Telescope: Improving Zero-Shot Detection of LLM-Generated Content by Measuring Token Repetition Probability

Christopher Nassif* Josh F. Cooper*

Virginia Tech — Dept. of Electrical & Computer Engineering / Dept. of Industrial & Systems Engineering (*equal contribution)

Code & datasets: github.com/ChrisNassif/Telescope

Motivation

Humans cannot reliably tell LLM text from human written text, so algorithmic detection is needed to curb misuse such as spam, disinformation, and academic dishonesty. LLM detection techniques must also keep pace with the frantic pace of new model releases. For this reason we focus on **zero-shot detection** which requires no retraining per target model and cannot overfit to a particular model or data.

The *Vestigial Heuristic* Hypothesis

Neural networks learn simple patterns early in training. In this work we ask the following question:

Do early pre-training pressures instill biases that are never unlearned?

Both target and reference models should carry these biases, so LLM text activates it more strongly than human text, which we can use as probe.

Telescope Perplexity

For model $\mathcal{M}$ and tokens $\vec{s} = (s_1, \ldots, s_L)$, Telescope Perplexity scores the model's likelihood of repeating the *last* token given its full context:

$$\text{Telescope Perplexity}_{\mathcal{M}}(\vec{s}) = -\frac{1}{L}\sum_{i=1}^{L} \log \mathcal{M}(s_i \mid s_{1:i})$$

Standard perplexity predicts the *next* token, $\mathcal{M}(s_i \mid s_{1:i-1})$. Conditioning on $s_{1:i}$ directly targets repetition likelihood.

To classify a piece of text as LLM generated, Telescope Perplexity outputs a score, Telescope Perplexity$_{\mathcal{M}}(\vec{s})$, and we flag the text as LLM-generated if and only if Telescope Perplexity$_{\mathcal{M}}(\vec{s}) > \tau$, where τ is a threshold chosen by the algorithm designer.

Experimental Setup

- **Baselines:** Binoculars (prior SOTA), Perplexity, DetectLLM-LRR, Fast-DetectGPT.
- **12 reference models**: Gemma 2, Llama 3.1, Falcon, SmolLM/2, GPT-J, GPT-Neo. Sizes range from 135M–9B parameters.
- **Metrics:** AUROC (threshold-free) + Transferability-F1 (threshold tuned on *other* sets).
- **Datasets:** We introduce/test on new datasets that utilize more modern target models than previous work such as GPT4o Mini and Deepseek V3, and we introduce new datasets that specifically target ESL (non native english speaker text) rewriting, which is a known llm text detector failure mode.

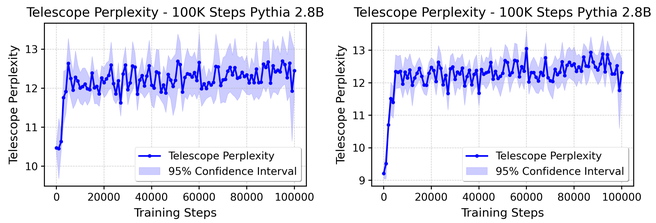
Acknowledgements

We thank Dr. Ming Jin, Virginia Tech Advanced Research Computing (ARC), and Mobius Logic for computational resources.

{chrisnassif, joshfcooper}@vt.edu

The Signature Emerges Early

Training Dynamics: On Pythia and Amber-7B checkpoints, Telescope Perplexity measured against a Pythia reference model and a SmolLM-360M reference model rises sharply early in training, then plateaus. This suggests that rising Telescope Perplexity is a foundational "vestigial" artifact developed early in training, not a capability that grows with scale.



Telescope Perplexity measured with Pythia-2.8B (left) and SmolLM-360M (right) on different Pythia-2.8B checkpoints.

Signature Locality: Telescope Perplexity, when all context is restricted to token bigrams only, loses less than 10% of its performance on the HC3 dataset with SmolLM-360M as its reference model, showing that it primarily leverages local signals.

Method	Full		Bigram
Telescope	**0.995**	→	**0.897**
Perplexity	0.991	→	0.761

Performance decrease of Telescope Perplexity and Perplexity when striping context down to the last context token.

LLMs can Detect LLM Text in Their *own* Training Corpus

SmolLM-360M can seperate human (FineWeb) from llm generated (Cosmopedia V2) data. The ability of models to separate their own training data is evidence that "vestigial" artifacts are generally *learned* biases rather than artifacts of memorized data.

Method	AUC	F1
Binoculars	0.769	0.733
Perplexity	0.952	0.906
Telescope	**0.996**	**0.987**

Separability of SmolLM-360M training corpus by technique.

Transferability Performance

Average Transferability F1-Score Across 12 Reference Models (threshold tuned on all other datasets)

Test Dataset	F1 Score				
	Telescope (ours)	Binoculars	Perplexity	DetectLLM	Fast-DetectGPT
GB Essay ChatGPT	**0.94317**	0.84594	0.93620	0.94249	0.76713
GB News ChatGPT	0.86115	**0.93763**	0.81355	0.87625	0.63212
GB Creative ChatGPT	**0.96255**	0.85858	0.89799	0.85935	0.72314
GB Essay GPT4o	**0.93922**	0.82922	0.93493	0.93341	0.73008
GB Creative GPT4o	**0.96269**	0.85953	0.73004	0.74011	0.72738
GB News Claude	0.69429	**0.79833**	0.76009	0.74903	0.60027
GB Creative Claude	**0.89028**	0.76001	0.69536	0.75085	0.62287
GB Essay Claude	0.87984	0.74559	0.86685	**0.88829**	0.69177
GB Essay Deepseek V3	0.94006	**0.94689**	0.93718	0.94351	0.71785
GB Creative Deepseek V3	0.93721	0.90241	**0.95466**	0.90851	0.69879

Detection Performance

Average AUROC Across 12 Reference Models

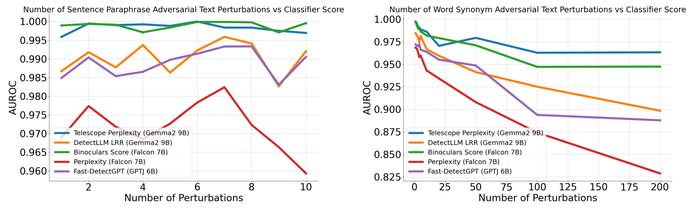
Dataset	AUROC				
	Telescope (ours)	Binoculars	Perplexity	DetectLLM	Fast-DetectGPT
Detect LLM Text	**0.99219**	0.76588	0.89307	0.92981	0.70085
AI vs Human	**0.95143**	0.86297	0.90743	0.90316	0.75608
HC3	0.99155	0.99441	**0.99471**	0.98436	0.95584
HC3 Plus	**0.98451**	0.88510	0.90999	0.87758	0.83575
ESL GPT4o Mini	**0.99983**	0.79637	0.82523	0.69051	0.60603
GB Essay ChatGPT	0.98628	0.88434	**0.99810**	0.99730	0.55624
GB News ChatGPT	0.90480	0.98773	0.98817	**0.99050**	0.91940
GB Creative ChatGPT	**0.99397**	0.91846	0.94990	0.91852	0.52336
GB Essay GPT4o	0.98136	0.85505	**0.99365**	0.99163	0.51477
GB Creative GPT4o	**0.99271**	0.91276	0.92303	0.87374	0.63813
GB News Claude	0.88038	**0.89263**	0.87211	0.86317	0.77787
GB Creative Claude	**0.96604**	0.82929	0.89304	0.87449	0.60276
GB Essay Claude	0.94223	0.77288	0.94310	**0.95988**	0.61633
GB Essay Deepseek V3	0.98484	0.99225	**0.99881**	0.99680	0.82763
GB Creative Deepseek V3	0.98199	**0.99569**	0.98852	0.96391	0.90439

SOTA or competitive across diverse datasets, excelling on more modern target models. On Detect LLM Text the error rate drops ∼15× vs. Perplexity; on ESL the gap exceeds 3 orders of magnitude.

Robustness

- **Short text:** Performs well even on shorter texts around 100–200 words.
- **Perturbations:** Resilient to a variety of attacks such as but not limited to synonym swaps, paraphrasing, small character/sentence/paragraph edits.
- **English as a Second Language Text:** Rerforms near-perfectly (0.99983 AUROC), which means our technique is significantly *less* prone to the non-native-speaker bias seen in prior work.
- **Adversarial:** Small drops under BERT "humanizer" (0.9896→0.9606 AUROC) and repeat-words prompt attacks (0.99993→0.996 AUROC).

Limitations: Stylized/formulaic genres (poetry, some news).



AUROC under sentence paraphrasing (left) and synonym swaps (right), on Ghostbusters datasets.

Takeaways

- Evidence for the **vestigial heuristic hypothesis** from testing for locality, emergence in training, and training data separability.
- **Simple, inexpensive, one-model** probe of an early-forming repetition bias.
- **SOTA/competitive** zero-shot llm text detection, which is robust to different reference models, different target models, threshold transferability, perturbation schemes, and adversarial attacks.