

# Evaluation and Benchmark for Machine-Generated Text Detection

## 机器生成文本检测的评估和基准

赵国宇

2024.08.23



中国科学院 信息工程研究所  
INSTITUTE OF INFORMATION ENGINEERING, CAS

# OUTLINE

---

## 01 DATASET & BENCHMARK

- 1.1 流行数据集介绍与对比
- 1.2 24年新数据集介绍 (M4、M4GT-Bench、RAID、TEXTMACHINA)

## 02 EVALUATION

- 2.1 评测维度
- 多场景评测: MAGE: Machine-generated Text Detection in the Wild (ACL 2024)
- 鲁棒性评测: Stumbling Blocks: Stress Testing the Robustness of Machine-Generated Text Detectors Under Attacks (ACL 2024)

## 03 CONCLUSION

# 1.1 流行数据集介绍与对比- Dataset

Summary of Detection Datasets for LLM-generated text detection.

Corpus	Use	Human LLMs	LLMs Type	Language	Attack	Domain	
HC3 (Guo et al. 2023)	train	~80k	~43k	ChatGPT	English, Chinese	-	Web Text, QA, Social Media
CHEAT (Yu et al. 2023a)	train	~15k	~35k	ChatGPT	English	Paraphrase	Scientific Writing
HC3 Plus (Su et al. 2023b)	train valid test	~95k ~10k ~38k	GPT-3.5-Turbo	English, Chinese	Paraphrase	News Writing, Social Media	
OpenLLMText (Chen et al. 2023a)	train, valid, test	~52k ~8k ~8k	~209k ~33k ~33k	ChatGPT, PaLM, LLaMA, GPT2-XL	English	-	Web Text
GROVER Dataset (Zellers et al. 2019b)	train	~24k	Grover-Mega	English	-	News Writing	
TweepFake (Fagni et al. 2021)	train	~12k	~12k	GPT-2, RNN, Markov, LSTM, CharRNN	English	-	Social Media
GPT-2 Output Dataset <sup>6</sup>	train test	~250k ~5k	~2000k ~40k	GPT-2 (small, medium, large, xl)	English	-	Web Text
ArguGPT (Liu et al. 2023c)	train valid test	~6k 700 700	GPT2-XL, Text-Babbage-001, Text-Curie-001, Text-Davinci-001, Text-Davinci-002, Text-Davinci-003, GPT-3.5-Turbo	English	-	Scientific writing	
DeepfakeTextDetect (Li et al. 2023c)	train valid test	~236k ~56k ~56k	GPT (Text-Davinci-002, Text-Davinci-003, GPT-Turbo-3.5), LLaMA (6B, 13B, 30B, 65B), GLM-130B, FLAN-T5 (small, base, large, xl, xxl), OPT(125M, 350M, 1.3B, 2.7B, 6.7B, 13B, 30B, iml1.3B, iml-30B), T0 (3B, 11B), BLOOM-7B1, GPT-J-6B, GPT-NeoX-20B)	English	Paraphrase	Social Media, News Writing, QA, Story Generation, Comprehension and Reasoning, Scientific writing	

- 大模型生成文本数据集非常重要
- 用于开发和校准detector
- 处于初级阶段，主要针对特定领域和特定模型

# 1.1 流行数据集介绍与对比- Dataset

Summary of Detection Datasets for LLM-generated text detection.

Corpus	Use	Human LLMs	LLMs Type	Language	Attack	Domain	
HC3 (Guo et al. 2023)	train	~80k	~43k	ChatGPT	English, Chinese	-	Web Text, QA, Social Media
CHEAT (Yu et al. 2023a)	train	~15k	~35k	ChatGPT	English	Paraphrase	Scientific Writing
HC3 Plus (Su et al. 2023b)	train	~95k		GPT-3.5-Turbo	English, Chinese	Paraphrase	News Writing, Social Media
	valid	~10k					
	test	~38k					
OpenLLMText (Chen et al. 2023a)	train, valid, test	~52k, ~8k, ~8k	~209k, ~33k, ~33k	ChatGPT, PaLM, LLaMA, GPT2-XL	English	-	Web Text
GROVER Dataset (Zellers et al. 2019b)	train	~24k		Grover-Mega	English	-	News Writing
TweepFake (Fagni et al. 2021)	train	~12k	~12k	GPT-2, RNN, Markov, LSTM, CharRNN	English	-	Social Media
GPT-2 Output Dataset <sup>6</sup>	train, test	~250k, ~5k	~2000k, ~40k	GPT-2 (small, medium, large, xl)	English	-	Web Text
ArguGPT (Liu et al. 2023c)	train, valid, test	~6k, 700, 700		GPT2-XL, Text-Babbage-001, Text-Curie-001, Text-Davinci-001, Text-Davinci-002, Text-Davinci-003, GPT-3.5-Turbo	English	-	Scientific writing
DeepfakeTextDetect (Li et al. 2023c)	train, valid, test	~236k, ~56k, ~56k		GPT (Text-Davinci-002, Text-Davinci-003, GPT-Turbo-3.5), LLaMA (6B, 13B, 30B, 65B), GLM-130B, FLAN-T5 (small, base, large, xl, xxl), OPT(125M, 350M, 1.3B, 2.7B, 6.7B, 13B, 30B, iml1.3B, iml-30B), T0 (3B, 11B), BLOOM-7B1, GPT-J-6B, GPT-NeoX-20B)	English	Paraphrase	Social Media, News Writing, QA, Story Generation, Comprehension and Reasoning, Scientific writing

HC3 (The Human ChatGPT Comparison Corpus)  
 最早的开源数据集，开创性贡献，包括收集人类和ChatGPT对相同问题的回答，计算机、金融、医学等领域。  
 Prompt缺乏多样性。例What、Why、How

# 1.1 流行数据集介绍与对比- Dataset

Summary of Detection Datasets for LLM-generated text detection.

Corpus	Use	Human LLMs	LLMs Type	Language	Attack	Domain
HC3 (Guo et al. 2023)	train	~80k	~43k	ChatGPT	-	Web Text, QA, Social Media
CHEAT (Yu et al. 2023a)	train	~15k	~35k	ChatGPT	Paraphrase	Scientific Writing
HC3 Plus (Su et al. 2023b)	train	~95k	-	GPT-3.5-Turbo	Paraphrase	News Writing, Social Media
	valid	~10k		English, Chinese		
	test	~38k				
OpenLLMText (Chen et al. 2023a)	train, valid, test	~52k, ~8k, ~8k	~209k, ~33k, ~33k	ChatGPT, PaLM, LLaMA, GPT2-XL	-	Web Text
GROVER Dataset (Zellers et al. 2019b)	train	~24k	-	Grover-Mega	-	News Writing
TweepFake (Fagni et al. 2021)	train	~12k	~12k	GPT-2, RNN, Markov, LSTM, CharRNN	-	Social Media
GPT-2 Output Dataset <sup>6</sup>	train, test	~250k, ~5k	~2000k, ~40k	GPT-2 (small, medium, large, xl)	-	Web Text
ArguGPT (Liu et al. 2023c)	train, valid, test	~6k, 700, 700	-	GPT2-XL, Text-Babbage-001, Text-Curie-001, Text-Davinci-001, Text-Davinci-002, Text-Davinci-003, GPT-3.5-Turbo	-	Scientific writing
DeepfakeTextDetect (Li et al. 2023c)	train, valid, test	~236k, ~56k, ~56k	-	GPT (Text-Davinci-002, Text-Davinci-003, GPT-Turbo-3.5), LLaMA (6B, 13B, 30B, 65B), GLM-130B, FLAN-T5 (small, base, large, xl, xxl), OPT(125M, 350M, 1.3B, 2.7B, 6.7B, 13B, 30B, iml1.3B, iml-30B), T0 (3B, 11B), BLOOM-7B1, GPT-J-6B, GPT-NeoX-20B)	Paraphrase	Social Media, News Writing, QA, Story Generation, Comprehension and Reasoning, Scientific writing

CHEAT  
检测ChatGPT生成的虚假学术内容的最大的公共可访问资源。关注的学科太少，忽视了跨领域。

# 1.1 流行数据集介绍与对比- Dataset

Summary of Detection Datasets for LLM-generated text detection.

Corpus	Use	Human LLMs	LLMs Type	Language	Attack	Domain
HC3 (Guo et al. 2023)	train	~80k	~43k	ChatGPT	-	Web Text, QA, Social Media
CHEAT (Yu et al. 2023a)	train	~15k	~35k	ChatGPT	Paraphrase	Scientific Writing
HC3 Plus (Su et al. 2023b)	train valid test	~95k ~10k ~38k	GPT-3.5-Turbo	English, Chinese	Paraphrase	News Writing, Social Media
OpenLLMText (Chen et al. 2023a)	train, valid, test	~52k ~8k ~8k	~209k ~33k ~33k	ChatGPT, PaLM, LLaMA, GPT2-XL	-	Web Text
GROVER Dataset (Zellers et al. 2019b)	train	~24k	Grover-Mega	English	-	News Writing
TweepFake (Fagni et al. 2021)	train	~12k	~12k	GPT-2, RNN, Markov, LSTM, CharRNN	-	Social Media
GPT-2 Output Dataset <sup>6</sup>	train test	~250k ~5k	~2000k ~40k	GPT-2 (small, medium, large, xl)	-	Web Text
ArguGPT (Liu et al. 2023c)	train valid test	~6k 700 700	GPT2-XL, Text-Babbage-001, Text-Curie-001, Text-Davinci-001, Text-Davinci-002, Text-Davinci-003, GPT-3.5-Turbo	English	-	Scientific writing
DeepfakeTextDetect (Li et al. 2023c)	train valid test	~236k ~56k ~56k	GPT (Text-Davinci-002, Text-Davinci-003, GPT-Turbo-3.5), LLaMA (6B, 13B, 30B, 65B), GLM-130B, FLAN-T5 (small, base, large, xl, xxl), OPT(125M, 350M, 1.3B, 2.7B, 6.7B, 13B, 30B, iml1.3B, iml-30B), T0 (3B, 11B), BLOOM-7B1, GPT-J-6B, GPT-NeoX-20B)	English	Paraphrase	Social Media, News Writing, QA, Story Generation, Comprehension and Reasoning, Scientific writing

## HC3 Plus

HC3的增强版，引入了一个新部分专门针对需要语义不变性的任务，例如摘要、翻译和释义，包含3个数据集。Prompt仍然缺乏多样性。

# 1.1 流行数据集介绍与对比- Dataset

Summary of Detection Datasets for LLM-generated text detection.

Corpus	Use	Human LLMs	LLMs Type	Language	Attack	Domain	
HC3 (Guo et al. 2023)	train	~80k	~43k	ChatGPT	English, Chinese	-	Web Text, QA, Social Media
CHEAT (Yu et al. 2023a)	train	~15k	~35k	ChatGPT	English	Paraphrase	Scientific Writing
HC3 Plus (Su et al. 2023b)	train	~95k		GPT-3.5-Turbo	English, Chinese	Paraphrase	News Writing, Social Media
	valid	~10k					
	test	~38k					
OpenLLMText (Chen et al. 2023a)	train, valid, test	~52k, ~8k, ~8k	~209k, ~33k, ~33k	ChatGPT, PaLM, LLaMA, GPT2-XL	English	-	Web Text
GROVER Dataset (Zellers et al. 2019b)	train	~24k		Grover-Mega	English	-	News Writing
TweepFake (Fagni et al. 2021)	train	~12k	~12k	GPT-2, RNN, Markov, LSTM, CharRNN	English	-	Social Media
GPT-2 Output Dataset <sup>6</sup>	train, test	~250k, ~5k	~2000k, ~40k	GPT-2 (small, medium, large, xl)	English	-	Web Text
ArguGPT (Liu et al. 2023c)	train, valid, test	~6k, 700, 700		GPT2-XL, Text-Babbage-001, Text-Curie-001, Text-Davinci-001, Text-Davinci-002, Text-Davinci-003, GPT-3.5-Turbo	English	-	Scientific writing
DeepfakeTextDetect (Li et al. 2023c)	train, valid, test	~236k, ~56k, ~56k		GPT (Text-Davinci-002, Text-Davinci-003, GPT-Turbo-3.5), LLaMA (6B, 13B, 30B, 65B), GLM-130B, FLAN-T5 (small, base, large, xl, xxl), OPT(125M, 350M, 1.3B, 2.7B, 6.7B, 13B, 30B, iml1.3B, iml-30B), T0 (3B, 11B), BLOOM-7B1, GPT-J-6B, GPT-NeoX-20B)	English	Paraphrase	Social Media, News Writing, QA, Story Generation, Comprehension and Reasoning, Scientific writing

OpenLLMText  
4种大语言模型。  
分为训练集、验证集和测试集。  
并没有完全捕捉到跨领域和多语言文本的细微差别。

# 1.1 流行数据集介绍与对比- Dataset

Summary of Detection Datasets for LLM-generated text detection.

Corpus	Use	Human LLMs	LLMs Type	Language	Attack	Domain	
HC3 (Guo et al. 2023)	train	~80k	~43k	ChatGPT	English, Chinese	-	Web Text, QA, Social Media
CHEAT (Yu et al. 2023a)	train	~15k	~35k	ChatGPT	English	Paraphrase	Scientific Writing
HC3 Plus (Su et al. 2023b)	train	~95k		GPT-3.5-Turbo	English, Chinese	Paraphrase	News Writing, Social Media
	valid	~10k					
	test	~38k					
OpenLLMText (Chen et al. 2023a)	train, valid, test	~52k, ~8k, ~8k	~209k, ~33k, ~33k	ChatGPT, PaLM, LLaMA, GPT2-XL	English	-	Web Text
GROVER Dataset (Zellers et al. 2019b)	train	~24k		Grover-Mega	English	-	News Writing
TweepFake (Fagni et al. 2021)	train	~12k	~12k	GPT-2, RNN, Markov, LSTM, CharRNN	English	-	Social Media
GPT-2 Output Dataset <sup>6</sup>	train	~250k	~2000k	GPT-2 (small, medium, large, xl)	English	-	Web Text
	test	~5k	~40k				
ArguGPT (Liu et al. 2023c)	train, valid, test	~6k, 700, 700		GPT2-XL, Text-Babbage-001, Text-Curie-001, Text-Davinci-001, Text-Davinci-002, Text-Davinci-003, GPT-3.5-Turbo	English	-	Scientific writing
DeepfakeTextDetect (Li et al. 2023c)	train, valid, test	~236k, ~56k, ~56k		GPT (Text-Davinci-002, Text-Davinci-003, GPT-Turbo-3.5), LLaMA (6B, 13B, 30B, 65B), GLM-130B, FLAN-T5 (small, base, large, xl, xxl), OPT(125M, 350M, 1.3B, 2.7B, 6.7B, 13B, 30B, iml1.3B, iml-30B), T0 (3B, 11B), BLOOM-7B1, GPT-J-6B, GPT-NeoX-20B)	English	Paraphrase	Social Media, News Writing, QA, Story Generation, Comprehension and Reasoning, Scientific writing

GROVER Dataset  
 主要关注新闻文章。

# 1.1 流行数据集介绍与对比- Dataset

Summary of Detection Datasets for LLM-generated text detection.

Corpus	Use	Human LLMs	LLMs Type	Language	Attack	Domain
HC3 (Guo et al. 2023)	train	~80k	~43k	ChatGPT	-	Web Text, QA, Social Media
CHEAT (Yu et al. 2023a)	train	~15k	~35k	ChatGPT	Paraphrase	Scientific Writing
HC3 Plus (Su et al. 2023b)	train valid test	~95k ~10k ~38k	GPT-3.5-Turbo	English, Chinese	Paraphrase	News Writing, Social Media
OpenLLMText (Chen et al. 2023a)	train, valid, test	~52k ~8k ~8k	~209k ~33k ~33k	ChatGPT, PaLM, LLaMA, GPT2-XL	-	Web Text
GROVER Dataset (Zellers et al. 2019b)	train	~24k	Grover-Mega	English	-	News Writing
TweepFake (Fagni et al. 2021)	train	~12k	~12k	GPT-2, RNN, Markov, LSTM, CharRNN	-	Social Media
GPT-2 Output Dataset <sup>6</sup>	train test	~250k ~5k	~2000k ~40k	GPT-2 (small, medium, large, xl)	-	Web Text
ArguGPT (Liu et al. 2023c)	train valid test	~6k 700 700	GPT2-XL, Text-Babbage-001, Text-Curie-001, Text-Davinci-001, Text-Davinci-002, Text-Davinci-003, GPT-3.5-Turbo	English	-	Scientific writing
DeepfakeTextDetect (Li et al. 2023c)	train valid test	~236k ~56k ~56k	GPT (Text-Davinci-002, Text-Davinci-003, GPT-Turbo-3.5), LLaMA (6B, 13B, 30B, 65B), GLM-130B, FLAN-T5 (small, base, large, xl, xxl), OPT(125M, 350M, 1.3B, 2.7B, 6.7B, 13B, 30B, iml1.3B, iml-30B), T0 (3B, 11B), BLOOM-7B1, GPT-J-6B, GPT-NeoX-20B)	English	Paraphrase	Social Media, News Writing, QA, Story Generation, Comprehension and Reasoning, Scientific writing

TweepFake Dataset  
5种大语言模型。  
用于分析Twitter上的虚假推文，  
这些推文来自真实和虚假账户。

# 1.1 流行数据集介绍与对比- Dataset

Summary of Detection Datasets for LLM-generated text detection.

Corpus	Use	Human LLMs	LLMs Type	Language	Attack	Domain	
HC3 (Guo et al. 2023)	train	~80k	~43k	ChatGPT	English, Chinese	-	Web Text, QA, Social Media
CHEAT (Yu et al. 2023a)	train	~15k	~35k	ChatGPT	English	Paraphrase	Scientific Writing
HC3 Plus (Su et al. 2023b)	train	~95k		GPT-3.5-Turbo	English, Chinese	Paraphrase	News Writing, Social Media
	valid	~10k					
	test	~38k					
OpenLLMText (Chen et al. 2023a)	train,	~52k	~209k	ChatGPT, PaLM, LLaMA,	English	-	Web Text
	valid,	~8k	~33k	GPT2-XL			
	test	~8k	~33k				
GROVER Dataset (Zellers et al. 2019b)	train	~24k		Grover-Mega	English	-	News Writing
TweepFake (Fagni et al. 2021)	train	~12k	~12k	GPT-2, RNN, Markov, LSTM, CharRNN	English	-	Social Media
GPT-2 Output Dataset <sup>6</sup>	train	~250k	~2000k	GPT-2 (small, medium, large, xl)	English	-	Web Text
	test	~5k	~40k				
ArguGPT (Liu et al. 2023c)	train	~6k		GPT2-XL, Text-Babbage-001,	English	-	Scientific writing
	valid	700		Text-Curie-001, Text-Davinci-001,			
	test	700		Text-Davinci-002, Text-Davinci-003, GPT-3.5-Turbo			
DeepfakeTextDetect (Li et al. 2023c)	train	~236k		GPT (Text-Davinci-002, Text-Davinci-003, GPT-Turbo-3.5),	English	Paraphrase	Social Media, News Writing, QA, Story Generation, Comprehension and Reasoning, Scientific writing
	valid	~56k		LLaMA (6B, 13B, 30B, 65B),			
	test	~56k		GLM-130B, FLAN-T5 (small, base, large, xl, xxl), OPT(125M, 350M, 1.3B, 2.7B, 6.7B, 13B, 30B, iml1.3B, iml-30B), T0 (3B, 11B), BLOOM-7B1, GPT-J-6B, GPT-NeoX-20B)			

GPT2-Output Dataset  
来自Web Text测试集的250k个文档。  
旨在进一步研究GPT-2模型的可探测性。

# 1.1 流行数据集介绍与对比- Dataset

Summary of Detection Datasets for LLM-generated text detection.

Corpus	Use	Human LLMs	LLMs Type	Language	Attack	Domain
HC3 (Guo et al. 2023)	train	~80k	~43k	ChatGPT	-	Web Text, QA, Social Media
CHEAT (Yu et al. 2023a)	train	~15k	~35k	ChatGPT	Paraphrase	Scientific Writing
HC3 Plus (Su et al. 2023b)	train valid test	~95k ~10k ~38k	GPT-3.5-Turbo	English, Chinese	Paraphrase	News Writing, Social Media
OpenLLMText (Chen et al. 2023a)	train, valid, test	~52k ~8k ~8k	~209k ~33k ~33k	ChatGPT, PaLM, LLaMA, GPT2-XL	-	Web Text
GROVER Dataset (Zellers et al. 2019b)	train	~24k	Grover-Mega	English	-	News Writing
TweepFake (Fagni et al. 2021)	train	~12k	~12k	GPT-2, RNN, Markov, LSTM, CharRNN	-	Social Media
GPT-2 Output Dataset <sup>6</sup>	train test	~250k ~5k	~2000k ~40k	GPT-2 (small, medium, large, xl)	-	Web Text
ArguGPT (Liu et al. 2023c)	train valid test	~6k 700 700	GPT2-XL, Text-Babbage-001, Text-Curie-001, Text-Davinci-001, Text-Davinci-002, Text-Davinci-003, GPT-3.5-Turbo	English	-	Scientific writing
DeepfakeTextDetect (Li et al. 2023c)	train valid test	~236k ~56k ~56k	GPT (Text-Davinci-002, Text-Davinci-003, GPT-Turbo-3.5), LLaMA (6B, 13B, 30B, 65B), GLM-130B, FLAN-T5 (small, base, large, xl, xxl), OPT(125M, 350M, 1.3B, 2.7B, 6.7B, 13B, 30B, iml1.3B, iml-30B), T0 (3B, 11B), BLOOM-7B1, GPT-J-6B, GPT-NeoX-20B)	English	Paraphrase	Social Media, News Writing, QA, Story Generation, Comprehension and Reasoning, Scientific writing

**ArguGPT Dataset**  
7种大语言模型。  
专门用于检测各种学术环境(如课堂练习、托福和GRE写作任务)中机器生成的文本。

# 1.1 流行数据集介绍与对比- Dataset

Summary of Detection Datasets for LLM-generated text detection.

Corpus	Use	Human LLMs	LLMs Type	Language	Attack	Domain
HC3 (Guo et al. 2023)	train	~80k	~43k	ChatGPT	-	Web Text, QA, Social Media
CHEAT (Yu et al. 2023a)	train	~15k	~35k	ChatGPT	Paraphrase	Scientific Writing
HC3 Plus (Su et al. 2023b)	train	~95k		GPT-3.5-Turbo	Paraphrase	News Writing, Social Media
	valid	~10k				
	test	~38k				
OpenLLMText (Chen et al. 2023a)	train,	~52k	~209k	ChatGPT, PaLM, LLaMA,	-	Web Text
	valid,	~8k	~33k	GPT2-XL		
	test	~8k	~33k			
GROVER Dataset (Zellers et al. 2019b)	train	~24k		Grover-Mega	-	News Writing
TweepFake (Fagni et al. 2021)	train	~12k	~12k	GPT-2, RNN, Markov, LSTM, CharRNN	-	Social Media
GPT-2 Output Dataset <sup>6</sup>	train	~250k	~2000k	GPT-2 (small, medium, large, xl)	-	Web Text
	test	~5k	~40k			
ArguGPT (Liu et al. 2023c)	train	~6k		GPT2-XL, Text-Babbage-001,	-	Scientific writing
	valid	700		Text-Curie-001, Text-Davinci-001, Text-Davinci-002, Text-Davinci-003, GPT-3.5-Turbo		
	test	700				
DeepfakeTextDetect (Li et al. 2023c)	train	~236k		GPT (Text-Davinci-002, Text-Davinci-003, GPT-Turbo-3.5), LLaMA (6B, 13B, 30B, 65B), GLM-130B, FLAN-T5 (small, base, large, xl, xxl), OPT(125M, 350M, 1.3B, 2.7B, 6.7B, 13B, 30B, iml1.3B, iml-30B), T0 (3B, 11B), BLOOM-7B1, GPT-J-6B, GPT-NeoX-20B)	Paraphrase	Social Media, News Writing, QA, Story Generation, Comprehension and Reasoning, Scientific writing
	valid	~56k				
	test	~56k				

DeepfakeTextDetect Dataset 用于深度伪造文本检测。10个不同数据集（新闻文章、故事、科学著作等）。27个大语言模型。

# 1.1 流行数据集介绍与对比- Potential Dataset

从0开始构建数据集非常繁琐  $\implies$  从现有人类编写的数据集，延伸出大模型检测的数据集

Summary of other potential datasets that can easily extended to LLM-generated text detection tasks.

Corpus	Size	Source	Language	Domain
XSum (Narayan, Cohen, and Lapata 2018)	42k	BBC	English	News Writing
SQuAD (Rajpurkar et al. 2016)	98.2k	Wiki	English	Question Answering
WritingPrompts (Fan, Lewis, and Dauphin 2018)	302k	Reddit WRITINGPROMPTS	English	Story Generation
Wiki40B (Guo et al. 2020)	17.7m	Wiki	40+ Languages	Web Text
PubMedQA (Jin et al. 2019)	211k	PubMed	English	Question Answering
Children's Book Corpus (Hill et al. 2016)	687k	Books	English	Question Answering
Avax Tweets Dataset (Muric, Wu, and Ferrara 2021)	137m	Twitter	English	Social Media
Climate Change Dataset (Littman and Wrubel 2019)	4m	Twitter	English	Social Media
Yelp Dataset (Asghar 2016)	700k	Yelp	English	Social Media
ELI5 (Fan et al. 2019)	556k	Reddit	English	Question Answering
ROCStories (Mostafazadeh et al. 2016)	50k	Crowdsourcing	English	Story Generation
HellaSwag (Zellers et al. 2019a)	70k	ActivityNet Captions, Wikihow	English	Question Answering
SciGen (Moosavi et al. 2021)	52k	arXiv	English	Scientific Writing, Question Answering
WebText (Radford et al. 2019)	45m	Web	English	Web Text
TruthfulQA (Lin, Hilton, and Evans 2022)	817	authors writtEnglish	English	Question Answering
NarrativeQA (Kočíský et al. 2018)	1.4k	Gutenberg3, web	English	Question Answering
TOEFL11 (Blanchard et al. 2013)	12k	TOEFL test	11 Languages	Scientific writing
Peer Reviews (Kang et al. 2018)	14.5k	NIPS 2013–2017, CoNLL 2016, ACL 2017 ICLR 2017, arXiv 2007–2017	English	Scientific Writing

- ① Q&A: 相同的问题, 回答
- ② Scientific Writing: 给定学术主题
- ③ Story Generation: 写故事
- ④ News Writing: 写新闻
- ⑤ Web Text: 网络文本数据, 比较广泛, 主要源于Wikipedia等
- ⑥ Social Media: 主观表达能力
- ⑦ Comprehension and Reasoning: 理论与推理

# 1.2 24年新数据集介绍-M4

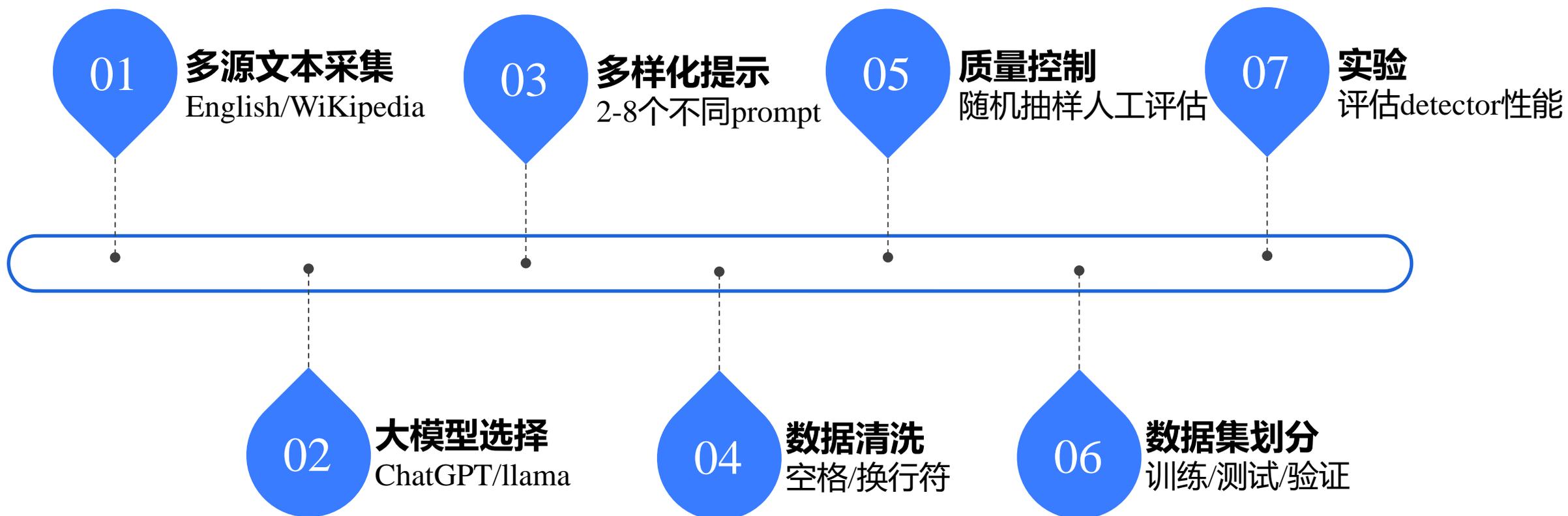
**亮点:** Multi-Generator, Multi-Domain, and Multi-Lingual

该数据集捕捉到了跨语言的微妙之处，包含了多种语言的内容。提升多样性。

Source/ Domain	Data License	Language	Total Human	Parallel Data							
				Human	Davinci003	ChatGPT	GPT4	Cohere	Dolly-v2	BLOOMz	Total
Wikipedia	CC BY-SA-3.0	English	6,458,670	3,000	3,000	2,995	3,000	2,336	2,702	3,000	20,033
Reddit ELI5	Huggingface	English	558,669	3,000	3,000	3,000	3,000	3,000	3,000	3,000	21,000
WikiHow	CC-BY-NC-SA	English	31,102	3,000	3,000	3,000	3,000	3,000	3,000	3,000	21,000
PeerRead	Apache license	English	5,798	5,798	2,344	2,344	2,344	2,344	2,344	2,344	19,862
arXiv abstract	CC0-public domain	English	2,219,423	3,000	3,000	3,000	3,000	3,000	3,000	3,000	21,000
Arabic-Wikipedia	CC BY-SA-3.0	Arabic	1,209,042	3,000	-	3,000	-	-	-	-	6,000
True & Fake News	MIT License	Bulgarian	94,000	3,000	3,000	3,000	-	-	-	-	9,000
Baike/Web QA	MIT license	Chinese	113,313	3,000	3,000	3,000	-	-	-	-	9,000
id_newspapers_2018	CC BY-NC-SA-4.0	Indonesian	499,164	3,000	-	3,000	-	-	-	-	6,000
RuATD	Apache 2.0 license	Russian	75,291	3,000	3,000	3,000	-	-	-	-	9,000
Urdu-news	CC BY 4.0	Urdu	107,881	3,000	-	3,000	-	-	-	-	6,000
<b>Total</b>				35,798	23,344	32,339	14,344	13,680	14,046	14,344	147,895

# 1.2 24年新数据集介绍-M4

## ■ 构建步骤



## 1.2 24年新数据集介绍-M4GT-Bench

M4数据集的扩展版，涉及**9种语言**、**6个领域**、**9个大语言模型**和**3个不同的任务**。

提出了基于M4数据集的新基准，包含三个任务：

- ① 二元MGT检测，是否由机器生成。
- ② 多路检测，识别由哪一个特定的模型生成文本。
- ③ **混合人机文本检测，识别人类编写文本和机器生成文本的边界。**

**(局限：本文假设混合文本首先由人编写，然后由机器继续编写，任务是检测变化的单一边界。但实际情况要复杂得多。)**

# 1.2 24年新数据集介绍-M4GT-Bench

任务一：二元检测。

- 数据集基于M4数据集**扩展**而来，包含65,177个人类写作的文本和73,288个机器生成的文本。
- 为了解决数据不平衡问题，对人类文本进行了**上采样**。
- 还包括了使用**GPT-4生成**的每个领域的文本，提升泛化能力。

Source Domain	Total	Human Upsample+	Parallel	Parallel Data					Total Machine	New test GPT-4
				davinci-003	ChatGPT	Cohere	Dolly-v2	BLOOMz		
<b>OUTFOX</b>	16,272	13,272	3,000	3,000	3,000	3,000	3,000	3,000	15,000	3,000
Wikipedia	14,333	11,997	2,336	3,000	2,995	2,336	2,702	2,999	14,032	3,000
Wikihow	15,999	13,000	2,999	3,000	5,557	3,000	3,000	3,000	17,557	3,000
Reddit ELI5	16,000	13,000	3,000	3,000	3,000	3,000	3,000	2,999	14,999	3,000
arXiv abstract	15,998	13,000	2,998	3,000	3,000	3,000	3,000	3,000	15,000	3,000
PeerRead	2,847	0	2,847	2,340	2,340	2,342	2,344	2,334	11,700	2,334
Total	65,177	50,997	14,180	14,340	16,892	13,678	14,046	14,332	73,288	14,344

Table 1: **Tasks 1 and 2 data statistics:** all data used for Task 1; data without upsampled human for Task 2. The first row (OUTFOX) and the last column (GPT-4) represent newly generated data added to the M4 (Wang et al., 2023).

# 1.2 24年新数据集介绍-M4GT-Bench

任务一：二元检测。

- 引入了**新的语言**（德语和意大利语），以及由ChatGPT和Jais-30B生成的阿拉伯语文本。

Source/ Domain	Data License	Language	Total Human	Parallel Data					
				Human	davinci-003	ChatGPT	Jais	LLaMA-2	Total
Arabic-Wikipedia	CC BY-SA-3.0	Arabic	1,209,042	3,000	–	3,000	–	–	6,000
True & Fake News	MIT License	Bulgarian	94,000	3,000	3,000	3,000	–	–	9,000
Baike/Web QA	MIT license	Chinese	113,313	3,000	3,000	3,000	–	–	9,000
id_newspapers_2018	CC BY-NC-SA-4.0	Indonesian	499,164	3,000	–	3,000	–	–	6,000
RuATD	Apache 2.0 license	Russian	75,291	3,000	3,000	3,000	–	–	9,000
Urdu-news	CC BY 4.0	Urdu	107,881	3,000	–	3,000	–	–	6,000
<b>News</b>	Apache 2.0	Arabic	1,000	1,000	–	1,000	100	–	2,100
<b>CHANGE-it News</b>	CC BY-NC-SA 4.0	Italian	127,402	3,000	–	–	–	3,000	6,000
<b>News</b>	CC BY-NC-SA-4.0	German	10,000	3,000	–	3,000	–	–	6,000
<b>Wikipedia</b>	CC BY-SA-3.0	German	2,882,103	3,000	–	3,000	–	–	6,000
<b>Total</b>	–	–	5,119,196	28,000	9,000	25,000	100	3,000	65,100

Table 2: **Task 1 Multilingual** introduced new languages: German, Italian, news for Arabic by ChatGPT and Jais-30B. LLaMA-2-70B used here for generating Italian texts is a fine-tuned Italian version, named *camoscio-70B*.

# 1.2 24年新数据集介绍-M4GT-Bench

## 任务二：多路检测

- 包括**六个生成器**：ChatGPT、davinci-003、GPT-4、Cohere、Dolly-v2和BLOOMz。
- 收集了一个**新的领域OUTFOX**，用于评估分类器在**学生论文**中的领域泛化能力。

Source Domain	Human Total=	Upsample+	Parallel	Parallel Data					Total Machine	New test GPT-4
				davinci-003	ChatGPT	Cohere	Dolly-v2	BLOOMz		
<b>OUTFOX</b>	16,272	13,272	3,000	3,000	3,000	3,000	3,000	3,000	15,000	3,000
Wikipedia	14,333	11,997	2,336	3,000	2,995	2,336	2,702	2,999	14,032	3,000
Wikihow	15,999	13,000	2,999	3,000	5,557	3,000	3,000	3,000	17,557	3,000
Reddit ELI5	16,000	13,000	3,000	3,000	3,000	3,000	3,000	2,999	14,999	3,000
arXiv abstract	15,998	13,000	2,998	3,000	3,000	3,000	3,000	3,000	15,000	3,000
PeerRead	2,847	0	2,847	2,340	2,340	2,342	2,344	2,334	11,700	2,334
Total	65,177	50,997	14,180	14,340	16,892	13,678	14,046	14,332	73,288	14,344

Table 1: **Tasks 1 and 2 data statistics:** all data used for Task 1; data without upsampled human for Task 2. The first row (OUTFOX) and the last column (GPT-4) represent newly generated data added to the M4 (Wang et al., 2023).

# 1.2 24年新数据集介绍-M4GT-Bench

## 任务三：混合人机文本检测

(本文假设混合文本首先由人编写，然后由机器继续编写，任务是检测变化的单一边界。)

- 特别为**学术论文评论** (PeerRead) 和**学生论文** (OUTFOX) 两个领域生成了混合文本，人类编写的比例从0-50%不等。
- 使用ChatGPT、GPT-4和LLaMA-2系列生成了5,676个和1,000个示例。

Domain	Generator	Train	Dev	Test	Total
PeerRead	ChatGPT	3,649 (232)	505 (23)	1,522 (89)	5,676 (344)
	LLaMA-2-7B*	3,649 (5)	505 (0)	1,035 (1)	5,189 (6)
	LLaMA-2-7B	3,649 (227)	505 (24)	1,522 (67)	5,676 (318)
	LLaMA-2-13B	3,649 (192)	505 (24)	1,522 (84)	5,676 (300)
	LLaMA-2-70B	3,649 (240)	505 (21)	1,522 (88)	5,676 (349)
OUTFOX	GPT-4	–	–	1,000 (10)	1,000 (10)
	LLaMA2-7B	–	–	1,000 (8)	1,000 (8)
	LLaMA2-13B	–	–	1,000 (5)	1,000 (5)
	LLaMA2-70B	–	–	1,000 (19)	1,000 (19)

Table 3: **Task 3 boundary identification data** based on GPT and LLaMA-2 series over domains of academic paper review (PeerRead) and student essay (OUTFOX). The number in “()” is the number of examples purely generated by LLMs, i.e., human and machine boundary index=0. LLaMA-2-7B\* and LLaMA-2-7B used different prompts.

# 1.2 24年新数据集介绍-M4GT-Bench

## 任务三：混合人机文本检测

Complete a partially *written peer review* about the paper: {paper\_title}

Here is the abstract of the paper: {paper\_abstract}

Here is the partial review: {partial\_review}

Make sure:

1. Continue to generate at least {num\_of\_words} words.

2. Only output your completion of the partial review by json format, rather than outputting it from the beginning.

Output:

-----  
Act as an experienced *essay writer*.

Given the following [problem statement]:

Explain the reasons why the Electoral College system is being opposed by some people and argue for or against its continuation in the United States presidential elections.

Please write an essay of at least **318** words with a clear opinion. The written essay should look like human.

Here is the [partial\_essay]:

The Electoral College is a system of electing our president created by our founding fathers when they were writing the Constitution. Lately though, people have opposed the Electoral College due to the election when Al Gore beat George Bush in the popular vote, but lost the electoral vote (and a seat as President I do not support the Electoral College

Please continue to write without any additional text:

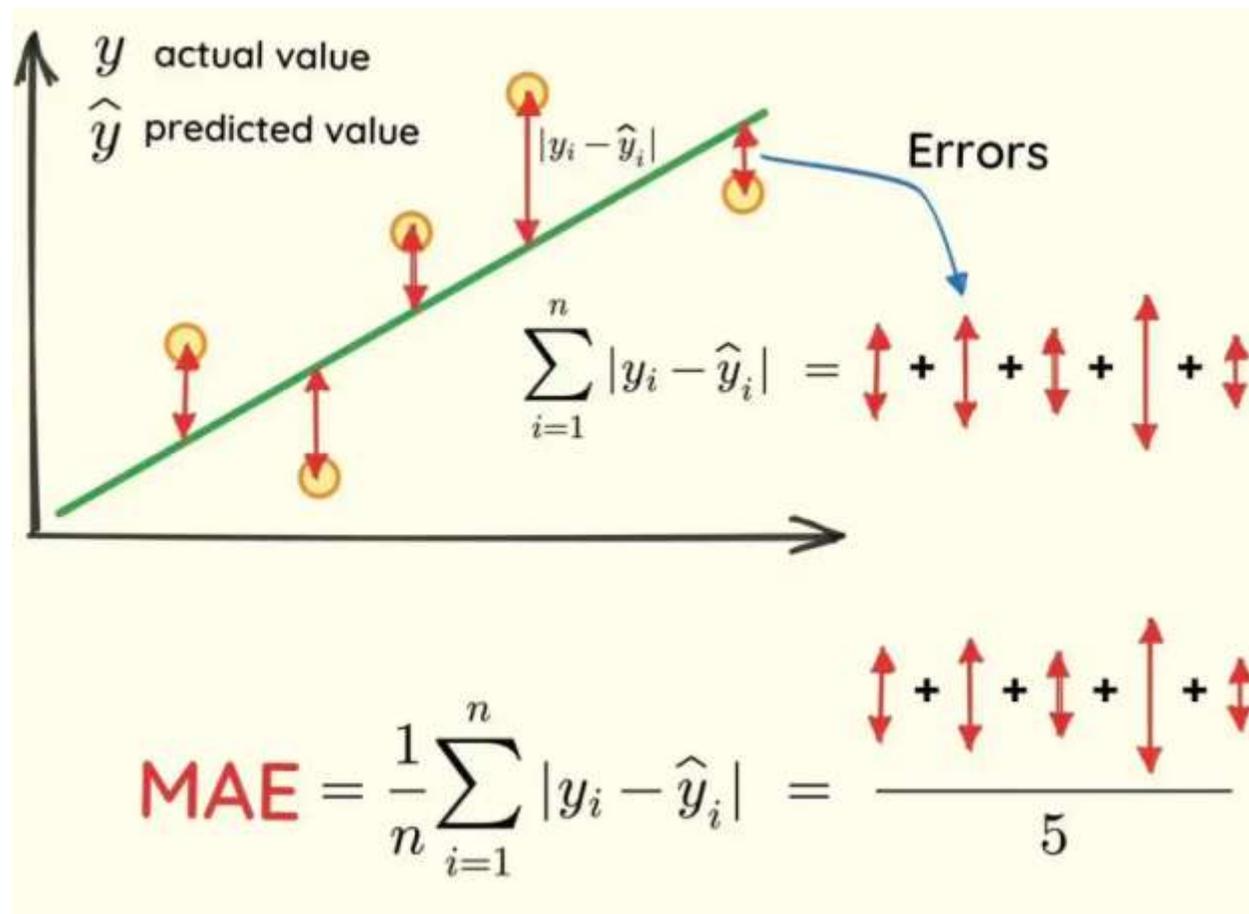
Figure 2: Task 3 prompt templates used to generate continuations of paper reviews and student essays.

# 1.2 24年新数据集介绍-M4GT-Bench

任务三：混合人机文本检测

## 平均绝对误差(MAE)

评价边界检测模型的性能。  
预测位置指数与实际变化点之间的平均绝对差值。



# 1.2 24年新数据集介绍-M4GT-Bench

## 任务三：混合人机文本检测

### ■ Peerread:

Detector	Train Data	Peerread LLaMA-2-7B*	Peerread ChatGPT	All Test
Longformer	All	$1.89 \pm 0.79$	$4.36 \pm 0.36$	$21.54 \pm 0.25$
	ChatGPT	$31.43 \pm 6.15$	$4.55 \pm 0.36$	$25.14 \pm 0.93$
	LLaMA-2-7B*	$1.94 \pm 0.07$	$51.379 \pm 0.72$	$53.62 \pm 1.60$
DeBERTa-v3	All	$0.57 \pm 0.23$	$2.63 \pm 0.20$	$15.55 \pm 2.60$
	ChatGPT	$14.96 \pm 2.19$	$2.53 \pm 0.09$	$19.67 \pm 1.05$
	LLaMA-2-7B*	$0.66 \pm 0.12$	$24.59 \pm 4.07$	$32.35 \pm 0.78$

Table 8: **Task 3 MAE** for Longformer and Deberta-v3 under (1) cross-generator setting for PeerRead, and (2) unseen domains with multiple generators (*All test*). Training data is PeerRead using LLaMA-2-7B\* and ChatGPT.

■ **OUTFOX:** 对于使用LLaMA-2数据训练的Longformer, 在所有测试中, MAE都大于53

# 1.2 24年新数据集介绍-RAID (Robust AI Detection)

**亮点:** 跨越11种模型, 8个域, **11种对抗性攻击和4种解码策略**, 并对12个detector(8个开源和4个闭源)进行基准测试, 具有对detector鲁棒性的检测, 目前**最大的数据集**。但**只有英语**。

Models			Domains		Decoding Strategy	
GPT-4	GPT-2 XL	GPT-3	Abstracts	Recipes	Greedy (temp. = 0)	
Cohere	Cohere (Chat)		Books	Reddit	Sampling (temp. = 1, p = 1)	
MPT-30B	MPT-30B (Chat)		News	Reviews		
Mistral-7B	Mistral-7B (Chat)		Poetry	Wikipedia		
ChatGPT	LLaMA 2 70B (Chat)					
<i>11 models</i>			<i>8 domains</i>		<b>Repetition Penalty</b>	
					With ✓ (rep = 1.2)	
					Without ✗ (rep = 1.0)	
Detectors			Adversarial Attacks			
Neural	Metric-Based	Commercial	Alternative Spelling	Homoglyph		
RoBERTa-B (GPT-2)	GLTR	GPTZero	Article Deletion	Number Swap		
RoBERTa-L (GPT-2)	Fast DetectGPT	Originality	Insert Paragraphs	Paraphrase		
RoBERTa-B (ChatGPT)	Binoculars	Winston	Upper Lower Swap	Synonym Swap		
RADAR	LLMDet	ZeroGPT	Zero-Width Space	Misspelling		
<i>12 detectors</i>			Whitespace Addition		<i>11 attacks</i>	

## 1.2 24年新数据集介绍-RAID (Robust AI Detection)

**亮点：** 跨越11种模型， 8个域， **11种对抗性攻击和4种解码策略**， 并对12个detector(8个开源和4个闭源)进行基准测试， 具有对detector鲁棒性的检测， 目前**最大的数据集**。 但**只有英语**。

Name	Size	Domain coverage?	Model coverage?	Sampling coverage?	Multilingual coverage?	Adversarial coverage?
TuringBench (Uchendu et al., 2021)	200k	✗	✓	✗	✗	✗
RuATD (Shamardina et al., 2022)	215k	✓	✓	✗	✗	✗
HC3 (Guo et al., 2023)	26.9k	✓	✗	✗	✓	✗
MGTBench (He et al., 2023)	2817	✓	✓	✗	✗	✓
MULTITuDE (Macko et al., 2023)	74.1k	✗	✓	✗	✓	✗
AuText2023 (Sarvazyan et al., 2023b)	160k	✓	✗	✗	✓	✗
M4 (Wang et al., 2023b)	122k	✓	✓	✗	✓	✗
CCD (Wang et al., 2023a)	467k	✗	✗	✗	✓	✓
IMDGSP (Mosca et al., 2023)	29k	✗	✓	✗	✗	✗
HC-Var (Xu et al., 2023)	145k	✓	✗	✗	✗	✗
HC3 Plus (Su et al., 2024)	210k	✓	✗	✗	✓	✗
MAGE (Li et al., 2024)	447k	✓	✓	✗	✗	✗
<b>RAID (Ours)</b>	6.2M	✓	✓	✓	✗	✓

# 1.2 24年新数据集介绍-RAID (Robust AI Detection)

Adversarial Attacks	
Alternative Spelling	Homoglyph
Article Deletion	Number Swap
Insert Paragraphs	Paraphrase
Upper Lower Swap	Synonym Swap
Zero-Width Space	Misspelling
Whitespace Addition	

11 attacks

鲁棒性



Decoding Strategy	
Greedy	(temp. = 0)
Sampling	(temp. = 1, p = 1)

Repetition Penalty	
With ✓	(rep = 1.2)
Without ✗	(rep = 1.0)

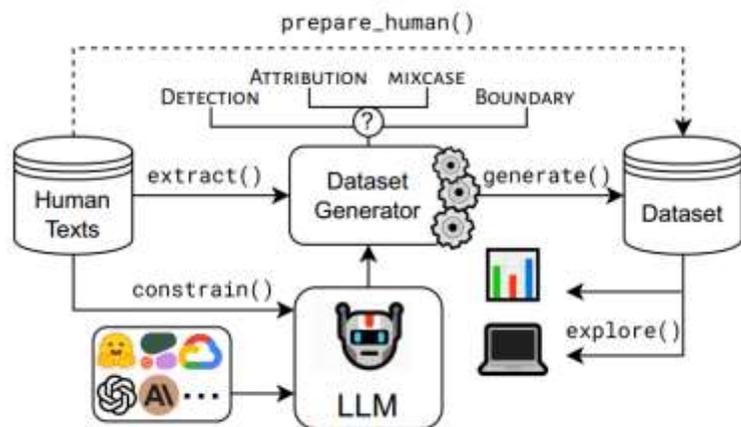
1. 替代拼写 (Alternative Spelling)
2. 冠词删除 (Article Deletion)
3. 段落插入 (Add Paragraph)
4. 大小写转换 (Upper-Lower Swap)
5. 零宽度空间 (Zero-Width Space)
6. 空白字符 (Whitespace)
7. 同形异义字 (Homoglyph)
8. 数字乱序 (Number Swap)
9. 改述替换 (Paraphrase)
10. 拼写错误 (Misspelling)
11. 同义词替换 (Synonym Swap)

为每一个prompt生成四个输出，对应四个解码策略：

1. **贪婪解码 (Greedy Decoding)**：生成文本时，选择概率最高的词作为下一个词。
2. **随机采样 (Random Sampling)**：随机选取词。
3. **重复惩罚 (Repetition Penalty)**：降低重复出现的文本，通过乘法因子 $\theta$ 来实现的， $\theta=1.2$ 。
4. **重复惩罚 (Repetition Penalty)**： $\theta=1.0$ 。

# 1.2 24年新数据集介绍-TEXTMACHINA

**亮点：**一个模块化和可扩展的Python框架，旨在帮助创建高质量、无偏的数据集，允许用户通过自定义配置来生成数据集。



```
from text_machina import get_generator
from text_machina import Config,
    InputConfig, ModelConfig

config = Config(
    input=InputConfig(...),
    model=ModelConfig(...),
    generation={...},
    task_type="detection",
)

generator = get_generator(config)
dataset = generator.generate()
```

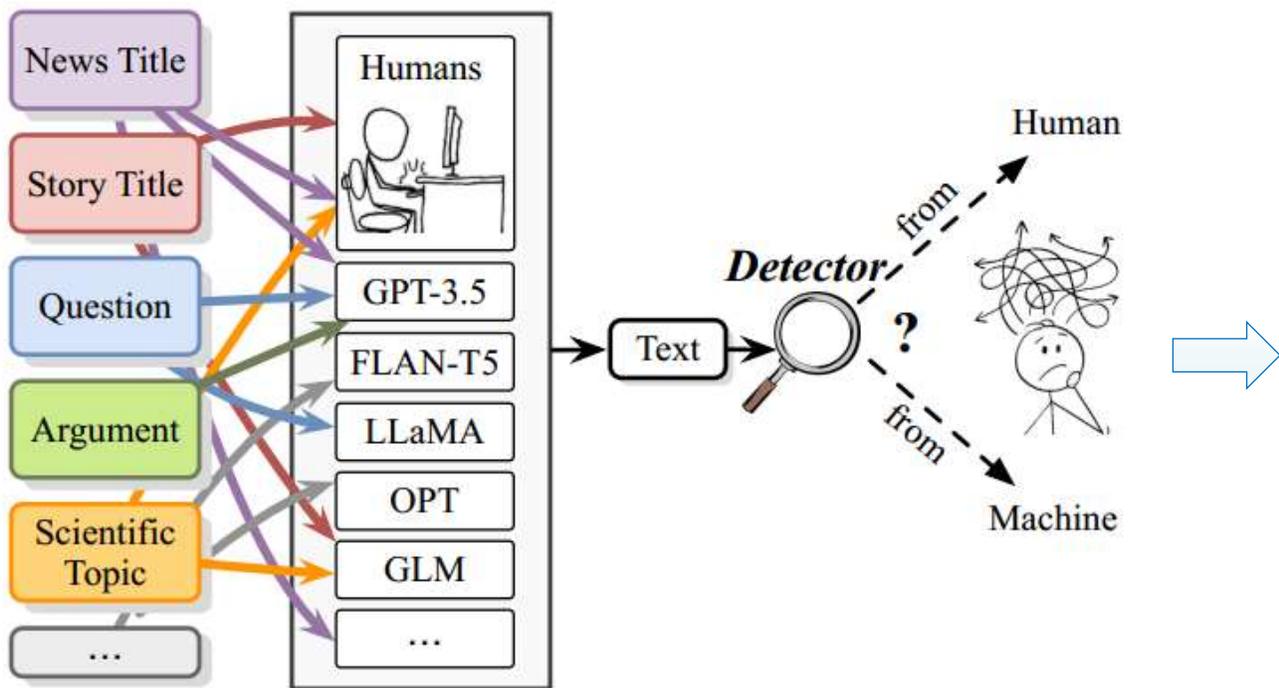
```
1 # Config for everything related to dataset generation inputs
2 input_config:
3 # Dataset metadata
4 domain: news
5 language: en
6
7 # Dataset generator parameters
8 quantity: 10
9 random_sample_human: true
10
11 # HuggingFace dataset params
12 dataset: xsum
13 dataset_text_column: document
14 dataset_params:
15 split: test
16
17 # Prompt template
18 template: >-
19 Write a news article whose summary is '{summary}',
20 using the entities: {entities}\n\nArticle:
21
22 # Extractor params
23 extractor: combined
24 extractors_list:
25 - auxiliary.Auxiliary
26 - entity_list.EntityList
27 max_input_tokens: 256
28
29 # Config for model instantiation
30 model_config:
31 provider: openai
32 model_name: gpt-3.5-turbo-instruct
33 api_type: completion
34 threads: 8
35 max_retries: 5
36 timeout: 120
37
38 # Decoding args
39 generation_config:
40 # Ignore use `max_tokens` to get automatic length estimation
41 # max_tokens: 100
42 temperature: 0.7
43 presence_penalty: 1.0
```

# 1.2 24年新数据集介绍-Data Challenges

---

- **multiple types of attacks**: 有助于确定检测方法的有效性、鲁棒性
- **diverse domains and varied tasks**: 对检测器的健壮性、可用性和可信度具有重要意义
- **multiple LLMs**: 检测多种模型
- **multiple languages**: 相同的问题不同的语言有不同的回答
- **temporal**: 更新数据库

## 2 评测维度-评测场景



现有研究多在**特定领域或特定语言模型**上评估检测方法，而实际应用中需要面对**未知来源的多样化文本**。

1. 现有的detector能否有效区分现实场景中，不同llm针对**不同任务**生成的文本？
2. 在**开放领域**设置中，无论主题或内容如何，人类编写的文本和机器生成的文本之间是否存在固有的区别？

## 2 评测维度-评测场景

构建了一个大规模的机器生成文本检测测试平台**MAGE** (MAchine-GEnerated text detection,):

- **7个**不同写作任务(如故事生成、新闻写作、科学写作和常识推理)的人类编写文本
- 使用**27个**llm(如ChatGPT、LLaMA和Bloom)生成相应的机器生成文本
- 将数据分类到**8个**评测场景中, 每个场景在分布方差和检测复杂性方面都表现出越来越高的wild水平。

检测难度加大

1. **固定领域和特定模型 (Fixed-domain & Model-specific):**
2. **任意领域和特定模型 (Arbitrary-domains & Model-specific):**
3. **固定领域和任意模型 (Fixed-domain & Arbitrary-models):**
4. **任意领域和任意模型 (Arbitrary-domains & Arbitrary-models):**
5. **未见模型 (Unseen Models):**
6. **未见领域 (Unseen Domains):**
7. **未见领域和未见模型 (Unseen-domains & Unseen-model):**
8. **改述攻击 (Paraphrasing Attack):**

## 2 评测维度-评测场景

Methods	Human/Machine	AvgRec	AUROC
FastText	94.72%/94.36%	94.54%	0.98
GLTR	90.96%/83.94%	87.45%	0.94
Longformer	97.30%/95.91%	<b>96.60%</b>	<b>0.99</b>
DetectGPT	91.68%/81.06%	86.37%	0.92

Table 2: (Testbed 1) White-box detection performance. “Human/Machine” denotes HumanRec and MachineRec, respectively.

Settings	Methods	Metrics			
		HumanRec	MachineRec	AvgRec	AUROC
Testbed 2,3,4: In-distribution Detection					
Arbitrary-domains & Model-specific	FastText (Joulin et al., 2017)	88.96%	77.08%	83.02%	0.89
	GLTR (Gehrmann et al., 2019)	75.61%	79.56%	77.58%	0.84
	Longformer (Beltagy et al., 2020)	95.25%	96.94%	<b>96.10%</b>	<b>0.99</b>
	DetectGPT* (Mitchell et al., 2023)	48.67%	75.95%	62.31%	0.60
Fixed-domain & Arbitrary-models	FastText (Joulin et al., 2017)	89.43%	73.91%	81.67%	0.89
	GLTR (Gehrmann et al., 2019)	37.25%	88.90%	63.08%	0.80
	Longformer (Beltagy et al., 2020)	89.78%	97.24%	<b>93.51%</b>	<b>0.99</b>
	DetectGPT* (Mitchell et al., 2023)	86.92%	34.05%	60.48%	0.57
Arbitrary-domains & Arbitrary-models	FastText (Joulin et al., 2017)	86.34%	71.26%	78.80%	0.83
	GLTR (Gehrmann et al., 2019)	12.42%	98.42%	55.42%	0.74
	Longformer (Beltagy et al., 2020)	82.80%	98.27%	<b>90.53%</b>	<b>0.99</b>
	DetectGPT* (Mitchell et al., 2023)	86.92%	34.05%	60.48%	0.57
Testbed 5,6: Out-of-distribution Detection					
Unseen Models	FastText (Joulin et al., 2017)	83.12%	54.09%	68.61%	0.74
	GLTR (Gehrmann et al., 2019)	25.77%	89.21%	57.49%	0.65
	Longformer (Beltagy et al., 2020)	83.31%	89.90%	<b>86.61%</b>	<b>0.95</b>
	DetectGPT* (Mitchell et al., 2023)	48.67%	75.95%	62.31%	0.60
Unseen Domains	FastText (Joulin et al., 2017)	54.29%	72.79%	63.54%	0.72
	GLTR (Gehrmann et al., 2019)	15.84%	97.12%	56.48%	0.72
	Longformer (Beltagy et al., 2020)	38.05%	98.75%	<b>68.40%</b>	<b>0.93</b>
	DetectGPT* (Mitchell et al., 2023)	86.92%	34.05%	60.48%	0.57

## 2 评测维度-评测场景

HumanRec	MachineRec	AvgRec	AUROC
Testbed 7: Unseen Domains & Unseen Model			
52.50%	99.14%	75.82%	0.94
88.78 <sup>†</sup>	84.12% <sup>†</sup>	86.54% <sup>†</sup>	0.94
Testbed 8: Paraphrasing Attack			
52.16%	81.73%	66.94%	0.75
88.78% <sup>†</sup>	37.05% <sup>†</sup>	62.92% <sup>†</sup>	0.75

Table 5: (Testbed 7-8) Detection performance of **Longformer** detector on the two challenging test sets. <sup>†</sup>denotes the refined decision boundary. Appendix G includes the performance of other detection methods.

## 2 评测维度-评测场景

---

### Result:

- 当文本来自单一领域或由有限范围的LLMs生成时，所有检测方法都有效。
- 但随着领域和模型多样性的增加，除了基于PLM（Pre-trained Language Model）的检测器外，其他方法性能显著下降。
- 在面对未知领域（OOD）测试组时，即使是最佳性能的检测器也难以准确分类。

## 2 评测维度-攻击方式

测试detector对恶意攻击的鲁棒性。

本文研究了**8种**MGT检测器在**12种**实际攻击下的鲁棒性，包括编辑、转述、提示、共同生成。

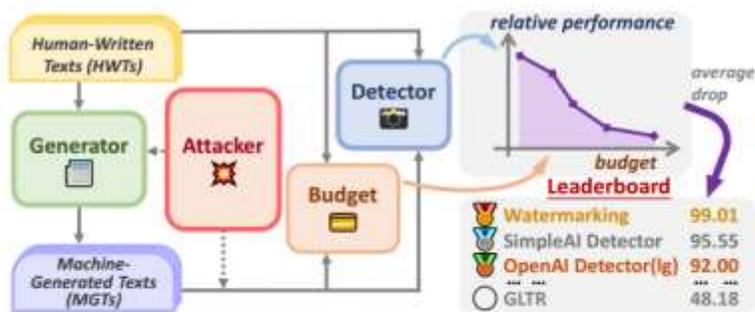


Figure 1: **Pipeline of the study.** The attacks are carried out on the machine-generated texts before, during, or after generation. Each attack is applied with different perturbation levels, denoted as budgets (§4).

### Budget:

每次攻击的扰动程度。

利用一系列文本生成评估指标作为攻击的预算。

Metric	Scale	Definition
Levenshtein Edit Distance (Levenshtein, 1965)	$\geq 0 \uparrow$	The minimum number of single-character edits (insertions, deletions or substitutions).
Jaro Similarity (Jaro, 1989)	$\geq 0 \downarrow$	A similarity metric based on matching characters and transpositions in two strings.
Perplexity (PPL)	$> 0 \leftrightarrow$	Apply Llama-7B-hf (Touvron et al., 2023b). MGTs to estimate the model distribution $Q$ and HWTs to estimate the target distribution $P$ . For attacked scenarios, the closer value to the unattacked scenario is favored.
MAUVE (Pillutla et al., 2021)	M2H $(0, 1] \leftrightarrow$ A2B $(0, 1] \downarrow$	MGTs (attacked) to estimate the model distribution $Q$ and MGTs (unattacked) to estimate the target distribution $P$ .
Cosine Similarity	$[-1, 1] \downarrow$	Utilize BART embedding (Lewis et al., 2020) to compare the similarity of texts after the attack to before the attack.
BERTScore (Zhang et al., 2019)	M2H $[0, 1] \leftrightarrow$ A2B $[0, 1] \downarrow$	MGTs as the candidates $\hat{x}$ and HWTs as the reference $x$ . For attacked scenarios, the closer value to the unattacked scenario is favored. MGTs (attacked) as the candidates $\hat{x}$ and MGTs (unattacked) as the reference $x$ .
BARTScore (Yuan et al., 2021)	M2H $< 0 \leftrightarrow$ A2B $< 0 \downarrow$	MGTs as the source $x$ and HWTs as the target $y$ . For attacked scenarios, the closer value to the unattacked scenario is favored. MGTs (attacked) as the source $x$ and MGTs (unattacked) as the target $y$ .

# 2 评测维度-攻击方式

➤ 编辑攻击：在字符级别进行小编辑，一些攻击可能会导致文本稍微失去质量和可读性。

1. Typo Insertion (错误插入)

2. Homoglyph Alteration (同形异义词替换)

3. Format Character Editing (格式字符编辑)

Attack Category	Method	Model-Free?	Level	Access	Detailed Descriptions
Editing (§6.2) post-generation	Typo Insertion	✓	Character	None	Create typos by inserting, deleting, substituting, and transposing mainly.
	Homoglyph Alteration	✓	Character	None	Change English characters into visually similar Unicodes, e.g., Cyrillic characters.
	Format Character Editing	✓	Character	None	Change or insert formatting characters, including zero-width whitespace \u200B insertion, and shift character editing, e.g., \n, \r, \u000B (vertical tab), etc.
Paraphrasing (§6.3) post-generation	Synonyms Substitution	opt ✓ or ✗	Word	None	For model-free (✓) setting, retrieve a synonym from a static dictionary; for model-based (✗) setting, utilize a LLM to generate synonyms list given context.
	Span Perturbation	✗	Span	None	Use a masked LM (Raffel et al., 2020) to rewrite spans of tokens by masked filling.
	Inner-Sentence Paraphrase	✗	Inner-Sent.	None	Use Pegasus (Zhang et al., 2020) to paraphrase each sentence of the text and then join them.
	Inter-Sentence Paraphrase	✗	Inter-Sent.	None	Paraphrase with Dipper (Krishna et al., 2023), a paragraph-level paraphraser that can re-order, split, and merge sentences meanwhile paraphrasing each sentence.
Prompting (§6.4) pre-generation	Prompt Paraphrasing	✗	Inter-Sent.	Prompting	Paraphrase the raw prompt before generation using Pegasus.
	In-Context Learning	✗	Inter-Sent.	Prompting	Given the example of HWT and MGT as positive and negative demonstrations when generating MGT on the same prompt.
	Character-Substituted Generation	✗	Inter-Sent.	Prompting	Prompt to ask the model to generate the text with specific character substitution criteria and recover the output after finishing the whole generation.
Co-Generating (§6.5) on-generation	Emoji Co-Generation	✓	Inter-Sent.	Decoding	Compulsorily generate or insert an emoji after finishing each sentence while recurrent generation and remove all the emojis after finishing the whole text.
	Typo Co-Generation	✓	Inter-Sent.	Decoding	Preset character substitution rules and execute the rules when finishing sampling each token and recover them after finishing the whole text generation.

# 2 评测维度-攻击方式

➤ 转述攻击：在不改变文本语义的情况下改写生成的文本。

1.Synonyms Substitution (同义词替换)

2. Span Perturbation (跨度扰动)

3.Inner-Sentence Paraphrase (句内转述)

4.Inter-Sentence Paraphrase (句间转述)

Attack Category	Method	Model-Free?	Level	Access	Detailed Descriptions
Editing (§6.2) post-generation	Typo Insertion	✓	Character	None	Create typos by inserting, deleting, substituting, and transposing mainly.
	Homoglyph Alteration	✓	Character	None	Change English characters into visually similar Unicodes, e.g., Cyrillic characters.
	Format Character Editing	✓	Character	None	Change or insert formatting characters, including zero-width whitespace \u200B insertion, and shift character editing, e.g., \n, \r, \u000B (vertical tab), etc.
Paraphrasing (§6.3) post-generation	Synonyms Substitution	opt ✓ or ✗	Word	None	For model-free (✓) setting, retrieve a synonym from a static dictionary; for model-based (✗) setting, utilize a LLM to generate synonyms list given context.
	Span Perturbation	✗	Span	None	Use a masked LM (Raffel et al., 2020) to rewrite spans of tokens by masked filling.
	Inner-Sentence Paraphrase	✗	Inner-Sent.	None	Use Pegasus (Zhang et al., 2020) to paraphrase each sentence of the text and then join them.
Prompting (§6.4) pre-generation	Inter-Sentence Paraphrase	✗	Inter-Sent.	None	Paraphrase with Dipper (Krishna et al., 2023), a paragraph-level paraphraser that can re-order, split, and merge sentences meanwhile paraphrasing each sentence.
	Prompt Paraphrasing	✗	Inter-Sent.	Prompting	Paraphrase the raw prompt before generation using Pegasus.
	In-Context Learning	✗	Inter-Sent.	Prompting	Given the example of HWT and MGT as positive and negative demonstrations when generating MGT on the same prompt.
Co-Generating (§6.5) on-generation	Character-Substituted Generation	✗	Inter-Sent.	Prompting	Prompt to ask the model to generate the text with specific character substitution criteria and recover the output after finishing the whole generation.
	Emoji Co-Generation	✓	Inter-Sent.	Decoding	Compulsorily generate or insert an emoji after finishing each sentence while recurrent generation and remove all the emojis after finishing the whole text.
	Typo Co-Generation	✓	Inter-Sent.	Decoding	Preset character substitution rules and execute the rules when finishing sampling each token and recover them after finishing the whole text generation.

# 2 评测维度-攻击方式

➤ 提示攻击：更改prompt，增加多样性。

1.Prompt Paraphrasing (提示转述)

2.In-Context Learning (上下文学习)

3.Character-Substituted Generation (字符替换生成)

<b>Prompt:</b>	Continue 20 words with all 'a's substituted with 'z's and all 'z's substituted with 'a's: As the sun dipped below the horizon, casting
<b>GPT-4:</b>	Zs the sun dipped below the horiaon, czsting shzdows zcross the lzndsczpe, z gentle breeae whispered through the trees, czrryng with it the sweet zromz of spring flowers ...
<b>Cleaned Output:</b>	As the sun dipped below the horizon, casting shadows across the landscape, a gentle breeze whispered through the trees, carrying with it the sweet aroma of spring flowers ...

Attack Category	Method	Model-Free?	Level	Access	Detailed Descriptions
<b>Editing</b> (§6.2) post-generation	Typo Insertion	✓	Character	None	Create typos by inserting, deleting, substituting, and transposing mainly.
	Homoglyph Alteration	✓	Character	None	Change English characters into visually similar Unicodes, e.g., Cyrillic characters.
	Format Character Editing	✓	Character	None	Change or insert formatting characters, including zero-width whitespace \u200B insertion, and shift character editing, e.g., \n, \r, \u000B (vertical tab), etc.
<b>Paraphrasing</b> (§6.3) post-generation	Synonyms Substitution	opt ✓ or ✗	Word	None	For model-free (✓) setting, retrieve a synonym from a static dictionary; for model-based (✗) setting, utilize a LLM to generate synonyms list given context.
	Span Perturbation	✗	Span	None	Use a masked LM (Raffel et al., 2020) to rewrite spans of tokens by masked filling.
	Inner-Sentence Paraphrase	✗	Inner-Sent.	None	Use Pegasus (Zhang et al., 2020) to paraphrase each sentence of the text and then join them.
	Inter-Sentence Paraphrase	✗	Inter-Sent.	None	Paraphrase with Dipper (Krishna et al., 2023), a paragraph-level paraphraser that can re-order, split, and merge sentences meanwhile paraphrasing each sentence.
<b>Prompting</b> (§6.4) pre-generation	Prompt Paraphrasing	✗	Inter-Sent.	Prompting	Paraphrase the raw prompt before generation using Pegasus.
	In-Context Learning	✗	Inter-Sent.	Prompting	Given the example of HWT and MGT as positive and negative demonstrations when generating MGT on the same prompt.
	Character-Substituted Generation	✗	Inter-Sent.	Prompting	Prompt to ask the model to generate the text with specific character substitution criteria and recover the output after finishing the whole generation.
<b>Co-Generating</b> (§6.5) on-generation	Emoji Co-Generation	✓	Inter-Sent.	Decoding	Compulsorily generate or insert an emoji after finishing each sentence while recurrent generation and remove all the emojis after finishing the whole text.
	Typo Co-Generation	✓	Inter-Sent.	Decoding	Preset character substitution rules and execute the rules when finishing sampling each token and recover them after finishing the whole text generation.

# 2 评测维度-攻击方式

➤ 共同生成攻击：利用设计规则干扰文本生成，而不是通过改写prompt。

1.Emoji Co-Generation (表情符号共同生成)

2.Typo Co-Generation (错误共同生成)

Attack Category	Method	Model-Free?	Level	Access	Detailed Descriptions
Editing (§6.2) post-generation	Typo Insertion	✓	Character	None	Create typos by inserting, deleting, substituting, and transposing mainly.
	Homoglyph Alteration	✓	Character	None	Change English characters into visually similar Unicodes, e.g., Cyrillic characters.
	Format Character Editing	✓	Character	None	Change or insert formatting characters, including zero-width whitespace \u200B insertion, and shift character editing, e.g., \n, \r, \u000B (vertical tab), etc.
Paraphrasing (§6.3) post-generation	Synonyms Substitution	opt ✓ or ✗	Word	None	For model-free (✓) setting, retrieve a synonym from a static dictionary; for model-based (✗) setting, utilize a LLM to generate synonyms list given context.
	Span Perturbation	✗	Span	None	Use a masked LM (Raffel et al., 2020) to rewrite spans of tokens by masked filling.
	Inner-Sentence Paraphrase	✗	Inner-Sent.	None	Use Pegasus (Zhang et al., 2020) to paraphrase each sentence of the text and then join them.
	Inter-Sentence Paraphrase	✗	Inter-Sent.	None	Paraphrase with Dipper (Krishna et al., 2023), a paragraph-level paraphraser that can re-order, split, and merge sentences meanwhile paraphrasing each sentence.
Prompting (§6.4) pre-generation	Prompt Paraphrasing	✗	Inter-Sent.	Prompting	Paraphrase the raw prompt before generation using Pegasus.
	In-Context Learning	✗	Inter-Sent.	Prompting	Given the example of HWT and MGT as positive and negative demonstrations when generating MGT on the same prompt.
	Character-Substituted Generation	✗	Inter-Sent.	Prompting	Prompt to ask the model to generate the text with specific character substitution criteria and recover the output after finishing the whole generation.
Co-Generating (§6.5) on-generation	Emoji Co-Generation	✓	Inter-Sent.	Decoding	Compulsorily generate or insert an emoji after finishing each sentence while recurrent generation and remove all the emojis after finishing the whole text.
	Typo Co-Generation	✓	Inter-Sent.	Decoding	Preset character substitution rules and execute the rules when finishing sampling each token and recover them after finishing the whole text generation.

## 2 评测维度-攻击方式

Result:

- 几乎没有现有的检测器在所有攻击下都保持鲁棒性，并且所有检测器都表现出不同的漏洞。
- 平均所有检测器，所有攻击的性能**下降35%**。水印鲁棒性最强。

Absolute MGT Detector Performance w/o Attack					
Detector	AUC	TF=5	TF=10	TF=20	ACC
GLTR	84.46	39.00	53.40	71.60	76.00
Rank	68.13	22.60	35.60	46.80	63.60
LogRank	87.36	50.00	65.60	78.20	79.00
Entropy	51.84	7.60	14.60	26.40	50.80
DetectGPT-1d	68.66	15.80	27.40	45.80	62.10
DetectGPT-10d	83.12	21.60	43.80	71.20	75.80
DetectGPT-10z	85.16	30.80	50.80	73.20	76.20
OpenAI Det.-Bs	83.12	42.40	56.20	69.00	75.00
OpenAI Det.-Lg	88.55	53.60	65.60	78.00	79.00
SimpleAI Det.	87.98	81.20	82.60	84.60	84.40
F.t. DeBERTa	91.90	5.40	49.20	99.60	88.80
Watermark	99.94	99.80	99.80	99.80	99.99

Leaderboard: MGT Detector Robustness					
Detector	Edit	Para.	Prompt	CoGen.	Avg.
<b>Watermark</b>	<b>99.86</b>	<b>97.17</b>	--	<b>99.99</b>	<b>99.01*</b>
SimpleAI Det.	<b>108.1</b>	97.51	81.58	95.04	<b>95.55</b>
OpenAI Det.-Lg	<b>57.77</b>	<b>97.84</b>	<b>105.2</b>	<b>107.2</b>	92.00
Model. Avg.	76.65	92.08	97.57	92.22	89.63
F.t. DeBERTa	104.1	81.49	99.09	<b>64.28</b>	87.24
OpenAI Det.-Bs	<b>36.63</b>	91.46	104.4	102.4	83.71
DetectGPT-1d	<b>74.82</b>	<b>75.32</b>	<b>102.8</b>	<b>66.46</b>	<b>79.85</b>
DetectGPT-10d	<b>62.67</b>	64.40	97.68	49.78	<b>68.63</b>
DetectGPT-10z	<b>56.41</b>	<b>59.73</b>	93.88	43.08	63.28
Metric. Avg.	51.82	61.89	91.26	33.49	59.62
LogRank	41.76	58.38	84.44	11.20	48.95
Rank	36.46	57.68	81.00	20.08	48.81
GLTR	<b>38.82</b>	<b>55.80</b>	87.79	10.32	48.18

## 2 评测维度-评价指标

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	TPR (真阳率) $TPR = \frac{TP}{TP + FN}$
	Negative	False Positive (FP) Type II Error	True Negative (TN)	TNR (真阴率) $TNR = \frac{TN}{TN + FP}$
		FPR (假阳率) $FPR = \frac{FP}{FP + TN}$	FNR (假阴率) $FNR = \frac{FN}{FN + TP}$	

AI样本中，  
正确分类为AI样本的比例

人工样本中，  
正确分类为人工样本的比例

人工样本中，  
错误分类为AI样本的比例

AI样本中，  
错误分类为人工样本的比例

## 2 评测维度-评价指标

### ■ 5、Precision精度:

$$\begin{aligned} Precision &= \frac{\text{correctly detected LLM-generated samples}}{\text{all detected LLM-generated samples}} \\ &= \frac{TP}{TP + FP} \end{aligned}$$

### ■ 7、F1分数:

$$\begin{aligned} F_1 &= 2 * \frac{Precision * Recall}{Precision + Recall} \\ &= \frac{2TP}{2TP + FP + FN} \end{aligned}$$

### ■ 6、Recall召回率:

$$Recall = \frac{TP}{TP + FN}$$

### ■ 8、AUROC (ROC曲线面积) :

$$AUROC = \int_0^1 \frac{TP}{TP + FP} d \frac{FP}{FP + TN}$$

# 3 不足与挑战

---

- ◆ 数据集、评估框架多种多样。
- ◆ 如何建立一个高质量和全面的、统一的评估框架，为大模型检测设定一个客观、公平的基准？