Learn-to-Distance: Distance Learning for Detecting LLM-Generated Text

Hongyi Zhou*[1] Jin Zhu*[2] Kai Ye[3] Ying Yang[1] Erhan Xu†[3] Chengchun Shi†[3]

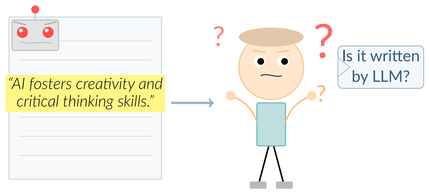
[1]Tsinghua University [2]University of Birmingham [3]London School of Economics and Political Science

*Equal contribution †Corresponding authors

Paper Code

Problem Statement

Modern LLMs (GPT-4o, Claude-3.5, Gemini, DeepSeek, ...) produce text nearly indistinguishable from human writing, raising urgent concerns about *misinformation*, *academic integrity*, and *intellectual property*.



Main Contributions on the Problem

Theoretically: Geometric framework explaining *why* rewrite-based methods work (Prop. 1), *why* they generalise to unseen prompts (Prop. 2), and *why* learning the distance outperforms fixing it (Motivating Proposition).

Methodologically: New ML-based, rewrite-based detector that *adaptively learns* the distance function, leading to better discrimination between human and LLM text.

Empirically: Across **24** datasets, **7** LLMs, **100+** settings, L2D outperforms **12** baselines; avg. relative gain **41.5%–75.4%**; adversarially robust; learning the distance gives **96%** relative improvement over using a fixed distance.

Demystifying Rewrite-Based Detectors

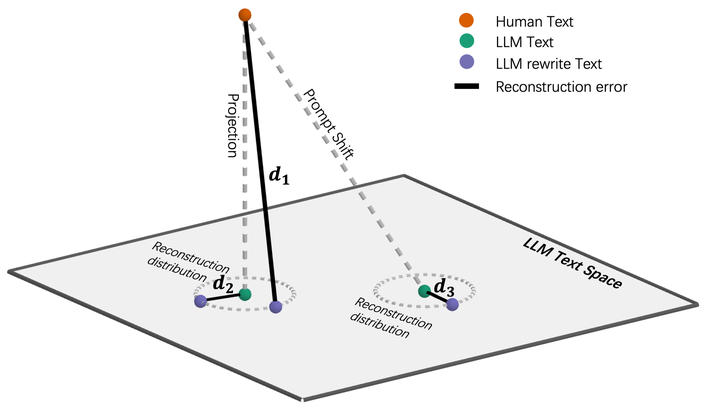


Figure 1. **Why rewrite-based detection works.** Brown dot = human text; green dots = LLM projection and LLM-generated text. Purple dots = their reconstructions. Since $d_1 > d_2$, the reconstruction error for *human* text is larger than for LLM-generated text (Proposition 1). Moreover, $d_1 > d_3$ shows prompt-robustness (Proposition 2).

Let $\mathcal{M}$ be the LLM subspace, p and q the human- and LLM-text distributions, $\Pi_{\mathcal{M}}$ the projection onto $\mathcal{M}$ such that $\Pi_{\mathcal{M}}(\boldsymbol{X}) \sim q$ when $\boldsymbol{X} \sim p$. Write the LLM's rewrite of $\boldsymbol{X}$ as $\mathcal{R}(\boldsymbol{X}) = \Pi_{\mathcal{M}}(\boldsymbol{X}) + e$ for small $e \in \mathcal{M}$.

Proposition 1 — Validity of Rewrite-Based Detection

$\mathbb{E}_{\boldsymbol{X}\sim p}\big[d^*(\boldsymbol{X}, \mathcal{R}(\boldsymbol{X}))\big] \geq \mathbb{E}_{\boldsymbol{X}\sim q}\big[d^*(\boldsymbol{X}, \mathcal{R}(\boldsymbol{X}))\big]$, with equality iff p is supported on $\mathcal{M}$.

Proposition 2 — Robustness to Unseen Prompts

Let q_{prompt} be the distribution under an unseen prompt and $|e| \leq \epsilon$ a.s. Then:

$$\mathbb{E}_{\boldsymbol{X}\sim p}[d^*(\boldsymbol{X}, \mathcal{R}(\boldsymbol{X}))] - \mathbb{E}_{\boldsymbol{X}\sim q_{\text{prompt}}}[d^*(\boldsymbol{X}, \mathcal{R}(\boldsymbol{X}))] \geq \mathbb{E}_{\boldsymbol{X}\sim p}|\boldsymbol{X} - \Pi_{\mathcal{M}}(\boldsymbol{X})| - O(\epsilon).$$

Our Proposal: Learn-to-Distance (L2D)

Motivating Proposition — Optimality of Learning the Distance

Within bounded distance functions $d : \mathcal{X} \times \mathcal{X} \to [0, M]$, the optimal assigns distance 0 when both texts lie in $\mathcal{M}$, and M when one is in $\mathcal{M}$ and the other in $\mathcal{H} \setminus \mathcal{M}$.

Since d_{opt} depends on the target LLM's subspace $\mathcal{M}$, a fixed distance cannot be universally optimal — motivating adaptive distance learning.

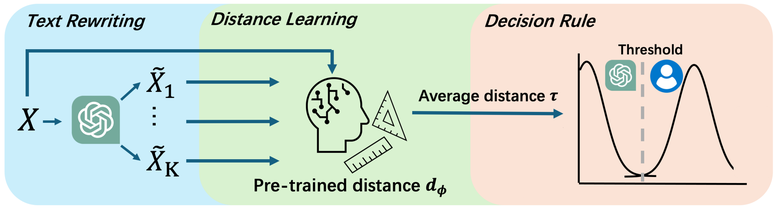


Figure 2. Workflow of L2D.

Learned Distance: Details

We parameterise the distance via a fine-tuned language model p_ϕ:

$$d_\phi(\boldsymbol{X}_1, \boldsymbol{X}_2) = \left| \frac{\log p_\phi(\boldsymbol{X}_1)}{\texttt{len}(\boldsymbol{X}_1)} - \frac{\log p_\phi(\boldsymbol{X}_2)}{\texttt{len}(\boldsymbol{X}_2)} \right|$$

Training objective (maximise reconstruction-error gap):

$$\max_\phi \underbrace{\mathbb{E}_{\boldsymbol{X}\sim\mathcal{D}_h}\Big[\tfrac{1}{K}\textstyle\sum_k d_\phi(\boldsymbol{X}, \widetilde{\boldsymbol{X}}_k)\Big]}_{\text{push human–rewrite distance} \uparrow} - \underbrace{\mathbb{E}_{\boldsymbol{X}\sim\mathcal{D}_m}\Big[\tfrac{1}{K}\textstyle\sum_k d_\phi(\boldsymbol{X}, \widetilde{\boldsymbol{X}}_k)\Big]}_{\text{pull LLM–rewrite distance} \downarrow}$$

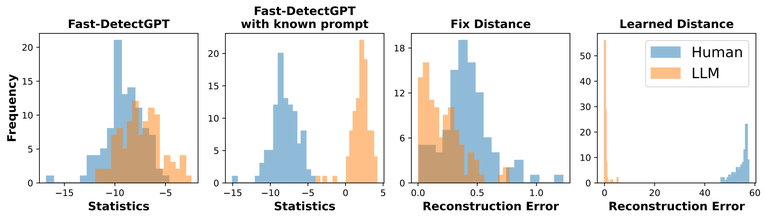


Figure 3. Effect of training distance. The learned d_ϕ provides clear separation in both cases (right two panels); a fixed distance does not.

Notable advantage — generalisation to unseen LLMs. The objective trains d_ϕ to assign large reconstruction errors to human-written text (≈40–60) and small errors to LLM-generated text (<5). Although switching to a different LLM or context slightly narrows this gap, human-text errors remain substantially larger — so d_ϕ generalises across models and prompts.

Table 1. AUCs across 27 settings (3 datasets × 3 LLMs × 3 prompt types). Best highlighted in cyan. The learned distance clearly outperforms a fixed distance (FD).

Dataset	Method	Claude-3.5				GPT-4o				Gemini			
		rewrite	polish	expand	Avg.	rewrite	polish	expand	Avg.	rewrite	polish	expand	Avg.
News	FD	0.541	0.539	0.576	0.552	0.525	0.515	0.579	0.540	0.576	0.613	0.645	0.611
	L2D	1.000	0.989	1.000	0.996	0.994	1.000	1.000	0.998	1.000	1.000	1.000	1.000
Wiki	FD	0.532	0.522	0.532	0.529	0.589	0.614	0.738	0.647	0.510	0.605	0.579	0.565
	L2D	0.955	0.942	0.953	0.950	0.963	0.987	0.993	0.981	0.983	0.982	0.988	0.984
Story	FD	0.612	0.647	0.728	0.662	0.683	0.821	0.892	0.799	0.641	0.800	0.856	0.766
	L2D	0.999	0.955	0.995	0.983	0.982	0.997	0.980	0.986	0.984	0.999	0.997	0.993

Experimental Results

Table 2. AUC under **unseen prompts**: 3 datasets × 3 target LLMs × 3 prompt types (rewrite, polish, expand) = 27 settings. L2D achieves best performance in nearly all cases. Cyan = best; Orange = second best. Last two rows per dataset: Abs. Gain and Rel. Gain (%) of L2D over the best baseline.

Dataset	Method	Claude-3.5				GPT-4o				Gemini			
		rewrite	polish	expand	Avg.	rewrite	polish	expand	Avg.	rewrite	polish	expand	Avg.
News	Likelihood	0.598	0.604	0.645	0.616	0.572	0.587	0.539	0.566	0.594	0.579	0.732	0.635
	LRR	0.594	0.626	0.636	0.619	0.633	0.620	0.559	0.604	0.656	0.601	0.717	0.658
	Binoculars	0.555	0.634	0.709	0.633	0.535	0.567	0.631	0.578	0.507	0.632	0.589	0.576
	IDE	0.606	0.686	0.726	0.673	0.577	0.736	0.696	0.670	0.608	0.672	0.716	0.665
	FDGPT	0.524	0.610	0.686	0.607	0.508	0.561	0.641	0.570	0.507	0.617	0.586	0.570
	BARTScore	0.728	0.583	0.563	0.625	0.653	0.526	0.549	0.576	0.567	0.606	0.671	0.615
	RoBERTa	0.544	0.524	0.546	0.538	0.509	0.532	0.568	0.536	0.501	0.566	0.567	0.545
	RADAR	0.744	0.805	0.912	0.821	0.774	0.966	0.994	0.911	0.807	0.858	0.920	0.862
	ADGPT	0.518	0.616	0.569	0.567	0.617	0.644	0.561	0.608	0.514	0.543	0.502	0.520
	RAIDAR	0.934	0.919	0.942	0.932	0.882	0.900	0.866	0.882	0.800	0.948	0.921	0.890
	ImBD	0.920	0.915	0.986	0.940	0.866	0.978	0.985	0.943	0.877	0.952	0.966	0.932
	L2D	1.000	0.989	1.000	0.996	0.994	1.000	1.000	0.998	1.000	1.000	1.000	1.000
	Abs. Gain (%)	6.6	7.0	1.4	5.6	11.2	2.2	1.5	5.5	12.3	4.8	3.4	6.8
	Rel. Gain (%)	100.0	86.6	100.0	94.0	94.7	100.0	100.0	96.3	100.0	100.0	100.0	100.0
Wiki	Likelihood	0.519	0.532	0.562	0.538	0.546	0.553	0.649	0.583	0.505	0.512	0.533	0.517
	LRR	0.532	0.508	0.540	0.527	0.541	0.612	0.695	0.616	0.522	0.508	0.536	0.522
	Binoculars	0.608	0.667	0.762	0.679	0.619	0.717	0.862	0.733	0.571	0.768	0.793	0.711
	IDE	0.565	0.621	0.613	0.600	0.584	0.712	0.682	0.659	0.573	0.642	0.699	0.638
	FDGPT	0.587	0.646	0.739	0.658	0.597	0.712	0.867	0.725	0.557	0.748	0.791	0.699
	BARTScore	0.760	0.634	0.520	0.638	0.785	0.592	0.529	0.635	0.605	0.590	0.615	0.603
	RoBERTa	0.635	0.659	0.759	0.684	0.565	0.590	0.522	0.559	0.638	0.740	0.782	0.720
	RADAR	0.533	0.507	0.620	0.553	0.541	0.814	0.933	0.763	0.550	0.564	0.680	0.598
	ADGPT	0.518	0.616	0.569	0.567	0.617	0.644	0.561	0.608	0.514	0.543	0.502	0.520
	RAIDAR	0.889	0.900	0.920	0.903	0.845	0.871	0.851	0.856	0.848	0.927	0.950	0.908
	ImBD	0.952	0.954	0.976	0.961	0.875	0.967	0.986	0.943	0.874	0.964	0.956	0.931
	L2D	0.955	0.942	0.953	0.950	0.963	0.987	0.993	0.981	0.983	0.982	0.988	0.984
	Abs. Gain (%)	0.2	–	–	–	8.8	1.9	0.6	3.8	10.9	1.8	3.2	5.3
	Rel. Gain (%)	5.0	–	–	–	70.6	59.1	45.8	66.4	86.3	49.6	72.6	76.9
Story	Likelihood	0.502	0.532	0.587	0.541	0.623	0.740	0.814	0.725	0.512	0.656	0.702	0.623
	LRR	0.556	0.540	0.596	0.564	0.570	0.728	0.739	0.679	0.504	0.563	0.632	0.566
	Binoculars	0.595	0.663	0.755	0.671	0.674	0.739	0.806	0.740	0.624	0.832	0.927	0.794
	IDE	0.616	0.610	0.632	0.619	0.575	0.650	0.673	0.633	0.580	0.579	0.609	0.589
	FDGPT	0.571	0.635	0.743	0.650	0.655	0.735	0.808	0.733	0.603	0.000	0.918	0.507
	BARTScore	0.767	0.706	0.566	0.680	0.724	0.754	0.685	0.721	0.708	0.733	0.674	0.705
	RoBERTa	0.588	0.586	0.660	0.611	0.540	0.504	0.539	0.527	0.571	0.569	0.657	0.599
	RADAR	0.597	0.614	0.510	0.574	0.507	0.756	0.827	0.697	0.560	0.513	0.619	0.564
	ADGPT	0.755	0.746	0.789	0.763	0.617	0.698	0.655	0.657	0.729	0.692	0.658	0.693
	RAIDAR	0.861	0.767	0.847	0.825	0.831	0.872	0.831	0.845	0.848	0.866	0.907	0.874
	ImBD	0.954	0.905	0.976	0.945	0.933	0.985	0.964	0.961	0.979	0.990	0.993	0.987
	L2D	0.999	0.955	0.995	0.983	0.982	0.997	0.980	0.986	0.984	0.999	0.997	0.993
	Abs. Gain (%)	4.5	5.0	1.9	3.8	4.9	1.2	1.6	2.6	0.4	0.9	0.4	0.6
	Rel. Gain (%)	97.8	53.0	81.2	69.6	73.3	79.4	44.8	65.4	21.4	87.1	62.4	46.4

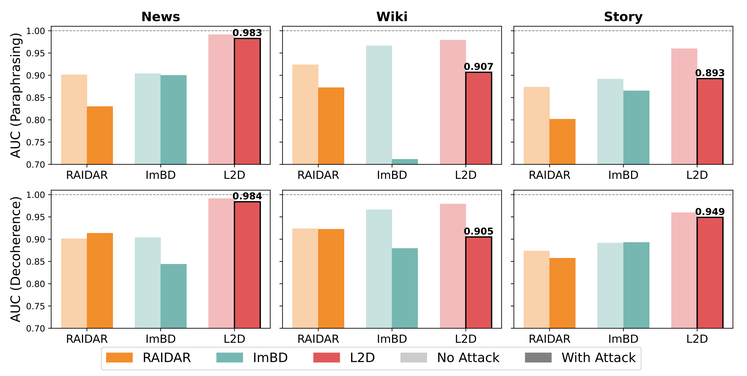


Figure 4. AUC under **paraphrasing** (top) and **decoherence** (bottom) attacks on News, Wiki, and Story datasets. Lighter/darker bar = without/under attack; bold edge = best under attack. **L2D's AUC remains stable or improves** in every setting.