Detecting ChatGPT: A Survey of the State of Detecting ChatGPT-Generated Text

Mahdi Dhaini, Wessel Poelman and Ege Erdogan

04.09.2023, RANLP-Stud-2023

Chair of Software Engineering for Business Information Systems (sebis)
Department of Computer Science
School of Computation, Information and Technology (CIT)
Technical University of Munich (TUM)
wwwmatthes.in.tum.de

Introduction

Remarkable abilities of LLMs

Potential societal impacts and risks

Introduction of ChatGPT

Misuse in various domains: education, scientific writing and medical fields.

Introduction

- Remarkable abilities of LLMs
- Potential societal impacts and risks
- Introduction of ChatGPT
- Misuse in various domains: education, scientific writing and medical fields.

Increasing attention for detecting machine-generated text

Introduction

Remarkable abilities of LLMs

Potential societal impacts and risks

Introduction of ChatGPT

Misuse in various domains: education, scientific writing and medical fields.

Increasing attention for detecting machine-generated text

CLIN33 (English & Dutch)
ALTA 2023 (English)
AUTEXTIFICATION (English & Spanish)
RuATD (English & Russian)

Related Work

What *general* approaches exist for machine-generated text detection?

Related Work on Detecting Machine-generated Text

Human judges are decent at spotting machine-generated text from 'older' LLMs such as GPT-2 (Ippolito et al., 2020; Dugan et al., 2020, 2023)

Related Work on Detecting Machine-generated Text

Human judges are decent at spotting machine-generated text from 'older' LLMs such as GPT-2 (Ippolito et al., 2020; Dugan et al., 2020, 2023)

Recent efforts inspired by Computer Vision methods: watermarking or finding model artifacts. (Kirchenbauer et al., 2023; Tay et al., 2020)

Related Work on Detecting Machine-generated Text

Human judges are decent at spotting machine-generated text from 'older' LLMs such as GPT-2 (Ippolito et al., 2020; Dugan et al., 2020, 2023).

Recent efforts inspired by Computer Vision methods: watermarking or finding model artifacts. (Kirchenbauer et al., 2023; Tay et al., 2020)

Access to log-probabilities of LLM essential in applicability of approaches. Works on analyzing probability curvatures or top-k most probable tokens. (Gehrmann et al., 2019; Ippolito et al., 2020; Mitchell et al., 2023)

Related Work on Detecting Machine-generated Text

Human-machine collaboration systems (Jawahar et al., 2020)

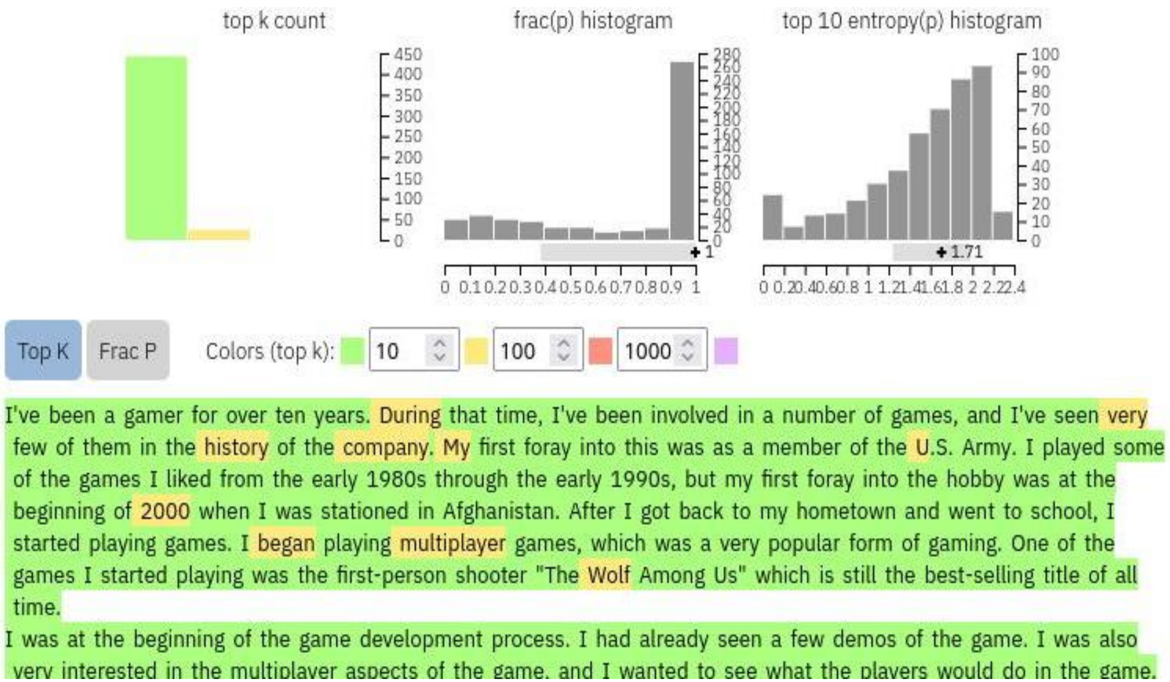


Image from GLTR tool (Gehrmann et al., 2019)

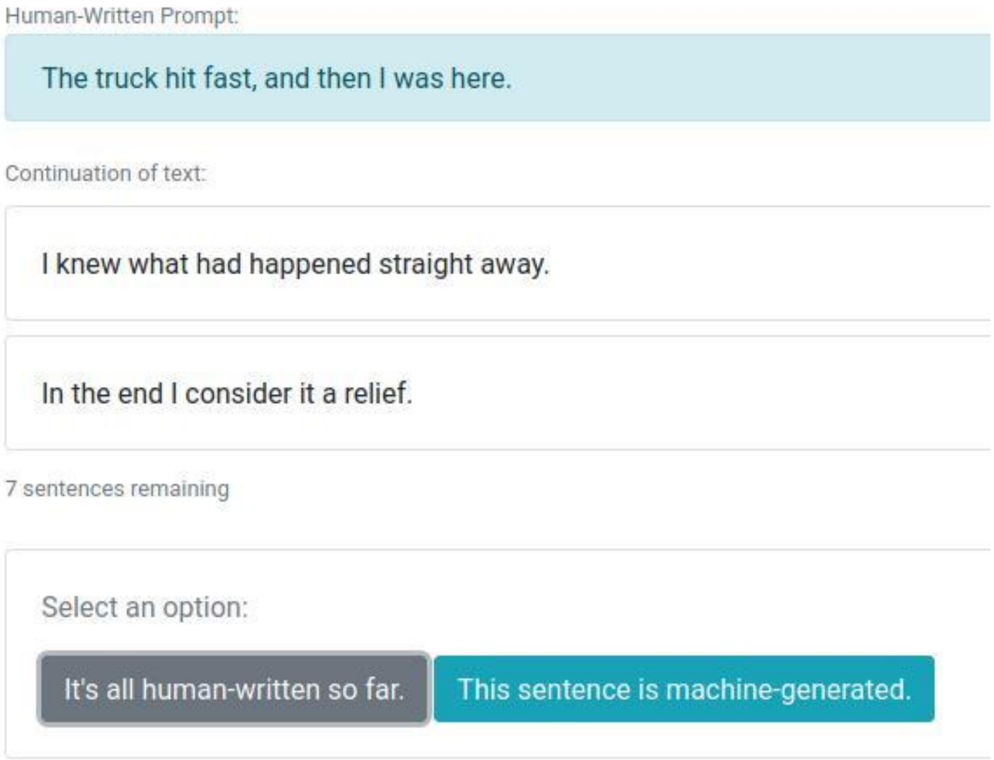
Related Work on Detecting Machine-generated Text

Human-machine collaboration systems (Jawahar et al., 2020)

Image from GLTR tool (Gehrmann et al., 2019)

Gamification of task (Dugan et al., 2020, 2023)

Detect possible transition point from human to machine-generated text to gain insights into characteristics.



Detecting ChatGPT-generated Text

Are these approaches applicable for ChatGPT?

Detecting ChatGPT-generated Text

We don't have access to model probabilities for ChatGPT!

Detecting ChatGPT-generated Text

We don't have access to model probabilities for ChatGPT!

“Black-box scenario”: Classification must be done purely on a piece of text alone.

Detecting ChatGPT-generated Text

We don't have access to model probabilities for ChatGPT!

“Black-box scenario”: Classification must be done purely on a piece of text alone.

How is this done? What datasets are created for this purpose? What insights can we learn from this task?

Detecting ChatGPT-generated Text

We don't have access to model probabilities for ChatGPT!

"Black-box scenario": Classification must be done purely on a piece of text alone.

How is this done? What datasets are created for this purpose? What insights can we learn from this task?

Aim of our contribution

Previous Surveys

Comparison with previous surveys	
Jawahar et al. (2020)	Great overview of general machine-generated text detection methods. (no ChatGPT)
Crothers et al. (2023)	Extensive overview of threat models of generated text, nice overview of comparing generation and detection strategies. (no ChatGPT)
Pegoraro et al. (2023)	Overview of open and closed source detection methods for various models, including ChatGPT.

Previous Surveys

Comparison with previous surveys		
Jawahar et al. (2020)	Great overview of general machine-generated text detection methods. (no ChatGPT)	**We focus on ChatGPT specifically.**
Crothers et al. (2023)	Extensive overview of threat models of generated text, nice overview of comparing generation and detection strategies. (no ChatGPT)	**We focus on datasets, methods and characteristics.**
Pegoraro et al. (2023)	Overview of open and closed source detection methods for various models, including ChatGPT.	**We focus on academic works**

Datasets for Detecting ChatGPT-generated Text

What/how different datasets have been constructed for detecting ChatGPT-generated text?

Datasets

Dataset	Domain	Public	OOD	Type	Human / ChatGPT Samples

Datasets

Dataset	Domain	Public	OOD	Type	Human / ChatGPT Samples
(Guo et al. 2023) HC3-English	Mixed		×	Q&A	58,546 / 26,903

Datasets

Dataset	Domain	Public	OOD	Type	Human / ChatGPT Samples
(Guo et al. 2023) HC3-English	Mixed		×	Q&A	58,546 / 26,903
(Guo et al. 2023) HC3-Chinese	Mixed		×	Q&A	22,259 / 17,522
(Yu et al. 2023) CHEAT	Scientific		✓	Abstracts	15,395 / 35,304
(He et al. 2023) MGTBench	Mixed		×	Q&A	2,817 / 2,817
(Liu et al. 2023) ArguGPT	Education		×	Essays	4,115 / 4,038
(Vasilatos et al. 2023)	Education	*	×	Q&A	960 / 960
(Mitrovic et al. 2023)	Restaurant reviews	*	✓	Reviews	1,000 / 395 + 1,000 rephrase
(Weng et al. 2023)	Scientific		×	Titles/abstracts	59,232 / 59,232
(Antoun et al. 2023)	Mixed		✓	Q&A	58,546 / 26,903 + 5,969 OOD
(Liao et al. 2023)	Medical		×	Abstracts and records	2,200 / 2,200

Datasets

Mixed	Education
HC3 (Guo et al. 2023) • English and Chinese. • Q&A pairs from OpenQA, Reddit ELI5, WikiQA, Medical Dialog, FiQA, and manual crawling of Wikipedia. **MGTBench (He et al. 2023)** • Q&A pairs from TruthfulQA, SQuaD1, NarrativeQA. • Prompting ChatGPT (+ other LLMs) with context. **(Antoun et al. 2023)** • Translates HC3 to French and adds ChatGPT/BingChat Q&A samples with questions from MFAQ, and sentences from the French Treebank dataset. • “Adversarial” examples written by humans to look like ChatGPT.	**ArguGPT (Liu et al. 2023)** • Essays of different English levels from WECCL, TOEFL, GRE with automated scores. • ChatGPT asked to write essay given the question. • Only ChatGPT text freely available. **(Vasilatos et al. 2023)** • Builds on (Ibrahim et al. 2023): metadata and Q&A pairs from university courses with different subjects. • Prompt ChatGPT directly with the question.

Datasets

Restaurant Reviews

(Mitrovic et al. 2023)

- Builds on the Kaggle restaurant reviews dataset.
- ChatGPT prompted to write reviews of different kinds (e.g., a *bad* review).
- Includes ChatGPT rephrasing of human-written reviews.

Medical

(Liao et al. 2023)

- Medical abstracts from Kaggle, radiology reports from MIMIC-III (Johnson et al. 2023).
- ChatGPT asked to continue writing given part of human-written text.

Scientific

(Weng et al. 2023)

- Builds on (Narechania et al. 2023)'s dataset of title/abstract pairs from data visualization papers.
- ChatGPT asked to directly write abstracts given the titles.

CHEAT (Yu et al. 2023)

- Abstracts from computer science papers.
- ChatGPT prompted in different ways:
  - *Generate:* Write abstract given the title and keywords.
  - *Polish:* "Polish" the given human-written abstract.
  - *Mix:* Text from human-written and polished abstracts mixed at the sentence level.

Constructing ChatGPT-generated Datasets

- Directly prompt with questions for Q&A datasets. Provide context to match human-written dataset.
  - Reddit ELI5: *"Explain like I am five, ___"* (Guo et al. 2023)
  - NarrativeQA: *"I will provide a context and a question to you. You need to answer me the question based on the context. The context is: ___. The question is: ___.* (He et al. 2023)

- Prompt in different ways to increase variety of samples. (Mitrovic et al. 2023)
  - *"Write me a two-line review about a restaurant that has* ***some good*** *aspects."*
  - *"Write me a review about a restaurant that has* ***some good and some bad*** *aspects."*

- Ask ChatGPT to rephrase human-written text (Yu et al. 2023; Mitrovic et al. 2023)

- Combine ChatGPT- and human-written text. (Liao et al. 2023)
  - Mix at the sentence-level, or continue human-written text with ChatGPT

- Translate ChatGPT text from another language (Antoun et al. 2023)

Methods for Detecting ChatGPT-generated Text

What methods have been proposed for detecting ChatGPT-generated text?

Methods

Paper	Dataset	Approaches	Explainability	Code
Mitrović et al. 2023	Mitrović et al. 2023	DistilBERT PBC	SHAP	×
Liao et al. 2023	Liao et al. 2023	BERT PBC XGBoost CART	transformer-interpret	×
Liu et al. 2023	ArguGPT	RoBERTa-large SVM	×	✓*
Guo et al. 2023	HC3	GLTR RoBERTa-single RoBERTa-QA	×	✓
Antoun et al. 2023a	Antoun et al. 2023a	CamemBERT CamemBERTa RoBERTa ELECTRA XLM-R	×	✓
Vasilatos et al. 2023	Ibrahim et al. 2023	PBC	×	×

Table 2: Methods proposed in the literature for detecting ChatGPT-generated text. PBC: Perplexity-based classifier. Publicly available models can be accessed by clicking on the ✓character. *Authors indicate it will be made available at a future date.

Methods

Paper	Dataset	Approaches	Explainability	Code
Mitrović et al. 2023	Mitrović et al. 2023	DistilBERT PBC	SHAP	×
Liao et al. 2023	Liao et al. 2023	BERT PBC XGBoost CART	transformer-interpret	×
Liu et al. 2023	ArguGPT	RoBERTa-large SVM	×	✓*
Guo et al. 2023	HC3	GLTR RoBERTa-single RoBERTa-QA	×	✓
Antoun et al. 2023a	Antoun et al. 2023a	CamemBERT CamemBERTa RoBERTa ELECTRA XLM-R	×	✓
Vasilatos et al. 2023	Ibrahim et al. 2023	PBC	×	×

Table 2: Methods proposed in the literature for detecting ChatGPT-generated text. PBC: Perplexity-based classifier. Publicly available models can be accessed by clicking on the ✓character. *Authors indicate it will be made available at a future date.

Methods

Paper	Dataset	Approaches	Explainability	Code
Mitrović et al. 2023	Mitrović et al. 2023	DistilBERT PBC	SHAP	×
Liao et al. 2023	Liao et al. 2023	BERT PBC XGBoost CART	transformer-interpret	×
Liu et al. 2023	ArguGPT	RoBERTa-large SVM	×	✓*
Guo et al. 2023	HC3	GLTR RoBERTa-single RoBERTa-QA	×	✓
Antoun et al. 2023a	Antoun et al. 2023a	CamemBERT CamemBERTa RoBERTa ELECTRA XLM-R	×	✓
Vasilatos et al. 2023	Ibrahim et al. 2023	PBC	×	×

Table 2: Methods proposed in the literature for detecting ChatGPT-generated text. PBC: Perplexity-based classifier. Publicly available models can be accessed by clicking on the ✓character. *Authors indicate it will be made available at a future date.

Methods

Paper	Dataset	Approaches	Explainability	Code
Mitrović et al. 2023	Mitrović et al. 2023	DistilBERT PBC	SHAP	×
Liao et al. 2023	Liao et al. 2023	BERT PBC XGBoost CART	transformer-interpret	×
Liu et al. 2023	ArguGPT	RoBERTa-large SVM	×	✓*
Guo et al. 2023	HC3	GLTR RoBERTa-single RoBERTa-QA	×	✓
Antoun et al. 2023a	Antoun et al. 2023a	CamemBERT CamemBERTa RoBERTa ELECTRA XLM-R	×	✓
Vasilatos et al. 2023	Ibrahim et al. 2023	PBC	×	×

Table 2: Methods proposed in the literature for detecting ChatGPT-generated text. PBC: Perplexity-based classifier. Publicly available models can be accessed by clicking on the ✓character. *Authors indicate it will be made available at a future date.

Methods

Paper	Dataset	Approaches	Explainability	Code
Mitrović et al. 2023	Mitrović et al. 2023	DistilBERT PBC	SHAP	×
Liao et al. 2023	Liao et al. 2023	BERT PBC XGBoost CART	transformer-interpret	×
Liu et al. 2023	ArguGPT	RoBERTa-large SVM	×	✓*
Guo et al. 2023	HC3	GLTR RoBERTa-single RoBERTa-QA	×	✓
Antoun et al. 2023a	Antoun et al. 2023a	CamemBERT CamemBERTa RoBERTa ELECTRA XLM-R	×	✓
Vasilatos et al. 2023	Ibrahim et al. 2023	PBC	×	×

Table 2: Methods proposed in the literature for detecting ChatGPT-generated text. PBC: Perplexity-based classifier. Publicly available models can be accessed by clicking on the ✓character. *Authors indicate it will be made available at a future date.

Methods

Paper	Dataset	Approaches	Explainability	Code
Mitrović et al. 2023	Mitrović et al. 2023	DistilBERT PBC	SHAP	×
Liao et al. 2023	Liao et al. 2023	BERT PBC XGBoost CART	transformer-interpret	×
Liu et al. 2023	ArguGPT	RoBERTa-large SVM	×	✓*
Guo et al. 2023	HC3	GLTR RoBERTa-single RoBERTa-QA	×	✓
Antoun et al. 2023a	Antoun et al. 2023a	CamemBERT CamemBERTa RoBERTa ELECTRA XLM-R	×	✓
Vasilatos et al. 2023	Ibrahim et al. 2023	PBC	×	×

Table 2: Methods proposed in the literature for detecting ChatGPT-generated text. PBC: Perplexity-based classifier. Publicly available models can be accessed by clicking on the ✓character. *Authors indicate it will be made available at a future date.

Methods

Paper	Dataset	Approaches	Explainability	Code
Mitrović et al. 2023	Mitrović et al. 2023	DistilBERT PBC	SHAP	×
Liao et al. 2023	Liao et al. 2023	BERT PBC XGBoost CART	transformer-interpret	×
Liu et al. 2023	ArguGPT	RoBERTa-large SVM	×	✓*
Guo et al. 2023	HC3	GLTR RoBERTa-single RoBERTa-QA	×	✓
Antoun et al. 2023a	Antoun et al. 2023a	CamemBERT CamemBERTa RoBERTa ELECTRA XLM-R	×	✓
Vasilatos et al. 2023	Ibrahim et al. 2023	PBC	×	×

Table 2: Methods proposed in the literature for detecting ChatGPT-generated text. PBC: Perplexity-based classifier. Publicly available models can be accessed by clicking on the ✓character. *Authors indicate it will be made available at a future date.

Analysis of Human and ChatGPT-Generated Text

What are the takeaways from the analyses of the textual characteristics of Human and ChatGPT-generated text for different domains and datasets?

Analysis of Human and ChatGPT-Generated Text

Domain	ChatGPT vs Human-written text
Medical	▪ Lower text perplexity ▪ More fluent, neutral, positive. ▪ More general in content and language style
English argumentative essay	▪ Syntactically more complex sentences than English language learners ▪ Lower lexical diversity
Multi-domain QA	▪ Organized and neutral way, offers less bias and harmful information ▪ Formal, less emotional, and more objective
Scientific abstracts	▪ Better choice of vocabulary ▪ More unique words, ▪ More connecting words, ▪ Fewer grammatical errors
Language-agnostic characteristics*	▪ Similar characteristics for ChatGPT-generated text in different languages (English, French, Chinese)

Analysis of Human and ChatGPT-Generated Text

Domain	ChatGPT vs Human-written text
Medical	▪ Lower text perplexity ▪ More fluent, neutral, positive. ▪ More general in content and language style
English argumentative essay	▪ Syntactically more complex sentences than English language learners ▪ Lower lexical diversity
Multi-domain QA	▪ Organized and neutral way, offers less bias and harmful information ▪ Formal, less emotional, and more objective
Scientific abstracts	▪ Better choice of vocabulary ▪ More unique words, ▪ More connecting words, ▪ Fewer grammatical errors
Language-agnostic characteristics*	▪ Similar characteristics for ChatGPT-generated text in different languages (English, French, Chinese)

Analysis of Human and ChatGPT-Generated Text

Domain	ChatGPT vs Human-written text
Medical	▪ Lower text perplexity ▪ More fluent, neutral, positive. ▪ More general in content and language style
English argumentative essay	▪ Syntactically more complex sentences than English language learners ▪ Lower lexical diversity
Multi-domain QA	▪ Organized and neutral way, offers less bias and harmful information ▪ Formal, less emotional, and more objective
Scientific abstracts	▪ Better choice of vocabulary ▪ More unique words, ▪ More connecting words, ▪ Fewer grammatical errors
Language-agnostic characteristics*	▪ Similar characteristics for ChatGPT-generated text in different languages (English, French, Chinese)

Analysis of Human and ChatGPT-Generated Text

Domain	ChatGPT vs Human-written text
Medical	▪ Lower text perplexity ▪ More fluent, neutral, positive. ▪ More general in content and language style
English argumentative essay	▪ Syntactically more complex sentences than English language learners ▪ Lower lexical diversity
Multi-domain QA	▪ Organized and neutral way, offers less bias and harmful information ▪ Formal, less emotional, and more objective
Scientific abstracts	▪ Better choice of vocabulary ▪ More unique words, ▪ More connecting words, ▪ Fewer grammatical errors
Language-agnostic characteristics*	▪ Similar characteristics for ChatGPT-generated text in different languages (English, French, Chinese)

230904 Dhaini et al. Detecting ChatGPT: A Survey of the State of Detecting ChatGPT-Generated Text

Analysis of Human and ChatGPT-Generated Text

Domain	ChatGPT vs Human-written text
Medical	▪ Lower text perplexity ▪ More fluent, neutral, positive. ▪ More general in content and language style
English argumentative essay	▪ Syntactically more complex sentences than English language learners ▪ Lower lexical diversity
Multi-domain QA	▪ Organized and neutral way, offers less bias and harmful information ▪ Formal, less emotional, and more objective
Scientific abstracts	▪ Better choice of vocabulary ▪ More unique words, ▪ More connecting words, ▪ Fewer grammatical errors
Language-agnostic characteristics*	▪ Similar characteristics for ChatGPT-generated text in different languages (English, French, Chinese)

General Insights

What general insights do we have on the state of detecting ChatGPT-generated text?

General Insights

Role of Explainable AI

- Understanding writing styles
- Debugging

Humans versus ChatGPT in the detection task

Robustness of detectors

- Perturbed data
- Out-of-domain

Impact of text length on detection

- Full text training vs short text evaluation

Lack of special prompts in ChatGPT-generated text

- General style and state
- Future work investigation

Perplexity-based detectors

- Open-source LLMs for calculating perplexity scores

Cost of constructing machine-generated datasets

- Need for large-scale ChatGPT-generated datasets

Multilinguality

- English dominance
- Detecting translated text.

General Insights

Role of Explainable AI

- Understanding writing styles
- Debugging

Humans versus ChatGPT in the detection task

Robustness of detectors

- Perturbed data
- Out-of-domain

Impact of text length on detection

- Full text training vs short text evaluation

Lack of special prompts in ChatGPT-generated text

- General style and state
- Future work investigation

Perplexity-based detectors

- Open-source LLMs for calculating perplexity scores

Cost of constructing machine-generated datasets

- Need for large-scale ChatGPT-generated datasets

Multilinguality

- English dominance
- Detecting translated text.

General Insights

Role of Explainable AI

- Understanding writing styles
- Debugging

Impact of text length on detection

- Full text training vs short text evaluation

Humans versus ChatGPT in the detection task

Lack of special prompts in ChatGPT-generated text

- General style and state
- Future work investigation

Robustness of detectors

- Perturbed data
- Out-of-domain

Perplexity-based detectors

- Open-source LLMs for calculating perplexity scores

Cost of constructing machine-generated datasets

- Need for large-scale ChatGPT-generated datasets

Multilinguality

- English dominance.
- Performance
- Less reliability in detecting translated text.

General Insights

Role of Explainable AI

- Understanding writing styles
- Debugging

Humans versus ChatGPT in the detection task

Robustness of detectors

- Perturbed data
- Out-of-domain

Impact of text length on detection

- Full text training vs short text evaluation

Lack of special prompts in ChatGPT-generated text

- General style and state
- Future work investigation

Perplexity-based detectors

- Open-source LLMs for calculating perplexity scores

Cost of constructing machine-generated datasets

- Need for large-scale ChatGPT-generated datasets

Multilinguality

- English dominance.
- Performance
- Less reliability in detecting translated text.

Future Work

Observations	Challenges	Next steps
Big variety in methods, datasets and insights.	No consistency in experimental setups (prompting, adversarial samples, out-of-domain tests, etc.)	**Test methods across datasets and datasets across methods.**

Future Work

Observations	Challenges	Next steps
Big variety in methods, datasets and insights.	No consistency in experimental setups (prompting, adversarial samples, out-of-domain tests, etc.)	**Test methods across datasets and datasets across methods.**
Critical domains such as health and education are covered.	No datasets found for news domain.	**Add additional critical domains.**

Future Work

Observations	Challenges	Next steps
Big variety in methods, datasets and insights.	No consistency in experimental setups (prompting, adversarial samples, out-of-domain tests, etc.)	**Test methods across datasets and datasets across methods.**
Critical domains such as health and education are covered.	No datasets found for news domain.	**Add additional critical domains.**
Data for English, French and Chinese.	English is by far the dominant language.	**Look into effect of language and create datasets for more diverse languages.**

Future Work

Observations	Challenges	Next steps
Big variety in methods, datasets and insights.	No consistency in experimental setups (prompting, adversarial samples, out-of-domain tests, etc.)	**Test methods across datasets and datasets across methods.**
Critical domains such as health and education are covered.	No datasets found for news domain.	**Add additional critical domains.**
Data for English, French and Chinese.	English is by far the dominant language.	**Look into effect of language and create datasets for more diverse languages.**
Most data is openly available.	Lack of reporting on when data was collected.	**Repeated testing of methods across time, ChatGPT is closed source; can change at any moment.**

Limitations

Rapid pace of work in this area

ChatGPT being a closed-source system

Reproducibility of results

Mahdi Dhaini, MSc
Wessel Poelman, MSc
Ege Erdogan, BSc

Technical University of Munich (TUM)
TUM School of CIT
Department of Computer Science (CS)
Chair of Software Engineering for Business
Information Systems (sebis)

Boltzmannstraße 3
85748 Garching bei München

+49.89.289.17132
matthes@in.tum.de
wwwmatthes.in.tum.de