



Large Language Models: Stylometric observations

George Mikros

Hamad Bin Khalifa University, Qatar



Overview

-
- Linguistic and stylometric profiling of the AI-writing
 - Human perception of AI writing
 - Linguistic and Statistical characteristics of the LLMs language
 - AI-writing detection
 - AI writing detection experiments
 - Human vs. ChatGPT
 - LLMs between themselves
 - Open Questions about the LLMs and the science of Linguistics.

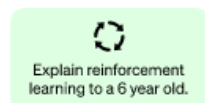
What is ChatGPT?

- 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.

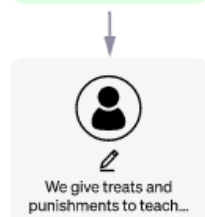
Step 1

Collect demonstration data and train a supervised policy.

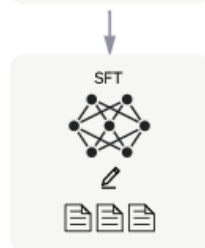
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



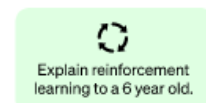
This data is used to fine-tune GPT-3.5 with supervised learning.



Step 2

Collect comparison data and train a reward model.

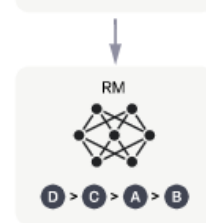
A prompt and several model outputs are sampled.



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.



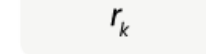
The policy generates an output.



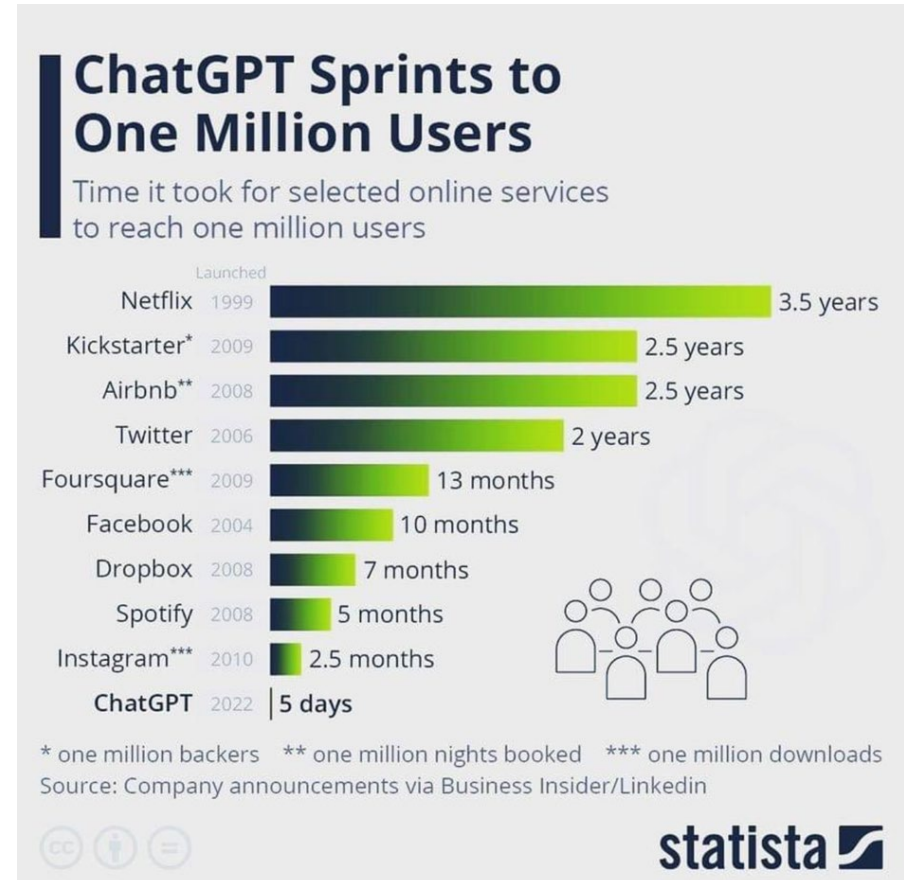
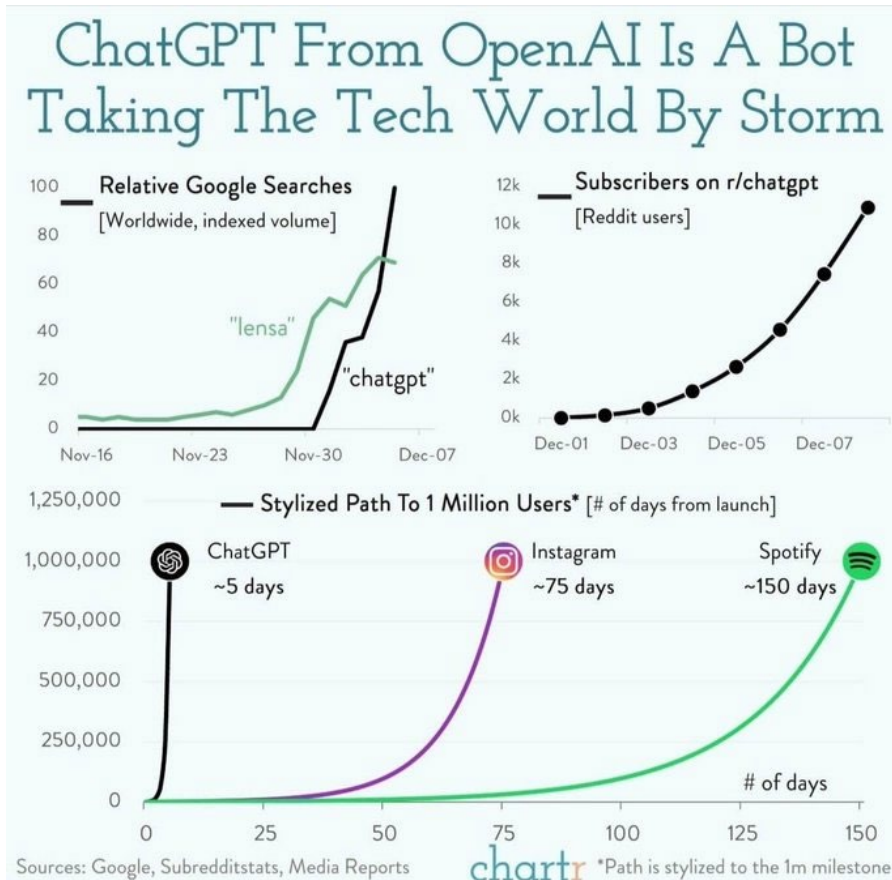
The reward model calculates a reward for the output.




The reward is used to update the policy using PPO.



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

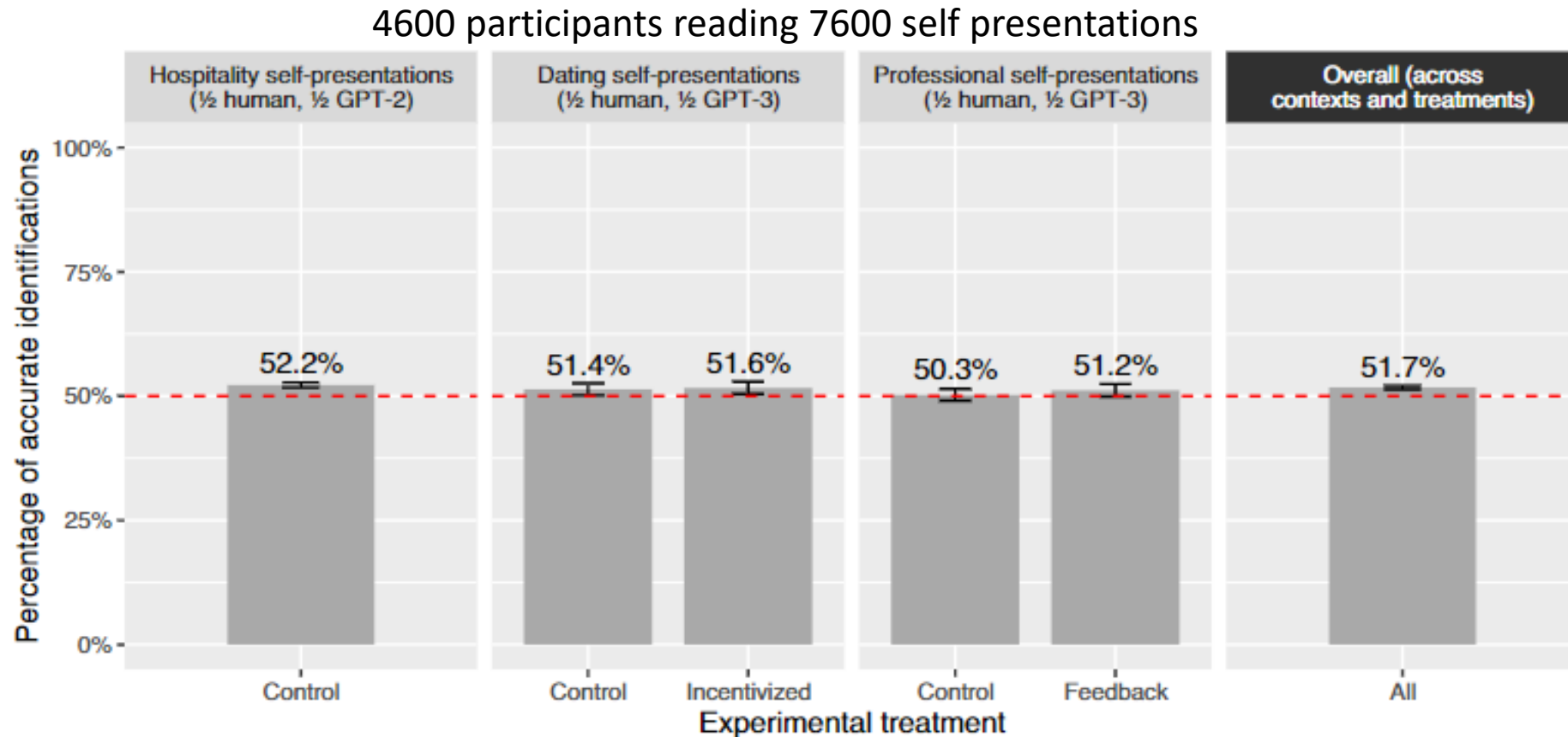




Summary of key capabilities, limitations, and concerns around ChatGPT and other LLMs

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

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>

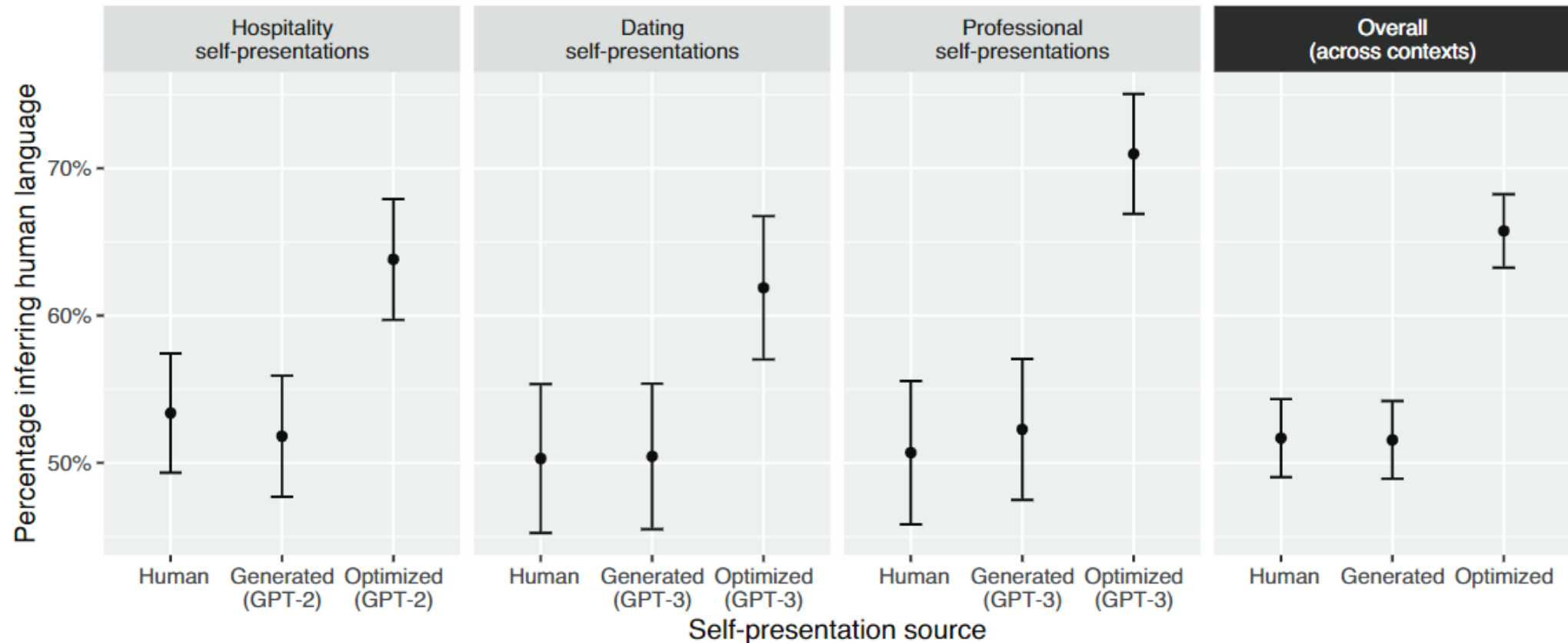
Features that make humans believe a text is written by AI

	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 †	1.105 ^{***} (1.085, 1.126)	1.233 ^{***} (1.169, 1.296)
Repetitive content †	1.083 ^{***} (1.059, 1.106)	1.470 ^{***} (1.379, 1.561)
Conversational words	0.947 ^{***} (0.925, 0.970)	0.898 ^{**} (0.829, 0.967)
<u>Misaligned features</u>		
Grammatical issues †	1.048 ^{***} (1.028, 1.069)	0.851 ^{***} (0.788, 0.913)
Rare bigrams	1.042 ^{***} (1.019, 1.065)	0.666 ^{***} (0.596, 0.736)
Long words	1.034 ^{**} (1.009, 1.059)	0.783 ^{***} (0.706, 0.861)
Contractions	0.947 ^{***} (0.924, 0.970)	1.134 ^{***} (1.065, 1.203)
<u>Nonindicative</u>		
Second-person pronouns	1.059 ^{***} (1.038, 1.079)	0.970 (0.908, 1.032)
Filler words	1.009 (0.990, 1.027)	1.119 [*] (1.021, 1.218)
Swear words	0.969 ^{**} (0.948, 0.989)	0.965 (0.905, 1.024)
Authentic words	0.946 ^{***} (0.921, 0.971)	0.945 (0.870, 1.021)
Focus on past	0.938 ^{***} (0.917, 0.959)	1.002 (0.940, 1.064)
First-person pronouns	0.925 ^{***} (0.886, 0.963)	0.992 (0.868, 1.117)
Family words	0.910 ^{***} (0.889, 0.932)	1.014 (0.950, 1.077)
Word count	0.904 ^{***} (0.874, 0.935)	1.076 (0.986, 1.165)
<i>Constant</i>	0.850 ^{***} (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: †manually labeled feature. *^{*}p^{***} 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.



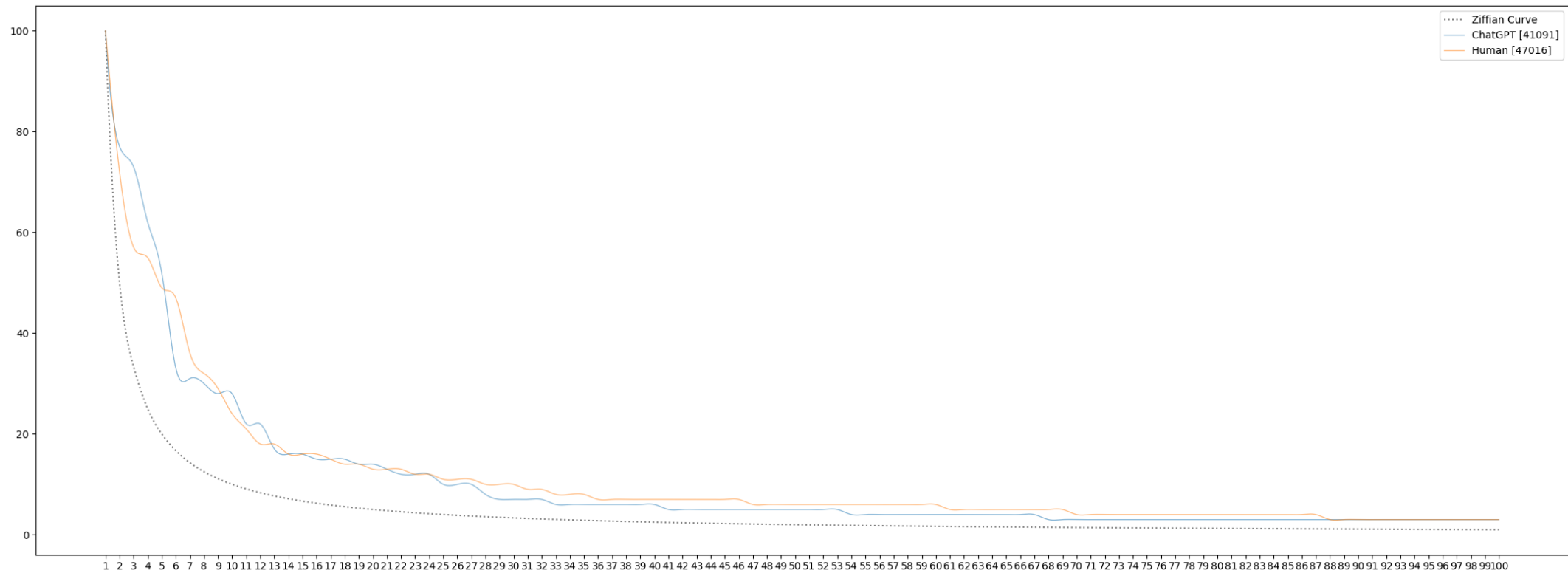
Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11), e2208839120. <https://doi.org/doi:10.1073/pnas.2208839120>

ChatGPT detection: Some experimental insights

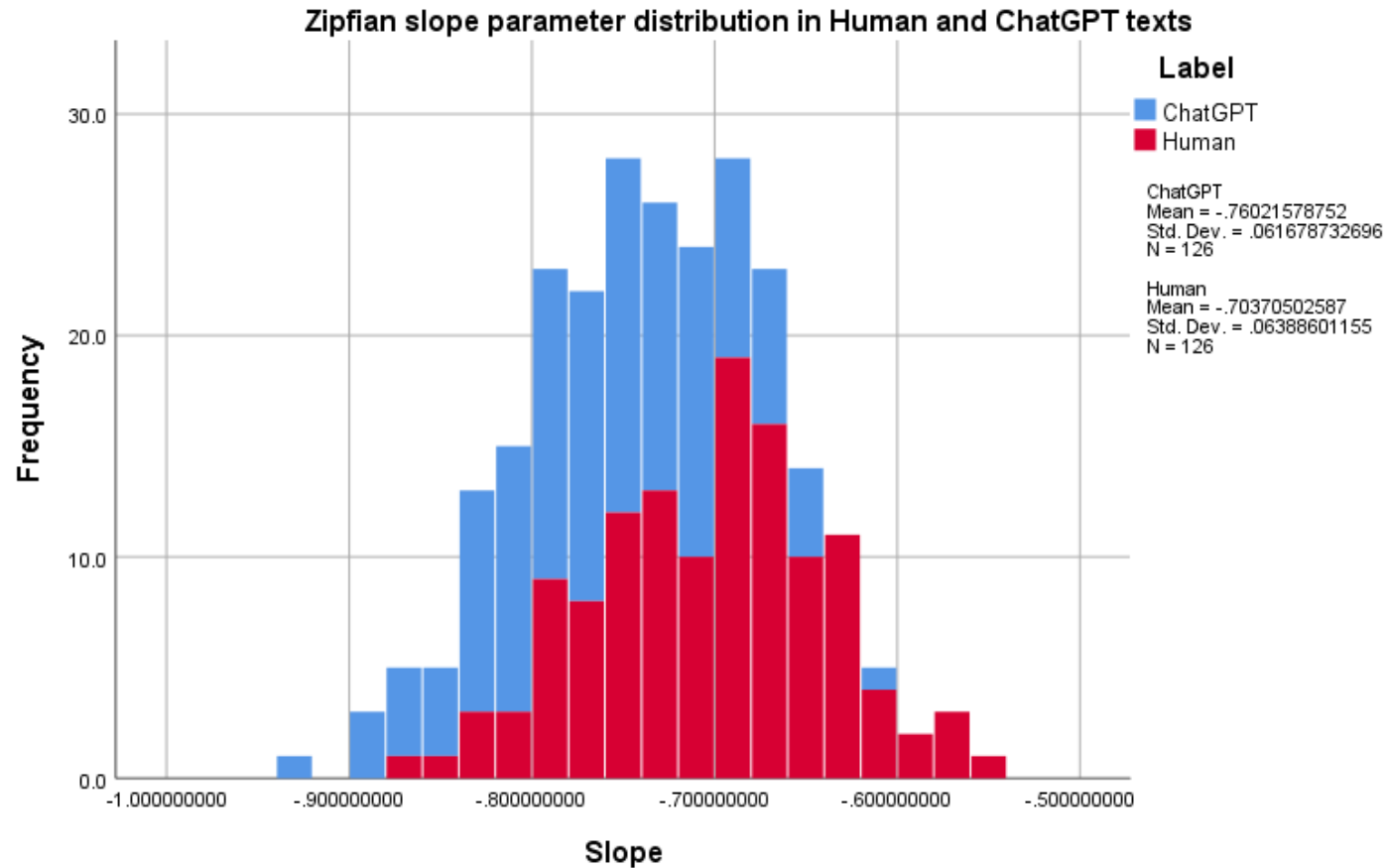
- Corpus developed by Shijaku and Canhasi (2023)
 - Size: 252 texts, of which humans wrote 126 and were part of a more comprehensive collection of TOEFL essays.
 - Each human's essay topic was given as a prompt to ChatGPT, and a machine-generated text was produced, resulting in another 126
 - AI-written texts that matched one to one the topics of the human essays.

	Texts	Words (N)	SD	Max	Min
ChatGPT	126	41,735	58.99	516	178
Human	126	47,633	105.11	658	187
Grand Total	252	89,368	88.23	658	178

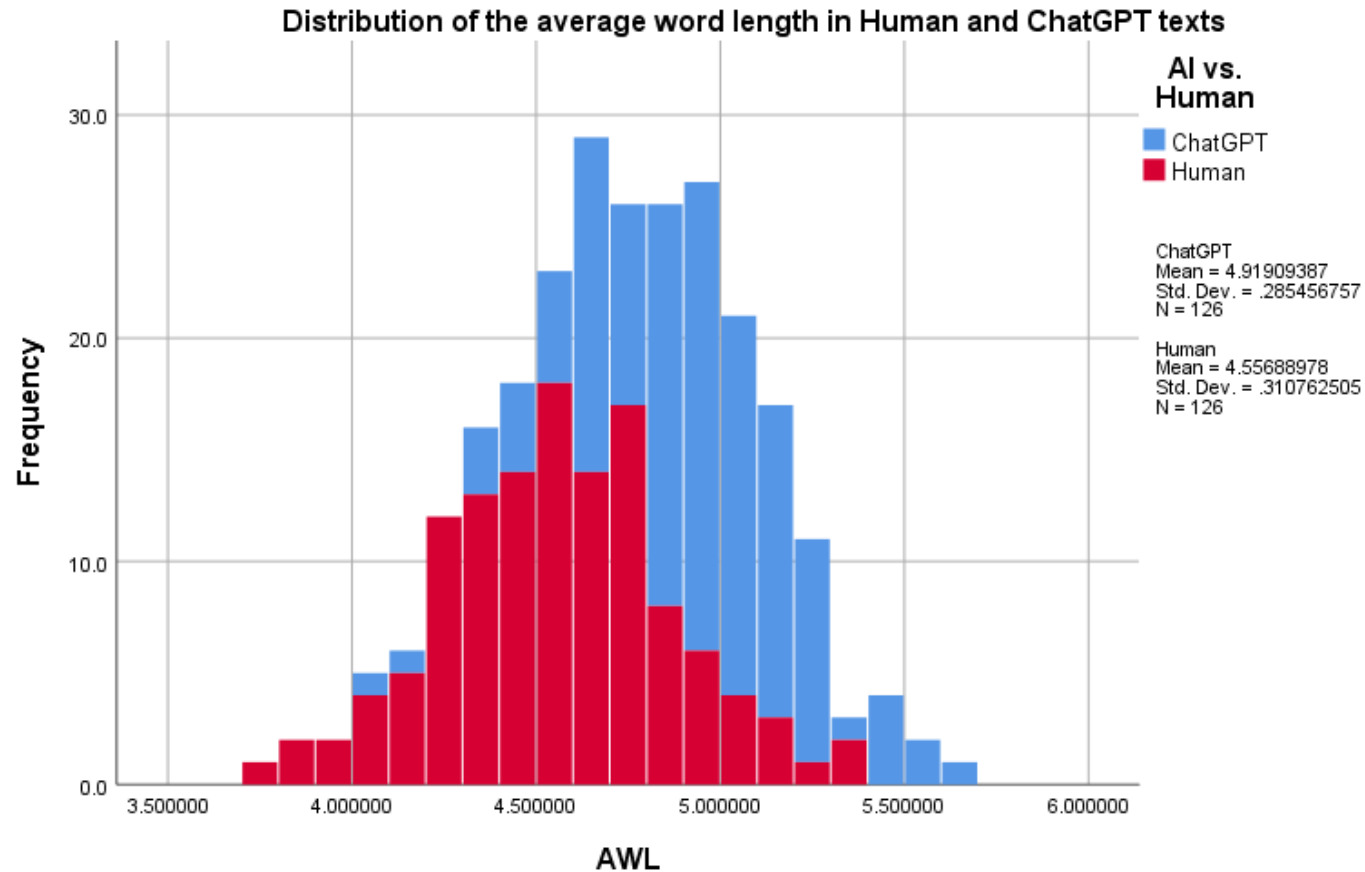
Statistical Characteristics of Language: Zipfian fit



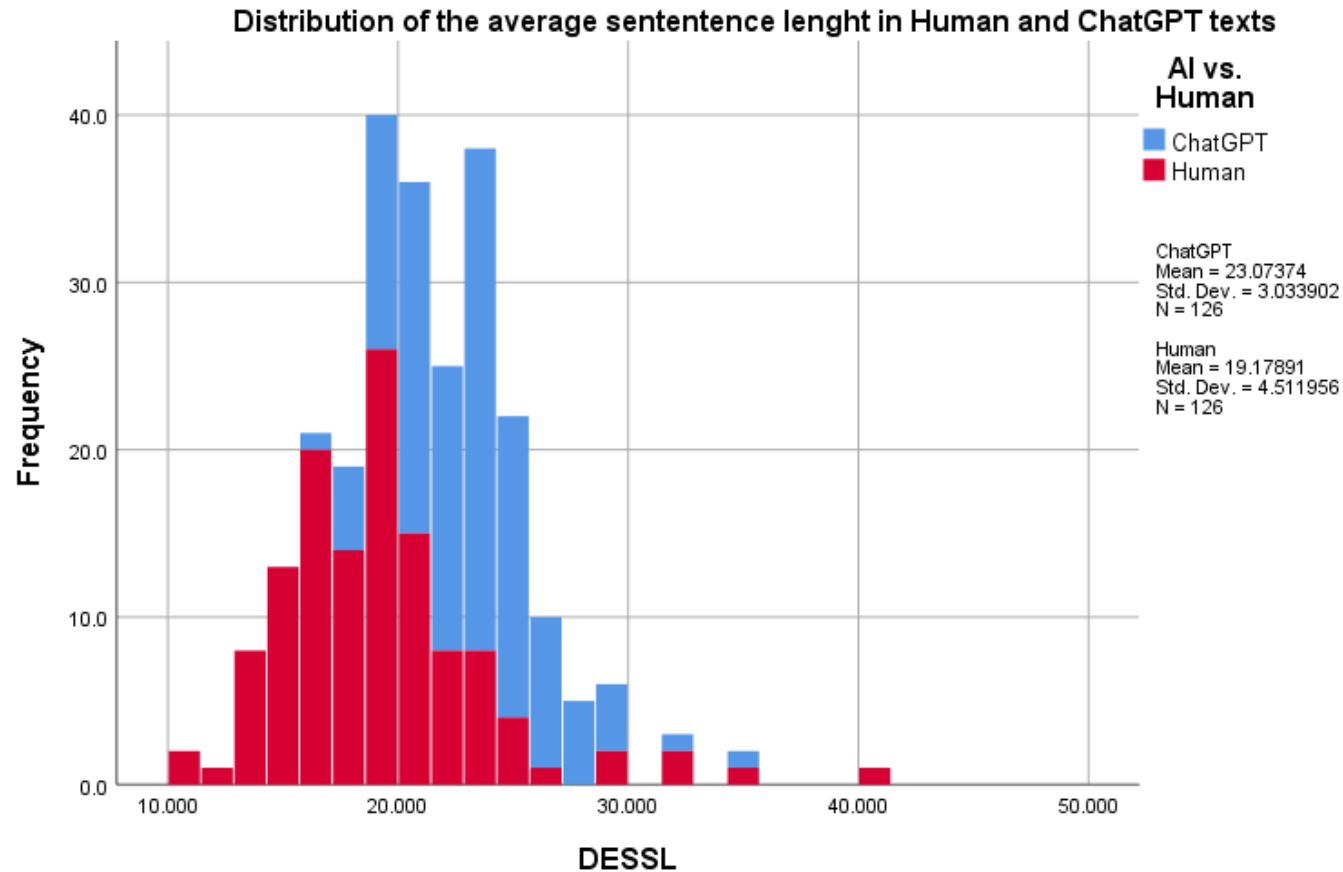
Statistical Characteristics of Language: Zipfian fit



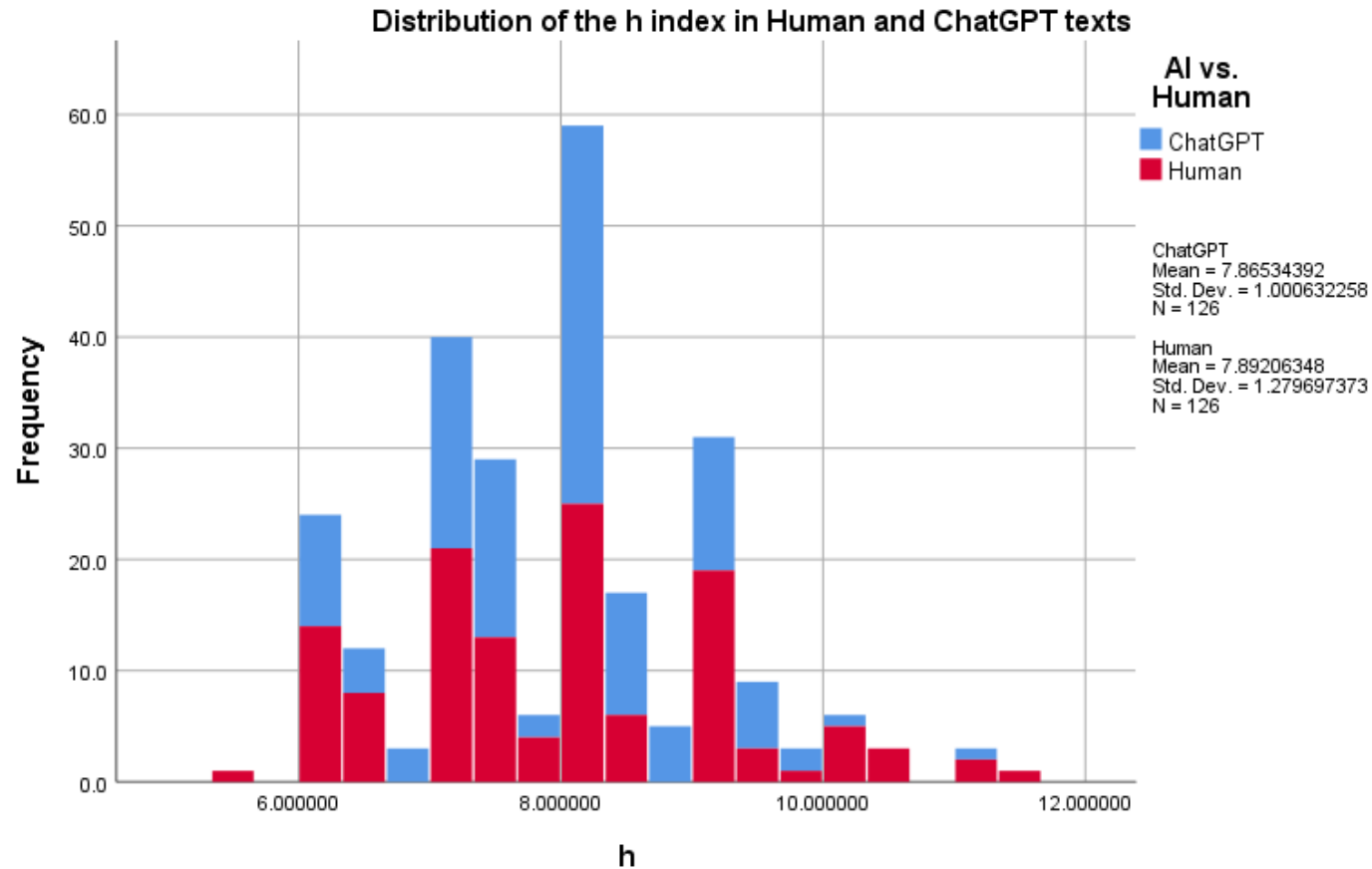
Statistical Characteristics of Language: Average Word Length



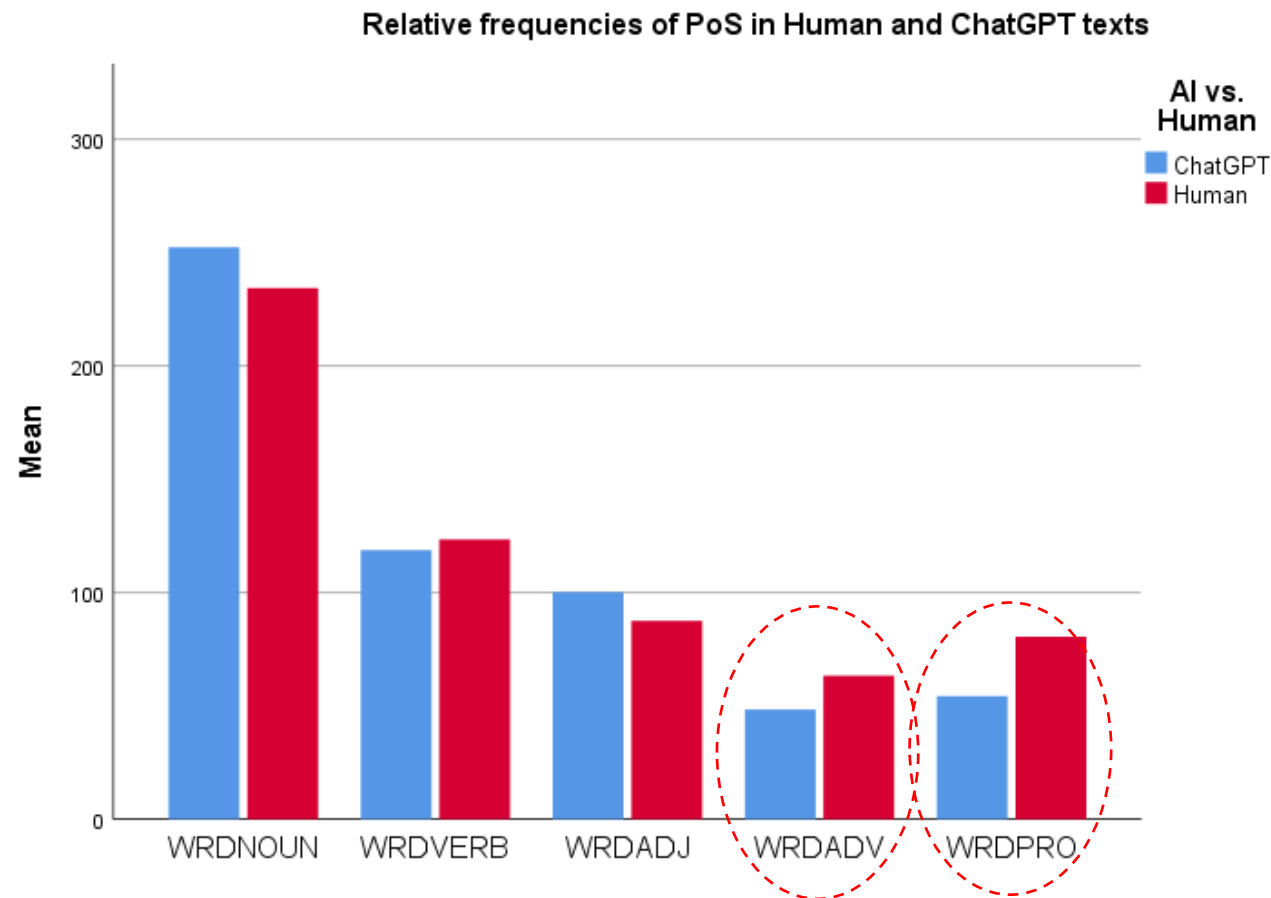
Statistical Characteristics of Language: Average Sentence Length



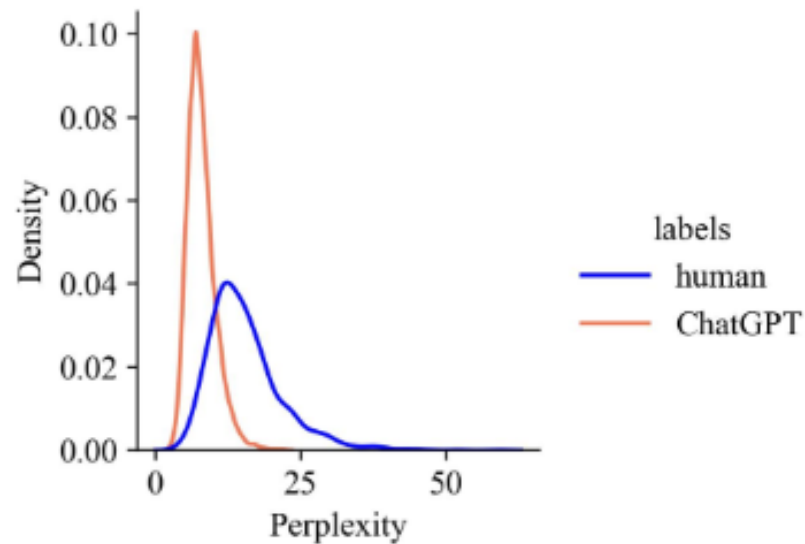
Statistical Characteristics of Language: h index



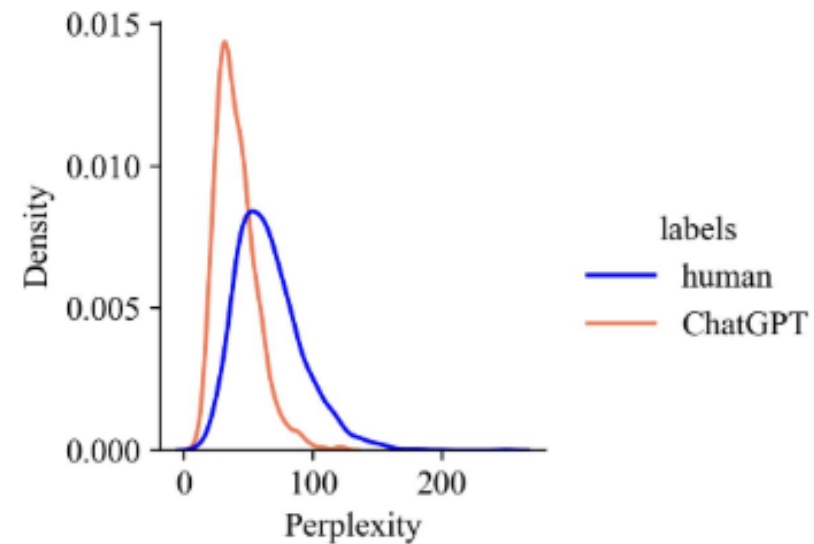
Statistical Characteristics of Language: PoS relative frequencies



Features discriminating AI-writing: Perplexity



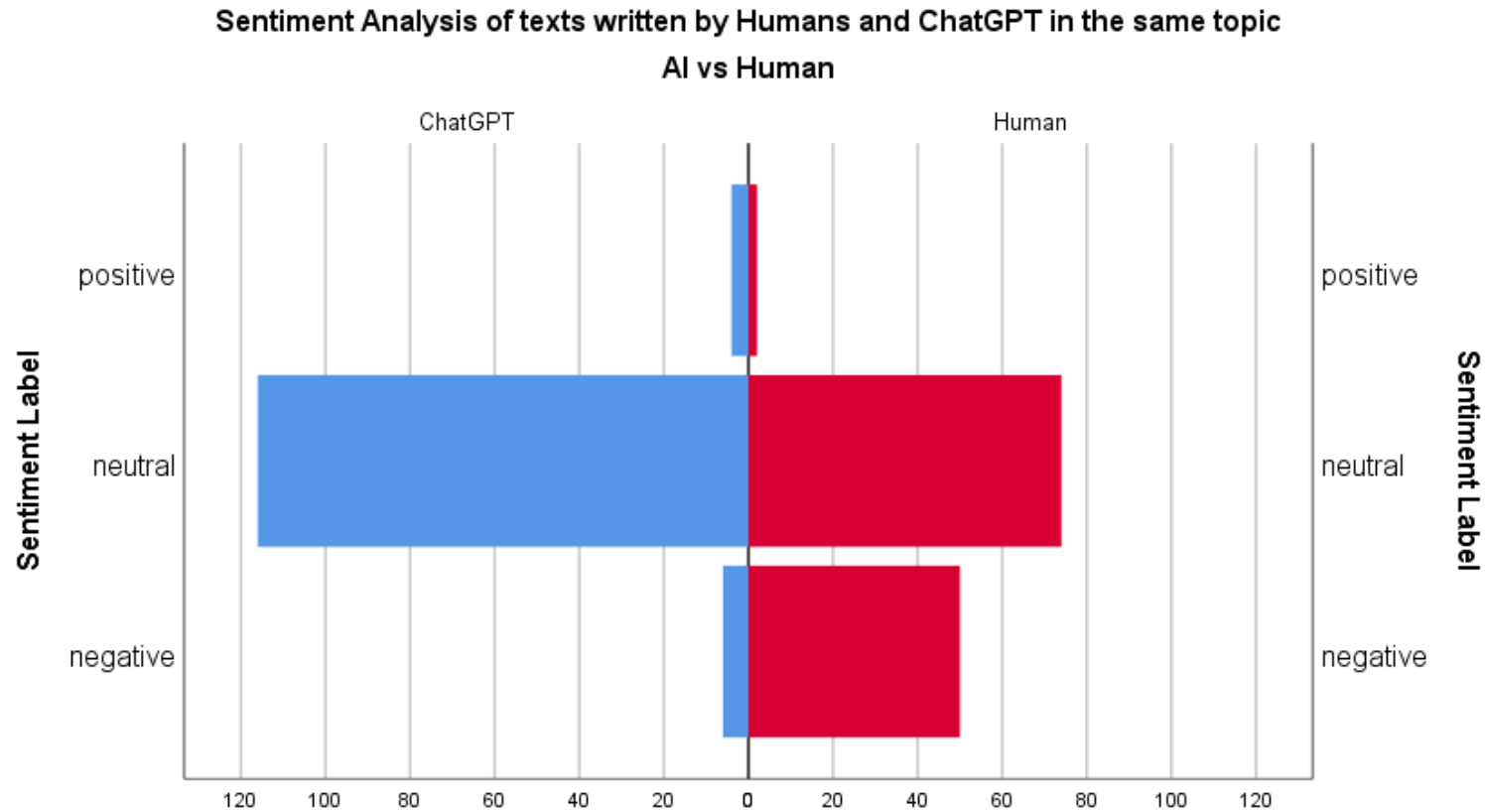
A. Text perplexity of medical abstract



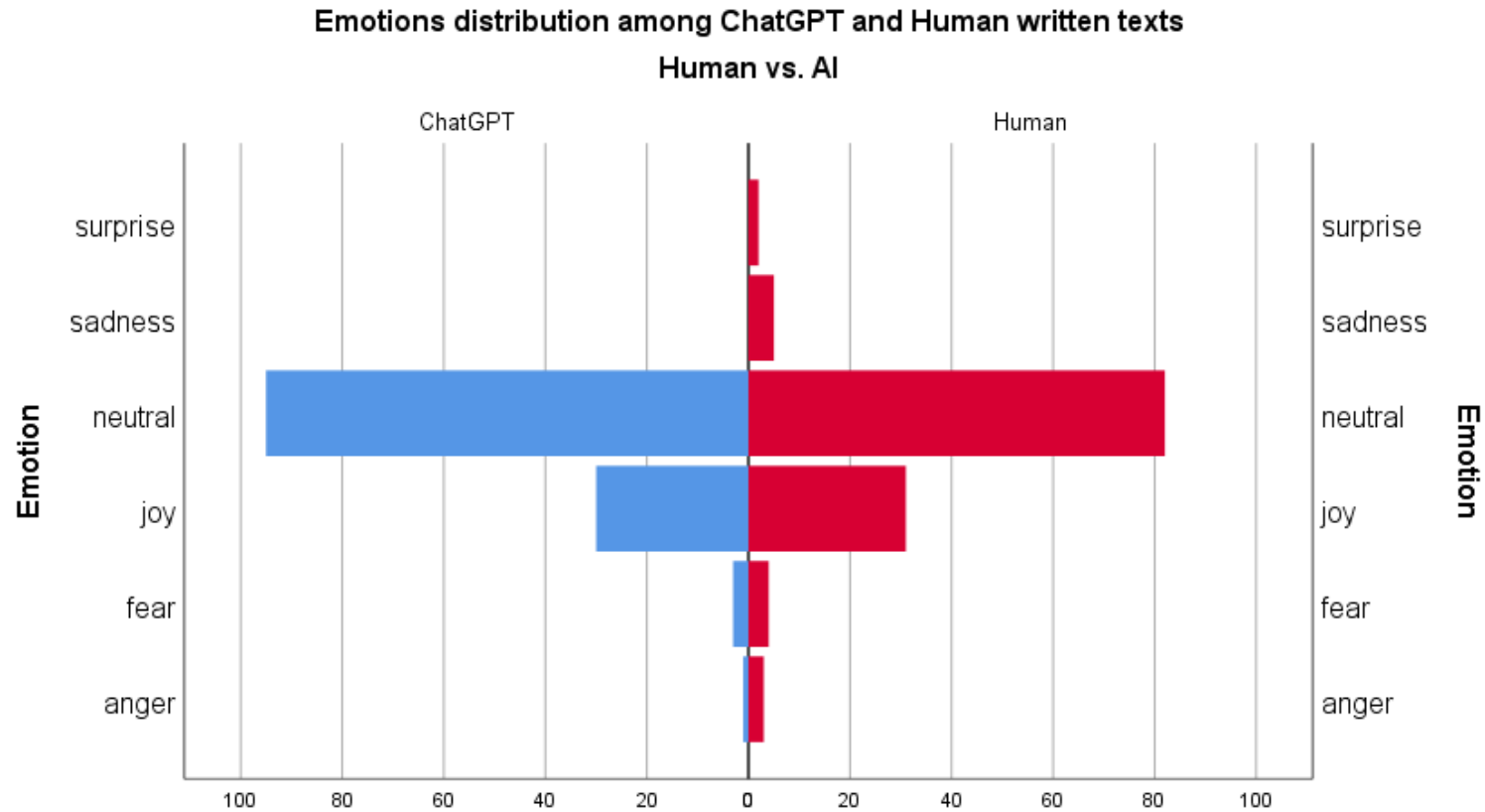
B. Text perplexity of radiology report

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

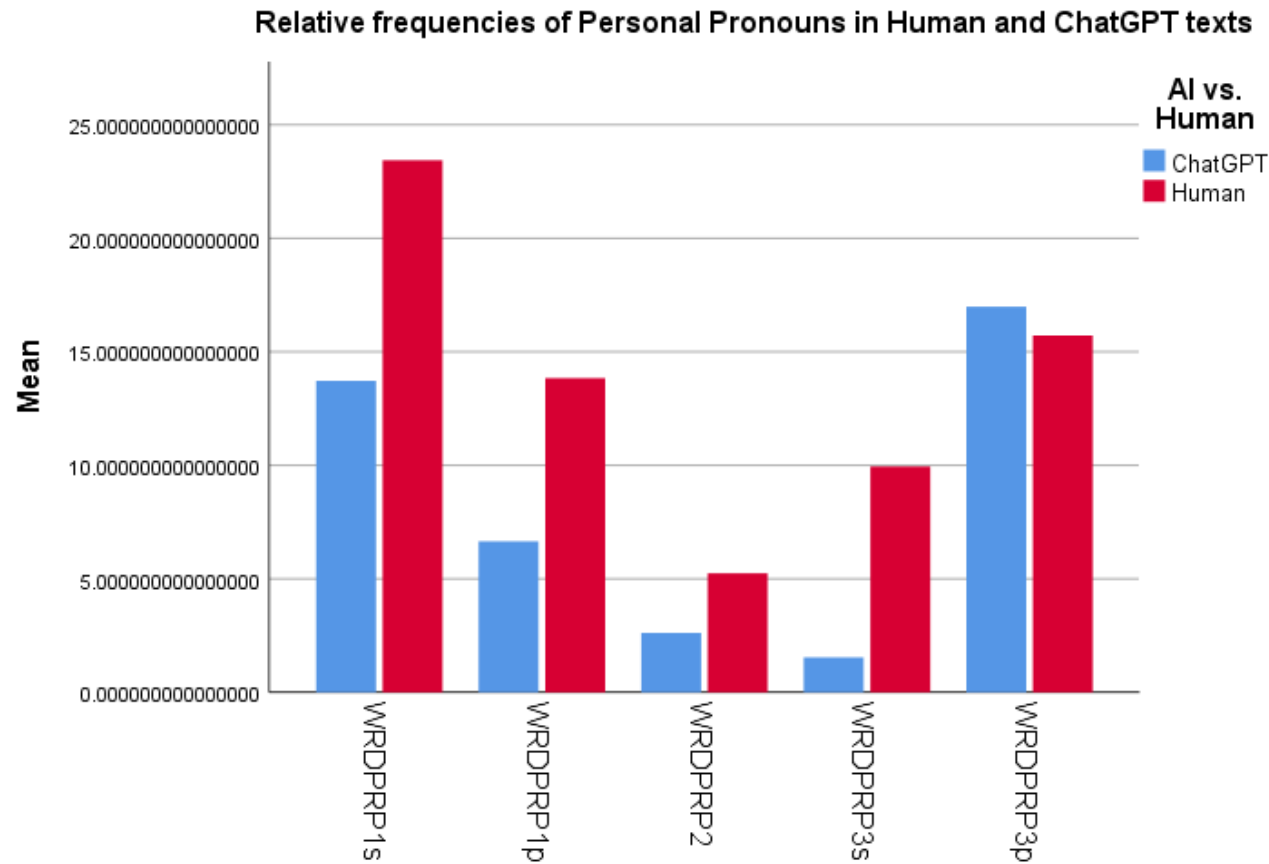
Features discriminating AI-writing: Sentiment



Features discriminating AI-writing: Emotions

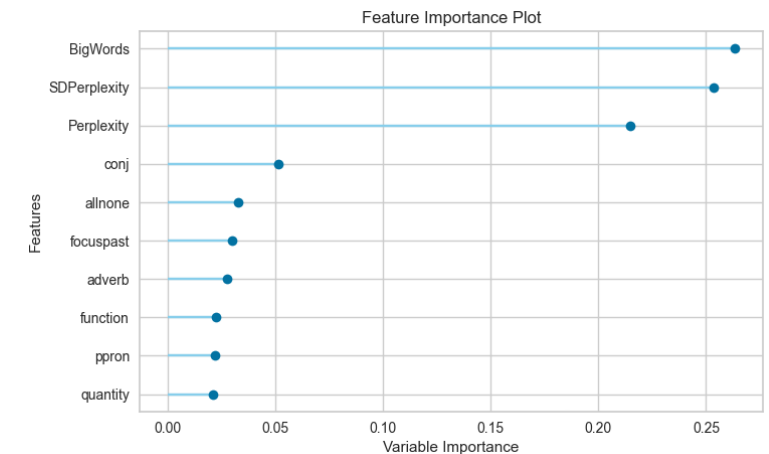
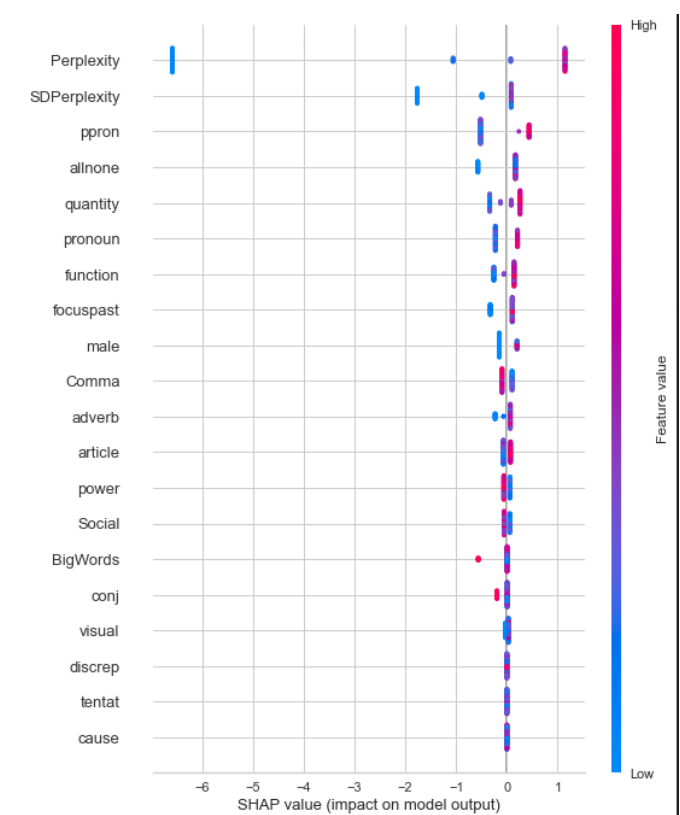


Features discriminating AI-writing: Personal Pronouns



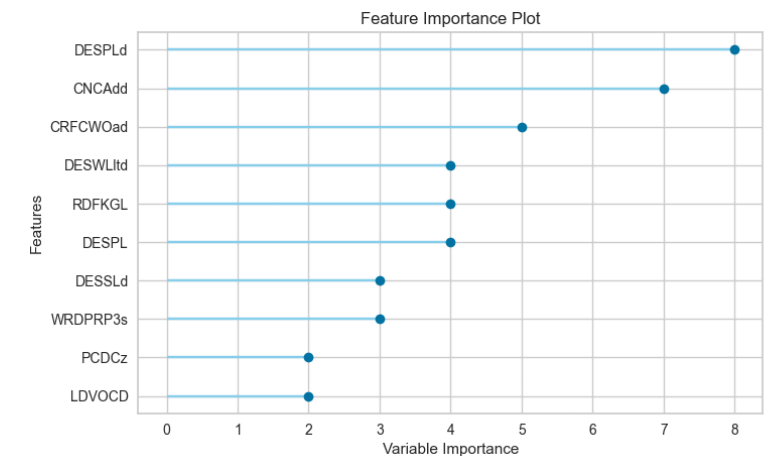
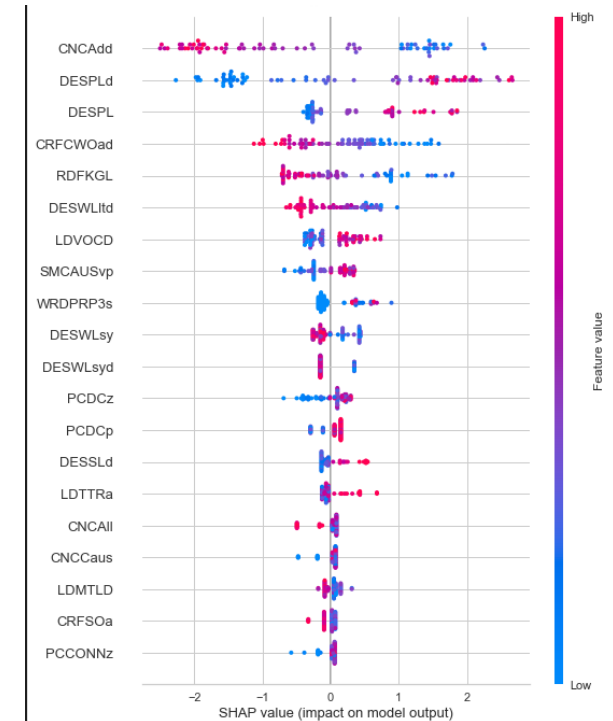
Features discriminating AI-writing: Quantitative Text Indices [1]

- 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-]
- CNCAdd: 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)

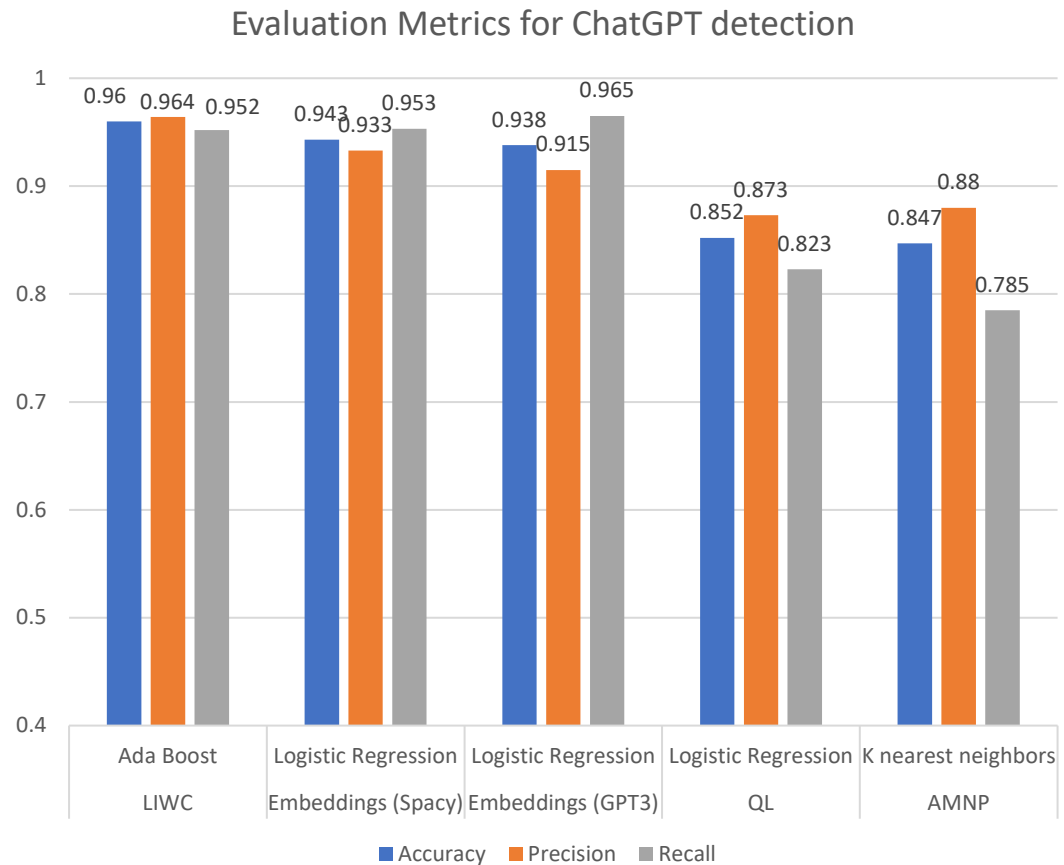
Developing a ChatGPT detector

Features

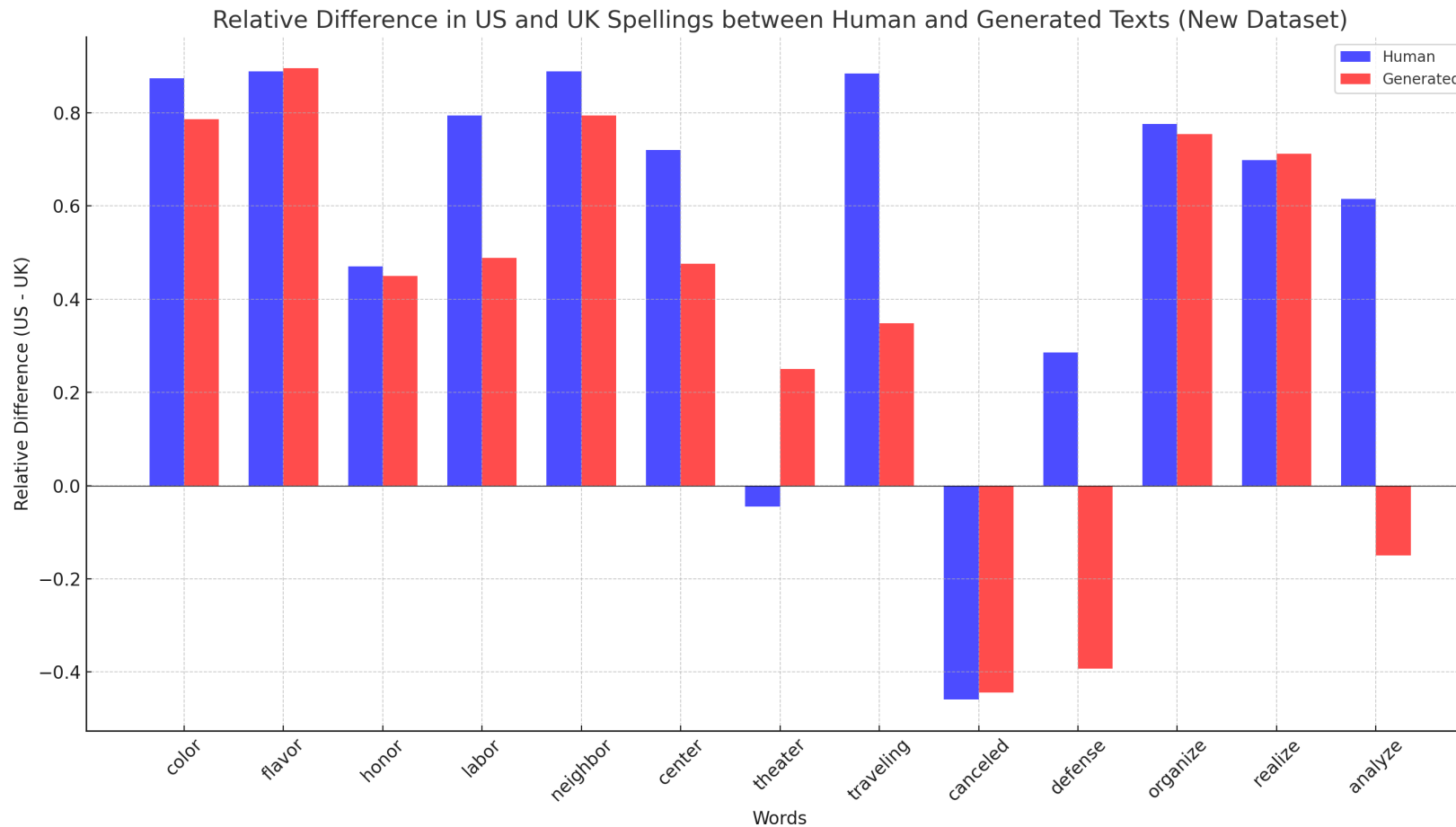
1. Author Multilevel Ngram Profiles (AMNP)
2. Embeddings (Spacy)
3. Embeddings (GPT3)
4. Linguistic Word Count Inquiry (LIWC)
5. Quantitative Linguistics (QL) indices were calculated by the software QUITA.

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.

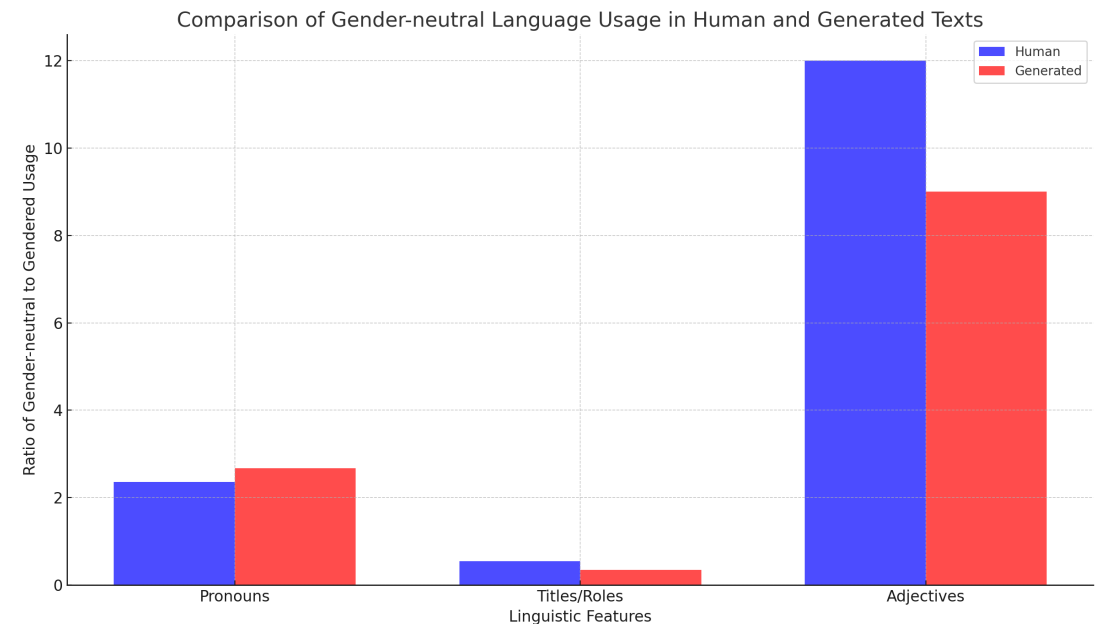


UK and British spelling variation in AI and Human-generated texts

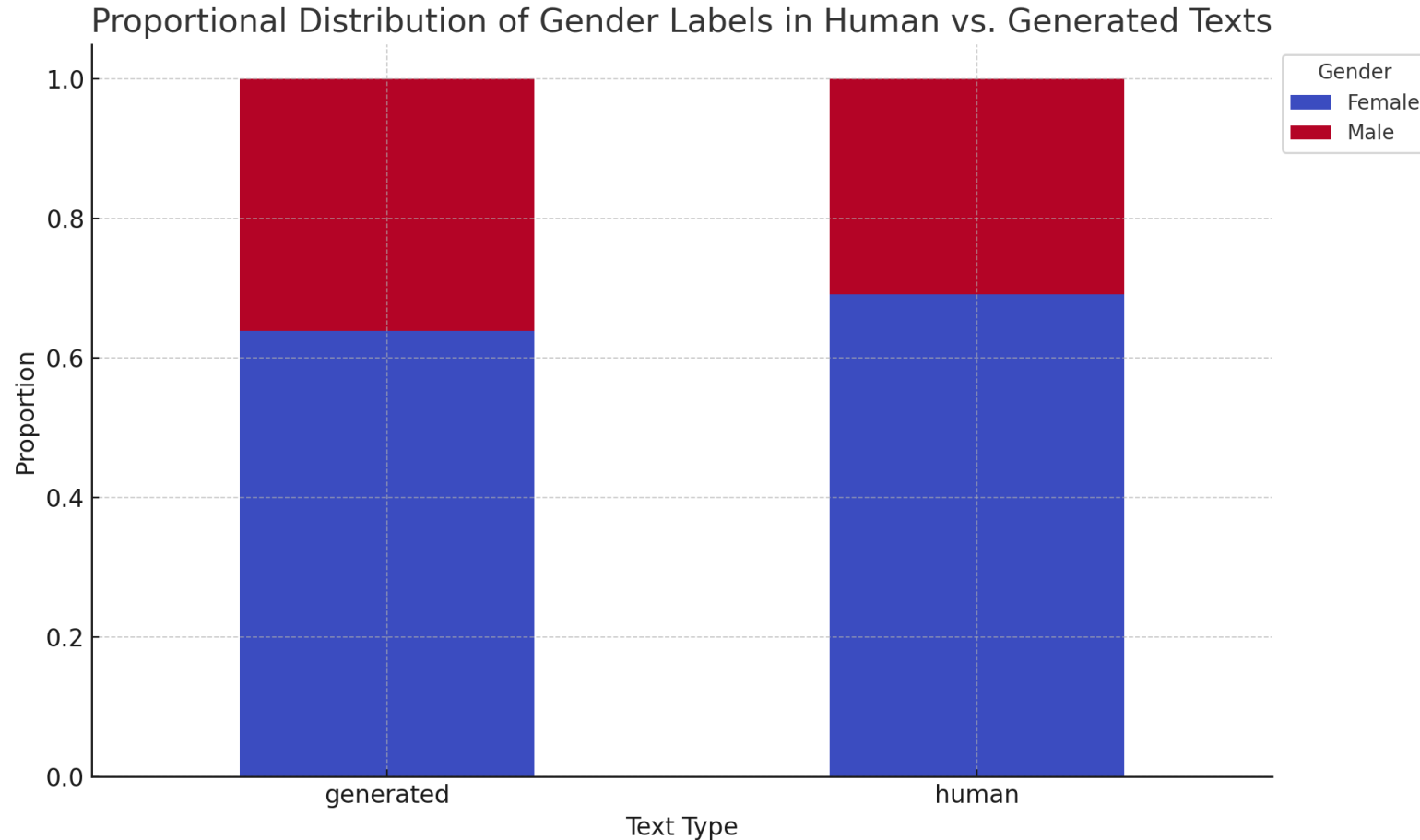


Gender-neutral language

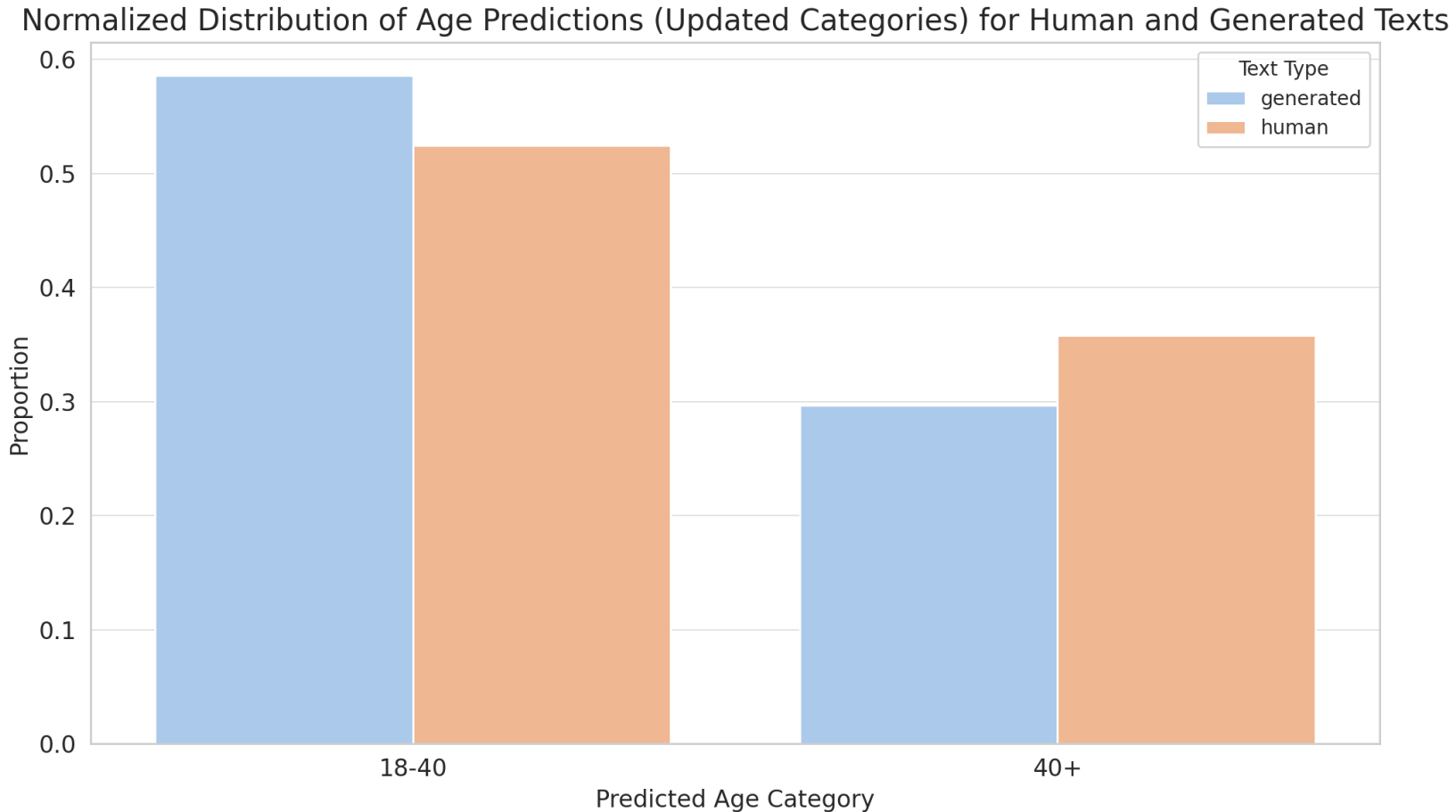
- Both 'human' and 'generated' texts prefer gender-neutral pronouns over gendered ones, with 'generated' texts showing a slightly higher preference.
- Human texts use more gender-neutral titles/roles than 'generated' texts.
- The use of gendered adjectives is relatively low, but 'human' texts have a slightly higher occurrence of such adjectives compared to 'generated' texts.



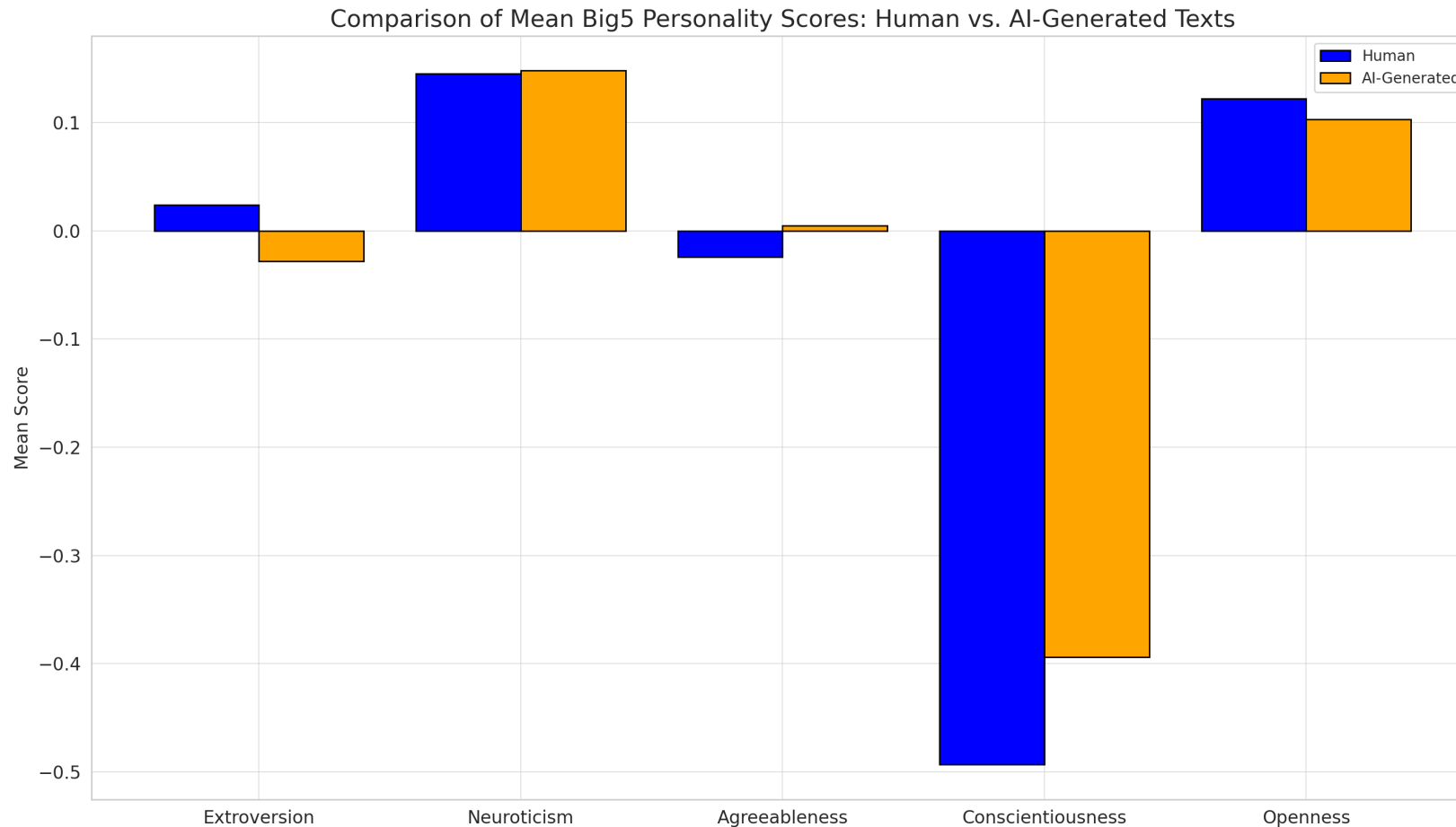
Gender profiling in human and AI-generated texts



Age profiling in human and AI-generated texts

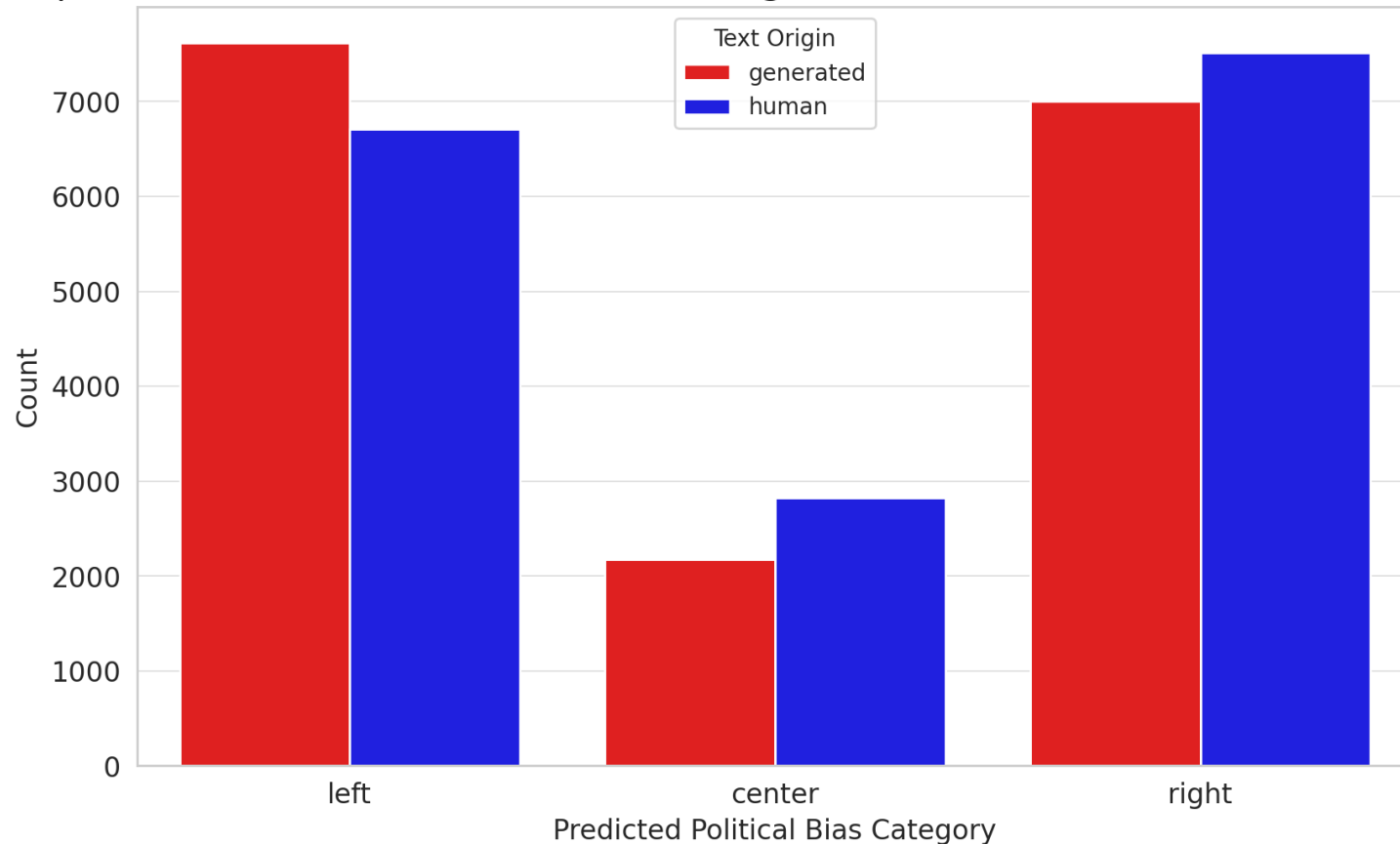


Personality profiling in human and AI-generated texts



Political bias profiling in human and AI-generated texts

Comparison of Predicted Political Bias Categories between Human and Generated Texts



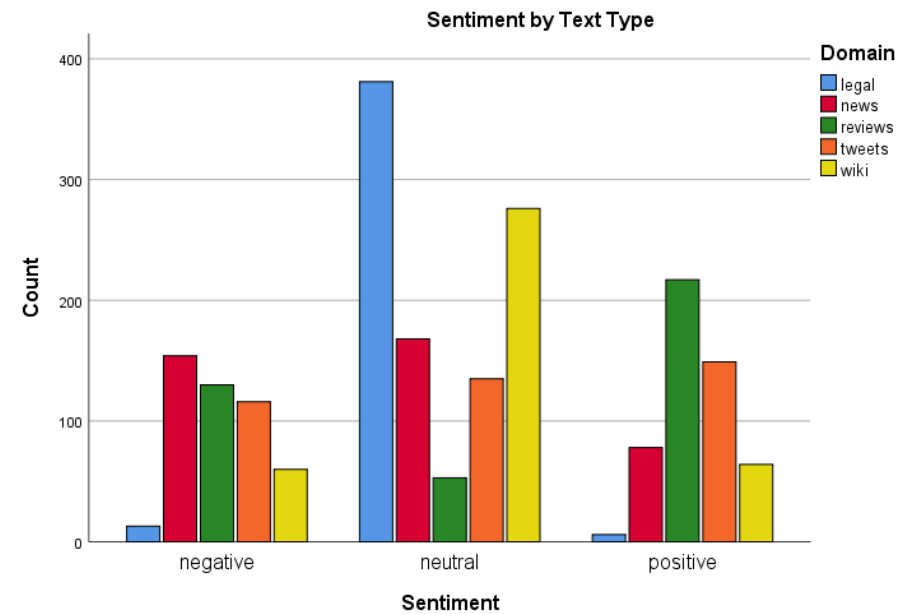
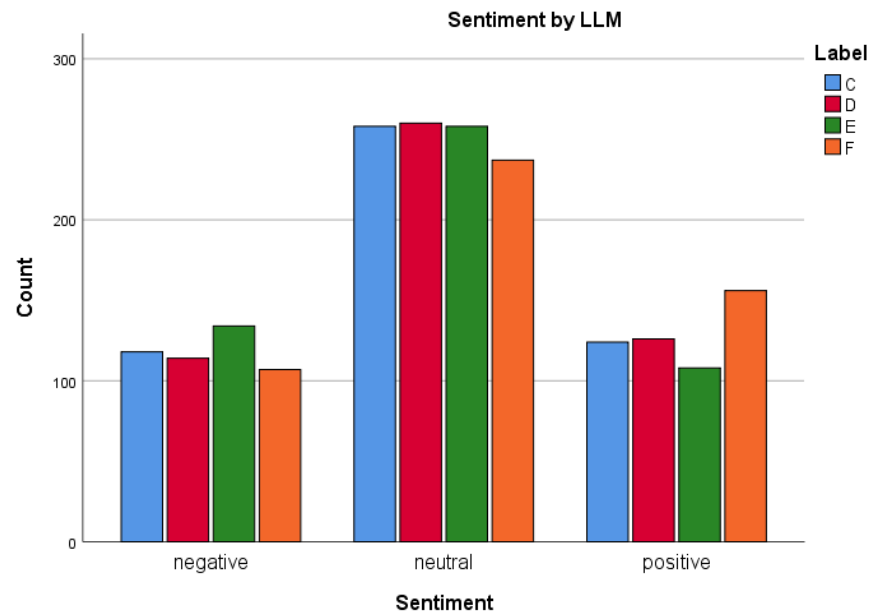
Comparing stylometric profiles of different LLMs

- Corpus compiled by AuTextTification: Shared task that will take place as part of IberLEF 2023, the 5th Workshop on Iberian Languages Evaluation Forum at the SEPLN 2023 Conference.

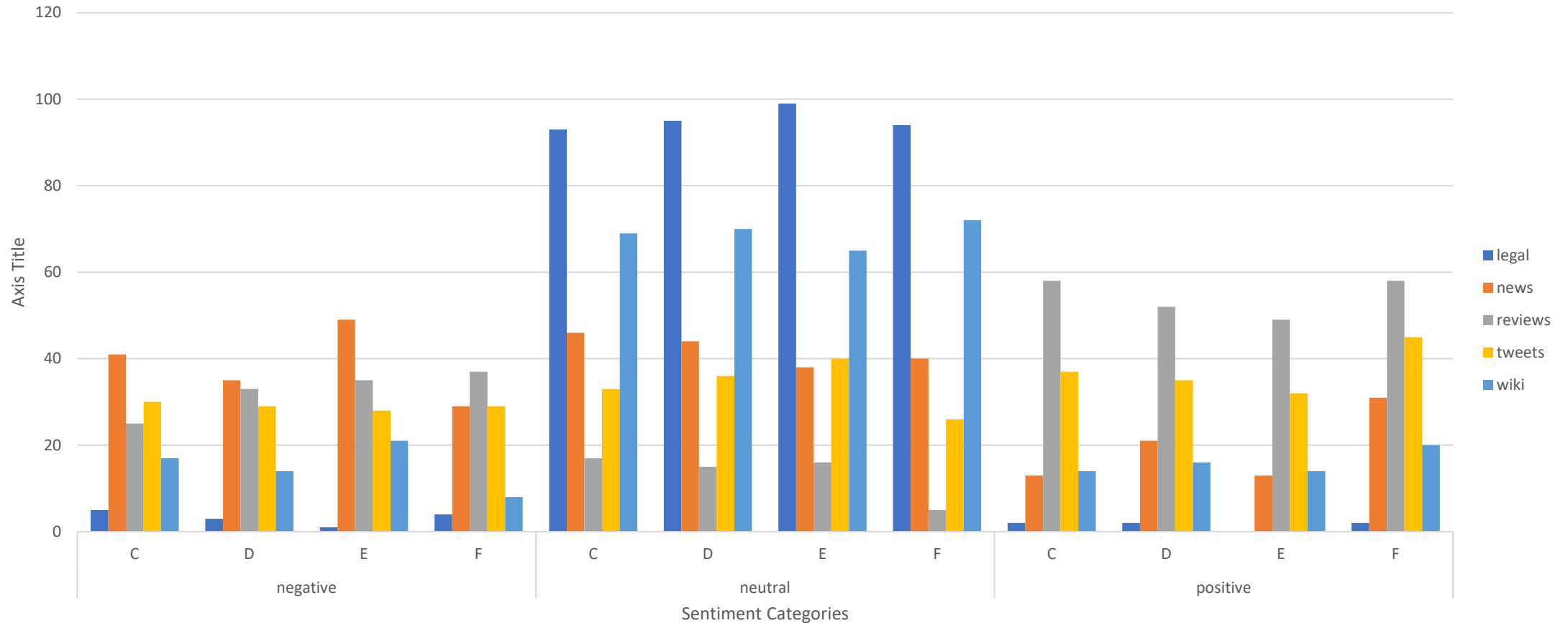
Descriptive Statistics	A	B	C	D	E	F	Total
N of texts	3,562	3,648	3,687	3,870	3,821	3,826	22,414
N of tokens	228,758	232,343	231,729	235,856	229,722	191,004	1,349,412
SD of N of tokens	25.97	25.93	26.48	23.96	24.45	24.85	25.73
Min N of tokens	2	2	2	3	2	2	2
Max N of tokens	97	96	97	97	97	94	97

- "A": "bloom-1b7", "B": "bloom-3b", "C": "bloom-7b1", "D": "Babbage 3b", "E": "curie 13b", "F": "text-davinci-003 175b"

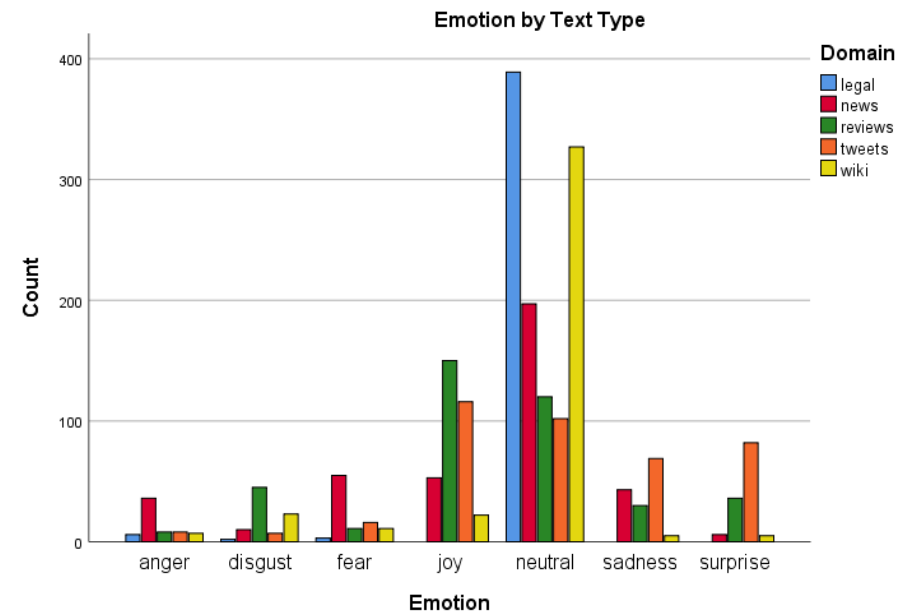
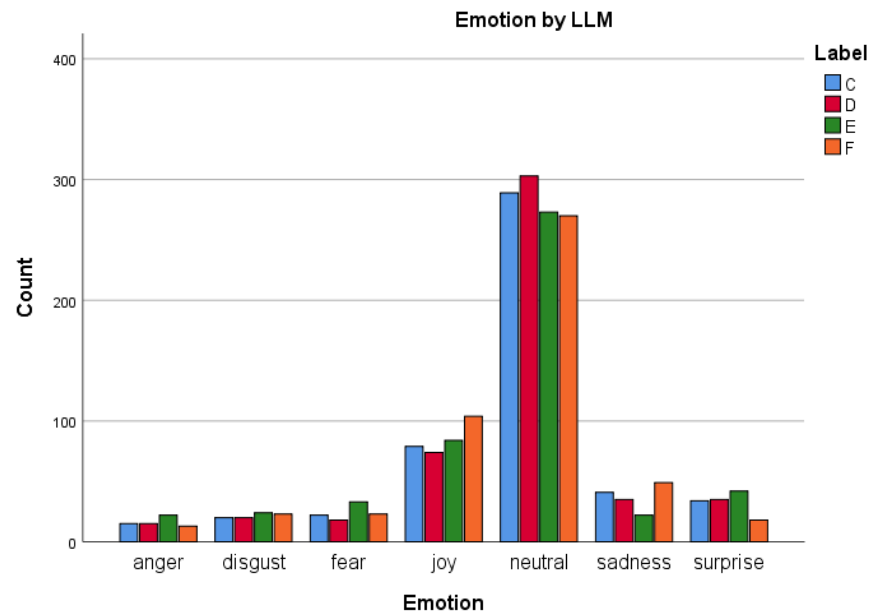
Comparing sentiment across different LLMs



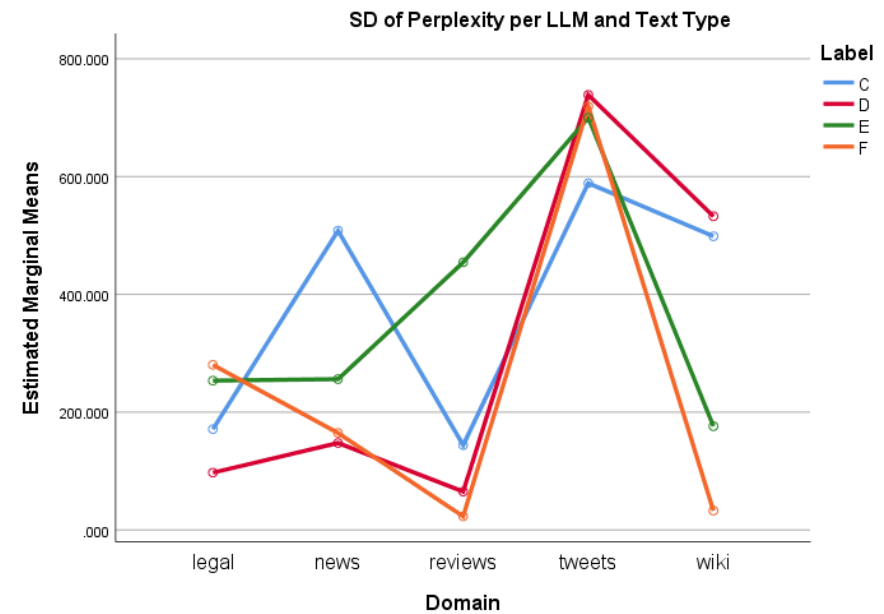
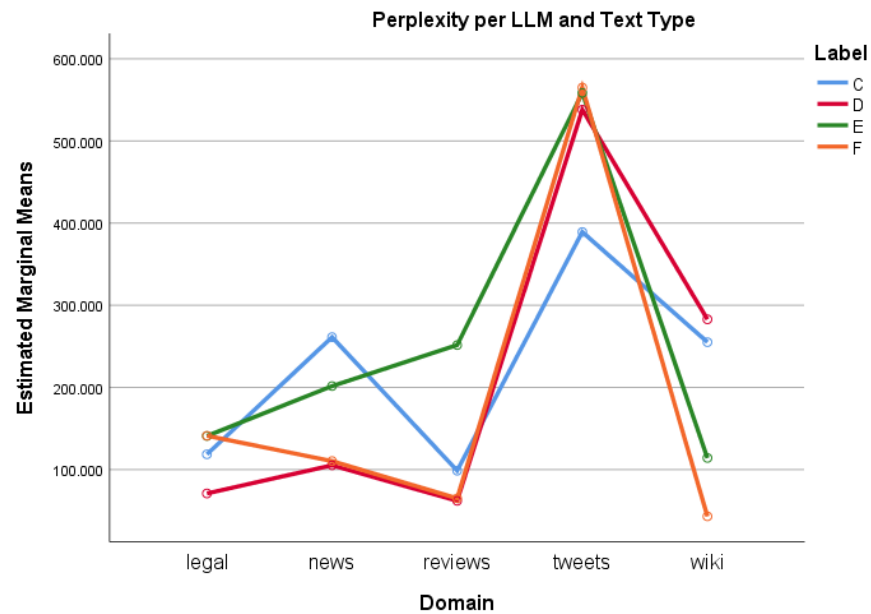
Sentiment per LLM and Text Type

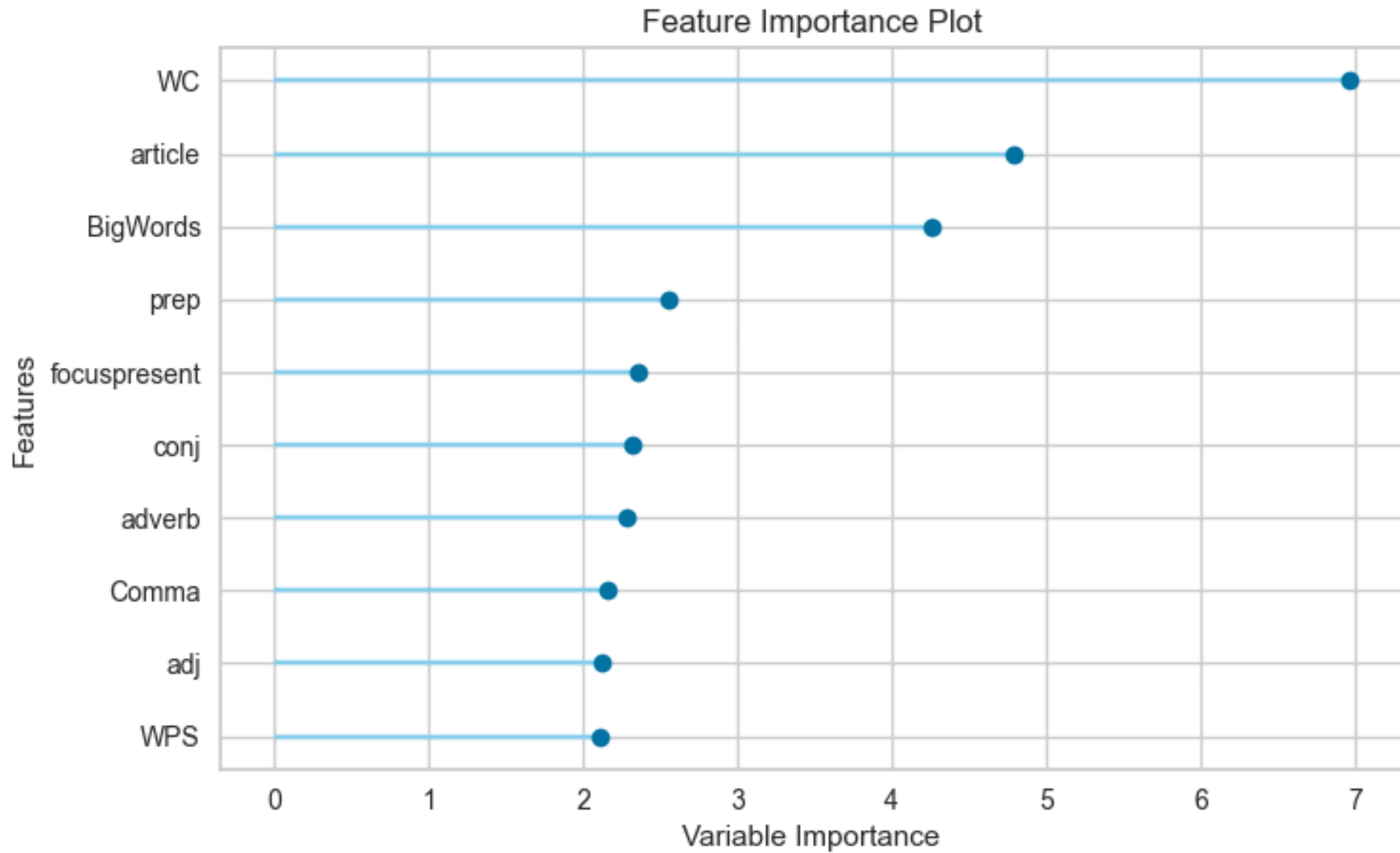


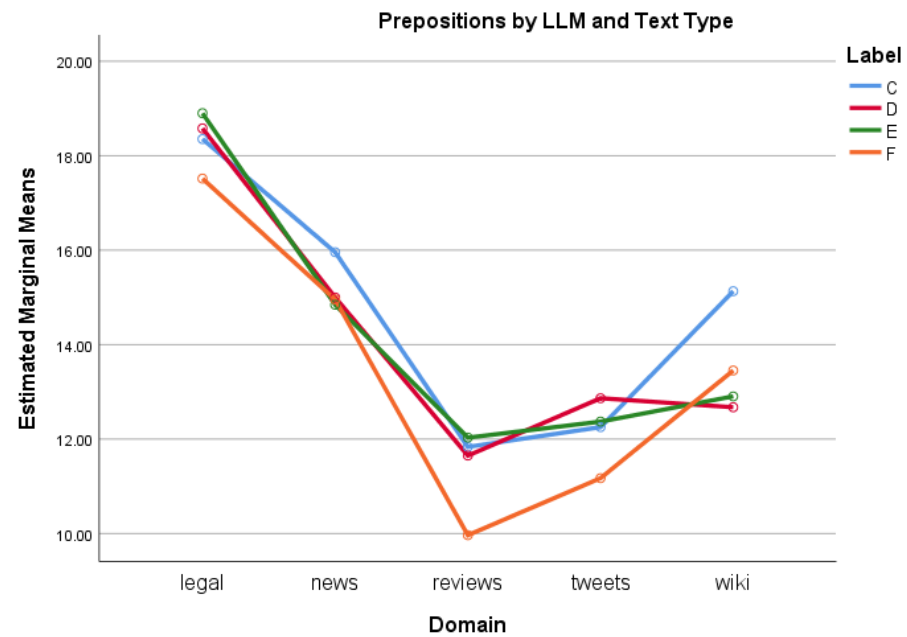
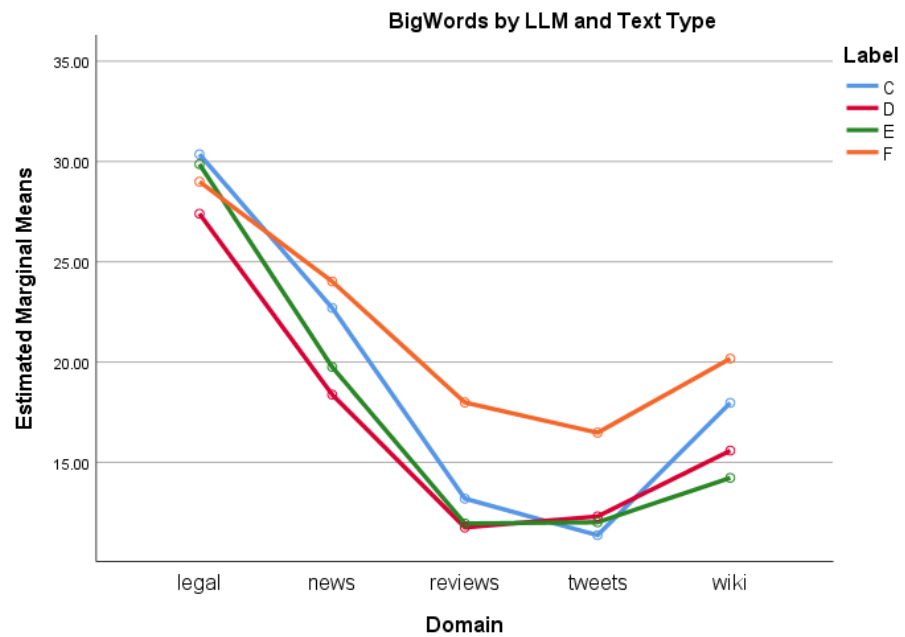
Comparing emotion across different LLMs

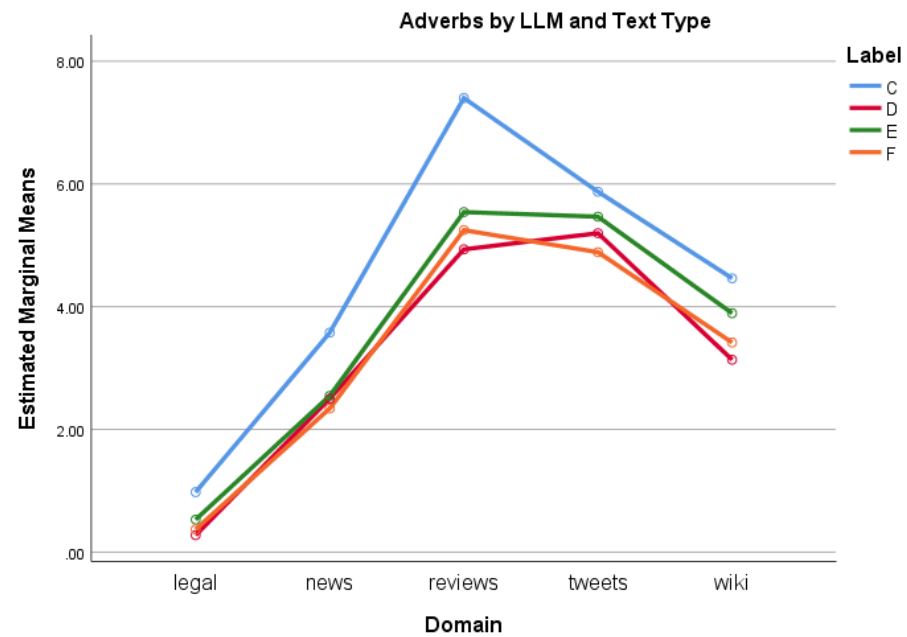
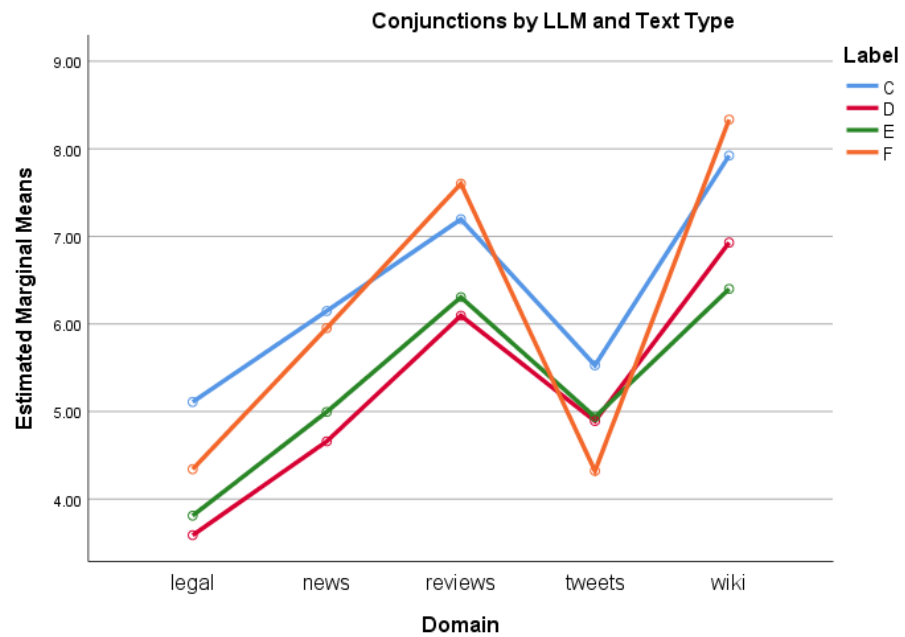


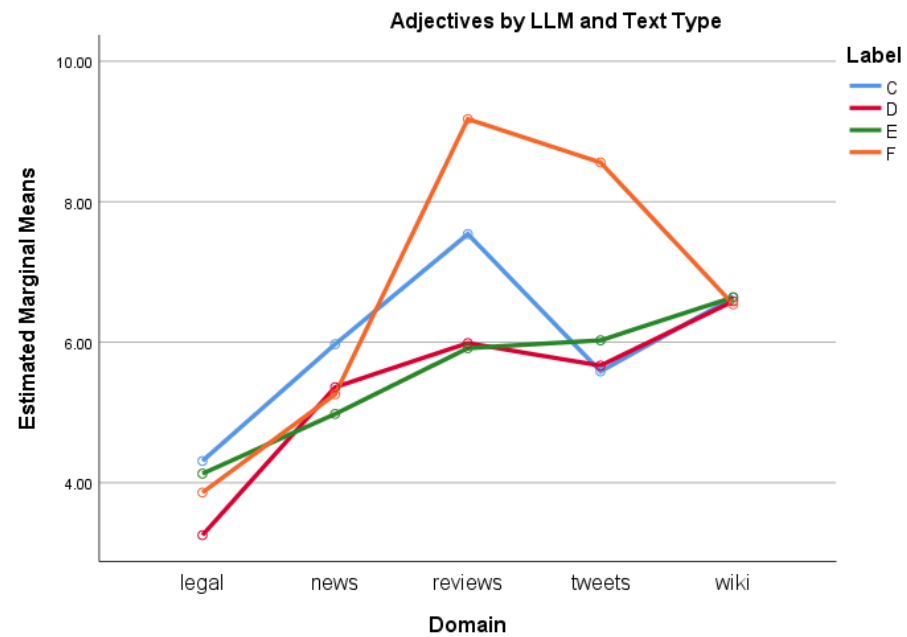
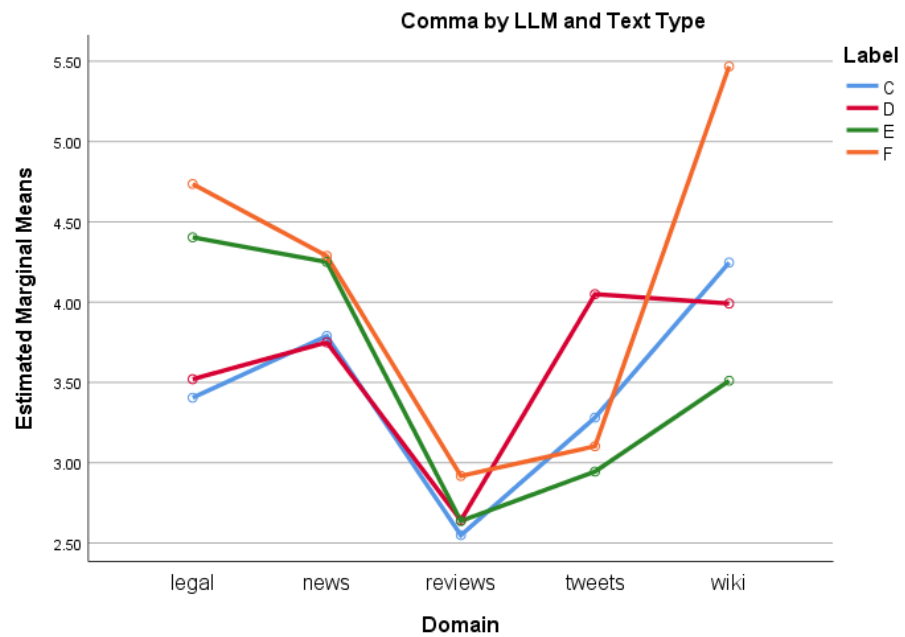
Mean and SD of Perplexity per LLM and Text Type





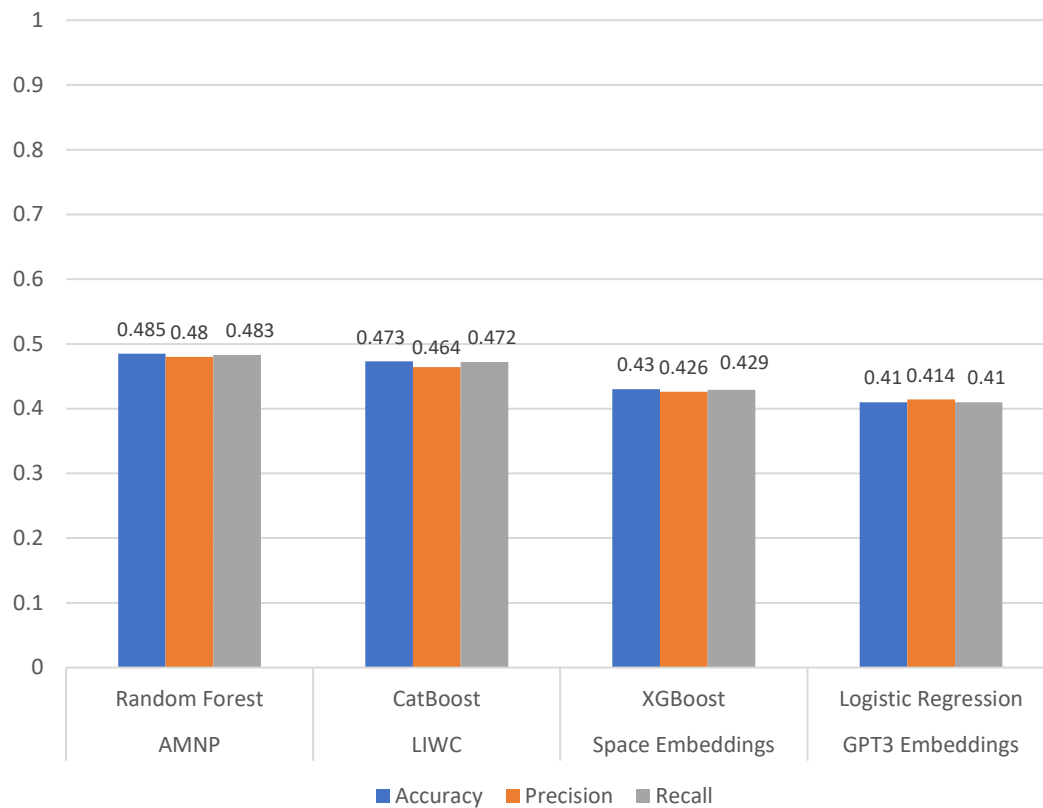




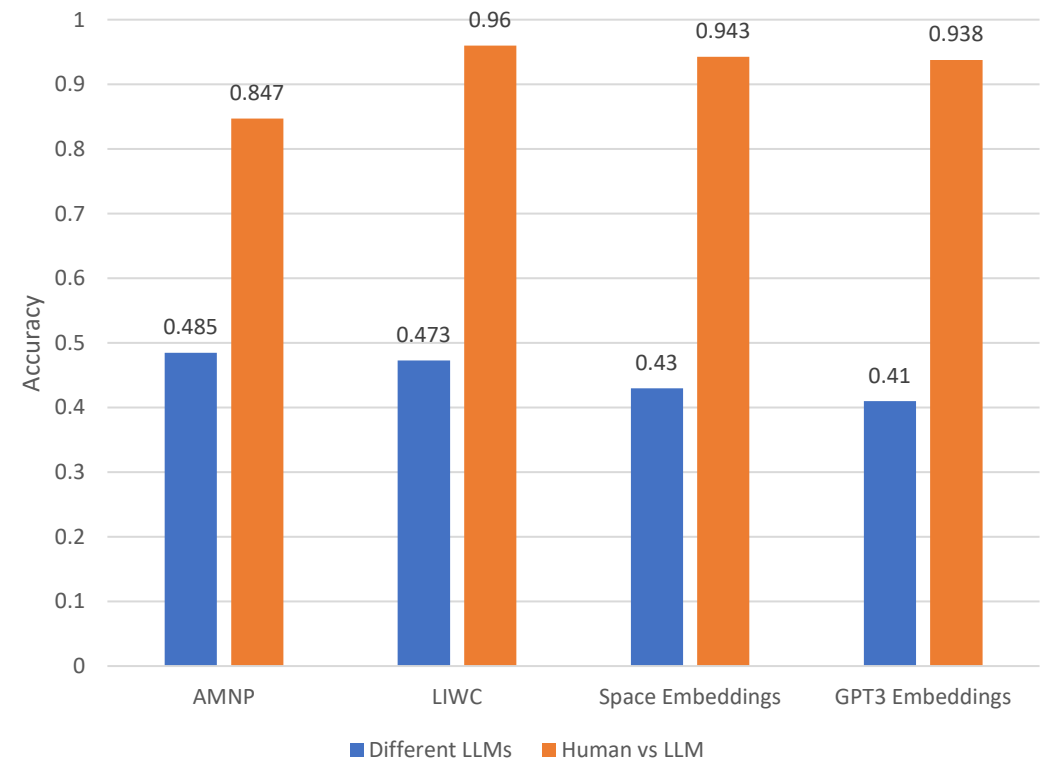


LLMs detection

Classification evaluation in LLMs detection



Classification Accuracy Comparison in Human vs LLM and between LLMs



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 99.999999999% 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.

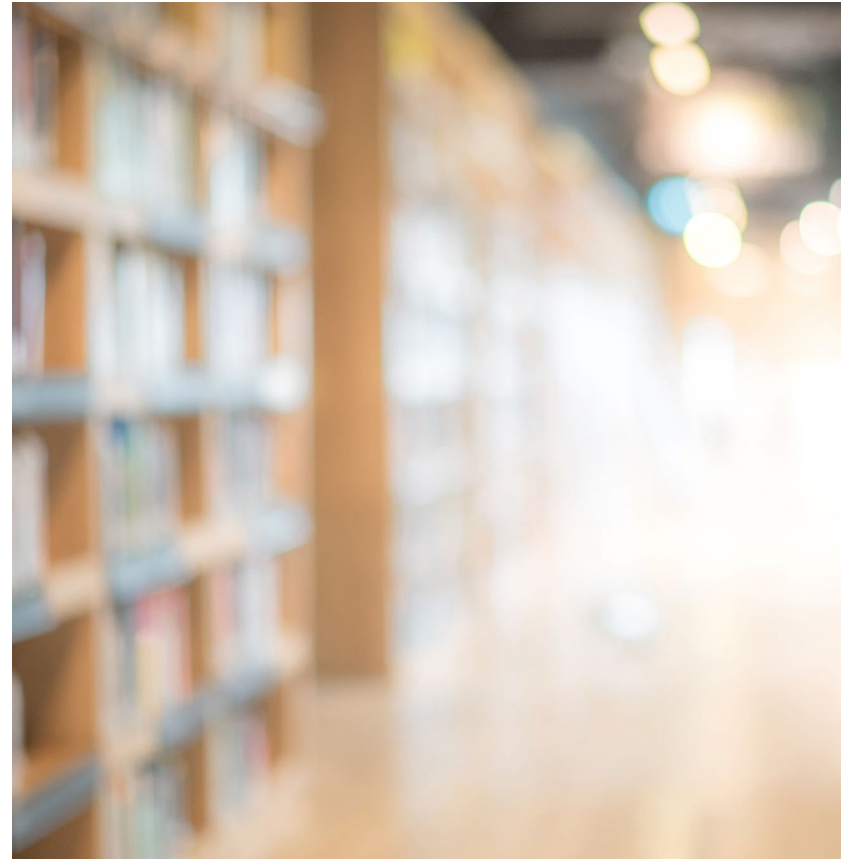
Some thought on using AI-writing detectors in education

- AI-writing detection is **NOT** possible, and it **WILL NOT** be possible as long as humans interact with the output.
 - 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

