澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

科技學院
Faculdade de Ciências e Tecnologia
Faculty of Science and Technology

NEURAL INFORMATION PROCESSING SYSTEMS

DetectRL: Benchmarking LLM-Generated Text Detection in Real-World Scenarios

DetectEval

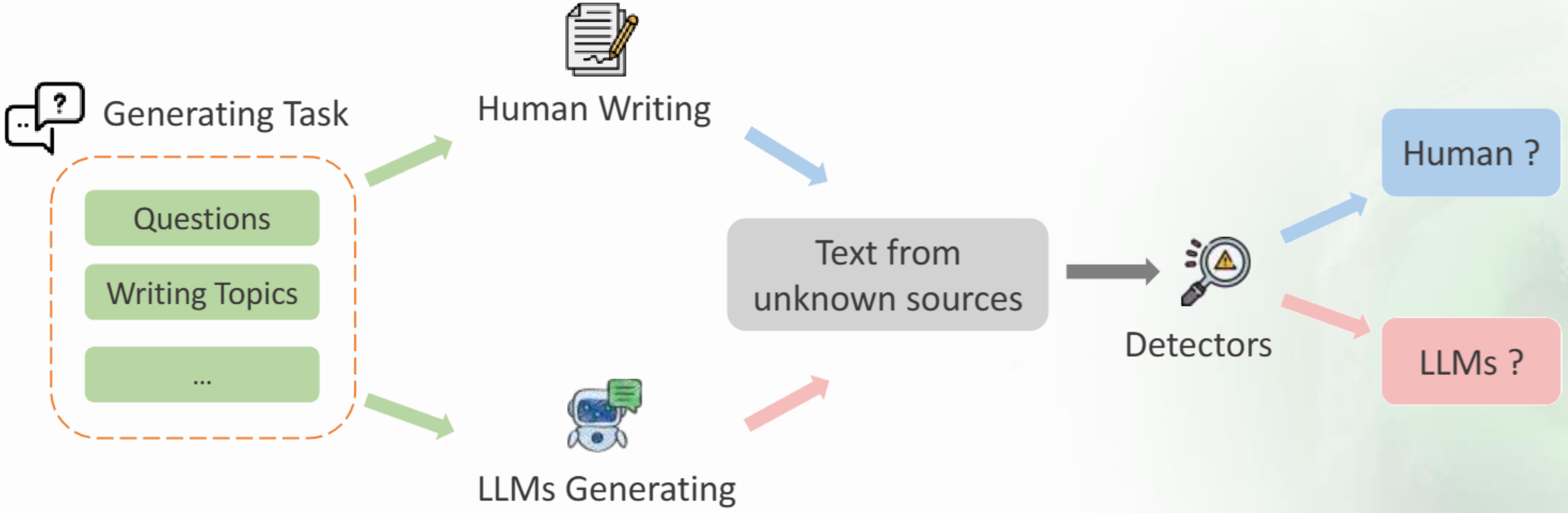
Junchao Wu, Runzhe Zhan, Derek F. Wong, Shu Yang,
Xinyi Yang, Yulin Yuan, Lidia S. Chao
NLP2CT Lab, University of Macau

NLP2CT
Natural Language Processing & Portuguese – Chinese Machine Translation Laboratory

RL stands for Real-world LLM & Reinforcement Learning, DetectRL aims to enhance the development of detectors that perform effectively in real-world scenarios, thereby improving their overall effectiveness, similar to the principles of reinforcement learning.

Background

- The **critical task** of detecting text generated by large language models.
- Detection capabilities of current detectors have reached **impressive** levels.



Motivation

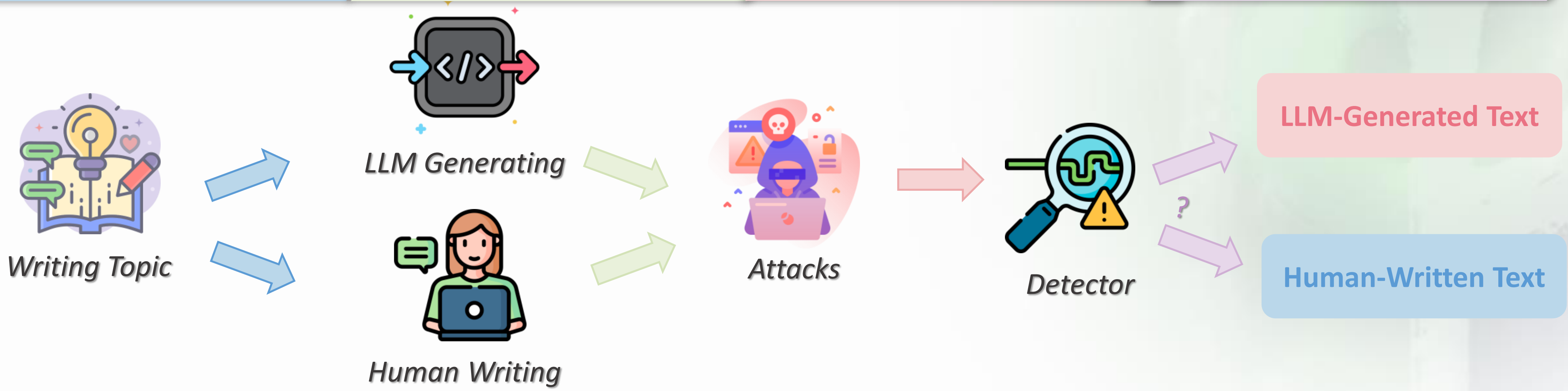
- Previous popular benchmarks primarily focused on **idealized test data**.
- The reliability of existing detectors in **real-world applications** remains **underexplored**.

Tasks are only required in in **high-risk** domains with potential for **misuse**.

Texts are generally generated by **powerful and popular** LLMs.

People would like to use **various prompts, edit or modify** text or find ways to **avoid detection**.

The detector needs to generalize to **varied lengths** of text for training and text.



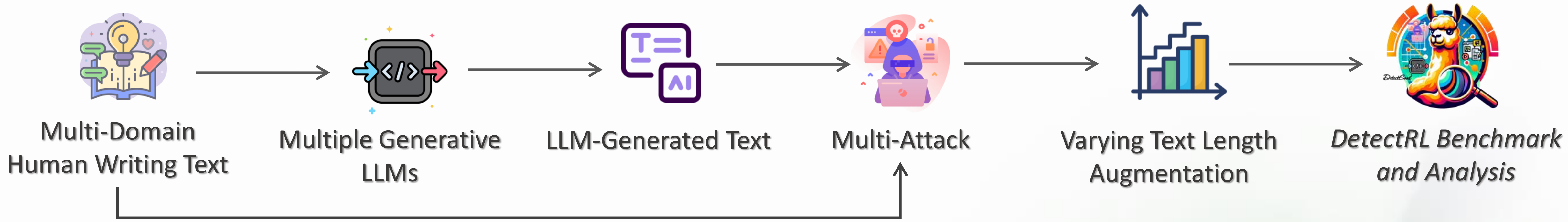
Research Questions

(1) How do **SOTA** LLM-generated text detectors perform in **real-world application scenarios**?

(2) What **real-world factors** influence the performance of detectors and to what extent?

We investigate these questions by introducing DetectRL, a novel benchmark for real-world LLM-generated text detection.

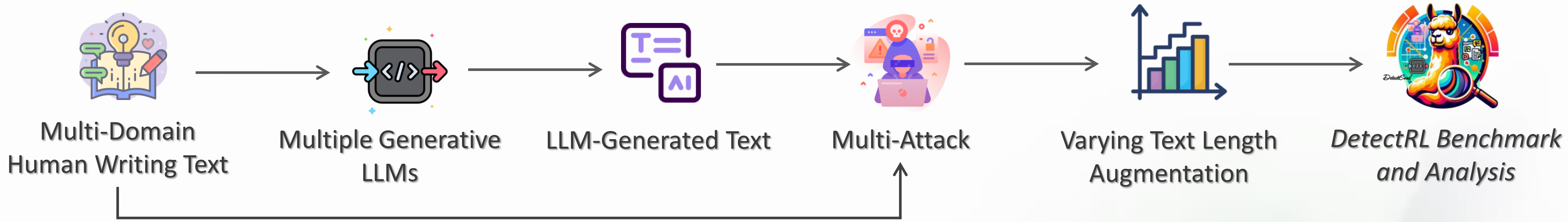
Our Benchmark: DetectRL



Pipeline of Benchmark Framework

- High-risk and abuse-prone writing **domain**
- Widely-used and powerful **LLMs**
- Various **Attacks** align with practical applications
- Text with **varying interval lengths**
- **Balanced sample distributions** across domains, LLMs, and attack types in all test scenarios.

Our Benchmark: DetectRL



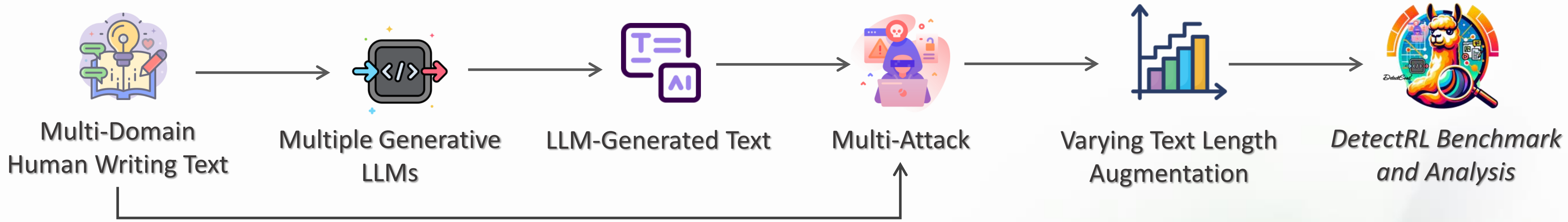
Data Sources

- arXiv Archive *(academic writing)*
- XSum Dataset *(news writing)*
- Writing Prompts *(creative writing)*
- Yelp Reviews *(social media)*

Generative Models

- GPT-3.5-Turbo
- PaLM-2-bison
- Claude-instant
- Llama-2-70b

Our Benchmark: DetectRL



Attacks Methods

Attacks Typts	Sub Types	Methods
Direct Prompt	Direct Prompt	Prompt
Prompt Attacks	Few-Shot Prompt	Prompt
	ICO Prompt	Prompt
Paraphrase Attacks	DIPPER Paraphrase	DIPPER Paraphraser
	Polish Using LLMs	Prompt
	Back Translation	Google Translation API
Perturbation Attacks	Character-Level Perturbation	TextFooler
	Word-Level Perturbation	DeepBugWord
	Sentence-Level Perturbation	TextBugger
Data Mixing	Multi-LLMs Mixing	Sentence Mixing
	LLM-Centered Mixing	Sentence Mixing

Various Prompts Usage

Human Revision

Writing Errors

Data Mixing

Benchmark Statistics and Task definition

Task	Setting	Sub Setting	Training Supervised	Training Zero-Shot	Test
Task 1	Multi-Domain	Academic	25,990	2,008	2,008
		News	25,992	2,008	2,008
		Creative	25,985	2,008	2,008
		Social Media	25,984	2,008	2,008
	Multi-LLM	GPT-3.5-turbo	25,987	2,008	2,008
		Claude-instant	25,990	2,008	2,008
		PaLM-2-bison	25,987	2,008	2,008
		Llama-2-70b	25,987	2,008	2,008
	Multi-Attack	Direct	20,384	2,016	2,016
		Prompt	31,568	2,032	2,032
		Paraphrase	42,767	2,016	2,016
		Perturbation	42,784	2,016	2,016
		Data Mixing	401,184	2,008	2,008
Task 2	Domain Generalization	Academic	25,990	2,008	6,024
		News	25,992	2,008	6,024
		Creative	25,985	2,008	6,024
		Social Media	25,984	2,008	6,024
	LLM Generalization	GPT-3.5-turbo	25,987	2,008	6,024
		Claude-instant	25,990	2,008	6,024
		PaLM-2-bison	25,987	2,008	6,024
		Llama-2-70b	25,987	2,008	6,024
	Attack Generalization	Direct	20,384	2,016	6,048
		Prompt	31,568	2,032	6,096
		Paraphrase	42,767	2,016	6,048
		Perturbation	42,784	2,016	6,048
		Data Mixing	401,184	2,008	6,024
Task 3	Varying Text Length	Training-Time	16,200	16,200	900
		Test-Time	900	900	16,200
Task 4	Human Writing	Direct	20,384	2,016	2,016
		Paraphrase	42,767	2,016	2,016
		Perturbation	42,784	2,016	2,016
		Data Mixing	42,788	2,012	2,012

Task 1: In-domain robustness

To evaluate the **foundational performance** of detectors in different domains, generators, and attack strategies.

Task 2: Generalization

To evaluate the detector's ability to handle **out-of-distribution** samples within each category.

Task 3: Varying text length

To evaluates how training-time and test-time **text length** affects the performance of detectors.

Task 4: Real-world human writing

To evaluates the impact of **human-written factors** on the performance of detectors.

Evaluation Metrics

AUROC

- considers both True Positive Rate (TPR) and False Positive Rate (FPR).

F1 Score

- considers both Precision and Recall.

Detection Methods

Zero-shot Methods

- Log-Likelihood *(Gehrmann et al., 2019)*
- Rank *(Gehrmann et al., 2019)*
- Log-Rank *(Gehrmann et al., 2019)*
- LRR *(Su et al., 2023)*
- NPR *(Su et al., 2023)*
- Revise-Detcet. *(Zhu et al., 2023)*
- DetectGPT *(Mitchell et al., 2023)*
- DNA-GPT *(Yang et al., 2024)*
- Binoculars *(Hans et al., 2024)*
- Fast-DetectGPT *(Bao et al., 2024)*

Supervised Classifiers

- *RoBERTa-Base* *(Liu et al., 2019)*
- *RoBERTa-Large* *(Liu et al., 2019)*
- *XLM-RoBERTa-Base* *(Conneau et al., 2019)*
- *XLM-RoBERTa-Large* *(Conneau et al., 2019)*

Discussion: Leaderboard

- Supervised detectors **consistently outperform** zero-shot detectors.
- For zero-shot detectors, **Binoculars** ranked highest.
- DetectGPT and similar advanced detectors are **unreliable**.

Leaderboard: LLM-Generated Text Detector in Real-World Scenarios

Tasks Settings → / Detectors ↓	Multi-Domain		Multi-LLM		Multi-Attack		Generalization			Time		Human Writing		Avg.
							Domain	LLM	Attack	Train	Test			
	AUROC	F_1	AUROC	F_1	AUROC	F_1	F_1	F_1	F_1	F_1	F_1	AUROC	F_1	F_1
Rob-Base	99.98	99.75	99.93	99.58	99.56	97.66	83.00	91.81	92.37	79.99	74.00	97.34	94.31	93.02
Rob-Large	99.78	98.87	95.16	90.03	99.87	99.03	77.20	82.85	83.96	86.08	85.23	96.68	94.63	91.49
X-Rob-Base	99.92	99.34	99.14	98.17	98.49	96.07	75.97	92.73	90.58	84.25	73.83	93.43	90.29	91.71
X-Rob-Large	99.01	97.44	97.40	93.47	99.31	97.75	76.14	85.89	73.42	86.35	79.83	97.21	94.43	90.59
Binoculars	83.95	78.25	83.30	74.83	85.05	78.53	77.47	74.10	74.70	73.82	74.34	90.68	85.98	79.61
Revise-Detect.	67.24	60.82	66.36	53.72	70.89	57.24	54.50	53.28	50.63	65.71	67.96	83.29	82.16	64.13
Log-Rank	64.43	57.53	63.75	54.18	68.52	55.15	55.10	52.78	51.28	57.44	59.74	88.46	83.85	62.48
LRR	65.47	55.45	64.93	53.01	68.53	57.99	54.61	52.73	57.41	57.09	58.15	85.99	80.56	62.46
Log-Likelihood	63.71	56.36	62.97	53.13	67.97	54.38	53.37	51.77	50.73	57.92	59.28	88.48	83.75	61.83
DNA-GPT	64.92	55.83	64.36	51.09	68.36	53.36	51.51	47.09	41.98	57.63	62.43	87.80	82.77	60.70
Fast-DetectGPT	58.52	48.07	59.58	46.55	60.70	50.63	48.35	36.56	49.47	61.31	55.08	76.03	68.47	55.33
Rank	51.34	44.97	50.33	42.06	57.08	48.83	42.61	41.49	38.84	41.67	46.65	83.86	80.00	51.52
NPR	48.37	41.41	47.27	40.04	53.49	45.22	38.58	38.83	36.10	37.60	42.17	80.03	75.98	48.08
DetectGPT	34.43	21.52	34.93	14.80	36.19	19.15	11.54	13.11	11.84	35.78	34.69	60.86	48.76	29.05
Entropy	46.02	27.40	46.97	34.25	43.75	24.69	25.06	31.07	16.53	13.38	15.99	22.39	16.60	28.01

Discussion: Significant Challenge

- **Incorporating a mix distribution** of domains, LLMs, and attack types increases the testing pressure of zero-shot method.

Leaderboard: LLM-Generated Text Detector in Real-World Scenarios

Tasks Settings →	Multi-Domain		Multi-LLM		Multi-Attack		Generalization			Time		Human Writing		Avg.
							Domain	LLM	Attack	Train	Test			
Detectors ↓	AUROC	F_1	AUROC	F_1	AUROC	F_1	F_1	F_1	F_1	F_1	F_1	AUROC	F_1	F_1
Rob-Base	99.98	99.75	99.93	99.58	99.56	97.66	83.00	91.81	92.37	79.99	74.00	97.34	94.31	93.02
Rob-Large	99.78	98.87	95.16	90.03	99.87	99.03	77.20	82.85	83.96	86.08	85.23	96.68	94.63	91.49
X-Rob-Base	99.92	99.34	99.14	98.17	98.49	96.07	75.97	92.73	90.58	84.25	73.83	93.43	90.29	91.71
X-Rob-Large	99.01	97.44	97.40	93.47	99.31	97.75	76.14	85.89	73.42	86.35	79.83	97.21	94.43	90.59
Binoculars	83.95	78.25	83.30	74.83	85.05	78.53	77.47	74.10	74.70	73.82	74.34	90.68	85.98	79.61
Revise-Detect.	67.24	60.82	66.36	53.72	70.89	57.24	54.50	53.28	50.63	65.71	67.96	83.29	82.16	64.13
Log-Rank	64.43	57.53	63.75	54.18	68.52	55.15	55.10	52.78	51.28	57.44	59.74	88.46	83.85	62.48
LRR	65.47	55.45	64.93	53.01	68.53	57.99	54.61	52.73	57.41	57.09	58.15	85.99	80.56	62.46
Log-Likelihood	63.71	56.36	62.97	53.13	67.97	54.38	53.37	51.77	50.73	57.92	59.28	88.48	83.75	61.83
DNA-GPT	64.92	55.83	64.36	51.09	68.36	53.36	51.51	47.09	41.98	57.63	62.43	87.80	82.77	60.70
Fast-DetectGPT	58.52	48.07	59.58	46.55	60.70	50.63	48.35	36.56	49.47	61.31	55.08	76.03	68.47	55.33
Rank	51.34	44.97	50.33	42.06	57.08	48.83	42.61	41.49	38.84	41.67	46.65	83.86	80.00	51.52
NPR	48.37	41.41	47.27	40.04	53.49	45.22	38.58	38.83	36.10	37.60	42.17	80.03	75.98	48.08
DetectGPT	34.43	21.52	34.93	14.80	36.19	19.15	11.54	13.11	11.84	35.78	34.69	60.86	48.76	29.05
Entropy	46.02	27.40	46.97	34.25	43.75	24.69	25.06	31.07	16.53	13.38	15.99	22.39	16.60	28.01

Discussion: In-domain Robustness

- Text with more formal **stylistic nature** poses a greater challenge.

Metrics →	AUROC	F_1	AUROC	F_1	AUROC	F_1	AUROC	F_1	AUROC	F_1	AUROC	F_1
Multi-Domain												
Domain Settings →	-		ArXiv		XSum		Writing		Review		Avg.	
Log-Likelihood	-		65.35	57.55	45.68	41.32	68.00	59.38	75.84	67.22	63.22	56.37
Entropy	-		48.39	29.71	67.84	57.23	39.06	20.55	28.82	02.14	46.53	27.66
Rank	-		57.17	54.62	36.87	22.47	56.26	50.90	55.08	51.90	51.09	44.97
Log-Rank	-		67.01	60.09	46.74	42.60	67.58	57.57	76.40	69.88	64.43	57.78
LRR	-		70.54	61.34	50.09	38.38	64.65	53.09	76.61	68.99	65.47	55.70
NPR	-		53.85	49.65	34.59	18.31	54.96	52.30	50.09	45.39	48.87	41.16
DetectGPT	-		22.15	00.00	12.21	00.00	58.95	50.83	44.43	35.25	34.44	21.02
DNA-GPT	-		67.41	58.30	64.22	45.09	69.04	58.25	78.17	69.28	69.71	57.23
Revise-Detect.	-		70.40	37.51	50.34	46.07	73.24	64.29	75.01	68.71	67.75	54.65
Binoculars	-		84.03	76.77	77.39	72.18	94.38	79.73	90.00	84.32	86.95	78.75
Fast-DetectGPT	-		43.69	24.46	39.19	28.39	74.21	67.84	77.02	71.62	58.03	48.08
Avg.	-		59.09	46.36	47.74	37.45	65.48	55.88	66.13	57.70	59.68	49.39
Rob-Base	-		100.0	100.0	99.99	99.85	99.99	99.65	99.97	99.50	99.99	99.75
Rob-Large	-		99.99	99.90	99.85	98.95	99.54	97.73	99.76	98.90	99.54	98.87
X-Rob-Base	-		100.0	100.0	99.97	99.55	99.84	98.76	99.88	99.05	99.92	99.59
X-Rob-Large	-		99.98	99.85	99.84	98.95	99.85	98.31	96.40	92.66	99.23	97.19
Avg.	-		99.99	99.93	99.91	99.32	99.80	98.61	99.00	97.52	99.67	98.85

Discussion: In-domain Robustness

- Difference in **statistical patterns of LLMs** pose significant challenges to detectors.

Multi-LLM											
LLM Settings →	-	GPT-3.5		Claude		PaLM-2		Llama-2		Avg.	
Log-Likelihood	-	62.89	57.80	43.32	28.10	70.03	60.73	75.65	65.90	62.47	53.63
Entropy	-	46.84	23.29	52.25	30.42	45.34	16.56	43.48	66.75	46.98	34.26
Rank	-	52.19	49.32	41.68	22.78	50.40	41.74	57.05	54.40	50.33	42.56
Log-Rank	-	62.84	56.87	43.32	30.12	70.89	63.09	77.97	66.66	63.76	54.68
LRR	-	61.61	52.12	43.30	18.91	71.17	65.51	83.65	75.51	64.43	53.01
NPR	-	50.29	43.81	41.64	32.91	44.64	34.77	52.53	48.68	47.78	40.54
DetectGPT	-	43.46	26.27	32.86	12.56	26.72	00.00	36.71	20.40	34.44	14.81
DNA-GPT	-	61.87	55.04	48.88	25.67	71.48	60.77	75.22	62.89	64.86	51.59
Revise-Detect.	-	70.10	62.72	49.87	27.28	69.84	59.03	75.65	65.87	66.87	53.73
Binoculars	-	88.14	82.50	55.15	39.35	93.30	88.20	96.64	92.30	83.31	75.59
Fast-DetectGPT	-	65.56	59.55	30.01	00.00	65.99	57.58	76.79	69.08	59.59	46.55
Avg.	-	60.52	51.75	43.84	24.37	61.80	49.81	68.30	62.58	58.62	47.35
Rob-Base	-	99.97	99.70	99.98	99.80	99.94	99.40	99.84	99.45	99.93	99.59
Rob-Large	-	99.77	98.86	96.23	92.48	97.93	92.64	86.72	76.17	95.66	90.54
X-Rob-Base	-	99.88	99.45	98.26	97.48	98.77	97.19	99.69	98.57	99.15	98.17
X-Rob-Large	-	99.55	97.56	91.67	84.24	98.73	94.43	99.66	97.67	97.65	93.73
Avg.	-	99.79	98.89	96.53	93.50	98.84	95.91	96.47	92.96	98.09	95.50

Discussion: In-domain Robustness

- **Perturbation attacks** represent the most significant threat to current detectors.

Multi Attack												
Attack Settings →	Direct		Prompt		Paraph.		Perturb		Mixing		Avg.	
Log-Likelihood	89.25	82.09	86.87	78.16	64.55	57.59	35.51	00.78	63.70	53.31	67.97	54.38
Entropy	26.47	00.00	26.18	00.00	48.12	26.01	68.62	68.95	49.37	28.52	43.75	24.69
Rank	83.50	76.27	81.21	72.86	60.60	52.60	08.04	00.00	52.05	42.46	57.08	48.83
Log-Rank	89.25	81.45	86.35	77.51	64.69	59.17	37.71	00.78	64.63	56.86	68.52	55.15
LRR	85.83	77.40	80.80	74.30	63.99	55.20	45.91	29.27	66.12	53.81	68.53	57.99
NPR	77.98	71.61	77.15	70.63	56.94	46.25	06.78	00.00	48.63	37.65	53.49	45.22
DetectGPT	52.84	40.90	51.83	37.98	31.79	16.89	18.21	00.00	26.28	00.00	36.19	19.15
DNA-GPT	88.01	80.78	85.62	77.47	65.61	54.94	40.45	02.73	62.14	50.89	68.77	53.76
Revise-Detect.	86.88	79.61	84.89	76.21	67.26	62.03	43.98	07.56	65.27	54.39	69.26	56.76
Binoculars	94.87	89.73	93.45	88.12	88.34	81.56	76.89	69.34	89.12	83.67	88.53	82.48
Fast-DetectGPT	79.56	72.45	78.43	70.34	70.12	62.89	49.56	41.23	67.23	59.78	68.58	61.34
Avg.	78.04	70.33	76.68	67.85	60.89	52.90	38.76	30.54	60.41	52.43	62.56	54.41
Rob-Base	99.87	99.60	99.78	99.47	99.67	99.12	98.32	97.45	99.12	98.76	99.35	98.88
Rob-Large	98.73	97.83	98.45	97.56	97.89	96.78	96.12	94.67	97.56	96.34	97.75	96.64
X-Rob-Base	99.56	99.12	99.23	99.01	98.89	98.34	98.56	97.89	99.01	98.56	98.85	98.58
X-Rob-Large	99.45	98.67	98.89	97.98	98.23	97.67	97.89	96.34	98.67	97.89	98.63	97.71
Avg.	99.40	98.80	99.09	98.50	98.67	97.98	97.22	96.09	98.34	97.89	98.54	97.85

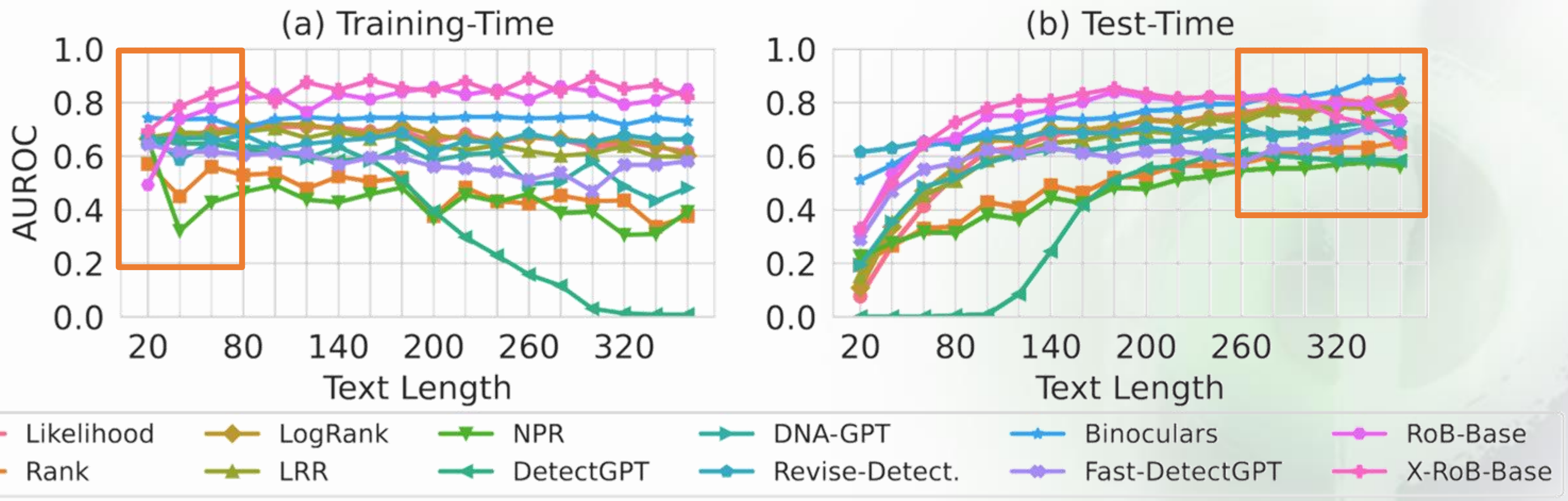
Discussion: Generalization of Detectors

- **Less formal stylistic data** to enhance generalization.
- Texts generated by LLMs with **similar statistical patterns** generally perform well with each other.
- **Perturbation attacks** poses the greatest challenge to generalization.

Detectors →	*LRR (Zero-shot)*					*Fast-DetectGPT (Zero-shot)*					*Rob-Base (Supervised)*				
						Multi-Domain									
Train ↓ Eval →	ArXiv	XSum	Writing	Review	Avg.	ArXiv	XSum	Writing	Review	Avg.	ArXiv	XSum	Writing	Review	Avg.
ArXiv	57.55	40.88	38.44	55.81	48.17	24.46	23.71	59.67	60.17	42.00	100.0	75.90	77.68	70.69	81.06
XSum	57.45	41.32	39.08	55.81	48.41	28.43	28.39	62.99	63.08	45.72	68.43	99.85	71.79	67.17	76.81
Writing	61.14	46.31	59.38	67.98	58.70	34.81	33.60	67.84	68.30	51.13	78.58	72.72	99.65	94.24	86.29
Review	61.49	47.02	57.12	67.22	58.21	40.70	37.66	68.25	71.62	54.55	82.64	84.15	85.10	99.50	87.84
						Multi-LLM									
Train ↓ Eval →	GPT-3.5	PaLM-2	Claude	Llama-2	Avg.	GPT-3.5	PaLM-2	Claude	Llama-2	Avg.	GPT-3.5	PaLM-2	Claude	Llama-2	Avg.
GPT-3.5	52.12	61.79	24.70	75.34	53.48	59.55	59.56	12.96	69.93	50.50	99.97	70.34	62.90	94.68	81.97
PaLM-2	52.36	65.51	26.23	75.58	54.92	55.77	57.58	08.20	68.43	47.49	99.25	99.40	93.43	99.25	97.83
Claude	45.73	57.66	18.91	72.67	48.74	00.19	00.00	00.00	01.18	00.34	96.83	83.92	99.80	89.77	92.58
Llama-2	52.14	62.23	25.25	75.51	53.78	56.28	57.74	08.65	69.08	47.93	99.45	93.02	87.56	99.45	94.87
						Multi-Attack									
Train ↓ Eval →	Prompt	Paraph.	Perturb	Mixing	Avg.	Prompt	Paraph.	Perturb	Mixing	Avg.	Prompt	Paraph.	Perturb	Mixing	Avg.
Direct	74.23	58.35	30.69	56.42	54.92	64.01	40.45	41.02	31.81	44.32	95.73	94.91	64.32	89.07	86.00
Prompt	74.30	58.35	30.81	56.42	54.97	64.00	39.94	40.40	31.25	43.89	97.18	94.98	86.18	92.92	92.81
Paraphrase	70.22	55.20	20.25	51.26	49.23	61.54	38.32	36.86	27.90	41.15	93.66	98.26	78.81	89.38	90.02
Perturb	71.81	58.22	29.27	55.19	53.62	64.01	40.45	41.14	31.93	44.38	87.01	91.46	98.66	91.38	92.12
Mixing	71.02	55.77	24.01	53.81	51.15	65.89	46.38	45.78	40.93	49.74	93.46	91.93	95.26	93.64	93.57

Discussion: Impact of text length

- **Shorter training samples** for stronger zero-shot detectors.
- **Longer test samples** for better zero-shot detection, but not too long for supervised methods.



Discussion: Impact of real-world human writing

- Paraphrase attacks and data mixing have **minimal impact** on zero-shot detectors, but paraphrase attacks can **confuse** supervised detectors.
- Perturbation attacks on human-written texts appeared to **enhance** the discernment capabilities of zero-shot detectors.

Settings → Detectors ↓	Direct		Paraphrase Attack		Perturbation Attack		Data Mixing		Avg.	
	AUROC	F_1	AUROC	F_1	AUROC	F_1	AUROC	F_1	AUROC	F_1
Zero-shot Detectors										
Log-Likelihood	89.25	82.09	76.77	74.28	99.53	97.76	88.40	80.88	88.48	83.75
Entropy	26.47	00.00	27.15	00.00	03.37	00.00	32.58	66.40	22.39	16.60
Rank	83.50	76.27	72.14	74.13	99.63	98.13	80.17	71.48	83.86	80.00
Log-Rank	89.25	81.45	76.78	75.17	99.49	97.57	88.32	81.23	88.46	83.85
LRR	85.83	77.40	76.05	74.46	98.09	94.78	83.99	75.60	85.99	80.56
NPR	77.98	71.61	69.82	70.60	98.35	95.51	73.97	66.22	80.03	75.98
DetectGPT	52.84	40.90	68.45	73.45	87.95	79.74	34.20	00.98	60.86	48.76
DNA-GPT	88.01	80.78	77.19	75.95	98.81	95.83	87.40	76.55	87.85	82.27
Revise-Detect.	86.88	79.61	65.39	73.65	98.96	95.48	85.52	77.37	84.18	81.52
Binoculars	94.75	88.10	80.00	74.76	98.26	94.87	93.80	88.32	91.70	86.51
Fast-DetectGPT	77.28	68.79	77.18	70.13	84.43	74.45	65.23	60.53	76.03	68.47
Avg.	77.45	67.90	69.72	66.96	87.89	84.01	73.96	67.77	77.25	71.66
Supervised Detectors										
Rob-Base	99.77	98.10	89.82	80.98	99.99	99.65	99.81	98.51	97.34	94.31
Rob-Large	99.77	98.95	87.01	80.42	99.99	99.95	99.95	99.20	96.68	94.63
X-Rob-Base	98.36	96.20	81.93	75.06	99.96	99.30	93.47	90.62	93.43	90.29
X-Rob-Large	99.79	98.31	89.07	80.32	99.99	99.90	99.82	99.20	97.21	94.43
Avg.	99.42	97.89	86.95	79.19	99.98	99.70	98.26	96.88	96.16	93.41

Conclusion

- DetectRL, a **novel benchmark** designed to evaluate the usability of detectors in scenarios that closely resemble real-world applications.
- Reveal the **primary reasons** why existing detectors for LLM-generated texts struggle in practical applications.
- Discussion of the **potential factors** influencing detector performance.
- Provides a **data curation framework**, which supports the rapid creation of an evolving, comprehensive benchmark aligns with real-world scenarios.

Thanks for listening!

Code & Data:

https://github.com/NLP2CT/DetectRL