Turing team at PAN 2026: Adversarial Data Fine-tuned LLM for AI-Generated Text Detection

Notebook for the PAN Lab at CLEF 2026

Filip Milak[1], Sofia Fornaroli[1] and Miquel Sánchez Queralt[1]

[1]University of Copenhagen, Karen Blixens Vej 8, 2300 København, Denmark

Abstract

We present the Turing team's submission to the PAN 2026 Voight-Kampff Generative AI Detection shared task. Our approach combines adversarial data augmentation with a two-stage fine-tuning framework to build robust detectors of AI-generated text. We first assemble a diverse training corpus from four established datasets supplemented by self-created obfuscated data, generated by prompting seven state-of-the-art LLMs to rewrite human-authored texts using eight stylistic prompts targeting linguistic markers of human writing. Five open-weight LLMs ranging from 4B to 14B parameters are then fine-tuned in two stages: supervised fine-tuning on standard data followed by pairwise metric fine-tuning on adversarially augmented data, explicitly pushing human and AI text representations apart in the embedding space. The three best-performing models are combined with a TF-IDF Random Forest classifier into a weighted ensemble with an uncertainty gate. Out-of-distribution evaluation on a held-out GPT-5-mini test set, using both seen and unseen rewriting prompts with human texts from an entirely different source, confirms that pairwise metric fine-tuning generalises to unseen generators and prompt strategies without prompt-specific overfitting. On the official PAN 2026 hidden test set our system achieves a Mean score of 0.892, ROC-AUC of 0.957, and F0.5u of 0.952.

Keywords

PAN2026, Voight-Kampff Generative AI Detection 2026, Large Language Models, Data Augmentation, Adversarial Data, Obfuscation, Machine-generated Text Detection, Pairwise Metric Learning

1. Introduction

As large language models (LLMs) evolve rapidly and continuously, expanding into domains such as essay writing, coding, messaging, and scientific publishing, they have come to excel in both text generation and rewriting [1]. Alongside the immense opportunities these tools provide, they also present risks for misuse, including vulnerability detection [2], the spread of unintentional falsehoods and biased information [3, 4], and plagiarism [5]. Simultaneously, it is becoming increasingly difficult to distinguish between human-authored and artificial intelligence (AI) written text, which threatens the perceived authenticity of human communication. To achieve robust classification, researchers have developed various preventive and curative strategies, watermarking serves as a primary preventive method [6], whereas the detection of AI-generated text (AIGT) represents a curative approach [7]. However, automated AI text detection remains imperfect. When faced with out-of-distribution (OOD) data or small adversarial perturbations, such as paraphrasing [8], spelling injections [9], or homoglyph attacks [10], the accuracy of subsequent classification can be severely impaired.

The *Voight-Kampff AI Detection* task at PAN 2026 [11][12] addresses the increasingly sophisticated challenge of authorship attribution in the era of LLMs. Formulated as a binary classification problem, the task requires participants to submit a model that distinguishes between human-authored text ($y = 0$) and machine-generated content ($y = 1$).

The task is organized as a "builder-breaker" collaboration between the *Voight-Kampff* task at PAN and the ELOQUENT Lab. Within this ecosystem, PAN participants act as *builders*, developing robust detection systems, while ELOQUENT participants serve as *breakers*, investigating generative methods and obfuscation strategies designed to bypass state-of-the-art classifiers.

CLEF 2026 Working Notes, September 21 – 24, Jena, Germany

kgf646@alumni.ku.dk (F. Milak); rvj326@alumni.ku.dk (S. Fornaroli); nbl458@alumni.ku.dk (M. S. Queralt)

This notebook describes our adversarial data augmentation and two-stage fine-tuning approach, in which five open-weight LLMs ranging from 4B to 14B parameters are first fine-tuned on a standard corpus and then trained with a pairwise metric objective on the adversarial dataset. The three best-performing LLMs are combined with a TF-IDF Random Forest classifier into a weighted ensemble with an uncertainty gate, yielding a final Mean score of 0.930 on the PAN 2026 official test set.

2. Related Work

Prior to the widespread adoption of large language models, AI-generated text detectors primarily targeted n-gram models and Markov chain generators, whose statistical regularity made them relatively straightforward to identify using classical machine learning approaches [13]. As transformer architectures matured, more capable detectors emerged alongside them, including the RoBERTa-based classifier [14] and the GROVER detector [15], both designed to identify output from increasingly fluent generators. The public release of ChatGPT at the end of 2022 marked a qualitative shift in the field: as LLM-generated text became indistinguishable from human writing to casual readers, the demand for reliable detectors intensified substantially. Today, both open-source and proprietary detection systems have emerged, including GPTZero, Originality AI, and CopyLeaks [16].

Detection approaches broadly fall into four categories: *watermarking* as a preventive strategy, and three remedial strategies, *statistical and stylometric*, *transformer-based*, and *LLM-based*, which are sometimes combined into hybrid pipelines. Stylometric and statistical methods represent the earliest remedial techniques, valued for their interpretability and low computational cost. They exploit measurable differences between human and machine writing, such as the tendency of LLMs to produce sequences of unusually high token probability, captured by perplexity-based methods [17], or their characteristically low burstiness, reflected in a preference for uniformly long sentences [18]. Transformer-based methods have since become dominant, leveraging contextual representations for entropy analysis and probability curvature estimation, with BERT [19] and RoBERTa [20] serving as the most widely adopted backbones. More recently, LLMs themselves have been repurposed as detectors, exploiting their own understanding of generation patterns to classify output [21].

Despite these advances, obfuscation remains the most critical vulnerability facing current detection systems. Deliberately altering AI-generated text to evade classifiers through paraphrasing, homoglyph substitution, or structural simplification substantially degrades detector performance [22]. Paraphrasing attacks in particular have proven highly effective: Krishna et al. [23] demonstrated that even lightweight paraphrasers can reduce detection rates from over 70% to near chance level with minimal semantic distortion. Beyond post-hoc modification, LLMs can also be guided at generation time to produce detection-resistant text through carefully designed prompts [24], removing the need for a separate paraphrasing step entirely. At the character level, deep classification pipelines are inherently fragile: frameworks such as *TextBugger* [25] and *DeepWordBug* [26] have shown that minor perturbations (e.g. zero-width space insertions, visual lookalike substitutions) are sufficient to severely impair inference accuracy. Supervised pairwise learning has been shown to be an effective auxiliary objective during the fine-tuning of pre-trained language models, improving generalisation and robustness to noisy training data compared to cross-entropy loss alone [27]. Our work directly addresses this threat surface by incorporating adversarially augmented training data and pairwise metric learning to build robustness against both character-level and semantic-level obfuscation.

3. Methodology

Our approach proceeds in three stages: assembling a diverse training corpus, generating adversarially augmented data through prompt-based rewriting, and constructing held-out out-of-distribution (OOD) evaluation sets to assess generalisation to unseen generators and prompt strategies.

3.1. Training Dataset

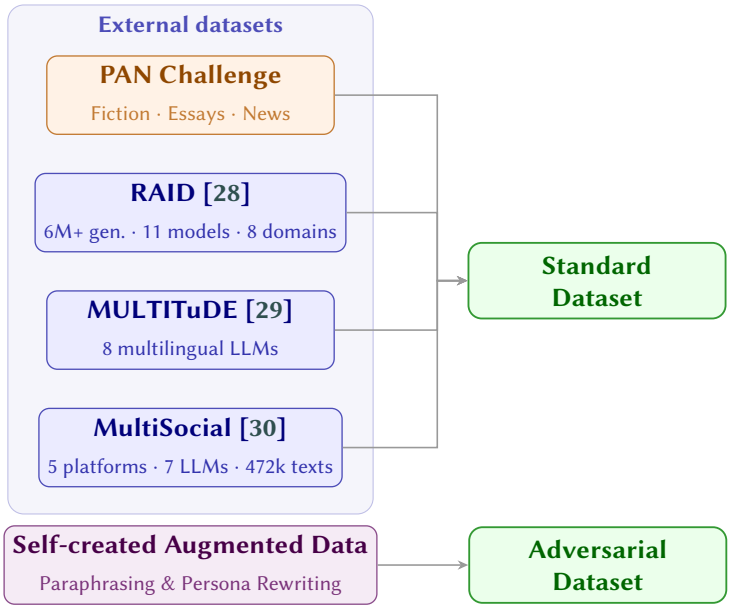

Figure 1: Training data sources.

Table 1
Dataset overview. OOD sets were not used in training.

Split	Dataset	Human	AIGT	Avg. words
Training	Challenge	10,353	0	619
	RAID	160,452	113,637	235
	MULTITuDE	1,299	9,039	233
	MultiSocial	6,320	44,436	28
	Adversarial Data	0	8,250	536
OOD held-out	Same-prompt	2,000	2,000	539
	Unseen-prompt	1,000	1,000	177

Building on last year's finding that base models overfit to the limited diversity of the provided validation data, we assembled a training corpus from four established datasets spanning multiple domains, writing styles, and sources, supplemented by self-created augmented data:

1. Human-written text from the PAN challenge corpus across three domains: fiction, essays, and news.
2. A subset of RAID [28], covering over 6 million generations from 11 models across 8 domains, including adversarial perturbations and mixed-content challenges.
3. English text from MULTITuDE [29], containing output from 8 multilingual LLMs.
4. English text from MultiSocial [30], a 472,097-text benchmark spanning 5 social media platforms and 7 multilingual LLMs.
5. Self-created adversarial dataset (see Section 3.2).

Figure 1 and Table 1 summarise the full corpus, including the held-out OOD evaluation sets described in Section 3.2.

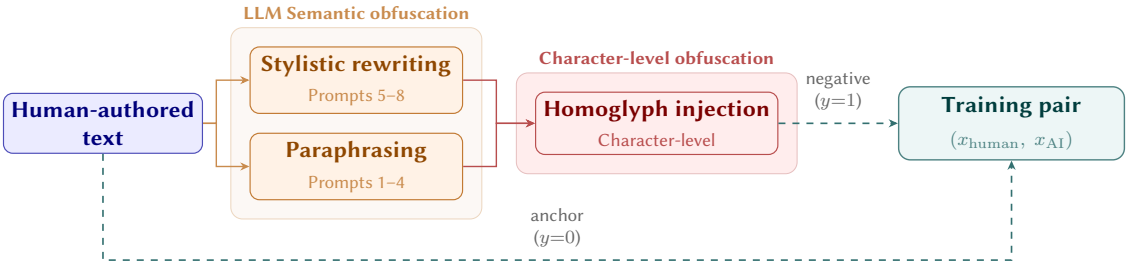

Figure 2: Adversarial augmentation pipeline for pairwise learning.

3.2. Adversarial Augmentation and OOD Evaluation

Our adversarial augmentation combines two complementary obfuscation strategies. First, *semantic rewriting*: to strengthen robustness against obfuscated content we generated adversarially augmented training data by prompting seven state-of-the-art LLMs (`Gemini 3 flash-preview`, `Claude Sonnet 4.6`, `Mistral Devstral 2 123B`, `Mistral Large 3`, `Amazon Nova Pro`, `Qwen3 VL 235B`, `DeepSeek V3.2`) to rewrite human-authored texts into naturalistic styles [31]. Text generation was performed using zero-shot prompting, in which the model received stylistic based indications without providing examples. The strategy was adapted following evidence of strong obfuscation performance in text generated by powerful language models using zero-shot prompting without the need for fine-tuning [32]. The stylistic indications covered a range of dimensions, such as syntactical simplicity, disrupted coherence and orthographic mistakes that have been found to be crucial factors in AI detection [33]. Prompts 1-4 focus on rewriting text to challenge these aspects, creating less structured and rigid content. Moreover, LLMs tend to exhibit lower emotional intensity and clearer presentation compared to human writing [34]. To address this bias, prompts 5-8 adopted a personification role, directing the language of the LLM towards a specific tone and adopting a different personality. The 8 prompts result in a final augmented dataset of 8,250 texts. Second, *character-level perturbation*: a stochastic homoglyph injection layer substitutes standard characters with visually identical Unicode equivalents during the pairwise metric fine-tuning stage, forcing the model to learn representations invariant to character-level attacks.

To evaluate generalisation beyond the training distribution, we additionally generated two held-out OOD sets using GPT-5-mini, a generator absent from all training data. The *same-prompt* set reuses the eight training prompts applied to PAN challenge human texts, isolating the effect of generator novelty. The *unseen-prompt* set applies four novel prompts (Prompts A–D) to human texts from the C4 realnewslike corpus [35], constituting a fully OOD condition across generator, prompt strategy, and human source domain simultaneously.

Table 2 lists all prompts with their role in the pipeline.

3.3. Ensemble Architecture

For our final detection pipeline, we employ an ensemble architecture combining three pairwise fine-tuned LLMs with a TF-IDF-based Random Forest classifier, as illustrated in Figure 3. Five candidate models were evaluated: Alibaba's Qwen3.5-4B and Qwen3.5-9B [36], two compact dense transformers from the Qwen3.5 series, Qwen3-14B [37], a 14B model competitive against models twice its size, Meta's Llama-3.1-8B [38], trained on over 15 trillion tokens, and Microsoft's phi-4 [39], a 14B model whose data-quality-focused training allows it to outperform significantly larger models on reasoning benchmarks. Based on in-distribution validation performance, the three strongest models were retained for the final ensemble: Qwen3-14B, Llama-3.1-8B, and phi-4.

Each retained model undergoes a two-stage fine-tuning procedure: first on the standard training dataset using cross-entropy loss to establish reliable in-distribution classification, then on the adversarially augmented dataset using the pairwise metric objective described below to optimise for adversarial robustness.

Table 2
Rewriting prompts used for adversarial augmentation (1–8) and unseen-prompt OOD evaluation (A–D).

ID	Role	Instruction
1	Training	Rewrite the text so the ideas are less organized, incorporating abrupt transitions between concepts.
2	Training	Rewrite the text as if recalling the content from memory, allowing for omissions and vague phrasing.
3	Training	Rewrite this as a casual message written quickly by a student in an informal setting.
4	Training	Rewrite the text by introducing human-like errors in grammar, punctuation, and sentence structure.
5	Training	Rephrase the text to mimic spontaneous, spoken expression. Insert natural tone, hesitation, and personality while avoiding rigid or systematic structures.
6	Training	Act as a veteran literary editor. Polish the text invisibly to eliminate stiffness, formulas, and tells, letting the prose breathe naturally while preserving intent.
7	Training	Act as a ghostwriter rewriting this text to sound like a human typing frankly. Introduce natural rhythms, small characteristics, and realistic imperfections.
8	Training	Act as a forensic linguist reconstructing this text to pass expert human scrutiny. Maintain its core meaning but rewrite it with the distinct register of an intuitive writer.
A	OOD only	Rewrite the following text in a completely different register (formal to informal or vice versa), while preserving the core meaning.
B	OOD only	Rewrite the following text at roughly half the length, keeping only the most essential information, as if the shorter version were the original.
C	OOD only	Rewrite the following text from a first-person perspective, as if the writer personally experienced or believes everything described.
D	OOD only	Rewrite the following text with a strong emotional tone of your choice, letting it influence word choice and rhythm throughout.

The Random Forest classifier operates on TF-IDF features extracted from the input text and was trained on the standard and adversarially augmented dataset only. Its inclusion in the ensemble was motivated by the strong lexical baseline performance observed in the previous year's PAN Voight-Kampff challenge, where TF-IDF representations proved competitive against more complex LLM-based systems. We note, however, that post-hoc OOD analysis revealed that the Random Forest's strong in-distribution performance is partially attributable to source-text memorisation rather than genuine AI detection (see Section 4).

Each model receives the input text and outputs a probability score indicating the likelihood of AI authorship. A weighted average is then computed across all four members, with weights determined by held-out validation performance: $w = 0.33$ for Qwen3-14B, $w = 0.30$ for Llama-3.1-8B, $w = 0.22$ for phi-4, and $w = 0.15$ for the Random Forest. The resulting ensemble probability is passed through an uncertainty gate: predictions falling within a ± 0.05 margin of 0.5 are assigned an abstention score of 0.5, improving reliability under low-confidence conditions. This design is well-suited to the shared task evaluation framework, which rewards abstention over incorrect confident predictions. The final output is therefore one of three values: 0 (human-authored), 1 (AI-generated), or 0.5 (uncertain).

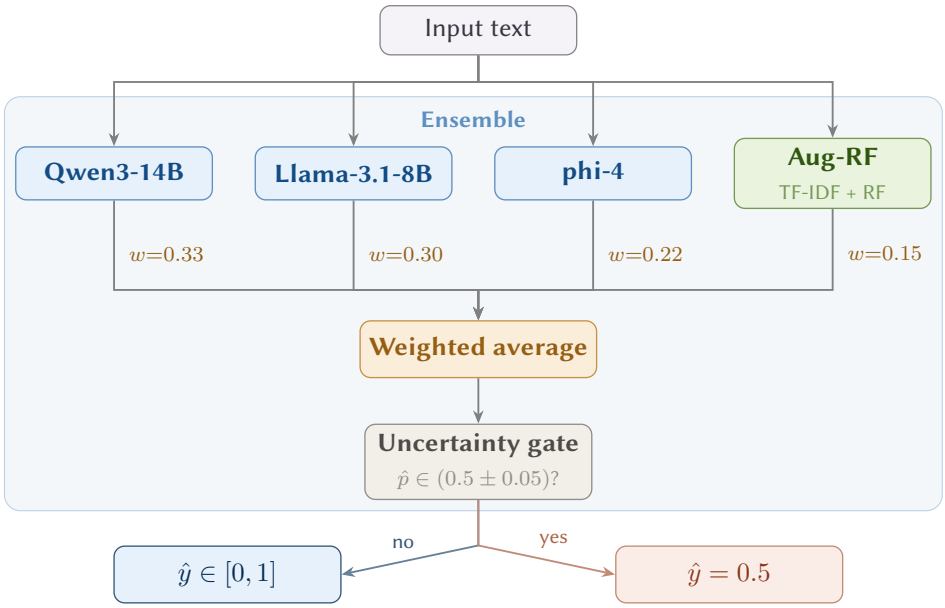

Figure 3: Ensemble inference pipeline.

3.4. Pairwise Metric Fine-Tuning

To mitigate the accuracy drop observed on adversarially augmented data, we inject a stochastic homoglyph perturbation layer directly into the training pipeline during the second fine-tuning stage. This forces the model to learn representations that are invariant to character-level noise and non-standard Unicode substitutions.

Standard cross-entropy training tends to produce anisotropic feature clusters near the human/AI decision boundary, which are easily disrupted by obfuscation. To address this, we augment the cross-entropy objective with a pairwise cosine embedding loss (specifically, the standard `CosineEmbeddingLoss` formulation), drawing on the principle of pairwise representation learning [40, 41]. We employ a Siamese network configuration in which an original human text (the anchor, h_a) and its AI-obfuscated counterpart (the negative, h_m) are processed simultaneously within the same batch. In all training pairs y is set to -1, so the loss exclusively pushes the two representations apart; the $y = 1$ case is included in Eq. (2) for completeness of the standard cosine-embedding-loss definition but is never instantiated during training. The combined objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \alpha \cdot \mathcal{L}_{\cos}(h_a,\, h_m,\, y) \tag{1}$$

where $\mathcal{L}_{\text{CE}}$ is the standard cross-entropy loss, α is a scaling coefficient, and $\mathcal{L}_{\cos}$ is the cosine embedding loss:

$$\mathcal{L}_{\cos}(h_a, h_m, y) = \begin{cases} 1 - \cos(h_a,\, h_m), & \text{if } y = 1 \\ \max\bigl(0,\, \cos(h_a,\, h_m) - m\bigr), & \text{if } y = -1 \end{cases} \tag{2}$$

With $y = -1$ for every pair, the model is trained to project human text representations and their AI-rewritten counterparts into maximally separated regions of the embedding space, reducing the vulnerability of the decision boundary to adversarial perturbation.

4. Results

All experiments were run and evaluated through TIRA [42], the shared task evaluation platform used by PAN 2026.

The results are evaluated using the following metrics:

- **ROC-AUC**: The area under the ROC curve.
- **F1**: The harmonic mean of precision and recall.
- **C@1**: A modified accuracy score that assigns non-answers (score = 0.5) the average accuracy of the remaining cases.
- **Mean**: The arithmetic mean of all the metrics above.

4.1. In-Distribution Results

Table 3
In-distribution Mean score across training stages on the standard and adversarial test sets. *Zero-shot*: next-token log-probability scoring with no task-specific training. *SFT*: supervised fine-tuning on the standard dataset. *Pairwise*: SFT followed by pairwise metric fine-tuning on adversarial data.

	Standard			Adversarial		
Model	Zero-shot	SFT	Pairwise	Zero-shot	SFT	Pairwise
Qwen3.5-4B	0.340	0.821	0.912	0.537	0.739	0.953
Qwen3.5-9B	0.332	0.860	0.891	0.546	**0.772**	0.947
phi-4	0.351	0.842	0.949	**0.581**	0.702	0.950
Qwen3-14B	**0.407**	0.916	0.952	0.558	0.684	**0.956**
Llama-3.1-8B	0.348	**0.937**	**0.953**	0.500	0.758	0.955
Mean	0.356	0.875	0.931	0.544	0.731	0.952

Table 4
In-distribution ensemble member and final ensemble evaluation. The Ensemble row combines all members via weighted averaging followed by an uncertainty gate ($\delta = 0.05$). LLM members are pairwise fine-tuned; Aug-RF is a TF-IDF Random Forest.

	Standard				Adversarial			
Model	ROC	F1	c@1	Mean[a]	ROC	F1	c@1	Mean[a]
Qwen3-14B	0.977	0.919	0.969	0.949	0.976	0.946	0.947	0.955
Llama-3.1-8B	0.978	0.923	0.971	0.952	0.970	0.946	0.947	0.953
phi-4	0.976	0.919	0.966	0.946	0.971	0.938	0.939	0.948
Aug-RF	0.933	0.589	0.683	0.696	0.992	0.880	0.910	0.913
Ensemble	**0.990**	**0.944**	**0.981**	**0.967**	**0.995**	**0.953**	**0.954**	**0.964**

[a]Mean computed over ROC-AUC, F1, c@1, and Brier complement after applying the uncertainty gate ($\delta = 0.05$).

Table 3 shows the three-stage progression across all models. Zero-shot performance is near chance on the standard test set (Mean ≈ 0.35), confirming that detection capability is entirely learned rather than inherent to model scale or architecture. On the adversarial set, zero-shot scores are slightly higher (≈ 0.54), not because the models detect AI text more reliably, but because the generative scoring mechanism is biased toward predicting AI for obfuscated text regardless of content.

Supervised fine-tuning substantially improves standard performance (Mean 0.875) but struggles to generalise to adversarial data (Mean 0.731), indicating overfitting to the standard distribution. Pairwise metric fine-tuning closes and reverses the gap: adversarial performance now exceeds standard performance by 0.021, down from a deficit of 0.144.

Table 4 shows that all three LLM members achieve strong and consistent performance, with Mean scores above 0.94 on both test sets after pairwise metric fine-tuning. Aug-RF stands out as an outlier: its ROC-AUC on the adversarial set (0.992) is the highest of any individual member, yet its F1 (0.880) and c@1 (0.910) are considerably lower, reflecting the tendency of TF-IDF representations to rank

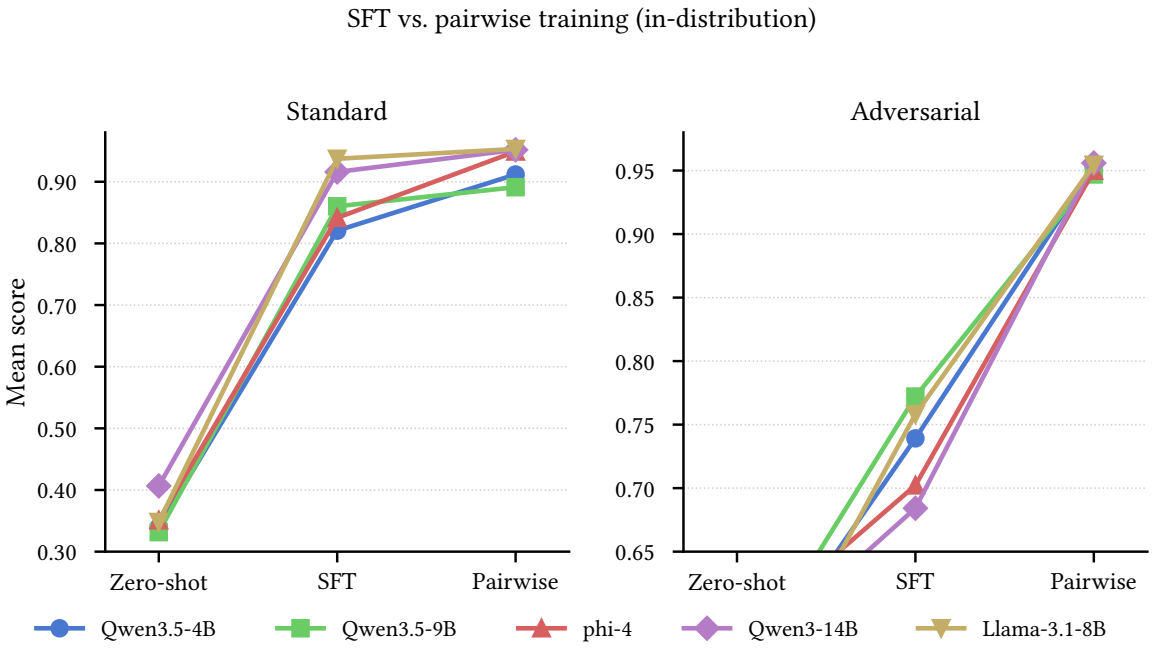

Figure 4: Mean score across training stages for all five models on the standard and adversarial test sets. The adversarial gap after SFT (Mean 0.875 vs 0.731) closes substantially after pairwise metric fine-tuning (0.931 vs 0.952).

confidently but miscalibrate predictions near the decision boundary. The ensemble outperforms all individual members on every metric across both test sets, confirming that the combination mitigates individual model weaknesses, most notably the RF calibration gap and the small per-model variance in c@1.

4.2. OOD Results

Figure 5 and Table 5 present results on both held-out conditions. Zero-shot performance remains near chance across all models and conditions (Mean ≈ 0.49 same-prompt, ≈ 0.39 unseen-prompt), confirming that detection capability is entirely learned rather than inherent to model scale. The steeper drop under unseen prompts indicates that the zero-shot scoring mechanism is sensitive to prompt-induced stylistic patterns, a signal the untrained models cannot interpret, rather than to the AI origin of the text itself.

Pairwise metric fine-tuning produces substantial gains under both conditions, with Mean scores increasing from chance level to $\approx$0.86 (same-prompt) and $\approx$0.90 (unseen-prompt). Notably, performance on unseen prompts meets or exceeds same-prompt performance for every model, confirming that pairwise metric fine-tuning learns genuine AI detection rather than prompt-specific surface patterns. The ensemble achieves the highest ROC-AUC with 0.963 in the same-prompt condition and offers the most consistent performance across members, but individual members outperform it on Mean in both OOD conditions.

The elevated false negative rate (FNR $= 0.321$ same-prompt, 0.251 unseen-prompt) reflects the conservative design of the uncertainty gate ($\delta = 0.05$), which abstains on borderline predictions rather than committing to a potentially incorrect label. This is an intentional precision–recall tradeoff: in forensic detection contexts, a false accusation of AI authorship carries higher cost than an abstention.

For classical models, Table 6 and Figure 6 reveal a striking divergence between the two OOD conditions. Aug-LR degrades modestly across conditions (ROC-AUC $0.879 \rightarrow 0.821$), consistent with a partial generalisation failure under distribution shift. Aug-RF, by contrast, achieves a suspiciously high ROC-

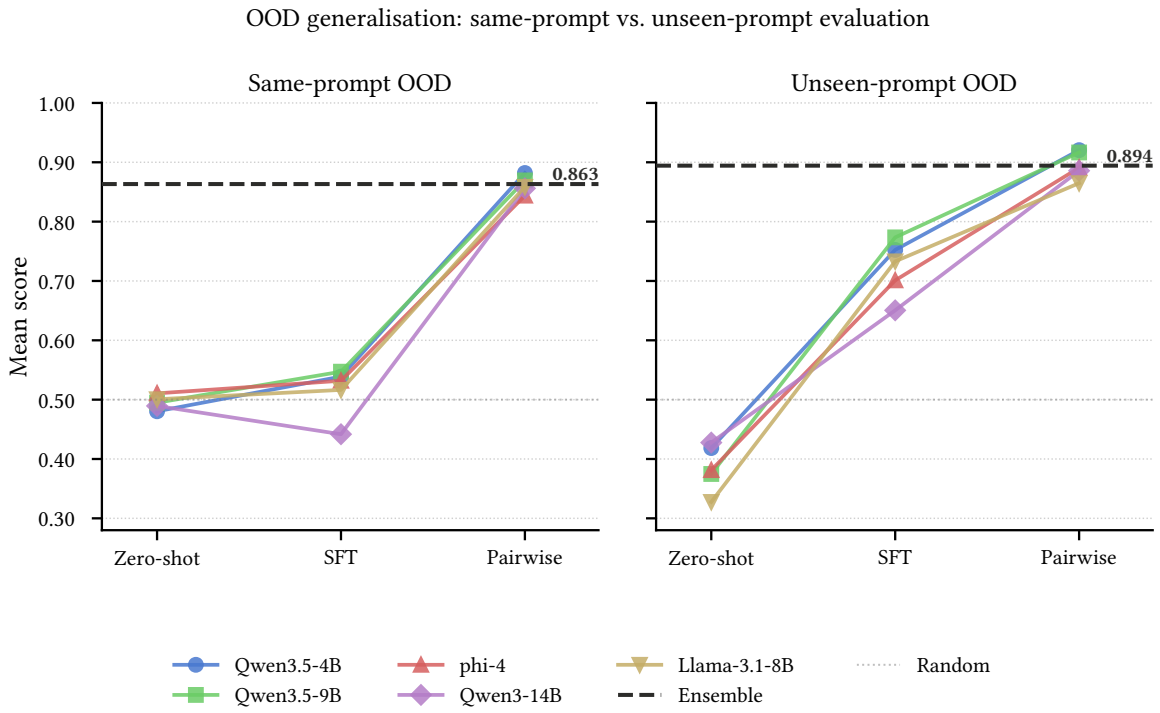

Figure 5: OOD generalisation under same-prompt and unseen-prompt conditions. Pairwise metric fine-tuned models hold up or improve under unseen prompts, confirming no prompt-specific overfitting. The ensemble achieves the best performance in both conditions.

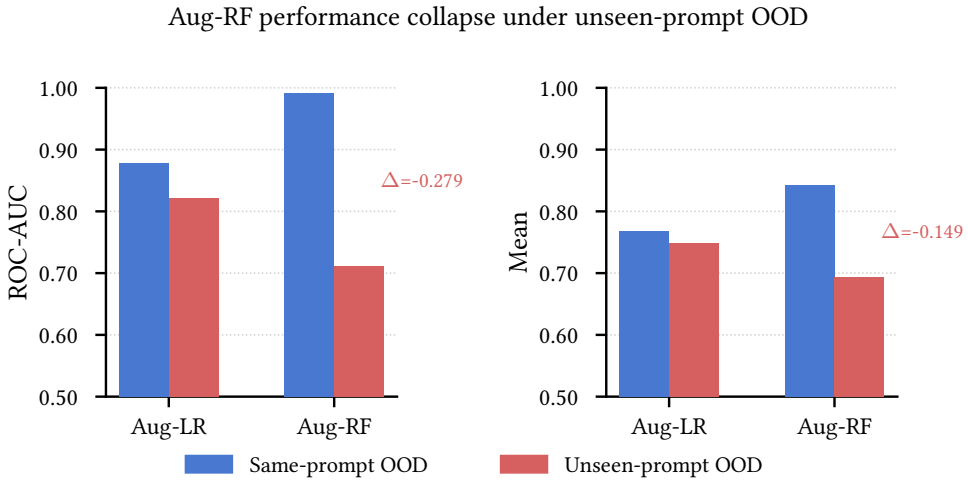

Figure 6: Aug-RF ROC-AUC and Mean score under same-prompt and unseen-prompt OOD conditions. The sharp collapse under unseen prompts confirms source-text memorisation rather than genuine AI detection.

AUC of 0.991 under same-prompt OOD with the highest of any individual model, but collapses to 0.712 under unseen-prompt OOD with new human source texts. This near-0.28 drop is the largest of any model under any condition and provides compelling evidence that Aug-RF exploits source-text identity rather than genuine AI generation patterns: when the human texts change, its discriminative signal disappears. This finding retrospectively justifies its conservative ensemble weight (w=0.15) and argues against its inclusion in future ensemble designs where full OOD robustness is required.

Table 5
OOD evaluation on the GPT-5-mini held-out set under two prompt conditions. *Same-prompt*: held-out texts generated with the same eight prompts used during adversarial training. *Unseen-prompt*: held-out texts generated with four novel prompts and human texts drawn from C4 (realnewslike), a source entirely absent from training. FNR highlights the precision/recall tradeoff of the uncertainty gate.

		Same-prompt OOD			Unseen-prompt OOD		
Model	Stage	ROC	Mean	FNR	ROC	Mean	FNR
Zero-shot (no training)							
Qwen3.5-4B	Zero-shot	0.457	0.480	—	0.452	0.418	—
Qwen3.5-9B	Zero-shot	0.475	0.495	—	0.400	0.375	—
phi-4	Zero-shot	0.504	0.511	—	0.419	0.381	—
Qwen3-14B	Zero-shot	0.473	0.489	—	0.416	0.428	—
Llama-3.1-8B	Zero-shot	0.520	0.501	—	0.287	0.328	—
After pairwise metric fine-tuning							
Qwen3.5-4B	Pairwise	0.905	**0.882**	**0.258**	**0.978**	**0.920**	0.169
Qwen3.5-9B	Pairwise	0.902	0.869	0.282	0.958	0.917	**0.165**
phi-4	Pairwise	0.895	0.844	0.325	0.949	0.891	0.219
Qwen3-14B	Pairwise	0.909	0.856	0.288	0.951	0.886	0.239
Llama-3.1-8B	Pairwise	0.893	0.859	0.280	0.918	0.865	0.243
Ensemble	Pairwise	**0.963**	0.863	0.321	0.963	0.894	0.251

Table 6
Classical model OOD evaluation under same-prompt and unseen-prompt conditions (Standard and adversarially augmented data training configuration only). The collapse of Aug-RF on unseen-prompt OOD (ROC-AUC 0.712) confirms source-text memorisation rather than genuine AI detection, justifying its conservative weight (w=0.15) in the ensemble.

	Same-prompt OOD			Unseen-prompt OOD		
Model	ROC	Mean	FNR	ROC	Mean	FNR
Aug-LR	0.879	0.769	0.439	**0.821**	**0.749**	0.325
Aug-RF	**0.991**	**0.843**	**0.372**	0.712	0.694	**0.038**

4.3. Official Results

Table 7
Official evaluation results on the PAN 2026 Voight-Kampff hidden test sets.

Dataset	ROC-AUC	Brier	C@1	F1	F0.5u	Mean
ELOQUENT 2026	0.984	0.903	0.887	0.912	0.961	0.930
PAN 2026	0.957	0.882	0.846	0.851	0.952	0.892

Table 7 reports the official evaluation results on the ELOQUENT 2026 and PAN 2026 hidden test sets [12]. On the ELOQUENT test set, the system achieves a Mean score of 0.930, confirming that the in-distribution performance reported in Section 4 generalises to the official evaluation condition. The ROC-AUC of 0.984 reflects strong discriminative ability across all decision thresholds, while the Brier complement of 0.903 indicates that the ensemble's probability outputs are well-calibrated rather than merely directionally correct. On the PAN 2026 test set, the system achieves a Mean score of 0.892, with a ROC-AUC of 0.957 and Brier of 0.882. Although it scores slightly lower than for the ELOQUENT set, the performance remains high across all metrics, indicating good overall robustness against hidden test sets. The ordering F0.5u $>$ F1 $>$ C@1 in both evaluations is consistent with the uncertainty gate design: the system is highly precise when confident, trades some recall for abstention on borderline cases, and

rewards that abstention through the C@1 metric. This tradeoff is particularly well-suited to the shared task evaluation framework, which penalises incorrect confident predictions more than abstentions

5. Conclusion

In this paper, we presented the Turing Team's submission to the PAN 2026 Voight-Kampff Generative AI Detection shared task. Our approach combined adversarial data augmentation and pairwise metric fine-tuning to increase robustness in AI-generated text detection.

The results demonstrate the importance of pairwise training across both in-distribution and OOD evaluations. Moreover, the ensemble architecture further improves the detection by combining large language models with a lexically based classifier. At the same time, the performance decrease of the TF-IDF Random Forest under unseen-prompts data reveals the limitations of lexical approaches in generalisation, highlighting the importance of OOD evaluations. Overall, the strong performance achieved on both the PAN 2026 and ELOQUENT 2026 hidden test sets suggests that our system generalises well across unseen data sets.

Acknowledgments

We gratefully acknowledge **DeiC Interactive HPC** for providing the computational resources that made this research possible.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude Sonnet 4.6 and partially Gemini 3 for grammar and spelling checks. Further, the author(s) used Claude Sonnet 4.6 for Figures 2, 1, and 3 in order to generate the TikZ foundation, which was further changed and adapted. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] M-A-P Team, X. Du, Y. Yao, K. Ma, et al., SuperGPQA: Scaling LLM evaluation across 285 graduate disciplines, arXiv preprint arXiv:2502.14739 (2025).

[2] Z. Sheng, Z. Chen, S. Gu, H. Huang, G. Gu, J. Huang, LLMs in software security: A survey of vulnerability detection techniques and insights, ACM Computing Surveys 58 (2025). doi:10.1145/3769082.

[3] J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, K.-W. Chang, Gender bias in contextualized word embeddings, 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), 2019, pp. 629–634.

[4] A. Azaria, T. Mitchell, The internal state of an LLM knows when it's lying, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 967–976. URL: https://aclanthology.org/2023.findings-emnlp.68/. doi:10.18653/v1/2023.findings-emnlp.68.

[5] S. Pudasaini, L. Miralles-Pechuán, D. Lillis, M. Llorens Salvador, Survey on AI-generated plagiarism detection: The impact of large language models on academic integrity, Journal of Academic Ethics 23 (2024) 1137–1170.

[6] X. Zhao, P. V. Ananth, L. Li, Y.-X. Wang, Provable robust watermarking for AI-generated text, arXiv preprint arXiv:2306.17439 (2023).

[7] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, S. Feizi, Can AI-generated text be reliably detected?, arXiv preprint arXiv:2303.11156 (2023).
[8] S. Rastogi, D. Pruthi, Revisiting the robustness of watermarking to paraphrasing attacks, in: Conference on Empirical Methods in Natural Language Processing, 2024. URL: https://api.semanticscholar.org/CorpusID:273901659.
[9] G. Huang, Y. Zhang, Z. Li, Y. You, M. Wang, Z. Yang, Are AI-generated text detectors robust to adversarial perturbations?, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 6005–6024. URL: https://aclanthology.org/2024.acl-long.327. doi:10.18653/v1/2024.acl-long.327.
[10] A. Creo, S. Pudasaini, SilverSpeak: Evading AI-generated text detectors using homoglyphs, in: Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect), International Conference on Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 1–46. URL: https://aclanthology.org/2025.genaidetect-1.1.
[11] J. Bevendorff, M. Fröbe, A. Greiner-Petter, A. Jakoby, M. Mayerl, P. Nakov, H. Plutz, M. Potthast, B. Stein, M. N. Ta, Y. Wang, E. Zangerle, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, A. Barrón-Cedeño, A. G. S. de Herrera, E. S. Salido, S. MacAvaney, J. M. Struß (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2026.
[12] J. Bevendorff, R. R. Gunti, J. Karlgren, M. Fr"obe, M. Potthast, B. Stein, Overview of the third "Voight-Kampff" Generative AI / LLM Detection Task at PAN and ELOQUENT 2026, in: A. Barrón-Cedeño, A. G. S. de Herrera, E. S. Salido, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2026.
[13] D. Beresneva, Computer-generated text detection using machine learning: A systematic review, in: International Conference on Applications of Natural Language to Information Systems, Springer, 2016, pp. 421–426.
[14] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, et al., Release strategies and the social impacts of language models, arXiv preprint arXiv:1908.09203 (2019).
[15] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, Y. Choi, Defending against neural fake news, Advances in Neural Information Processing Systems 32 (2019).
[16] A. A. Alethary, A. H. Aliwy, A systematic review of AI-generated text detection: Approaches, tools, and datasets, Al-Salam Journal for Engineering and Technology 5 (2026) 145–165.
[17] C. Vasilatos, M. Alam, T. Rahwan, Y. Zaki, M. Maniatakos, HowkGPT: Investigating the detection of ChatGPT-generated university student homework through context-aware perplexity analysis, arXiv preprint arXiv:2305.18226 (2023).
[18] F. Habibzadeh, GPTZero performance in identifying artificial intelligence-generated medical texts: A preliminary study, Journal of Korean Medical Science 38 (2023).
[19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: North American Chapter of the Association for Computational Linguistics, 2019. URL: https://api.semanticscholar.org/CorpusID:52967399.
[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[21] A. Bhattacharjee, H. Liu, Fighting fire with fire: Can ChatGPT detect AI-generated text?, ACM SIGKDD Explorations Newsletter 25 (2024) 14–21.
[22] D. Macko, R. Móro, A. Uchendu, I. Srba, J. S. Lucas, M. Yamashita, N. I. Tripto, D. Lee, J. Simko, M. Bielikova, Authorship obfuscation in multilingual machine-generated text detection, arXiv preprint arXiv:2401.07867 (2024).
[23] K. Krishna, Y. Song, M. Karpinska, J. Wieting, M. Iyyer, Paraphrasing evades detectors of AI-

generated text, but retrieval is an effective defense, in: Advances in Neural Information Processing Systems, 2023.
[24] N. Lu, S. Liu, R. He, K. Tang, Large language models can be guided to evade AI-generated text detection, arXiv preprint arXiv:2305.10847 (2023).
[25] J. Li, S. Ji, T. Du, B. Li, T. Wang, TEXTBUGGER: Generating adversarial text against real-world applications, in: Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS), 2019. doi:10.14722/ndss.2019.23138.
[26] J. Gao, J. Lanchantin, M. L. Soffa, Y. Qi, Black-box generation of adversarial text sequences to evade deep learning classifiers, in: Proceedings of the 2018 IEEE Security and Privacy Workshops (SPW), IEEE, 2018, pp. 92–101.
[27] B. Gunel, J. Du, A. Conneau, V. Stoyanov, Supervised contrastive learning for pre-trained language model fine-tuning, in: International Conference on Learning Representations, 2021. URL: https://arxiv.org/abs/2011.01403.
[28] L. Dugan, A. Hwang, F. Trhlík, A. Zhu, J. M. Ludan, H. Xu, D. Ippolito, C. Callison-Burch, RAID: A shared benchmark for robust evaluation of machine-generated text detectors, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 12463–12492. URL: https://aclanthology.org/2024.acl-long.674/. doi:10.18653/v1/2024.acl-long.674.
[29] D. Macko, R. Moro, A. Uchendu, J. Lucas, M. Yamashita, M. Pikuliak, I. Srba, T. Le, D. Lee, J. Simko, et al., MULTITuDE: Large-scale multilingual machine-generated text detection benchmark, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 9960–9987.
[30] D. Macko, J. Kopál, R. Moro, I. Srba, MultiSocial: Multilingual benchmark of machine-generated text detection of social-media texts, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 727–752. URL: https://aclanthology.org/2025.acl-long.36/. doi:10.18653/v1/2025.acl-long.36.
[31] J. Bevendorff, D. Dementieva, M. Fröbe, B. Gipp, A. Greiner-Petter, J. Karlgren, M. Mayerl, P. Nakov, A. Panchenko, M. Potthast, et al., Overview of PAN 2025: Voight-Kampff generative AI detection, multilingual text detoxification, multi-author writing style analysis, and generative plagiarism detection, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2025, pp. 388–411.
[32] S. Utpala, S. Hooker, P.-Y. Chen, Locally differentially private document generation using zero shot prompting, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, 2023, pp. 8442–8457. URL: https://aclanthology.org/2023.findings-emnlp.566/. doi:10.18653/v1/2023.findings-emnlp.566.
[33] S. M. Seals, V. L. Shalin, Long-form analogies generated by ChatGPT lack human-like psycholinguistic properties, arXiv preprint arXiv:2306.04537 (2023). doi:10.48550/arXiv.2306.04537.
[34] J. Wu, S. Yang, R. Zhan, Y. Yuan, L. S. Chao, et al., A survey on LLM-generated text detection: Necessity, methods, and future directions, Computational Linguistics 51 (2025) 275–338. doi:10.1162/coli_a_00549.
[35] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67.
[36] Qwen Team, Qwen3.5: Towards native multimodal agents, 2026. URL: https://qwen.ai/blog?id=qwen3.5.
[37] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, et al., Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025).
[38] A. Grattafiori, et al., The Llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).
[39] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, et al., Phi-4 technical report, arXiv preprint arXiv:2412.08905 (2024).
[40] P. Khosla, P. Teterwak, C. W. Wang, A. Sarna, Y. Tian, P. Isola, D. Erhan, C. Liu, D. Krishnan,

Supervised contrastive learning, in: Advances in Neural Information Processing Systems (NeurIPS), volume 33, 2020, pp. 18661–18673.

[41] T. Gao, X. Yao, D. Chen, SimCSE: Simple contrastive learning of sentence embeddings, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021, pp. 6894–6910.

[42] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241.