10^{-4}$, paired bootstrap). Lexical and embedding baselines show high recall but low precision; they tend to classify all pairs as paraphrases. Embedding baselines perform even more poorly; this is discussed in Appendix B.

Model	Acc	F1	P	R	ΔAcc
BLEU	.65	.69	.60	.81	$-.20$
METEOR	.52	.65	.50	.94	$-.32$
Levenshtein	.63	.67	.58	.80	$-.22$
BERTScore_{110M}	.53	.65	.51	.91	$-.31$
$\text{Sentence-BERT}_{110M}$	.48	.65	.48	**1.00**	$-.36$
SmolLM2_{135M}	.52	.00	.40	.00	$-.32$
$\text{Qwen2.5}_{1.5B}$	.71	.74	.65	.85	$-.13$
Mistral_{7B}	.73	.75	.67	.86	$-.11$
Aya-expanse_{32B}	**.86**	**.87**	**.82**	.92	$+.02$
Mini-PLUIE_{114M}	.84	.85	.78	.94	—

Table 3: Aggregated performance on the full test set. All bootstrap 95% confidence intervals (Acc, F1, P, R) are $< 10^{-2}$. LLM systems (SmolLM2, Qwen2.5, Mistral, aya-32B) are used in ParaPLUIE. ΔAcc is the difference between the model accuracy and Mini-PLUIE (p-value $< 10^{-4}$).

4.3 Agreement with the teacher

We measure how closely each system reproduces the teacher's ranking and binary decisions, using Spearman ρ, Kendall τ, and Cohen's κ (Table 4). Mini-PLUIE reaches $r = .89$, $\tau = .69$,

$\kappa = .79$, well ahead of the next-best baseline (Mistral: $r = .77$, $\tau = .57$, $\kappa = .62$). Mini-PLUIE is the only system to combine high ranking correlation and high classification agreement, surpassing Mistral and Qwen2.5 on both axes despite orders-of-magnitude fewer parameters. Lexical and embedding-based metrics score near zero on κ despite non-negligible rank correlations, confirming that their score distributions are too misaligned with the teacher's to be binarised with any fixed threshold. Disagreements between teacher and Mini-PLUIE are discussed in Appendix C.

System	r	ρ	τ	κ
BLEU	+.34	+.37	+.25	+.27
METEOR	+.14	+.16	+.11	+.07
Levenshtein	+.35	+.40	+.27	+.25
BERTScore_{110M}	+.26	+.32	+.21	+.11
$\text{Sentence-BERT}_{110M}$	+.21	+.35	+.24	+.00
SmolLM2_{135M}	−.01	−.02	−.01	−.00
$\text{Qwen2.5}_{1.5B}$	+.74	+.75	+.55	+.57
Mistral_{7B}	+.77	+.77	+.57	+.62
Mini-PLUIE_{114M}	**+.89**	**+.88**	**+.69**	**+.79**

Table 4: Agreement with the teacher across all systems. Pearson r and Spearman ρ measure correlation with the teacher's continuous scores; Kendall τ measures ranking concordance; Cohen κ measures classification agreement.

4.4 Inference speed

We conduct a strict iso-hardware comparison on a single NVIDIA H200 NVL. The teacher can only be evaluated at batch size 1 and scores 12.8 pairs per second. Mini-PLUIE at batch size 1 already achieves 346 pairs/s, a $\times\mathbf{27}$ speedup. At its optimal batch size of 64 (Figure 1), it reaches 2,231 pairs/s, a $\times\mathbf{175}$ speedup over the teacher. Crucially, this comparison is conservative by design: the H200 NVL is among the most favourable hardware available for large-LLM inference. On commodity hardware the gap would be substantially wider. This motivates its adoption in large-scale evaluation pipelines. Results for all models are available in Appendix D.

4.5 Carbon cost

We quantify the environmental trade-off of Mini-PLUIE following Lannelongue et al. (2021). Calculation details are available in Appendix E. The total distillation cost has two components: scoring all pairs with the teacher and fine-tuning Mini-PLUIE. The break-even point is $N^* \approx$ 1,58M pairs (Figure 2). Beyond N^*, every evaluation run with Mini-PLUIE instead of ParaPLUIE represents a carbon saving, a threshold quickly reached in a large-scale evaluation scenario. Concretely, scoring 1M pairs costs 1.14 $kgCO_2eq$ with aya-expanse-32b and only 6.5 gCO_2eq with Mini-PLUIE, a reduction of $\times\mathbf{175}$.

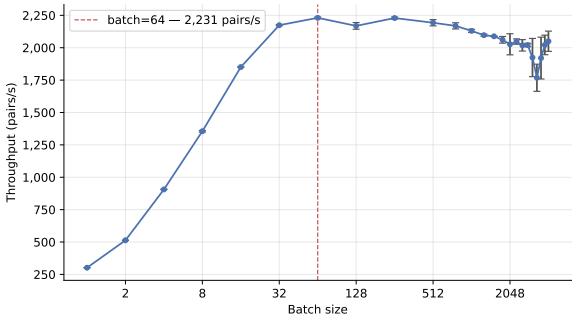

Figure 1: Throughput of Mini-PLUIE as a function of batch size on an NVIDIA H200 NVL. Saturation occurs at batch 64. Error bars show 95% confidence intervals.

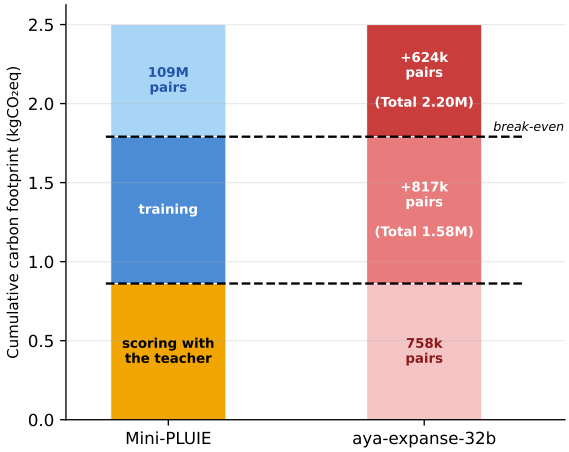

Figure 2: Carbon cost of Mini-PLUIE compared to its teacher. The dashed lines mark the boundary between the two fixed-cost components of Mini-PLUIE. The number of pairs scored in the carbon budget is indicated; counts are cumulative across sections.

5 Conclusion

We introduce Mini-PLUIE, a lightweight distilled approximation of ParaPLUIE that is two orders of magnitude faster than its teacher while remaining competitive, and release a ready-to-use distillation framework to create task-specific versions. A systematic ablation identifies the encoder, not the decoder, as the bottleneck of LLM-judge distillation.

Future work includes using Mini-PLUIE as a training signal for paraphrase generation models.

Limitations

The present study reports results at the training budget used during the ablation sweep (up to 300 epochs, early-stopped on the development loss). Several configurations, most visibly `TF Enc` with `BERT-base` unfrozen, show test-loss curves that have not fully plateaued at the point of early stopping (see Appendix F). Extending training for these configurations could slightly change the results; however, all encoders follow the same training trends, so the relative ordering of configurations is unlikely to shift.

Teacher ceiling. As with any distillation approach, the quality of `Mini-PLUIE` is bounded by that of its teacher. `Aya-expanse-32b` reaches 0.86 accuracy on our test set, meaning 14% of its decisions deviate from the reference labels, and `Mini-PLUIE` inherits both the compression loss and these teacher errors. Interpreting this figure is not straightforward: the test set is dominated by `PAWS-Wiki-Noisy`, whose labels are automatically generated rather than human-annotated, and the teacher achieves higher accuracy on the human-annotated `PAWS-Wiki` split than on the noisy one; so some apparent errors may in fact reflect label noise rather than genuine mistakes. A detailed breakdown (Appendix C) shows that 9.5% of test pairs are misclassified by both systems, directly reflecting the teacher ceiling, while on the 10.3% of pairs where they disagree, the teacher is correct in 59% of cases and `Mini-PLUIE` in 41%. Improving the teacher, through a stronger model or better prompting, could directly raise the student's ceiling.

Language restriction. All training and evaluation data are drawn from English paraphrase benchmarks with specific distributional properties (adversarial word-order permutations in `PAWS`, crowd-sourced sentence pairs in `MRPC`). Performance on other languages or specialised domains such as biomedical, legal, or code is unknown and could degrade.

Training objective. The goal of `Mini-PLUIE` is to distil the teacher's semantic knowledge of paraphrase similarity, including its graded confidence; fine-tuning directly on reference labels could increase the observed accuracy. Accuracy against human labels is therefore only a partial measure; what matters more is how faithfully the student mirrors the teacher's ordering and agreement structure across the full score range. To isolate the effect of the loss function from that of the training data, we train the best configuration (`MLP`, `BERT-base`, no freeze, one layer) under two alternative regimes and find that MSE distillation seems to transfer the teacher's semantic structure more faithfully than direct binary supervision (Appendix G).

Ethical considerations

Reproducibility. `ParaPLUIE` baselines were scored using the official implementation available at https://gitlab.inria.fr/expression/paraphrase-generation-evaluation-powered-by-an-llm-a-semantic-metric-not-a-lexical-one-coling-2025, the full prompt is available in Appendix H. All underlying models were downloaded on 2026-05-11 from the Hugging Face Hub: `Aya-expanse-32b` (Dang et al., 2024), `Mistral-7B-Instruct-v0.3` (Jiang et al., 2023), `Qwen2.5-1.5B-Instruct` (Qwen et al., 2025), and `SmolLM2-135M-Instruct` (Allal et al., 2025). Sentence encoder checkpoints (`BERT-base-uncased`, `all-MiniLM-L6-v2`, `all-mpnet-base-v2`, `ModernBERT-base`) were retrieved from the same hub under their respective licences. Training was conducted on a single AMD MI300A; speed benchmarks were run on a single NVIDIA H200 NVL. All `Mini-PLUIE` configurations trained in this study are released[2].

Carbon impact. LLMs are now widely used across scientific fields, serving as baselines, automatic annotators, and data generators. Their environmental cost is significant, and we believe the research community should report it systematically, not only model training, but the full study cost: fine-tuning, baseline inference, and scoring. We therefore provide a detailed carbon accounting of each computational phase of this study in Appendix E.

References

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.

[2] https://anonymous.url

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, 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.

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.

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

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.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65. Association for Computational Linguistics.

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.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6894–6910. Association for Computational Linguistics.

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.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Real-time inference in multi-sentence tasks with deep pretrained transformers. volume abs/1905.01969.

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.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. *Preprint*, arXiv:2310.08491.

Loïc Lannelongue, Jason Grealey, and Michael Inouye. 2021. Green algorithms: Quantifying the carbon footprint of computation. *Advanced Science*, 8(12):2100707.

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.

Quentin Lemesle, Léane Jourdan, Daisy Munson, Pierre Alain, Jonathan Chevelu, Arnaud Delhay, and Damien Lolive. 2026. *-pluie: Personalisable metric with llm used for improved evaluation. *Preprint*, arXiv:2602.15778.

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

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130. Association for Computational Linguistics.

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. *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.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster cheaper and lighter. volume abs/1910.01108.

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

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *CoRR*, abs/2103.15316.

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.

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. 30.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2023. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *Preprint*, arXiv:2306.05087.

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 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.

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.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *Preprint*, arXiv:2310.17631.

A Ablation study

We trained 36 configurations of `Mini-PLUIE` crossing three decoder architectures (`MLP`, `TF Enc`, `TF Enc ZIP`), four pretrained encoders (`BERT-base-uncased`, `ModernBERT-base`, `all-mpnet-base-v2`, `all-MiniLM-L6-v2`), two fine-tuning regimes (frozen vs. fully fine-tuned encoder) and two decoder depths (1 or 5 layers).

All three decoders operate on the token-level representations produced by the encoder. `MLP` concatenates the two sentence representations into a $2d$-dimensional vector, passes it through a feed-forward network (hidden size h) followed by layer normalisation, then reduces to a scalar via mean pooling and a linear projection. `TF Enc` replaces the feed-forward network with one or more standard Transformer encoder layers (8 attention heads, feed-forward dimension h), applied to the concatenated sequence. `TF Enc ZIP` uses the same Transformer encoder but the two token sequences are intertwined before processing, so that attention can attend across both sentences at every layer. For `BERT-base`, `ModernBERT` and `mpnet`: $d = 768$, $h = 2{,}048$; for `MiniLM-L6`: $d = 384$, $h = 1{,}024$.

Per-configuration results are reported in Table 5. The key factor is whether the encoder is fine-tuned: freezing it costs up to 30 accuracy points. When

Decoder	Encoder	Freeze	Layers	HC	LLM	MCPG	MRPC	PAWS	PAWS-W	PAWS-W-N	Acc	F1	P	R	r
MLP	BERT-base	—	1	0.55	0.61	0.75	**0.71**	0.90	0.87	0.84	0.84	0.85	0.78	0.94	0.89
TF Enc		—	1	0.35	0.61	0.75	0.70	0.90	0.87	0.84	0.84	0.85	0.78	0.93	0.89
TF Enc ZIP		—	1	0.35	0.64	0.75	0.70	0.90	0.87	0.84	0.84	0.85	0.78	0.93	**0.90**
MLP		✓	1	0.65	0.46	0.69	0.55	0.68	0.70	0.61	0.62	0.58	0.62	0.54	0.50
TF Enc		✓	1	0.55	0.63	**0.88**	0.62	0.74	0.73	0.67	0.68	0.67	0.66	0.68	0.61
TF Enc ZIP		✓	1	0.45	0.54	**0.88**	0.60	0.72	0.73	0.67	0.68	0.67	0.66	0.68	0.61
MLP		✓	5	0.55	0.42	0.75	0.55	0.70	0.71	0.61	0.63	0.59	0.63	0.55	0.51
TF Enc		✓	5	**0.75**	0.53	0.44	0.55	0.60	0.68	0.55	0.56	0.44	0.57	0.36	0.35
TF Enc ZIP		✓	5	**0.75**	0.56	0.62	0.56	0.62	0.69	0.58	0.60	0.53	0.60	0.48	0.40
MLP	ModernBERT	—	1	0.55	0.44	0.81	0.56	0.86	0.83	0.79	0.79	0.79	0.77	0.81	0.80
TF Enc		—	1	0.30	0.47	0.75	0.56	0.86	0.82	0.77	0.78	0.78	0.75	0.81	0.77
TF Enc ZIP		—	1	0.35	0.44	0.56	0.52	0.76	0.80	0.69	0.70	0.63	0.77	0.53	0.72
MLP		✓	1	0.50	0.36	0.75	0.54	0.71	0.73	0.66	0.67	0.63	0.68	0.58	0.57
TF Enc		✓	1	0.50	0.58	0.62	0.61	0.76	0.75	0.69	0.70	0.70	0.67	0.74	0.63
TF Enc ZIP		✓	1	0.60	0.49	0.69	0.58	0.74	0.75	0.69	0.70	0.70	0.68	0.72	0.62
MLP		✓	5	0.55	0.39	0.50	0.50	0.72	0.74	0.66	0.67	0.62	0.68	0.57	0.57
TF Enc		✓	5	0.65	0.44	0.38	0.55	0.60	0.70	0.57	0.59	0.47	0.62	0.37	0.42
TF Enc ZIP		✓	5	0.70	0.34	0.50	0.51	0.65	0.71	0.59	0.60	0.46	0.66	0.35	0.44
MLP	MiniLM-L6	—	1	0.50	0.59	**0.88**	0.69	0.90	0.86	0.84	**0.85**	**0.86**	0.78	**0.95**	0.88
TF Enc		—	1	0.40	0.63	0.81	0.67	0.88	0.86	0.83	0.83	0.84	0.77	0.93	0.86
TF Enc ZIP		—	1	0.35	0.61	0.81	0.67	0.88	0.87	0.84	0.84	0.85	0.78	0.93	0.87
MLP		✓	1	0.55	0.47	0.81	0.51	0.68	0.68	0.62	0.63	0.60	0.63	0.57	0.47
TF Enc		✓	1	0.50	0.49	0.69	0.50	0.70	0.69	0.63	0.64	0.60	0.63	0.57	0.51
TF Enc ZIP		✓	1	0.60	0.37	0.62	0.52	0.68	0.69	0.63	0.63	0.60	0.63	0.57	0.51
MLP		✓	5	0.50	0.46	0.75	0.54	0.66	0.68	0.60	0.61	0.57	0.60	0.54	0.42
TF Enc		✓	5	0.60	0.39	0.69	0.59	0.69	0.70	0.66	0.67	0.66	0.65	0.67	0.51
TF Enc ZIP		✓	5	0.60	0.42	0.75	0.56	0.71	0.69	0.66	0.66	0.65	0.64	0.66	0.51
MLP	mpnet	—	1	0.55	0.64	0.75	0.68	**0.91**	0.87	**0.85**	**0.85**	**0.86**	**0.79**	0.94	**0.90**
TF Enc		—	1	0.45	0.66	0.75	0.67	0.90	**0.88**	0.84	**0.85**	0.85	**0.79**	0.93	**0.90**
TF Enc ZIP		—	1	0.40	**0.68**	0.75	0.68	0.90	**0.88**	**0.85**	**0.85**	**0.86**	**0.79**	0.93	0.89
MLP		✓	1	0.55	0.41	0.50	0.49	0.61	0.60	0.54	0.54	0.46	0.54	0.40	0.32
TF Enc		✓	1	0.50	0.49	0.81	0.55	0.61	0.61	0.55	0.56	0.51	0.54	0.48	0.34
TF Enc ZIP		✓	1	0.50	0.41	0.75	0.52	0.63	0.60	0.55	0.56	0.51	0.55	0.48	0.36
MLP		✓	5	0.65	0.44	0.50	0.44	0.56	0.62	0.51	0.52	0.38	0.50	0.31	0.22
TF Enc		✓	5	0.35	0.46	0.56	0.51	0.56	0.62	0.50	0.52	0.33	0.49	0.25	0.18
TF Enc ZIP		✓	5	0.70	0.42	0.50	0.47	0.56	0.63	0.50	0.52	0.31	0.49	0.22	0.18

Table 5: Per-test subdataset accuracy and aggregated metrics for the 36 Mini-PLUIE configurations. Freeze: ✓ = encoder frozen, — = encoder fine-tuned. r is the Pearson correlation with the teacher (Aya-expanse-32b). The configuration (MLP, BERT-base, unfrozen, 1 layer) is highlighted in Table 2. PAWS-W and PAWS-W-N stand for PAWS-Wiki and PAWS-Wiki-Noisy, respectively. All bootstrap 95% confidence intervals (Acc, F1, P, R) are $< 10^{-2}$.

fine-tuned, all three decoders are statistically indistinguishable on aggregated accuracy. mpnet configurations achieve the highest scores (Acc up to .85, r up to .90), slightly above BERT-base and MiniLM-L6. MiniLM-L6 ($d = 384$) reaches performance comparable to BERT-base ($d = 768$) while operating at half the embedding dimension, making it an appealing lighter alternative. However, all-mpnet-base-v2 was pre-trained on datasets that overlap with PAWS (Reimers and Gurevych, 2019), raising a data contamination risk; we therefore use BERT-base-uncased as the encoder throughout the main paper.

Table 6 reports throughput for all configurations at batch size 1. MiniLM-L6-based models are the fastest owing to their smaller embedding dimension ($d = 384$); ModernBERT-based models are the slowest. Within each encoder group, the three decoder architectures and two depth settings have limited impact on throughput: the bottleneck is the encoder forward pass.

B Per-dataset results for all baselines

Table 7 reports per-subdataset accuracy and aggregated metrics for all baselines and Mini-PLUIE. Thresholds for lexical and embedding metrics are tuned on the training split (best F1); ParaPLUIE LLM judges and Mini-PLUIE use the sign of the score.

Several patterns emerge, lexical metrics collapse on HC and LLM (accuracy $\leq$ .40): paraphrases in these subsets do not share surface form with their reference, making n-gram overlap uninformative. Embedding baselines ($\text{BERTScore}_{\text{ModernBERT}}$, $\text{Sentence-BERT}_{\text{MiniLM-L6}}$, $\text{Sentence-BERT}_{\text{mpnet-base}}$) exhibit a degenerate threshold on PAWS (R= 1.00, P$\approx$ .48): the train-tuned threshold classifies every pair as a paraphrase, as already discussed in Appendix B. SmolLM2 degenerates in the opposite direction (F1= .00,

Decoder	Encoder	Freeze	Layers	pairs/s	× teacher
MLP	BERT-base	—	1	346	×27
TF Enc		—	1	267	×21
TF Enc ZIP		—	1	270	×21
MLP		✓	1	285	×22
TF Enc		✓	1	289	×23
TF Enc ZIP		✓	1	278	×22
MLP		✓	5	284	×22
TF Enc		✓	5	254	×20
TF Enc ZIP		✓	5	276	×22
MLP	ModernBERT	—	1	185	×15
TF Enc		—	1	157	×12
TF Enc ZIP		—	1	167	×13
MLP		✓	1	182	×14
TF Enc		✓	1	151	×12
TF Enc ZIP		✓	1	147	×12
MLP		✓	5	143	×11
TF Enc		✓	5	151	×12
TF Enc ZIP		✓	5	154	×12
MLP	MiniLM-L6	—	1	**567**	×**44**
TF Enc		—	1	426	×33
TF Enc ZIP		—	1	521	×41
MLP		✓	1	505	×40
TF Enc		✓	1	477	×37
TF Enc ZIP		✓	1	522	×41
MLP		✓	5	429	×34
TF Enc		✓	5	403	×32
TF Enc ZIP		✓	5	406	×32
MLP	mpnet	—	1	302	×24
TF Enc		—	1	245	×19
TF Enc ZIP		—	1	315	×25
MLP		✓	1	321	×25
TF Enc		✓	1	265	×21
TF Enc ZIP		✓	1	312	×24
MLP		✓	5	253	×20
TF Enc		✓	5	243	×19
TF Enc ZIP		✓	5	266	×21

Table 6: Throughput of all 36 `Mini-PLUIE` configurations at batch size 1 on a single NVIDIA H200 NVL. Freeze: ✓ = encoder frozen, — = encoder fine-tuned. Speedup is relative to `Aya-expanse-32b` (13 pairs/s).

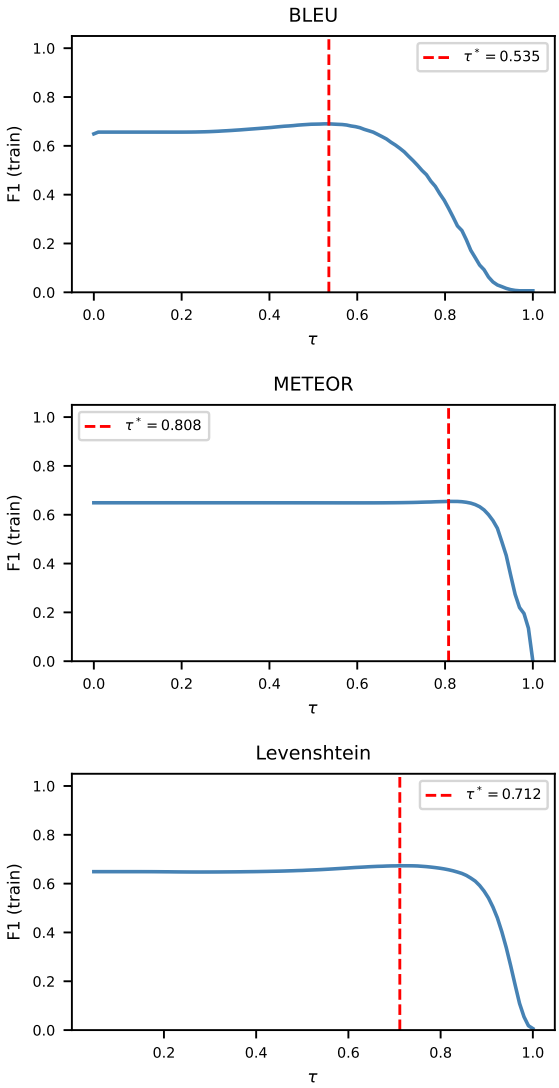

Figure 3: F1 vs. τ on the training split for lexical baselines (`BLEU`, `METEOR`, `Levenshtein`). The red dashed line marks τ^*, applied unchanged at test time in Table 3.

R= .00), abstaining from the paraphrase label entirely. Among LLMs, `Mistral` reaches perfect accuracy on `MCPG` and `Aya-expanse` on `HC`, while both struggle on `LLM` (.78), suggesting that LLM-generated paraphrases remain challenging even for large judges.

Threshold tuning of baselines. For the seven calibratable baselines (`BLEU`, `METEOR`, `Levenshtein`, `BERTScore` on `BERT-base-uncased` and `ModernBERT-base`, `Sentence-BERT` on `MiniLM-L6` and `mpnet`), the decision threshold τ^* used at test time is selected on the training split by maximising F1. The resulting curves are reported in Figures 3–5: each one shows F1 as a function of the decision threshold τ on train, with τ^* marked in red. The same τ^* is then applied unchanged on the test dataset to produce the rows of Table 3. For `Sentence-BERT`, the tuned threshold collapses to the minimum of the score range, classifying all pairs as paraphrases; this degenerate outcome reflects the absence of discriminative signal in cosine similarity on adversarial pairs.

C Teacher-Student Disagreement

To better understand the relationship between `Mini-PLUIE` and its teacher (`Aya-expanse-32b`), we perform a breakdown of their predictions against the reference labels on the full test set (75,824 pairs).

Shared errors (9.5%). The "both wrong" category directly reflects the teacher ceiling discussed in the Limitations section. Errors the teacher makes are inherited by the student through distillation.

Model	HC	LLM	MCPG	MRPC	PAWS	PAWS-W	PAWS-W-N	Acc	F1	P	R	r
BLEU	.35	.20	.75	.49	.64	.45	.68	.65	.69	.60	.81	+.34
METEOR	.35	.25	.75	.49	.50	.36	.54	.52	.65	.50	.94	+.14
Levenshtein	.40	.24	.75	.52	.62	.45	.65	.63	.67	.58	.80	+.35
$\text{BERTScore}_{\text{BERT-base}}$	.50	.31	.75	.47	.52	.37	.55	.53	.65	.51	.91	+.26
$\text{BERTScore}_{\text{ModernBERT}}$	.50	.75	.75	.69	.44	.32	.50	.48	.65	.48	**1.00**	+.21
$\text{Sentence-BERT}_{\text{MiniLM-L6}}$	.70	.80	.69	.70	.44	.32	.50	.48	.65	.48	**1.00**	+.15
$\text{Sentence-BERT}_{\text{mpnet-base}}$	.60	.78	.69	.69	.44	.32	.50	.48	.65	.48	**1.00**	+.21
SmolLM2	.50	.22	.31	.33	.56	.68	.50	.52	.00	.40	.00	−.01
Qwen2.5	.90	**.83**	.94	.70	.72	.73	.71	.71	.74	.65	.85	+.74
Mistral	.85	.78	**1.00**	**.71**	.75	.74	.73	.73	.75	.67	.86	**+.77**
Aya-expanse	**1.00**	.78	.94	**.71**	**.91**	**.91**	**.86**	**.86**	**.87**	**.82**	.92	—

Table 7: Per-test subdataset accuracy and aggregated metrics for all baselines. LLM systems (SmolLM2, Qwen2.5, Mistral, aya-32B) are used in ParaPLUIE. Thresholds for calibratable metrics (BLEU, METEOR, Levenshtein, BERTScore, Sentence-BERT) are tuned on the training split (best F1); ParaPLUIE and Mini-PLUIE use the sign of the score. r is the Pearson correlation with the teacher (Aya-expanse-32b). PAWS-W and PAWS-W-N stand for PAWS-Wiki and PAWS-Wiki-Noisy, respectively. All bootstrap 95% confidence intervals (Acc, F1, P, R) are $< 10^{-2}$.

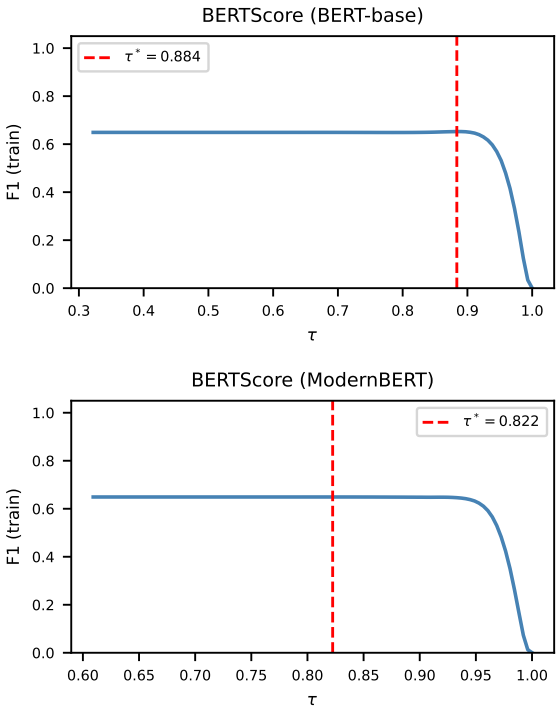

Figure 4: F1 vs. τ on the training split for BERTScore baselines. The red dashed line marks τ^*.

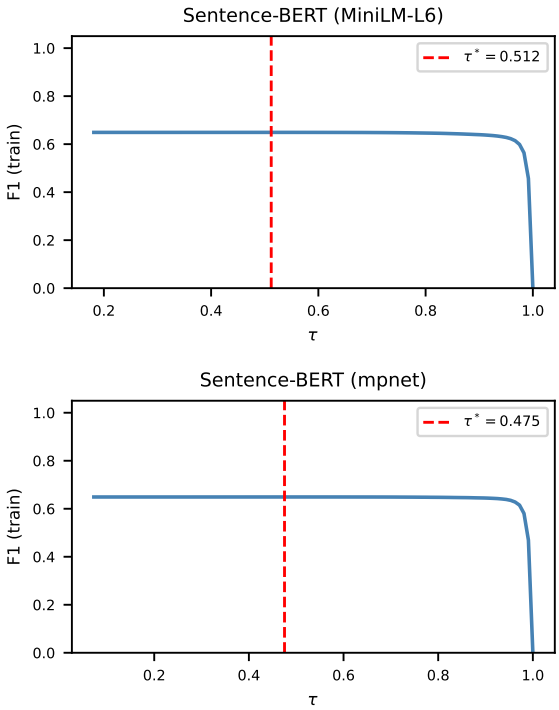

Figure 5: F1 vs. τ on the training split for Sentence-BERT baselines. The red dashed line marks τ^*.

Disagreements (10.3%). Among the 7,841 pairs where the two systems differ, the teacher is correct in 59% of cases and Mini-PLUIE in 41%. The modest gap, consistent with the accuracy differential (0.86 vs. 0.84), shows that Mini-PLUIE is not a noisy clone of the teacher. It seems to capture independent signal that occasionally lets it recover cases the teacher fails on.

Continuous agreement. The Pearson correlation between the continuous scores is $r = 0.894$ (p-value $< 10^{-4}$), closely matching the Spearman $\rho = 0.89$ reported in the main results and confirming that the score distributions are nearly linearly related.

Outcome	Count	%
Both correct	60,760	80.1%
Both wrong	7,223	9.5%
Teacher ✓ / Mini-PLUIE ×	4,627	6.1%
Teacher × / Mini-PLUIE ✓	3,214	4.2%
Total disagreements	7,841	10.3%

Table 8: Four-way breakdown of teacher vs. Mini-PLUIE predictions against human labels. The Mini-PLUIE configuration used is the MLP decoder with 1 layer and unfrozen BERT-base encoder.

D Inference speed of all systems

All measurements are conducted on a single NVIDIA H200 NVL following the protocol described in Section 4. Throughput is reported in pairs per second; higher is better. Lexical baselines run on CPU; all other systems run on GPU. At batch size 1, Mini-PLUIE (×27) already outperforms all GPU embedding baselines except Sentence-BERT$_{\text{MiniLM-L6}}$. BERTScore is surprisingly slow (×4–6); slower than SmolLM2$_{135\text{M}}$; due to the token-level alignment it performs on top of the encoder. LLM throughput scales poorly with model size: even the smallest judge (SmolLM2, 135M parameters) only reaches ×7, and Mistral$_{7\text{B}}$ drops to ×3.

System	pairs/s	× teacher
Lexical (CPU)		
Levenshtein	**741,881**	**×58,149**
BLEU	18,498	×1,450
METEOR	541	×42
Embedding (GPU, batch = 1)		
Sentence-BERT$_{\text{MiniLM-L6}}$	399	×31
Sentence-BERT$_{\text{mpnet}}$	209	×16
BERTScore$_{\text{BERT-base}}$	72	×6
BERTScore$_{\text{ModernBERT}}$	50	×4
LLM (GPU, batch = 1)		
SmolLM2$_{135\text{M}}$	86	×7
Qwen2.5$_{1.5\text{B}}$	77	×6
Mistral$_{7\text{B}}$	40	×3
Aya-expanse$_{32\text{B}}$	13	×1
Mini-PLUIE *(GPU)*		
Mini-PLUIE$_{114\text{M}}$ batch = 1	346	×27
Mini-PLUIE$_{114\text{M}}$ batch = 64	2,231	×175

Table 9: Throughput of all systems on a single NVIDIA H200 NVL. Speedup is relative to the teacher (Aya-expanse-32b, 13 pairs/s). Lexical baselines run on CPU; all other systems run on GPU. The Mini-PLUIE configuration used is the MLP unfrozen with 1 layer. All 95% confidence intervals on mean latency are below 0.1 ms per pair (worst case: METEOR, 57 μs).

E Carbon footprint details

To model the carbon footprint we used the formula proposed by Lannelongue et al. (2021):

$$C_{\text{total}} = t \times (P_{\text{GPU}} \times u + P_{\text{mem}}) \times \text{PUE} \times \text{CI}$$

where P_{GPU} is the GPU TDP (core power draw), u the utilisation rate, P_{mem} the memory power draw (0.3725 W/GB, Green Algorithms default), and PUE (Power Usage Effectiveness) the datacenter efficiency. All measures use the French grid Carbon Intensity (CI) (51.3 gCO_2eq/kWh) and a PUE of 1.56.

Break-even details. C_{distill} comprises two one-time costs measured on the hardware actually used. Fine-tuning Mini-PLUIE ran on a single AMD MI300A (TDP 350 W) for 156 epochs ($\approx$ 29.2 h), giving $C_{\text{train}} = 0.929$ kgCO_2eq. Scoring all 758,175 pairs (train, dev and test) with the teacher ran on an NVIDIA H200 NVL (TDP 600 W), giving $C_{\text{score}} = 0.862$ kgCO_2eq. The total distillation cost is $C_{\text{distill}} = C_{\text{train}} + C_{\text{score}} = 1.791$ kgCO_2eq.

At inference on the H200 NVL, the teacher costs $c_{\text{teacher}} = 1{,}136.5$ μgCO_2eq per pair and Mini-PLUIE costs $c_{\text{student}} = 6.5$ μgCO_2eq per pair. The break-even formula is:

$$N^* = \frac{C_{\text{distill}}}{c_{\text{teacher}} - c_{\text{student}}}$$

All study cost estimation. Table 10 reports the carbon footprint of each computational phase of this study. Training costs are estimated on the AMD MI300A (TDP 350 W) used for all runs; inference costs on the NVIDIA H200 NVL (TDP 600 W). The ablation over 38 configurations (36 configurations are described in Appendix A and the 2 others in Appendix G) dominates the total (46.4 out of 47.9 kgCO_2eq, i.e. 97%). The scoring of the 38 Mini-PLUIE configurations on the test set is negligible ($< 0.4\%$).

F Training curves

For each of the four pretrained encoders, we report the mean per-epoch train loss (dashed) and dev loss (solid) for all three architectures (MLP, TF Enc, TF Enc ZIP) crossed with the three fine-tuning regimes: encoder unfrozen with 1 layer (red, production configuration), encoder frozen with 1 layer (blue), and encoder frozen with 5 layers (green).

Phase	Hardware	$kgCO_2eq$	Hours
38 Mini-PLUIE configs (training)	MI300A	46.373	1457.8
Teacher scoring (758,175 pairs)	H200	0.862	16.5
Embedding baselines scoring (682,358 pairs)	H200	0.407	7.8
Other ParaPLUIE LLMs scoring (75,824 pairs)	H200	0.054	1.0
38 Mini-PLUIE configs (75,824 pairs)	H200	0.162	3.1
Total		**47.858**	**1486.3**

Table 10: Carbon footprint and runtime in hours per computational phase.

The dot marks the epoch at which the best test loss is reached (used for weight selection). Figures 7–10 show the MSE curves for each encoder; Figure 6 shows the BCE and BCE-filtered variants (Appendix G).

G Training objective

The goal of Mini-PLUIE is to distil the teacher's semantic knowledge of paraphrase similarity, including its graded confidence. Accuracy against labels is therefore only a partial measure; what matters is how faithfully the student mirrors the teacher's ordering and agreement structure across the full score range. To isolate the effect of the loss function from that of the training data, we train the best configuration (MLP, BERT-base, unfreeze, one layer) under two alternative regimes.

BCE (direct supervision). We replace the MSE loss with a binary cross-entropy loss trained directly on reference binary labels, keeping all other hyperparameters identical. The resulting model achieves higher aggregated accuracy (0.92 ± 0.002, F1 $= 0.92$) than the Mini-PLUIE (MSE) (0.84 ± 0.003, F1 $= 0.85$) (Table 11). However, this accuracy advantage does not reflect better knowledge transfer from the teacher: the BCE model's Spearman ranking correlation with the teacher drops to $\rho_S = +0.74$ (Kendall $\tau = +0.53$, Cohen $\kappa = +0.70$) vs. $\rho_S = +0.88$ for MSE ($\tau = +0.69$, $\kappa = +0.79$) (Table 12). BCE scores higher on labels precisely because it was trained on them, but in doing so it could diverge from the teacher's semantic structure.

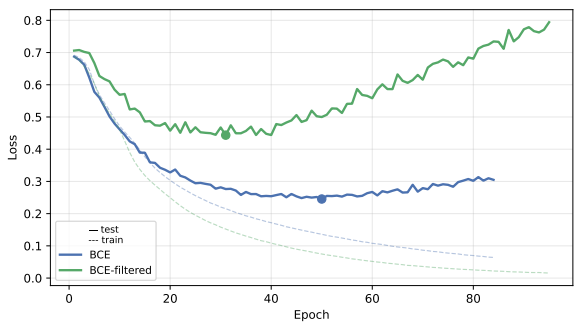

Figure 6: Training curves for the BCE and BCE-filtered variants (MLP, BERT-base, unfrozen, 1 layer). Both converge faster than MSE, as BCE optimises directly on binary labels rather than continuous teacher scores.

Model	Acc	F1	P	R
Mini-PLUIE MSE	.84	.85	.78	.94
Mini-PLUIE BCE	.92	.92	.90	.95
Mini-PLUIE BCE-filtered	.88	.88	.82	.95

Table 11: Classification performance of the three supervision regimes (human labels, threshold $= 0$, $n = 75{,}824$). BCE variants gain on aggregated accuracy but at the cost of semantic knowledge transfer (Table 12). The Mini-PLUIE configuration used is the MLP decoder with 1 layer and unfrozen BERT-base encoder.

Model	r	ρ	τ	κ
Mini-PLUIE MSE	$+.\mathbf{89}$	$+.\mathbf{88}$	$+.\mathbf{69}$	$+.\mathbf{79}$
Mini-PLUIE BCE	$+.77$	$+.74$	$+.53$	$+.70$
Mini-PLUIE BCE-filtered	$+.80$	$+.78$	$+.57$	$+.73$

Table 12: Teacher-agreement metrics for the three supervision regimes. MSE outperforms both BCE variants on all four metrics, even after controlling for label noise (BCE-filtered). The Mini-PLUIE configuration used is the MLP decoder with 1 layer and unfrozen BERT-base encoder.

BCE-filtered (fair comparison). To disentangle the loss function from label noise, we train another model with BCE loss on the subset of training pairs where the teacher's binary decision agrees with the human label, making supervision consistent with the teacher in both cases. Filtering partially closes the ranking gap ($\rho_S = +0.78$, $\tau = +0.57$, $\kappa = +0.73$) and narrows the accuracy advantage (0.88 ± 0.002) (Tables 11 and 12), confirming that label noise explains part of the BCE divergence. Yet all three ranking metrics remain substantially below MSE even after the noise is removed. Using BCE discards the very signal that encodes the teacher's semantic knowledge, its degrees of confidence. The MSE student inherits all of this and reproduces its errors consistently, not correcting them toward a different ground truth.

Pearson correlation with the teacher's continuous scores confirms the picture: $r = +0.89$ for

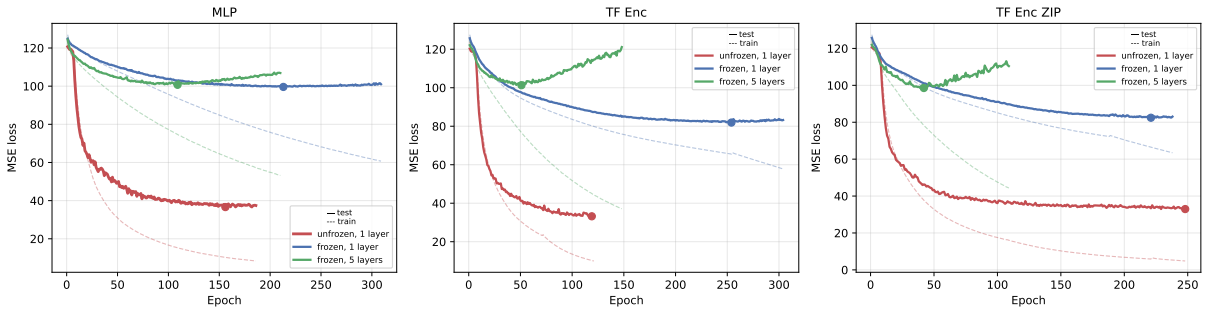

Figure 7: Training curves for the `BERT-base` encoder. The unfrozen variant (red) converges fastest and reaches the lowest test loss across all three architectures.

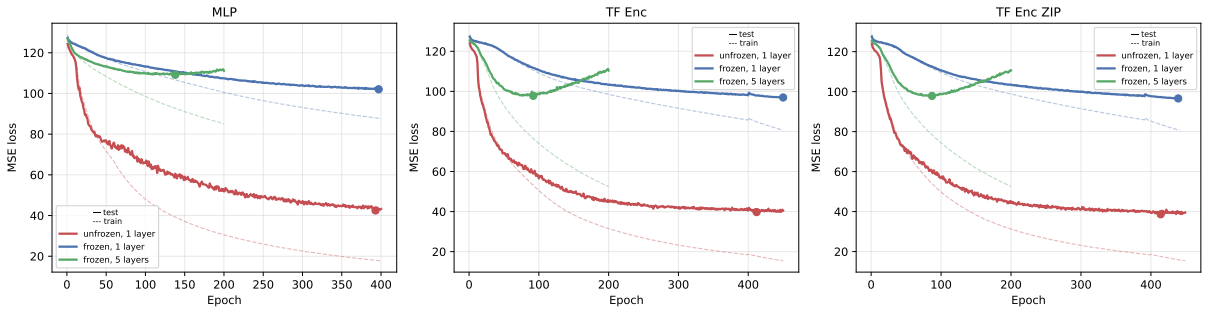

Figure 8: Training curves for the `MiniLM-L6` encoder.

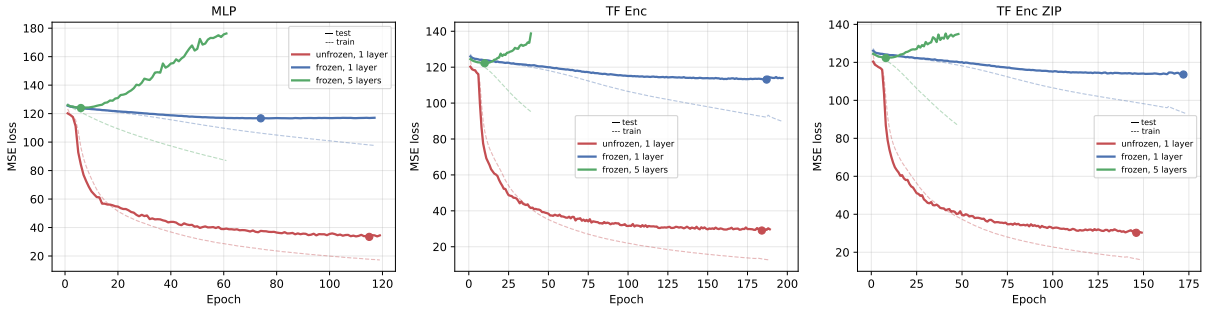

Figure 9: Training curves for the `mpnet` encoder.

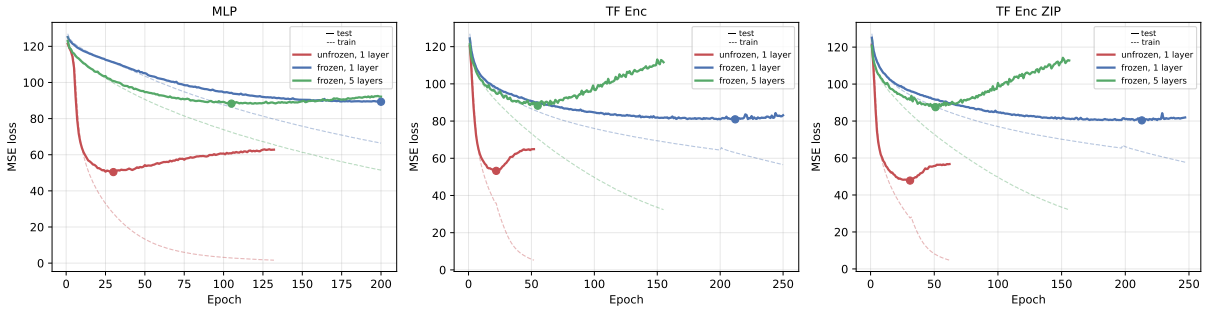

Figure 10: Training curves for the `ModernBERT` encoder. The instability of the unfrozen variant is visible across all architectures, consistent with the representation collapse reported in Section 4.

MSE vs. $+0.77$ and $+0.80$ for BCE (all labels and filtered). The gap persists after controlling for label noise, showing that soft-score distillation transfers the teacher's full semantic geometry, not merely its decision boundary.

H Prompt Template

The full prompt used for PARAPLUIE inference is provided below.

Prompt segment 1: PARAPLUIE prompt template

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

assistant="""Please provide the data 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: "{source}"; B: "{paraphrase}"""""
```