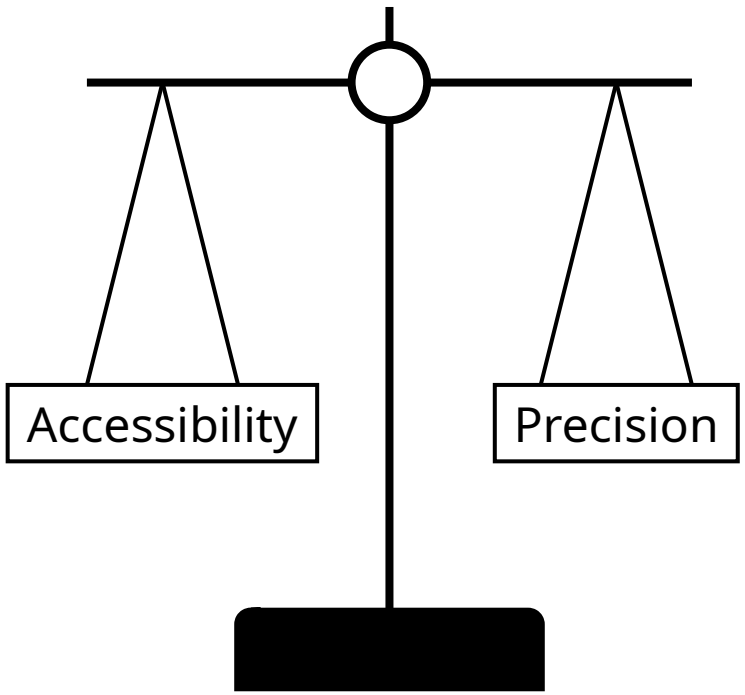
Homoglyph-Based Attacks: Circumventing LLM Detectors

Aldan Creo

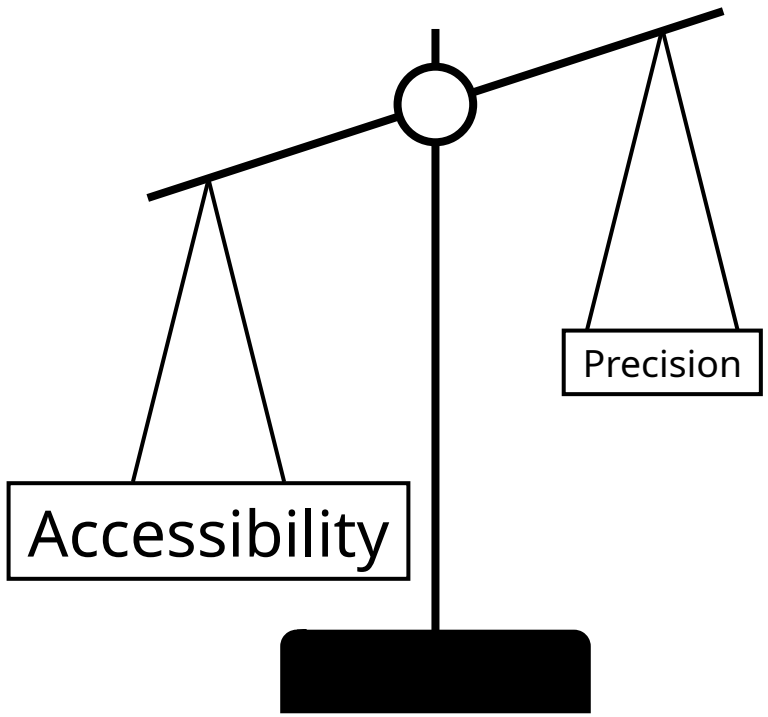
Technology Research Specialist

Some notes...

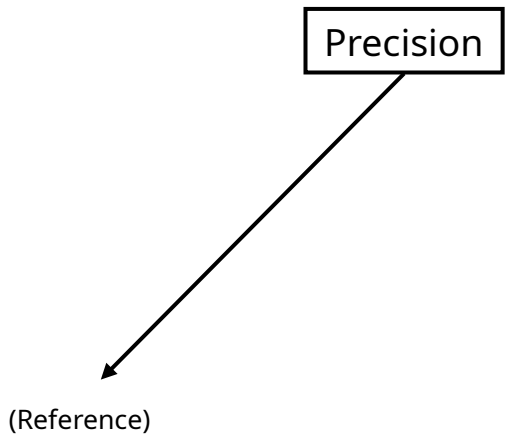
1



1



2



3

Takeaway

Bingo cards – explanation later

Let’s get
started!

Who am I?


development
science
research
ai-generated
european
software
graph
knowledge
chatgpt
llm
business
fundamentals
artificial intelligence
entrepreneurship
publishing
academic
event
paper
model
bootcamp
texts
global
summarization
award
international
microsoft
founder
young
maker
data
search
students
conference
google
python
website
excellence
foundation
first
best
exchange
digital
europe
youth
code
valedictorian
matlab
parliament
leaders
member
present
founded
student
devfest
summer
scholarship
program
detection
leadership
new
countries
madrid
summit
local
faire
web
club
legal
organized
catch
developed
djangocon
developers
represent
internal
university
dialogue
society
seminar
documents
graphs
innovation
association
usage
world
selected
president
contributed
intern
technologies
without
collaboration
scholarships
school
intelligent
language
distributed
management
recipient
forum
education
learning
performance
computer
negotiation
technology
teams
like
spearheaded
essentials
organizer
chatbot
hacking
council
events
group
azure
human
kindle
platforms
models
getting
obtained
managed
project
global
hackathon
hackupc
specialist
processing
embeddings
fully
caught
prize
speed
raising
scientific
features
webmaster
british
manager
contest
funding
akademia
communications
online
session
led
top
iris
centre
course
utilized
laudable
rights
implemented
federalist
university
statutes
organizing
world
sorbonne
université
iaprendizaxe
create
data
applied
teo
grants
electronic
public
natural
highest
repair
day
system
replicant

research

The 31st International
Conference on Computational
Linguistics

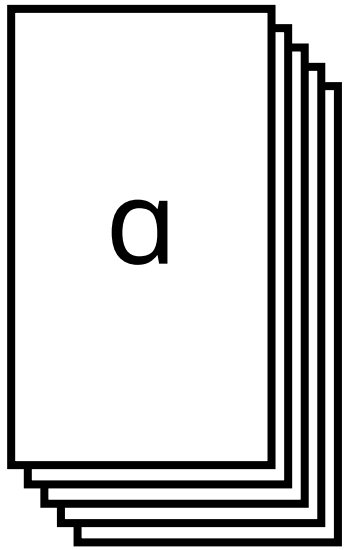
Evading AI-Generated Content Detectors using Homoglyphs

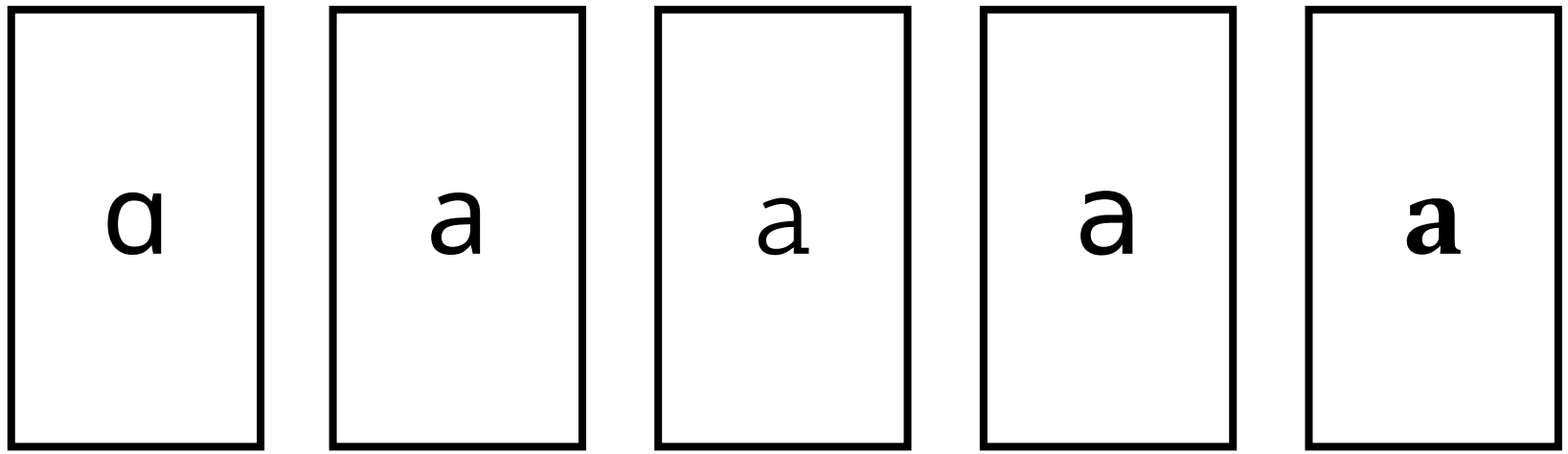
Aldan Creo, Shushanta Pudasaini

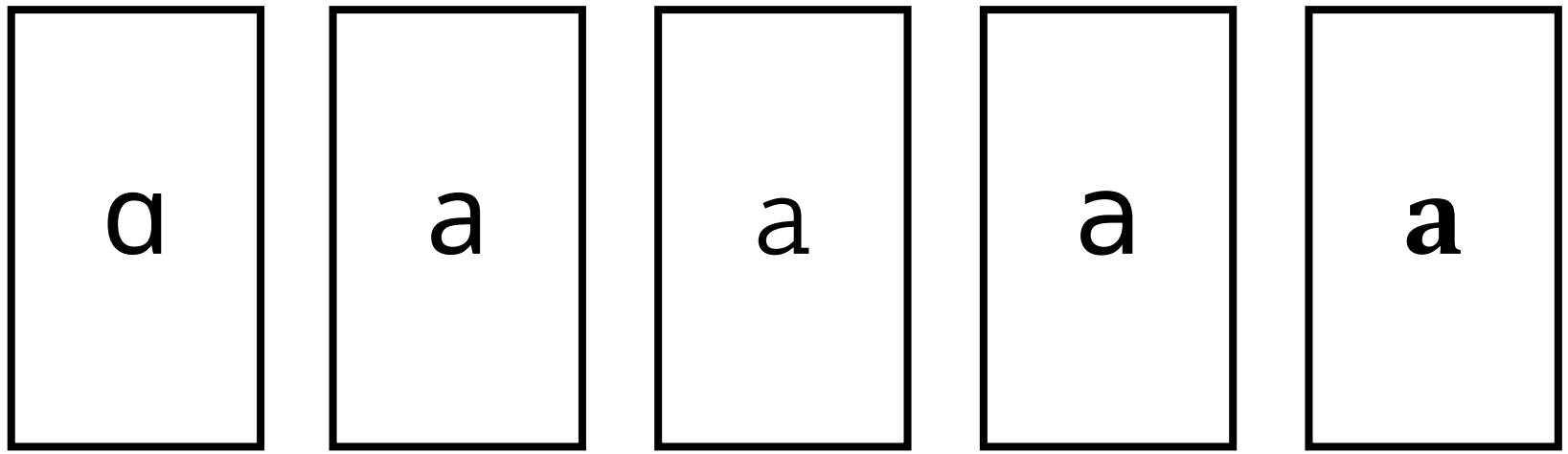
Structure

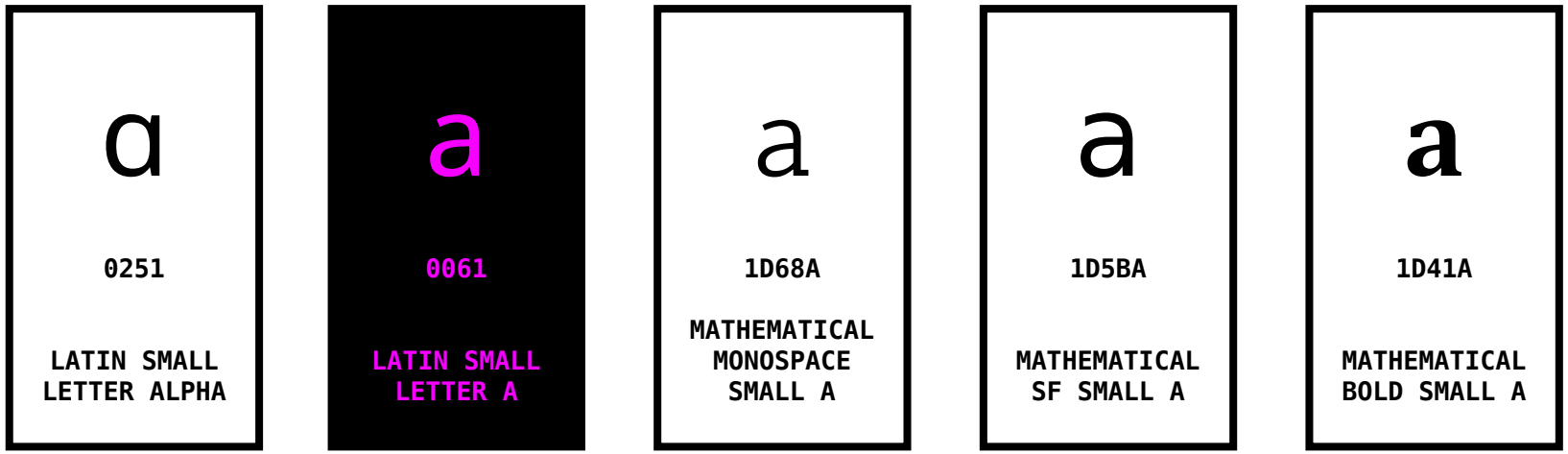
What are homoglyphs?	AI-generated text detectors	Homoglyph-based attacks
Technical analysis	Effectiveness	Implications

What are homoglyphs?


α


ɑ
a
a
a
a


ɑ
a
a
a
a

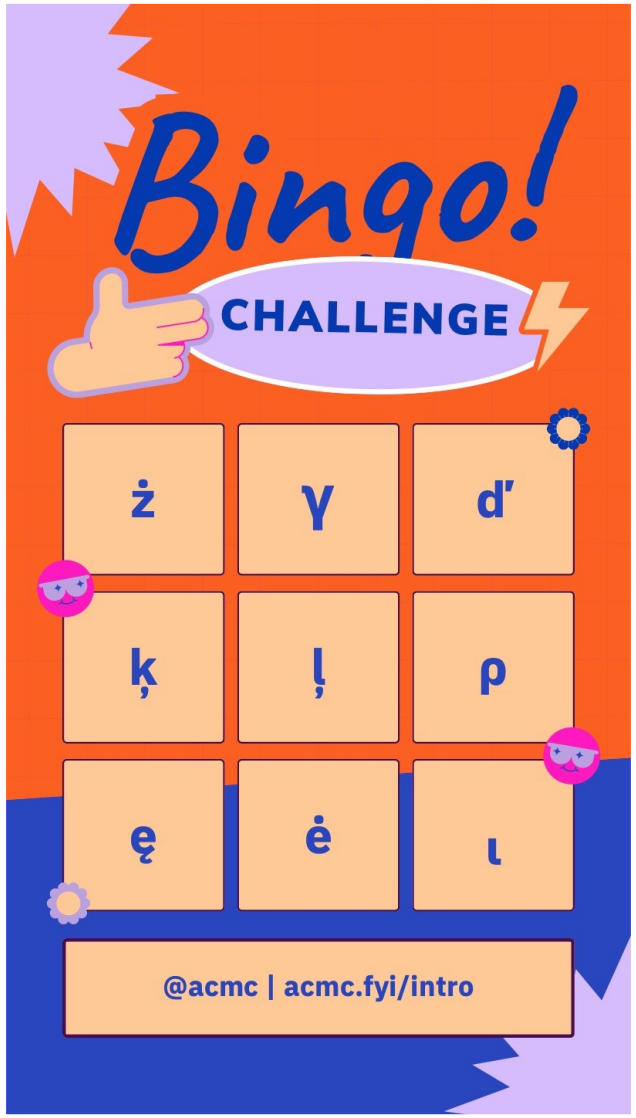

ɑ
0251
LATIN SMALL LETTER ALPHA
a
0061
LATIN SMALL LETTER A
a
1D68A
MATHEMATICAL MONOSPACE SMALL A
a
1D5BA
MATHEMATICAL SF SMALL A
a
1D41A
MATHEMATICAL BOLD SMALL A

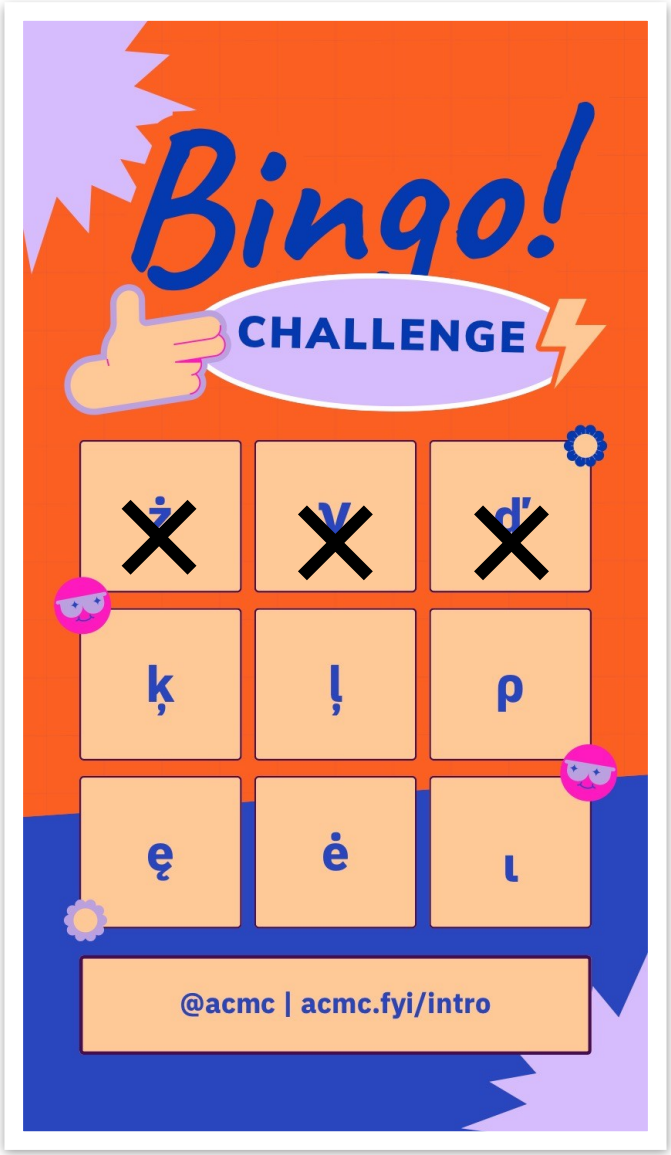
Homoglyphs

Characters that seem identical
but have different meanings

Homoglyphs

Characters that seem identical
but have different encodings


Bingo!
CHALLENGE
ż
ɣ
ď
ķ
ļ
ƿ
ę
ė
ɩ
@acmc | acmc.fyi/intro


Bingo!
CHALLENGE
ż
v
d'
ķ
ļ
ρ
ę
ė
ι
@acmc | acmc.fyi/intro

⚠ not all homoglyphs are in the slides!

Past applications

URL attacks

URL attacks

Click here to reset your password: https://paypal.com

(Boucher et al. May 2022)

URL attacks

Click here to reset your password: https://xn--paypa-n6a.com/

(Boucher et al. May 2022)

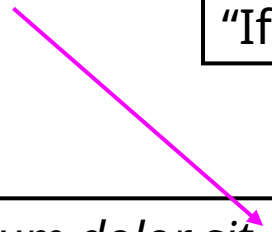
Hiding information

(Boucher et al. May 2022)

Steganography

Hiding data in pɩain sight
to avoid detection

“If the first ‘a’ is an ‘α’, trigger the exploit”



“If the first ‘a’ is an ‘ɑ’, trigger the exploit”

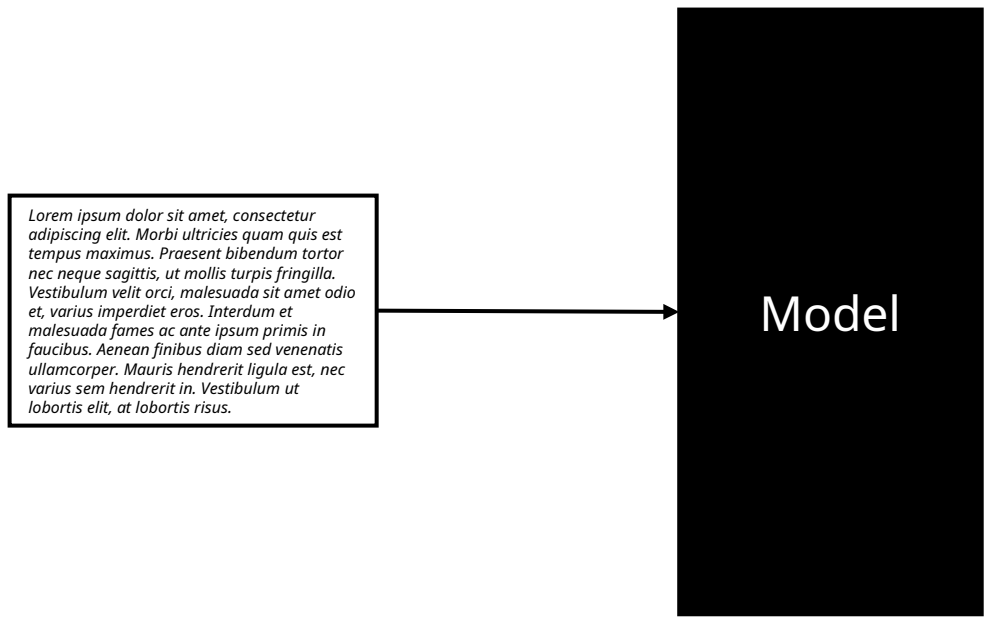
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi ultricies quam quis est tempus maximus. Praesent bibendum tortor nec neque sagittis, ut mollis turpis fringilla. Vestibulum velit orci, malesuada sit amet odio et, varius imperdiet eros. Interdum et malesuada fames ac ante ipsum primis in faucibus. Aenean finibus diam sed venenatis ullamcorper. Mauris hendrerit ligula est, nec varius sem hendrerit in. Vestibulum ut lobortis elit, at lobortis risus.

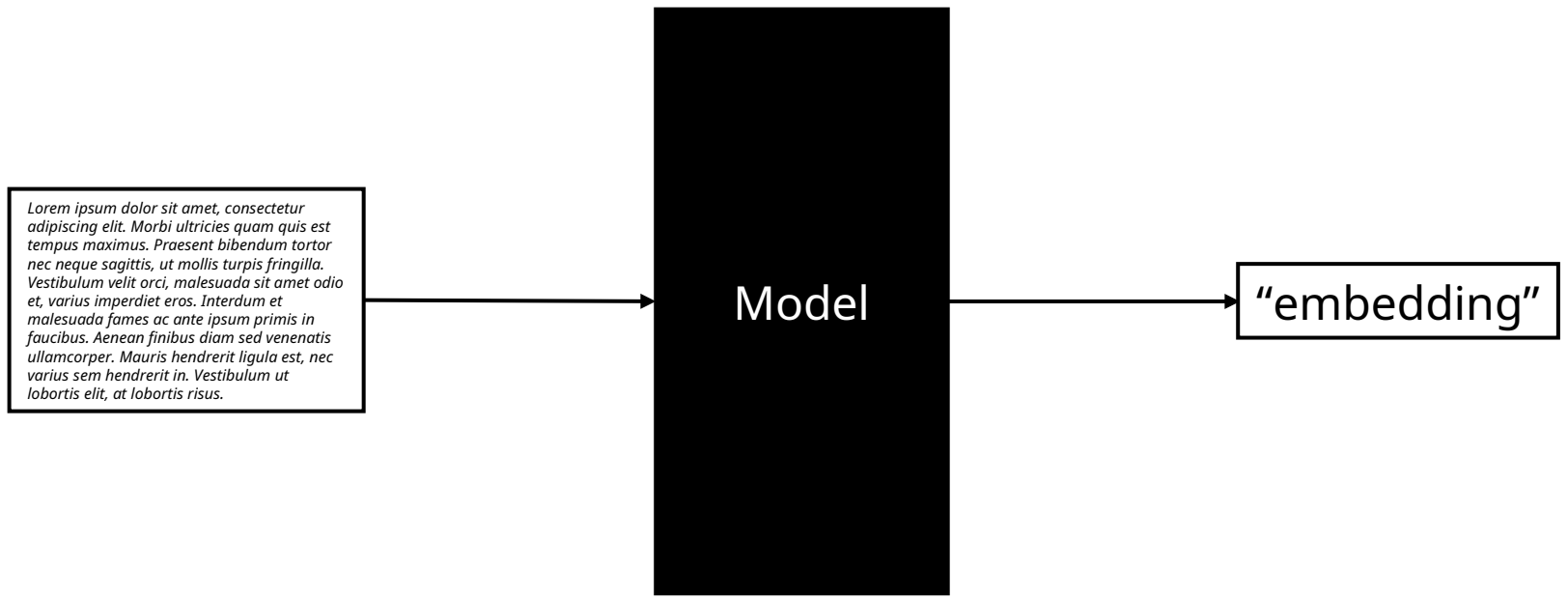
Homoglyphs: characters we can’t discern

AI-generated text detectors

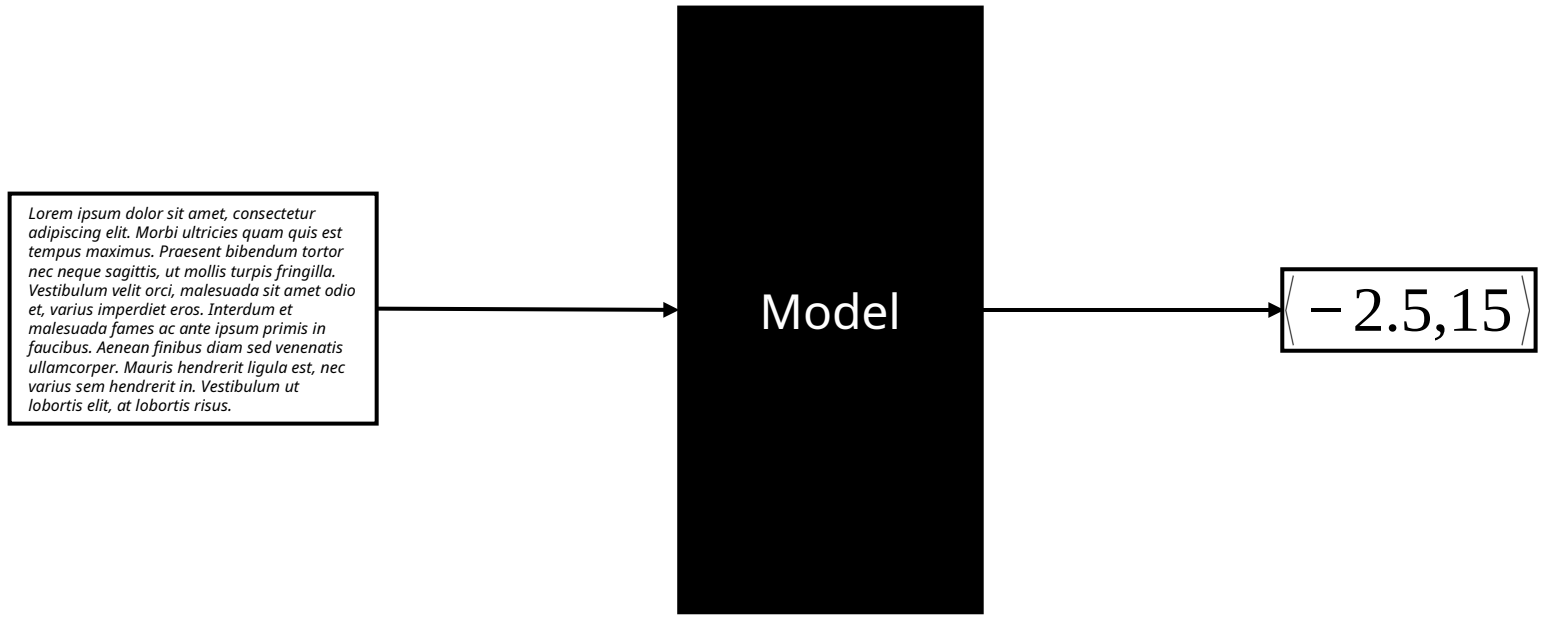
Classifiers

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi ultricies quam quis est tempus maximus. Praesent bibendum tortor nec neque sagittis, ut mollis turpis fringilla. Vestibulum velit orci, malesuada sit amet odio et, varius imperdiet eros. Interdum et malesuada fames ac ante ipsum primis in faucibus. Aenean finibus diam sed venenatis ullamcorper. Mauris hendrerit ligula est, nec varius sem hendrerit in. Vestibulum ut lobortis elit, at lobortis risus.

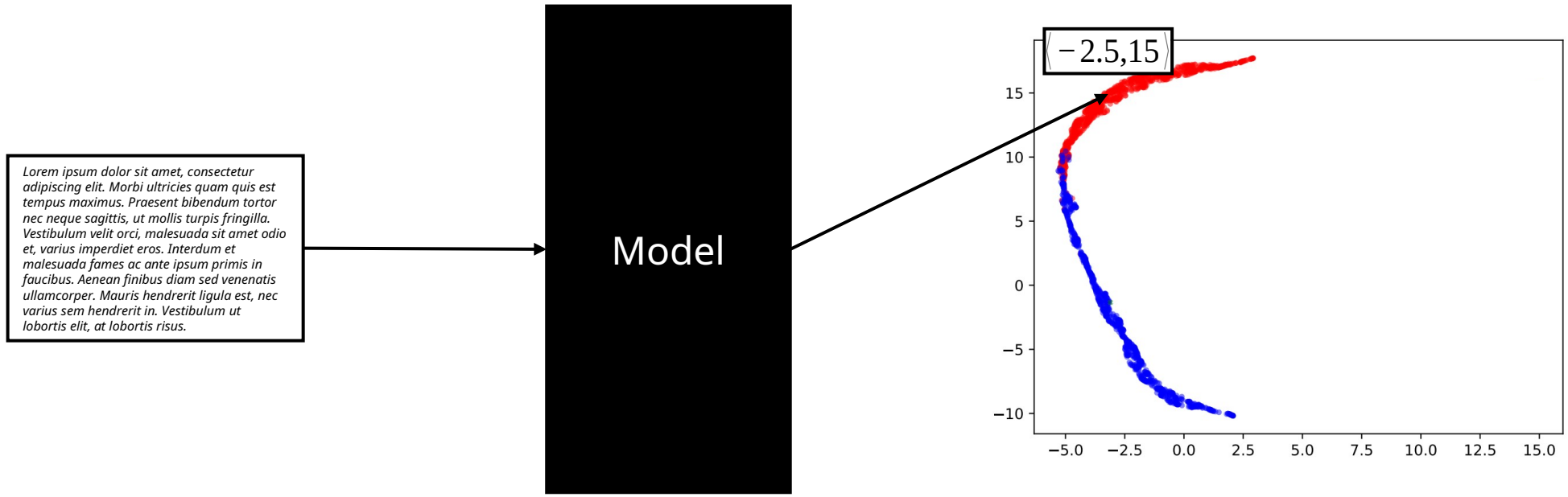

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi ultricies quam quis est tempus maximus. Praesent bibendum tortor nec neque sagittis, ut mollis turpis fringilla. Vestibulum velit orci, malesuada sit amet odio et, varius imperdiet eros. Interdum et malesuada fames ac ante ipsum primis in faucibus. Aenean finibus diam sed venenatis ullamcorper. Mauris hendrerit ligula est, nec varius sem hendrerit in. Vestibulum ut lobortis elit, at lobortis risus.
Model



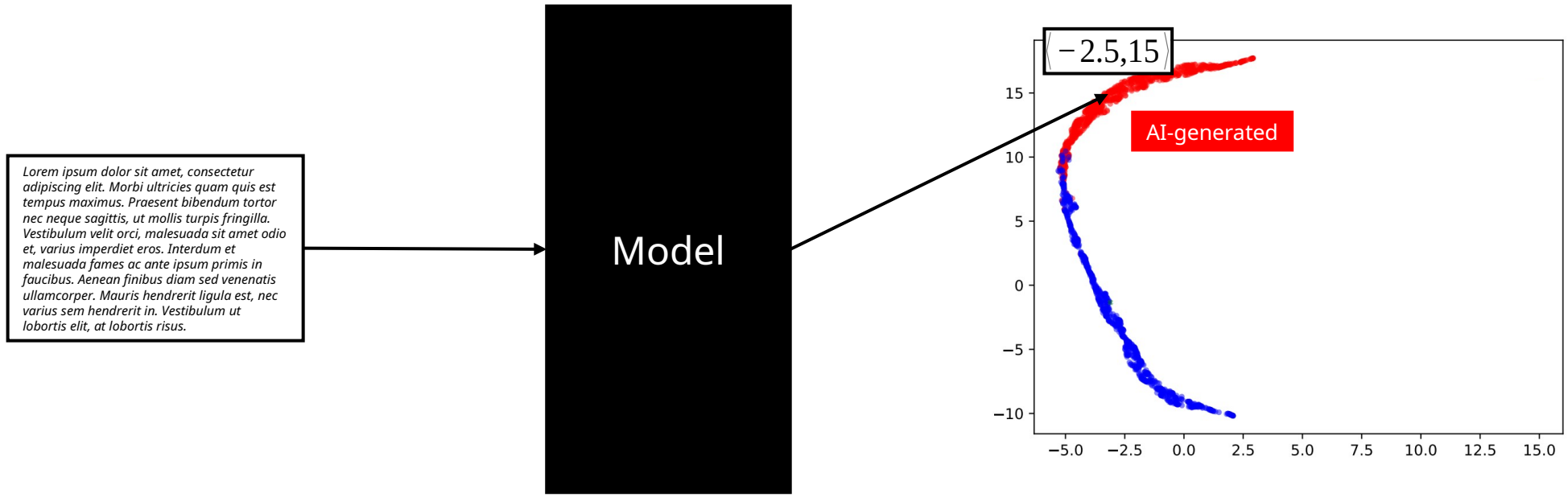
(Creo, Pudasaini 2024)



(Creo, Pudasaini 2024)



(Creo, Pudasaini 2024)

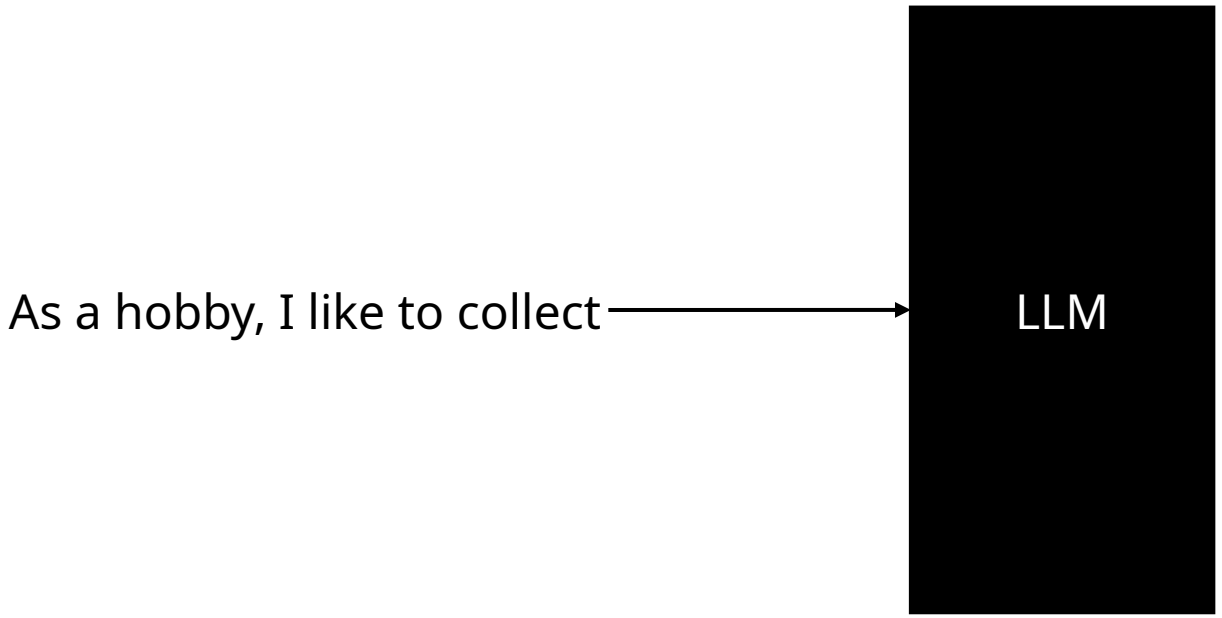


(Creo, Pudasaini 2024)

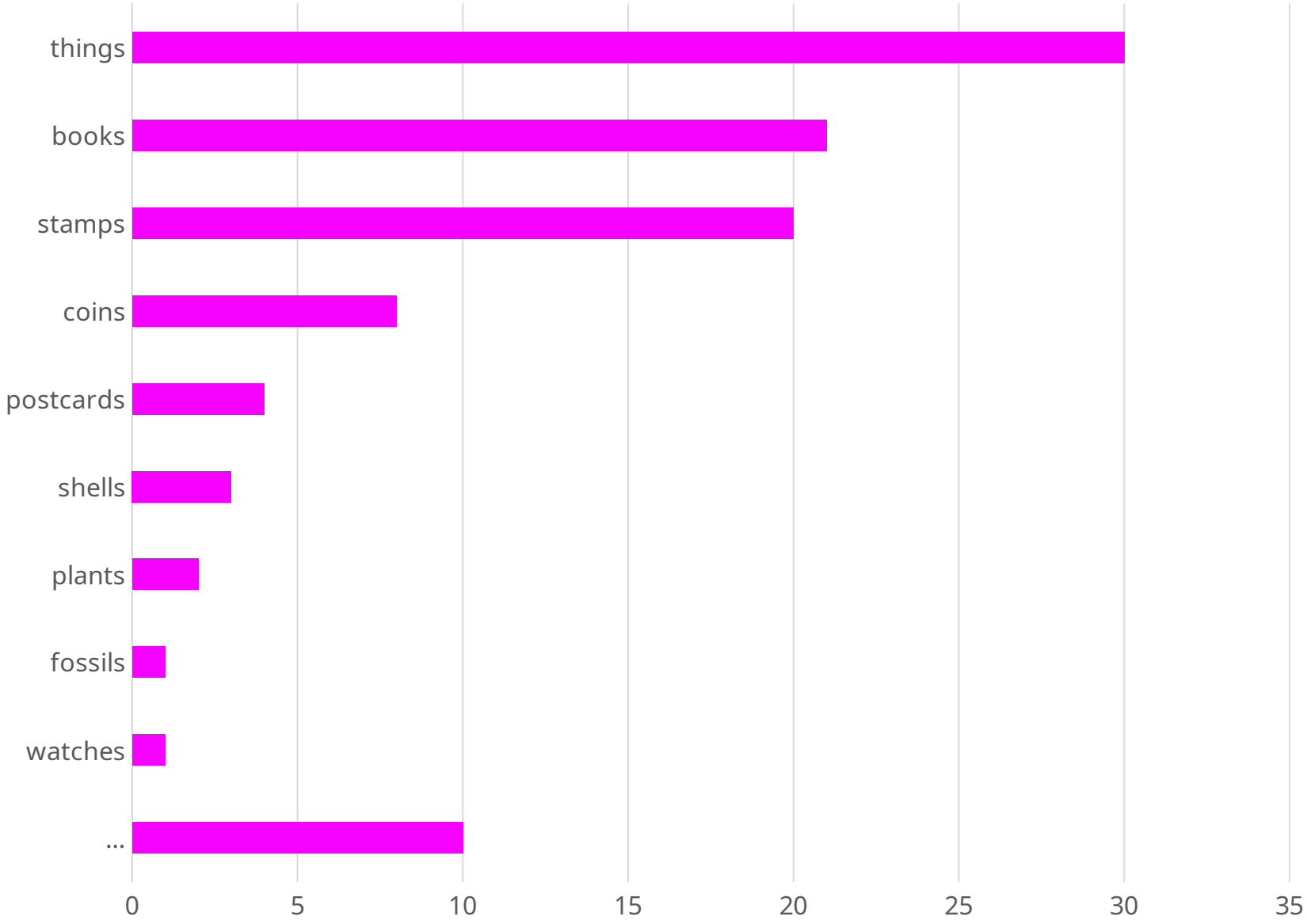
Perplėxity-based detectors

How do LLMs work?

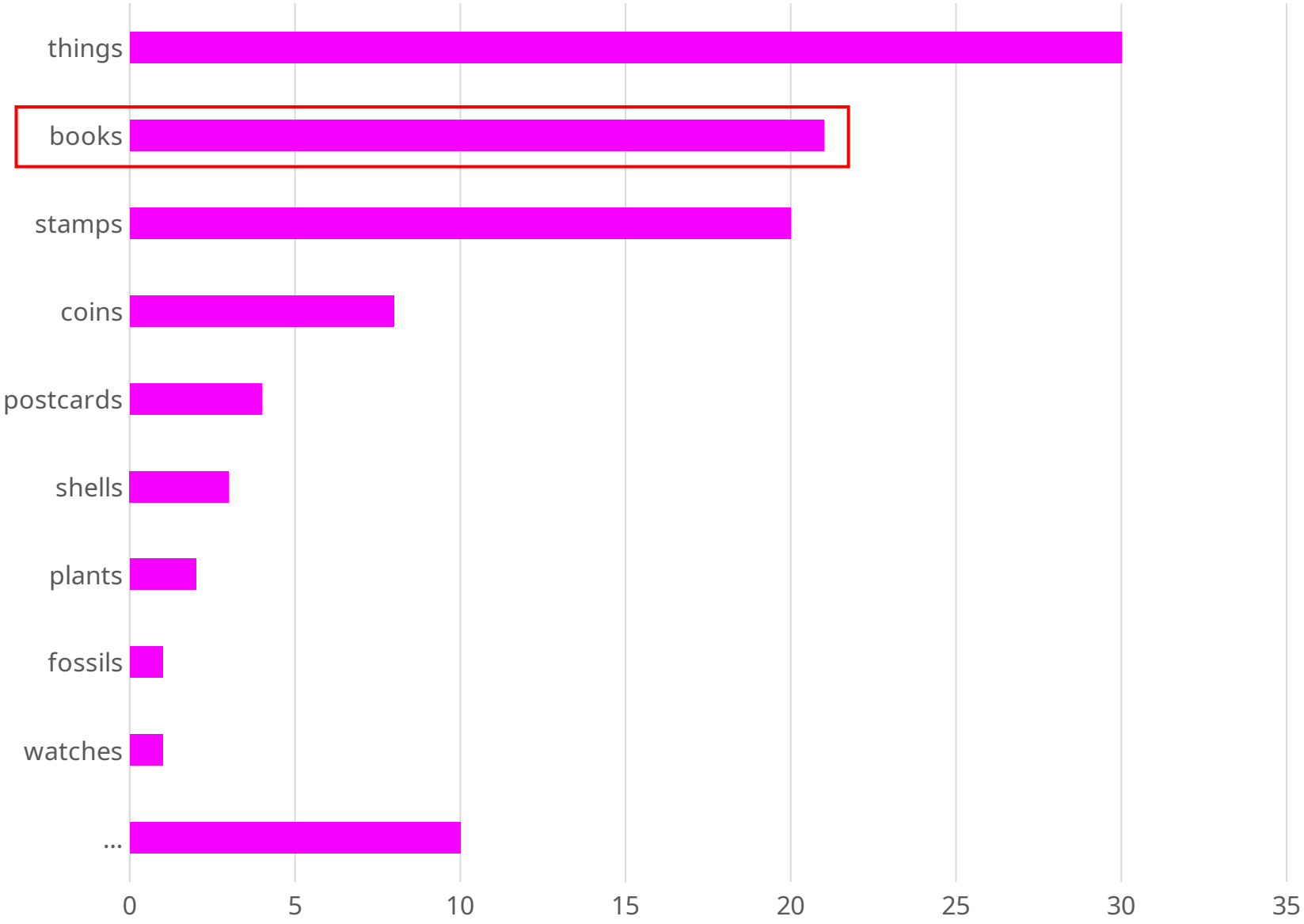
As a hobby, I like to collect _____


As a hobby, I like to collect
LLM

As a hobby, I like to collect



As a hobby, I like to collect



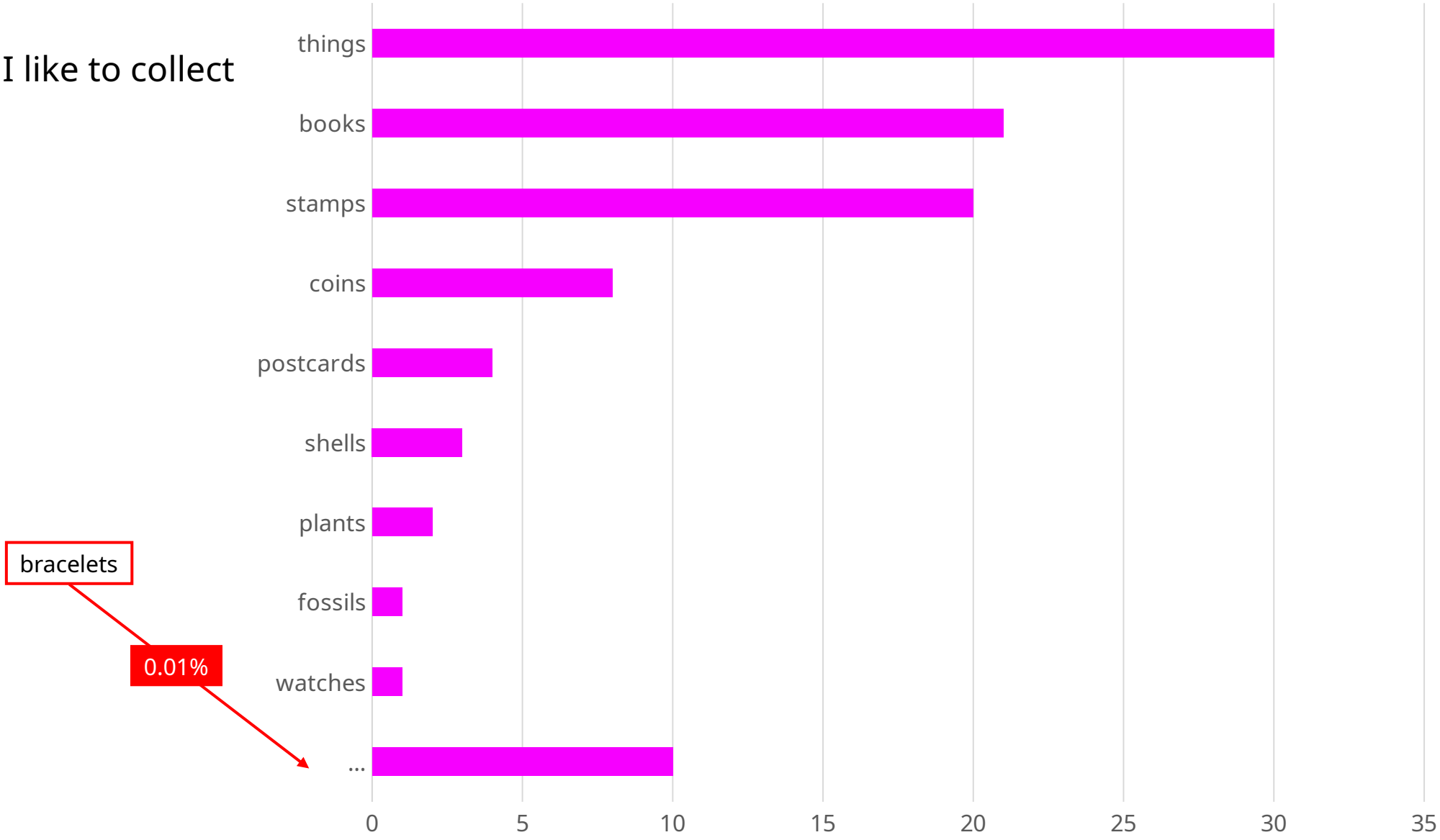
As a hobby, I like to collectbooks

AI-generated

As a hobby, I like to collectbracelets

Human-written

As a hobby, I like to collect



As a hobby, I like to collect

books

bracelets

↑ probability

↓ probability

As a hobby, I like to collect

books ↑ probability ↓ perplexity

bracelets ↓ probability ↑ perplexity

As a hobby, I like to collect

books

↑ probability ↓ perplexity AI-generated

bracelets

↓ probability ↑ perplexity human-written

Repeat for the whole text 👉
prediction

Watermarks

OpenAI has built a text watermarking method to detect ChatGPT-written content

COMPANY HAS MULLED ITS RELEASE OVER THE PAST YEAR

OpenAI has the tech to watermark ChatGPT text—it just won't release it

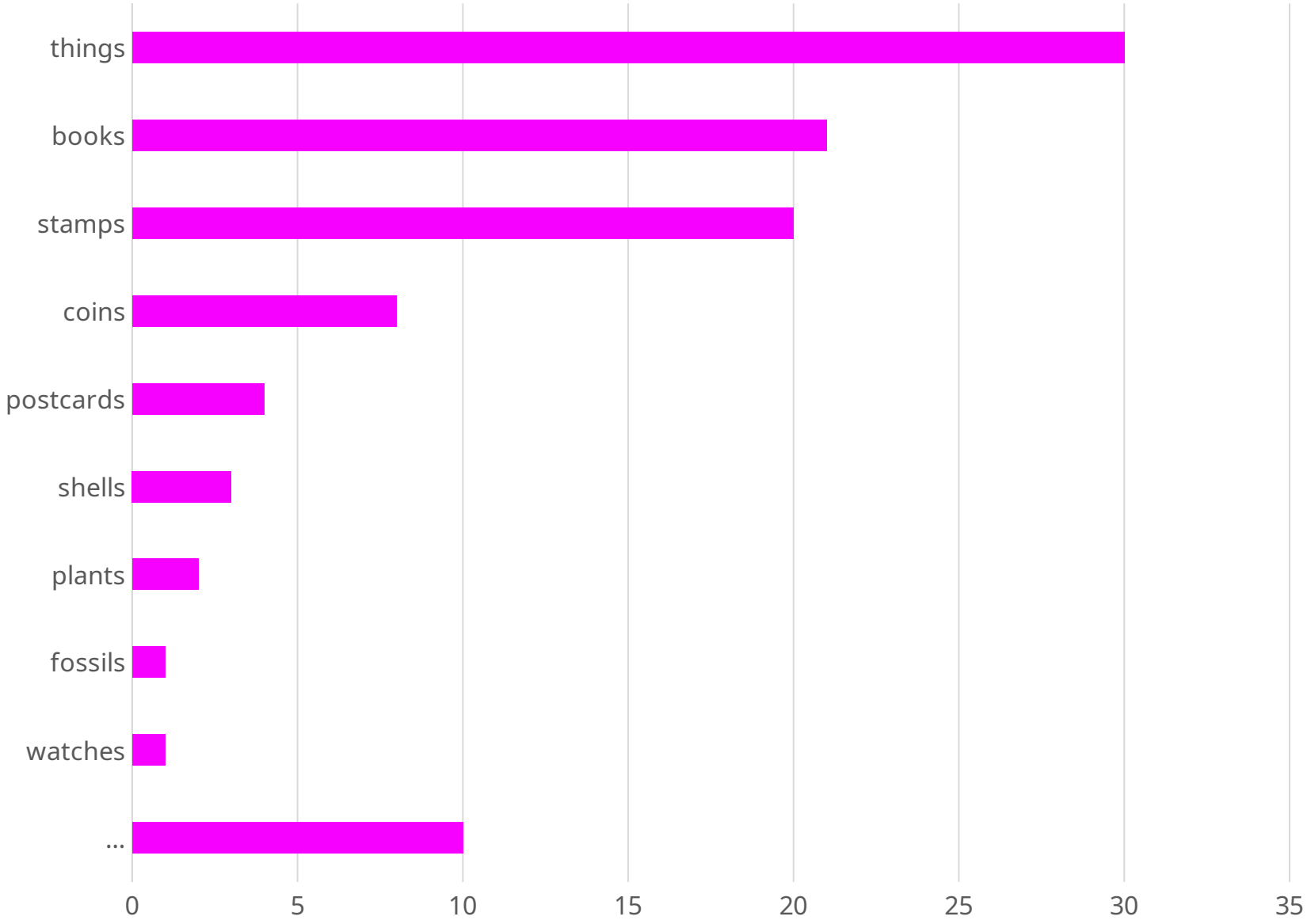
some say watermarking is the responsible thing to do...

OpenAI sitting on tool to watermark AI-generated content

The company has yet to release it, because it fears losing users

(Jowi Morales 2024, Samuel Axon 2024, Jacob Wulff Wold 2024)

As a hobby, I like to collect



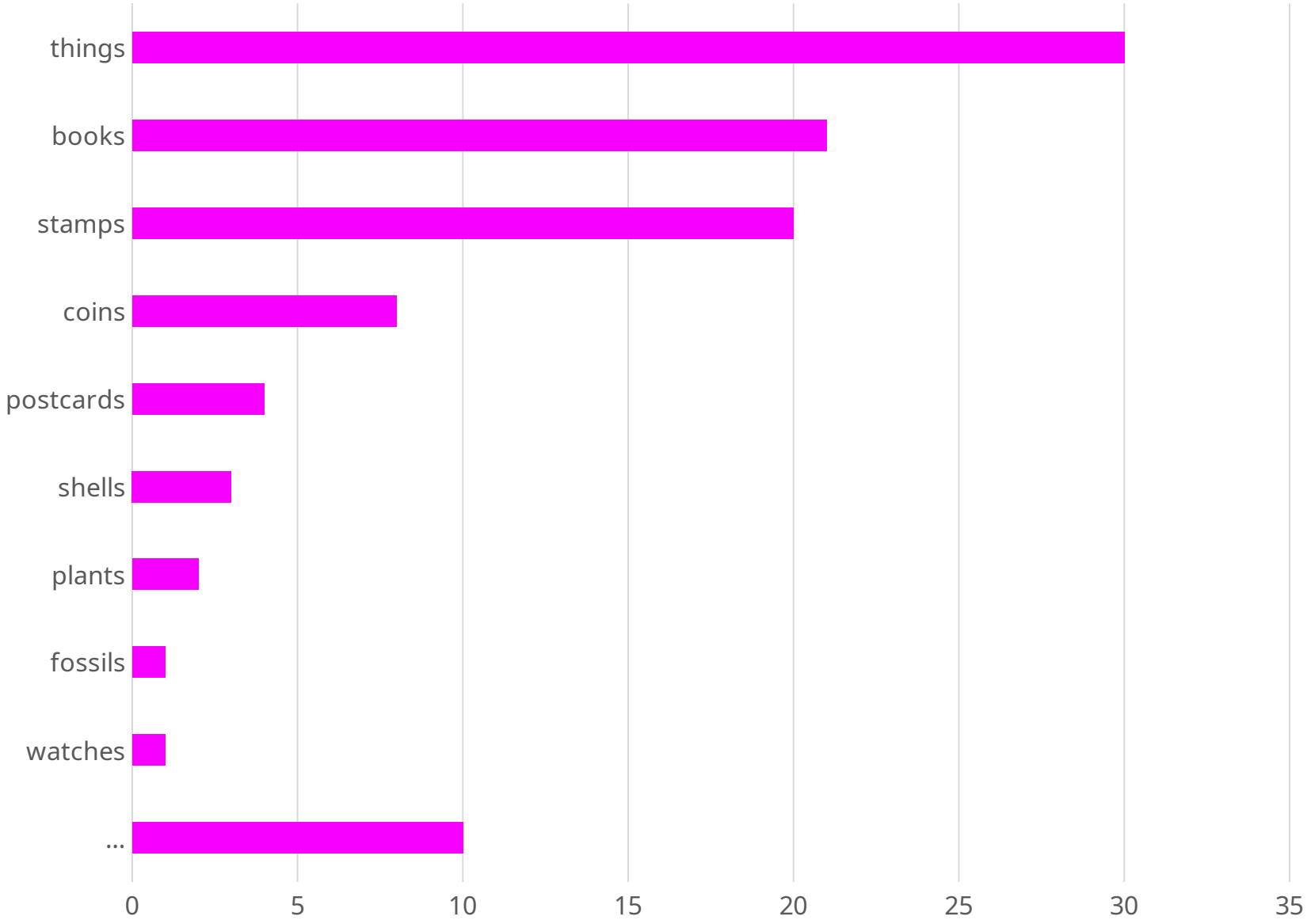
“Pick the N-th word. N is the number of letters in the previous word”

“Pick the N-th word. N is the number of letters in the previous word”

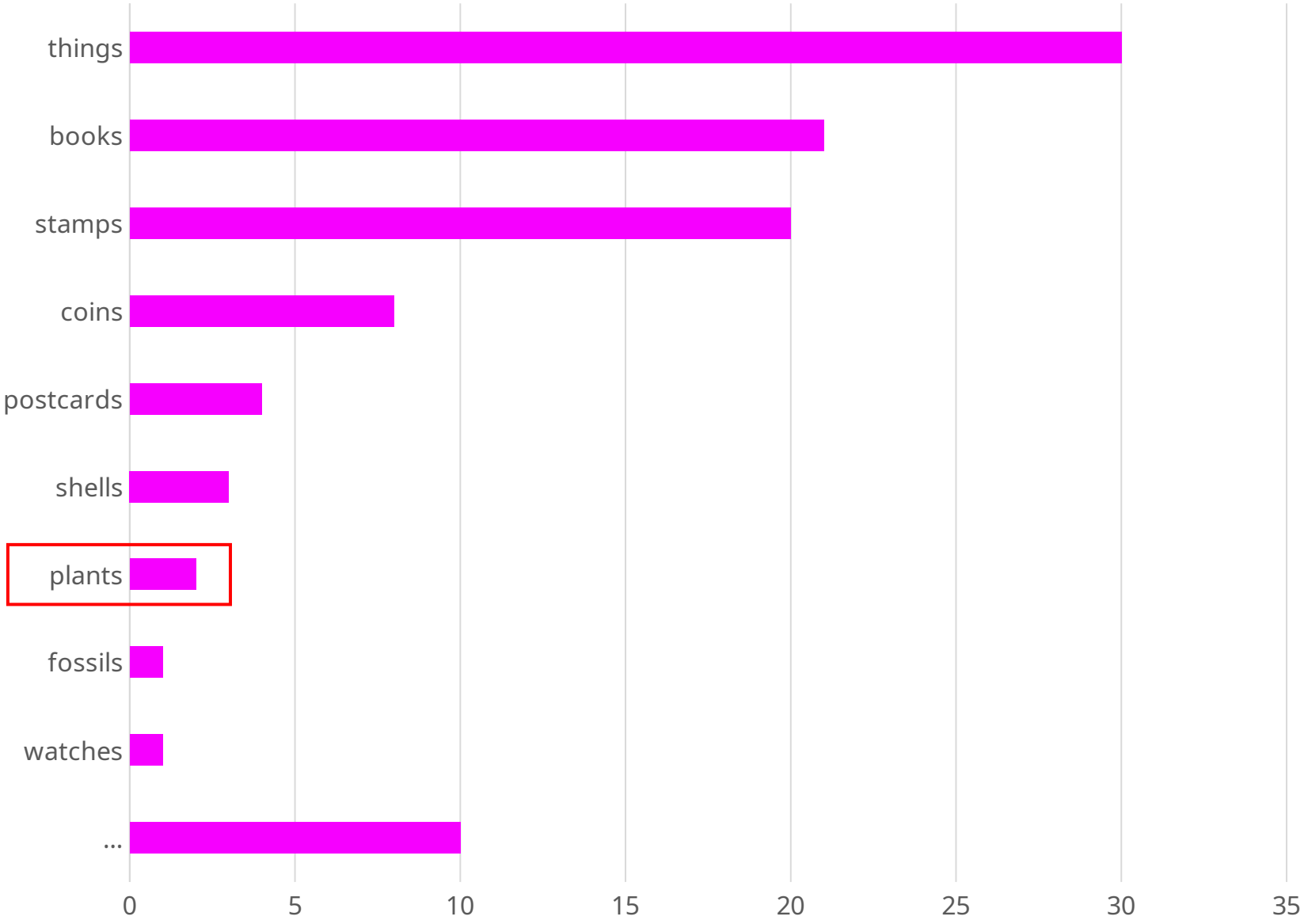
As a hobby, I like to collect

7 letters

As a hobby, I like to collect



As a hobby, I like to collect



As a hobby, I like to collectplants

Watermarked

Repeat for all of the words that we write

If a lot of words follow the rule 👉 AI-generated

* side
note

Detection of AI-generated texts: different techniques

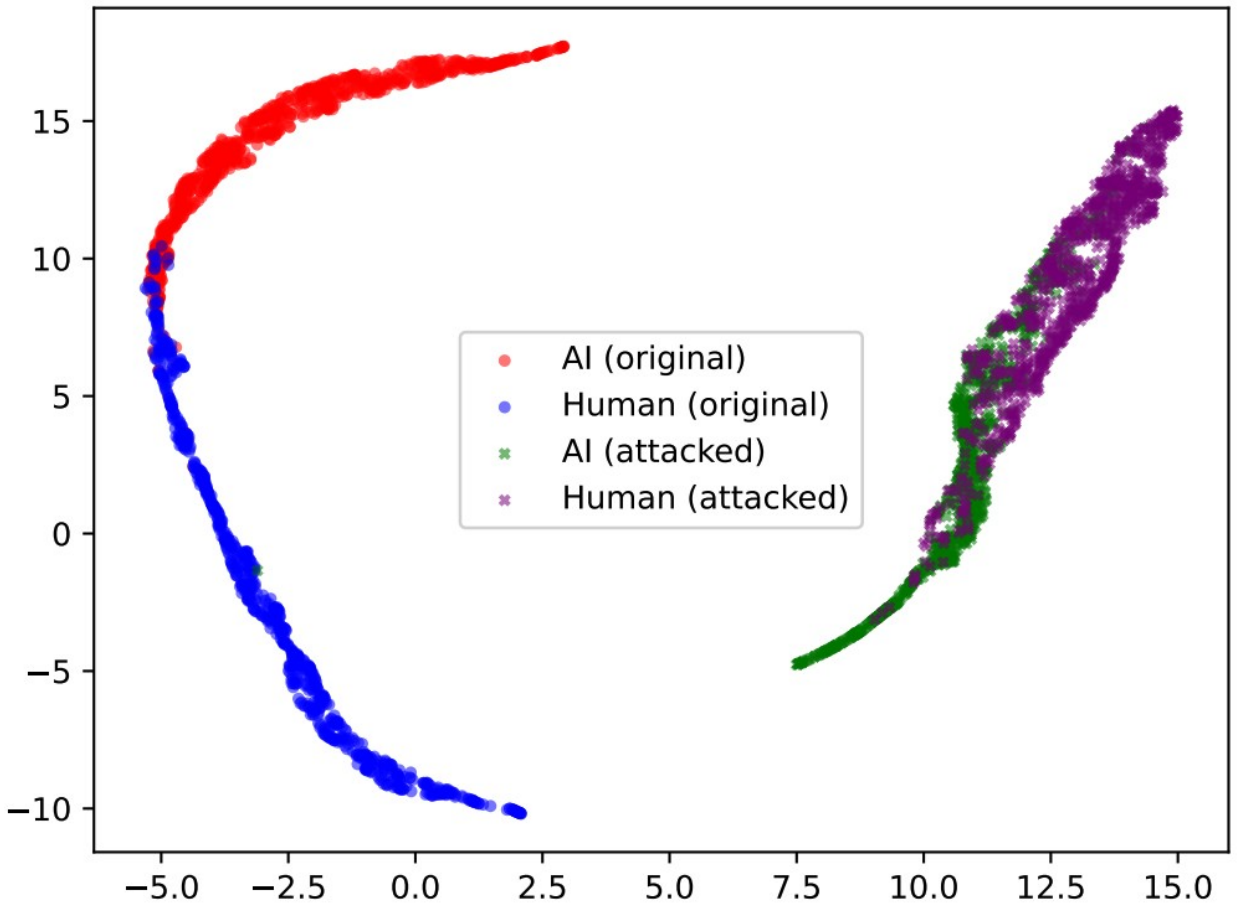
Homoglyph-based attacks

Demo

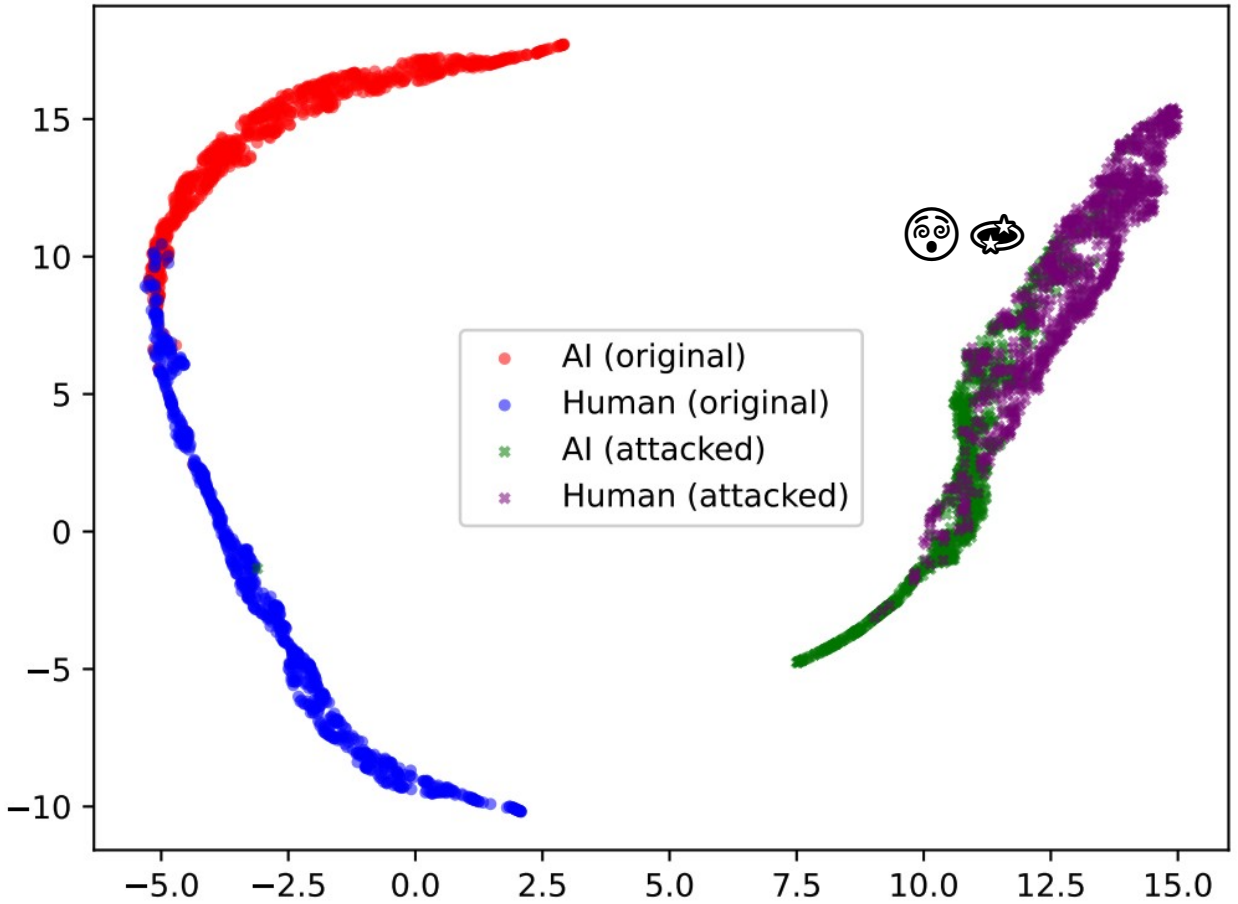
Homoglyph-based attacks: text becomes unrecognizable

Technical analysis

Classifiers



(Creo, Pudasaini 2024)



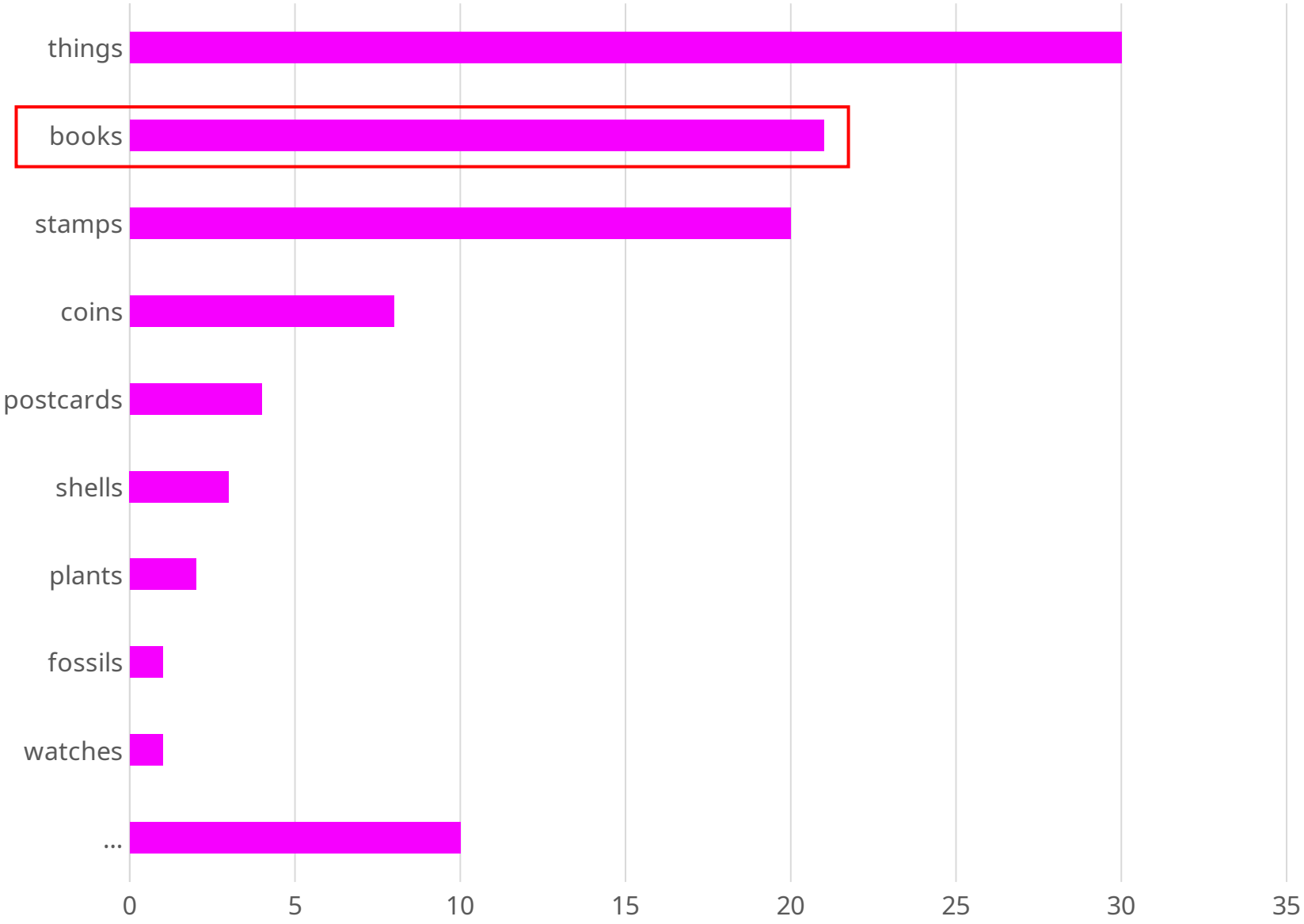
(Creo, Pudasaini 2024)

Perplexity-based detectors

As a hobby, I like to collectbooks

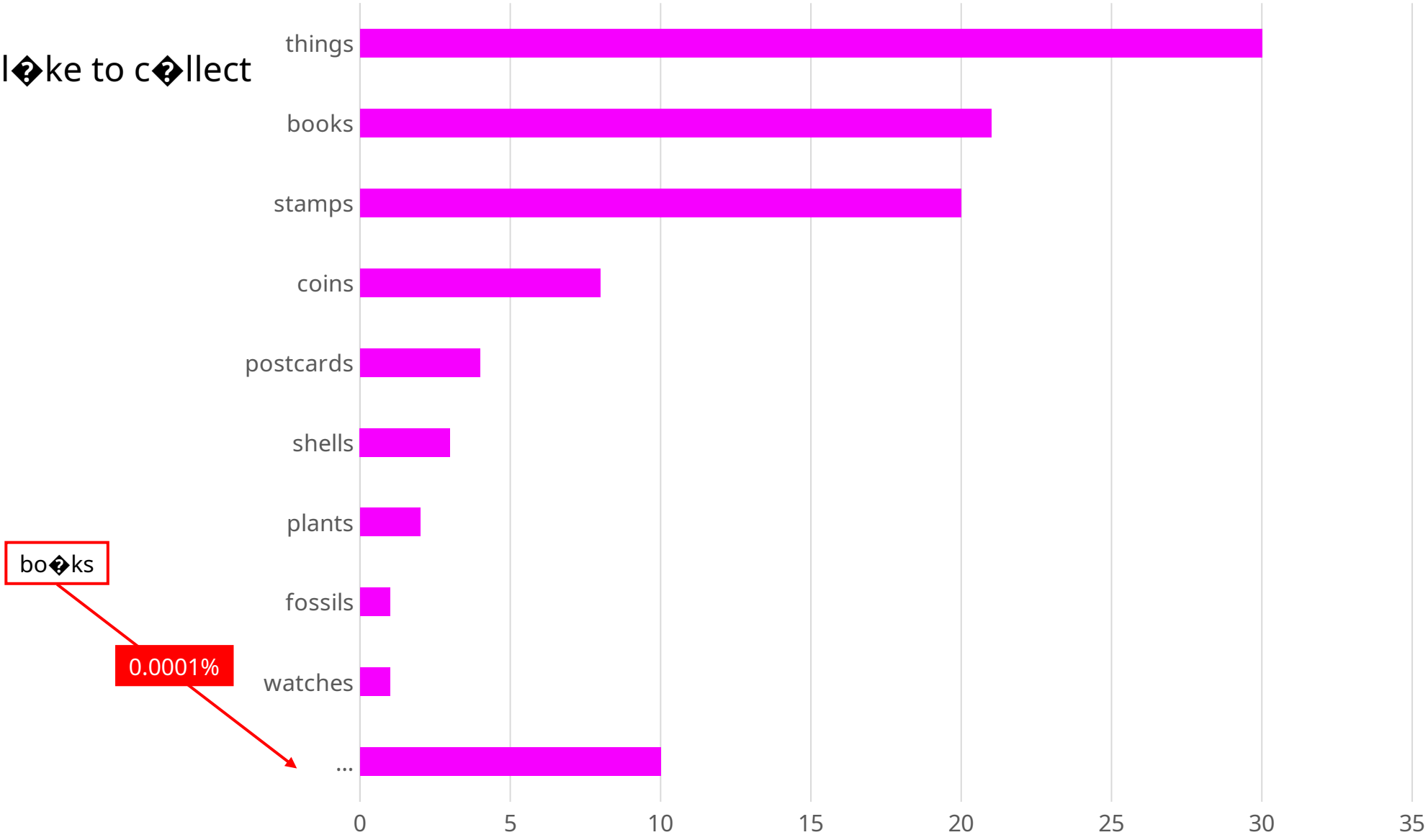
AI-generated

As a hobby, I like to collect



As a hobb�, I l�ke to c�llect b�oks

As a hobb�, I l�ke to c�llect



As a hobb�, I l�ke to c�llect b�oks

Human-written

Watermarks

As a hobby, I like to collect plants

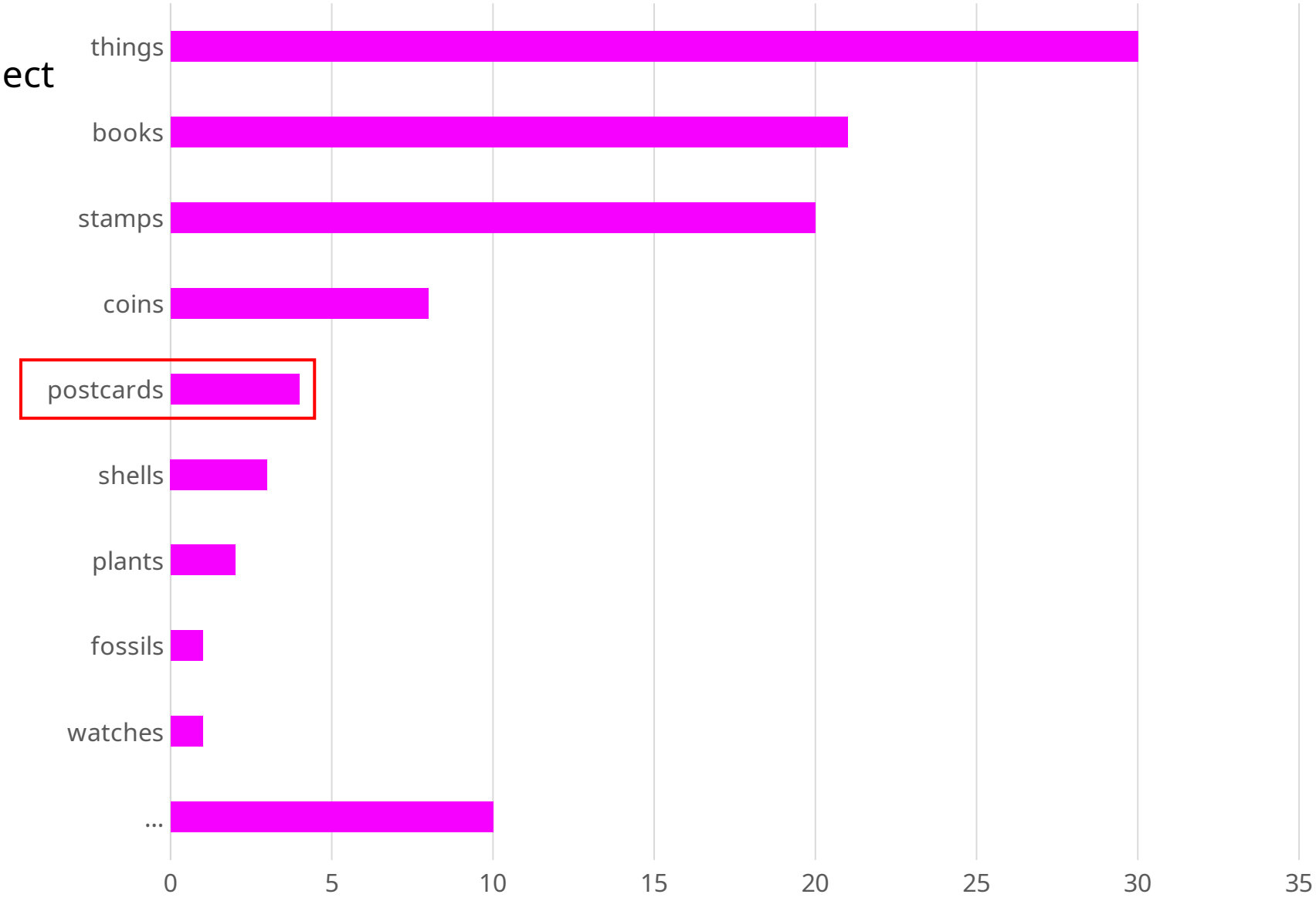
As a hobb�, I l�ke to c�llect plants

“Pick the N-th word. N is the number of letters in the previous word.”

As a hobb�, I l�ke to c�llect plants

5

As a hobb�, I l�ke to c�llect

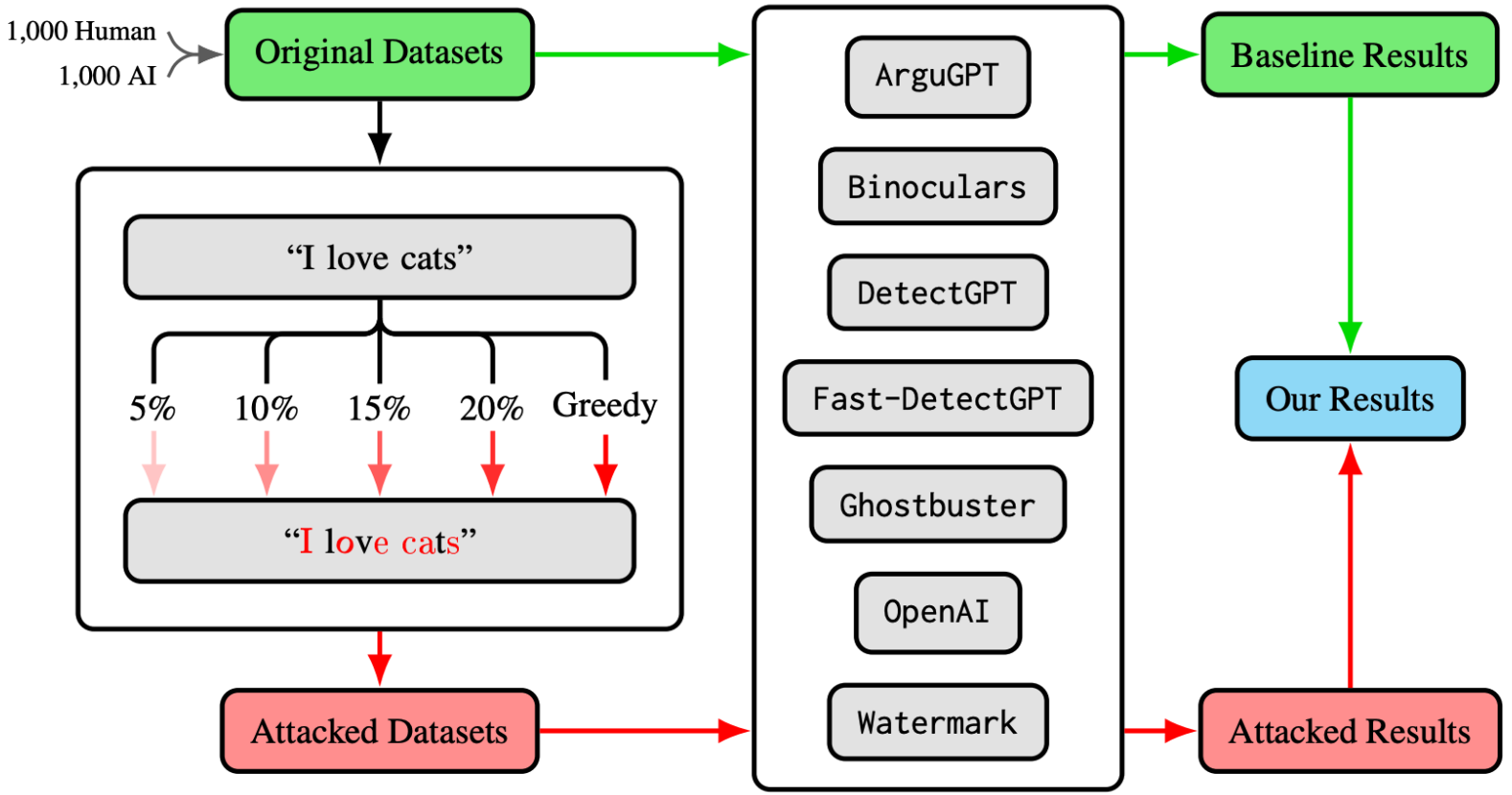


Doesn’t follow the rule 👉 human-written

Different mechanisms of action – all exploit “confusion”

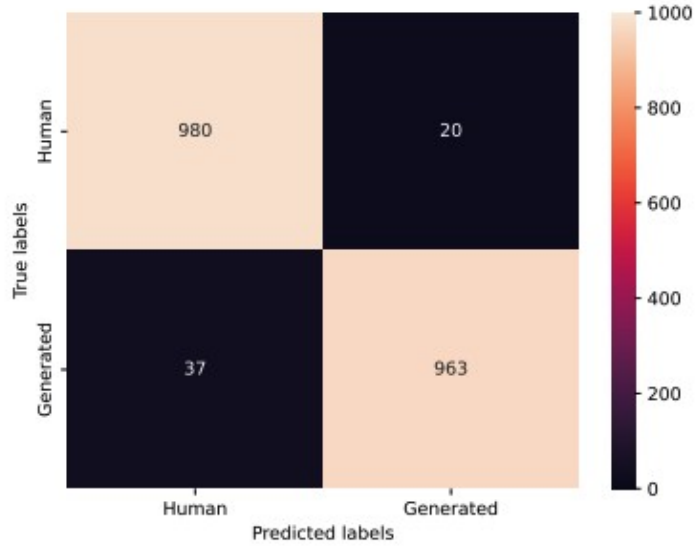
Effectiveness

Experimental approach



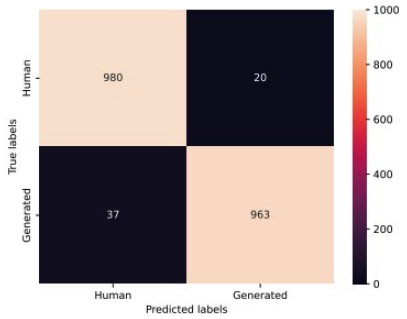
(Creo, Pudasaini 2024)

Results

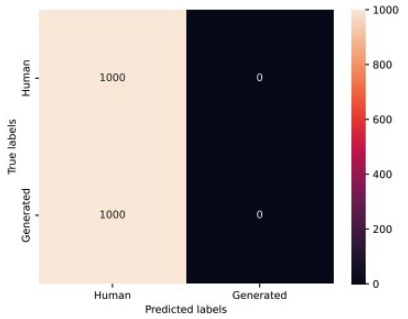


(a) No attack

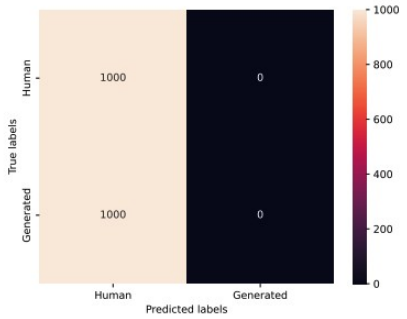
(Creo, Pudasaini 2024)



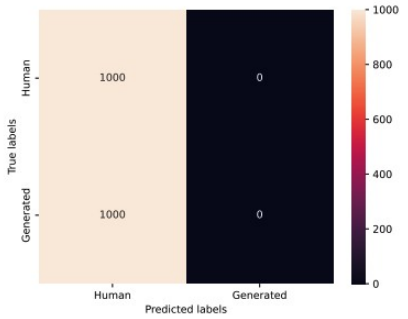
(a) No attack



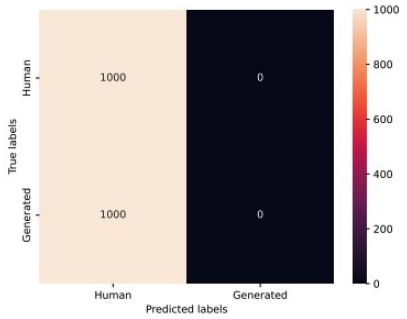
(b) Random attack (5%)



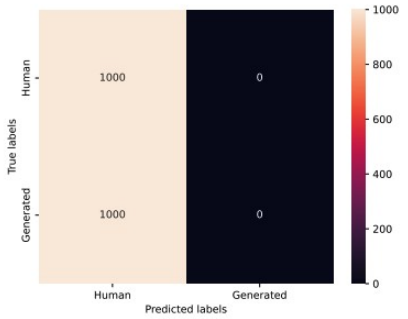
(c) Random attack (10%)



(d) Random attack (15%)



(e) Random attack (20%)



(f) Greedy attack

(Creo, Pudasaini 2024)

Dataset	Detector	Original	5%	10%	15%	20%	Greedy
	ArguGPT	0.94	0.0	0.0	0.0	0.0	0.0

(Creo, Pudasaini 2024)

Dataset	Detector	Original	5%	10%	15%	20%	Greedy
CHEAT	ArguGPT	0.94	0.0	0.0	0.0	0.0	0.0
	Binoculars	0.93					
	DetectGPT	0.14					
	Fast-DetectGPT	0.9					
	Ghostbuster	0.64					
	OpenAI	0.47					
essay	ArguGPT	0.92					
	Binoculars	0.91					
	DetectGPT	0.24					
	Fast-DetectGPT	0.88					
	Ghostbuster	0.92					
	OpenAI	-0.21					
reuter	ArguGPT	0.92					
	Binoculars	0.8					
	DetectGPT	0.23					
	Fast-DetectGPT	0.92					
	Ghostbuster	0.93					
	OpenAI	0.27					
writing prompts	ArguGPT	0.39					
	Binoculars	0.85					
	DetectGPT	0.44					
	Fast-DetectGPT	0.79					
	Ghostbuster	0.88					
	OpenAI	-0.05					
realnewslike	Watermark	0.92					
	Average	0.64					
	Standard deviation	0.36					

(Creo, Pudasaini 2024)

Dataset	Detector	Original	5%	10%	15%	20%	Greedy
CHEAT	ArguGPT	0.94	0.0	0.0	0.0	0.0	0.0
	Binoculars	0.93	0.37	0.11	0.04	0.02	0.13
	DetectGPT	0.14	-0.02	0.03	0.13	0.06	0.0
	Fast-DetectGPT	0.9	0.23	0.04	0.02	0.0	-0.01
	Ghostbuster	0.64	0.41	0.32	0.12	0.06	0.02
	OpenAI	0.47	0.0	0.0	0.0	-0.02	0.0
essay	ArguGPT	0.92	0.0	0.0	0.0	0.0	0.0
	Binoculars	0.91	0.22	0.05	0.0	0.0	0.05
	DetectGPT	0.24	-0.01	0.11	0.21	0.08	0.0
	Fast-DetectGPT	0.88	0.22	0.04	0.0	0.0	-0.08
	Ghostbuster	0.92	0.73	0.51	0.13	0.0	0.0
	OpenAI	-0.21	0.0	0.0	0.0	0.0	0.03
reuter	ArguGPT	0.92	0.0	0.0	0.0	0.0	0.0
	Binoculars	0.8	0.22	0.07	0.03	0.02	0.08
	DetectGPT	0.23	0.0	0.03	0.34	0.14	0.0
	Fast-DetectGPT	0.92	0.28	0.1	0.02	0.0	0.04
	Ghostbuster	0.93	0.61	0.51	0.16	0.04	0.0
	OpenAI	0.27	0.0	-0.04	-0.09	-0.11	-0.06
writing prompts	ArguGPT	0.39	0.0	0.0	0.0	0.0	0.0
	Binoculars	0.85	0.2	0.0	0.0	0.0	-0.04
	DetectGPT	0.44	0.04	0.01	0.02	0.02	0.0
	Fast-DetectGPT	0.79	0.3	0.05	-0.03	0.0	-0.33
	Ghostbuster	0.88	0.42	0.64	0.33	0.09	0.0
	OpenAI	-0.05	-0.04	-0.05	-0.13	-0.11	0.01
realnewslike	Watermark	0.92	0.18	-0.01	0.0	-0.03	0.0
Average		0.64	0.17	0.1	0.05	0.01	-0.01
Standard deviation		0.36	0.21	0.19	0.11	0.05	0.08

(Creo, Pudasaini 2024)

Highly effective, renders all detectors ineffective

Implications

Twitter
bots

Academic misconduct

Fraud, social engineering

Identity
theft

Fake news, loss of trust in media

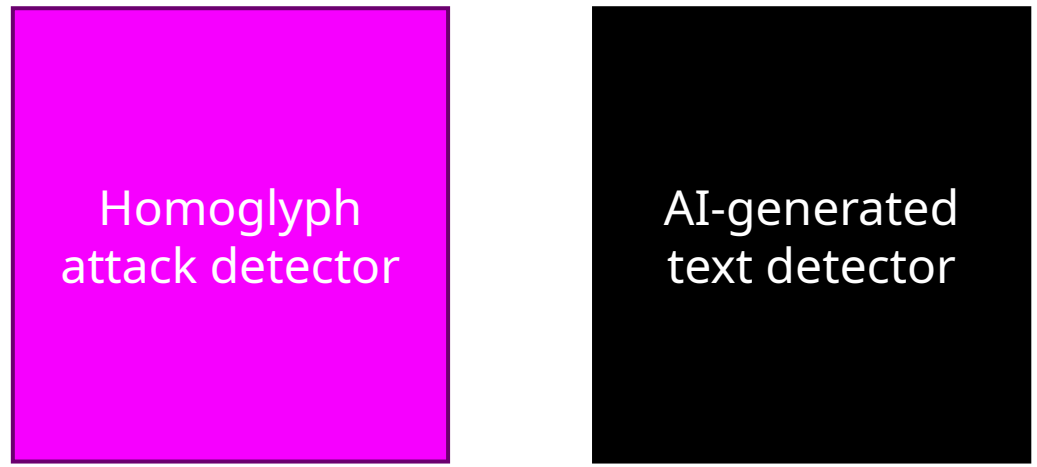
Lower access barrier: increased risk

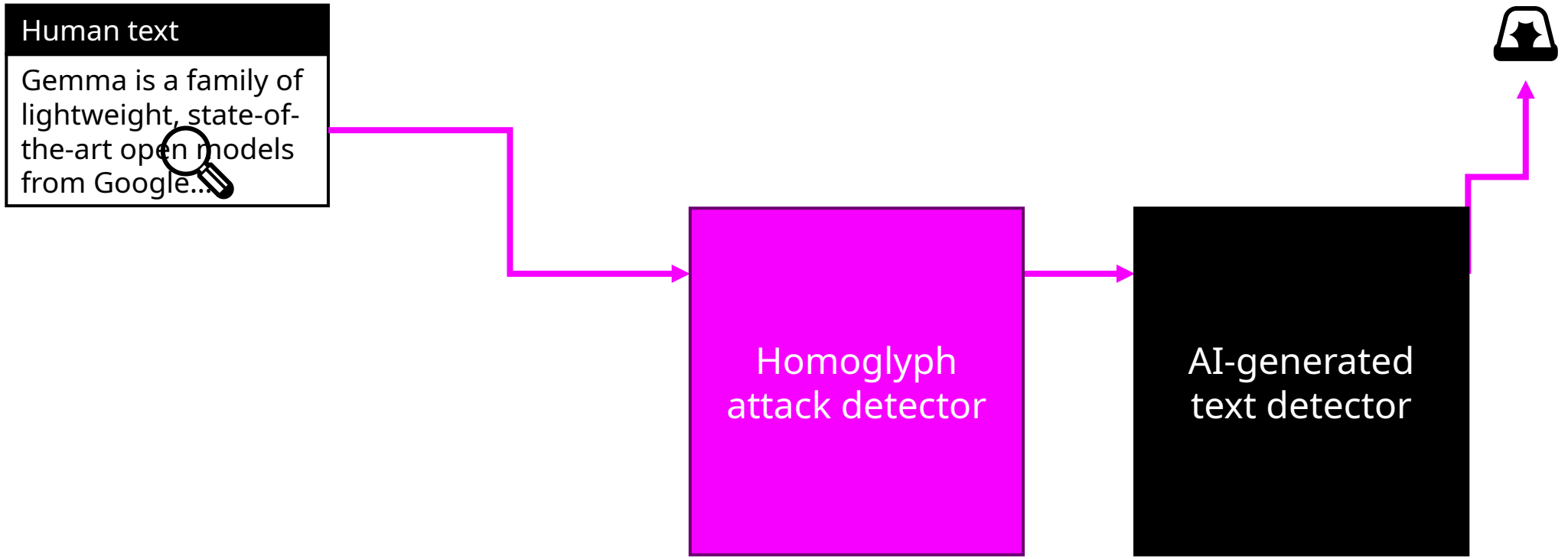
Safeguards

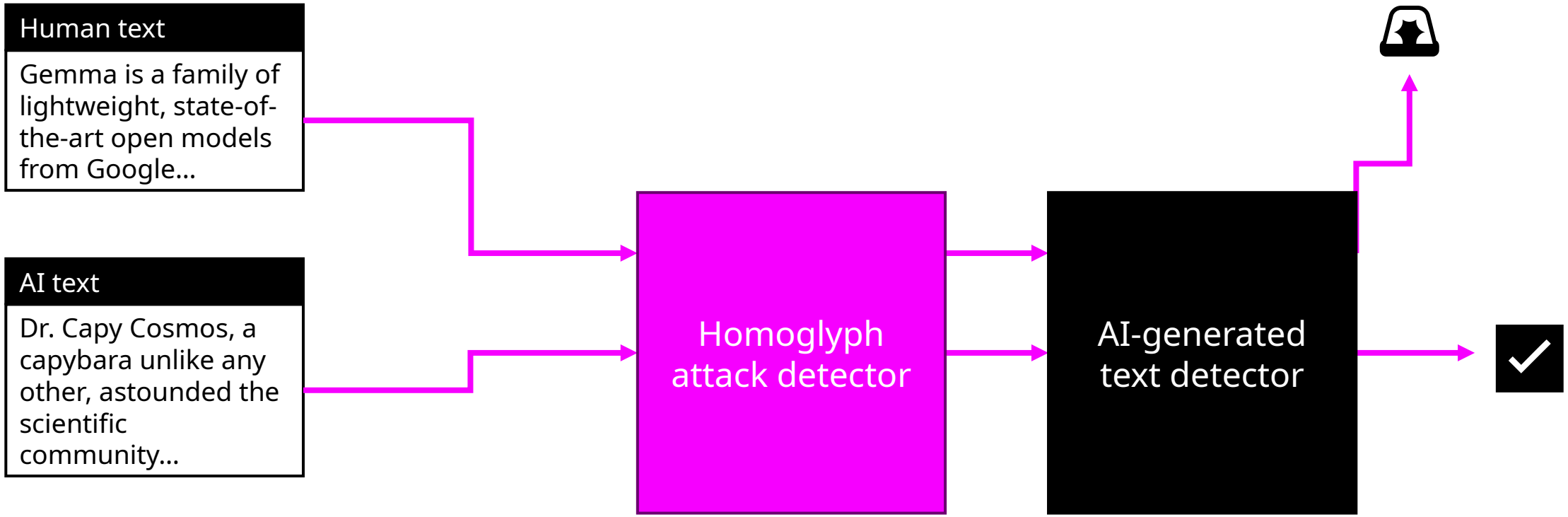
Safeguards

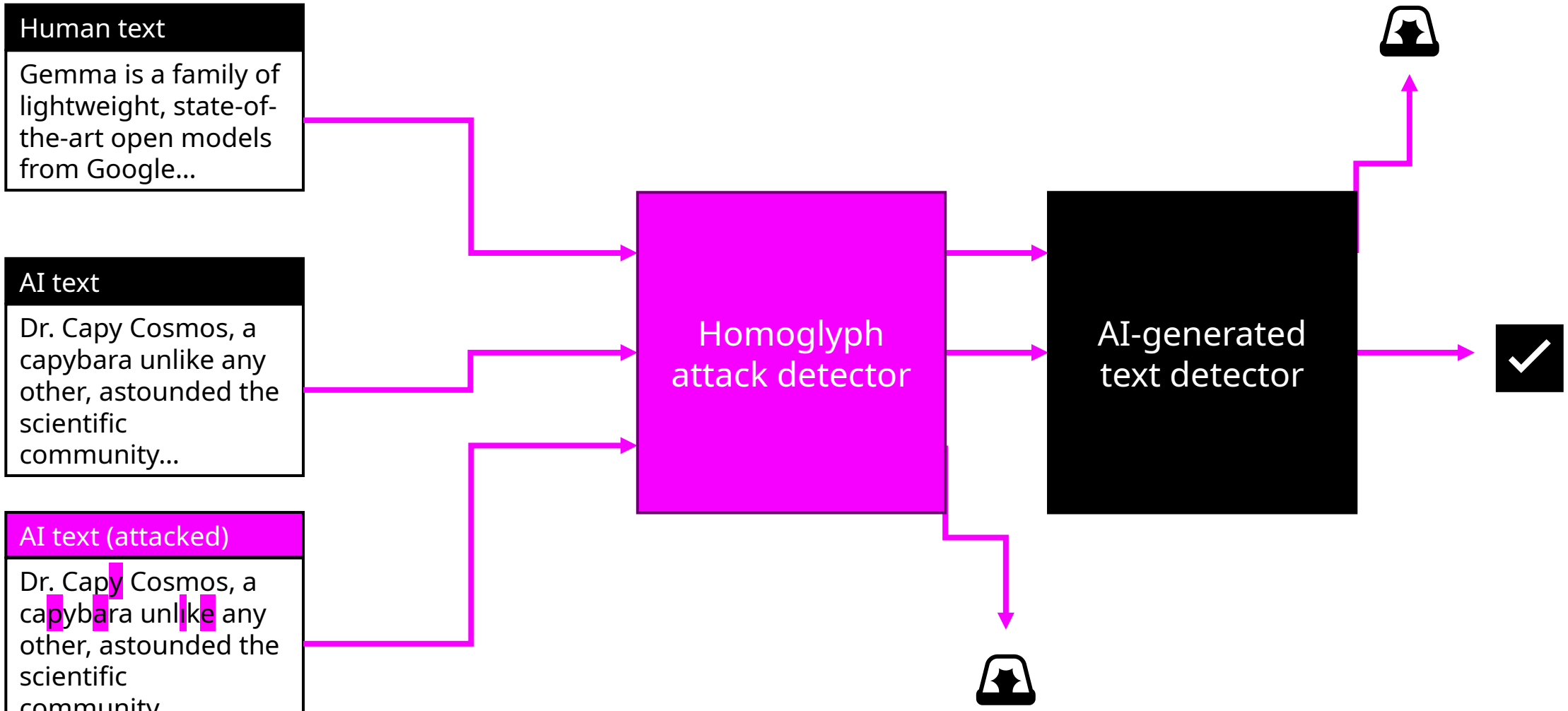
Input constraints

AI-generated
text detector


Homoglyph
attack detector
AI-generated
text detector


Human text
Gemma is a family of lightweight, state-of-the-art open models from Google...
Homoglyph attack detector
AI-generated text detector


Human text
Gemma is a family of lightweight, state-of-the-art open models from Google...
AI text
Dr. Capy Cosmos, a capybara unlike any other, astounded the scientific community...
Homoglyph attack detector
AI-generated text detector


Human text
Gemma is a family of lightweight, state-of-the-art open models from Google...
AI text
Dr. Capy Cosmos, a capybara unlike any other, astounded the scientific community...
AI text (attacked)
Dr. Capy Cosmos, a capybara unlike any other, astounded the scientific community...
Homoglyph attack detector
AI-generated text detector

Maybe not
enough...

Maybe not enough…

Specifically, considering that LLMs utilize decoder-only structures, the distribution for the i-th token at the final layer L, denoted as p_i^L, depends on the past i tokens. This represents the probability of any token in the vocabulary becoming the $(i+1)$-th token. The distribution difference between the target token and the alternative token, $p_i^L(w_t) - p_i^L(w_a)$, can be attributed to the cumulative contribution of the first i tokens. Similarly, $p_{i-1}^L(w_t) - p_{i-1}^L(w_a)$ reflects the collective contribution of the initial $i-1$ tokens.

(Feng et al. 2024)

Maybe not enough...

Specifically, considering that LLMs utilize decoder-only structures, the distribution for the i-th token at the final layer L, denoted as p_i^L, depends on the past i tokens. This represents the probability of any token in the vocabulary becoming the $(i+1)$-th token. The distribution difference between the target token and the alternative token, $p_i^L(w_t) - p_i^L(w_a)$, can be attributed to the cumulative contribution of the first i tokens. Similarly, $p_{i-1}^L(w_t) - p_{i-1}^L(w_a)$ reflects the collective contribution of the initial $i-1$ tokens.

(Feng et al. 2024)

Maybe not
enough...

Specifically, considering that LLMs utilize decoder-only structures, the distribution for the i-th token at the final layer L, denoted as p_i^L, depends on the past i tokens. This represents the probability of any token in the vocabulary becoming the $(i+1)$-th token. The distribution difference between the target token and the alternative token, $p_i^L(w_t) - p_i^L(w_a)$, can be attributed to the cumulative contribution of the first i tokens. Similarly, $p_{i-1}^L(w_t) - p_{i-1}^L(w_a)$ reflects the collective contribution of the initial $i-1$ tokens.

False positive

(Feng et al. 2024)

We need to get
smarter

Rule-based approach

Specifically, considering that LLMs utilize decoder-only structures, the distribution for the i-th token at the final layer L, denoted as p_i^L, depends on the past i tokens. This represents the probability of any token in the vocabulary becoming the $(i+1)$-th token. The distribution difference between the target token and the alternative token, $p_i^L(w_t) - p_i^L(w_a)$, can be attributed to the cumulative contribution of the first i tokens. Similarly, $p_{i-1}^L(w_t) - p_{i-1}^L(w_a)$ reflects the collective contribution of the initial $i-1$ tokens.

Percentage of words that mix homoglyphs

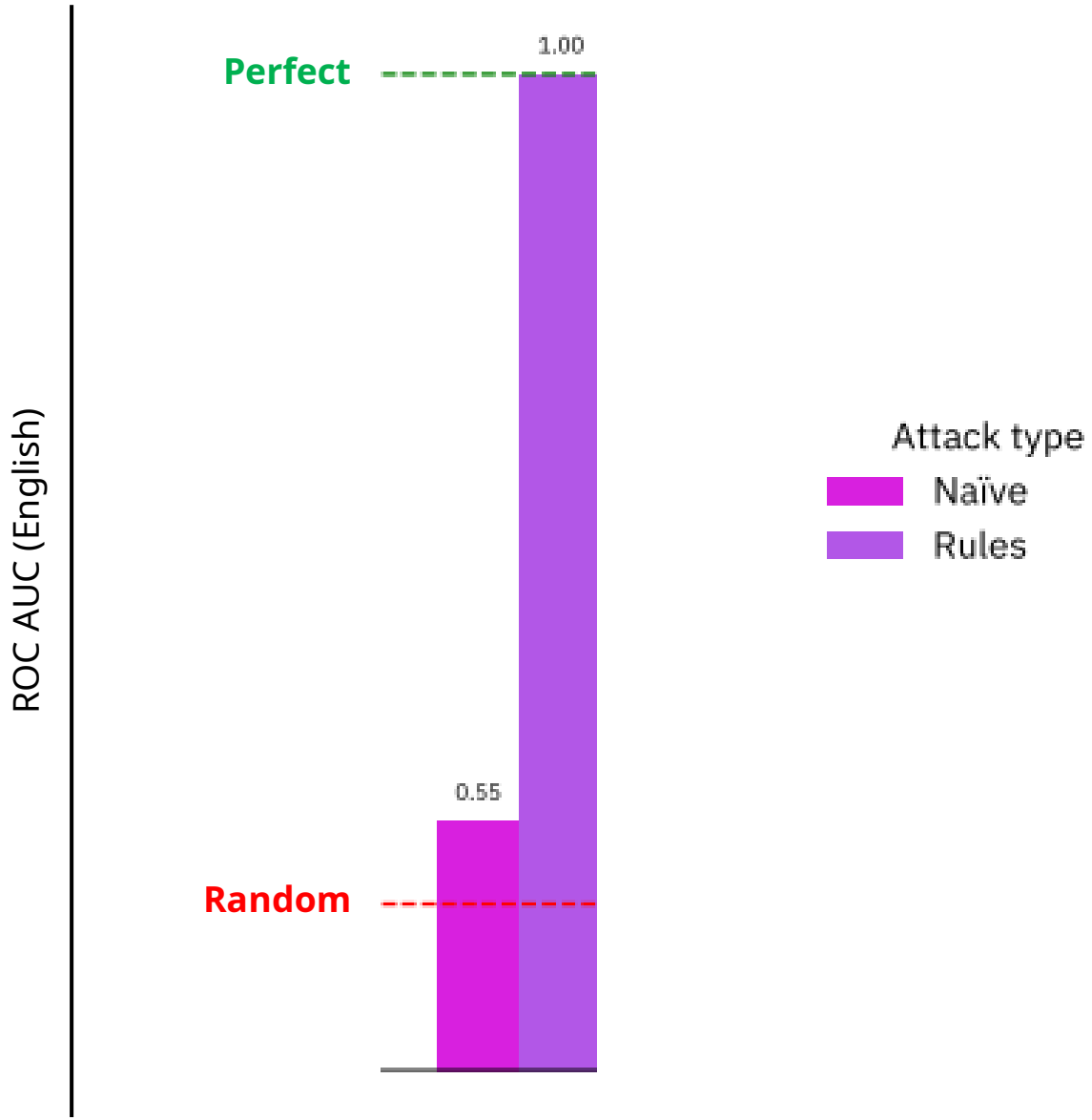
(Feng et al. 2024)

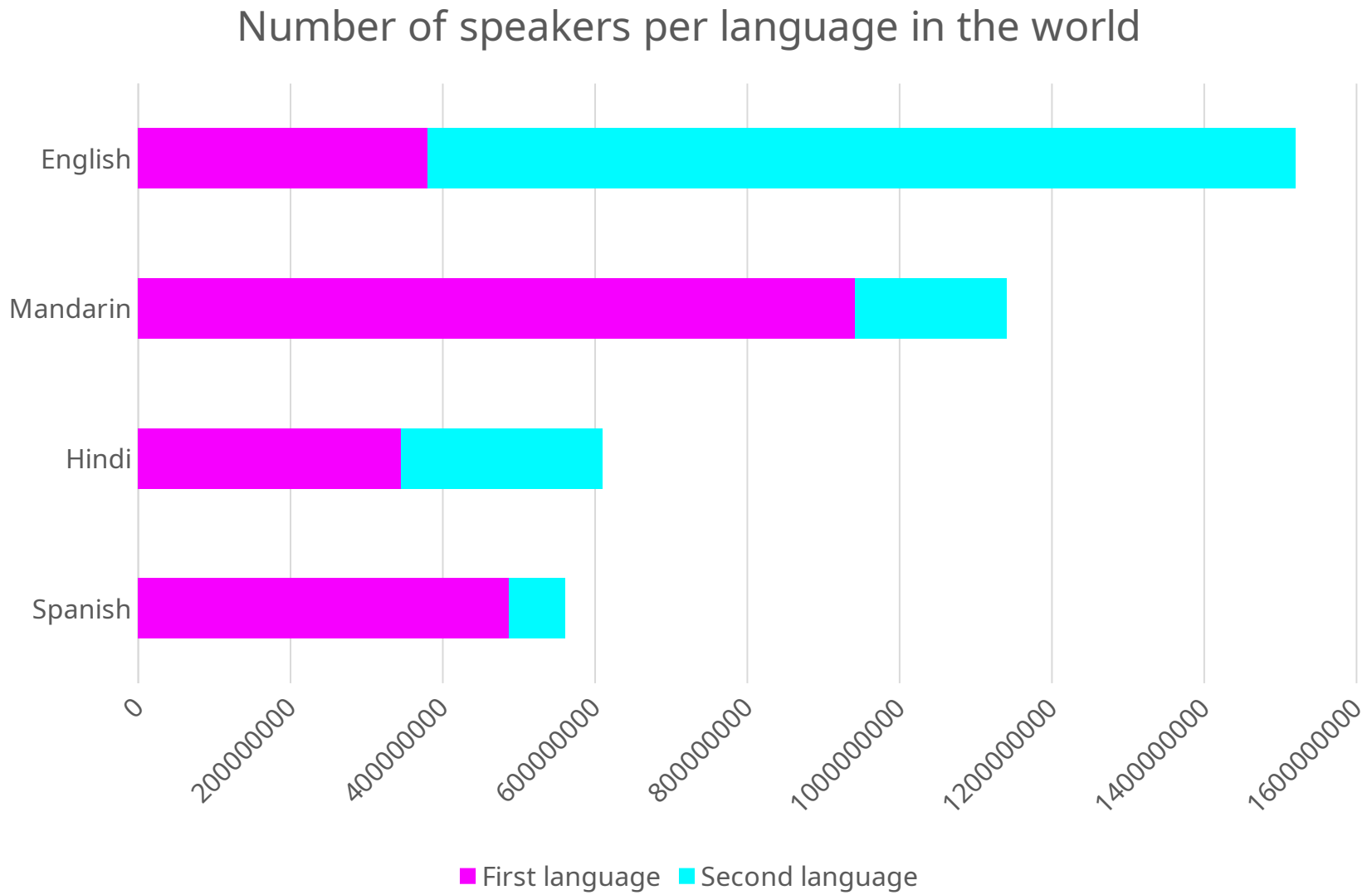
Rule-based approach

Specifically, considering that LLMs utilize decoder-only structures, the distribution for the i-th token at the final layer L, denoted as p_i^L, depends on the past i tokens. This represents the probability of any token in the vocabulary becoming the $(i+1)$-th token. The distribution difference between the target token and the alternative token, $p_i^L(w_t) - p_i^L(w_a)$, can be attributed to the cumulative contribution of the first i tokens. Similarly, $p_{i-1}^L(w_t) - p_{i-1}^L(w_a)$ reflects the collective contribution of the initial $i-1$ tokens.

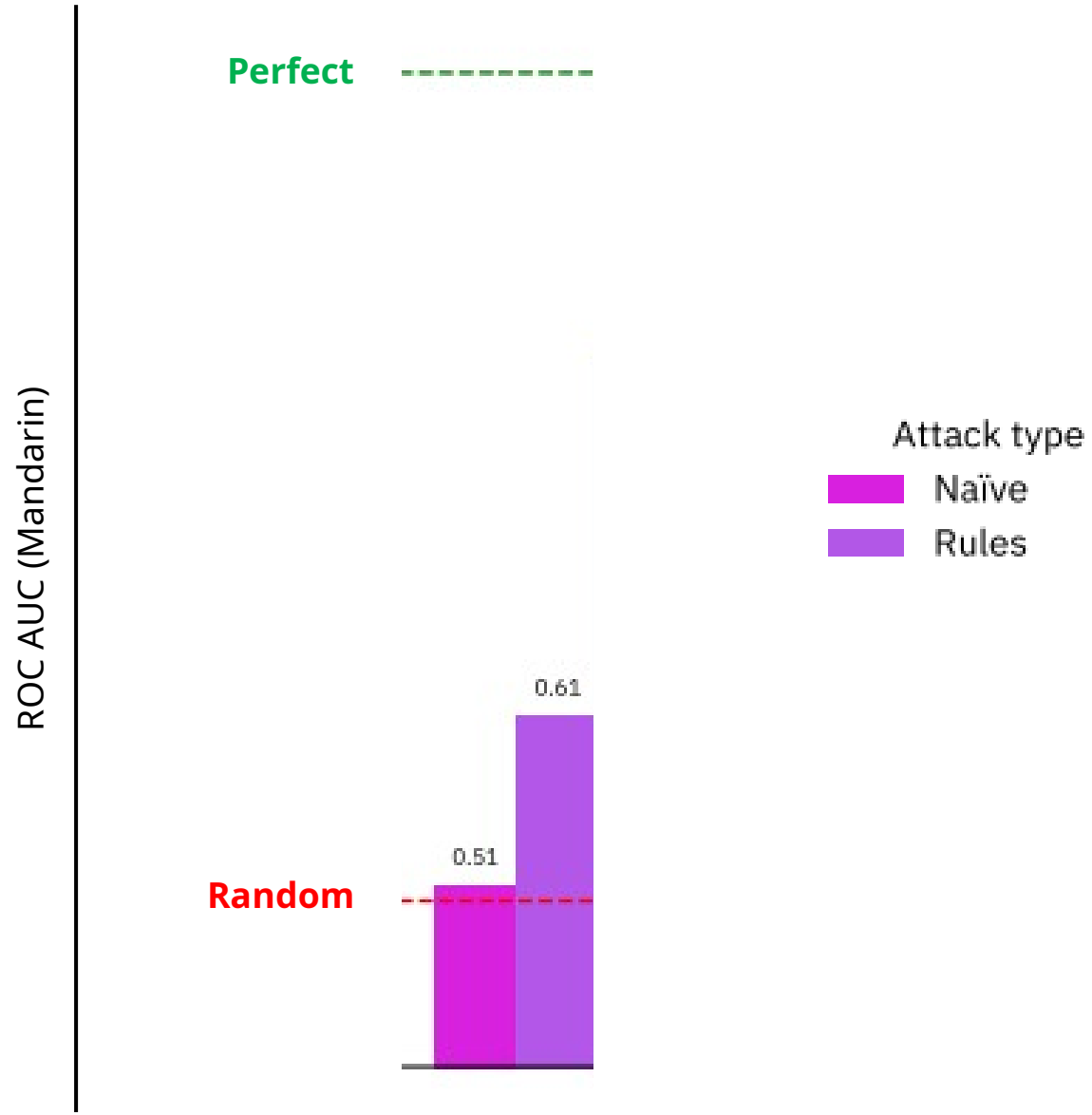
All good

(Feng et al. 2024)


ROC AUC (English)
Perfect
1.00
0.55
Random
Attack type
Naïve
Rules



(Eberhard et al. 2022)

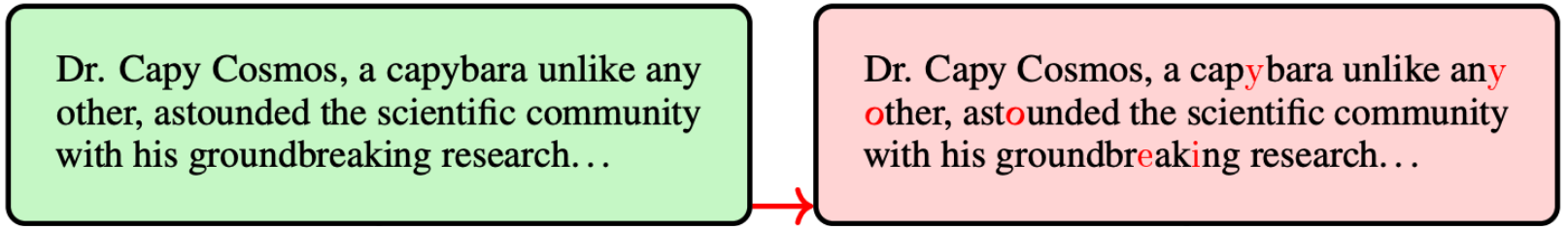

ROC AUC (Mandarin)
Perfect
Random
0.61
0.51
Attack type
Naïve
Rules

What else can we do?

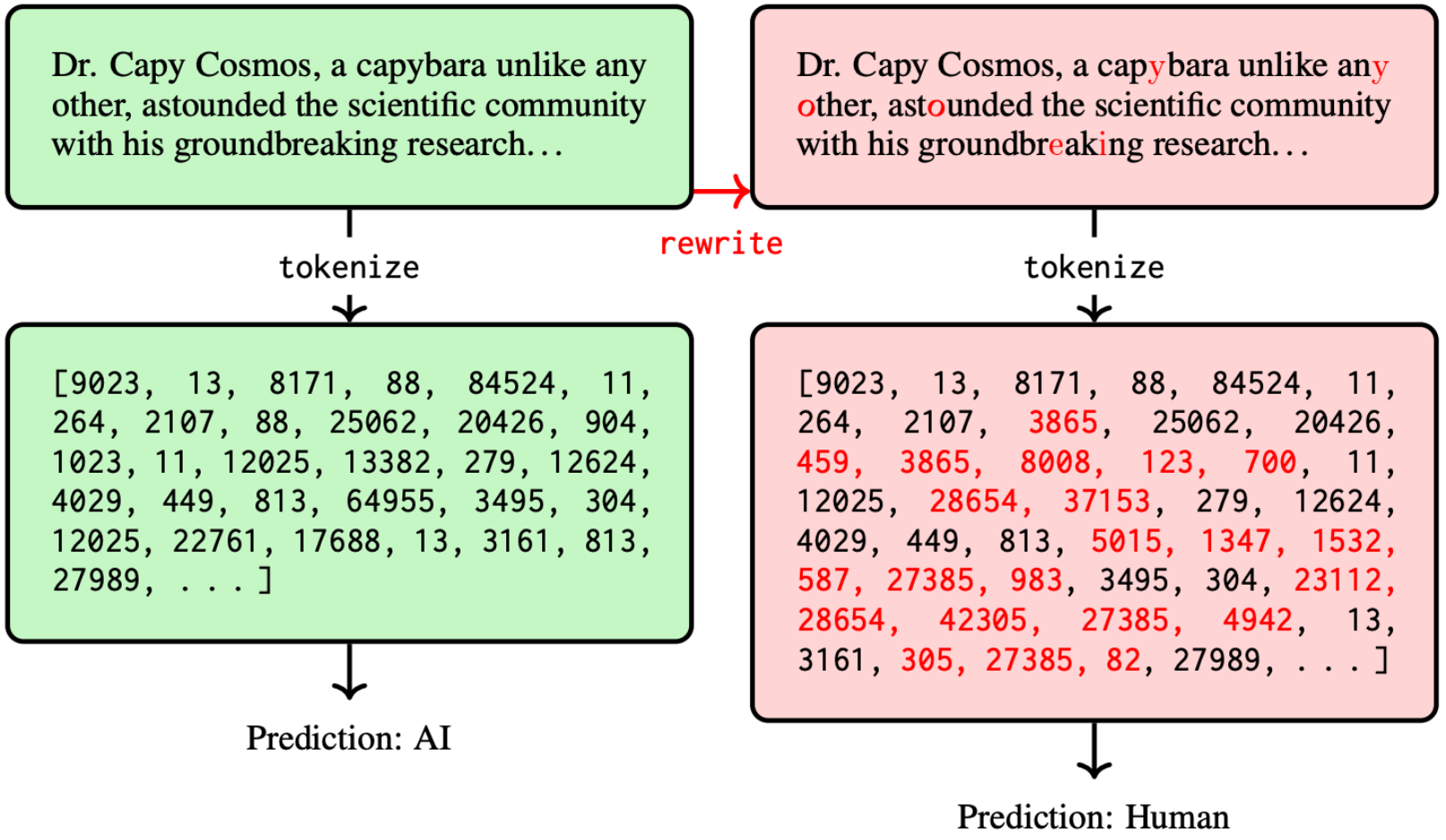
What else can we do?

2 idęas

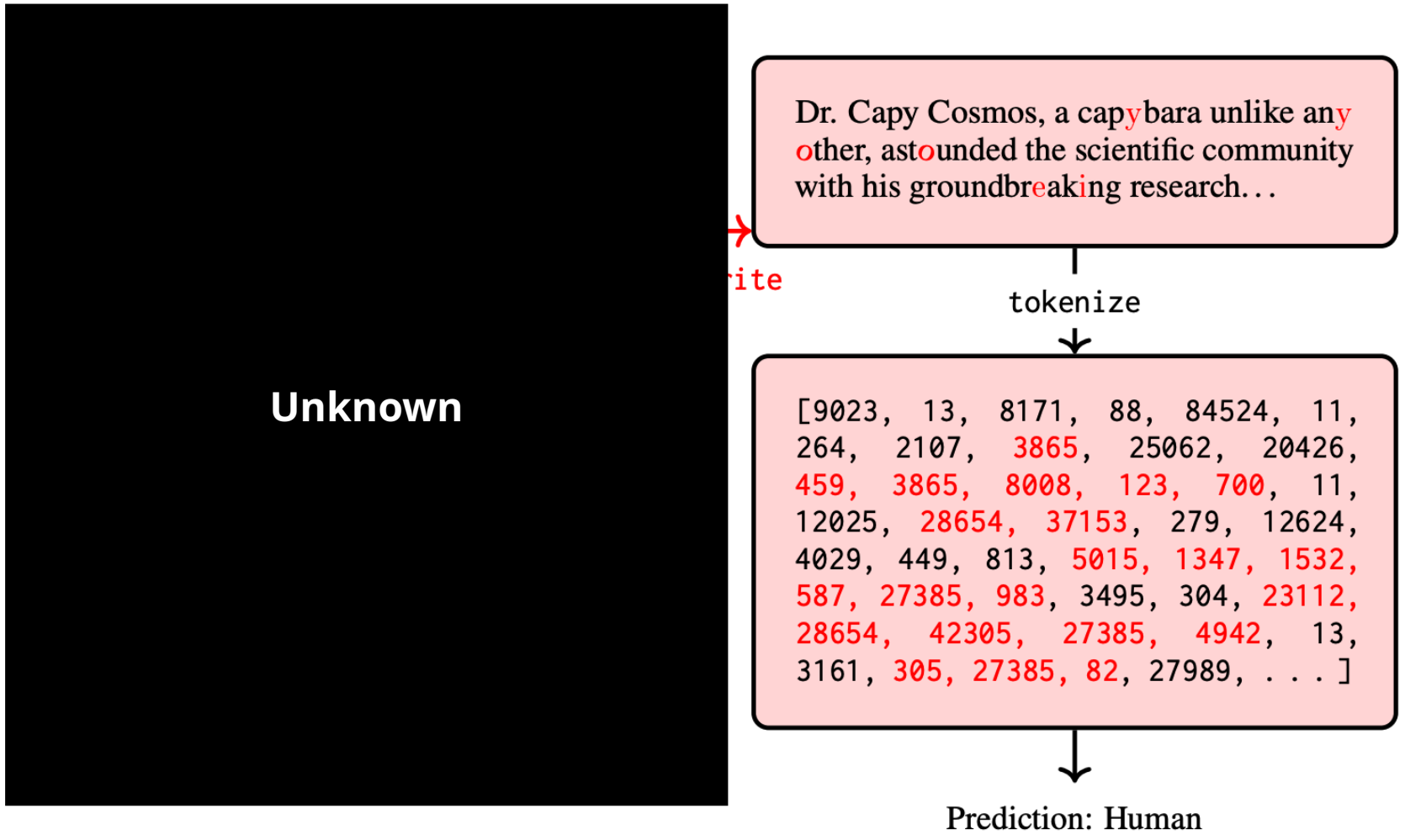
Idea 1: Tokenization



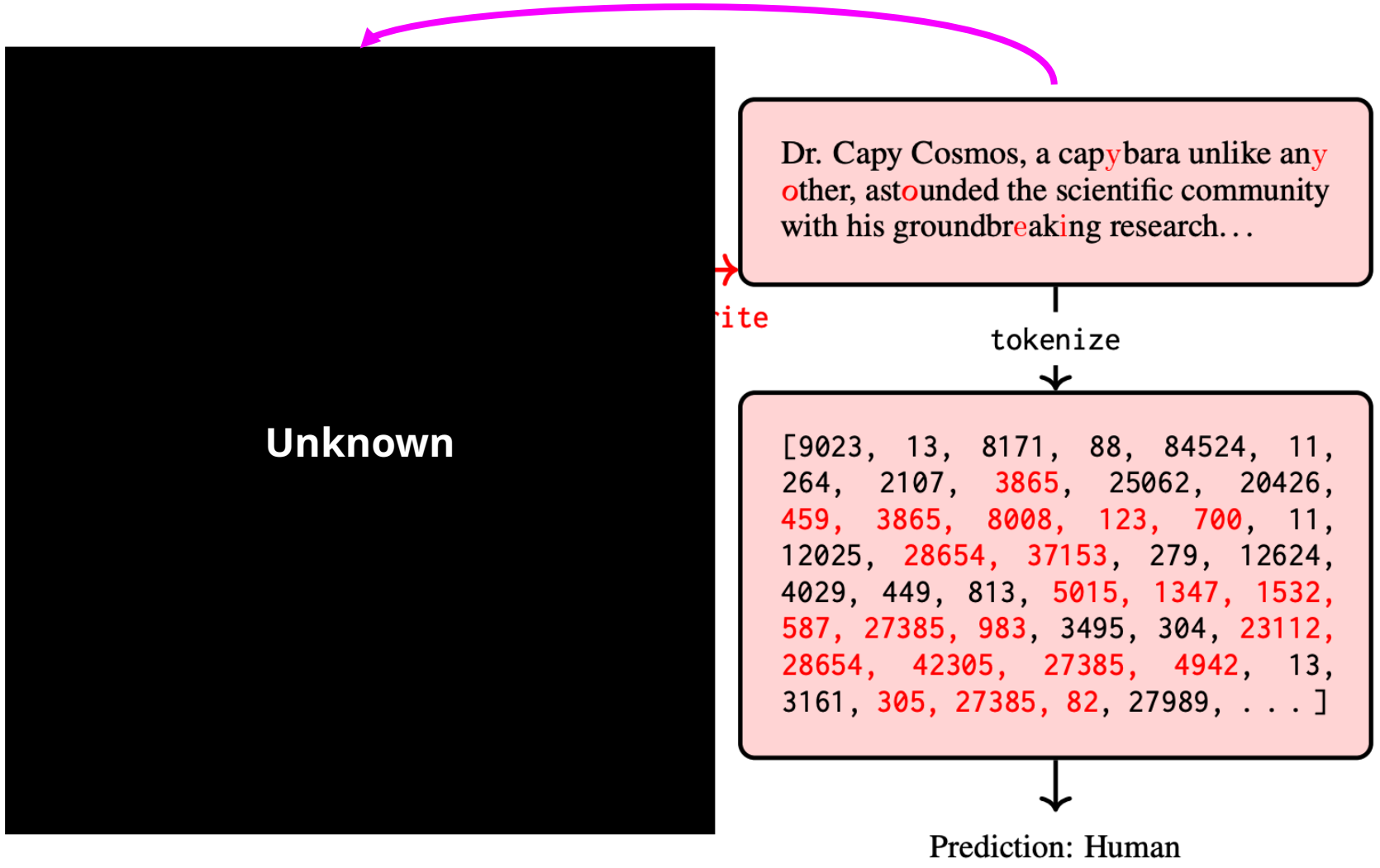
(Creo, Pudasaini 2024)



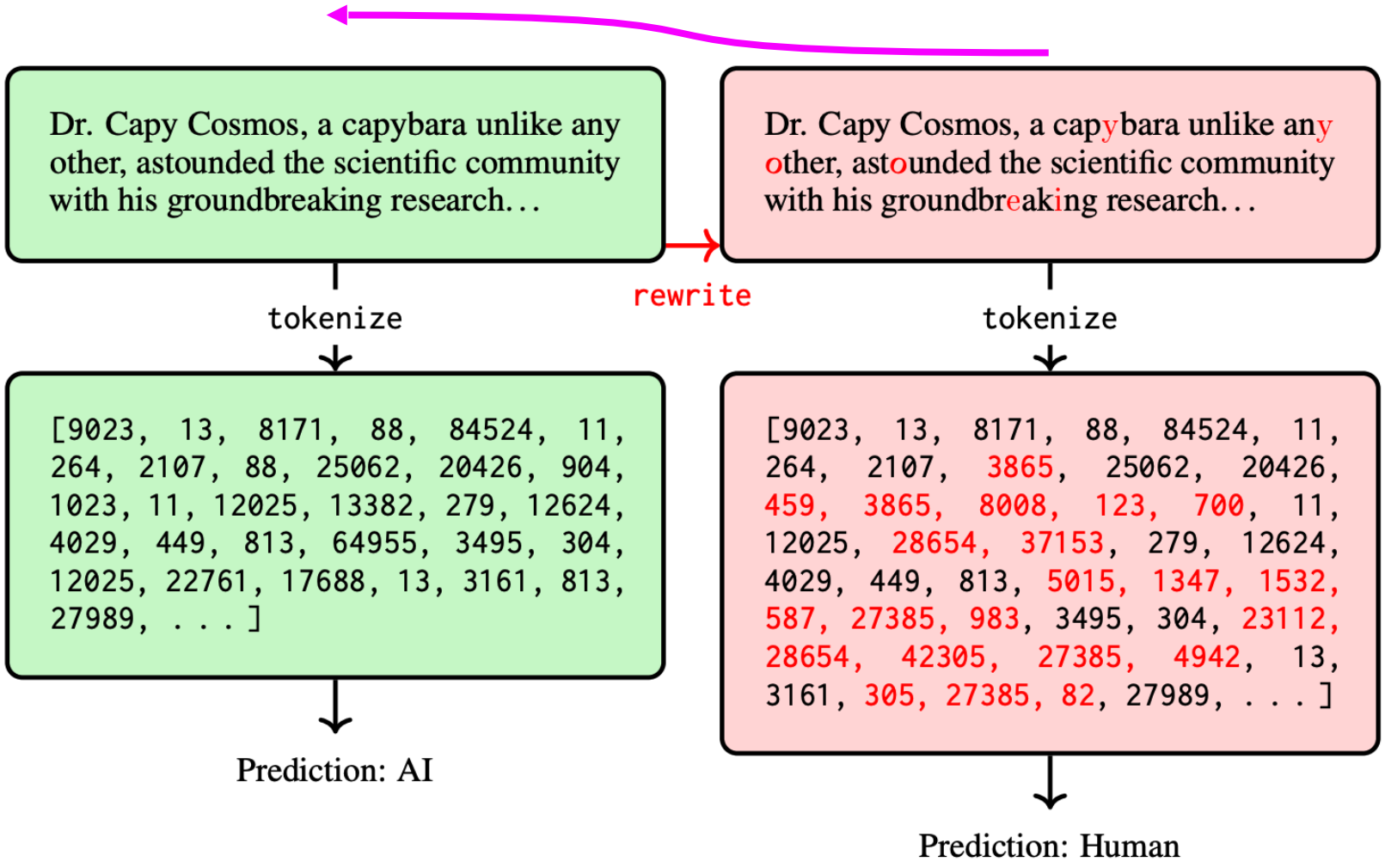
(Creo, Pudasaini 2024)



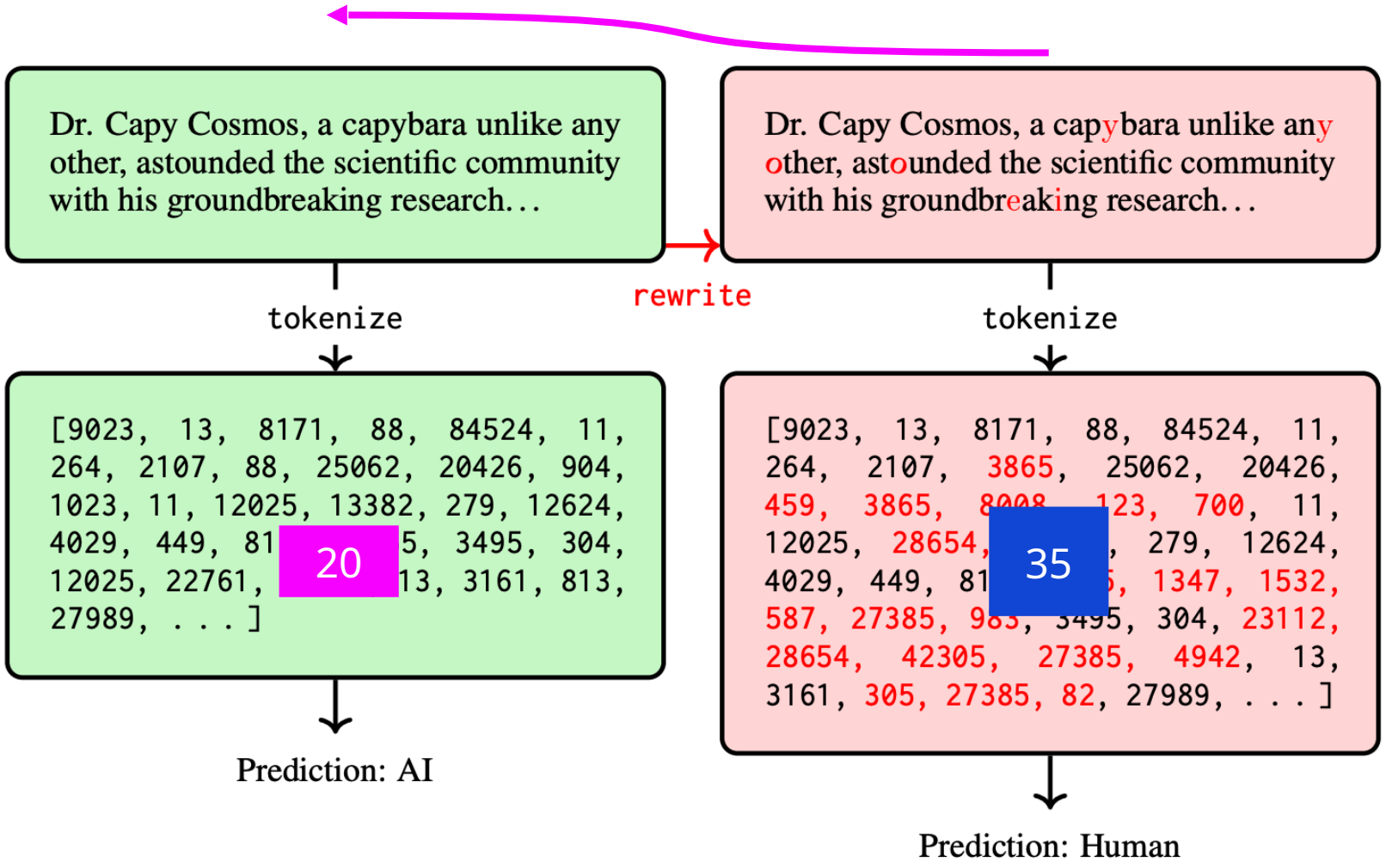
(Creo, Pudasaini 2024)



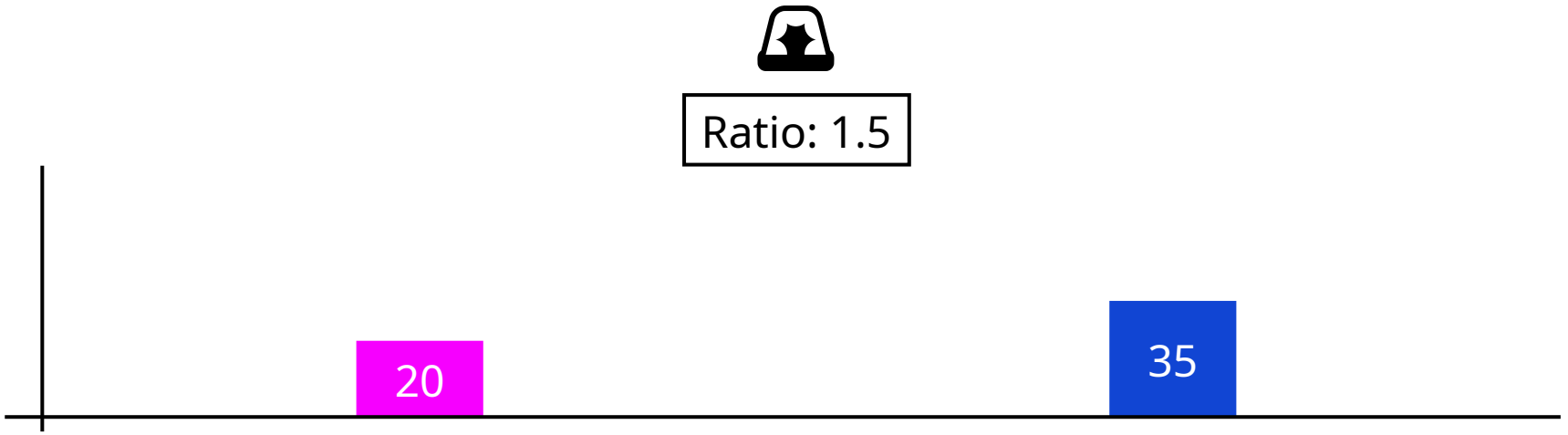
(Creo, Pudasaini 2024)



(Creo, Pudasaini 2024)

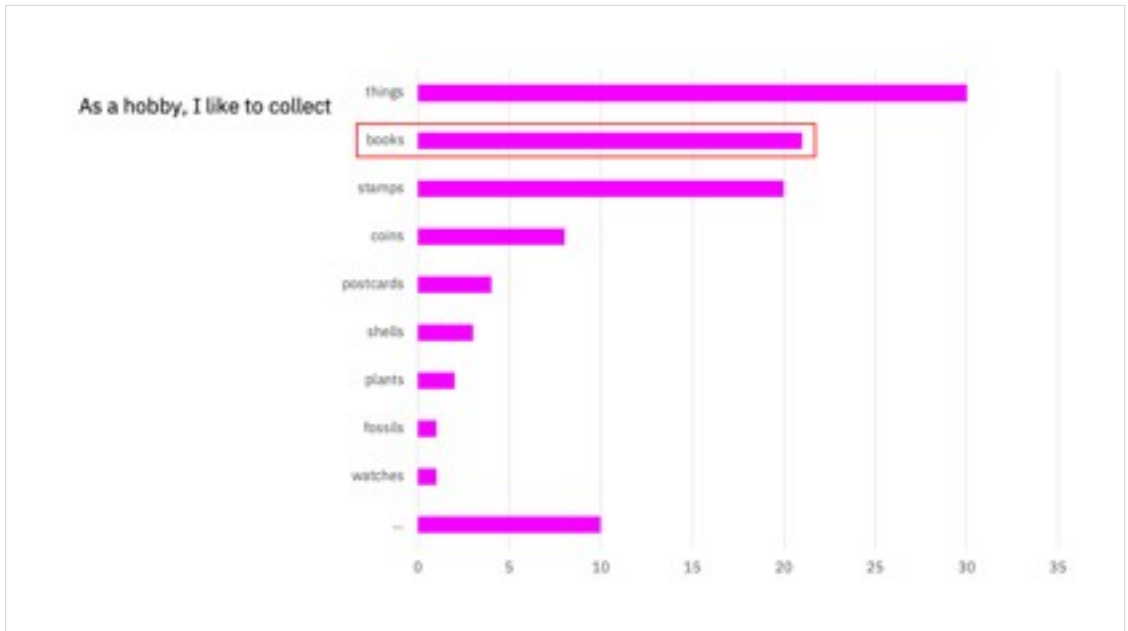


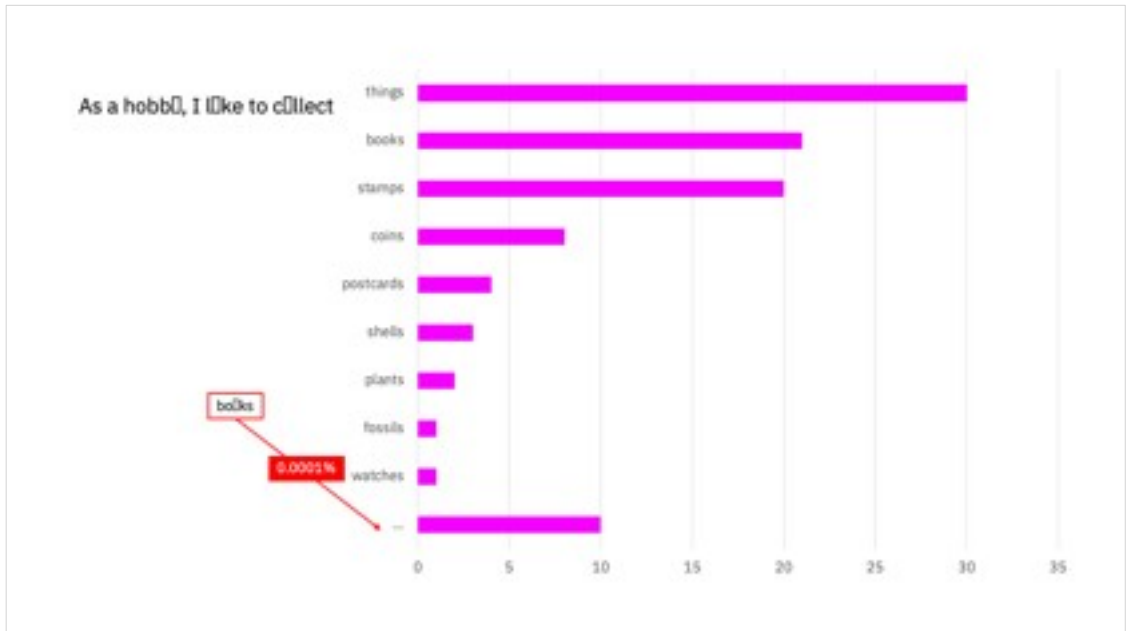
(Creo, Pudasaini 2024)



(Creo, Pudasaini 2024)

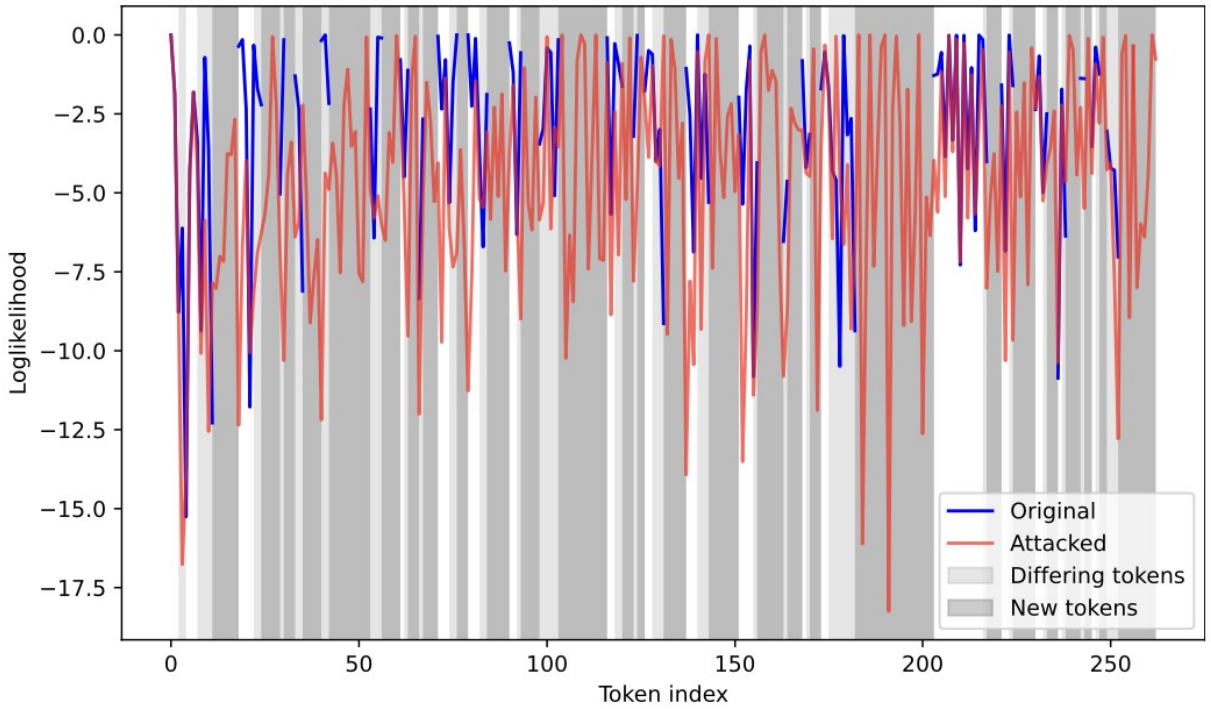
Idea 2: Perplexity


As a hobby, I like to collect
things
books
stamps
coins
postcards
shells
plants
fossils
watches
...
0
5
10
15
20
25
30
35


As a hobb□, I l□ke to c□llect
things
books
stamps
coins
postcards
shells
plants
fossils
watches
...
bo□ks
0.0001%
0
5
10
15
20
25
30
35

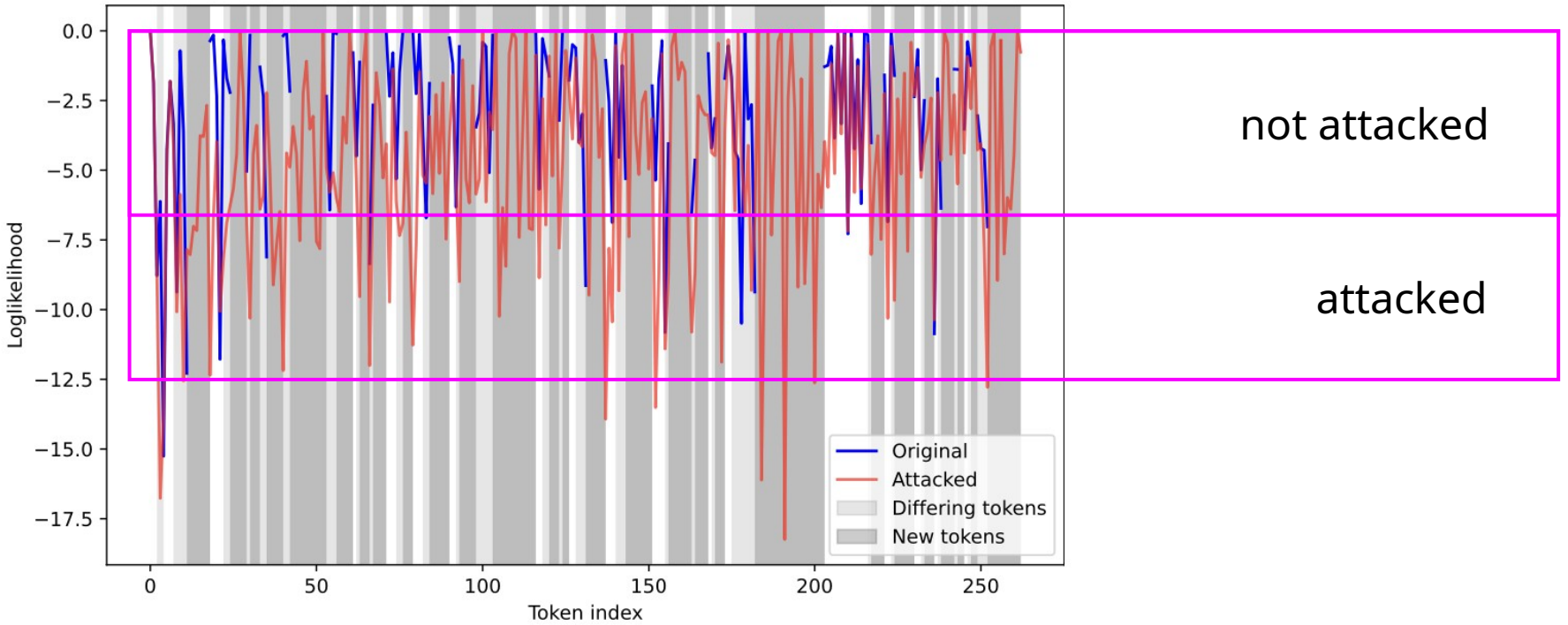
Dr. Capy Cosmos, a capybara unlike any other, astounded the scientific community with his groundbreaking research...

Dr. Capy Cosmos, a capybara unlike any other, astounded the scientific community with his groundbreaking research...



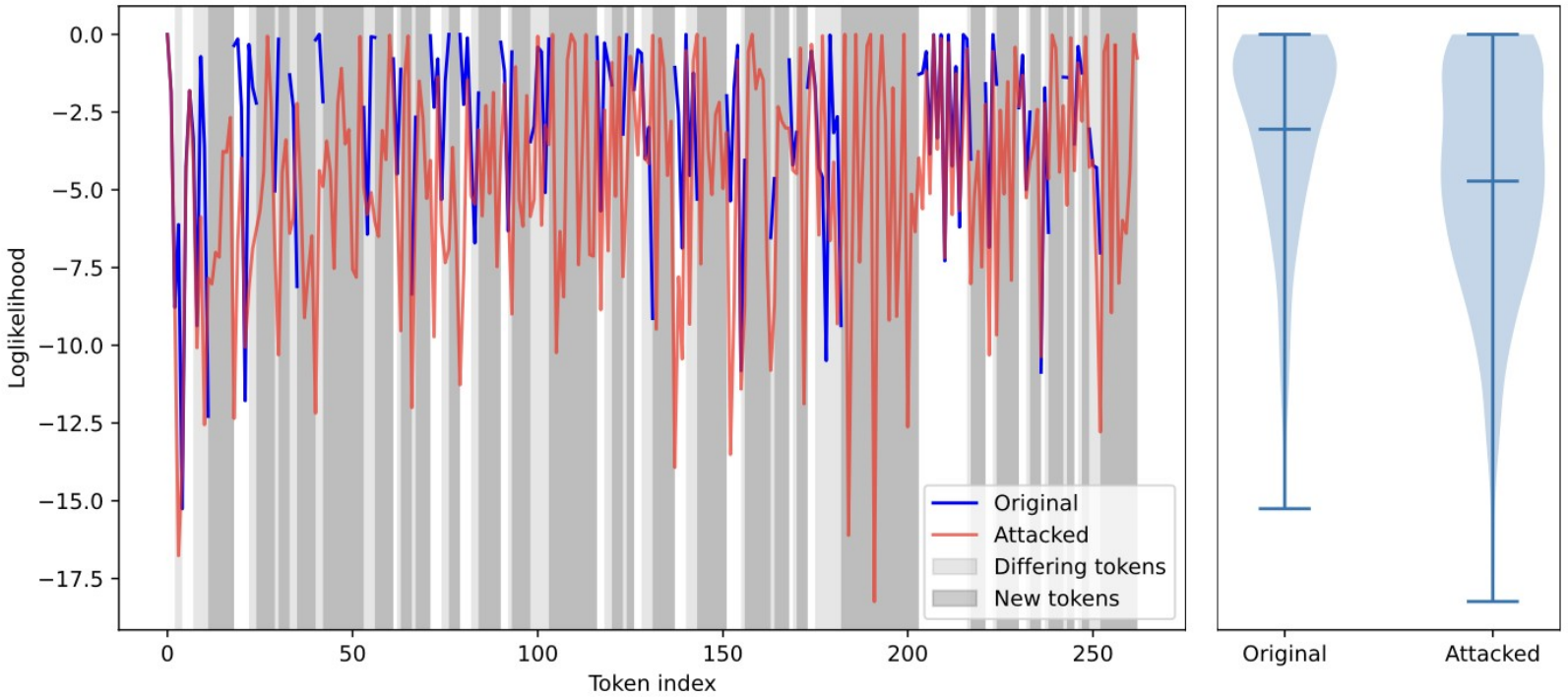
(a) Differences in loglikelihood per token

(Creo, Pudasaini 2024)



(a) Differences in loglikelihood per token

(Creo, Pudasaini 2024)

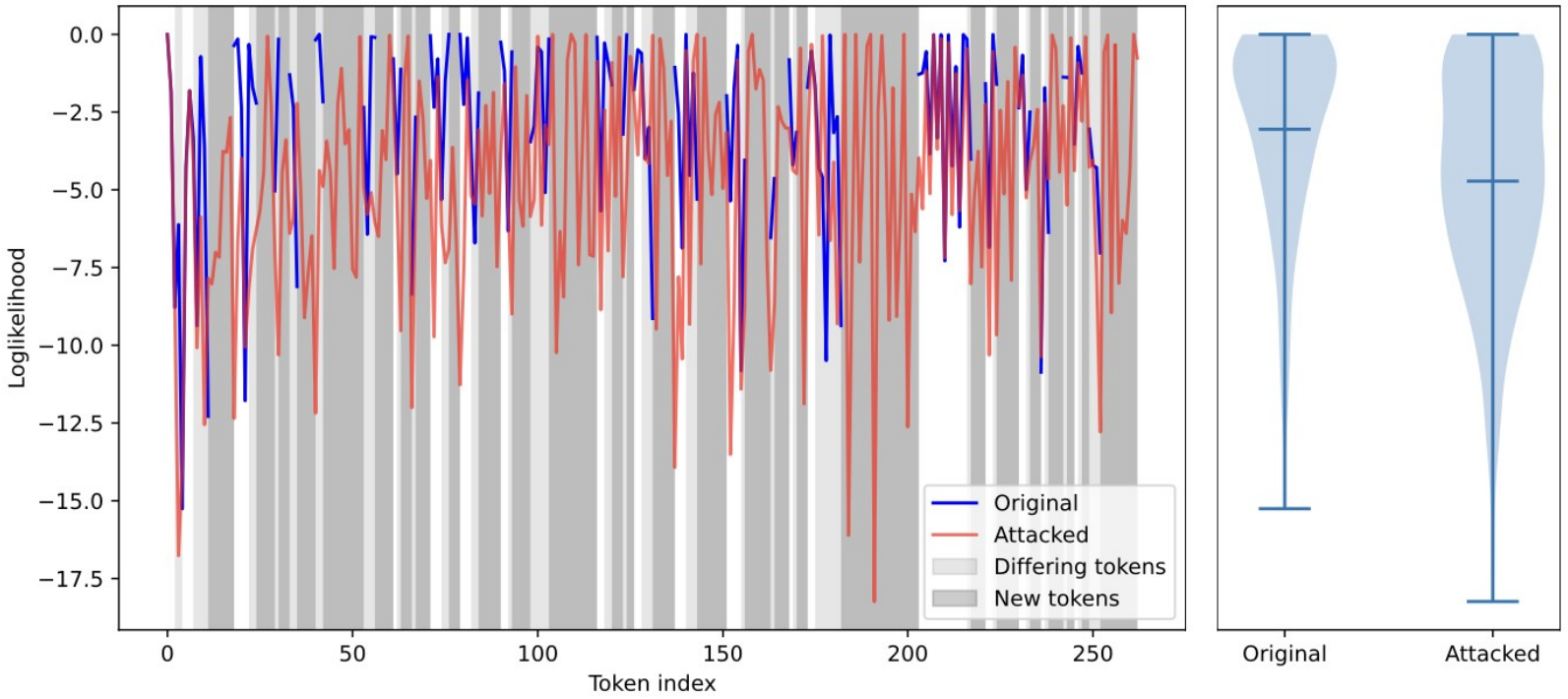


(a) Differences in loglikelihood per token

(b) Distribution

(Creo, Pudasaini 2024)

Perplexity values are not enough

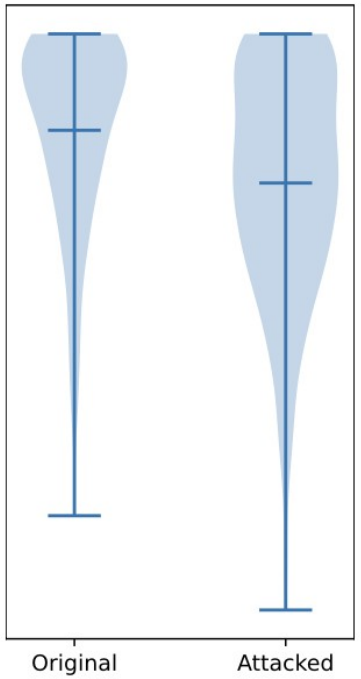


(a) Differences in loglikelihood per token

(b) Distribution

(Creo, Pudasaini 2024)

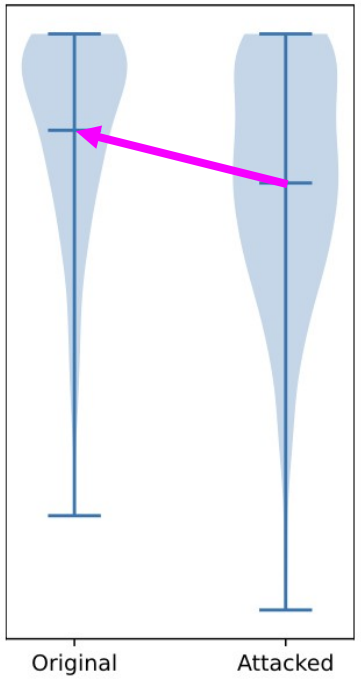
Perplexity values are not enough



(b) Distribution

(Creo, Pudasaini 2024)

Perplexity values are not enough

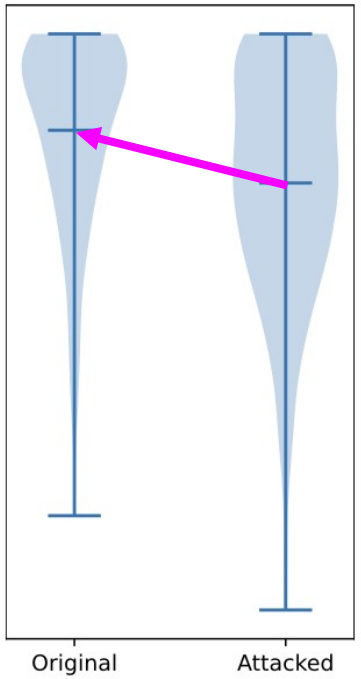


(b) Distribution

Normalize

(Creo, Pudasaini 2024)

Perplexity ratios are enough



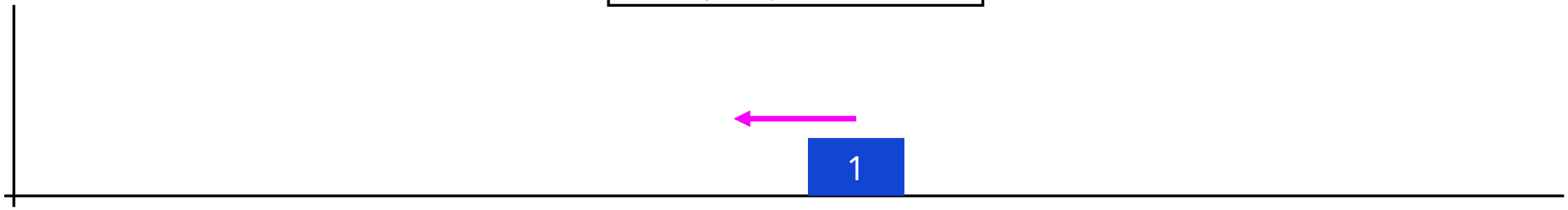
(b) Distribution

Ratio: 1.3

(Creo, Pudasaini 2024)

US Constitution

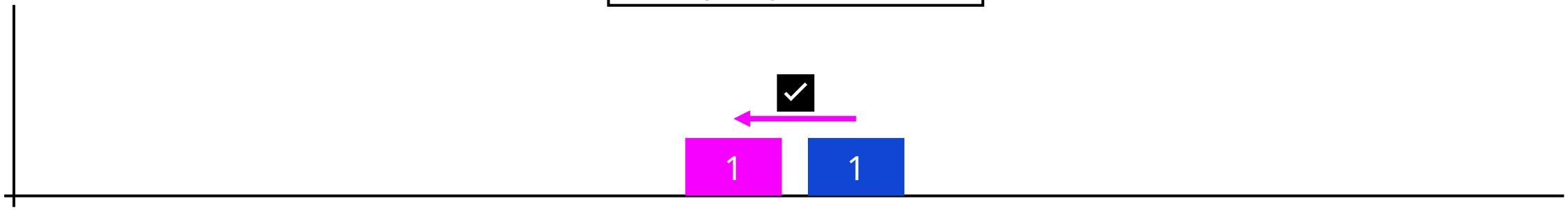
We the People of the United States, in Order to form a more perfect Union, establish Justice, insure domestic Tranquility...



Ratio: 1.3

US Constitution

We the People of the United States, in Order to form a more perfect Union, establish Justice, insure domestic Tranquility...



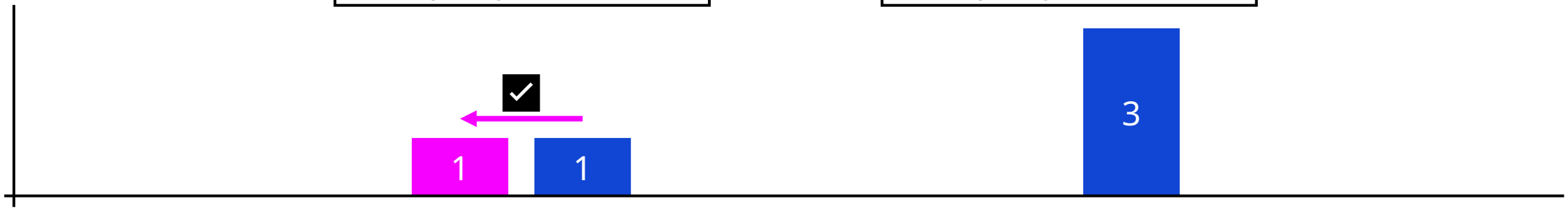
Ratio: 1.3

US Constitution

We the People of the United States, in Order to form a more perfect Union, establish Justice, insure domestic Tranquility...

US Constitution attack

We the People of the United States, in Order to form a more perfect Union, establish Justice, insure domestic Tranquility...



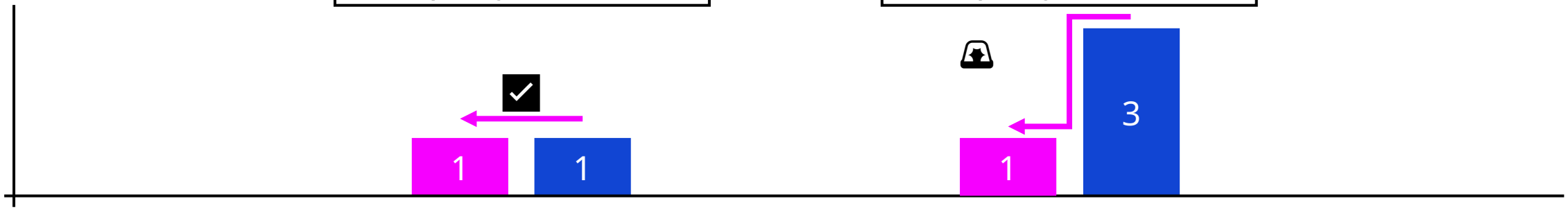
Ratio: 1.3

US Constitution

We the People of the United States, in Order to form a more perfect Union, establish Justice, insure domestic Tranquility...

US Constitution attack

We the People of the United States, in Order to form a more perfect Union, establish Justice, insure domestic Tranquility...

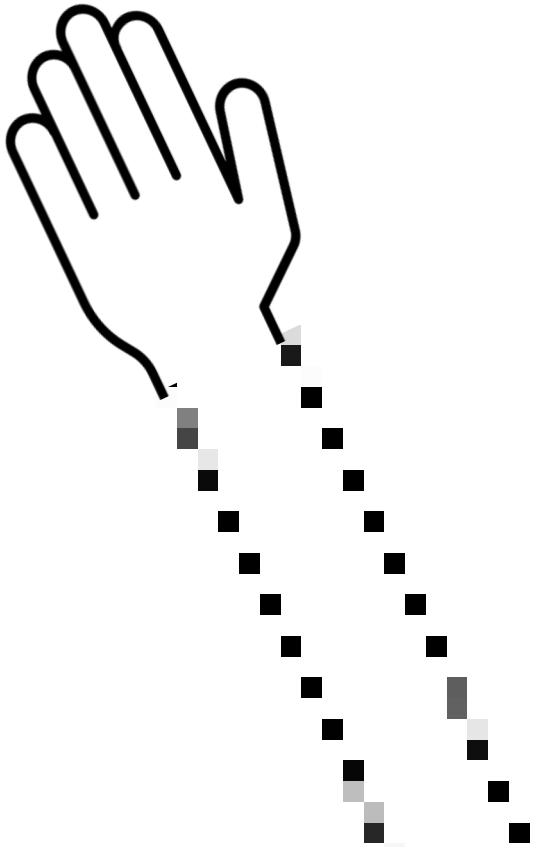


Ratio: 1.3

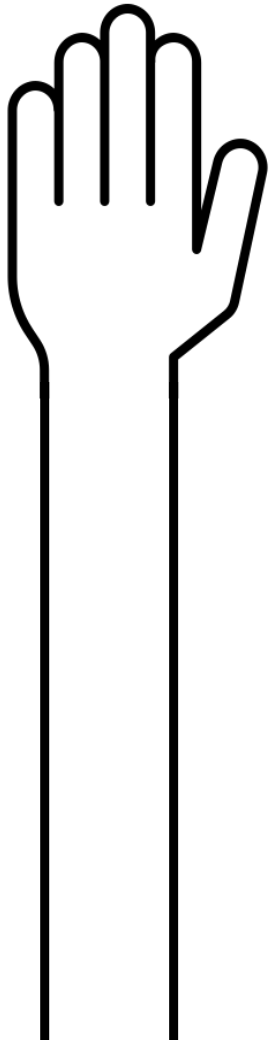
Demo
!

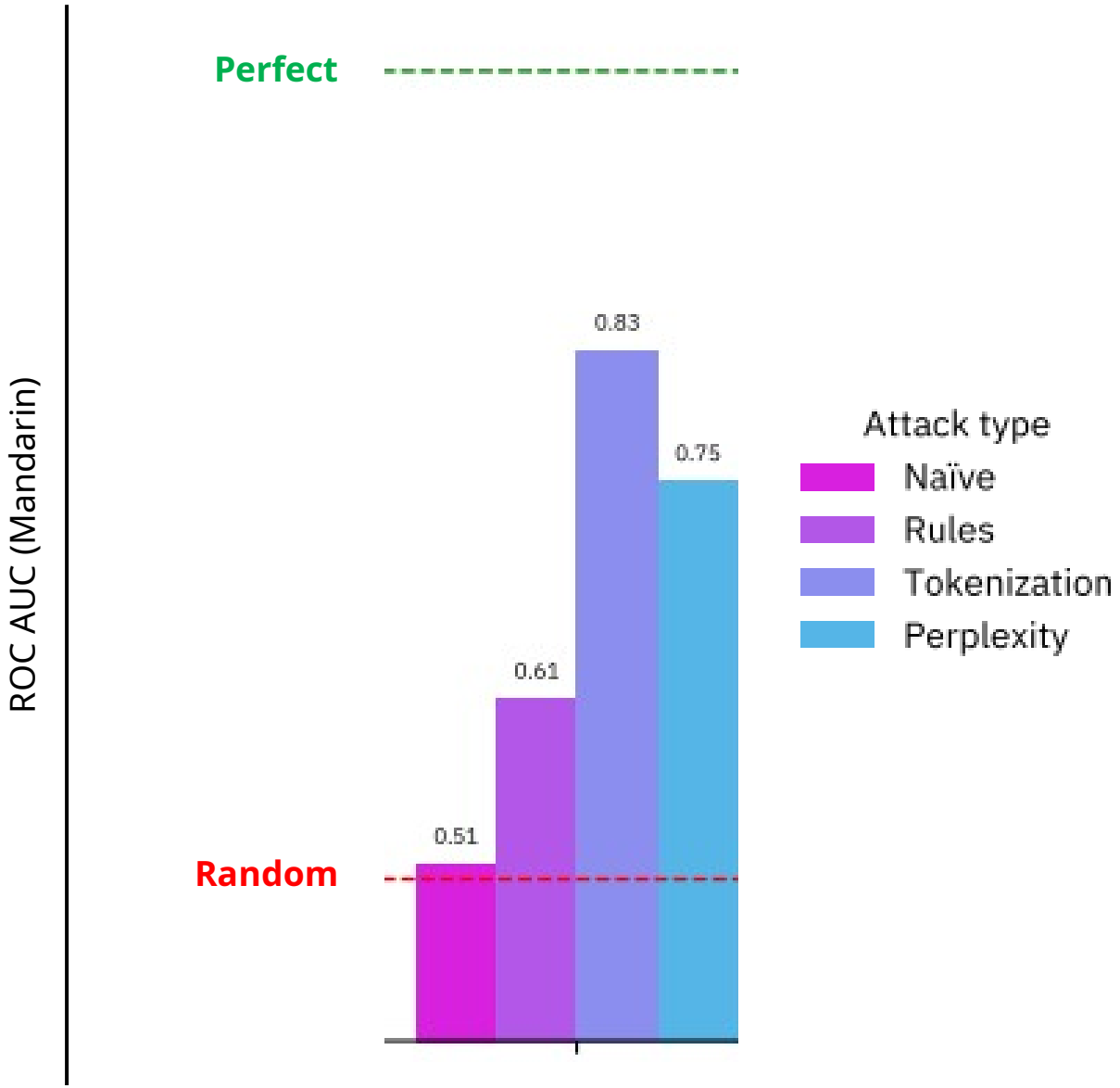
Tokenization or Perplexity?

Tokenization

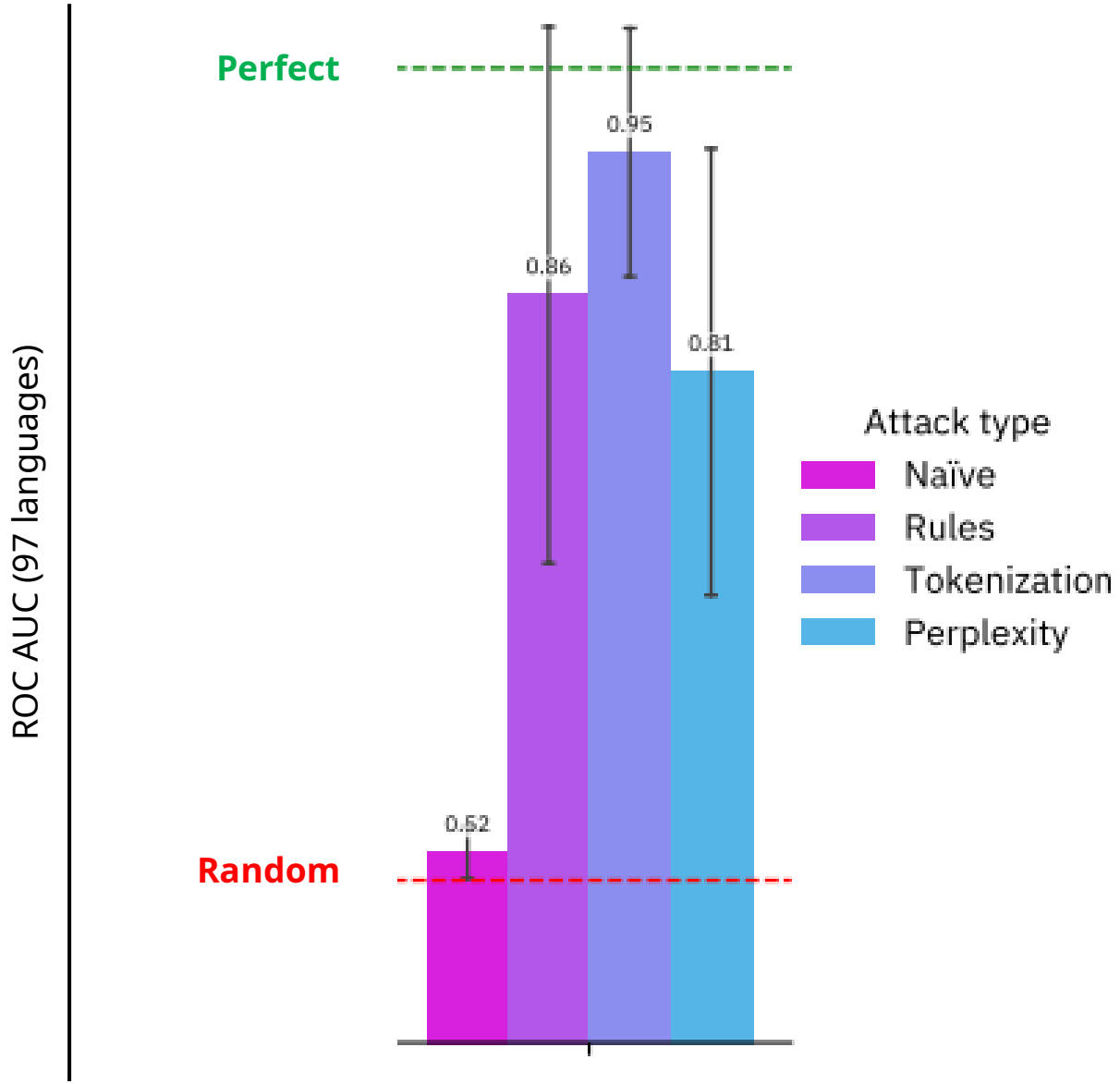


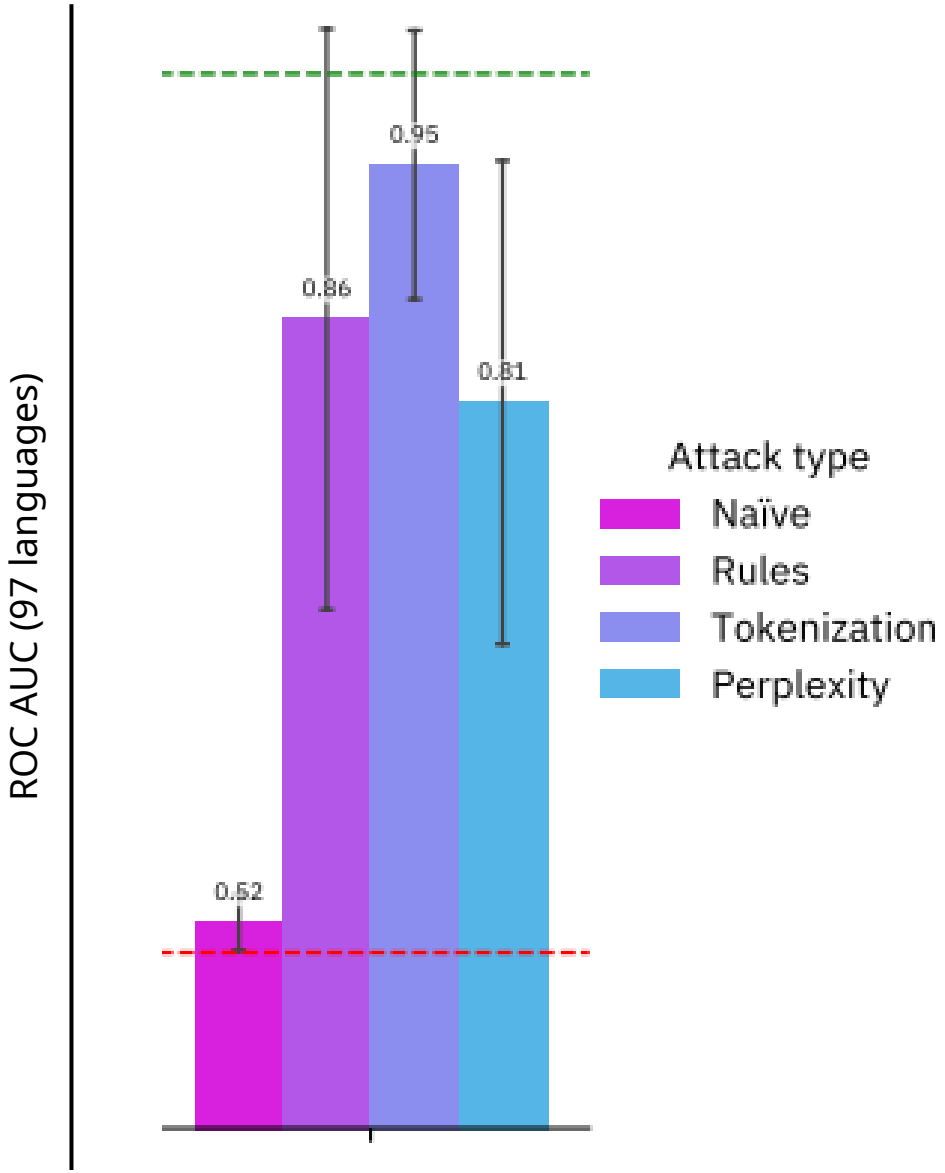
Perplexity

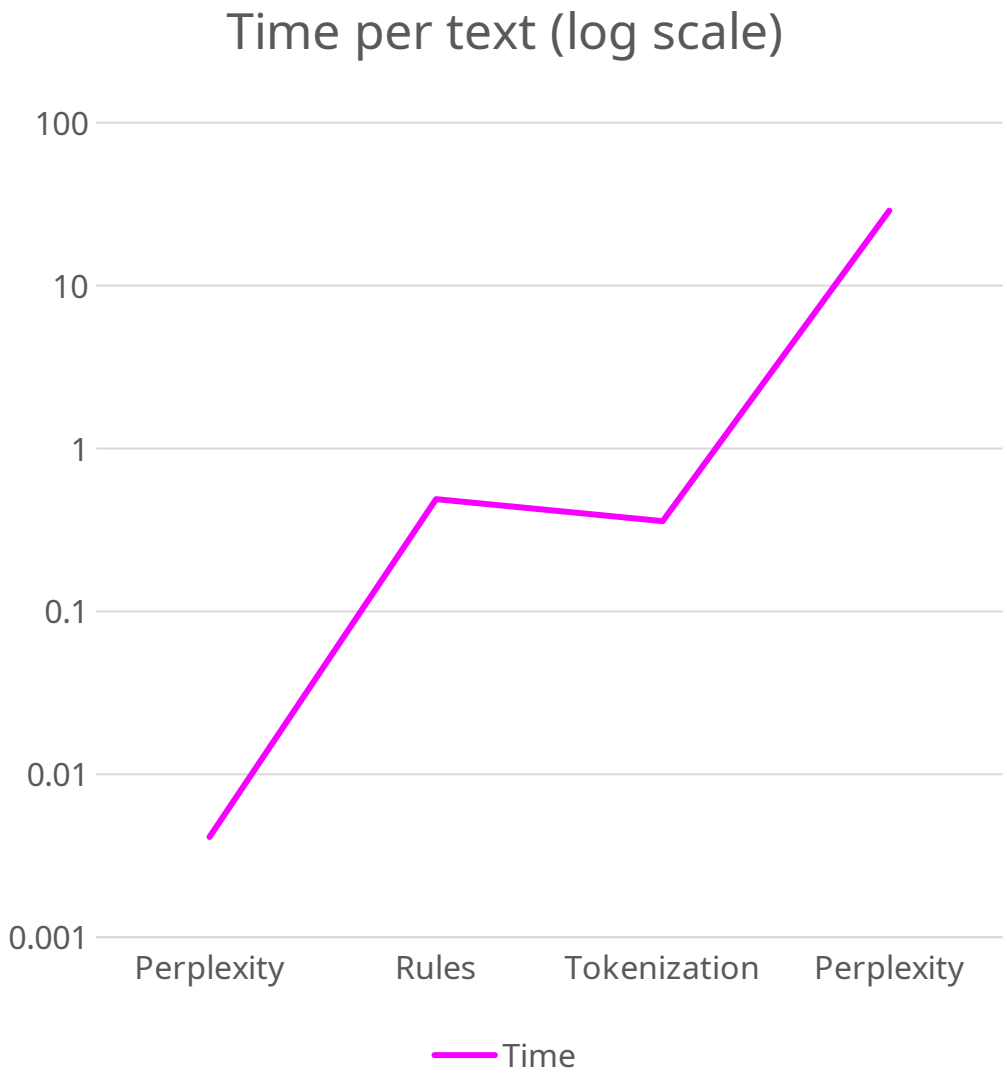


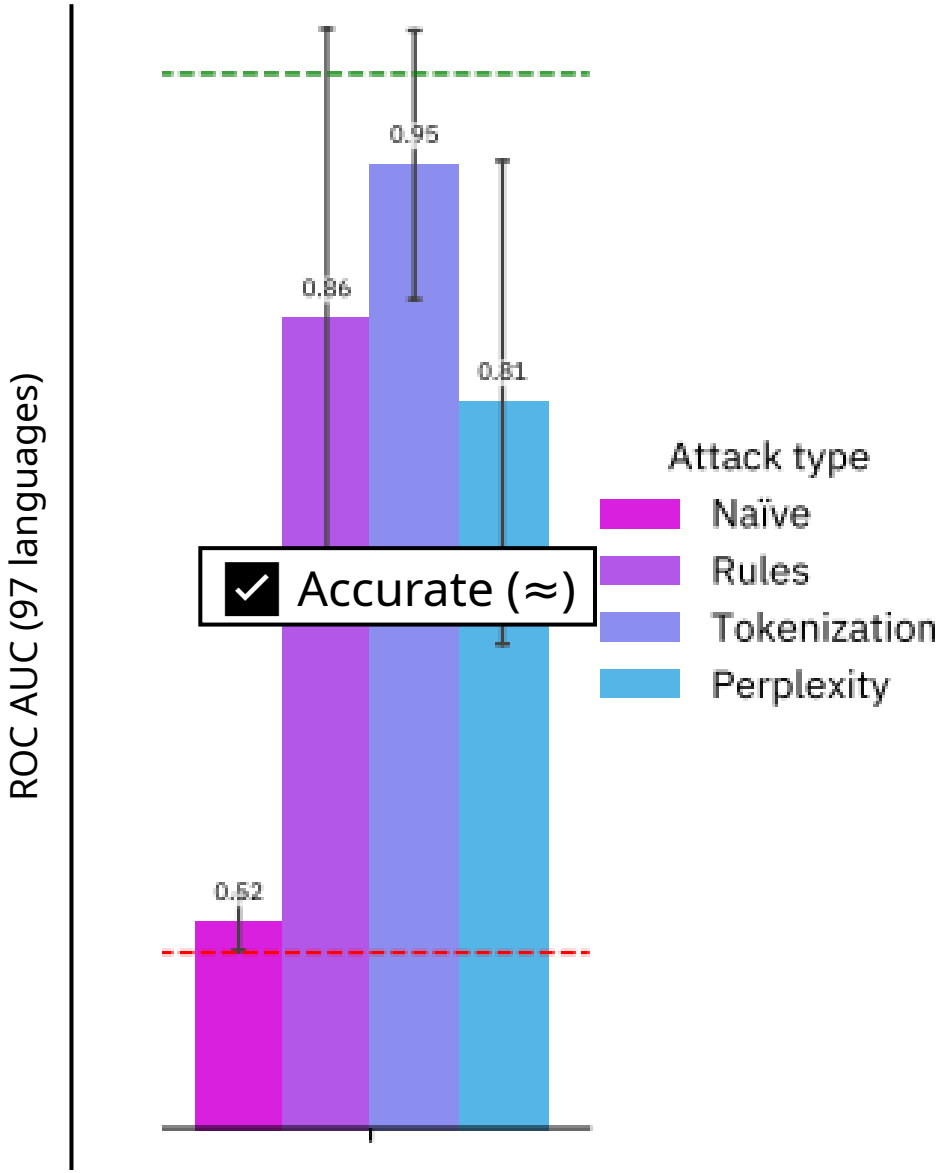

ROC AUC (Mandarin)
Perfect
Random
0.51
0.61
0.83
0.75
Attack type
Naïve
Rules
Tokenization
Perplexity

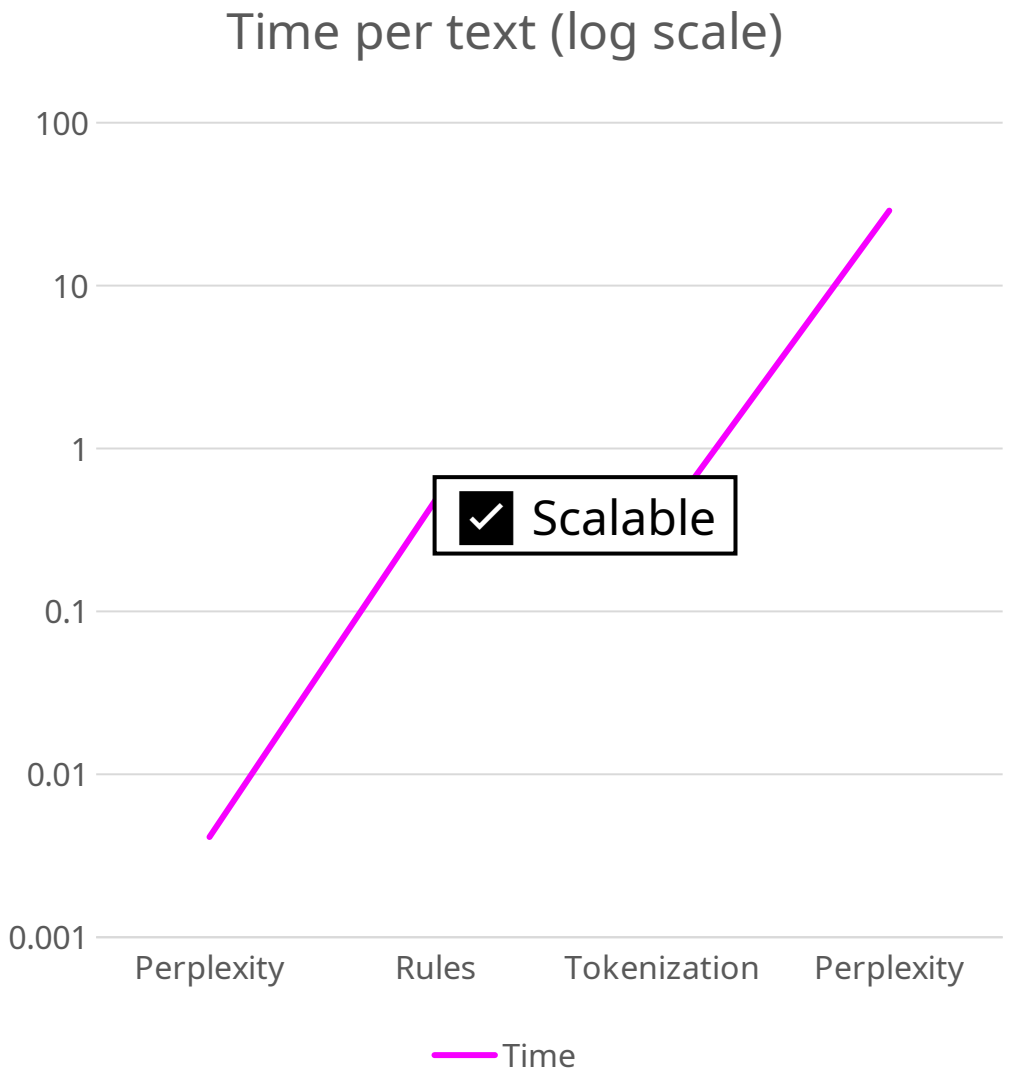
Does this extend to other languages?

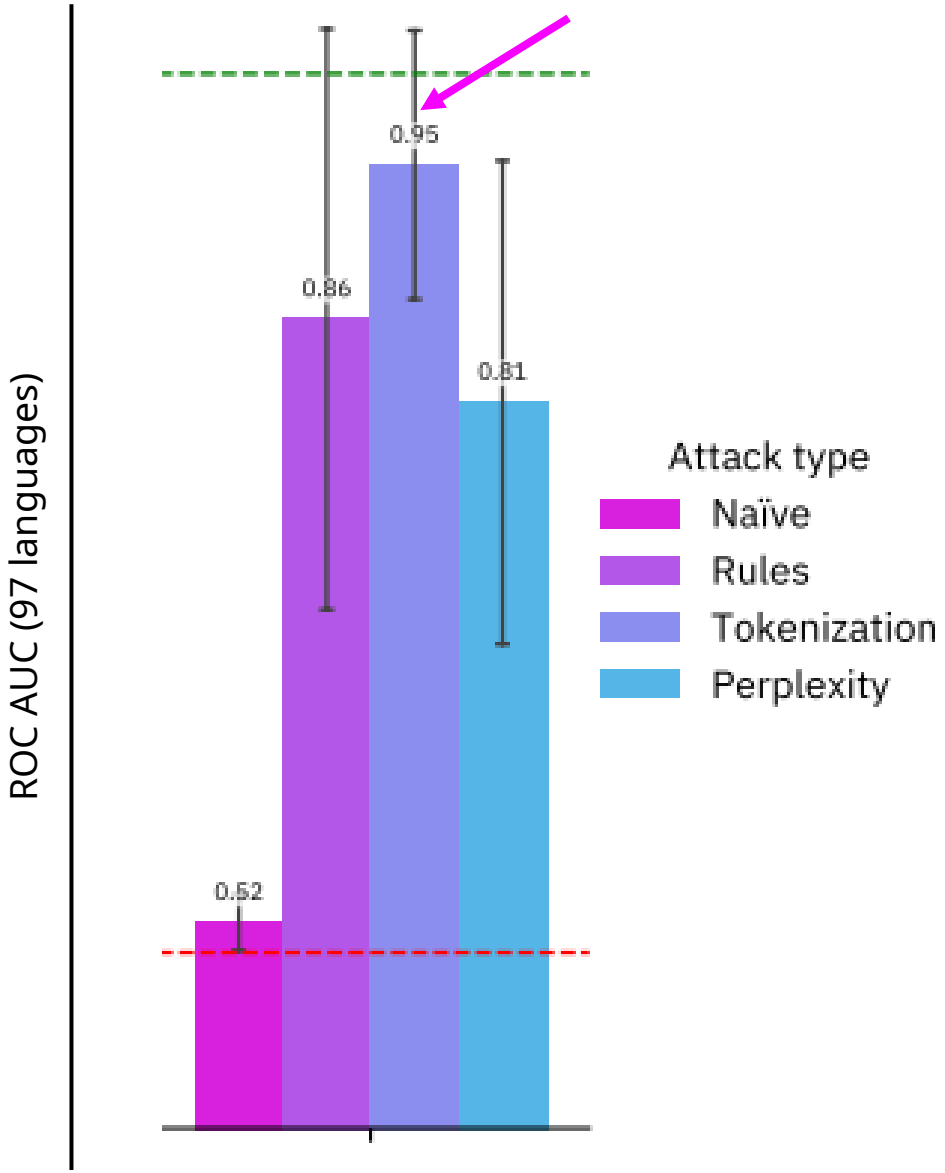

ROC AUC (97 languages)
Perfect
Random
0.52
0.86
0.95
0.81
Attack type
Naïve
Rules
Tokenization
Perplexity

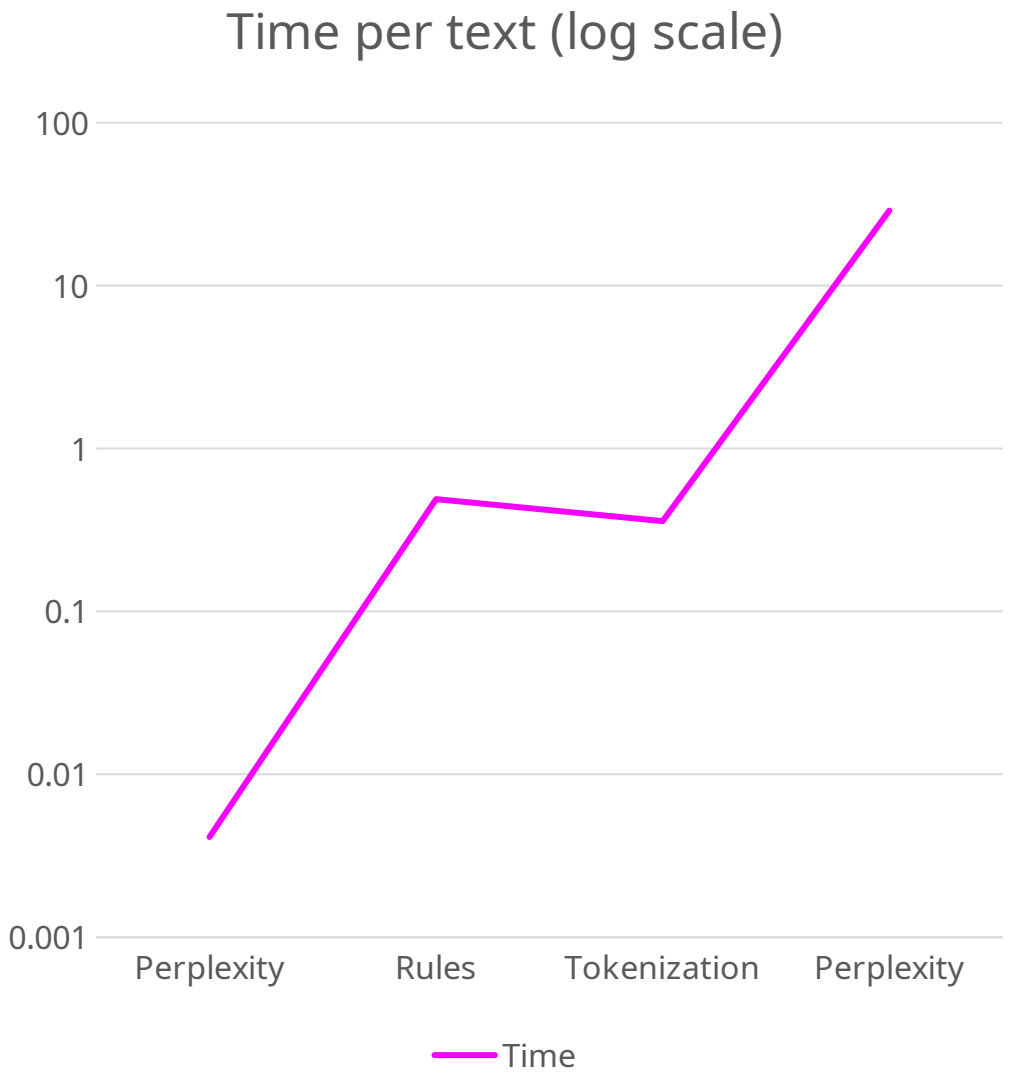

ROC AUC (97 languages)
0.52
0.86
0.95
0.81
Attack type
Naïve
Rules
Tokenization
Perplexity


Time per text (log scale)
100
10
1
0.1
0.01
0.001
Perplexity
Rules
Tokenization
Perplexity
Time


ROC AUC (97 languages)
0.95
0.86
0.81
0.52
Accurate (≈)
Attack type
Naïve
Rules
Tokenization
Perplexity


Time per text (log scale)
100
10
1
0.1
0.01
0.001
Scalable
Perplexity
Rules
Tokenization
Perplexity
Time


ROC AUC (97 languages)
0.95
0.86
0.81
0.52
Attack type
Naïve
Rules
Tokenization
Perplexity


Time per text (log scale)
100
10
1
0.1
0.01
0.001
Perplexity
Rules
Tokenization
Perplexity
Time

Work to do is on us – let’s keep innovating!

Bingo!

Steganography

Hiding data in plain sight
to avoid detection

URL attacks

Click here to reset your password: https://paypal.com

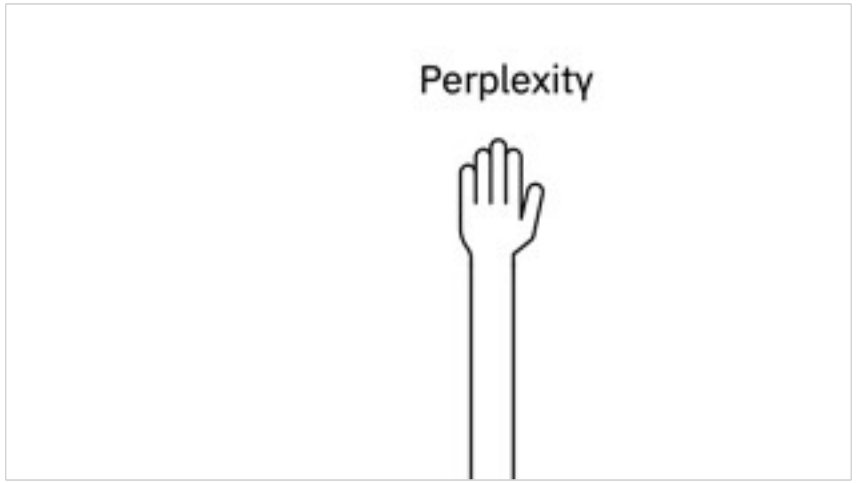
Perplėxity-based detectors

"Pick the N-th word. N is the number of letters in the previous word"

What else can we do?

2 idėjas

Idea 2: Perplexity


Perplexity


Bingo!
CHALLENGE
ż
ą
@acmc | acmc.fyi/intro

Recap

Homoglyphs: characters we can’t discern

Homoglyphs: characters we can’t discern

Detection of AI-generated texts: different techniques

Homoglyphs: characters we can’t discern

Detection of AI-generated texts: different techniques

Homoglyph-based attacks: text becomes unrecognizable

Homoglyphs: characters we can’t discern

Detection of AI-generated texts: different techniques

Homoglyph-based attacks: text becomes unrecognizable

Different mechanisms of action – all exploit “confusion”

Homoglyphs: characters we can’t discern

Detection of AI-generated texts: different techniques

Homoglyph-based attacks: text becomes unrecognizable

Different mechanisms of action – all exploit “confusion”

Highly effective, renders all detectors ineffective

Homoglyphs: characters we can’t discern

Detection of AI-generated texts: different techniques

Homoglyph-based attacks: text becomes unrecognizable

Different mechanisms of action – all exploit “confusion”

Highly effective, renders all detectors ineffective

Lower access barrier: increased risk

Homoglyphs: characters we can’t discern

Detection of AI-generated texts: different techniques

Homoglyph-based attacks: text becomes unrecognizable

Different mechanisms of action – all exploit “confusion”

Highly effective, renders all detectors ineffective

Lower access barrier: increased risk

Some ideas we can try

Homoglyphs: characters we can’t discern

Detection of AI-generated texts: different techniques

Homoglyph-based attacks: text becomes unrecognizable

Different mechanisms of action – all exploit “confusion”

Highly effective, renders all detectors ineffective

Lower access barrier: increased risk

Some safeguards we can try

Work to do is on us – let’s keep innovating!

Work to do is on us – let’s keep innovating!

***get the bibliography & connect!**

https://forms.gle/JzkJTLGNNiCAt4h99