Counter Turing Test (CT^2): Investigating AI-Generated Text Detection for Hindi

PAPER

DATASET

Ishan Kavathekar[1], Anku Rani[2*], Ashmit Chamoli[1], Ponnurangam Kumaraguru[1], Amit Sheth[3] and Amitava Das[3]

1. IIIT, Hyderabad
2. Massachusetts Institute of Technology
3. AI Institute, University of South Carolina

**Work was done when the author was at the AIISC*

ABSTRACT

Many AI-generated text detection (AGTD) techniques have recently emerged for English, showing strong results.

But how well do they adapt to other languages?

In this study, we evaluate Watermarking, RADAR, J-Guard, ConDA, and RAIDAR for Hindi, uncovering the fragility of these methods in non-English contexts

MODEL SELECTION

Model	No output	Gibberish output	English output	Code-switching
Bloom-560M	✓	–	✓	✓
Bloom-3B	✓	–	–	✓
Bloom-7B	✓	–	✓	✓
Bloomz-560M	✓	–	–	–
Bloomz-1B	✓	✓	–	–
Bloomz-3B	✓	–	–	✓
Bloomz-7B	✓	–	–	–
GPT-2 Base	–	✓	✓	✓
GPT-2 Medium	–	✓	✓	✓
GPT-2 Large	–	✓	✓	✓
GPT-2 XL	–	✓	✓	✓
mGPT-1.3B	–	✓	–	–
Mistral-7B	–	✓	–	–
mT0 models	✓	✓	–	–
mT5 models	✓	✓		

AG_{hi} DATASET

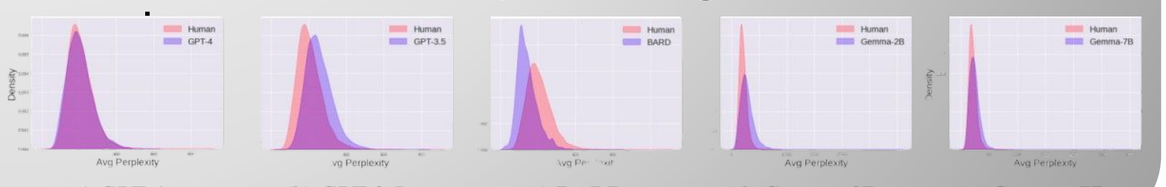
Data Sources	Human Written News Articles	AI Generated News Articles
BBC	1762	7390
NDTV	5281	22237
Total	7043	29627

	Data source	
Model	BBC	NDTV
GPT-4	1762	5280
GPT-3.5	1762	5280
BARD	1762	5280
Gemma-2B	468	1715
Gemma-7B	1636	4679
Total	7390	22234

Hindi AI Detectability Index (ADI_hi)

The Hindi AI Detectability Index (ADI_hi) ranks LLMs by their detectability, revealing a range from "easy-to-detect" to "nearly undetectable" models.

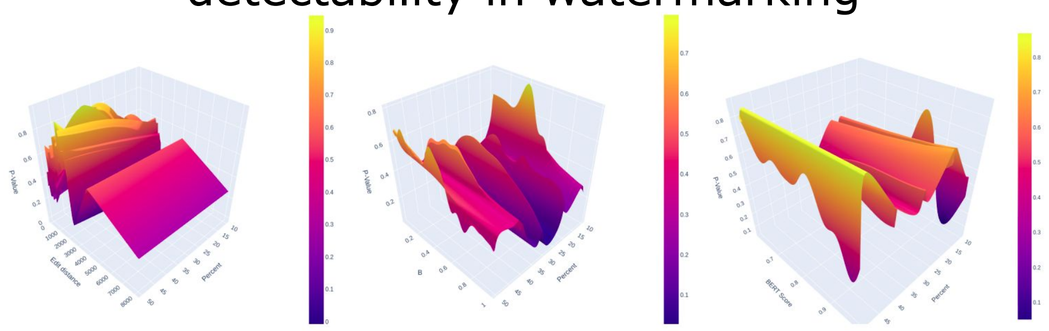
Newer models exhibit perplexities similar to human-written text, making them unreliable



(a) GPT-4 (b) GPT-3.5 (c) BARD (d) Gemma-2B (e) Gemma-7B

RESULTS

Testing the tradeoffs of distortion vs. detectability in watermarking

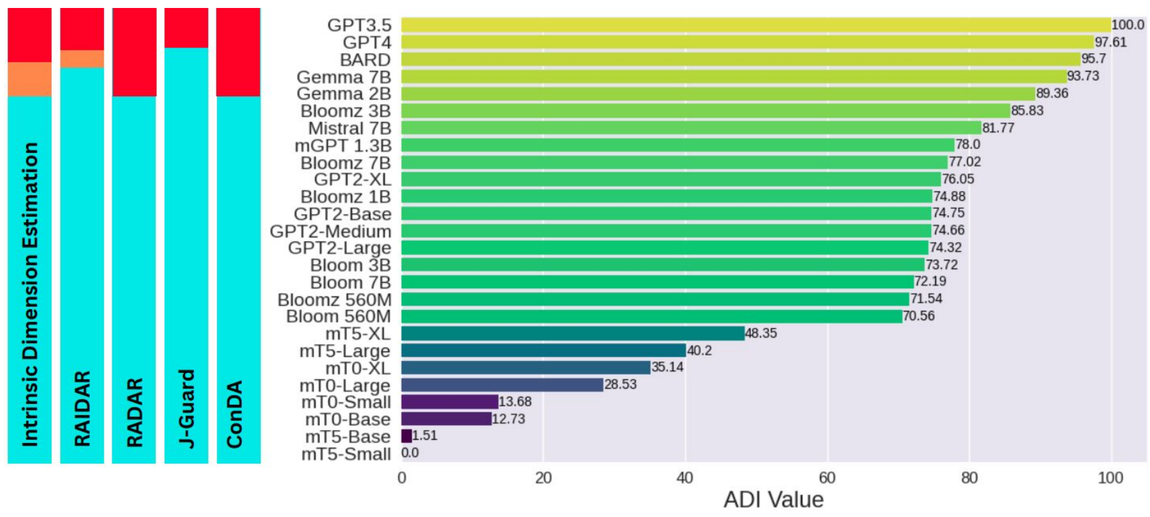


(a) Edit Distance (b) BLEU Score (c) BERTScore

Detection Techniques	Models	News Source 1 [BBC Data] Accuracy	Precision	Recall	F1-score	News Source 2 [NDTV Data] Accuracy	Precision	Recall	F1-score
RADAR	GPT-4	37.486	2.985	0.795	1.255	49.205	25.862	0.852	1.650
	GPT-3.5	37.089	0.000	0.000	0.000	48.959	12.838	0.360	0.699
	BARD	72.211	14.634	50.000	22.642	53.380	79.024	9.203	16.486
	Gemma-2B	50.748	52.448	16.026	24.55	53.557	61.423	19.125	29.169
	Gemma-7B	62.103	71.616	40.098	51.411	64.415	80.287	38.213	51.781
J-Guard	GPT-4	98.440	99.718	97.245	98.466	99.242	99.229	99.229	99.229
	GPT-3.5	99.291	99.128	99.417	99.272	98.958	99.606	98.249	98.923
	BARD	99.007	99.709	98.281	98.990	99.290	99.505	99.016	99.260
	Gemma-2B	99.467	99.454	99.454	99.454	99.344	98.996	99.711	99.352
	Gemma-7B	99.237	99.246	99.246	99.246	99.733	99.529	99.947	99.738
ConDA	GPT-4	43.445	42.212	35.528	38.582	51.856	52.736	35.770	42.627
	GPT-3.5	45.658	45.099	39.955	42.371	50.587	50.899	33.232	40.211
	BARD	47.645	47.456	43.927	45.623	55.245	57.030	42.548	48.736
	Gemma-2B	53.739	53.700	54.274	53.985	59.679	63.607	45.248	52.879
	Gemma-7B	52.353	52.323	52.995	52.657	57.491	59.537	46.762	52.382
RAIDAR	GPT-4	66.147	65.833	67.134	66.48	69.584	67.814	74.551	71.023
	GPT-3.5	64.589	64.345	65.439	64.888	60.549	60.43	61.116	60.771
	BARD	74.22	74.085	74.504	74.294	89.64	88.582	91.012	89.781
	Gemma-2B	98.404	98.925	97.872	98.396	96.939	96.532	97.376	96.952
	Gemma-7B	98.476	99.688	97.256	98.457	94.712	95.146	94.231	94.686
		MLE		PHD		MLE		PHD	
Intrinsic Dimension	Human written	10.016		6.967		9.592		6.781	
	GPT-4	9.541		7.002		9.416		6.900	
	GPT-3.5	9.796		6.882		9.549		6.720	
	BARD	7.272		3.120		7.061		3.105	
	Gemma-2B	4.368		3.004		4.537		3.118	
	Gemma-7B	5.354		3.597		5.577		3.744	

$$ADI_x = \frac{1}{U} * \sum_{j=1}^{U} \left[\{ \sum_{i=1}^{|V|} JSD_j(P_h^i \,||\, P_{ai}^i) \} * \frac{1}{|V|} \right]$$

High ADI scores indicate the need for advanced, robust detection methods as AI models increasingly produce human-like text.



TAKEAWAYS

- Existing AGTD techniques struggle with complex LLM and non-English settings, highlighting the need for language-specific considerations in AGTD.
- As LLMs evolve, so must AGTD, ensuring robust safeguards in multilingual contexts against misinformation and other risks.
- The ADI spectrum will serve as a crucial tool for assessing the detectability of LLMs.