

# Detection of AI writing: Using AI against AI

George Mikros

Hamad Bin Khalifa University, Qatar



# Overview

- 
- Motivation
  - Introduction to Large Language Models
  - Linguistic and stylometric profiling of the AI-writing
    - Human perception of AI writing
    - Linguistic and Statistical characteristics of the LLMs language
  - AI-writing detection
    - Current AI-writing Detectors
    - Reliability stress-tests
  - Conclusion

# Motivation: why AI-writing detection is important for Parliaments

- **Manipulation and Disinformation:** AI-generated texts can be used to spread false information or manipulate public opinion about parliamentary processes or the actions of lawmakers.
- **Impersonation of Public Officials:** AI can create realistic texts impersonating public officials, including lawmakers. This can be used maliciously to make false statements or commitments in the name of a public official, causing confusion or misrepresentation of their positions. For instance, an AI-generated email or social media post might falsely claim that a lawmaker supports a certain policy or has taken a certain action, which can have serious consequences on their reputation and the political process.
- **Automated Lobbying Efforts:** AI-generated texts can be employed in automated lobbying efforts, where numerous messages are sent to lawmakers to influence their opinions or votes on specific issues. Such messages might be indistinguishable from genuine constituent communications, making it difficult for lawmakers to discern real public opinion from manufactured campaigns. This could skew the perceived public sentiment on important issues, and potentially lead lawmakers to make decisions based on manipulated data.



GPT-3

The first (nearly...) AGI model

Natural Language Processing Model

Non-deterministic

NLP task-agnostic

Requires minimum fine-tuning

# What makes GPT-3 so magical?

## It is really big

With 175 billion parameters, it's the largest language model ever created

## Minimum fine-tuning

It only requires few-shot demonstrations via textual interaction with the model.

## Extraordinary - Supermodel

You can ask GPT-3 to be a translator, a programmer, a poet, or a famous author, and it can do it with its user (you) providing fewer than 10 training examples.

## Custom language tasks without training

Task-agnostic NLP model

# GPT-3

## a few-shot learner

- 
- **Size matters:** GPT-3 shows that language model performance scales as a power-law of model size, dataset size, and the amount of computation.
  - **Universal Learning Machine:** GPT-3 demonstrates that a language model trained on enough data can solve NLP tasks that it has never encountered. That means that GPT-3 studies the model as a general solution for many downstream jobs without fine-tuning.
  - **Expensive:** The cost of AI is increasing exponentially. Training GPT-3 would cost over \$4.6M using a Tesla V100 cloud instance.
  - **Black box:** Although there is a clear performance gain from increasing the model capacity, it is not clear what is really going on under the hood. Especially, it remains a question of whether the model has learned to do reasoning, or simply memorizes training examples in a more intelligent way.

# GPT-3 training data

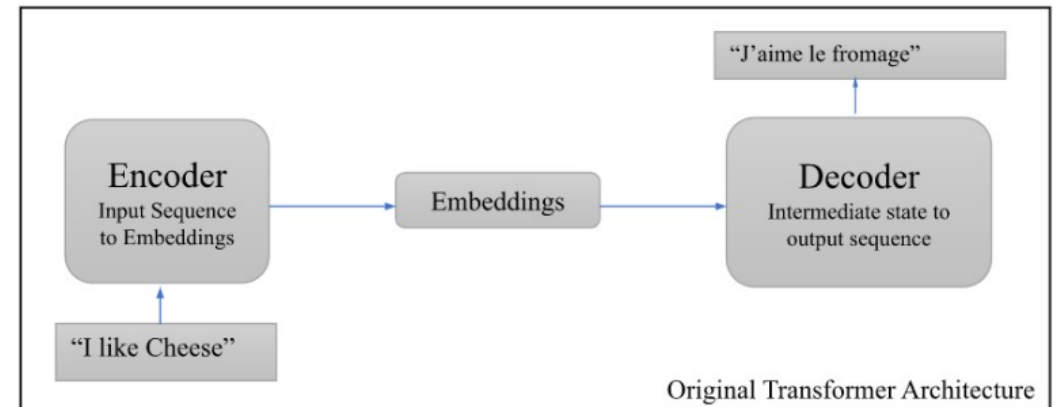
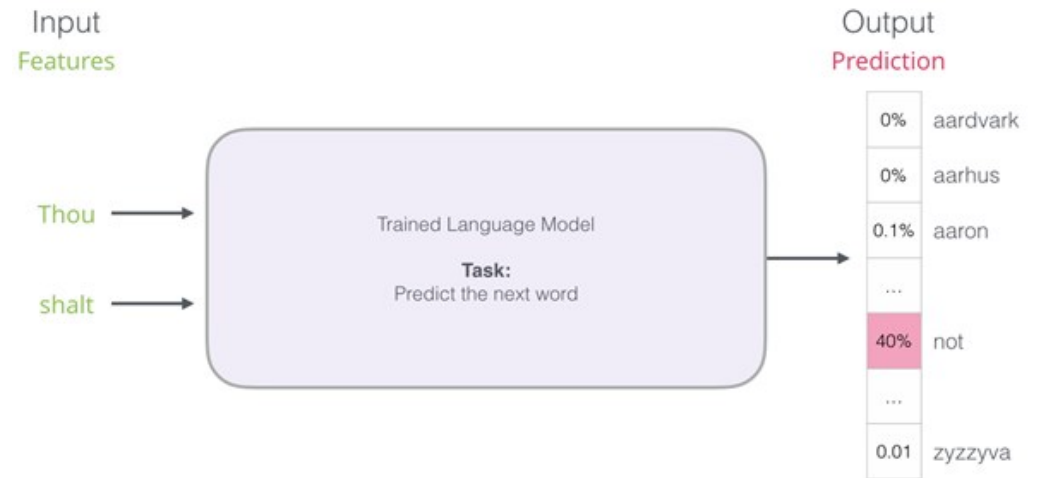
---

Dataset	# Tokens (Billions)
Total	499
Common Crawl (filtered by quality)	410
WebText2	19
Books1	12
Books2	55
Wikipedia	3

- It is a weighted mix of Common Crawl, WebText2 (a larger version of the original), two book corpora, and English Wikipedia.
- Some components (e.g., Wikipedia), were completely sampled 3+ times during training, while others, like the Common Crawl, weren't even completely sampled. The authors claim that this is to help raise the overall quality of the corpus by prioritizing known-good datasets.
- Altogether, the filtered/cleaned dataset is 500 billion tokens or 700GB.

# How GPT-3 works?

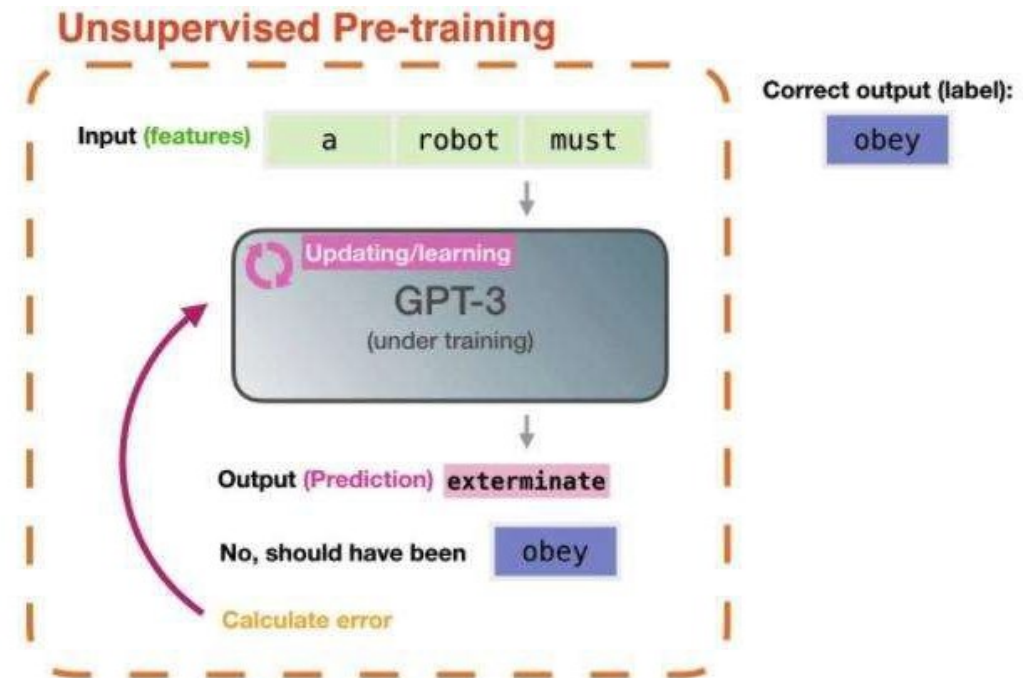
- GPT- 3 language model generates text. We can pass some text as input and generate text as output.
- The output generated by a language model like ChatGPT is based on patterns and relationships learned from the vast amounts of text it was trained on.
- The purpose of this type of models is to be able to predict a word or sequence of words given a text. The text is used as input, and the model generates a probability distribution over the dictionary of the words it knows and chooses based on it.
- It is important to note that the model doesn't have personal opinions, emotions, or consciousness, and its responses are limited to what it was trained on.





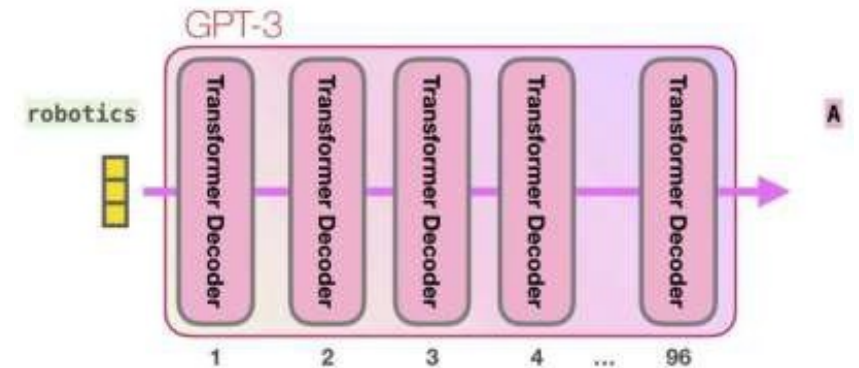
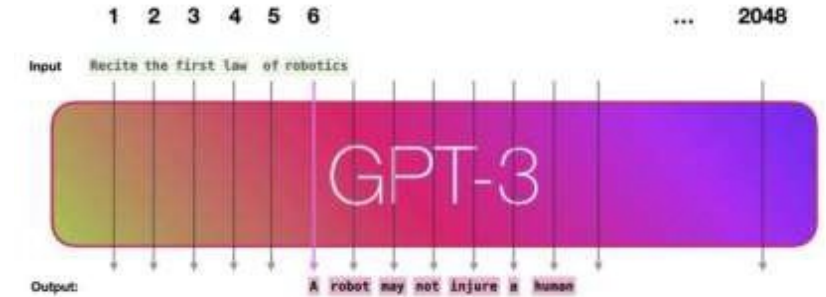
# How GPT-3 works?

- It works by analyzing and understanding patterns in massive text data. This process is called pre-training (<https://jalammar.github.io/how-gpt3-works-visualizations-animations/>).
- During this pre-training, GPT-3 is fed a large corpus of text and learns to predict the next word in a sentence based on the context of the previous words. This pre-training stage helps GPT-3 develop a general understanding of the language and its structure.
- Once the pre-training is complete, GPT-3 can then be fine-tuned for specific tasks, such as answering questions, translating text, or generating creative writing. The fine-tuning process involves training GPT-3 on smaller, task-specific datasets to perform even better for those specific tasks.
- **ChatGPT is, in a sense, a fine-tuning application of the GPT-3 model.**



# How GPT-3 works?

- GPT3 actually generates output one token at a time (let's assume a token is a word).
- The important calculations of the GPT3 occur inside its stack of 96 transformer decoder layers. This is the “depth” of “deep learning”.
- Each of these layers has its own 1.8B parameter to make its calculations. That is where the “magic” happens.
- GPT3 is 2048 tokens-wide. That is its “context window.” That means it has 2048 tracks along which tokens are processed.



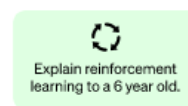
# What is ChatGPT?

- A GPT-3 model that has been trained to interact conversationally and now belongs to the GPT 3.5 series.
- The dialogue format makes it possible for ChatGPT to answer follow up questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.
- The model was trained using **Reinforcement Learning from Human Feedback (RLHF)**.
  - Human AI trainers provided conversations in which they played both sides—the user and an AI assistant. The trainers had access to model-written suggestions to help them compose their responses and transform the exchanges into a dialogue format.
  - They created a reward model for reinforcement learning. To collect this data, conversations that AI trainers had with the chatbot were used. They randomly selected a model-written message, sampled several alternative completions, and had AI trainers rank them. Using these reward models, they fine-tune the model using Proximal Policy Optimization. This process was repeated many times.

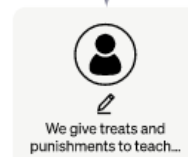
## Step 1

Collect demonstration data and train a supervised policy.

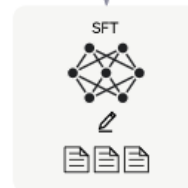
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



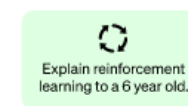
This data is used to fine-tune GPT-3.5 with supervised learning.



## Step 2

Collect comparison data and train a reward model.

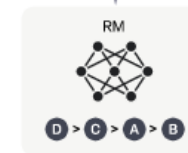
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



## Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.



The PPO model is initialized from the supervised policy.



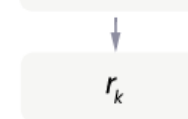
The policy generates an output.



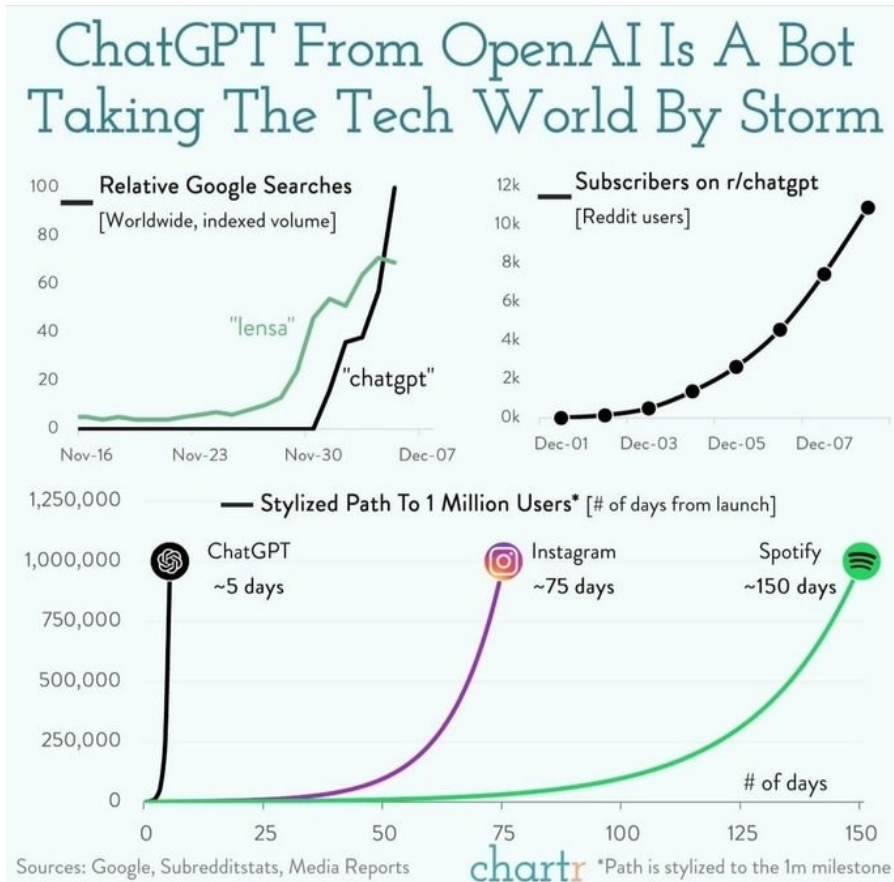
The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



# ChatGPT adoption rates: Fastest ever recorded in the history of digital platforms



## ChatGPT Sprints to One Million Users

Time it took for selected online services to reach one million users




\* one million backers \*\* one million nights booked \*\*\* one million downloads

Source: Company announcements via Business Insider/LinkedIn



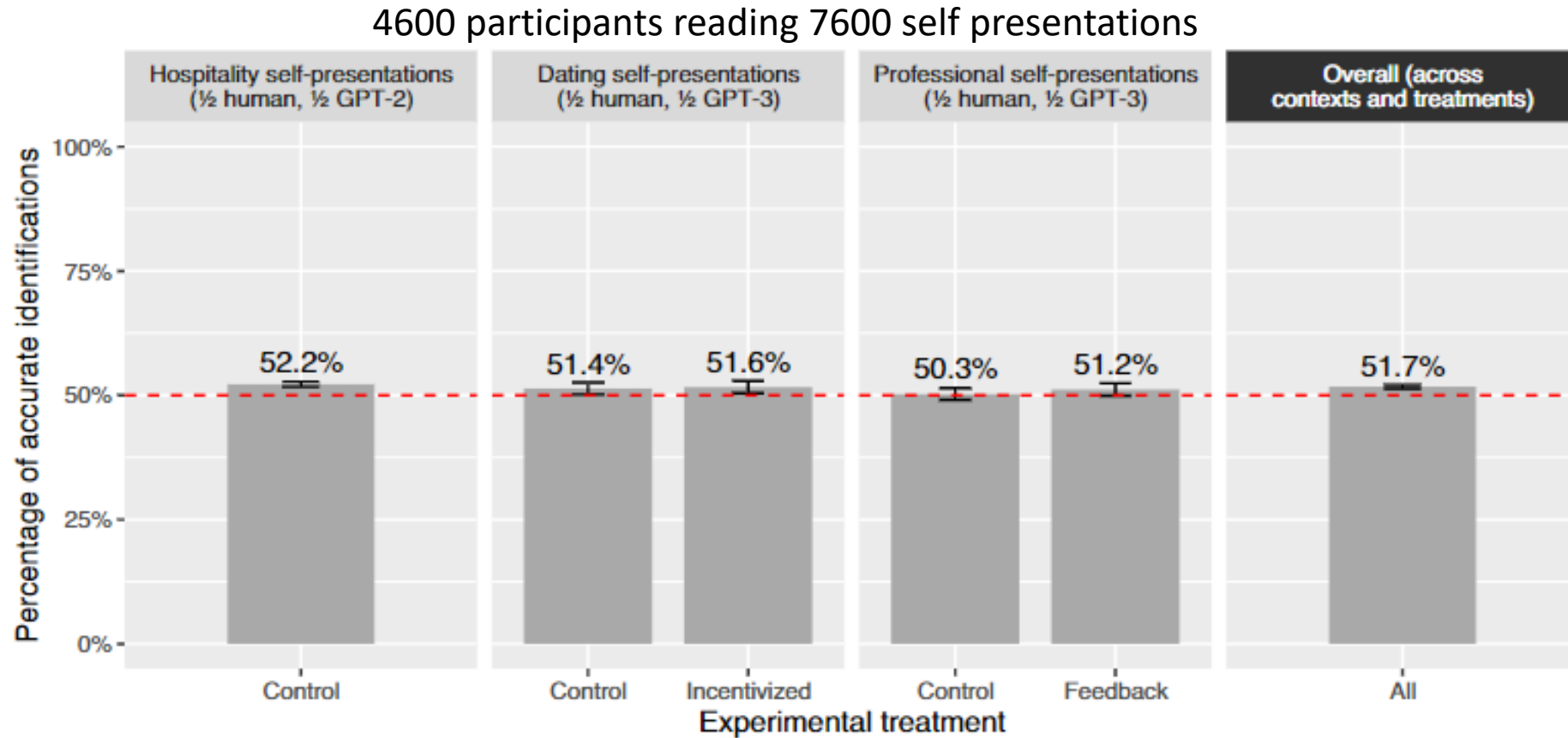
statista



## Summary of key capabilities, limitations, and concerns around ChatGPT and other LLMs

Capabilities	Limitations	Concerns
<ul style="list-style-type: none"><li>• It can write plausible sounding text on any topic.</li><li>• It can generate answers to a range of questions, including coding, maths-type problems and multiple choice.</li><li>• It is getting increasingly accurate and sophisticated with each release.</li><li>• It generates unique text each time you use it.</li><li>• It's great at other tasks like text summarisation.</li></ul>	<ul style="list-style-type: none"><li>• It can generate plausible but incorrect information.</li><li>• ChatGPT is only trained on information up until Sept 2021 (but those with the paid ChatGPT Plus service have access to a version that can access the internet)</li><li>• Limited ability to explain the sources of information for its responses (this varies between Chatbots)</li></ul>	<ul style="list-style-type: none"><li>• It can and does produce biased output (culturally, politically etc)</li><li>• It can generate unacceptable output.</li><li>• It has a high environmental impact, concerns around human impact and ownership of training material.</li><li>• Security and privacy concerns around the way users' data is used to train the models.</li><li>• There is a danger of digital inequity.</li></ul>

# Can humans realize if a text has been written by AI?



Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11), e2208839120. <https://doi.org/doi:10.1073/pnas.2208839120>



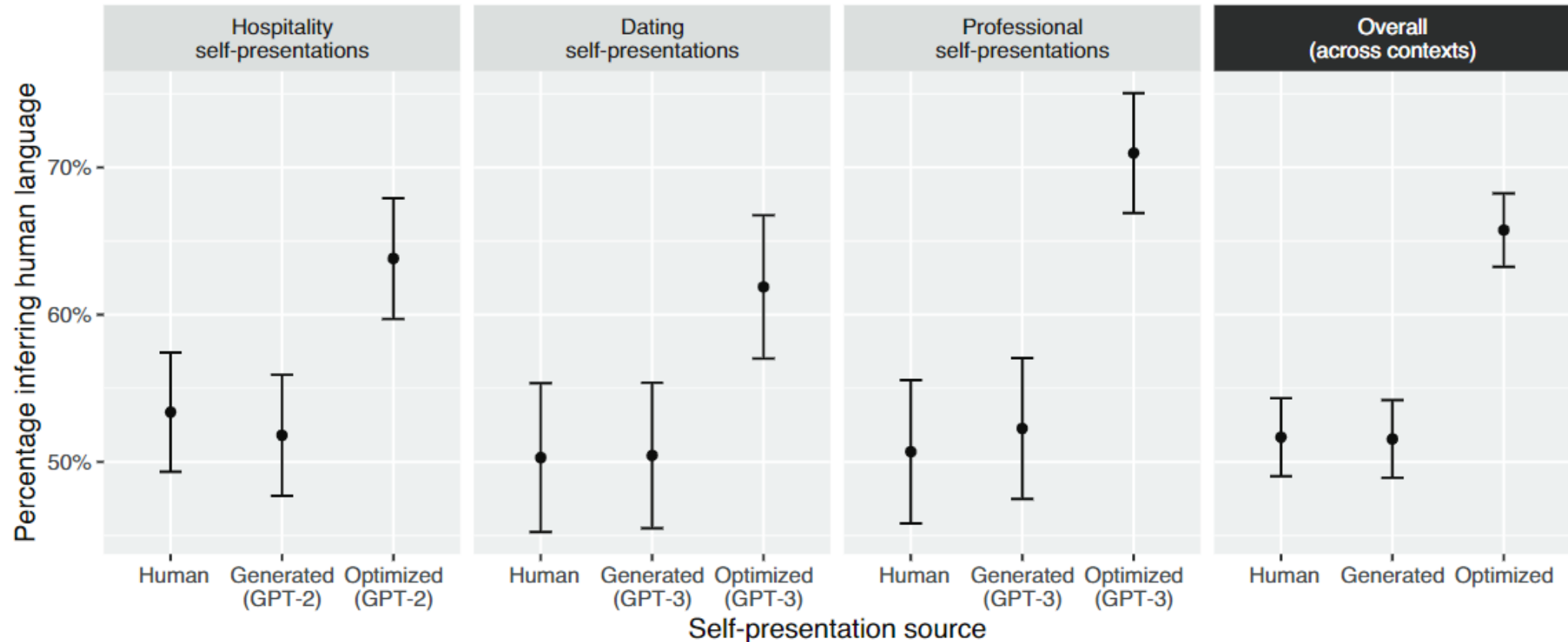
# Features that make humans believe a text is written by AI

	Dependent variable	
	(1) Perceived as AI-generated (odds ratios with 95% CI)	(2) Actually AI-generated (odds ratios with 95% CI)
<u>Aligned features</u>		
Nonsensical content <sup>†</sup>	1.105 <sup>***</sup> (1.085, 1.126)	1.233 <sup>***</sup> (1.169, 1.296)
Repetitive content <sup>†</sup>	1.083 <sup>***</sup> (1.059, 1.106)	1.470 <sup>***</sup> (1.379, 1.561)
Conversational words	0.947 <sup>***</sup> (0.925, 0.970)	0.898 <sup>**</sup> (0.829, 0.967)
<u>Misaligned features</u>		
Grammatical issues <sup>†</sup>	1.048 <sup>***</sup> (1.028, 1.069)	0.851 <sup>***</sup> (0.788, 0.913)
Rare bigrams	1.042 <sup>***</sup> (1.019, 1.065)	0.666 <sup>***</sup> (0.596, 0.736)
Long words	1.034 <sup>***</sup> (1.009, 1.059)	0.783 <sup>***</sup> (0.706, 0.861)
Contractions	0.947 <sup>***</sup> (0.924, 0.970)	1.134 <sup>***</sup> (1.065, 1.203)
<u>Nonindicative</u>		
Second-person pronouns	1.059 <sup>***</sup> (1.038, 1.079)	0.970 (0.908, 1.032)
Filler words	1.009 (0.990, 1.027)	1.119 <sup>*</sup> (1.021, 1.218)
Swear words	0.969 <sup>**</sup> (0.948, 0.989)	0.965 (0.905, 1.024)
Authentic words	0.946 <sup>***</sup> (0.921, 0.971)	0.945 (0.870, 1.021)
Focus on past	0.938 <sup>***</sup> (0.917, 0.959)	1.002 (0.940, 1.064)
First-person pronouns	0.925 <sup>***</sup> (0.886, 0.963)	0.992 (0.868, 1.117)
Family words	0.910 <sup>***</sup> (0.889, 0.932)	1.014 (0.950, 1.077)
Word count	0.904 <sup>***</sup> (0.874, 0.935)	1.076 (0.986, 1.165)
Constant	0.850 <sup>***</sup> (0.830, 0.870)	1.007 (0.947, 1.068)
Observations	38,866	4,690
Log likelihood	-26,318.460	-3,029.542
Akaike Inf. Crit.	52,670.930	6,093.085

Note: <sup>†</sup>manually labeled feature, <sup>\*</sup>p<sup>\*\*\*</sup> P < 0.001.

Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11), e2208839120. <https://doi.org/doi:10.1073/pnas.2208839120>

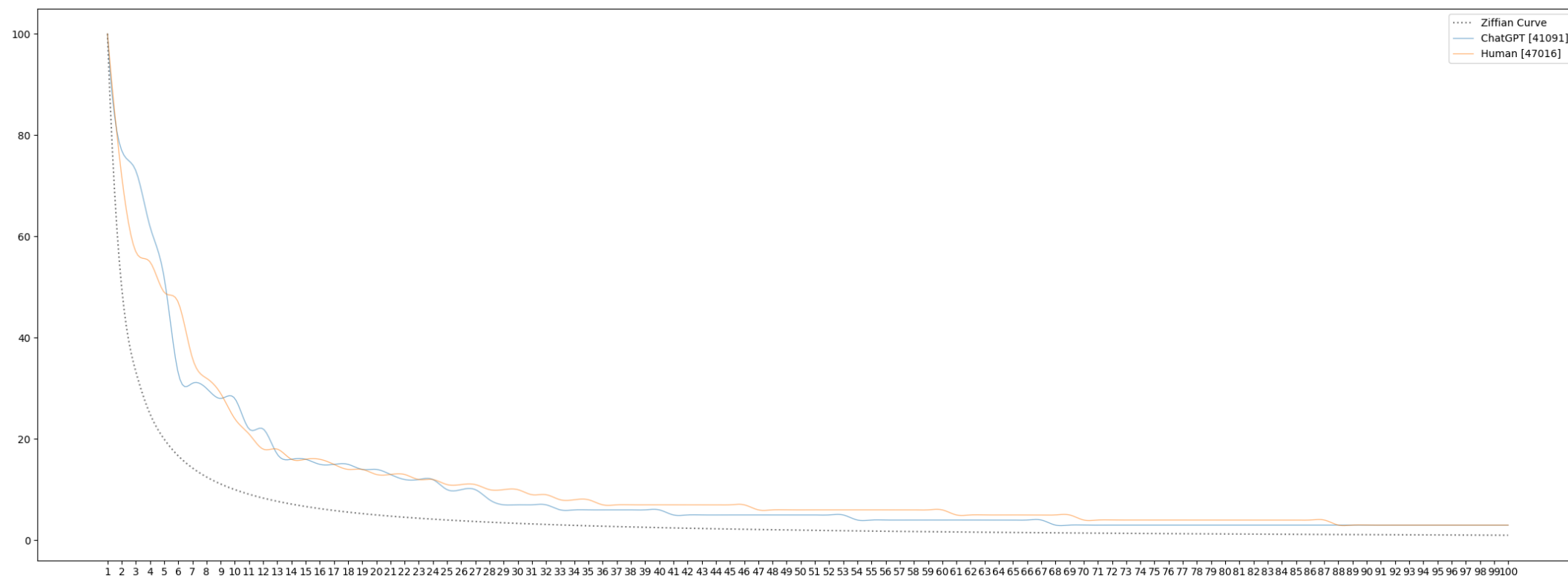
# AI models can be taught to sound more “human” than human.



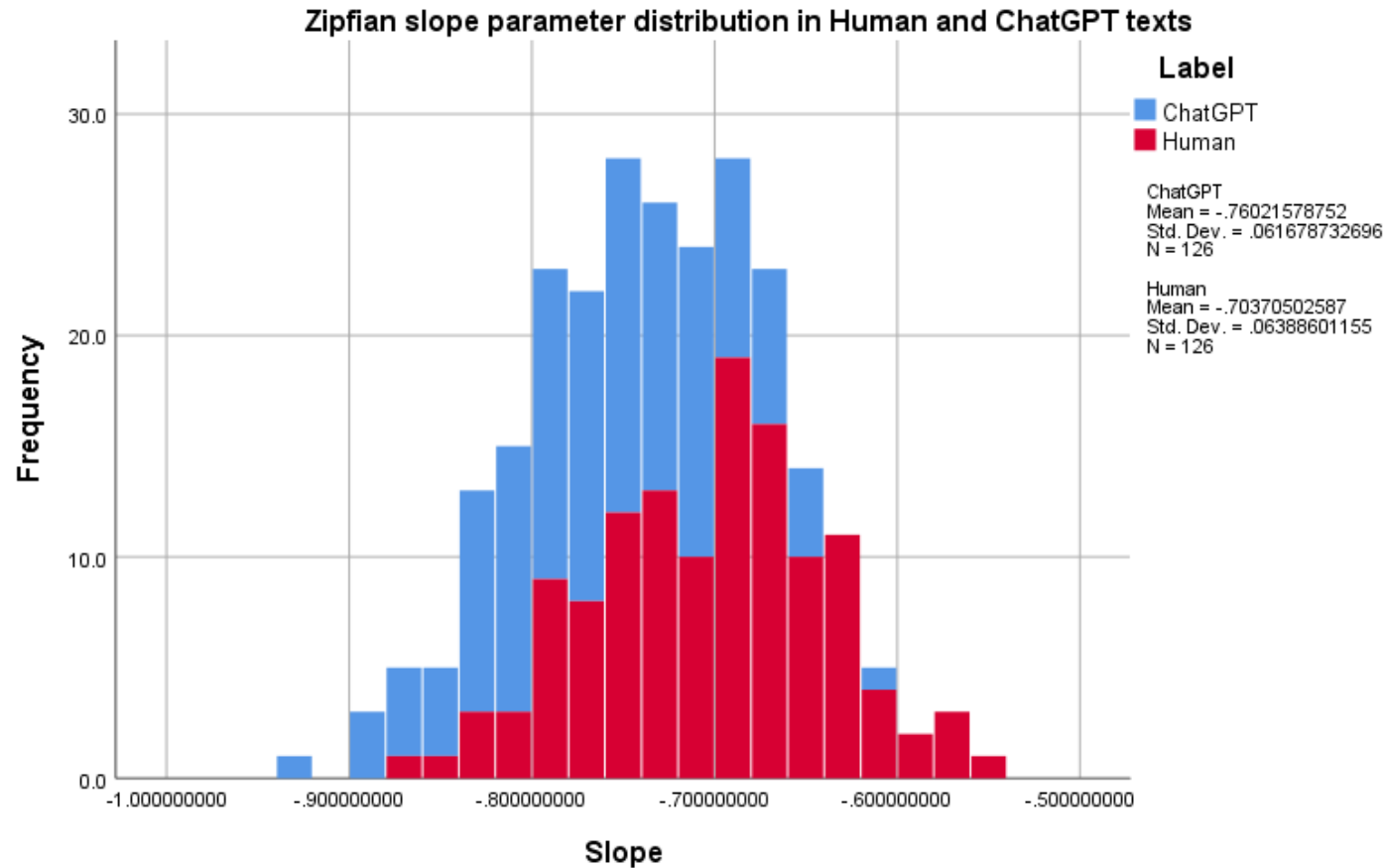
Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11), e2208839120. <https://doi.org/doi:10.1073/pnas.2208839120>



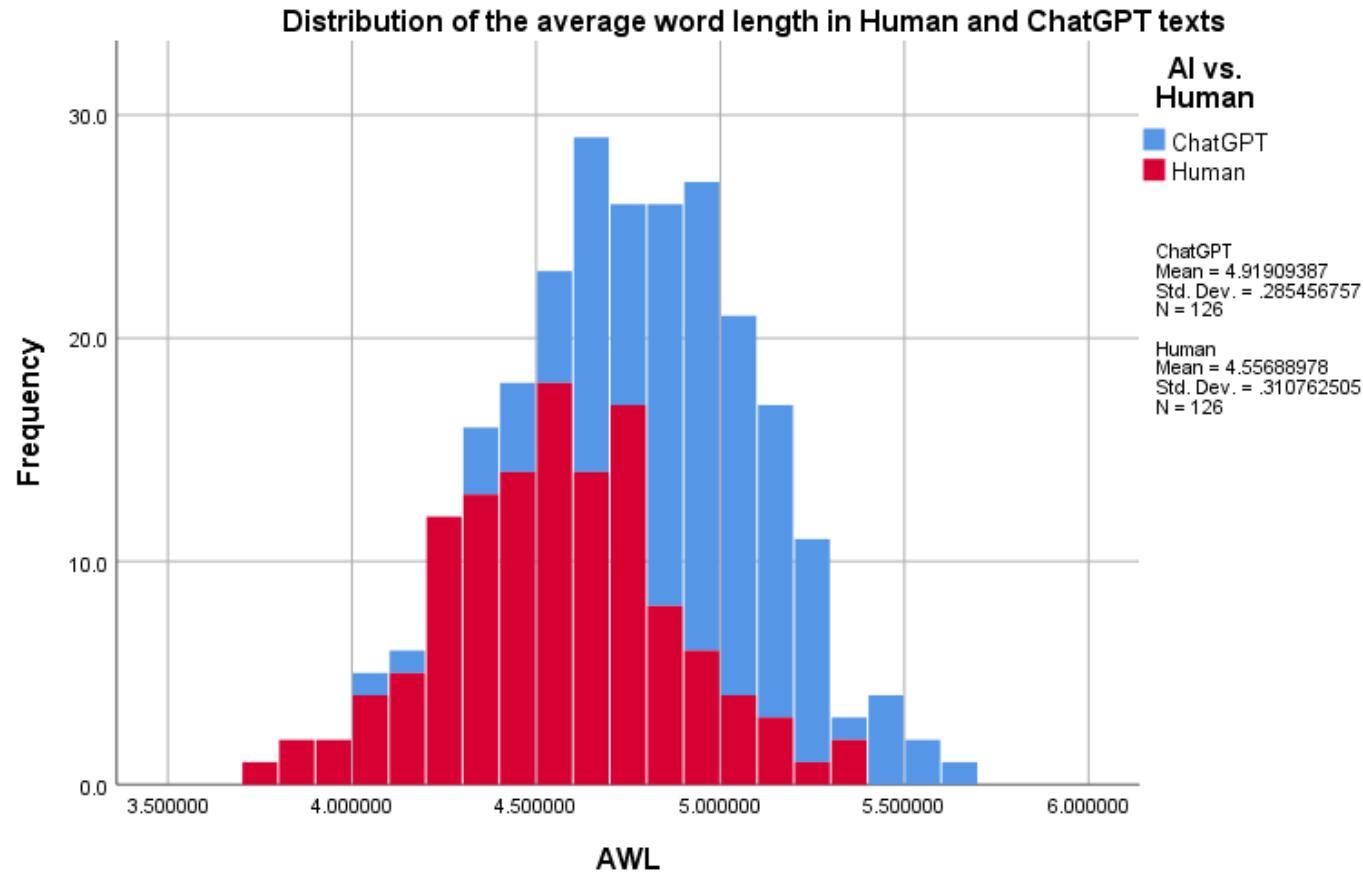
# Statistical Characteristics of Language: Zipfian fit



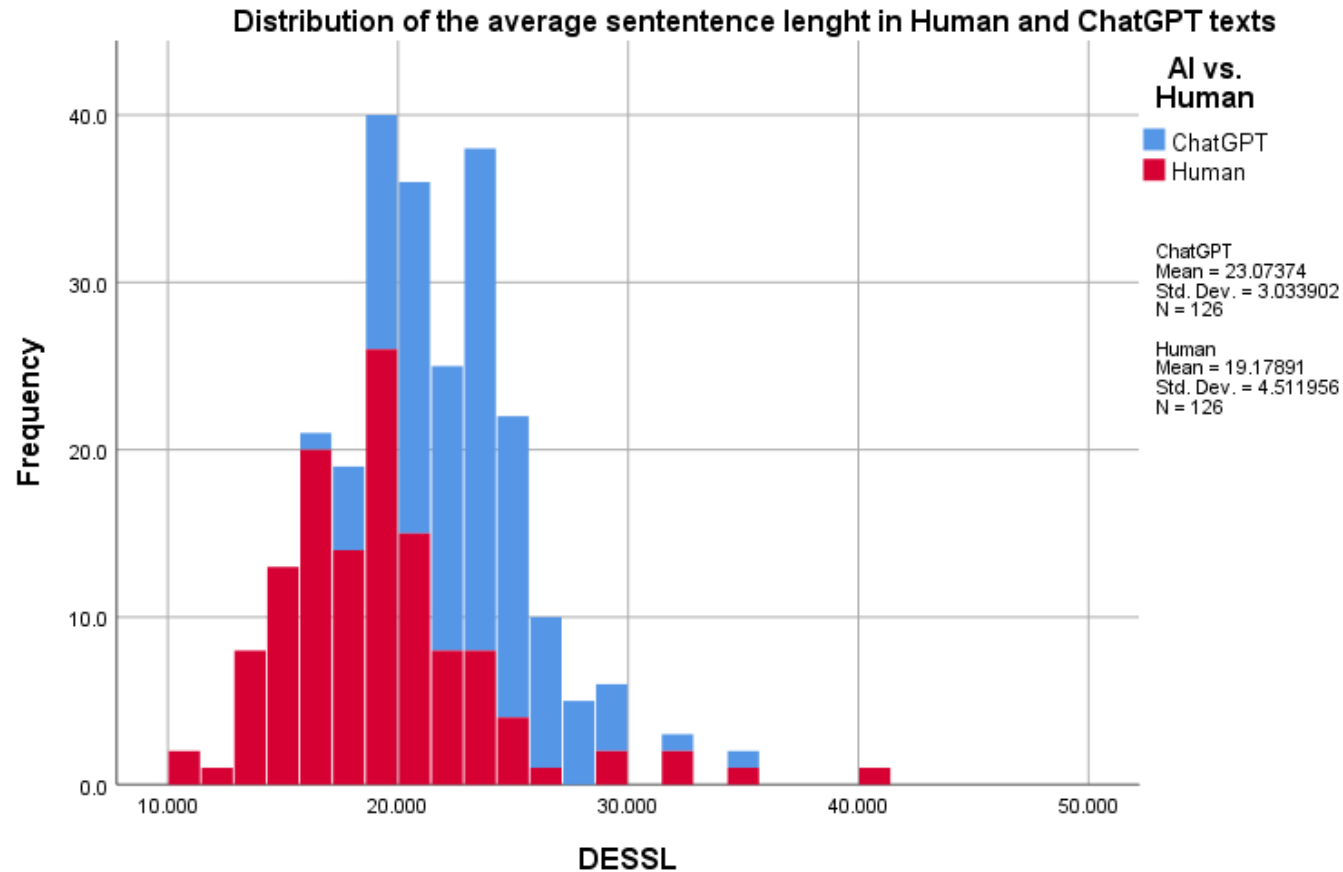
# Statistical Characteristics of Language: Zipfian fit



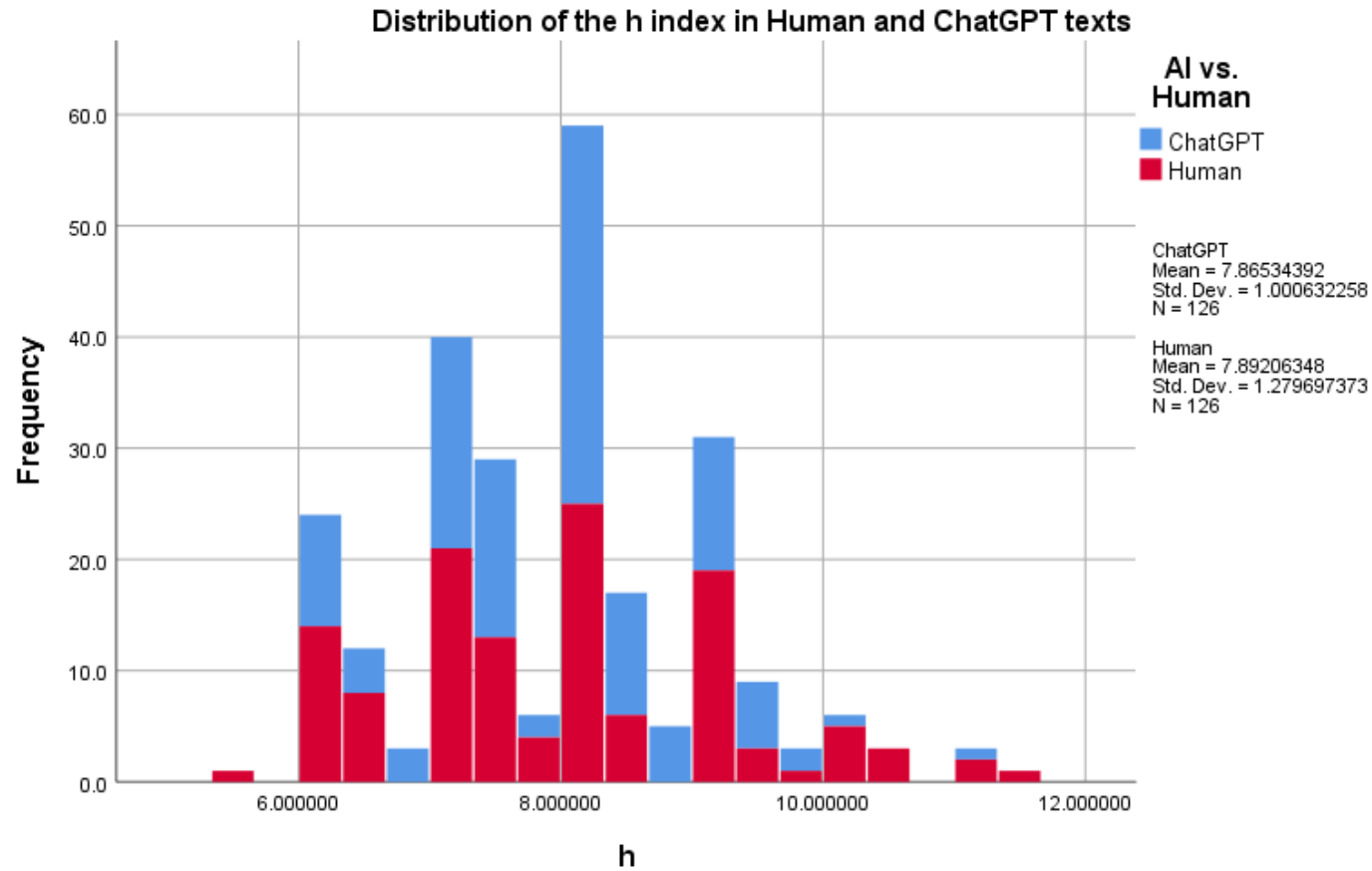
# Statistical Characteristics of Language: Average Word Length



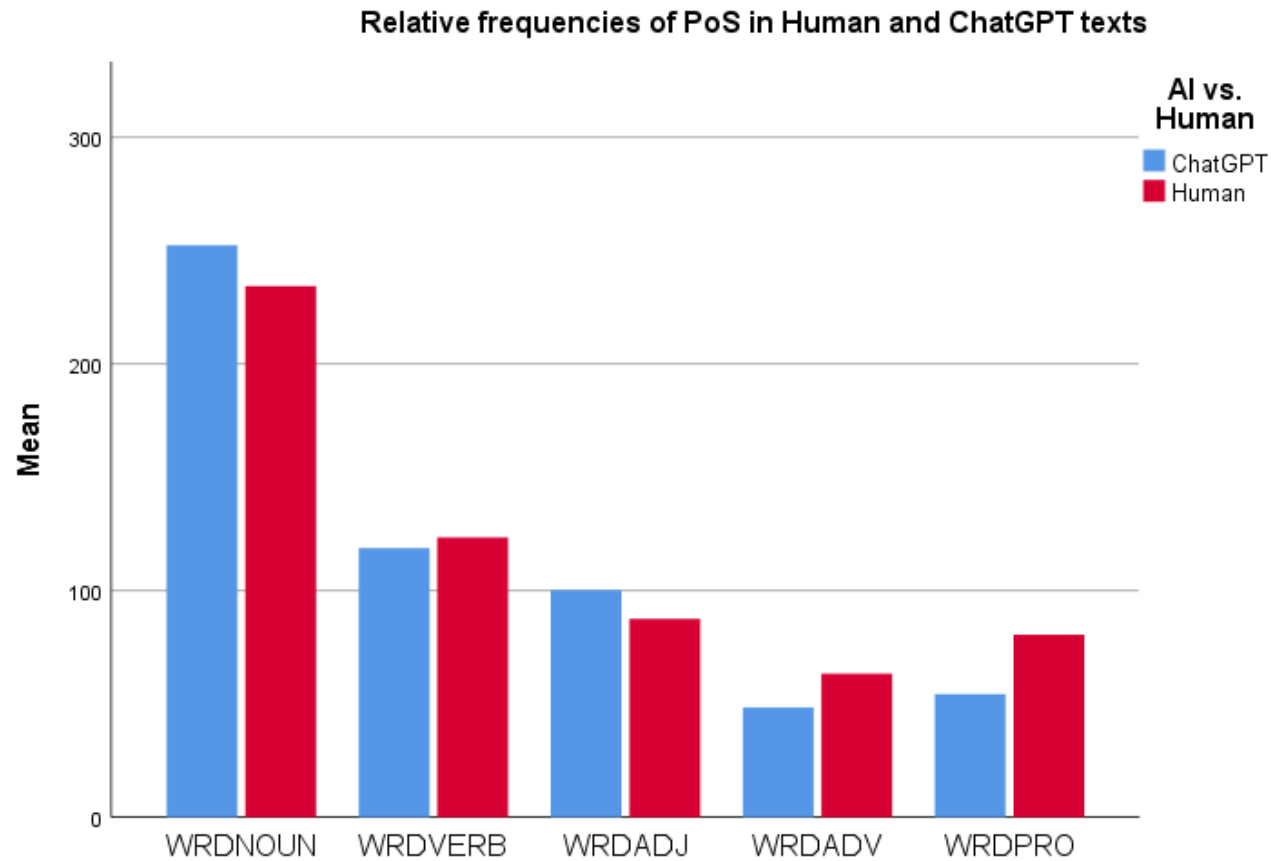
# Statistical Characteristics of Language: Average Sentence Length



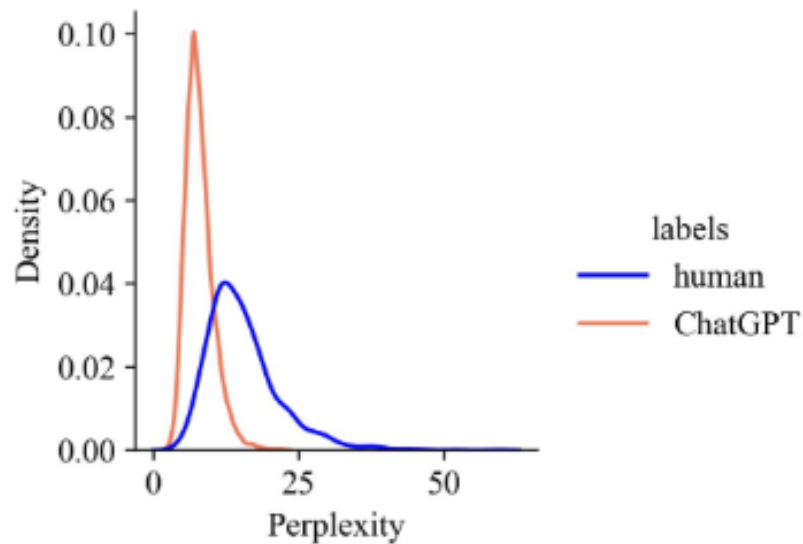
# Statistical Characteristics of Language: h index



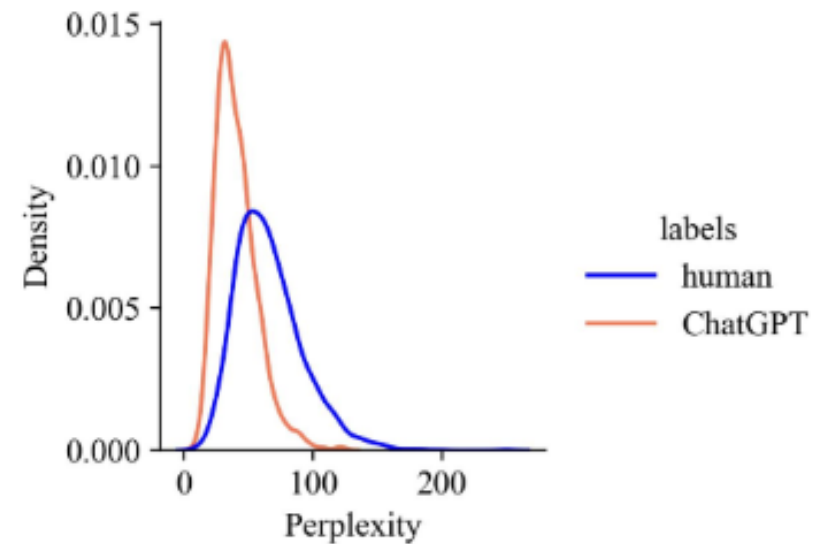
# Statistical Characteristics of Language: PoS relative frequencies



# Features discriminating AI-writing: Perplexity



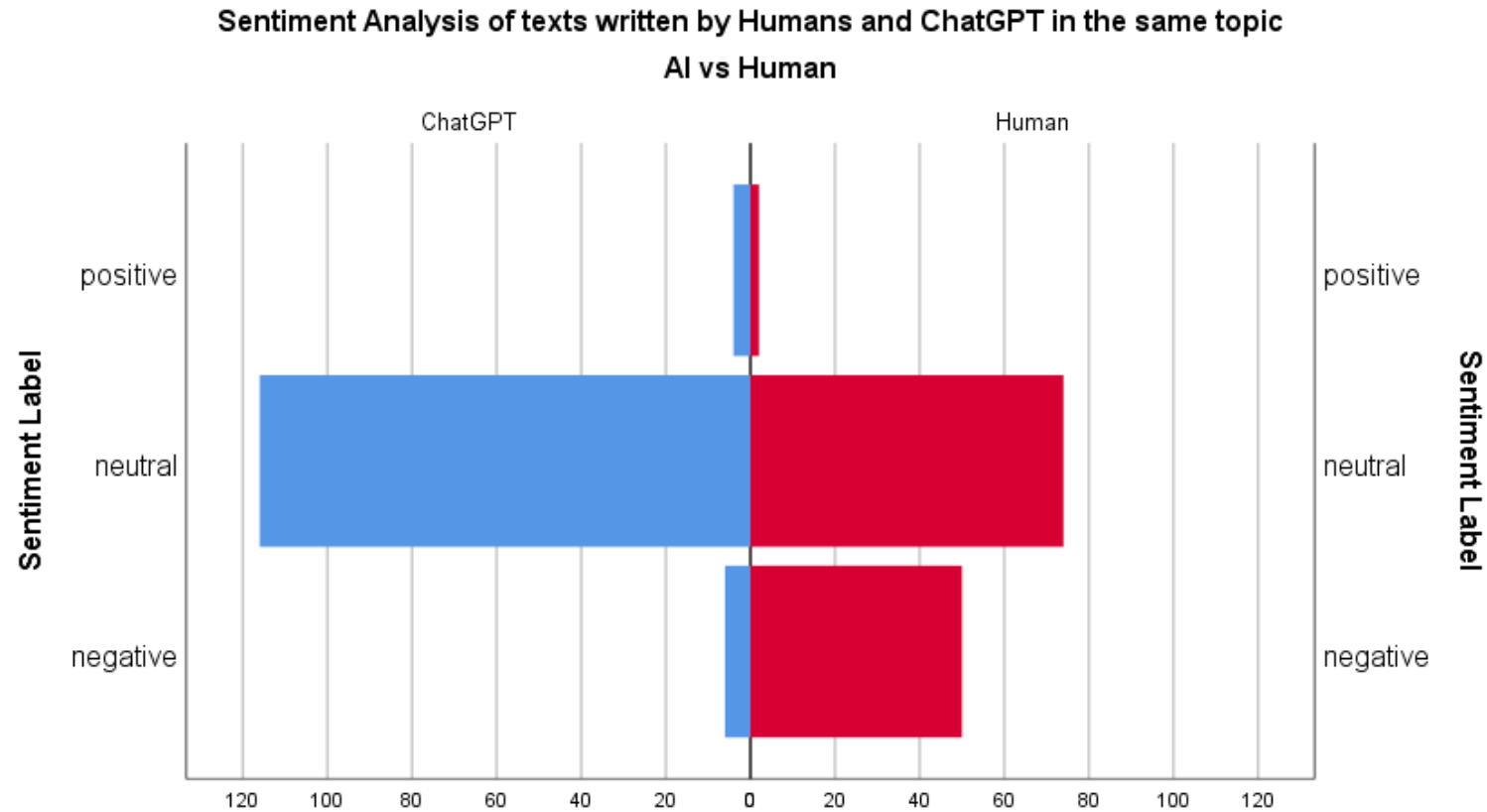
A. Text perplexity of medical abstract



B. Text perplexity of radiology report

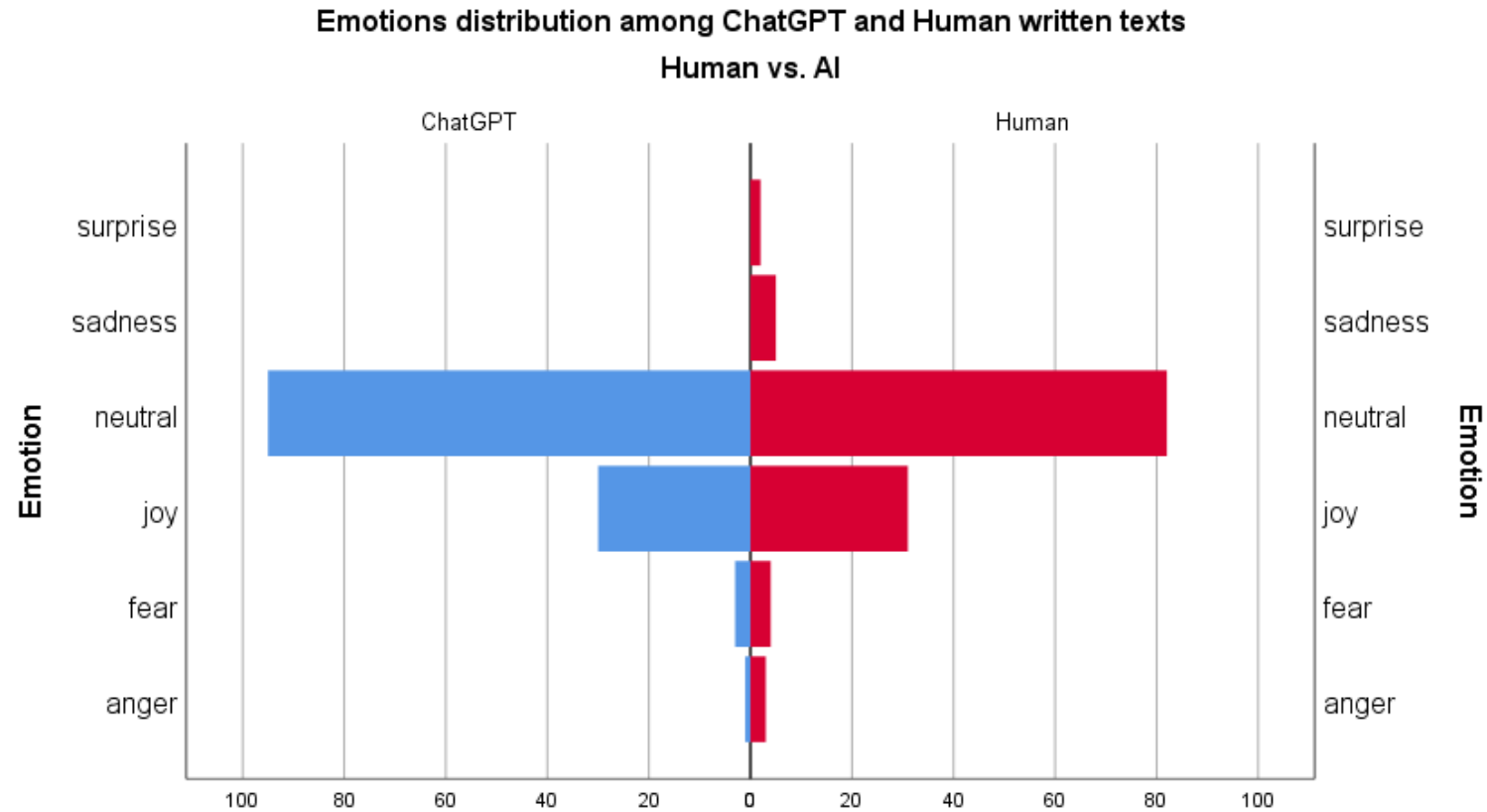
Liao, W., Liu, Z., Dai, H., Xu, S., Wu, Z., Zhang, Y., Huang, X., Zhu, D., Cai, H., Liu, T., & Li, X. (2023). Differentiate ChatGPT-generated and Human-written Medical Texts. *arXiv pre-print server*. <https://doi.org/None> arxiv:2304.11567

# Features discriminating AI-writing: Sentiment

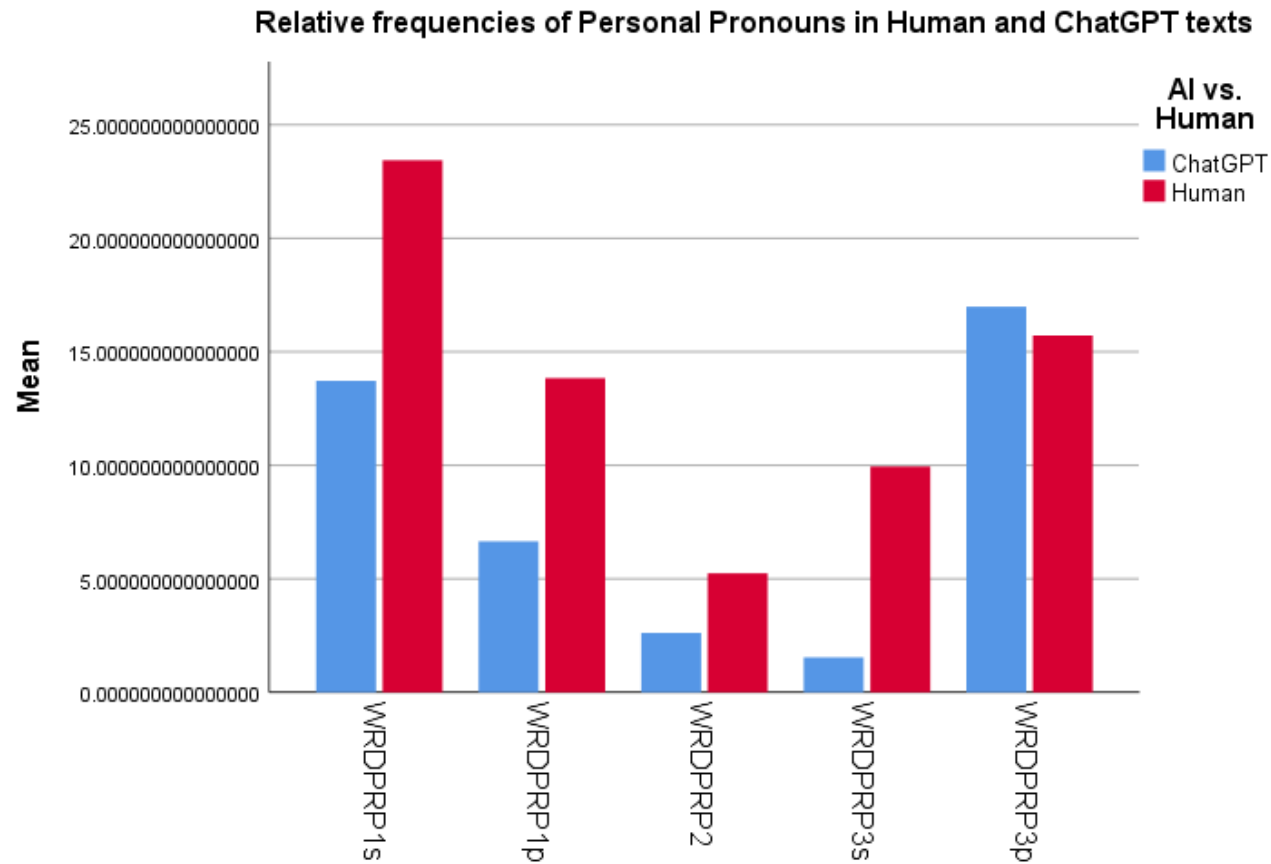




# Features discriminating AI-writing: Emotions

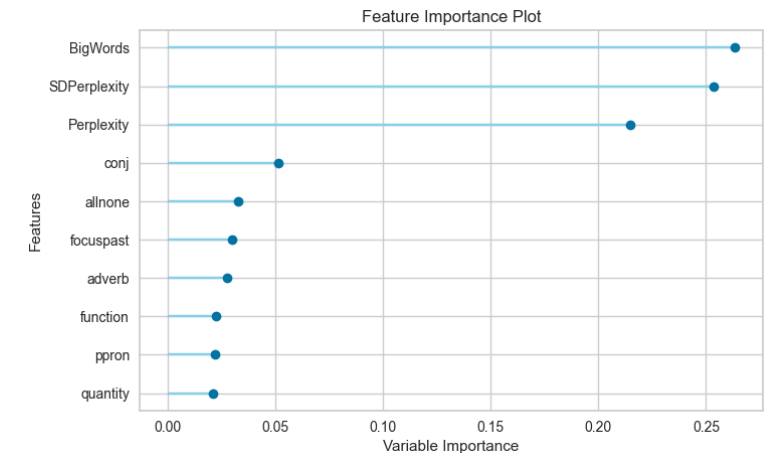
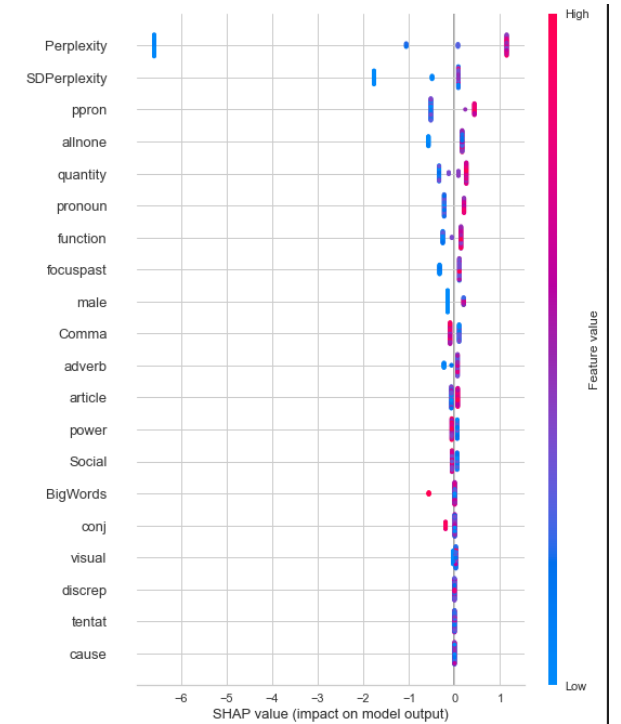


# Features discriminating AI-writing: Personal Pronouns



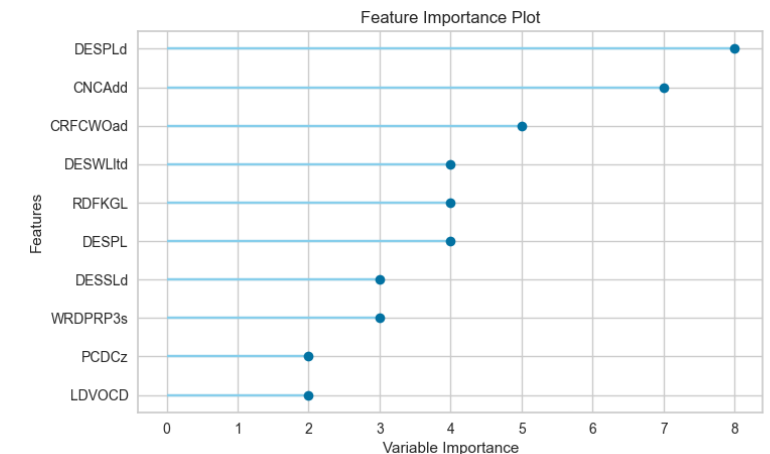
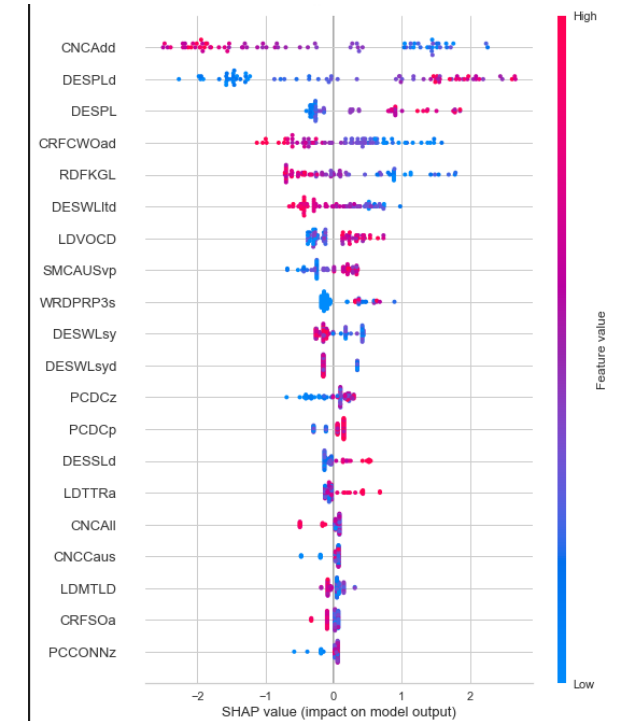
# Features discriminating AI-writing: Quantitative Text Indices [1]

- BigWords: % of words 7 characters or longer [AI+]
- SDPerplexity: Standard deviation of the perplexity [AI-]
- Perplexity [AI-]
- conj: Conjunctions [AI+]
- allnone: All or none (all, no, never, always) [AI-]
- focuspast: Past focus (was, had, were, been) [AI-]
- adverb: Adverbs [AI-]
- function: Total function words [AI-]
- ppron: Personal pronouns [AI-]
- quantity: Quantities (all, one, more, some) [AI-]



# Features discriminating AI-writing: Quantitative Text Indices [2]

- DESPLd: SD of the mean length of paragraphs [AI-]
- CNCAAdd: Additive connectives (“and,” “moreover”) [AI+]
- CRFCWOad: Content word overlap [AI+]
- DESWLltd: SD of the mean number of characters in words [AI+]
- RDFKGL: Flesch-Kincaid Grade Level [AI+]
- DESPL: Mean length of paragraphs (in sentences) [AI-]
- DESSLd: SD of the mean length of sentences [AI-]
- WRDPRP3s: Third-person singular pronoun [AI-]
- PCDCz: Deep cohesion. This dimension reflects the degree to which the text contains causal and intentional connectives when there are causal and logical relationships within the text [AI-]
- LDVOCd: Lexical Diversity. VOCD [AI-]



# Summary of the AI vs. Human discriminating features

Features that display higher values in AI-written texts [AI+]	Features that display lower values in AI-written texts [AI-]
BigWords: % of words 7 characters or longer	Perplexity (mean and SD)
Frequency of conjunctions	Paragraph length (mean and SD)
Additive connectives	SD of the average sentence length
Content words overlap	Frequency of adverbs and personal pronouns
SD of the average word length	Past focus
Flesch-Kincaid Grade Level	Quantities and contrasts of quantities (all or none)
	Deep cohesion
	Lexical Diversity (VOCD)

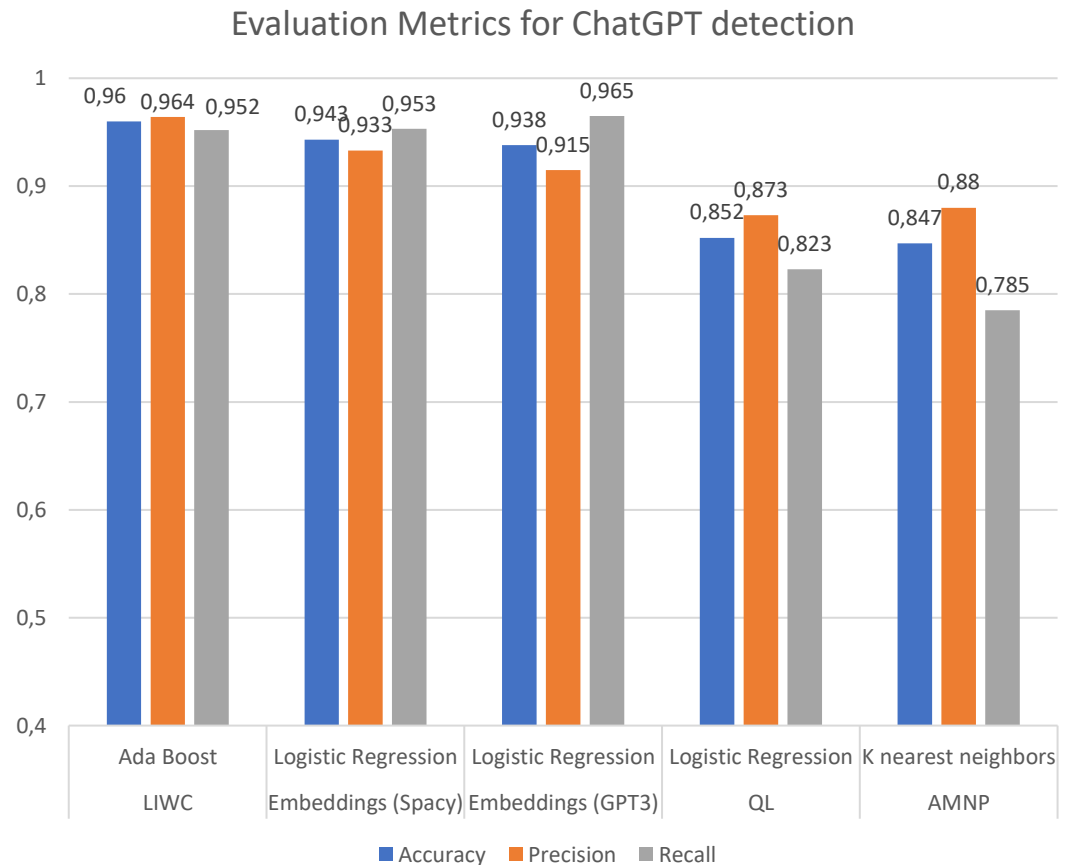
# Developing a ChatGPT detector

## Features

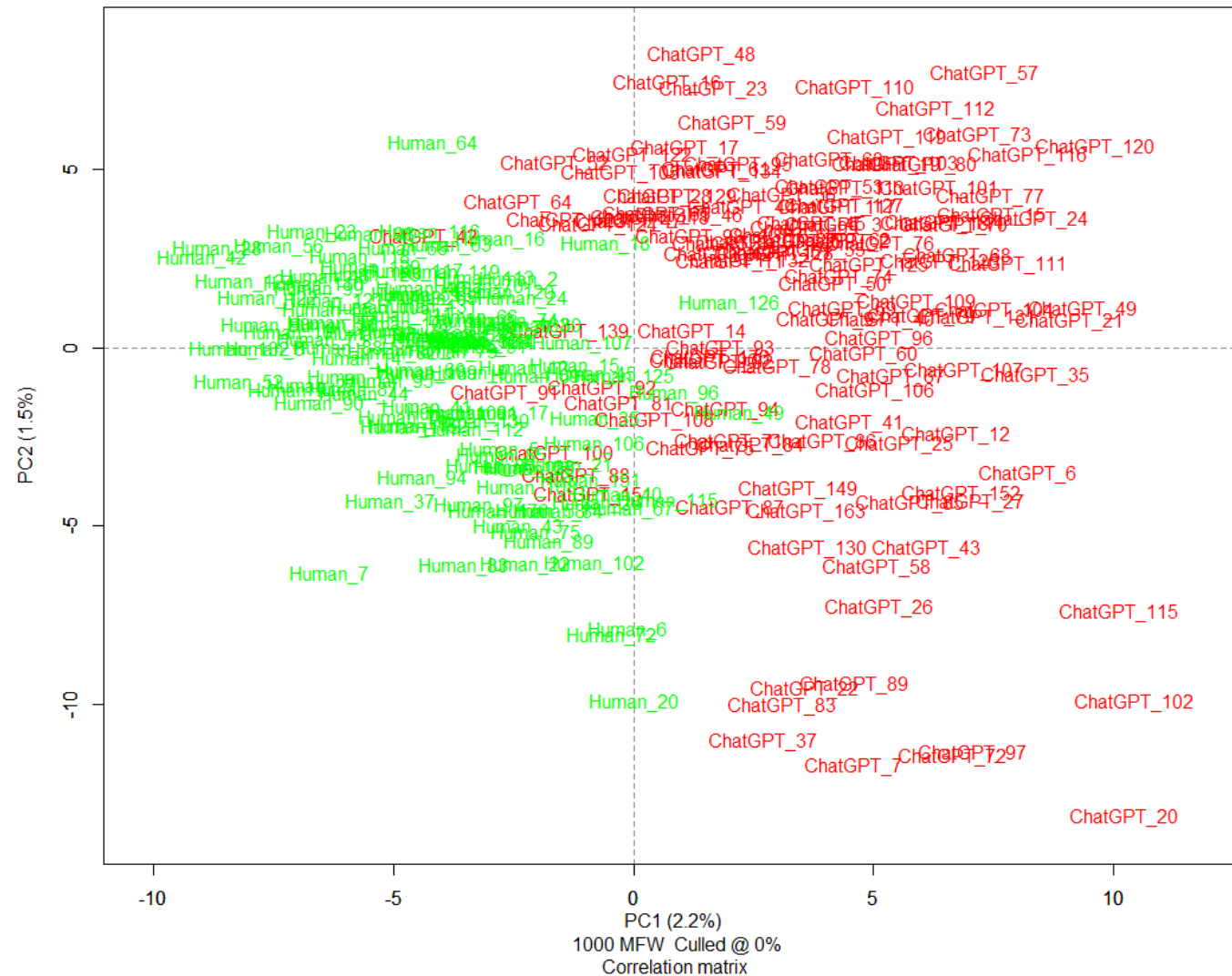
1. Author Multilevel Ngram Profiles (AMNP)
2. Embeddings (Spacy)
3. Embeddings (GPT3)
4. Linguistic Word Count Inquiry (LIWC)
5. Quantitative Linguistics (QL) indices were calculated by the software QUITA.

## Conclusions

- Standard stylometric feature groups such as the AMNP and the QL are not providing enough detection power. Although they work very well distinguishing human stylometric profiles, they can't detect ChatGPT writing efficiently.
- Word embeddings are powerful feature groups for detecting AI writing, but they exhibit significantly higher recall over detecting AI writing and provide many false positives.
- The most accurate feature group was the LIWC vocabulary, which focuses on various aspects of the expressions of the emotional and psychological states of the authors.



Documents  
Principal Components Analysis



# Approaches in AI-writing detection

- **Stylometry**

- We can compute thousand of stylometric features that capture well the authorship signal and use them for identifying the AI author. Works well for human texts. However...
  - In AI writing this approach is defeated easily. You can ask ChatGPT to write in different styles (write like Hemingway, write like a 10-year-old style, write like Trump etc.)

- **Transfer Learning**

- Use another LLM or even the same LLM to recognize its output. Fine-tune a transformer's model with labeled data (texts with ground truth information whether they have been written by AI or humans) and let the LLM to adjust its network weights so it can automatically classify a text (OpenAI's approach).
  - This approach suffers from the same issues of Stylometry. Even small changes in the LLM output can fool the detector.

- **Watermarking**

- One of the most active and prominent research areas. Watermarking involves dividing a dictionary of potential words into two sets based on an algorithm: a 'green set' which the AI will mainly use, and a 'red set' which the AI mostly won't use. When the AI generates text, it predominantly uses words from the 'green set'. A human reader wouldn't notice this distinction if the word division is done meticulously. Hence, if a piece of text primarily consists of 'green set' words, it's highly likely that it was written by an AI, since the probability of a human consistently choosing words from the 'green set' is extremely low.
  - The watermarking algorithm will have to be developed from the same company that developed the LLM
  - It can be easily fooled by paraphrasing tools
  - A new market will emerge for AI generators with no watermarking
  - Vulnerable to spoofing attacks

+

## 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.



# AI-writing detection

- Open AI text classifier: <https://platform.openai.com/ai-text-classifier>
  - GPTZero: <https://gptzero.me/>
  - Turnitin: <https://www.turnitin.com/blog/sneak-preview-of-turnitins-ai-writing-and-chatgpt-detection-capability>
  - Originality.ai: <https://app.originality.ai/>
  - CheckforAi: <https://checkforai.com/>
  - ChatGPT Detector: <https://huggingface.co/spaces/Hello-SimpleAI/chatgpt-detector-ling>
  - AI Detector: <https://contentatscale.ai/ai-content-detector/>
- 
- AI Content Detector (Writer.Com): <https://writer.com/ai-content-detector/>
  - GPT-2 Output Detector: <https://openai-openai-detector.hf.space/>
  - Giant Language model Test Room: <http://gltr.io/dist/index.html>
  - AI Content Detector (Copyleaks): <https://copyleaks.com/features/ai-content-detector>
  - AI Content Detector (Corrector): <https://corrector.app/ai-content-detector/>
  - Kazan SEO Content Checker: <https://kazanseo.com/>
  - On-Page: <https://on-page.ai/>

# GPTZero – logo evolution

**GPTZero**

Detect AI Plagiarism. Accurately

**GPTZero**

Humans Deserve the Truth

**GPTZero**

The World's **#1 AI Detector** with  
over 1 Million Users



What is the impact of AI writing in literacy development of kids?



As of my knowledge cutoff in September 2021, the impact of AI writing tools on the literacy development of children was a relatively new and under-researched area. However, we can draw upon related research and expert opinions to hypothesize several potential impacts.



1. **Assistance with spelling and grammar:** AI writing tools can correct spelling and grammatical errors, helping children improve their writing skills. This feedback can be immediate and personalized, potentially accelerating learning.
2. **Improving vocabulary:** AI writing tools can suggest more complex or varied vocabulary, encouraging children to expand their lexical repertoire.
3. **Supporting structure and coherence:** Some advanced AI writing tools can provide feedback on the structure and coherence of a text, helping children to develop these higher-level writing skills.
4. **Facilitating creative writing:** AI can provide prompts or continue a story that a child has started, potentially stimulating creativity and making writing a more enjoyable activity.

However, there are also potential risks and challenges:

1. **Over-reliance on AI:** If children rely too heavily on AI for spelling and grammar corrections, they may not develop these skills independently.
2. **Accuracy of AI feedback:** While AI has become quite advanced, it can still make errors or provide feedback that is misleading or unhelpful.
3. **Data privacy and safety:** Using AI writing tools often involves sharing personal data, which raises concerns about privacy and safety, especially for children.
4. **Equity issues:** Access to AI writing tools may be limited by socioeconomic factors, potentially exacerbating educational inequalities.

It's important to note that the impact of AI writing tools on children's literacy development will also depend on how these tools are used. For example, the role of teachers in integrating AI tools into the classroom and providing human guidance and feedback will likely be crucial. Further research will be needed to fully understand these impacts and how to maximize the benefits while mitigating the risks.

For the most recent research and studies conducted after my knowledge cutoff, I can perform a search. Would you like me to do so?

# GPTZero

## GPTZero

The World's **#1 AI Detector** with  
over 1 Million Users

[Try Origin](#), our **Chrome extension**

Detect ChatGPT, GPT3, GPT4, Bard, and other AI models. Try it for yourself 🚀

GPT3

GPT4

CHATGPT

BARD

HUMAN

AI + HUMAN

As of my knowledge cutoff in September 2021, the impact of AI writing tools on [the literacy development of children](#) was a relatively new and under-researched area. However, we can draw upon related research and expert opinions to hypothesize several potential impacts.

Assistance with spelling and grammar: AI writing tools can correct spelling and grammatical errors, helping children improve their writing skills. This feedback can be immediate and personalized, potentially accelerating learning.

Improving vocabulary: AI writing tools can suggest more complex or varied vocabulary, encouraging children to expand their lexical repertoire.

Supporting structure and coherence: Some advanced AI writing tools can provide feedback on the structure and coherence of a text.

2188/5000

or, choose a file to upload

CHOOSE FILE

No file chosen

Accepted file types: pdf, docx, txt

☒ I agree to the [terms of service](#)

GET RESULTS

**Your text is likely to be written  
entirely by a human**

What is the impact of AI writing in literacy development of kids?

As of my knowledge cutoff in September 2021, the impact of AI writing tools on the literacy development of children was a relatively new and under-researched area. However, we can draw upon related research and expert opinions to hypothesize several potential impacts.

1. **Assistance with spelling and grammar:** AI writing tools can correct spelling and grammatical errors, helping children improve their writing skills. This feedback can be immediate and personalized, potentially accelerating learning.
2. **Improving vocabulary:** AI writing tools can suggest more complex or varied vocabulary, encouraging children to expand their lexical repertoire.
3. **Supporting structure and coherence:** Some advanced AI writing tools can provide feedback on the structure and coherence of a text, helping children to develop these higher-level writing skills.
4. **Facilitating creative writing:** AI can provide prompts or continue a story that a child has started, potentially stimulating creativity and making writing a more enjoyable activity.

However, there are also potential risks and challenges:

1. **Over-reliance on AI:** If children rely too heavily on AI for spelling and grammar corrections, they may not develop these skills independently.
2. **Accuracy of AI feedback:** While AI has become quite advanced, it can still make errors or provide feedback that is misleading or unhelpful.
3. **Data privacy and safety:** Using AI writing tools often involves sharing personal data, which raises concerns about privacy and safety, especially for children.
4. **Equity issues:** Access to AI writing tools may be limited by socioeconomic factors, potentially exacerbating educational inequalities.

It's important to note that the impact of AI writing tools on children's literacy development will also depend on how these tools are used. For example, the role of teachers in integrating AI tools into the classroom and providing human guidance and feedback will likely be crucial. Further research will be needed to fully understand these impacts and how to maximize the benefits while mitigating the risks.

For the most recent research and studies conducted after my knowledge cutoff, I can perform a search. Would you like me to do so?

## AI Text Classifier

The AI Text Classifier is a fine-tuned GPT model that predicts how likely it is that a piece of text was generated by AI from a variety of sources, such as ChatGPT.

This classifier is available as a free tool to spark discussions on AI literacy. For more information on ChatGPT's capabilities, limitations, and considerations in educational settings, please visit [our documentation](#).

### Current limitations:

- Requires a minimum of 1,000 characters, which is approximately 150 - 250 words.
- The classifier isn't always accurate; it can mislabel both AI-generated and human-written text.
- AI-generated text can be edited easily to evade the classifier.
- The classifier is likely to get things wrong on text written by children and on text not in English, because it was primarily trained on English content written by adults.

### Try the classifier

To get started, choose an example below or paste the text you'd like to check. Be sure you have appropriate rights to the text you're pasting.

### Examples

- Human-Written   AI-Generated   Misclassified Human-Written

### Text

As of my knowledge cutoff in September 2021, the impact of AI writing tools on the literacy development of children was a [relatively new](#) and under-researched area. However, we can draw upon related research and expert opinions to hypothesize several potential impacts.

Assistance with spelling and grammar: AI writing tools can correct spelling and grammatical errors, helping children improve their writing skills. This feedback can be immediate and personalized, potentially accelerating learning.

Improving vocabulary: AI writing tools can suggest more complex or varied vocabulary, encouraging children to expand their lexical repertoire.

Supporting structure and coherence: Some advanced AI writing tools can provide feedback on the structure

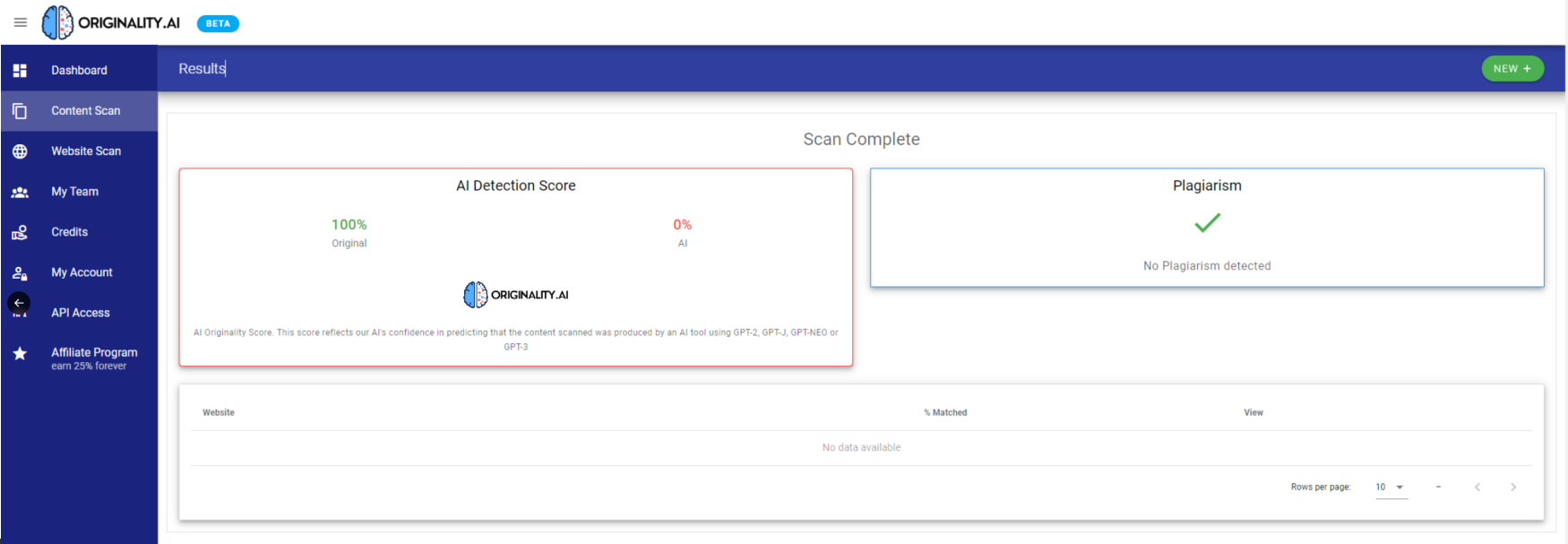
By submitting content, you agree to our [Terms of Use](#) and [Privacy Policy](#). Be sure you have appropriate rights to the content before using the AI Text Classifier.


- [Submit](#) [Clear](#)

The classifier considers the text to be **unlikely** AI-generated.


# OpenAI's own detector!

# Originality AI





Write a story of a boy and a boat. Try to alternate easy and rare words in each sentence.



Once upon a time, there was a young lad named Timmy who lived near a tranquil lake. He spent his days meandering along the shoreline, gazing at the picturesque scenery and dreaming of one day venturing out onto the water.

Timmy's father, a seasoned sailor, noticed his son's yearning and decided to construct a skiff for him. Timmy was ecstatic as he helped his father with the menial tasks, his excitement palpable.

Finally, the day arrived when the skiff was completed, and Timmy's father took him out on the water for the first time. Timmy was awestruck by the vast expanse of the lake and the feeling of the wind in his hair as they glided across the surface.

They spent many hours on the water together, Timmy's father imparting his nautical knowledge and Timmy soaking it all up like a sponge. Timmy's confidence grew with each excursion, and before long, he was adept at handling the skiff on his own.



# DetectGPT – open source



Write 200 words about a boy and a boat.



Once upon a time, there was a young boy named Jack who lived in a small coastal town. Jack had a passion for the sea and loved to spend his free time fishing and exploring the shoreline. One day, he decided to build a small boat from scratch, using only materials he found on the beach. He worked tirelessly for weeks, carefully piecing together the pieces of driftwood and sturdy rope he had collected.

Finally, the day arrived when Jack's boat was ready to set sail. He paddled out into the open sea, feeling free and alive as the wind filled the sails. The water was calm and the sun was shining, making it the perfect day for adventure. Jack sailed for hours, visiting distant islands and discovering new sights and sounds. He even caught a large fish, which he brought back to shore to share with his family.

From that day on, Jack made it his mission to explore as much of the sea as he could. He became known in the town as the boy with the boat, and many of the townspeople would often ask him to take them on voyages. Jack's boat brought him happiness and adventure, and he was grateful for the endless possibilities it opened up for him.



## Try Out DetectGPT

build a small boat from scratch, using only materials he found on the beach. He worked tirelessly for weeks, carefully piecing together the pieces of driftwood and sturdy rope he had collected.

Finally, the day arrived when Jack's boat was ready to set sail. He paddled out into the open sea, feeling free and alive as the wind filled the sails. The water was calm and the sun was shining, making it the perfect day for adventure. Jack sailed for hours, visiting distant islands and discovering new sights and sounds. He even caught a large fish, which he brought back to shore to share with his family.

Detect

**DetectGPT predicts that your text is likely to be from GPT-2.**

(because the z-score is above 1)

Results computed using 100 perturbations of your text.

Original log-probability minus average perturbed log-probability: 0.101

Standard deviation of perturbed log-probabilities: 0.069

**Z-score: 1.470**

35

Original

# DetectGPT – open source

QuillBot PREMIUM Paraphraser

QuillBot for Chrome | Write like a pro, everywhere you write. [Add to Chrome. It's free.](#)

Modes: **Standard** Fluency Formal Simple Creative Expand Shorten Synonyms:

Once upon a time, there was a young boy named Jack who lived in a small coastal town. Jack had a passion for the sea and loved to spend his free time fishing and exploring the shoreline. One day, he decided to build a small boat from scratch, using only materials he found on the beach. He worked tirelessly for weeks, carefully piecing together the pieces of driftwood and sturdy rope he had collected.

Finally, the day arrived when Jack's boat was ready to set sail. He paddled out to the open sea, feeling free and alive as the wind filled the sails. The water was calm and the sun was shining, making it the perfect day for adventure. Jack sailed for hours, visiting distant islands and discovering new sights and sounds. He even caught a large fish, which he brought back to shore to share with his family.

9 Words

**Rephrase**

1/9 Sentences • 145 Words

Changed Words Structural Changes Longest Unchanged

Try Out DetectGPT

The time had finally come for Jack's yacht to leave port. As the wind filled the sails, he set off into the wide open sea, feeling alive and free. The weather was ideal for adventure: the water was calm, and the sun was beaming. Jack spent hours at sea, exploring far-off islands and taking in fresh sights and sounds. He even caught a sizable fish, which he shared with his family after returning to shore.

Detect

**DetectGPT predicts that your text is very unlikely to be from GPT-2.**

(because the z-score is less than 0.25)

Results computed using 120 perturbations of your text.  
Original log-probability minus average perturbed log-probability: 0.008  
Standard deviation of perturbed log-probabilities: 0.080  
**Z-score: 0.100**

30

25

Original  
Avg perturbed

# False Positives

## COLLEGE OF HUMANITIES AND SOCIAL SCIENCES

[Research Clusters](#)[Research Highlights](#)[Access Hub](#)

### Overview

College of Humanities and Social Sciences is one of the six colleges in Hamad Bin Khalifa University (HBKU). As part of an innovative research-driven university, CHSS aims at researching and disseminating knowledge related to the forefront issues in humanities and social sciences. Our values are excellence, diversity, society orientation, collaboration, innovation, and integrity. Also, our overarching theme is social justice that we attempt to integrate into our research activities. As a result, we aim to address social justice and achieve core values through interdisciplinary research.

### Research Clusters

The four research clusters and their related research themes aim to organize research in the CHSS. Each research cluster can help the faculty and students to position themselves as researchers in the CHSS. Through these research clusters and research themes, the faculty and students get the chance to address relevant issues and produce new knowledge. The clusters are designed to be interdisciplinary and address a wide range of topics in humanities and social sciences. There is the hope that these research clusters create research spaces for cross-department and cross-college research collaborations.

### AI Text Classifier

The AI Text Classifier is a fine-tuned GPT model that predicts how likely it is that a piece of text was generated by AI from a variety of sources, such as ChatGPT.

This classifier is available as a tool to spark discussions on AI literacy. For more information on ChatGPT's capabilities, limitations, and considerations in educational settings, please visit [this page](#).

#### Current limitations:

- Requires a minimum of 1,000 characters, which is approximately 150 - 250 words.
- The classifier isn't always accurate; it can mislabel both AI-generated and human-written text.
- AI-generated text can be edited easily to evade the classifier.
- The classifier is likely to get things wrong on text written by children and on text not in English, because it was primarily trained on English content written by adults.

#### Try the classifier

To get started, choose an example below or paste the text you'd like to check. Be sure you have appropriate rights to the text you're pasting.

#### Examples

[Human-Written](#)[AI-Generated](#)[Misclassified Human-Written](#)

#### Text

College of Humanities and Social Sciences is one of the six colleges in Hamad Bin Khalifa University (HBKU). As part of an innovative research-driven university, CHSS aims at researching and disseminating knowledge related to the forefront issues in humanities and social sciences. Our values are excellence, diversity, society orientation, collaboration, innovation, and integrity. Also, our overarching theme is social justice that we attempt to integrate into our research activities. As a result, we aim to address social justice and achieve core values through interdisciplinary research.

The four research clusters and their related research themes aim to organize research in the CHSS. Each research cluster can help the faculty and students to position themselves as researchers in the CHSS. Through these research clusters and research themes, the faculty and students get the chance to address relevant issues and produce new knowledge. The clusters are designed to be interdisciplinary and address a wide range of topics in humanities and social sciences. There is the hope that these research clusters create research spaces for cross-department and cross-college research collaborations.

By submitting content, you agree to our [Terms of Use](#) and [Privacy Policy](#). Be sure you have appropriate rights to the content before using the AI Text Classifier.

[Submit](#)[Clear](#)

The classifier considers the text to be **possibly** AI-generated.



---

Turnitin claims a “less than 1/100 false positive rate,” but is that accurate?

- “Our AI writing preview has been trained on academic writing with high efficacy rates and can identify 97% of AI writing”
- But the company has shared no data at all, let alone external peer reviewed studies
- They didn’t test their system on the most sophisticated, recent AI software, ChatGPT running GPT-4.



+



*The  
Washington  
Post* found an  
example of a  
false positive

## We tested a new ChatGPT-detector for teachers. It flagged an innocent student.

Five high school students helped our tech columnist test a ChatGPT detector coming from Turnitin to 2.1 million teachers. It missed enough to get someone in trouble.



Analysis by [Geoffrey A. Fowler](#)  
Columnist | + **Follow**

Updated April 3, 2023 at 9:47 a.m. EDT | Published April 3, 2023 at 6:00 a.m. EDT



On April 17, 2023, I retested a [popular Reddit experiment](#) using GPTZero.me.

## GPTZero labeled the Bill of Rights “likely AI”

# Your text is likely to be written entirely by AI

The nature of AI-generated content is changing constantly. As such, these results should not be used to punish students. While we build more robust models for GPTZero, we recommend that educators take these results as one of many pieces in a holistic assessment of student work. See our [FAQ](#) for more information.

Amendment I Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances.

Amendment II A well regulated Militia, being necessary to the security of a free State, the right of the people to keep and bear Arms, shall not be infringed.

Amendment III No Soldier shall, in time of peace be quartered in any house, without the consent of the Owner, nor in time of war, but in a manner to be prescribed by law.

Free software  
explicitly  
designed to  
get around AI  
detectors

**[ud]** UNDETECTABLE.AI

Playground Documents Help Pricing <sup>SALE</sup> API Earn <sup>NEW</sup>



v.2023.04.17 - Looking for full authorship replication? [Read me.](#)

## We create truly undetectable AI content.

We turn your flagged AI content into high-quality writing, that matches your brand, and is indistinguishable from human-written text.

Readability University Purpose Essay

Type in the text you want to humanize

Character Limit: 0/50000 ≈ 0 words

☐ I agree to the terms of service  
(no academic misconduct)

Check for AI

Humanize

# Some thought on using AI-writing detectors in education

- AI-writing detection is **NOT** possible, and it **WILL NOT** be possible.
  - AI-writing detectors capture statistical characteristics of the linguistic output of the LLMs **BUT** since the generation of this output is stochastic, the statistical profiling is changing everytime. We chase a moving target.
  - Typical anti-plagiarism software is based on evidence. Any software of this kind calculates the similarity index based on the percentage of copied text from known source (e.g. Wikipedia). This means that the plagiarism cases can be supported by the source documents and are indisputable.
  - AI-writing detectors give a probabilistic interpretation of the written output they examine. A 90% index means practically nothing as there is no source document to support and make a case for plagiarism.
  - An unsubstantiated false positive result will destroy the trust relationship in the education community and create distrust and disbelief among its members.

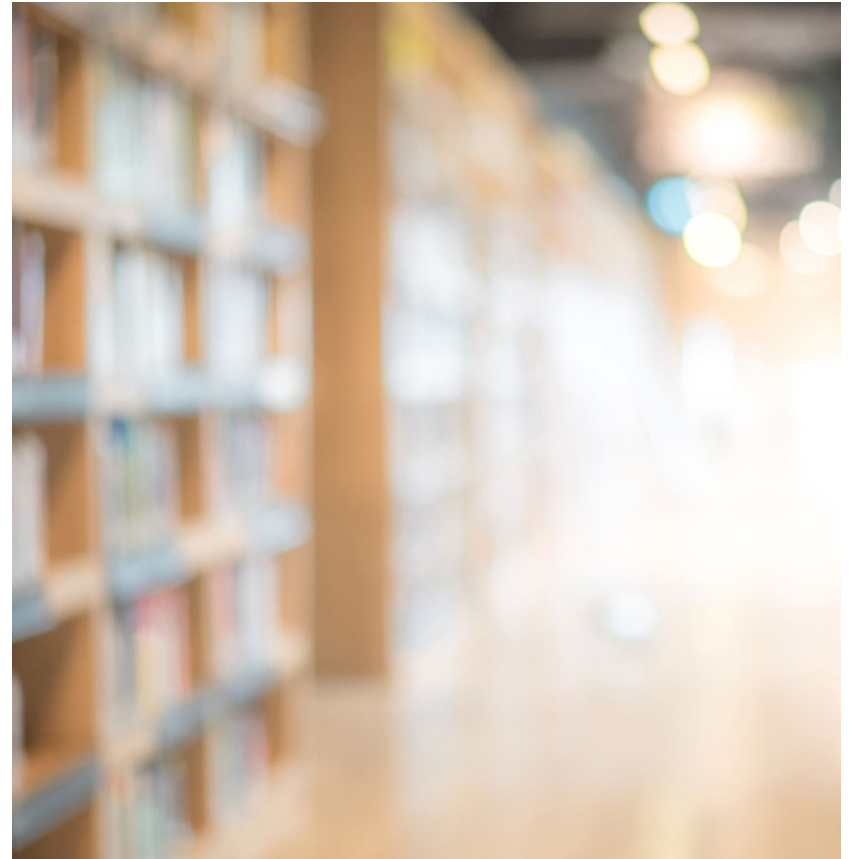




# The road ahead...

---

- 2022 will be the last year in the human history that we were sure that texts were written exclusively by humans.
- Prepare for mass flow of AI-written texts in the web the next years in the Web.
- In Science and Education hybrid writing will be the norm. Policies of academic integrity already have been updated to all institutions to reflect that. Citation standards to LLMs are already in place for APA and MLA.
- Retrospective detection could be applied to a degree if companies keep a database of outputs to certify whether a particular text sequence has ever been auto-generated.



Thank you!

gmikros@gmail. com

