COLING 2025 • Abu Dhabi

The 31st International Conference on Computational Linguistics

Benchmarking AI Text Detection: Assessing Detectors Against New Datasets, Evasion Tactics, and Enhanced LLMs

- Presenter: Shushanta Pudasaini
- Dr Marisa Llorens Salvador
- Dr Luis Miralles-Pechuán
- Dr David Lillis

Research Question: How effective are AIGC Detectors?

Are the current technical solutions for AI-generated text detection reliable enough?

12 Jan 2024

GENEVA INTERNET PLATFORM
digwatch

TOPICS TECHNOLOGIES PROCESSES TRENDS EVENTS

Home | Updates | AI-detection tools fail at universities worldwide

12 Jan 2024

AI-detection tools fail at universities worldwide

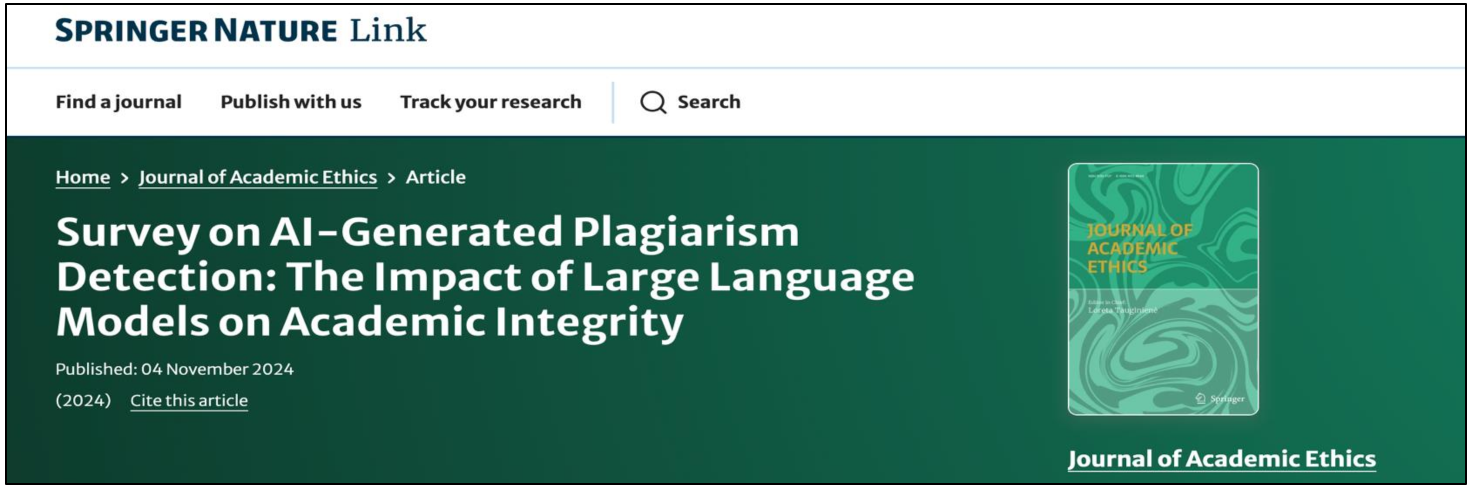
8 Aug 2024

AI STUDIES

University of Wisconsin-Madison Study Finds That Originality.ai Effectively Identifies Student-Written College Coursework From AI-Generated Text

Background Research

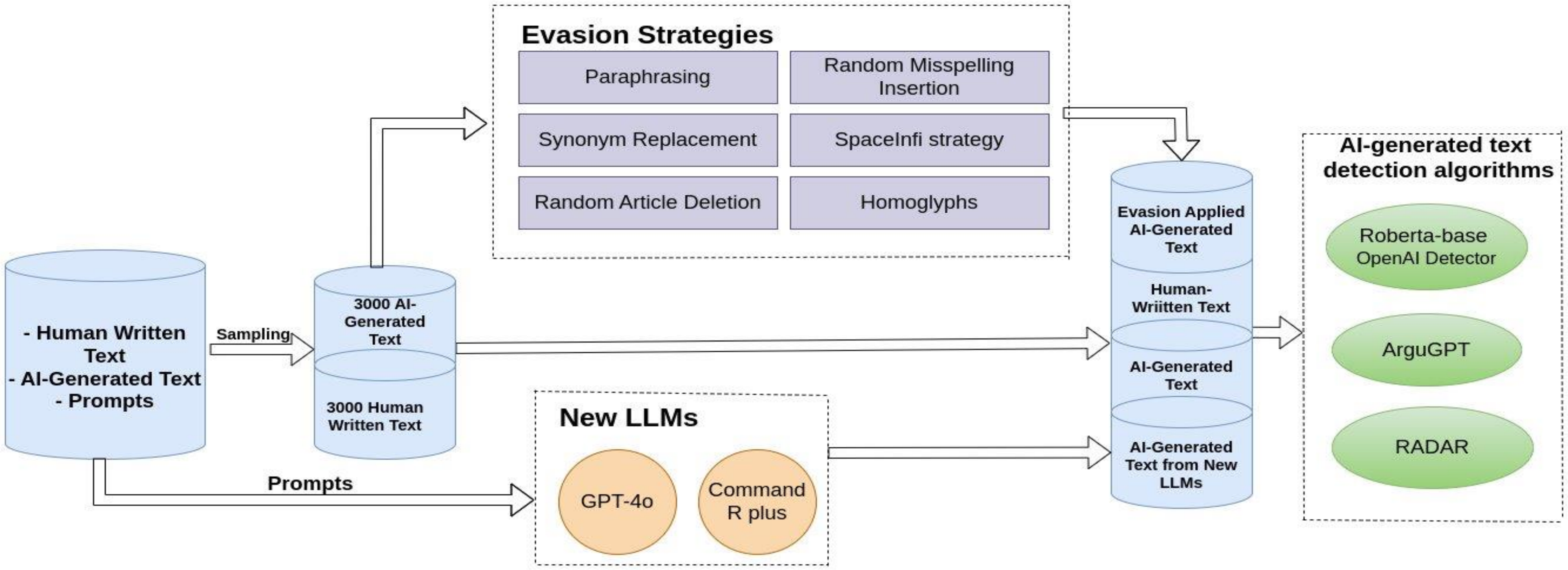
1. Datasets Used for AIGC Detection
2. Models Developed for AIGC Detection using several Approaches
   - Watermarking based approach
   - Zero-shot based approach
   - Training classifier-based approach
3. Tools Developed for AIGC detection



Evasion Techniques Developed to Fool AIGC Detectors

Survey done from Jan 2024 – Sep 2024

Experiment methodology



Block diagram of the methodology applied for benchmarking experiment

Results (OpenAI Detector)

Dataset	Acc (%)	F1	FN	FP
HC3	98.1	0.9806	113	1
M4	89.17	0.8804	607	43

Benchmarking results on **multiple domains/datasets**

LLMs Tested	Acc (%)	F1
GPT-3.5	91.0	0.9022
Command R plus	67.5	0.5255
GPT4-o	52.0	0.0943

Benchmarking results on **multiple generators/LLMs**

Experiment Type	Acc (%)	F1
No Evasion Applied	98.1	0.9806
Whitespace Insertion	92.02	0.9133
Synonym Replacement	83.43	0.8015
Paraphrasing	69.32	0.5576
Homoglyphs	66.93	0.506
Article Deletion	60.33	0.3425
Misspelling Insertion	51.12	0.0443

Benchmarking results on **multiple evasion techniques**

Discussion and Analysis

PATTERN ANALYSIS

1. The **same pattern was observed in all the models tested** in the experiment
   - OpenAI Detector
   - RADAR
   - Argu GPT
2. Performance **drops significantly** for **newer LLMs** (GPT-4o, Command R Plus).
3. AI detectors trained with **adversarial learning** (covering multiple evasion strategies) **perform better**.

MODEL ANALYSIS

OpenAI Detector:

- OpenAI Detector performs **poorly** against evasion techniques (e.g., paraphrasing, synonym replacement).

RADAR :

- Performs better against **paraphrasing & synonym replacement** but fails on **homoglyphs & article deletion**.

ArguGPT :

- Good at detecting non-evasive text.
- Fails under homoglyphs and misspellings.

CONCLUSION

AIGC detection models claim high accuracy. However, these models fail when subjected to testing on:

- Evasion applied AI-generated text
- AI-generated text from recent LLMs
- Texts from diverse datasets and domains

THANK YOU

Models	LLM tested against	Acc. (%)	F1-Score	FN (Out of 100)
OpenAI Detector	baseline	91	0.9022	17
	Command R plus	67.5	0.5255	64
	GPT-4o	0.52	9.43	95
RADAR	baseline	97.5	0.9751	2
	Command R plus	82	0.7882	33
	GPT-4o	0.6	36.51	77
ArguGPT	baseline	94	0.9434	0
	Command R plus	93.5	0.9384	1
	GPT-4o	90.5	0.9073	7

Benchmarking results on multiple generators/LLMs

Model	Dataset (6000 Samples)	Acc.(%)	F1	FN	FP	Prec	Rec
OpenAI Detector	M4	89.17	0.8804	607	43	0.982	0.797
	HC3	98.09	0.9806	113	1	0.999	0.962
RADAR	M4	94.13	0.9413	177	175	0.943	0.941
	HC3	89.18	0.8994	96	553	0.84	0.968
ArguGPT	M4	92	0.9257	8	472	0.863	0.997
	HC3	97.41	0.9748	0	155	0.951	1

Benchmarking results on multiple domains/dataset

Model	Dataset	Experiment Type	Acc. (%)	F1	FN (Out of 3000)
OpenAI Detector	M4 Dataset	non-evasive	89.17	0.8804	607
		evasion whitespace	79.63	0.7488	1179
		evasion removed articles	51.95	0.0999	2840
		evasion misspell text	51.77	0.0934	2851
		evasion homoglyph	61.53	0.3891	2265
		evasion synonym replaced	74.55	0.6651	1484
		evasion paraphrase	68.67	0.553	1837
	HC3 Dataset	non-evasive	98.1	0.9806	113
		evasion whitespace	92.02	0.9133	478
		evasion removed articles	60.33	0.3425	2379
		evasion misspell text	51.12	0.0443	2932
		evasion homoglyph	50.6	0.6693	1983
		evasion synonym replaced	83.43	0.8015	993
		evasion paraphrase	55.76	0.6932	1840
RADAR	M4 Dataset	non-evasive	94.13	0.9413	177
		evasion whitespace	95.10	0.9515	119
		evasion removed articles	71.47	0.6309	1537
		evasion misspell text	47.10	0.0006	2999
		evasion homoglyph	47.15	0.0025	2996
		evasion synonym replaced	94.27	0.9427	169
		evasion paraphrase	95.70	0.9576	83
	HC3 Dataset	non-evasive	89.18	0.8995	96
		evasion whitespace	89.82	0.9059	58
		evasion removed articles	82.06	0.8215	523
		evasion misspell text	41.70	0.0305	2945
		evasion homoglyph	40.92	0.0045	2991
		evasion synonym replaced	88.98	0.8974	108
		evasion paraphrase	90.22	0.9100	34
ArguGPT	M4 Dataset	non-evasive	92	0.9257	8
		evasion whitespace	91.93	0.9251	12
		evasion removed articles	89.20	0.8971	176
		evasion misspell text	42.13	0.0000	3000
		evasion homoglyph	42.13	0.0000	3000
		evasion synonym replaced	91.87	0.9244	16
		evasion paraphrase	90.55	0.9111	95
	HC3 Dataset	non-evasive	97.42	0.9748	0
		evasion whitespace	97.40	0.9747	1
		evasion removed articles	97.23	0.9730	11
		evasion misspell text	47.42	0.0000	3000
		evasion homoglyph	47.42	0.0000	2999
		evasion synonym replaced	97.37	0.9743	3
		evasion paraphrase	96.58	0.9664	50

Benchmarking results on multiple evasion techniques

Is AI-generated text detection even possible?

- Large number of solutions have been developed to solve the problem
- Most of the commercial tools and algorithms claim they have above 95% accuracy but they can be easily fooled
- Major challenge is to develop robust algorithms capable of detecting modified text and text generated from new powerful LLMs

COLING 2025 • Abu Dhabi

The 31st International Conference on Computational Linguistics

Benchmarking AI Text Detection: Assessing Detectors Against New Datasets, Evasion Tactics, and Enhanced LLMs

- Presenter: Shushanta Pudasaini
- Dr Marisa Llorens Salvador
- Dr Luis Miralles-Pechuán
- Dr David Lillis