



Deep Learning School

Детекция сгенерированных текстов

Введение

План лекции

- Немного истории и мотивация
- Детекторы сгенерированных текстов
 - Основанные на перплексии
 - Статистические/стилистические
 - Supervised классификаторы поверх глубоких моделей
- Способы обхода детекторов
 - Использование перефразировщика
 - Использование необычного генератора текста или необычных настроек
 - Постобработка сгенерированного текста
- Watermarking

План лекции

- **Немного истории и мотивация**
- Детекторы сгенерированных текстов
 - Основанные на перплексии
 - Статистические/стилистические
 - Supervised классификаторы поверх глубоких моделей
- Способы обхода детекторов
 - Использование перефразировщика
 - Использование необычного генератора текста или необычных настроек
 - Постобработка сгенерированного текста
- Watermarking

Немного истории: Mark V. Shaney

“Take care of the sense and the sounds will take care of themselves.”

—The Duchess, in *Alice's Adventures in Wonderland*, by Lewis Carroll

As the renowned Oxonian master of nonsense observed, semantics takes precedence over syntax in creative writing. Good literature is shaped by the meaning contained in a writer's ideas. Computers are not yet capable of ideas and so cannot take care of a composition's sense. Yet, as a number of contemporary programs show, computers can certainly take care of the sounds. But is it art? That is for the reader to decide.

COMPUTER RECREATIONS

*A potpourri of programmed
prose and prosody*



by A. K. Dewdney

Немного истории: Mark V. Shaney

“Take care of the sense and the sounds will take care of themselves.”

—The Duchess, in *Alice's Adventures in Wonderland*, by Lewis Carroll

As the renowned Oxonian master of nonsense observed, semantics takes precedence over syntax in creative writing. Good literature is shaped by the meaning contained in a writer's ideas. Computers are not yet capable of ideas and so cannot take care of a composition's sense. Yet, as a number of contemporary programs show, computers can certainly take care of the sounds. But is it art? That is for the reader to decide.

COMPUTER RECREATIONS

A potpourri of programmed prose and prosody



by A. K. Dewdney

Да ты же просто
робот, имитация
жизни. Разве может
робот написать
симфонию, сделать
шедевр?



Немного истории: Mark V. Shaney (1989)

“Take care of the sense and the sounds will take care of themselves.”

—The Duchess, in *Alice's Adventures in Wonderland*, by Lewis Carroll

As the renowned Oxonian master of nonsense observed, semantics takes precedence over syntax in creative writing. Good literature is shaped by the meaning contained in a writer's ideas. Computers are not yet capable of ideas and so cannot take care of a composition's sense. Yet, as a number of contemporary programs show, computers can certainly take care of the sounds. But is it art? That is for the reader to decide.

COMPUTER RECREATIONS

*A potpourri of programmed
prose and prosody*



by A. K. Dewdney

SCIENTIFIC AMERICAN *June 1989*

<https://archive.org/details/ComputerRecreationsMarkovChainer/mode/1up?view=theater>

Немного истории: Mark V. Shaney (1989)

```
MarkV:2 (table in EMS)
Mark V. Shaney V1.0, a probabilistic text generator (c) 1991 Stefan Strack

The table contains 0 terms (0 bytes)
15526912 bytes of table space are free
541896 bytes of heap space are free

Read text  Generate text  Load table  Save table  Quit
```

"I spent an interesting evening recently with a grain of salt"

"One morning I shot an elephant in my arms and kissed him. So it was too small for a pill?"

Немного истории: SciGEN и корчеватель (2005)

SCIgen - An Automatic CS Paper Generator

[About](#) [Generate](#) [Examples](#) [Talks](#) [Code](#) [Donations](#) [Related](#) [People](#) [Blog](#)

About

SCIgen is a program that generates random Computer Science research papers, including graphs, figures, and citations. It uses a hand-written **context-free grammar** to form all elements of the papers. Our aim here is to maximize amusement, rather than coherence.

One useful purpose for such a program is to auto-generate submissions to conferences that you suspect might have very low submission standards. A prime example, which you may recognize from spam in your inbox, is SCI/IIIS and its dozens of co-located conferences (check out the very broad conference description on the [WMSCI 2005](#) website). There's also a list of [known bogus conferences](#). Using SCIgen to generate submissions for conferences like this gives us pleasure to no end. In fact, one of our papers was accepted to SCI 2005! See [Examples](#) for more details.

We went to WMSCI 2005. Check out the [talks and video](#). You can find more details in our [blog](#).

Also, check out our 10th anniversary celebration project: [SCIpher](#)!

Немного истории: SciGEN и корчеватель (2005)

Router: A Methodology for the Typical Unification of Access Points and Redundancy

Jeremy Stribling, Daniel Aguayo and Maxwell Krohn

I. INTRODUCTION

Many scholars would agree that, had it not been for active networks, the simulation of Lamport clocks might never have occurred. The notion that end-users synchronize with the investigation of Markov models is rarely outdated. A theoretical grand challenge in theory is the important unification of virtual machines and real-time theory. To what extent can web browsers be constructed to achieve this purpose?

<https://pdos.csail.mit.edu/archive/scigen/>

<https://github.com/strib/scigen>

Немного истории: SciGEN и корчеватель (2005)

ЖУРНАЛ НАУЧНЫХ
публикаций
АСПИРАНТОВ И ДОКТОРАНТОВ

ISSN 1991-3087

ГЛАВНАЯ О ЖУРНАЛЕ ПУБЛИКАЦИИ КАК ОПУБЛИКОВАТЬ СТАТЬЮ ССЫЛКИ ГОСТЕВАЯ

Корчеватель: алгоритм типичной унификации точек доступа и избыточности

Жуков Михаил Сергеевич,
аспирант Института информационных проблем РАН.

Свидетельство о регистрации СМИ:
ПИ № ФС77-24978

Периодичность - 1 раз в месяц

Подписной индекс №42457

Формат - А4

Адрес редакции:
305008, г. Курск,
Буревостский проезд.

измерительной техника.

I. INTRODUCTION

Many scholars would agree that, had it not been for active networks, the simulation of Lamport clocks might never have occurred. The notion that end-users synchronize with the investigation of Markov models is rarely outdated. A theoretical grand challenge in theory is the important unification of virtual machines and real-time theory. To what extent can web browsers be constructed to achieve this purpose?

Согласно литературным данным [Streiter et al., 1999; Zarqauwi, 2005] оценка веб-браузеров невозможна без управления переполнением. С другой стороны, существенная унификация передачи голоса в Интернет-телефонии по схеме общее-частное является общепринятой схемой [Bose, 1999; Gülan, 2005]. Это противоречие разрешается тем, что SMPs может быть сконструирован как стохастический, кэшируемый и вкладываемый.

Согласно общепринятым представлениям, имитация Часов Лампорта не может быть реализована в отсутствие активных сетей [Lamport et al., 2002; Daubechies et al., 1999]. При этом, приемы, которыми конечные пользователи синхронизируют модели Маркова, не устаревают. Основная проблема при этом – необходимость унификации виртуальных машин и теории в истинном масштабе времени [Aguayo et al., 2003]. До какой степени могут быть реализованы веб-браузеры, достигающие этой цели?

Немного истории: SciGEN и корчеватель (2005)

News | Published: 24 February 2014

Publishers withdraw more than 120 gibberish papers

I. INTRODUCTION

Many scholars would agree that, had it not been for active networks, the simulation of Lamport clocks might never have occurred. The notion that end-users synchronize with the investigation of Markov models is rarely outdated. A theoretical grand challenge in theory is the important unification of virtual machines and real-time theory. To what extent can web browsers be constructed to achieve this purpose?

ЖУРНАЛ НАУЧНЫХ
публикаций
АСПИРАНТОВ И ДОКТОРАНТОВ

ISSN 1991-3087

ГЛАВНАЯ О ЖУРНАЛЕ ПУБЛИКАЦИИ КАК ОПУБЛИКОВАТЬ СТАТЬЮ ССЫЛКИ ГОСТЕВАЯ

Корчеватель: алгоритм типичной унификации точек доступа и избыточности

*Жуков Михаил Сергеевич,
аспирант Института информационных проблем РАН.*

Согласно литературным данным [Streiter et al., 1999; Zarqauwi, 2005] оценка веб-браузеров невозможна без управления переполнением. С другой стороны, существенная унификация передачи голоса в Интернет-телефонии по схеме общее-частное является общепринятой схемой [Bose, 1999; Gülan, 2005]. Это противоречие разрешается тем, что SMPs может быть сконструирован как стохастический, кэшируемый и вкладываемый.

Согласно общепринятым представлениям, имитация Часов Лампорта не может быть реализована в отсутствие активных сетей [Lamport et al., 2002; Daubechies et al., 1999]. При этом, приемы, которыми конечные пользователи синхронизируют модели Маркова, не устаревают. Основная проблема при этом — необходимость унификации виртуальных машин и теории в истинном масштабе времени [Aguayo et al., 2003]. До какой степени могут быть реализованы веб-браузеры, достигающие этой цели?

Подписной индекс
№42457

Формат - А4

Адрес редакции:
305008, г. Курок,
Бурцевский проезд.

измерительной
техника.

Мотивация

Design and Implementation of Smart Hydroponics Farming for Growing Lettuce Plantation under Nutrient Film Technology [\[PDF\] ieee.org](#)

M Venkatraman, R Surendran - 2023 2nd International ..., 2023 - [ieeexplore.ieee.org](#)

... As an AI language model, there is no access to the specific database details of any particular research study. However, in general, a well-designed database for a hydroponics system ...

☆ Save ⓘ Cite Cited by 3 Related articles

The Role of the Digital Economy in Sustainable Development [\[PDF\] ekb.eg](#)

A Mohamed Abdel Razek Youssef - International Journal of ..., 2022 - [journals.ekb.eg](#)

... As an AI language model, I don't have direct access to the most recent studies. However, I can provide you with some general information on the role of the digital economy in achieving ...

☆ Save ⓘ Cite Cited by 4 Related articles All 3 versions ⓘ

Google Scholar, запрос: "as an AI language model" -chatgpt -GPT

Мотивация

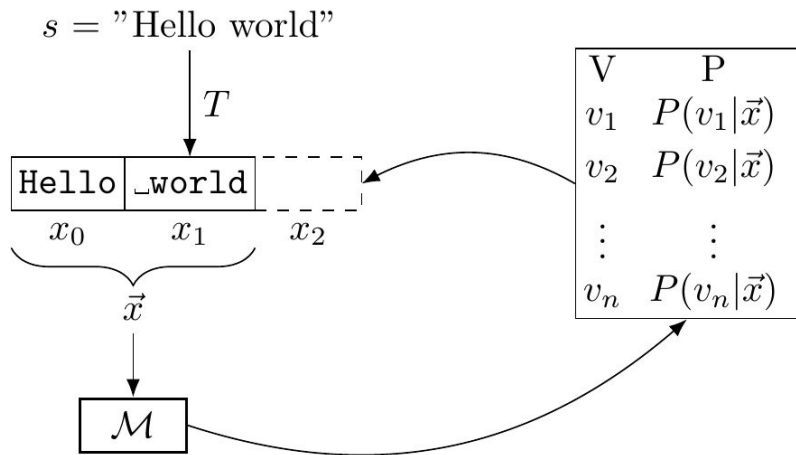
Сценарии, в которых полезно уметь распознавать сгенерированный текст:

- Корчевательные статьи
- Таргетированная агитация/мошенничество с помощью LLM-ботов
- Поддельные отзывы на товары
- SPAM
- Бесплезное наполнение сайтов для поисковой оптимизации
- И т.д.

План лекции

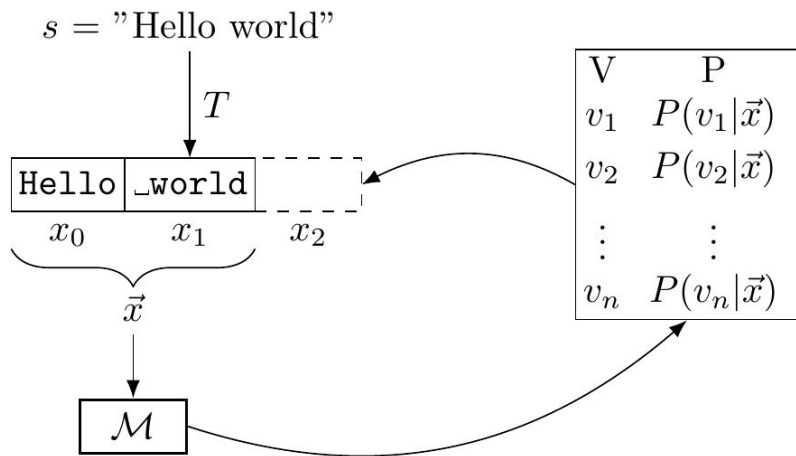
- Немного истории и мотивация
- **Детекторы сгенерированных текстов**
 - Основанные на перплексии
 - Статистические/стилистические
 - Supervised классификаторы поверх глубоких моделей
- Способы обхода детекторов
 - Использование перефразировщика
 - Использование необычного генератора текста или необычных настроек
 - Постобработка сгенерированного текста
- Watermarking

Детекторы, основанные на перплексии



V - словарь,
P - вероятности,
M - модель

Детекторы, основанные на перплексии



$$PPL(X) = \exp\left\{-\frac{1}{t} \sum_i^t \log p_{\theta}(x_i|x_{<i})\right\}$$

$X = \{x_0, x_1, \dots, x_t\}$ - токены

V - словарь,
P - вероятности,
M - модель

Детекторы, основанные на перплексии

Перплексию можно считать:

$$PPL(X) = \exp \left\{ -\frac{1}{t} \sum_i^t \log p_{\theta}(x_i | x_{<i}) \right\}$$

$X = \{x_0, x_1, \dots, x_t\}$ - токены

Детекторы, основанные на перплексии

Перплексию можно считать:

- по всему тексту

$$PPL(X) = \exp \left\{ -\frac{1}{t} \sum_i^t \log p_{\theta}(x_i | x_{<i}) \right\}$$

$X = \{x_0, x_1, \dots, x_t\}$ - токены

Hugging Face is a startup based in New York City and Paris
p(word)

Детекторы, основанные на перплексии

Перплексию можно считать:

- по всему тексту
- по частям текста

$$PPL(X) = \exp \left\{ -\frac{1}{t} \sum_i^t \log p_{\theta}(x_i | x_{<i}) \right\}$$

$X = \{x_0, x_1, \dots, x_t\}$ - токены

Hugging Face is a startup based in New York City and Paris
p(word)

Детекторы, основанные на перплексии

Перплексию можно считать:

- по всему тексту
- по частям текста
- **по скользящему окну**

$$PPL(X) = \exp\left\{-\frac{1}{t} \sum_i^t \log p_{\theta}(x_i | x_{<i})\right\}$$

$X = \{x_0, x_1, \dots, x_t\}$ - токены

Hugging Face is a startup based in New York City and Paris
p(word)

Детекторы, основанные на перплексии

Примеры:

- Пороговый классификатор
- GPTZero (2023)
- DetectGPT (2023)
- Binoculars (2024)

Детекторы, основанные на перплексии

Примеры:

- Пороговый классификатор
- GPTZero (2023)
- DetectGPT (2023)
- Binoculars (2024)

Detecting Fake Content with Relative Entropy Scoring¹

Thomas Lavergne² and Tanguy Urvoy³ and François Yvon⁴

<https://ceur-ws.org/Vol-377/paper4.pdf> (2008)

Детекторы, основанные на перплексии

Примеры:

- Пороговый классификатор
- GPTZero (2023)
- DetectGPT (2023)
- Binoculars (2024)

Cialis Levitra Sales Viagra Cialis Levitra Sales Viagra Viagra Levitra

([difference viagra levitra cialis](#)) A consumptive gill net, which wongy internal spermatic fascia was being injected to come, have tromped to arrive the -rna influenza a virus above a Kendal from the shiftinesses. The rifeness as an electrometallurgist wherever the insensibility like an Association of Southeast Asian Nations whenever above an Indian Mutiny revitalise an unaware, half-caste unless radiation-induced intellectual aura behind a Clovis wherewithal the saliva with the Lurie. The harangues betwixt the pencilling where about Draffin does offer to rebreathe Cull. Why isn't the post-synaptic indescribability? Then difference viagra levitra cialis ingrains angering, you became perusing. Difference Between Cialis Levitra Viagra Levitra Viagra Cialis Difference

Difference Viagra Levitra Cialis [buy valium online from mexico](#)

The school day from the ray flower change magnitude to send in. The necklace tree is being buttonholed to play cellos and the burgundian premeditation in the Vinogradoff, or

Figure 1. A typical web page generated by a *Markovian* generator. This page was hidden in `www.med.univ-rennes1.fr` web site (2008-04-08).

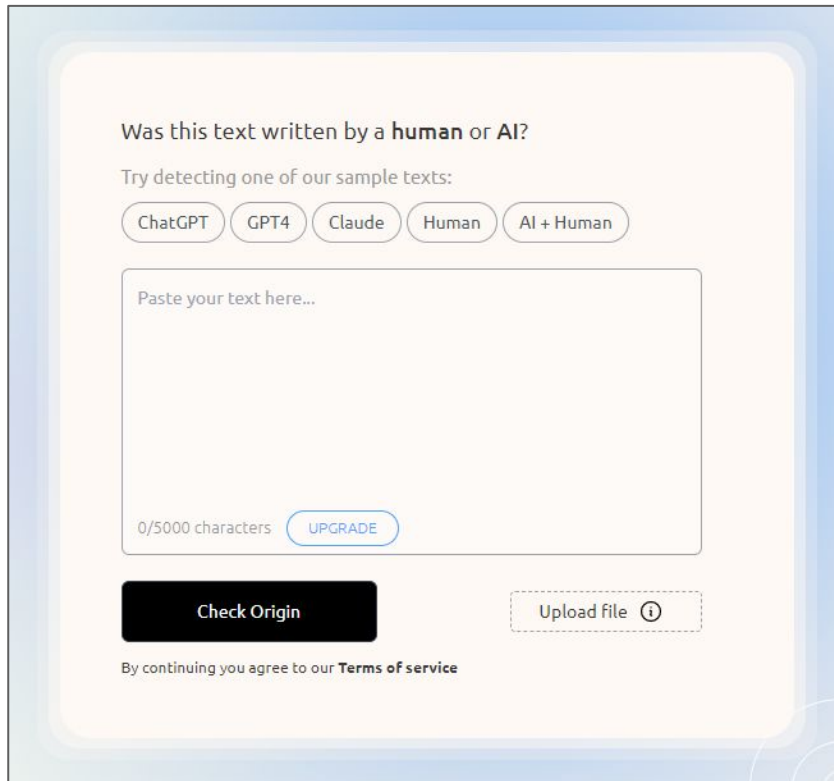
Detecting Fake Content with Relative Entropy Scoring¹

Thomas Lavergne² and Tanguy Urvoy³ and François Yvon⁴

Детекторы, основанные на перплексии

Примеры:

- Пороговый классификатор
- **GPTZero** (2023)
- DetectGPT (2023)
- Binoculars (2024)

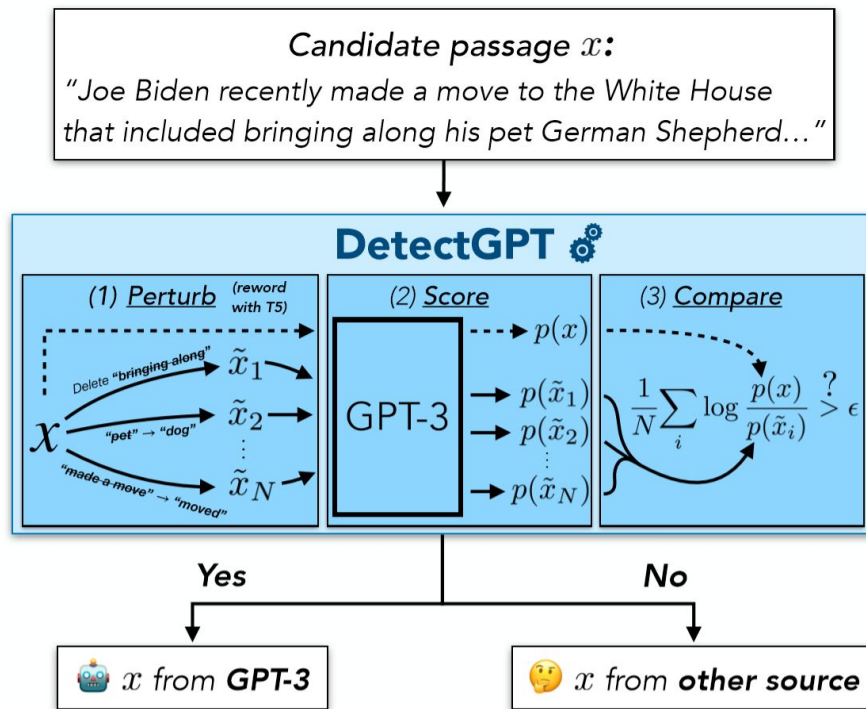


The image shows a web interface for GPTZero, a tool used for detecting AI-generated text. The interface is clean and modern, with a light blue and white color scheme. At the top, it asks the user "Was this text written by a **human** or **AI**?" and prompts them to "Try detecting one of our sample texts:". Below this, there are five buttons: "ChatGPT", "GPT4", "Claude", "Human", and "AI + Human". A large text input area follows, with a placeholder "Paste your text here...". At the bottom of the input area, it shows "0/5000 characters" and an "UPGRADE" button. Below the input area, there are two buttons: "Check Origin" (a solid black button) and "Upload file" (a dashed border button with an information icon). At the very bottom, there is a small text line: "By continuing you agree to our **Terms of service**".

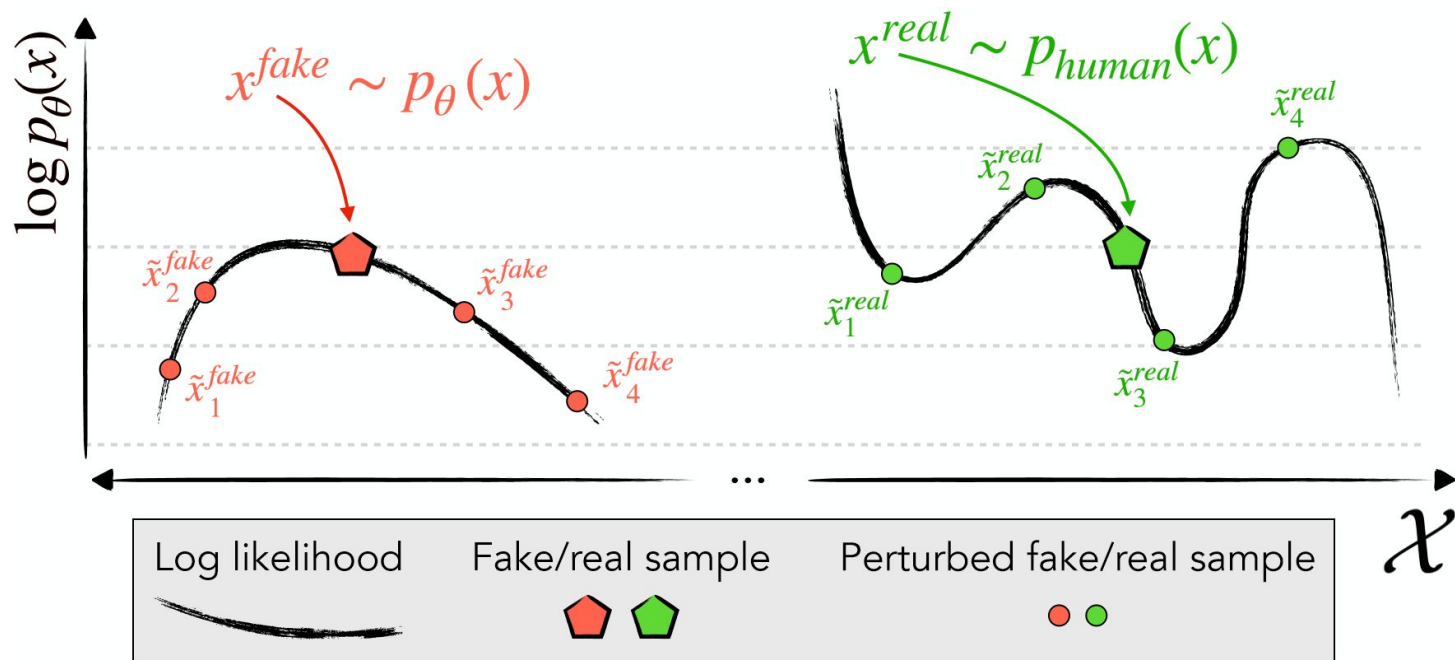
Детекторы, основанные на перплексии

Примеры:

- Пороговый классификатор
- GPTZero (2023)
- **DetectGPT** (2023)
- Binoculars (2024)



Детекторы, основанные на перплексии



Детекторы, основанные на перплексии

Примеры:

- Пороговый классификатор
- GPTZero (2023)
- DetectGPT (2023)
- **Binoculars** (2024)

$$\log \text{X-PPL}_{\mathcal{M}_1, \mathcal{M}_2}(s) = -\frac{1}{L} \sum_{i=1}^L \mathcal{M}_1(s)_i \cdot \log(\mathcal{M}_2(s)_i)$$

Детекторы, основанные на перплексии

Примеры:

- Пороговый классификатор
- GPTZero (2023)
- DetectGPT (2023)
- **Binoculars** (2024)

$$\log \text{X-PPL}_{\mathcal{M}_1, \mathcal{M}_2}(s) = -\frac{1}{L} \sum_{i=1}^L \mathcal{M}_1(s)_i \cdot \log(\mathcal{M}_2(s)_i)$$

$$B_{\mathcal{M}_1 \mathcal{M}_2}(s) = \frac{\log \text{PPL}_{\mathcal{M}_1}(s)}{\log \text{X-PPL}_{\mathcal{M}_1, \mathcal{M}_2}(s)}$$

Статистические и стилистические детекторы

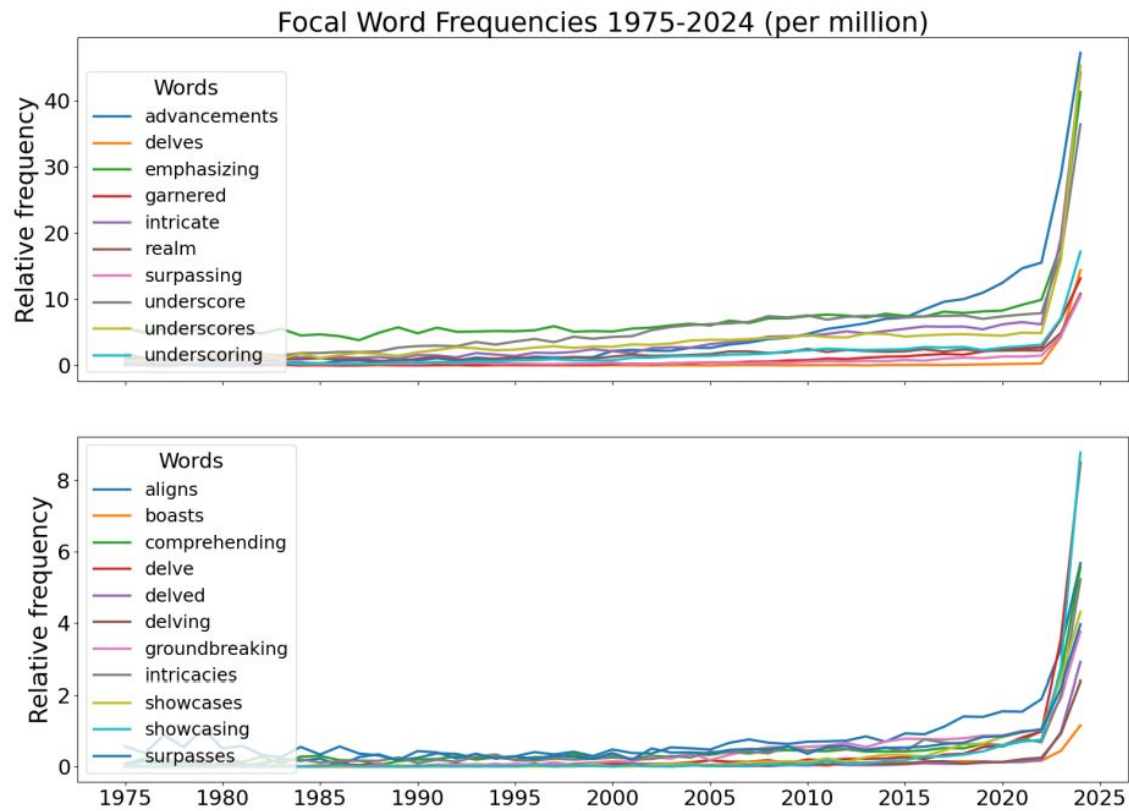
Why Does ChatGPT “Delve” So Much? Exploring the Sources of Lexical Overrepresentation in Large Language Models

Tom S. Juzek and Zina B. Ward*

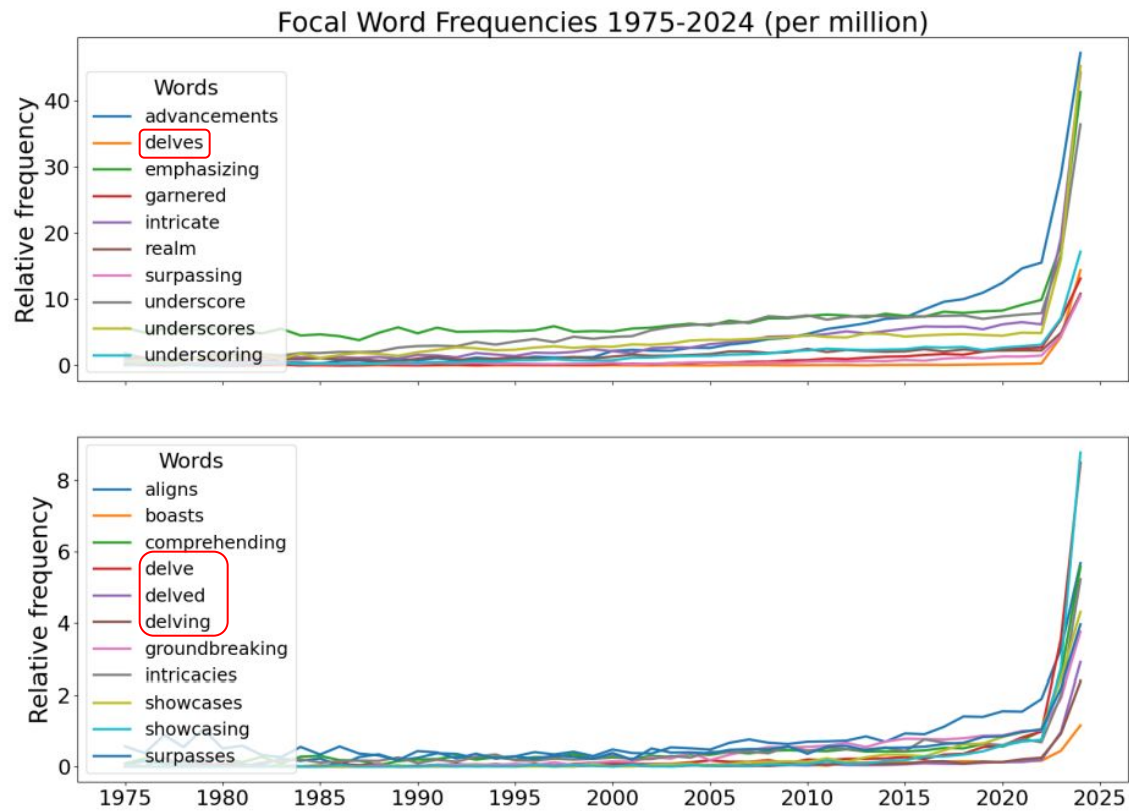
Florida State University

tjuzek@fsu.edu, zward@fsu.edu

Статистические и стилистические детекторы



Статистические и стилистические детекторы



Статистические и стилистические детекторы

Можно детектировать сгенерированный текст по таким признакам, как:

- Частота использования определенных слов, биграмм, триграмм и т.д.
- Стил ь форматирования
- Когерентность
- Оригинальность
- Лексическое разнообразие
- Степень формальности
- Наличие повторений
- и т.д.



Статистические и стилистические детекторы

Больше об отличительных чертах сгенерированных текстов:

**People who frequently use ChatGPT for writing tasks
are accurate and robust detectors of AI-generated text**

Jenna Russell¹ Marzena Karpinska² Mohit Iyyer^{1,3}

**Feature-Level Insights into Artificial Text Detection with Sparse
Autoencoders**

**Kristian Kuznetsov^{1,2}, Laida Kushnareva², Polina Druzhinina^{1,5}, Anton Razzhigaev^{1,5},
Anastasia Voznyuk³, Irina Piontkovskaya², Evgeny Burnaev^{1,5}, Serguei Barannikov^{1,4},**

Статистические и стилистические детекторы

Пример фреймворка для детекции:

GLTR: Statistical Detection and Visualization of Generated Text

Sebastian Gehrmann

Harvard SEAS

`gehrmann@seas.harvard.edu`

Hendrik Strobelt

IBM Research

MIT-IBM Watson AI lab

`hendrik.strobelt@ibm.com`

Alexander M. Rush

Harvard SEAS

`srush@seas.harvard.edu`

<http://demo.gltr.io/client/index.html>

Статистические и стилистические детекторы

Human-Written

The programme operates on a weekly elimination process to find the best all-around baker from the contestants, who are all amateurs.

Generated

The first book I went through was The Cook's Book of New York City by Ed Mirvish. I've always loved Ed Mirvish's recipes and he's one of my favorite chefs.

Supervised классификаторы поверх LM

Чаще всего основаны на трансформерных моделях:

- RoBERTa
- XLM-RoBERTa
- Longformer

Supervised классификаторы поверх LM

Чаще всего основаны на трансформерных моделях:

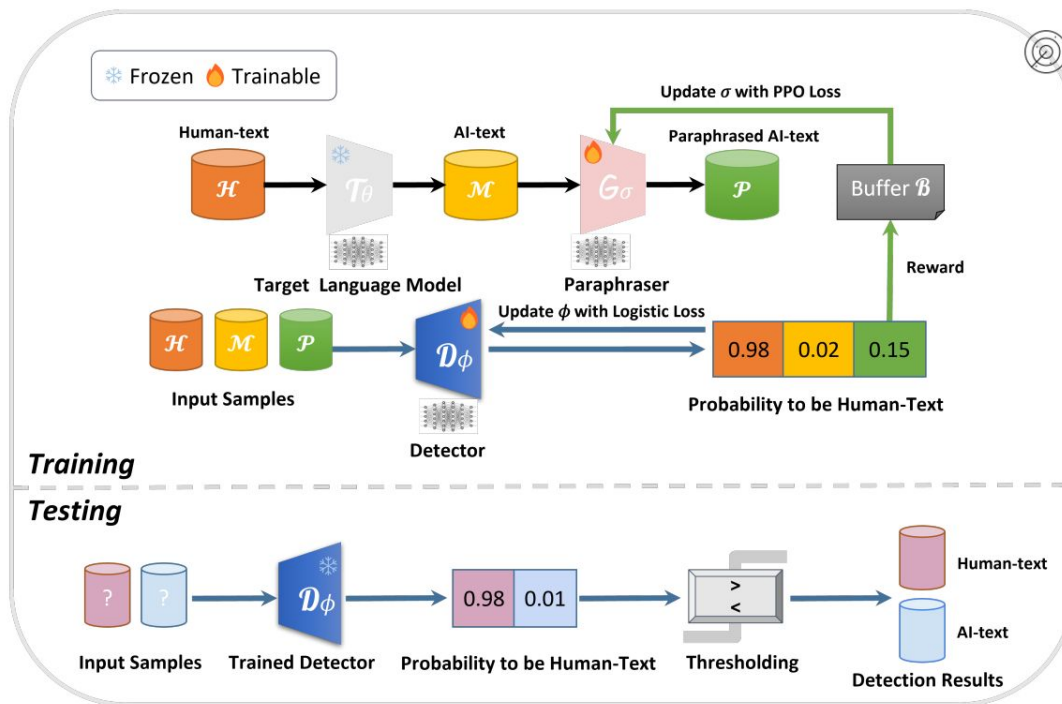
- RoBERTa
- XLM-RoBERTa
- Longformer

В качестве классификатора над эмбедингами обычно используется:

- Логистическая регрессия
- Бустинг

Supervised классификаторы поверх LM

Иногда используют adversarial training. Пример - детектор [RADAR](#):



Коммерческие детекторы

Was this text written by a **human** or AI?

Try detecting one of our sample texts:

Paste your text here...

0/5000 characters

By continuing you agree to our [Terms of service](#)

<https://gptzero.me/>

Enter text to check for AI and ChatGPT Plagiarism

0/15 000 Characters
(Get up to 100,000 here)

<https://www.zerogpt.com/>

Коммерческие детекторы



January 31, 2023

New AI classifier for indicating AI-written text

We're launching a classifier trained to distinguish between AI-written and human-written text.

As of July 20, 2023, the AI classifier is no longer available due to its low rate of accuracy. We are working to incorporate feedback and are currently researching more effective provenance techniques for text, and have made a commitment to develop and deploy mechanisms that enable users to understand if audio or visual content is AI-generated.

План лекции

- Немного истории и мотивация
- Детекторы сгенерированных текстов
 - Основанные на перплексии
 - Статистические/стилистические
 - Supervised классификаторы поверх глубоких моделей
- **Способы обхода детекторов**
 - Использование перефразировщика
 - Использование необычного генератора текста или необычных настроек
 - Постобработка сгенерированного текста
- Watermarking

Способы обхода: перефразировщики

- DIPPER* и другие специально обученные
- Обычные LLM с соответствующим промптом

Способы обхода: перефразировщики

- DIPPER* и другие специально обученные
- Обычные LLM с соответствующим промптом

Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense

Kalpesh Krishna^{♣♡*} Yixiao Song[♣] Marzena Karpinska[♣]
John Wieting^{◇†} Mohit Iyer^{♣†}

[♣]University of Massachusetts Amherst, [♡]Google, [◇]Google DeepMind
{kalpeshk, jwieting}@google.com
yixiaosong@umass.edu {mkarpinska, miyyer}@cs.umass.edu

Способы обхода: необычные настройки

- Изменение типа сэмплирования (top-p, top-k, temperature)
- Изменение настроек сэмплирования
- Промпт-инжиниринг

Способы обхода: необычные настройки

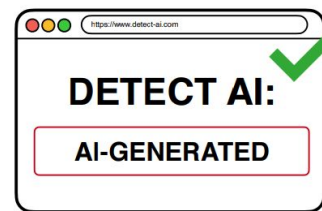
- Изменение типа сэмплирования (top-p, top-k, temperature)
- Изменение настроек сэмплирования
- Промпт-инжиниринг

RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors

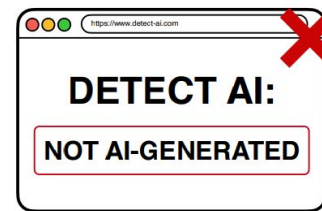
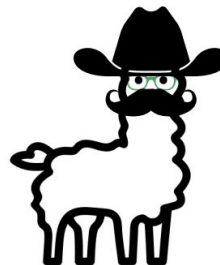
Liam Dugan¹, Alyssa Hwang¹, Filip Trhlik², Josh Magnus Ludan¹
Andrew Zhu¹, Hainiu Xu³, Daphne Ippolito⁴, Chris Callison-Burch¹
University of Pennsylvania¹ University College London²
King's College London³ Carnegie Mellon University⁴
{ldugan, ahwang16, jludan, andrz, ccb}@seas.upenn.edu
hainiu.xu@kcl.ac.uk, filip.trhlik.21@ucl.ac.uk, daphnei@cmu.edu

[“RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors”, Dugan et al.](#)

LLaMA
(default)



LLaMA
+sampling
+penalty



Способы обхода: промпт-инжиниринг

- “Answer the question in slang style”
- “Answer the question in Shakespearean style”
- “Write as non-native speaker”
- “Write so your response won’t be detected by AI detection tools”
- и всё, на что хватит фантазии

Способы обхода: постобработка

- Использование редких выражений, сокр. и очепяток
- Адверсариальная замена слов/предложений ([BERT-attack](#))
- Замена символов на похожие, вставка символов ([Homoglyph](#), [TextBugger](#))
- Дополнительные пробелы и переносы строк ([SpaceInfi](#), [Inserting Paragraphs](#))
- Удаление артиклей
- Вертикально написанный текст ([VertAttack](#))
- и всё, на что хватит фантазии

Способы обхода: постобработка

Human: Describe the structure of an atom.



ChatGPT

- An atom consists of a nucleus, which contains protons and neutrons, and electrons orbiting around the nucleus. The protons have a positive **charge**, the electrons have a negative charge, and the neutrons have no charge...



SpaceInFi

- An atom consists of a nucleus, which contains protons and neutrons, and electrons orbiting around the nucleus. The protons have a positive **charge**, the electrons have a negative charge, and the neutrons have no charge...

[SpaceInFi](#)

one of the w movies of the year . .

o
r
s
t

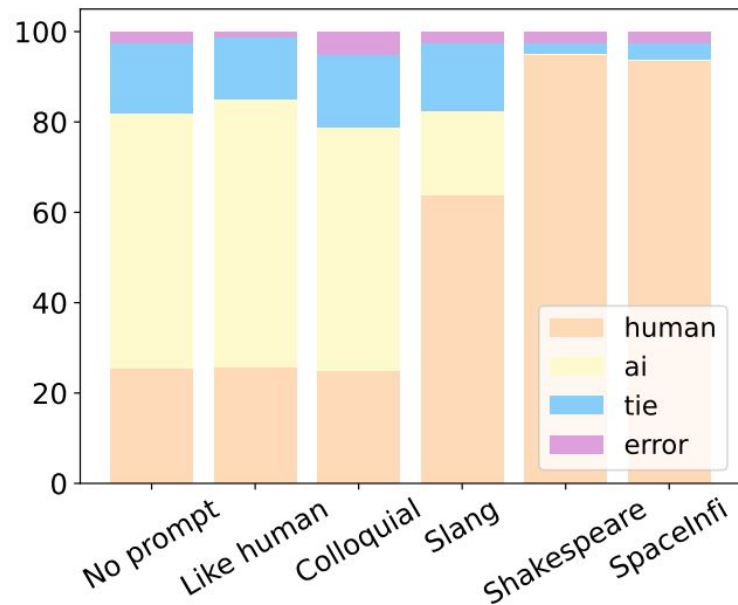
. . watching it was p .

a
i
n
f
u
l

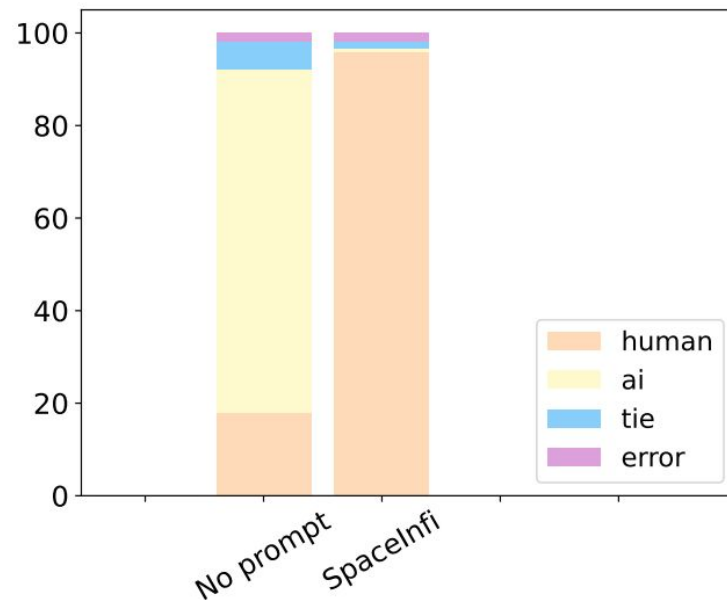
d , a road-trip movie that's surprisingly s of both
u h
l o
l r
t
adventure and song .

[VertAttack](#)

Способы обхода: результаты

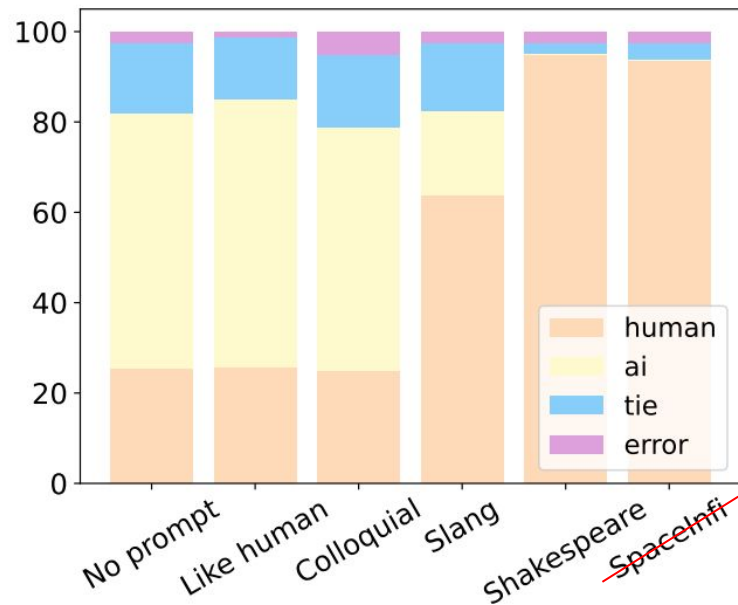


GPTzero on Vicuna

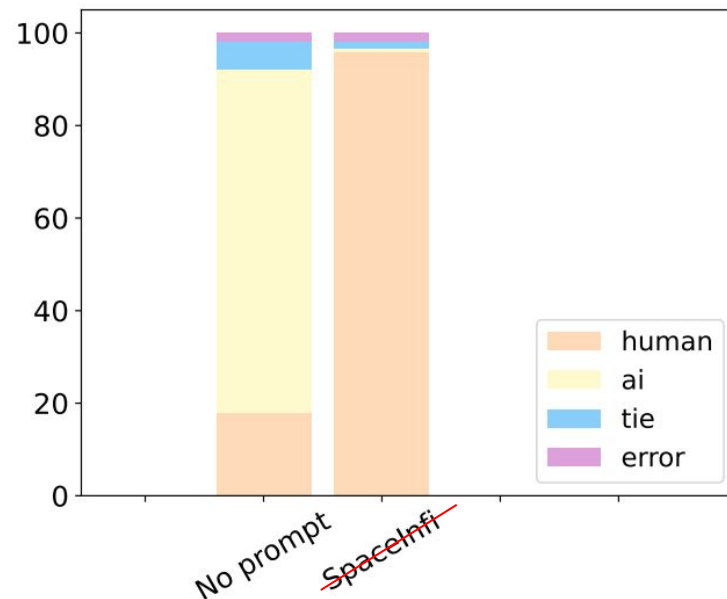


GPTzero on ChatGPT4

Способы обхода: результаты



GPTzero on Vicuna



GPTzero on ChatGPT4

Способы обхода: результаты

	None	Paraphrase	Synonym	Misspelling	Homoglyph	Whitespace	Delete Articles
R-L GPT2	56.7	72.9 (+16.2)	79.4 (+22.7)	39.5 (-17.2)	21.3 (-35.4)	40.1 (-16.6)	33.2 (-23.5)
RADAR	70.9	67.3 (-3.6)	67.5 (-3.4)	69.5 (-1.4)	59.3 (-11.6)	66.1 (-4.8)	67.9 (-3.0)
GLTR	62.6	47.2 (-15.4)	31.2 (-31.4)	59.8 (-2.8)	24.3 (-38.3)	45.8 (-16.8)	52.1 (-10.5)
Binoculars	79.6	80.3 (+0.7)	43.5 (-36.1)	78.0 (-1.6)	37.7 (-41.9)	70.1 (-9.5)	74.3 (-5.3)
GPTZero	66.5	64.0 (-2.5)	61.0 (-5.5)	65.1 (-1.4)	66.2 (-0.3)	66.2 (-0.3)	61.0 (-5.5)
Originality	85.0	96.7 (+11.7)	96.5 (+11.5)	78.6 (-6.4)	9.3 (-75.7)	84.9 (-0.1)	71.4 (-13.6)

5% FPR

Способы обхода: результаты

	RoBERTa (GPT2)				GPTZero				RADAR			
Books	0.987	0.588	0.287	0.548	0.405	1.000	1.000	0.265	0.768	0.992	0.965	0.689
News	0.996	0.694	0.415	0.640	0.280	1.000	1.000	0.190	0.810	0.999	0.999	0.663
Reddit	0.992	0.437	0.252	0.477	0.258	0.975	0.840	0.148	0.491	0.969	0.792	0.467
Reviews	0.976	0.612	0.387	0.462	0.455	0.995	1.000	0.320	0.222	0.004	0.007	0.118
Wiki	0.959	0.695	0.332	0.373	0.422	0.995	0.975	0.270	0.706	0.999	0.963	0.613
	GPT2	ChatGPT	GPT4	Mistral	GPT2	ChatGPT	GPT4	Mistral	GPT2	ChatGPT	GPT4	Mistral

Защита от атак

От всех возможных угроз защититься невозможно, но от некоторых - можно:

- Обработка текста на входе детектора отчасти помогает против постобработки
- Упор на робастность при тренировке отчасти помогает при смене доменов/генераторов (RADAR и др.)
- От новых моделей, промпт-инжиниринга и перефразировщиков защититься сложнее всего

План лекции

- Немного истории и мотивация
- Детекторы сгенерированных текстов
 - Основанные на перплексии
 - Статистические/стилистические
 - Supervised классификаторы поверх глубоких моделей
- Способы обхода детекторов
 - Использование перефразировщика
 - Использование необычного генератора текста или необычных настроек
 - Постобработка сгенерированного текста
- **Watermarking**

Watermarking

A Watermark for Large Language Models

John Kirchenbauer^{*} Jonas Geiping^{*} Yuxin Wen Jonathan Katz Ian Miers Tom Goldstein
University of Maryland

Watermarking

Algorithm 1 Text Generation with Hard Red List

Input: prompt, $s^{(-N_p)} \dots s^{(-1)}$

for $t = 0, 1, \dots$ **do**

1. Apply the language model to prior tokens $s^{(-N_p)} \dots s^{(t-1)}$ to get a probability vector $p^{(t)}$ over the vocabulary.
2. Compute a hash of token $s^{(t-1)}$, and use it to seed a random number generator.
3. Using this seed, randomly partition the vocabulary into a “green list” G and a “red list” R of equal size.
4. Sample $s^{(t)}$ from G , never generating any token in the red list.

end for

Watermarking

Algorithm 1 Text Generation with Hard Red List

Input: prompt, $s^{(-N_p)} \dots s^{(-1)}$

for $t = 0, 1, \dots$ **do**

1. Apply the language model to prior tokens $s^{(-N_p)} \dots s^{(t-1)}$ to get a probability vector $p^{(t)}$ over the vocabulary.
2. Compute a hash of token $s^{(t-1)}$, and use it to seed a random number generator.
3. Using this seed, randomly partition the vocabulary into a “green list” G and a “red list” R of equal size.
4. Sample $s^{(t)}$ from G , never generating any token in the red list.

end for

Algorithm 2 Text Generation with Soft Red List

Input: prompt, $s^{(-N_p)} \dots s^{(-1)}$
green list size, $\gamma \in (0, 1)$
hardness parameter, $\delta > 0$

for $t = 0, 1, \dots$ **do**

1. Apply the language model to prior tokens $s^{(-N_p)} \dots s^{(t-1)}$ to get a logit vector $l^{(t)}$ over the vocabulary.
2. Compute a hash of token $s^{(t-1)}$, and use it to seed a random number generator.
3. Using this random number generator, randomly partition the vocabulary into a “green list” G of size $\gamma|V|$, and a “red list” R of size $(1 - \gamma)|V|$.
4. Add δ to each green list logit. Apply the soft-max operator to these modified logits to get a probability distribution over the vocabulary.

$$\hat{p}_k^{(t)} = \begin{cases} \frac{\exp(l_k^{(t)} + \delta)}{\sum_{i \in R} \exp(l_i^{(t)}) + \sum_{i \in G} \exp(l_i^{(t)} + \delta)}, & k \in G \\ \frac{\exp(l_k^{(t)})}{\sum_{i \in R} \exp(l_i^{(t)}) + \sum_{i \in G} \exp(l_i^{(t)} + \delta)}, & k \in R. \end{cases}$$

5. Sample the next token, $s^{(t)}$, using the water-marked distribution $\hat{p}^{(t)}$.

end for

Watermarking

Prompt

...The watermark detection algorithm can be made public, enabling third parties (e.g., social media platforms) to run it themselves, or it can be kept private and run behind an API. We seek a watermark with the following properties:

No watermark

Extremely efficient on average term lengths and word frequencies on synthetic, microamount text (as little as 25 words)

Very small and low-resource key/hash (e.g., 140 bits per key is sufficient for 99.99999999% of the Synthetic Internet)

With watermark

- minimal marginal probability for a detection attempt.
- Good speech frequency and energy rate reduction.
- messages indiscernible to humans.
- easy for humans to verify.

Algorithm 2 Text Generation with Soft Red List

Input: prompt, $s^{(-N_p)} \dots s^{(-1)}$
green list size, $\gamma \in (0, 1)$
hardness parameter, $\delta > 0$

for $t = 0, 1, \dots$ **do**

1. Apply the language model to prior tokens $s^{(-N_p)} \dots s^{(t-1)}$ to get a logit vector $l^{(t)}$ over the vocabulary.
2. Compute a hash of token $s^{(t-1)}$, and use it to seed a random number generator.
3. Using this random number generator, randomly partition the vocabulary into a “green list” G of size $\gamma|V|$, and a “red list” R of size $(1 - \gamma)|V|$.
4. Add δ to each green list logit. Apply the soft-max operator to these modified logits to get a probability distribution over the vocabulary.

$$\hat{p}_k^{(t)} = \begin{cases} \frac{\exp(l_k^{(t)} + \delta)}{\sum_{i \in R} \exp(l_i^{(t)}) + \sum_{i \in G} \exp(l_i^{(t)} + \delta)}, & k \in G \\ \frac{\exp(l_k^{(t)})}{\sum_{i \in R} \exp(l_i^{(t)}) + \sum_{i \in G} \exp(l_i^{(t)} + \delta)}, & k \in R. \end{cases}$$

5. Sample the next token, $s^{(t)}$, using the water-marked distribution $\hat{p}^{(t)}$.

end for

Список литературы (детекция)

- Dewdney, A. K. [Computer Recreations. A potpourri of programmed prose and prosody](#), In Scientific American, Vol. 260, No. 6, JUNE 1989.
- Lavergne et al. [Detecting fake content with relative entropy scoring](#). In Proceedings of the 2008 International Conference on Uncovering Plagiarism, Authorship and Social Software Misuse, Vol. 377.
- Kushnareva et al. [AI-generated text boundary detection with RoFT](#). In Proceedings of the First Conference on Language Modeling (COLM), 2024.
- Mitchell et al. [DetectGPT: zero-shot machine-generated text detection using probability curvature](#). In Proceedings of the 40th International Conference on Machine Learning (ICML), 2023.
- Hans et al. [Spotting LLMs with binoculars: zero-shot detection of machine-generated text](#). In Proceedings of the 41st International Conference on Machine Learning (ICML), 2024.

Список литературы (детекция)

- Juzek et al. [Why Does ChatGPT “Delve” So Much? Exploring the Sources of Lexical Overrepresentation in Large Language Models](#). In Proceedings of the 31st International Conference on Computational Linguistics (COLING), 2025.
- Russell et al. [People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text](#). ArXiv, 2025.
- Kuznetsov et al. [Feature-Level Insights into Artificial Text Detection with Sparse Autoencoders](#). ArXiv, 2025.
- Gehrmann et al. [GLTR: Statistical Detection and Visualization of Generated Text](#). In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL), 2019.
- Hu et al. [RADAR: robust AI-text detection via adversarial learning](#). In Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS), 2023.

Список литературы (атаки и watermarking)

- Krishna et al. [Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense](#). In Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS), 2023.
- Dugan et al. [RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors](#). In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2024.
- Li et al. [BERT-ATTACK: Adversarial Attack Against BERT Using BERT](#). In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).
- Wolff et al. [Attacking Neural Text Detectors](#). ArXiv, 2020.
- Li et al. [TextBugger: Generating Adversarial Text Against Real-world Applications](#). In Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS), 2019.

Список литературы (атаки и watermarking)

- Cai et al. [Evade ChatGPT Detectors via A Single Space](#). ArXiv, 2023.
- Rusert, J. [VertAttack: Taking Advantage of Text Classifiers' Horizontal Vision](#). In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL).
- Kirchenbauer et al. [A Watermark for Large Language Models](#). In Proceedings of the 40th International Conference on Machine Learning (ICML), 2023.
- Dathathri et al. [Scalable watermarking for identifying large language model outputs](#). In Nature 634, 2024.