

How to use ChatGPT without getting caught

Aldan Creo

How to use ChatGPT without getting caught

What are LLMs?

Common techniques

Circumvention



What are LLMs?

Common techniques

Circumvention



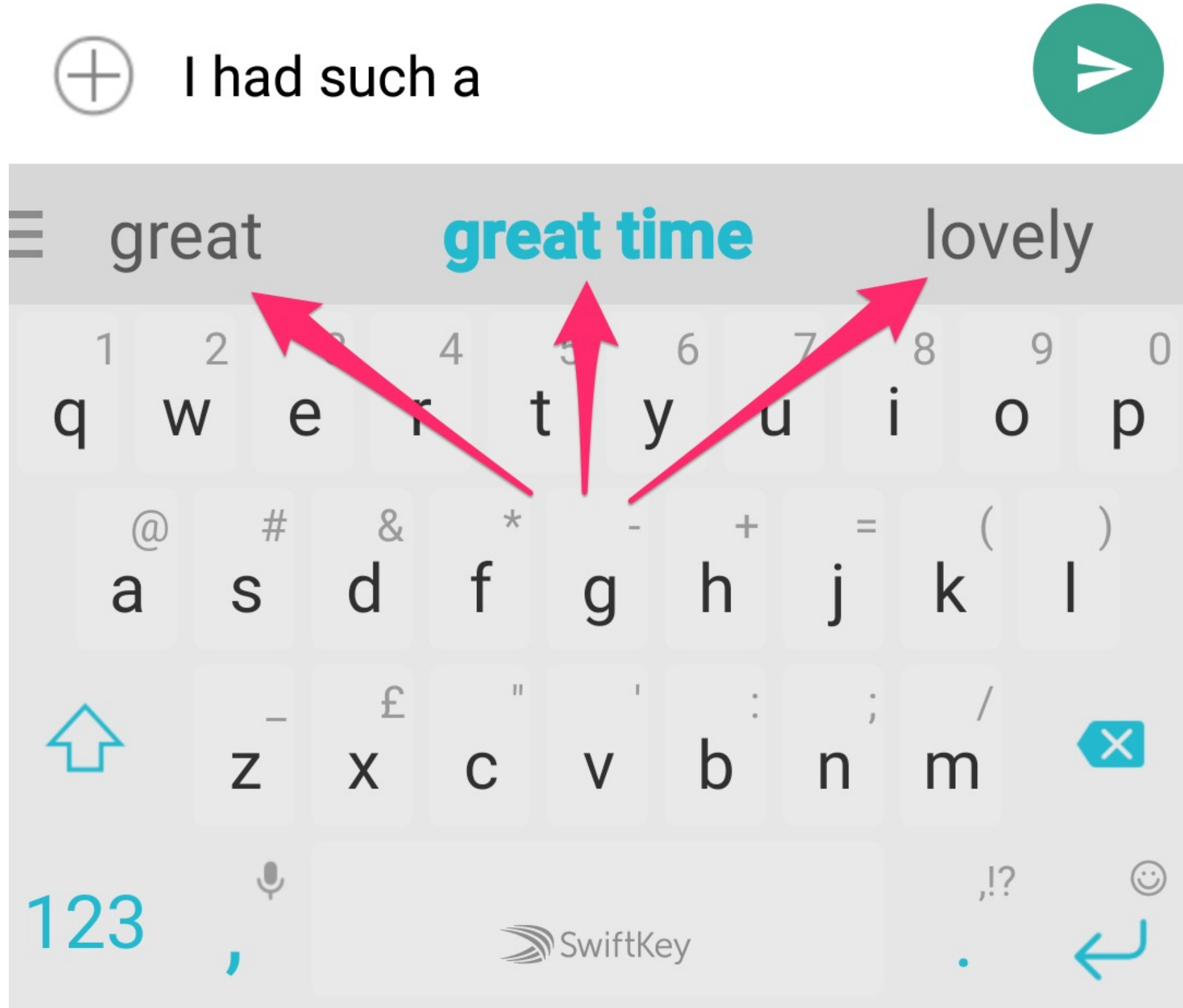
What are LLMs?



LLMs are text predictors



LLMs are text predictors





LLMs are text predictors

in out

We need to stop

We need to stop anthrop

We need to stop anthropomorph

We need to stop anthropomorphizing

We need to stop anthropomorphizing Chat

We need to stop anthropomorphizing ChatG

We need to stop anthropomorphizing ChatGPT

We need to stop anthropomorphizing ChatGPT.

LLMs are text predictors





LLMs are text predictors

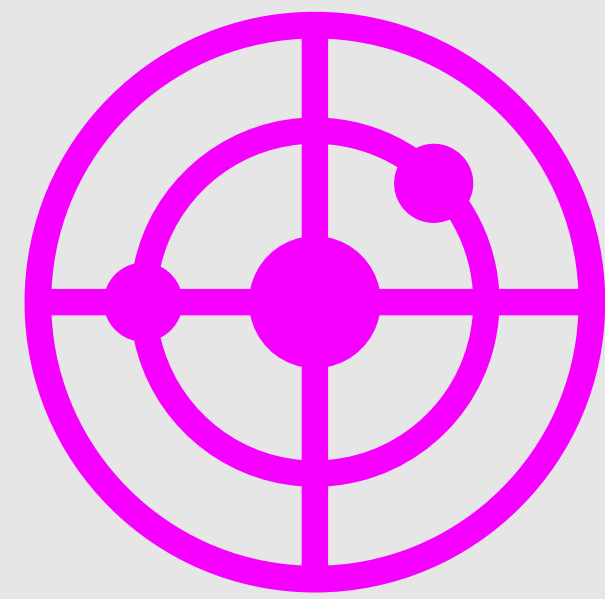




Common techniques

Circumvention

Ethics



Common techniques

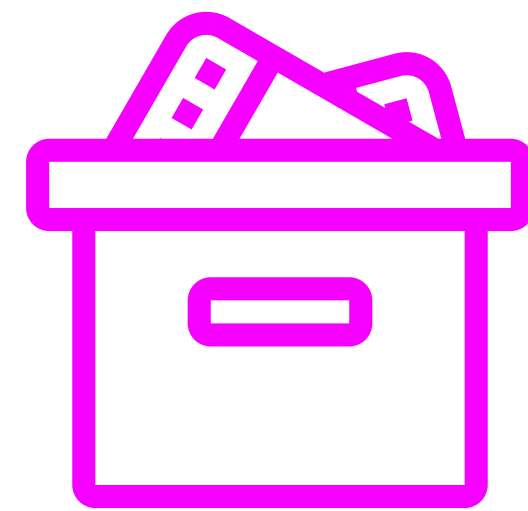
Classifiers

What's a classifier?

What's a classifier?

Yes / No

What's a classifier?



Training examples

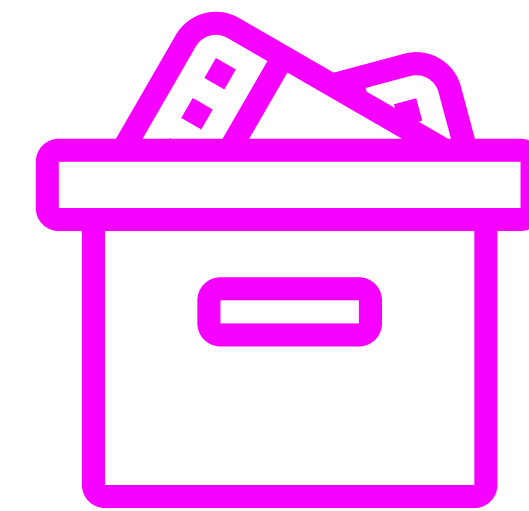


Test examples

What's a classifier?



Training examples

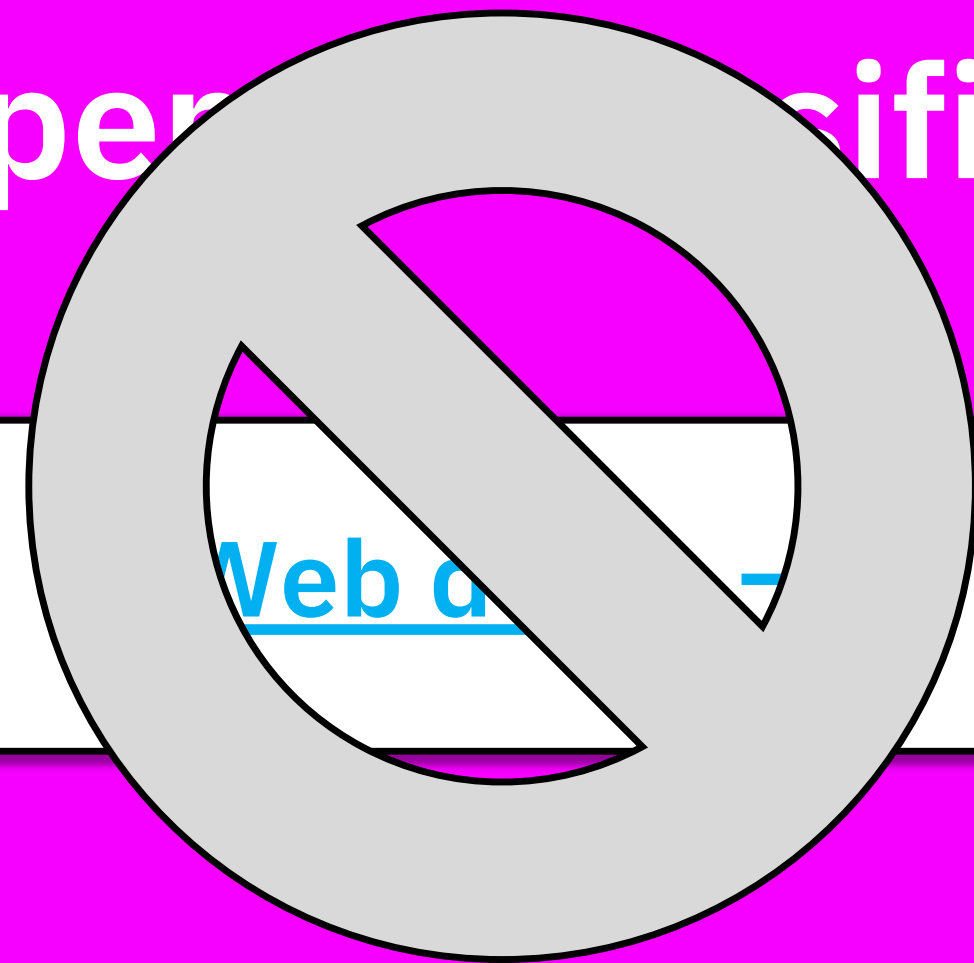
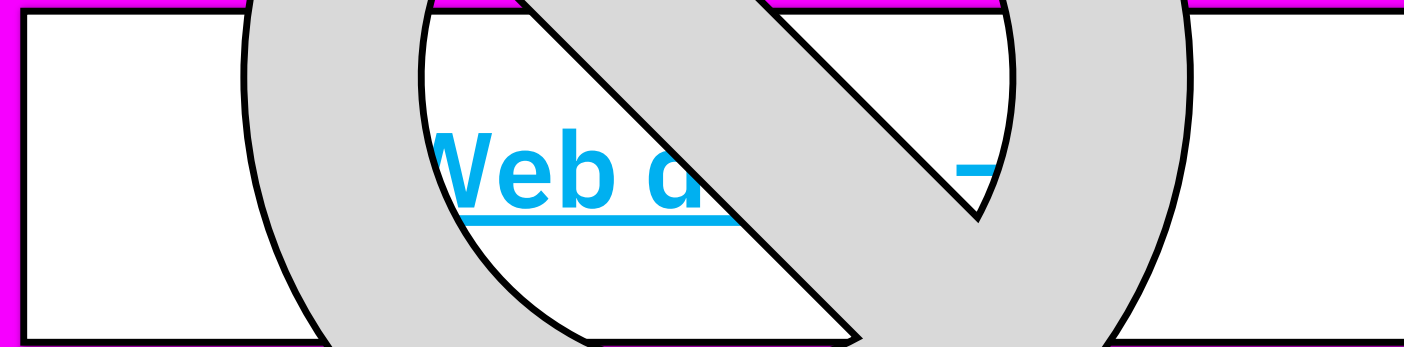


Test examples

OpenAI classifier

[Web demo →](#)

Open specific



Web d

Ghostbuster

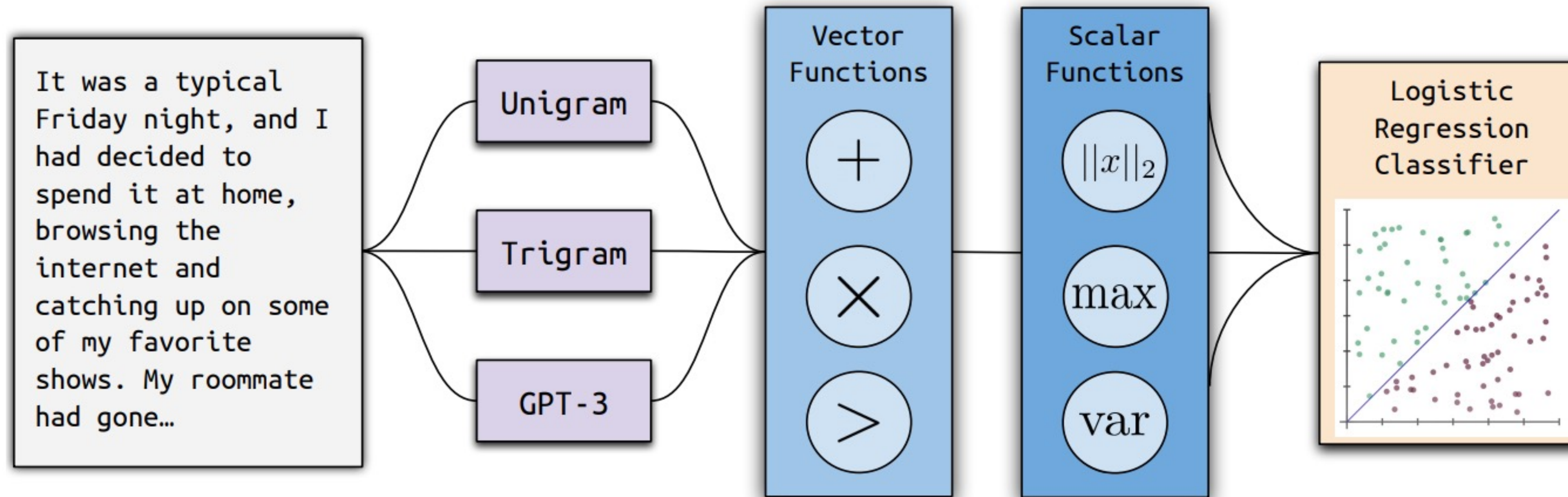
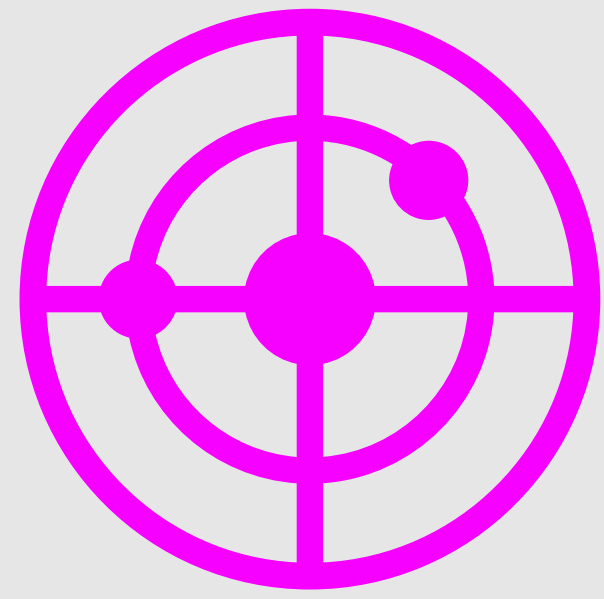


Figure 1: An outline of our model training procedure. First, we fed each document into a series of weaker language models to obtain token probabilities. Then, we ran a structured search over combinations of the model outputs and trained a linear classifier on the selected features.

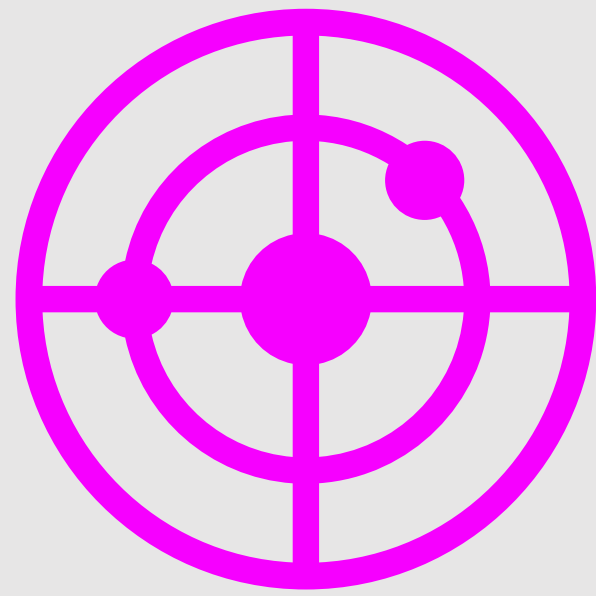
Ghostbuster

[Web demo →](#)



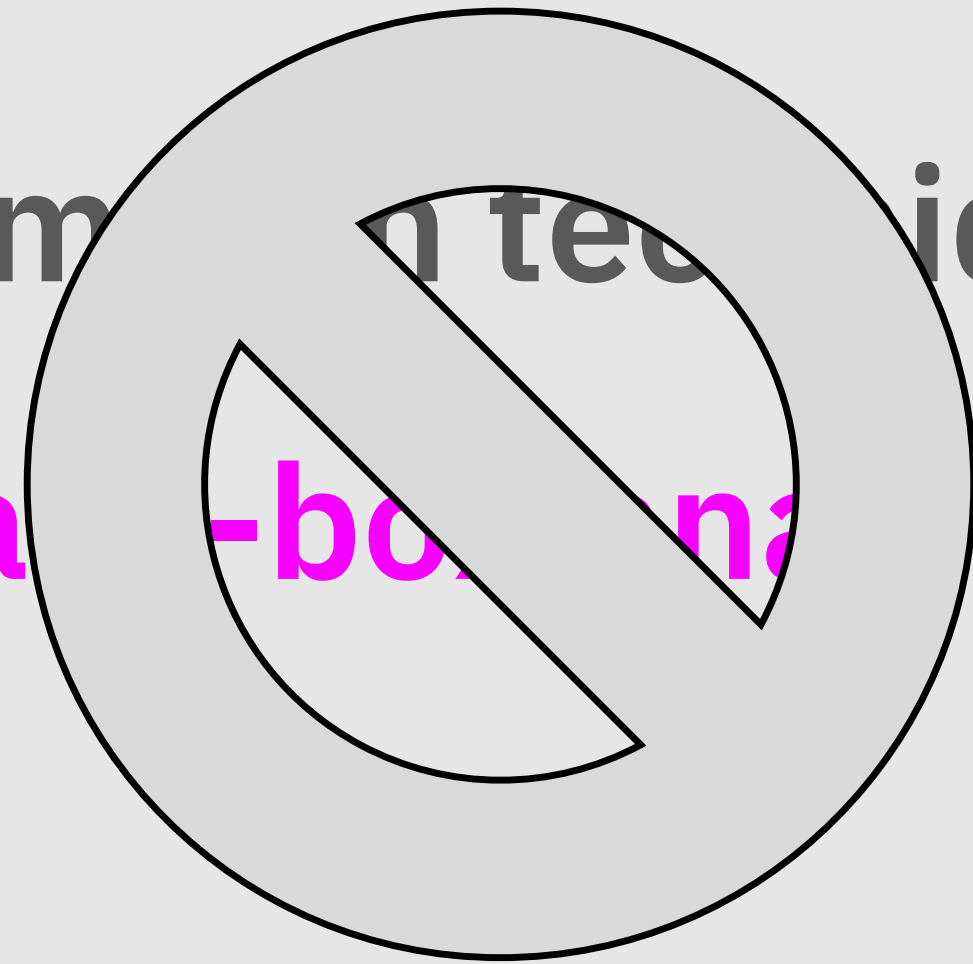
Common techniques

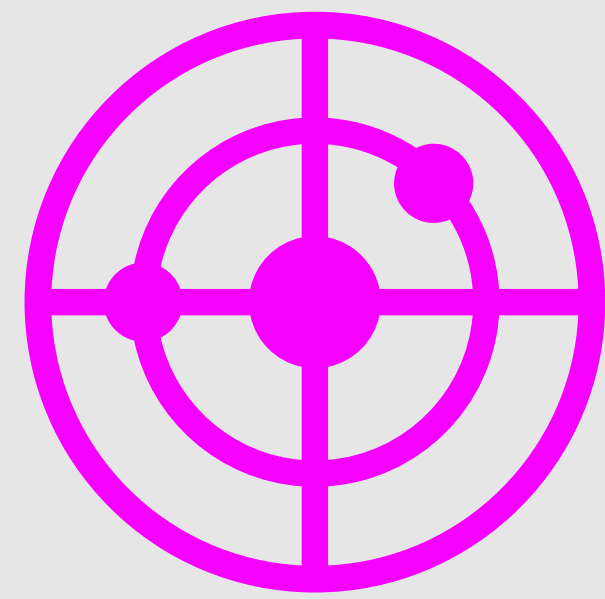
Black-box analysis



Common techniques

Black-box analysis



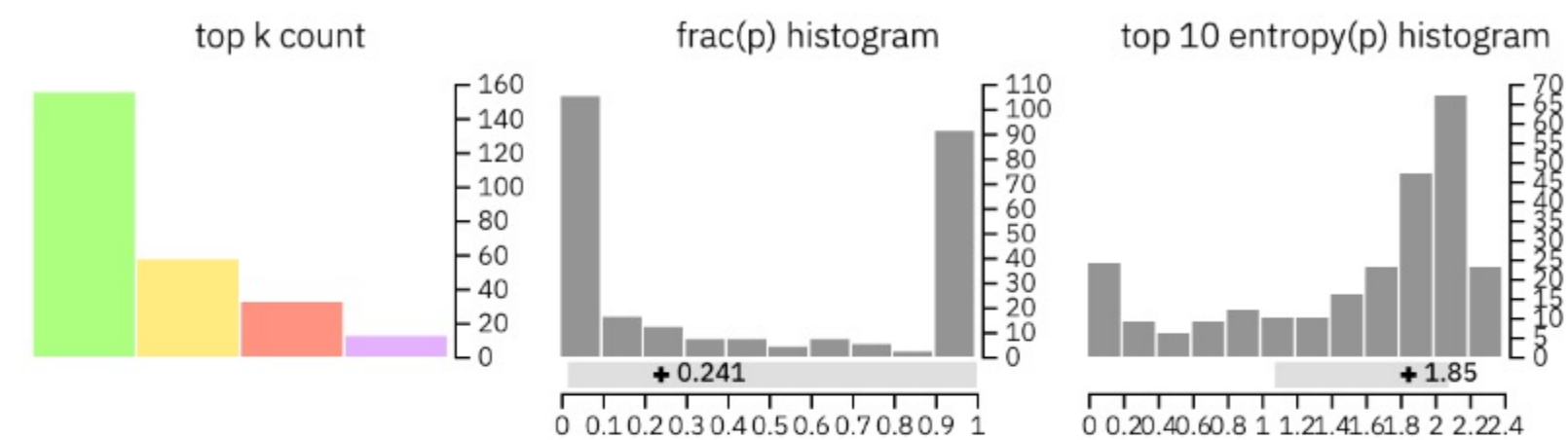


Common techniques

White-box analysis

GLTR

+ Real text



Top K: 10 100 1000
Frac P: 0.241
Colors (top k): 10 100 1000

Following Cas9 cleavage, DNA repair without a donor template is generally considered stochastic, heterogeneous and impractical beyond gene disruption. Here, we show that template-free Cas9 editing is predictable and capable of precise repair to a predicted genotype, enabling correction of disease-associated mutations in humans. We constructed a library of 2,000 Cas9 guide RNAs paired with DNA target sites and trained inDelphi, a machine learning model that predicts genotypes and frequencies of 1- to 60-base-pair deletions and 1-base-pair insertions with high accuracy ($r=0.87$) in five human and mouse cell lines. inDelphi predicts that 11% of Cas9 guide RNAs targeting the human genome are precise-50, yielding a single genotype comprising greater than or equal to 50% of all major editing products. We experimentally confirmed precise-50 insertions and deletions in 195 human disease-relevant alleles, including correction in primary patient-derived fibroblasts of pathogenic alleles to wild-type genotype for Hermansky-Pudlak syndrome and Menkes disease. This study establishes an approach for precise, template-free genome editing.

+ GPT-2

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge PÃ©rez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. PÃ©rez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

PÃ©rez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said PÃ©rez.

PÃ©rez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them - they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. PÃ©rez stated, "We can see, for example, that they have a common language, something like a dialect or dialectic."

Dr. PÃ©rez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to PÃ©rez, "In South America, such incidents seem to be quite common."

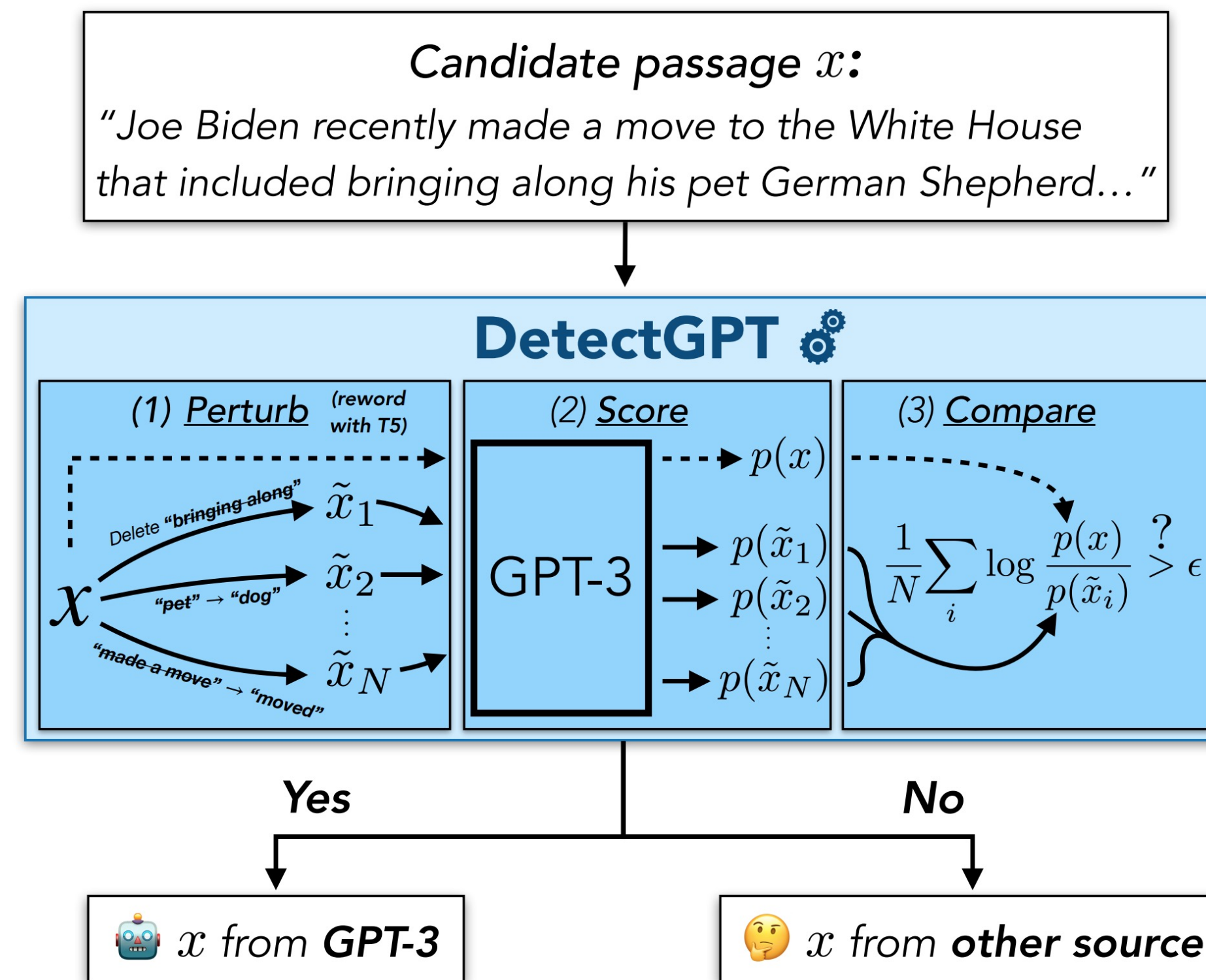
However, PÃ©rez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. "But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization," said the scientist.

GLTR

[Web demo →](#)

Exploiting probability distributions

DetectGPT



DetectGPT

The “copybara problem”

The “capybara problem”

User prompt

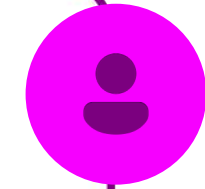
Can you write a few sentences about a capybara that is an astrophysicist?



The “capybara problem”

User prompt

Can you write a few sentences about a capybara that is an astrophysicist?



The “capybara problem”

User prompt

Can you write a few sentences about a capybara that is an astrophysicist?

ChatGPT

Dr. Capy Cosmos, a capybara unlike any other, astounded the scientific community with his groundbreaking research in astrophysics. With his keen sense of observation and unparalleled ability to interpret cosmic data, he uncovered new insights into the mysteries of black holes and the origins of the universe.

The “capybara problem”

User prompt

Can you write a few sentences about a capybara that is an astrophysicist?

ChatGPT

Dr. Capy Cosmos, a capybara unlike any other, astounded the scientific community with his groundbreaking research in astrophysics. With his keen sense of observation and unparalleled ability to interpret cosmic data, he uncovered new insights into the mysteries of black holes and the origins of the universe.

Conditioned probability

The “capybara problem”

Surprising!

ChatGPT



Dr. Capy Cosmos, a capybara unlike any other, astounded the scientific community with his groundbreaking research in astrophysics. With his keen sense of observation and unparalleled ability to interpret cosmic data, he uncovered new insights into the mysteries of black holes and the origins of the universe.

Exploiting probability distributions

Binoculars

SPOTTING LLMs WITH BINOCULARS: ZERO-SHOT DETECTION OF MACHINE-GENERATED TEXT

Abhimanyu Hans*
University of Maryland

Avi Schwarzschild*
Carnegie Mellon University

Valeriia Cherepanova
University of Maryland

Hamid Kazemi
University of Maryland

Aniruddha Saha
University of Maryland

Micah Goldblum
New York University

Jonas Geiping
ELLIS Institute & MPI for Intelligent Systems,
Tübingen AI Center

Tom Goldstein
University of Maryland

Exploiting probability distributions

Binoculars

$$\frac{\text{perplexity}(\text{text})}{\text{cross-perplexity}(\text{text})}$$

Exploiting probability distributions

Binoculars

How surprising a text is to M1



$$\frac{\text{perplexity}(\text{text})}{\text{cross-perplexity}(\text{text})}$$



How surprising the predictions of M2 are to M1

Exploiting probability distributions

Binoculars

Something can be surprising...

$$\frac{\text{perplexity}(\text{text})}{\text{cross-perplexity}(\text{text})}$$

...so we normalize it by the expected surprise of an LLM on that text

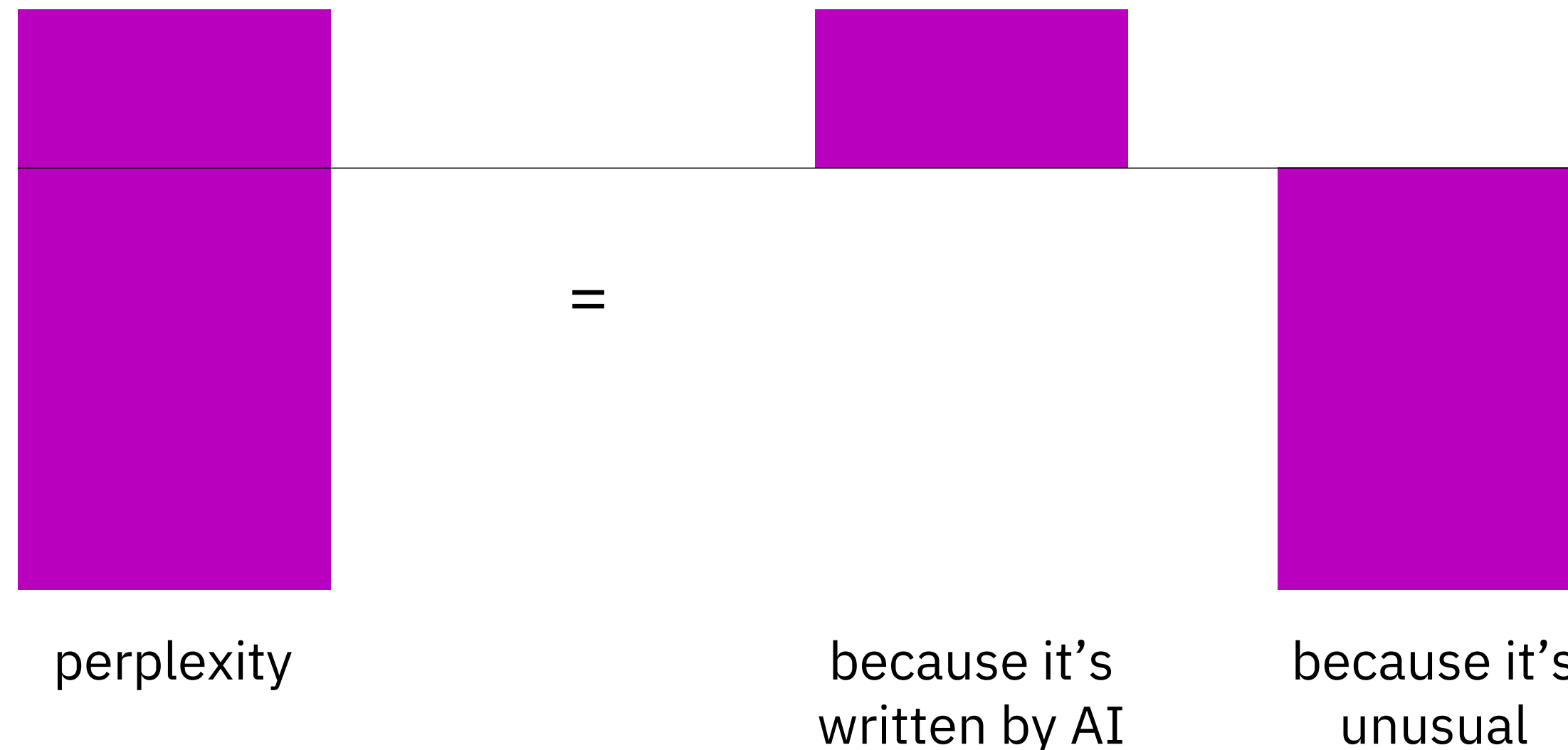
Dr. Capy Cosmos, a capybara unlike any other, astounded the scientific community with his groundbreaking research in astrophysics. With his keen sense of observation and unparalleled ability to interpret cosmic data, he uncovered new insights into the mysteries of black holes and the origins of the universe.

Dr. Capy Cosmos, a capybara unlike any other, astounded the scientific community with his groundbreaking research in astrophysics. With his keen sense of observation and unparalleled ability to interpret cosmic data, he uncovered new insights into the mysteries of black holes and the origins of the universe.



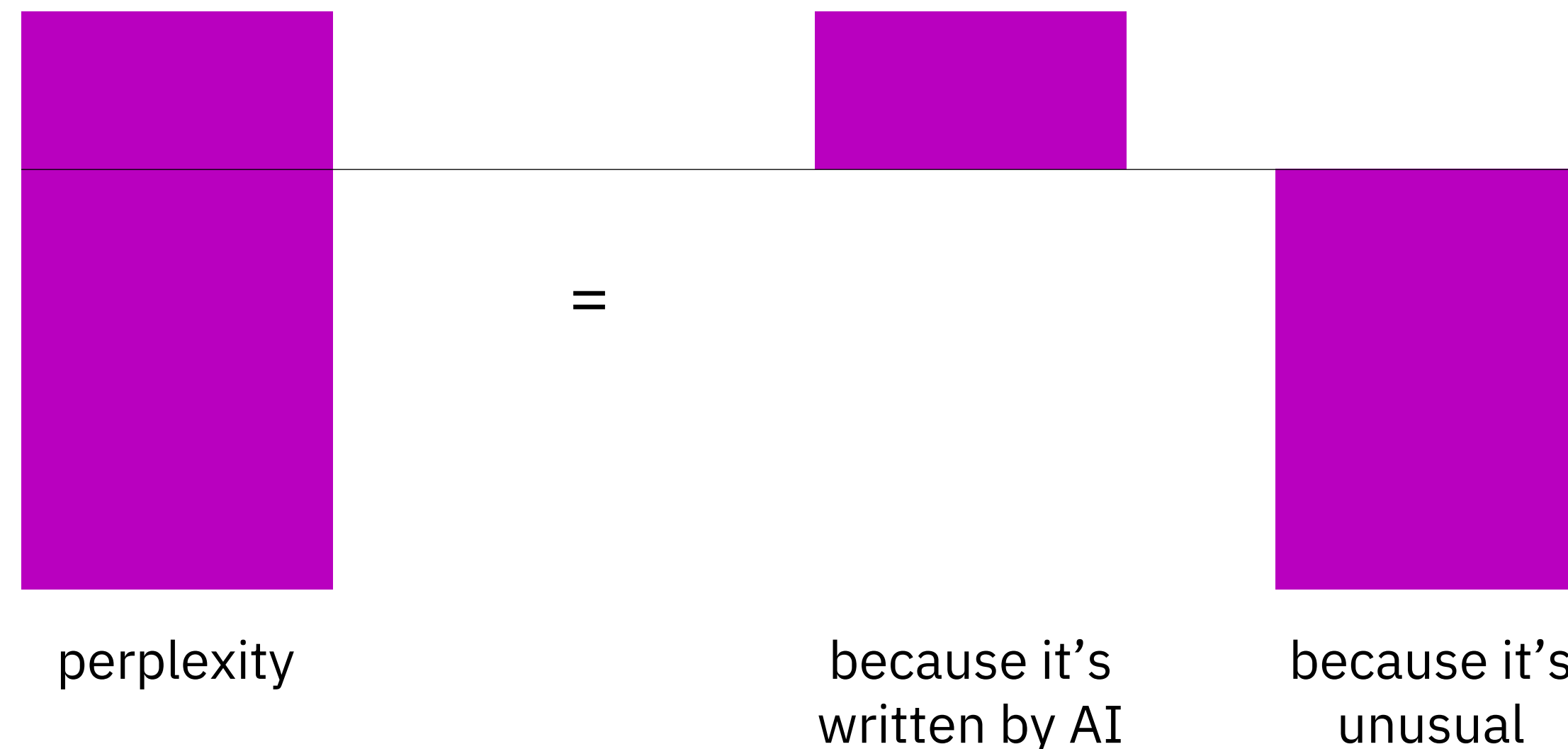
perplexity

Dr. Capy Cosmos, a capybara unlike any other, astounded the scientific community with his groundbreaking research in astrophysics. With his keen sense of observation and unparalleled ability to interpret cosmic data, he uncovered new insights into the mysteries of black holes and the origins of the universe.



Dr. Capy Cosmos, a capybara unlike any other, astounded the scientific community with his groundbreaking research in astrophysics. With his keen sense of observation and unparalleled ability to interpret cosmic data, he uncovered new insights into the mysteries of black holes and the origins of the universe.

...so we normalize it by the expected surprise of an LLM on that text



Exploiting probability distributions

Binoculars

Something can be surprising...

$$\frac{\text{perplexity}(\text{text})}{\text{cross-perplexity}(\text{text})}$$

...so we normalize it by the expected surprise of an LLM on that text

Exploiting probability distributions

Binoculars

$$\frac{\text{perplexity}(\text{text})}{\text{cross-perplexity}(\text{text})}$$

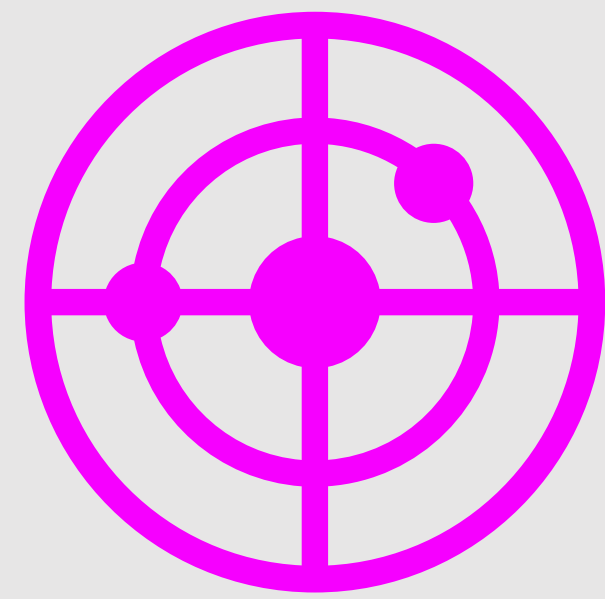
Magic ratio: 0.85

human



AI

Binoculars



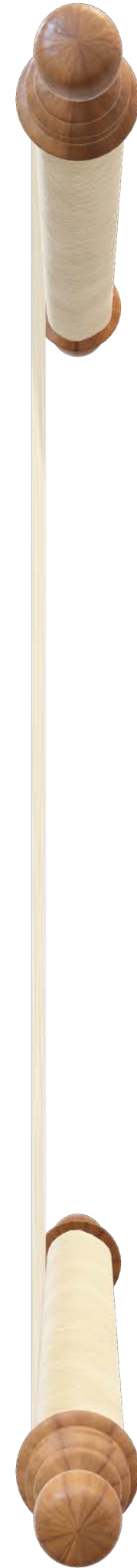
Common techniques

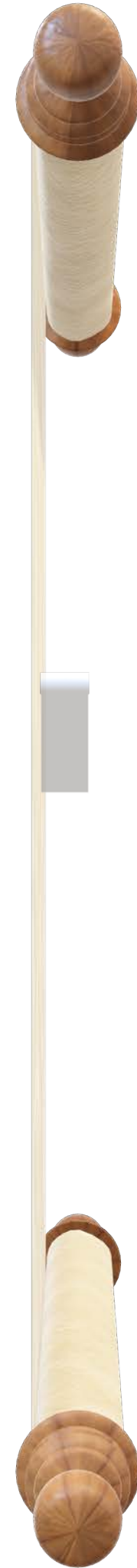
Watermarking





TOP SECRET







TOP SECRET

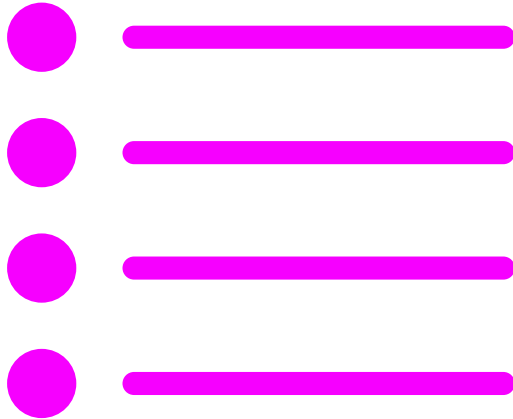
Strong watermarking

Strong watermarking

Red / green list

Strong watermarking

Red / green list



Red list



Green list

Strong watermarking

Red / green list

- Tree
- House
- Park
- Fountain
- Obama
- Note

- Car
- Street
- River
- Pizza
- Barack
- London

Strong watermarking

Red / green list

- Tree
- House
- Park
- Fountain
- **Obama**
- Note

- Car
- Street
- River
- Pizza
- **Barack**
- London

Weak watermarking

Red / green list

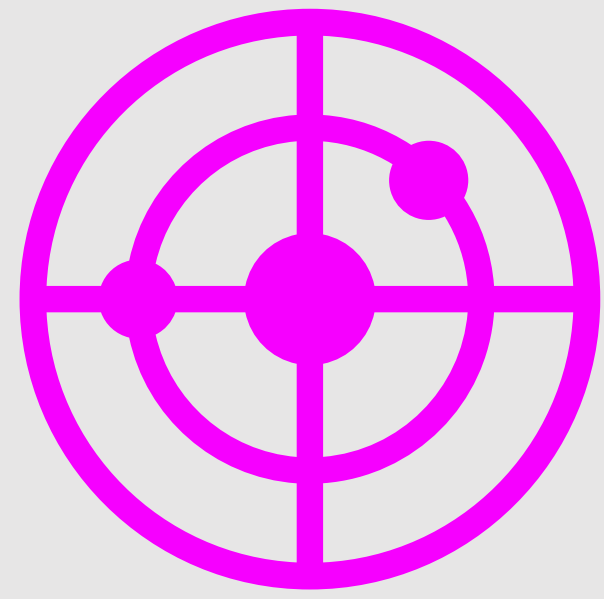
- Tree
 - House
 - Park
 - Fountain
 - Obama
 - Note
- Car
 - Street
 - River
 - Pizza
 - Barack
 - London

Weak watermarking

Red / green list... whenever you can

- Tree
- House
- Park
- Fountain
- **Obama**
- Note
- Car
- Street
- River
- Pizza
- **Barack**
- London

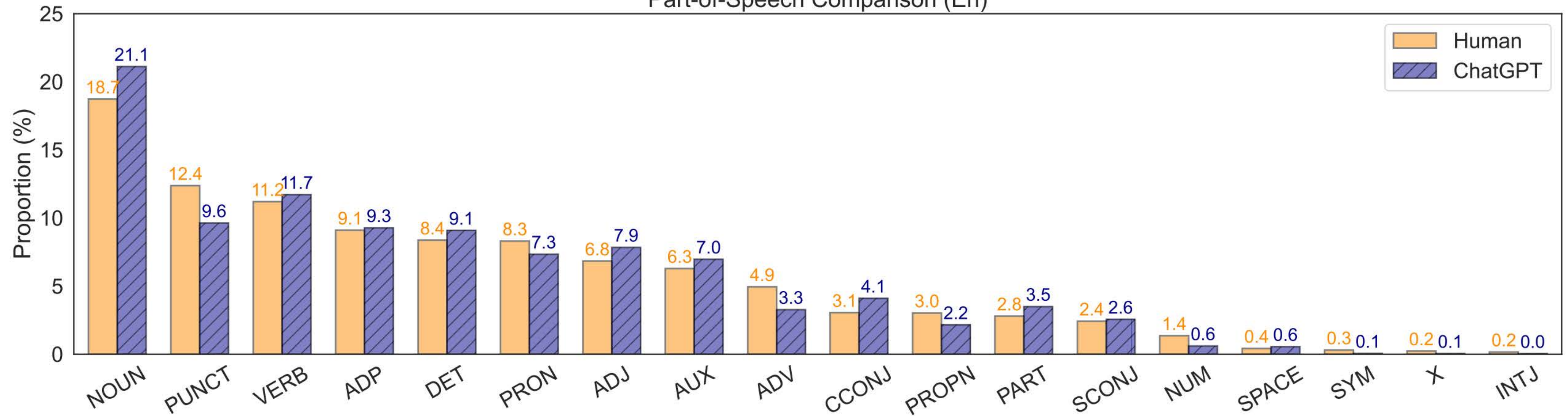
~~The “copybara problem”~~



Common techniques

Manual techniques

Part-of-Speech Comparison (En)



Writing style

Dialects

Typos

*The sun was shining and birds were singing. My sister was playing the piano when the **pone** rang.*

Hallucinations

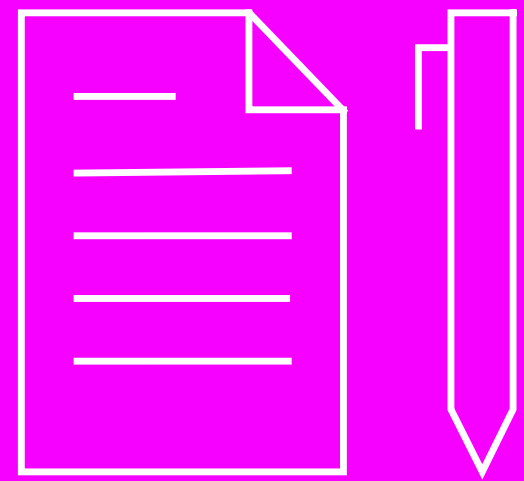
The sun was shining that night and birds were singing. My sister was playing the piano when the phone rang.

ERROR TYPE	DEFINITION	EXAMPLE
Language Errors		
Grammar and Usage	Missing, extra, incorrect, or out of order words	...explaining how cats feel emoticons ...
Off-Prompt	Generation is unrelated to or contradicts prompt	PROMPT: Dogs are the new kids. GENERATION: Visiting the dentist can be scary
Redundant	Lexical, semantic, or excessive topical repetition	Merchants worry about poor service or service that is bad ...
Self-Contradiction	Generation contradicts itself	Amtrak plans to lay off many employees, though it has no plans cut employee hours.
Incoherent	Confusing, but not any error type above	Mary gave her kids cheese toast but drew a map of it on her toast.
Factual Errors		
Bad Math	Math or conversion mistakes	... it costs over £1,000 (\$18,868) ...
Encyclopedic	Facts that annotator knows are wrong	Japanese Prime Minister Justin Trudeau said Monday ...
Commonsense	Violates basic understanding of the world	The dress was made at the spa.
Reader Issues		
Needs Google	Search needed to verify claim	Jose Celana, an artist based in Pensacola, FL, ...
Technical Jargon	Text requires expertise to understand	... an 800-megawatt photovoltaic plant was built ...



Human annotators are only slightly better than random guessing at identifying machine-generated texts.

[Deepfake Text Detection in the Wild](#)

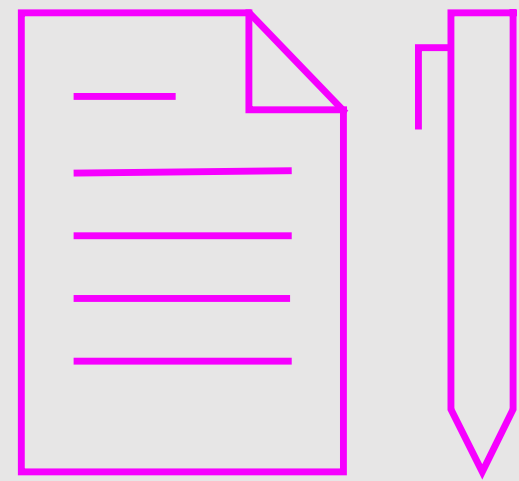


Circumvention



Ideas?





Circumvention

Paraphrasing

In a shocking finding, scientist discovered a herd of unicorns living in a remote valley.

(watermarked tokens in green, benign tokens in red)

 **OpenAI**
ChatGPT / GPT3

They have never been known to mingle with humans. Today, it is believed these unicorns live in an unspoilt environment which is surrounded by mountains. Its edge is protected by a thick wattle of wattle trees, giving it a majestic appearance. Along with their so-called miracle of multicolored' coat, their golden coloured feather makes them look like mirages. Some of them are rumored to be capable of speaking a large amount of different languages. They feed on elk and goats as they were selected from those animals that possess a fierceness to them, and can "eat" them with their long horns.

Watermark detector

prediction:
AI-written
(z=6.3)

DIPPER (our paraphraser)

There were never any reports of them mixing with people. It is believed they live in an unspoiled environment surrounded by mountains and protected by a thick clump of wattle. The herd has a regal look to it, with the magic, rainbow-colored coat and golden feathers. Some of them are said to be capable of speaking many languages. They eat deer and goats, because they are the descendants of those animals that sprang from fierce, dangerous animals and have horns long enough to "eat" these animals.

Watermark detector

prediction:
Unclear
(z=1.8)



Text	# tokens	# green tokens	Accuracy	Perplexity
Original	19042	11078	97%	6.7
Paraphrase with PEGASUS	16773	7412	████	10.2
Paraphrase with T5	15164	6493	████	16.7
Paraphrase with fine-tuned T5	14913	6107	████	18.7

Can AI-Generated Text be Reliably Detected?



Text	# tokens	# green tokens	Accuracy	Perplexity
Original	19042	11078	97%	6.7
Paraphrase with PEGASUS	16773	7412	80%	10.2
Paraphrase with T5	15164	6493	64%	16.7
Paraphrase with fine-tuned T5	14913	6107	57%	18.7

Can AI-Generated Text be Reliably Detected?

Quillbot

Grammarly

AutoWriter

GoCopy

Copy.ai

DIPPER

T5

DeepL Translator

DeepL Write

[Web demo →](#)



Circumvention

Manual tricks

« write as if you were... »

Using active voice

Very specific data

Avoid quotes

Outline the structure

Write the beginning of the answer

Use other LLMs

Don't use English

Transform type	Example	Sample references
Any token replacement	<i>dog</i> → <i>!og</i>	[13, 20–22]
Adjacent keyboard typos	<i>achieve</i> → <i>axhieve</i>	[23, 24]
Common misspellings	<i>achieve</i> → <i>acheive</i>	[23]
Number to word form	<i>12</i> → <i>twelve</i>	[25]
Leetspeak conversions	<i>leet</i> → <i>l33t</i>	[26]
Vowel removal, letter shuffling, etc	<i>letters</i> → <i>ltetrs</i>	[27]
Phonetic replacements	<i>fair</i> → <i>fare</i>	[28, 29]
Homograph adjective manipulation	<i>short lead</i> → <i>hot lead</i>	[30]
Unicode-based replacements	<i>a</i> → <i>à</i>	[31]
Visual character mapping	<i>cl</i> → <i>d</i>	[32, 33]
Invisible characters	<i>abc</i> → <i>abc<U+200B></i>	[8]
Inflectional perturbations	<i>run</i> → <i>running</i>	[34]
Thesaurus-based synonyms	<i>happy</i> → <i>joyful</i>	[35–37]
Word to emoji	<i>love</i> → <i>♡</i>	[38]
Word embedding-based	<i>fast</i> → <i>swift</i>	[24, 39, 40]
Counter-fitted synonyms	<i>happy</i> → <i>content</i>	[6, 41]
Context-aware embeddings	<i>bank</i> → <i>credit union</i>	[42]
Masked language model infill	<i>go [MASK](here) fast</i> → <i>go very fast</i>	[6, 19, 42–45]
Phrase replacements	<i>quickly run</i> → <i>swiftly sprint</i>	[38, 45, 46]
Predefined parse template filling	<i>[Noun] runs [Adverb]</i> → <i>The dog runs quickly</i>	[46]
Derived rule-based	<i>What NOUN</i> → <i>Which NOUN</i>	[47]

Table 2

Examples of different transformation types, and some examples of methods where they were used.

Transform type	Example	Sample references
Any token replacement	<i>dog</i> → <i>!og</i>	[13, 20–22]
Adjacent keyboard typos	<i>achieve</i> → <i>axhieve</i>	[23, 24]
Common misspellings	<i>achieve</i> → <i>acheive</i>	[23]
Number to word form	<i>12</i> → <i>twelve</i>	[25]
Leetspeak conversions	<i>leet</i> → <i>l33t</i>	[26]
Vowel removal, letter shuffling, etc	<i>letters</i> → <i>ltetrs</i>	[27]
Phonetic replacements	<i>fair</i> → <i>fare</i>	[28, 29]
Homograph adjective manipulation	<i>short lead</i> → <i>hot lead</i>	[30]
Unicode-based replacements	<i>a</i> → <i>à</i>	[31]
Visual character mapping	<i>cl</i> → <i>d</i>	[32, 33]
Invisible characters	<i>abc</i> → <i>abc<U+200B></i>	[8]
Inflectional perturbations	<i>run</i> → <i>running</i>	[34]
Thesaurus-based synonyms	<i>happy</i> → <i>joyful</i>	[35–37]
Word to emoji	<i>love</i> → <i>♡</i>	[38]
Word embedding-based	<i>fast</i> → <i>swift</i>	[24, 39, 40]
Counter-fitted synonyms	<i>happy</i> → <i>content</i>	[6, 41]
Context-aware embeddings	<i>bank</i> → <i>credit union</i>	[42]
Masked language model infill	<i>go [MASK](here) fast</i> → <i>go very fast</i>	[6, 19, 42–45]
Phrase replacements	<i>quickly run</i> → <i>swiftly sprint</i>	[38, 45, 46]
Predefined parse template filling	<i>[Noun] runs [Adverb]</i> → <i>The dog runs quickly</i>	[46]
Derived rule-based	<i>What NOUN</i> → <i>Which NOUN</i>	[47]

Table 2

Examples of different transformation types, and some examples of methods where they were used.

50 first words

Undetectable

[Web demo →](#)



Circumvention

Gonna get harder over time



as these models improve over time, the generated text looks increasingly similar to human text. [...]

even the most effective detector [is ineffective] when dealing with a sufficiently advanced language model.

[Can AI-Generated Text be Reliably Detected?](#)

You can detect *your* texts

Really hard, gonna be an issue in the future



<https://forms.gle/39ULnvq3rSBpmN538>



Further reading...

- [Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense](#)
- [Can AI-Generated Text be Reliably Detected?](#)
- [A Watermark for Large Language Models](#)
- [DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature](#)
- [New AI classifier for indicating AI-written text](#)
- [Ghostbuster: Detecting Text Ghostwritten by Large Language Models](#)
- [Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text](#)

