Dissecting the Linguistic Fingerprint: A Comparative Analysis of AI and Human-Written News Articles Across Domains

Mason Lonoff

New York University

ml9542@nyu.edu

1. Introduction

Today, Large Language Models (LLMs) are an everyday part of people's lives, but it was not that long ago where life existed without these models. Only in the last few years did LLMs gain such rapid adoption. To put this rapid adoption into perspective, ChatGPT had the fastest adoption to 100 million users in history (Wallisch, 2025). ChatGPT's user base has continued to grow dramatically. Today, according to Sam Altman, ChatGPT has a user base of nearly ~10% of the entire world; this number equates to roughly 800 million people (Paris, 2025). In the US alone, 52% of adults have reported that they use LLMs (Report last link)

What makes LLMs like ChatGPT so popular? The answer to this boils down to how flexible these models are. For example, using ChatGPT's GPT-4o model can produce text, images, and audio to user prompts (OpenAI, 2024). Users take advantage of the endless possibilities LLMs offer. Uses range from creating activities like poetry or songs to planning a trip to writing code (Imagining the Digital Future Center, 2025). Many users use LLMs to learn. For example, in a study by SSRS, 68% of people reported that they use LLMs to learn informally on their own. Additionally, 34% of people said that they use LLMs to get news and political information. As of today, May 9th, 2025, ChatGPT's GPT-4o model states that its cutoff date is June 2024 (Imagining the Digital Future Center, 2025). What happens when someone tries to

learn about what's happening in the world, but the event of interest happened after the model's knowledge cutoff date?

This paper will examine the stylistic behavior of how GPT-4o writes about news events that it hasn't been trained on. Factual hallucinations are a well-known limitation of LLMs. This paper will not analyze *what* the model says, but *how* it says it. The two main research questions this paper will address are:

1. How does GPT-4o write about current news stories that occur after its knowledge cutoff date, and how does this generated content differ from human written news articles?
2. What specific, linguistic, stylistic, and/or structural features contribute most to the differences between AI generated content and human written content?

The paper will proceed as follows: Section 2 described the methodology that was used to gather and generate the data. Section 3 presents and describes the stylistic metrics that were analyzed. Section 4 provides comparisons on results. Section 5 discusses important features, and Section 6 concludes the paper.

2. Data and Methodology

This paper compares human written articles to articles generated by GPT-4o about topics outside of its training range. The goal is to isolate and analyze differences in style or structure between these two sources while at the same time, controlling for topic and prompt consistency. Acquiring my data was done in two portions.

2.1. Collecting Human Written Articles

To build out the human dataset, I used GDELT Doc API to scrape news article metadata across a range of various categories. These categories are Economics, Politics, Health, Law, Artificial Intelligence, Climate and Conflicts. Each of those categories were further drilled down into specific topics. For example, “stock market” and “inflation” were topics in the “Economics” category. Every article was pulled between 10/01/24 and 4/01/2025 which is beyond GPT-4o's cutoff point. I limited results to domains from only trustworthy sources that were not behind a paywall such as Reuters, AP News, and BBC. I made sure to filter for titles that had a length of at least 30 characters, more than five words, and were written in English in order to provide the LLM a suitable headline to generate text about. Additionally, I also filtered out headlines that were too opinionated, headlines that were clickbait, reviews, transcripts, briefings, and advice.

The scraping script worked on pulling data one day at a time, over the course of weekly intervals to stay under GDELT’s API limits. To extract article content, I used the trafilatura library. To ensure that the articles had a coherent structure, I only scraped articles that had at least a length of 100 characters and at least 3 sentences. After collecting the articles, I removed any duplicates by title which resulted in a clean set of human written topics across categories and sources.

2.2. AI Generated Content

For each cleaned human written article, I used an API through OpenAI to use GPT-4o to generate corresponding AI-written versions. For each headline, I prompted the model with:

> "You are a journalist writing a full news article based only on the headline below. The event occurred after April 2024, and you have no access to real-world information about

what actually happened. Do not include a byline or dateline — just write the full article text. "

This prompt was designed to ensure that the model generates plausible articles based on the style of news reporting without relying on information it knows to be true. I looped through each title, called the model with temperature = 0.7. The max_tokens variable was set somewhere in between the mean and median token counts per article for the human articles grouped by category. I stored each generated GPT-4o article along with the original headline, URL, and human content. I made sure to add a time delay between the API calls, and to store intermediate progress (every 50 rows) to prevent data loss. Once the data was generated, I saved it to a csv. At the end, I had seven csv files that had all the information for a specific category.

News articles were chosen as the format to mimic because they provide a standardized and structured form of writing. News articles typically follow certain conventions in tone, structure, and style. By aligning both sources to this, I was able to ensure that any stylistic differences could be attributed to the source of the authorship.

2.3. Finalized Data

To ultimately prepare my data for analysis, I added a category column to each csv saved from the last step. I followed the same procedure for the next csv and then concatenated them together. I did this for each of the seven csv files so that I would have one large dataset. My final dataset consisted of 3295 rows and 6 columns across all seven categories (6590 different articles).

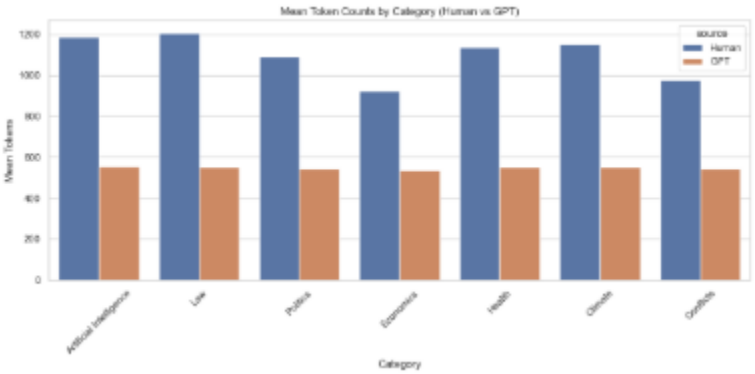
3. Presentation of Results

3.1. Overview of Approach

The goal of this paper was to identify stylistic, structural, and linguistic differences between AI and human news articles. This section extracts 20+ interpretable features from the text that reflects writing behavior and style. To keep the section focused on understandable, I group the findings into distinct analytical domains.

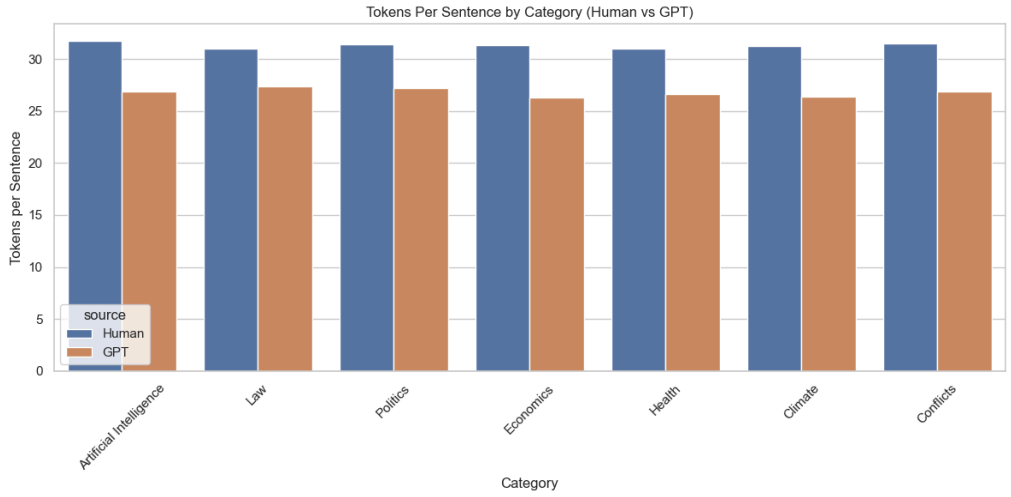
3.2. Structural Features

One of the major differences between human and AI articles can be found in the basic structure of the articles. The human generated articles were much larger than the gpt-generated articles. Even though the max tokens parameter I set was usually set above 850, the results show a consistent token count around 575. The image below shows how much smaller the average



token count per article is. Similarly, the average sentence count for each article has a similar behavior. GPT seemingly maxes out at ~20 sentences for each article while human articles are in the range of 30-40 sentences. This comparison isn't statistically valid as a straight up comparison since the lengths of each article is not normalized. However, this analysis showed underlying characteristics of GPT-4o to automatically cap itself at certain lengths regardless of input or category. Additionally, the tokens per sentence values for AI articles are much more similar to the counts for human articles. The image below shows how similar these counts are. It appears as though much of the cause for uneven token and sentence counts stems from GPT-4o implicitly setting maximum token counts.Upon analyzing how title length impacts token usage in AI articles, it was found that title

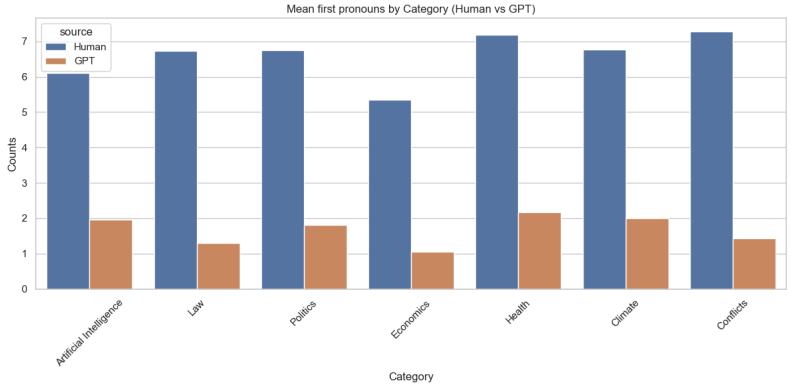
length and token counts for AI articles have a correlation of 0.08, so it's very unlikely that the headline prompt is responsible for that.



3.3 Parts of Speech

GPT-4o and human articles also vary on their specific use of parts of speech and stylistic devices. Human articles use nouns, verbs, and adverbs more while the AI articles use adjectives more often. These differences were all found to be statistically significant with p-values less than 0.005 except for the Economic category in the verb use percentage analysis. I used theWilcoxon rank sum test. To test the effect size of the statistical differences, I used a Wilcoxon Effect Size test. Adjectives were shown to have a large effect size with the minimum being 0.775 (r link).

Pronoun usage was also revealing. As seen in the image below, human authors used first person pronouns much more often than AI authors did. This is backed up by every category showing statistically significant results between sources. Human authors also used more third



person pronouns than GPT-4o did. The difference in usage is not as drastic as first person pronouns, but it still indicates that GPT-4o tries to avoid referencing any hypothetical person even indirectly.

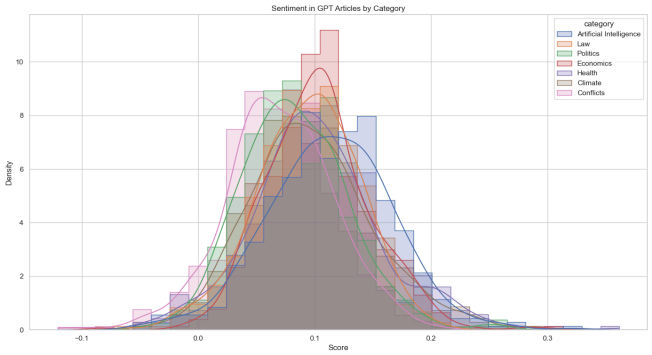
Use of passive terms was also slightly higher in human articles, but not every category showed statistical significance between sources. The only category that AI authors actually used more passive terms in was in Economics. Contraction counts were also much higher in the gpt-generated articles. This is a peculiar result considering the reading complexity score analysis that is coming up.

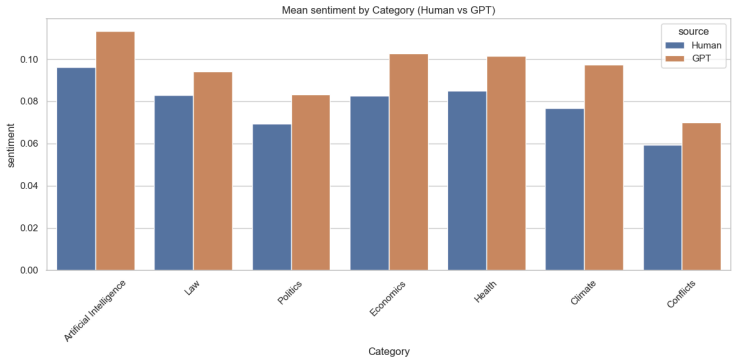
3.4. Named Entity Usage

This section analyzes how often authors are specifically mentioning important organizations, locations, or people. I used a transformer-based NER model to count mentions of the respective entities across each article. I made sure to normalize the data by making the entity counts per 1000 tokens as article length was a confounder. Across each entity, human articles saw a much higher usage rate. It's understandable that AI authors don't have the context to know all relevant organizations, locations, or people for a story just based off of a headline. However, it does align with the previous findings where the AI articles shied away from using pronouns, so it's further evidence that the model does not like to reference specific people.

3.5. Tone and Subjectivity

The images below display how sentiment in GPT articles is generally more positive. While the tone is still generally neutral, GPT-4o's sentiment is statistically significant across each category. GPT-4o wrote about AI topics in a more positive way than it did for any other category. Unsurprisingly, Conflicts also had the lowest sentiment score. For subjectivity (how





opinionated the text is), AI always had higher subjectivity than human generated content with all the categorical differences being statistically significant. The GPT-4o models surprisingly had higher subjectivity for AI articles compared to other categories. It's interesting that AI models write about AI articles in an opinionated and positive way.

3.6. Readability Metrics

I used the Flesch-Kincaid Grade Level (FKGL) and the Gunning Fog Index to analyze how complex the articles are. Both these scores estimate reading difficulty based on sentence length and word complexity. The GPT-4o models consistently ranked higher in complexity than human content. This indicates that while the articles are shorter in length and content, the actual articles are more dense. This is further supported by GPT-4o articles having a higher TTR ratio (lexical richness) score than human articles.

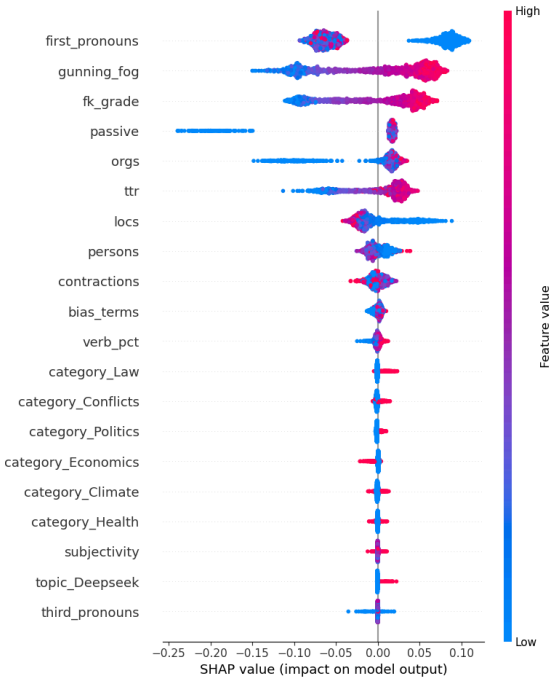
4. Comparative Analysis

This category will focus on inter-category analysis. Firstly, in tokens per sentence, Law and Politics have wider, and thus, higher tokens per sentence. This could be due to these industries being more inherently verbose. However, in contractions, Law and Politics see the highest average count. High contractions indicate that an article isn't verbose or as complex.

5. Classification & Feature Interpretation

In this section, I used a random forest model to classify the source of each article. As features, I used each of the 20+ features created previously. Those features include the readability metrics, the TTR ratio, named entities, contractions, pronoun usage, passive language usage, parts of speech usage, sentiment, subjectivity, bias term counts, and dummy variables for each topic and category. I did not include the token columns since they're skewed by article length. If all human articles are longer than GPT-4o articles, it's clear which source it came from.

The model performed extremely well. It has an F-1 score of 0.99 which is tremendous. Through cross validation, I was able to prevent overfitting as well. Once the model was done, to better understand which specific attributes distinguish GPT generated content from humans, I will use SHAP values to interpret the model's predictions. SHAP is a . It assigns importance values to each feature for each individual prediction. The results are in the table below. It shows several patterns. First, readability scores were some of the most predictive features. Also first person pronoun use and specific entity naming were also predictive. This all makes sense and tracks with the logic that gpt articles don't like to be specific. At its core, reading complexity and directness of the writing are the most definitive features to tell if an article is human generated or not.



6. Conclusion/Implications

Through this analysis I was tasked with analyzing how GPT-4o writes about news stories it hadn't been

trained on. I was also finding which features distinguish articles from being written by a human or a person. My results found that GPT articles are shorter than the human articles. Human articles also show higher specificity in what they write about. They reference people, places, and organizations directly. GPT articles are more positive, more opinionated, and more complex. It's a surprising discovery to see that AI articles are more emotional and complex. AI articles may not be as useful for easy learning as people may hope for. The main implication here is that factuality aside, GPT generated test is more advanced, dense, and emotionally charged than human written content. For journalists, they need to be cautious about using LLMs to generate text for their articles.

In potential future work, I would like to expand more on what types of articles cause these behaviors. For example, I could tag each article as "opinion" or "analysis". That way I could drill into exactly what type of articles cause these behaviors.

References

Imagining the Digital Future Center. (2025, March 12). Close encounters of the AI kind: Methodology and topline findings. Elon University. https://imaginingthedigitalfuture.org/wp-content/uploads/2025/03/ITDF-LLM-Report-topline-3-12-25.pdf

OpenAI. (2024, May 13). Hello GPT-4o. https://openai.com/index/hello-gpt-4o/

Paris, M. (2025, April 12). ChatGPT hits 1 billion users? 'Doubled in just weeks' says OpenAI CEO. Forbes. https://www.forbes.com/sites/martineparis/2025/04/12/chatgpt-hits-1-billion-users-openai-ceo-says-doubled-in-weeks/Forbes+2Forbes+2Forbes+2

Wallisch P. (2025). Week 2: Big Data [PowerPoint slides]. NYU Brightspace. https://brightspace.nyu.edu/d2l/le/lessons/398308/topics/10730384