SPRING 2024

DATATHON EXAMPLE REPORT

WRITTEN BY DEPAVIA. A, 2023

1ST PLACE - PHD DATATHON 2023

CONFIDENTIAL

clickbAIt

HOW WELL CAN AI GENERATE A GOOD HEADLINE?

Abstract

"AI won't replace people–but people who use AI will replace people who don't" declares IBM's report on AI-augmented work [1]. In a historical turning point, it is now becoming possible for everyday workers to develop and employ state-of-the-art machine learning methods to handle tasks of surprising complexity. As media firms look to the future, they must decide not *whether* they should incorporate AI methods in their workflow, but rather *how* they will do so. Rather than replacing human writers entirely, one attractive pathway is augmenting human creativity with suggestions and revisions from large language models. In this report, we explore how the edits provided by AI methods compare to those of expert human writers. The Upworthy Research Archive is a unique resource for studying revisions proposed by humans.

We find that the most significant differences between ChatGPT headlines and human-written headlines lies in their linguistic complexity, rather than differences in vocabulary or tone of voice. Moreover, these differences are nuanced: while ChatGPT uses fewer words than human writers, the headlines it produces are more difficult to read. We also examine how instructions given to ChatGPT influence the headlines produced. We find that specific instructions can change the style and tone of resulting headlines, but that the differences in linguistic complexity between ChatGPT- and human-written headlines persists across different prompts. These findings suggest that human writers still have an edge over large language models in creating easy-to-read headlines that can be quickly digested, and indicate that media firms should pay attention to linguistic complexity when choosing large language models to incorporate into their business model.

Introduction

One of the most revolutionary developments in the past year has been the popularization and mainstreaming of large language models, such as ChatGPT [5]. Automated language production enabled by large language models is much cheaper than employing large teams of human writers–but do current AI models produce work which is comparable in quality to human writing? Given the exploding popularity of ChatGPT, this report focuses on output from ChatGPT 3.5.

We study these questions using headlines from the Upworthy Research Archive. Upworthy.com emerged as dominant internet content host due to innovative A/B testing practices and a clear vision for attractive content [2]. They popularized a distinctive article style, often compared to so-called clickbait, which became iconic. Their company goals evolved significantly over time, partially in response to outside pressure from other social media platforms, including moves from Facebook.com to suppress content which resembled Upworthy's distinctive style [2, 4].

Nontechnical Summary

Main Questions

Our investigations follow two central questions:

1. *How do headlines produced by ChatGPT differ from human-written headlines?*
2. *What is an effective way to prompt ChatGPT to design a "human-like" headline?*

Key Findings

We find that headlines produced by ChatGPT primarily differ from human-produced headlines in their linguistic complexity, rather than in writing style or tone. We develop several measures that capture headlines' (a) style, such as punctuation and capitalization, (b) tone, e.g. positive, negative, and neutral sentiment, (c) and complexity, including reading-level difficulty and word length. We use logistic regression to predict whether a headline was produced by a human or by ChatGPT using these metrics. All of the most significant features identified by the model are all complexity measures, i.e. features from category (c).

We then test several different "prompts" and examine how they influence the headlines produced by ChatGPT. Our prompts fall under

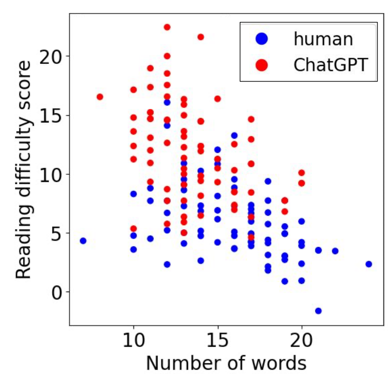

Fig. 1: Human-written headlines use more words but are linguistically simpler. Comparing human-produced headlines from the Upworthy Research Archive (blue) to corresponding headlines generated by ChatGPT (red) by word count and reading level. For full discussion see Technical Summary.

four categories: a generic headline request, a request for a "clickbait" article headline, a request for a headline using emotional language, and a request for a headline "in the style of Upworthy.com." We find that different prompts produce statistically significant changes in the headlines generated: using a permutation test, we find that switching from a generic prompt to a clickbait prompt produces a headline with higher word count and–perhaps unsurprisingly–more exclamations.

Comparing the ChatGPT headlines to those from Upworthy.com, we find that across all prompts, the ChatGPT headlines exhibited significant differences from those found on Upworthy.com. In particular, all prompts produce headlines which are particularly distinct in complexity from those on Upworthy.com, and specifically asking the large language model to mimic Upworthy's style does not reduce this gap. When we compare headlines from Upworthy.com to the "Upworthy.com style" ChatGPT prompt headlines, we find that while Upworthy's articles use more words than the ChatGPT generated ones, the headlines Upworthy.com produced target lower reading levels.

Technical Summary

Data Exploration and Preprocessing

The data in the Upworthy Research Archive describe A/B tests conducted on Upworthy.com between January 2013 to April 2015. Initial exploration had several goals, including:

- To understand and validate the representations for each "headline-image" pair tested.
- To examine how often headline edits produced dramatic changes in users' interactions with articles.
- To look for large-scale changes in the data over time.

To contextualize results, initial data exploration was complemented by reading outside sources, including David Karpf's book "Analytical Activism" and the announcement of the Upworthy Research Archive on Nature.com, to better understand Upworthy.com's A/B testing practices and larger internet context during this time period [2, 4]. For a full list of data features and their explanations, we refer readers to The Upworthy Research Archive webpage.

Data on the archive are organized into *packages*, which are "bundles of headlines and images that were randomly assigned to people on the website," and *tests* which are collections of packages [4].

However, exploration quickly revealed that there is not a one-to-one correspondence between entries in the dataset and unique interventions studied during A/B tests: on Upworthy.com, the only attributes which were displayed as part of an A/B test were the image and headline. However, in the baseline dataset from the Upworthy Research Archive, multiple distinct entries appear which differ only along attributes which were *not* part of the test, such as the leading sentence of the article or the internal link used to represent the article on Upworthy.com. In order to focus only on the impacts of A/B testing, entries with the same headline, image, and test ID were aggregated into a single unique entry, though standard deviations in impressions and click counts were retained. We also subselected for tests across which all images were held the same, as the Upworthy Research Archive does not contain image files or descriptions of images, and so this analysis was unable to assess their impact.

In order to normalize for varying number of page views, which exhibit long-scale fluctuations over time, we defined *click-through percentage*, capturing the rate at which users who were (potentially) exposed to a headline/image pair clicked on that article. In general while we found that most headline edits produced small changes, a nontrivial population achieved outstanding improvement in click-through percentage which is in line with anecdotal reports about early days of Upworthy A/B testing practices [2].

In the years from 2013 to 2015 the landscape of Upworthy.com articles changed dramatically over time, both due to shifts in internal policies at Upworthy and to external social media actors, including Facebook [2]. In order to control for these confounding variables, in later sections we focus our analysis on A/B tests conducted during the third quarter of 2014, the quarter with the most tests recorded in the Upworthy Research Archive.

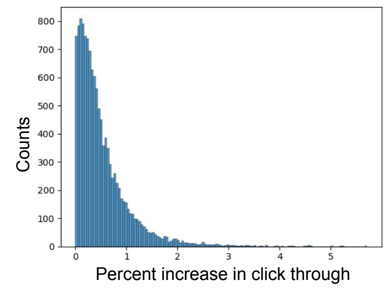

Fig. 2: Histogram displaying the increase in click through percent between article headlines and the best-performing headline in their test. Most headline revisions produced relatively small changes in user's tendency to click on articles, but significant outliers exist.

Headline Embeddings

The research questions we were interested in required that we first develop a quantitative method for comparing headlines. A *vector embedding* of a headline converts a given article headline into a series of numbers: the key question is how to generate those numbers and what they should represent. In order to reason about our research questions, we required embeddings with the following properties.

- Interpretability. One should be able to describe intuitively how the vector embedding of an article headline relates to the text.
- Deterministic. Every time we compute the embedding of a fixed headline, we should receive the same vector representation.

- Reasonable dimensionality. If the number of dimensions in our embedding is large compared to the number of available samples, additional statistical complexity is introduced. Given our desire to study data augmented by ChatGPT output, which may be time-consuming to produce, we are even further restricted in the number of samples at our disposal.

Given these criteria, we identified two families of embedding methods suitable to our needs. The first was a bag-of-words model which produced an embedding based on pre-trained `word2vec` vectors, and the second was a feature engineering approach based on textual and sentiment analysis. We experimented with embeddings from both families. However, the first approach had a significant drawback: it was incapable of capturing word order or grammatical information. For this reason we ultimately opted for latter approach.

We based our feature engineering on existing methods, initially designed for use with news articles. We first sub-selected for features with well-documented interpretations, and then refined our embedding by discarding any features along which our data had zero or trivial variance. Our final embedding consisted of 15 features, broadly grouped into categories reflecting headline style, headline tone, and headline complexity, as summarized below in Table 1.

Category	Feature(s)	Description
Style	'exclaim'	Exclamations
	'quotes','allpunc', 'allcaps'	Character features
Complexity	'flesch_kincaid','coleman_liau','smog'	Reading level
	'ttr', 'avg_wordlen', 'word_count'	Word features
Tone	'bias_words', 'vadneg', 'vadneu', 'vadpos'	Sentiment

Table 1: Embedding features

Producing Headlines with ChatGPT

One of the most interesting challenges in our research process was: how could we produce "article headlines" from ChatGPT which would be comparable in topics to those on the Upworthy Research Archive? The topic of an article has a dramatic influence over the vocabulary and style of its headline. To give two examples common in the Upworthy Research Archive, an article about a cancer patient is likely to use very different language than an article about cute dogs. Topic modelling is a vast field, and the question of how to detect, categorize, and summarize the relevant topics reflected in the Upworthy Research Archive would constitute its own fascinating research project. Moreover, even controlling for topic, the angle of an article is also a dramatic confounder for the headline style: consider for example an article about a cute dog with cancer as opposed to an article about a cute dog who plays sports.

We chose our solution based on the use cases we wanted to study. In this report, we ask what the role of large language models will be in *augmenting* human article headlines, not replacing human writers. As a proof of concept, we analyzed a simple method for incorporating ChatGPT into headline production: we provide ChatGPT with an existing (human-written) article headline from the Upworthy Research Archive, and then ask ChatGPT to suggest an improvement of that headline. This method allows for a one-to-one comparison between human versus AI generated articles, and opened the door further to questions about how the improvements suggested by ChatGPT compare to successful human revisions.

Creating the Dataset

Due to our limited timespan we decided to manually create the dataset of gpt-augmented headlines, as we were unable to identify any free and tractable methods of automating queries to ChatGPT. We followed the following steps to produce our datasets:

1. We subselected for packages in Q3 of 2014, and further subselected for tests in which at least two distinct packages occurred.
2. We randomly subsampled tests fitting these criteria and compiled their packages into a single dataset, ultimately containing 78 distinct headlines. We created a csv file containing the test IDs and headlines of these packages.
3. We decided on four distinct prompts to ChatGPT, discussed below. For each prompt, we made a separate copy of the csv file and added a transformed column containing the full text of the prompt to be input to ChatGPT.
4. Using ChatGPT3.5, we sequentially copied over the exact input text for each csv entry, and recorded ChatGPT's response for each headline.

The restricted size of these gpt-augmented datasets, as a result of manual creation, is a major limiting factor of our analysis. An important future research direction would be to compile much larger datasets of ChatGPT-improved headlines for investigation.

Choosing Prompts

ChatGPT output can be extremely sensitive to the "prompt" or input text received by the model. Effective use of large language models in augmenting human writing must involve careful engineering of prompts, so we next investigate how changing the prompt to ChatGPT influenced the resultant article headlines.

Based on outside reading and initial data exploration, we developed three different hypotheses for the question, "What prompts will produce headlines most similar to those on the Upworthy Research Archive?" The first, based on Upworthy's well-documented relationship to "clickbait" content, was that telling ChatGPT to produce titles "in the style of clickbait" would produce headlines more similar to the Upworthy Research Archive compared to some generic baseline instruction. The second hypothesis, based on data exploration with the sentiment features, was that telling ChatGPT to "use emotional language" would produce more similar article titles. The third hypothesis was that, given ChatGPT's access to internet data and familiarity with popular culture phenomena, telling ChatGPT directly to imitate the style of Upworthy.com would produce the most similar headlines.

We developed the following baseline input to ChatGPT. To create different prompt types, the ____ was replaced with specific instructions, as detailed below in Table 2.

> Suggest an improved title for the following headline ____. [ARTICLE TITLE GOES HERE].

For the 'generic' prompt the ____ was omitted altogether.

Prompt Type	Value of ____
generic	N/A
clickbait	"in the style of clickbait"
emotional	"using emotional language"
Upworthy-style	"in the style of Upworthy.com"

Table 2: Prompt inputs to ChatGPT

The resultant datasets are included in the supplementary material.

How do headlines produced by ChatGPT differ from human-written headlines?

The first question we aimed to answer was to establish whether the headlines produced by ChatGPT are meaningfully different from human-produced headlines. We first restricted our attention to comparing the 'generic prompt' ChatGPT headlines with human-written headlines. We collected the original human-produced headlines used as input to ChatGPT, as well as the resultant gpt-augmented headlines output by ChatGPT, and vectorized both sets using the feature embeddings described earlier in this report. With these vector embeddings, we now consider the distribution of the ChatGPT headlines versus the human headlines in feature space, and ask whether there is a statistically signficant difference between these two distributions in the feature space.

As we did not have evidence to support a belief that the distributions of headlines in feature space followed a standard distribution, we employed a non-parametric method (a permutation test) to examine whether our gpt-produced headlines differed significantly from those written by humans.

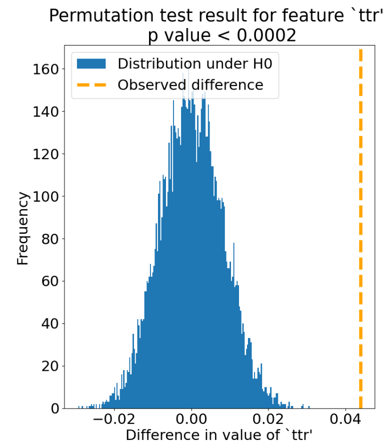

Fig. 3: 'ttr' encodes the ratio of distinct words used to total words in a headline. The observed difference along this feature between ChatGPT and human-written headlines is much more extreme than we would expect if both famlies of headlines came from the same distribution. The positive value indicates that headlines produced by ChatGPT have a significantly *higher* ratio of distinct words used to total words compared to human-written headlines.

Using Permutation Tests to Compare Headlines

The intuition underlying permutation tests is very simple and powerful: if two datasets X and Y are very similar, then if I take a random sample of data from X and mix it with a random sample of data from Y, the resultant dataset should look very similar to both X and Y. However, if X and Y are very different, then random mixtures of their data will not be very similar to the original X and Y.

More precisely, given two datasets, the null hypothesis H_0 is that samples in both datasets are drawn from the same distribution. A *permutation test* is a simple, non-parametric method which compares the differences observed in our data to the differences that we would expect assuming the null hypothesis is true [6]. The permutation test simulates the distribution of the test statistic under the null by repeatedly taking random reshufflings of the samples between the two datasets and computing the test statistic.

We performed permutation tests to compare the feature embeddings of human-written headlines to those of headlines produced by ChatGPT in response to the generic prompt. Our test statistic was the difference in average feature embeddings (each a numeric vector of length 15). This effectively ran a permutation test simultaneously along every feature in our embedding. Testing along every feature introduces concerns about multiple-testing, which we address below.

Our permutation tests concluded that enough evidence existed to reject the null hypothesis, namely to conclude that ChatGPT headlines follow a distinct distribution compared to human-written headlines. We performed permutation tests with 10,000 re-sampling steps and found statistically significant differences between ChatGPT headlines and human-written headlines along 7 of our 15 features. Interestingly, all of these were complexity features, and all had high statistical significance ($p < 0.0002$). A single style feature did have liminal statistical significance ($p < 0.01$), but because testing every feature simultaneously raises the possibility of spurious findings due to multiple testing, we used an adjusted threshold to determine significance ($p < 0.05/15 = 0.0033$) and so discarded this discovery. Figure 3 on the right illustrates the result of the permutation test on feature 'ttr,' which measures the ratio of distinct words used in a headline to total number of words used. A histogram of the simulated distribution of the difference under the null hypothesis is plotted in blue, where as the true difference in 'ttr' value observed in the data is plotted in orange. The fact that the orange line falls far outside of the blue distribution indicates that our ChatGPT headlines tend to have a higher fraction of distinct words.

Which Features Best Predict GPT versus Human Headlines?

Having identified statistically significant differences between ChatGPT headlines and human-written headlines, a natural question arose: given the feature embedding of an article, which features are most useful for predicting whether or not it was generated by ChatGPT? In order to answer this question, we needed to select a model for classifying articles as "ChatGPT" versus "human-written" based on their embedding. As this prediction task is one of binary classification, we selected a logistic regression model. We randomly split the data into training and test sets of equal size, and fit logistic regression models on the training set, and then evaluated their performance by measuring the accuracy of their classifications on the test set.

When given the full feature embedding, logistic regression succeeded at predicting labels with high accuracy, achieving $84.6 \pm 0.4\%$ test accuracy and $90.3 \pm 0.3\%$ training accuracy, aggregated over 100 random test/train splits. A visualization of these predictions is displayed in Figure 4, in which the correspondence between the true labels and the model predictions can be visually assessed.

More than simply predicting whether a headline was or was not gpt-generated, we were particularly interested in seeing which features would have significant coefficients in the logistic regression model. To assess the significance of the effect of each feature on the prediction, we report the p-value associated with testing the point null that the corresponding regression coefficient is zero. We found that the logistic regression model identified one significant feature and four additional features of liminal significance:

- 'word_count' with $p < 0.003$
- 'smog_index' with $p = 0.019$
- 'coleman_liau_index' with $p = 0.053$
- 'avg_wordlen' with $p = 0.064$
- 'allpunc' with $p = 0.0928$

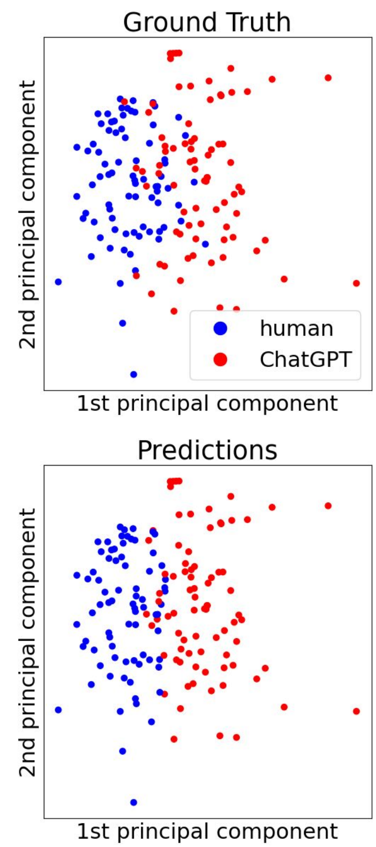

Fig. 4: Using logistic regression to predict whether a headline was generated by ChatGPT or written by a human. Each dot represents one headline, visualized along the first two principal components of the feature embedding matrix. In the ground truth figure, one can distinguish that the distribution of human-written headlines is distinct from those produced by ChatGPT, sitting slightly to the left along the first principal component.

Interestingly, we found that the trend identified by our permutation test repeated; the most significant features leveraged by the logistic regression model again fall under the category of measuring linguistic complexity (‘word_count’, ‘smog_index’, ‘coleman_liau_index’, ‘avg_wordlen’). With this pattern robustly established, we began to ask: “Would these trends persist if we changed the input instructions to ChatGPT?”

How do Different Prompts Affect ChatGPT Output?

In the above analysis, we addressed our first main research question while controlling for input to ChatGPT: we examined differences between headlines produced by the ‘generic’ prompt and human-written headlines. Our second main research question asks how changing input instructions to ChatGPT influences the resultant headlines. We consider four types of prompts: ‘generic,’ ‘clickbait,’ ‘emotional language,’ and ‘Upworthy.com-style’, as as discussed in the earlier section “Choosing Prompts.”

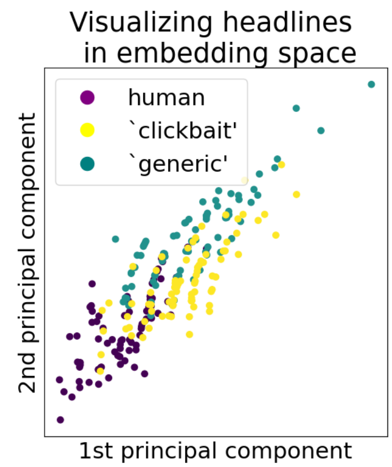

Fig. 5: Visualizing headlines produced by different prompts along principal components of the feature embedding matrix. Three different clusters are visible, suggesting that ChatGPT outputs remain distinct from human headlines even under different prompts. Our permutation tests elucidate these differences.

Comparing ‘generic’ and ‘clickbait’ Prompts

The first intervention we considered was comparing the results of the ‘generic’ prompt to those output by the ‘clickbait’ prompt. Interestingly, a new pattern emerged: whereas the distinction between human and ‘generic’ ChatGPT headlines had been predominently in complexity features, the distinctions between ‘clickbait’ and ‘generic’ headlines involved more measures of writing style. We used permutation tests to compare the distribution of ChatGPT ‘generic’ feature embedings to ‘clickbait’ feature embeddings. We found there was enough evidence to reject the null under our adjusted threshold ($p < 0.05/15 = 0.0033$ to account for multiplicity) on five attributes:

- ‘exclaim’, $p < 0.0002$
- ‘allpunc’, $p < 0.0002$
- ‘allcaps’, $p < 0.0002$
- ‘word_count’, $p < 0.0002$
- ‘avg_wordlen’, $p = 0.0028$

In contrast with our previous findings, three of the four most significant features were measures of writing style (‘exclaim’, ‘allpunc’, ‘allcaps’). Moreover, the measures of complexity which registered significance (‘word_count’, ‘avg_wordlen’) were measures of number and length of words. This lies in sharp contrast with the most significant features distinguishing human headlines from ChatGPT headlines, which were overwhelmingly measures of linguistic complexity rather than writing style. Figure 5 provides a visualization that complements these results: using principal component analysis to produce a 2-dimensional visualization of headlines feature embeddings and coloring points according to origin, three distinct clusters appear. This suggests while changing the instructions to ChatGPT produces different kinds of headlines, the distinction between human writing and ChatGPT output persists over different prompts.

Comparing Human to ChatGPT Headlines Across Prompts

We now wished to extend the above analysis to compare ChatGPT output to human headlines across varying prompt styles. We hypothesized that some of the prompts, particularly the 'Upworthy.com-style' instruction, would cause ChatGPT to produce output more similar to the headlines from the Upworthy Research Archive.

To our surprise, the distinctions between ChatGPT's output and the true human-written headlines remained extremely robust to input instructions, particularly among measures of complexity. For each prompt output, we ran permutation tests comparing the ChatGPT output to the human-written headlines following the procedure described earlier. Figure 6 provides a visual summary of our permutation test results; for each prompt output, we indicate the strength of the p-value by feature returned by the permutation test. We use a renormalized colorscale to indicate more- and less- significant p-values, so brighter colors should be interpreted as showing less evidence that a given prompt output differs from human writing along that feature.

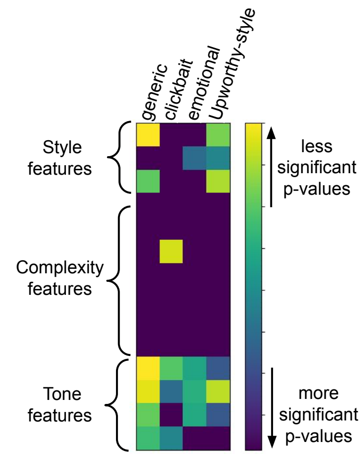

Fig. 6: Comparing headlines produced by ChatGPT under different prompts to human-generated headlines across features. Darker values indicate more evidence that the ChatGPT output is statistically distinct from human headlines. Across all prompt styles, almost all complexity features differ substantially between ChatGPT output and human writing.

To our surprise, we found no evidence to support our hypothesis: when comparing the 'Upworthy.com-style' output to the true human written headlines, we did not observe significant evidence that the 'Upworthy.com-style' output was more similar to human headlines than other prompt outputs. In particular, along features measuring linguistic complexity, almost all prompt outputs displayed marked divergence from human headlines.

The Nuanced Role of Complexity

Given the robust divergence in complexity features, one might expect to find a simplistic pattern such as "human written headlines have more/less linguistic complexity than ChatGPT output." To our surprise, we found that was not the case: in particular, human output scored higher in total word count, while scoring lower in measures of reading difficulty. One example of this phenomenon is plotted in Figure 7, a reproduction of a figure appearing early in this report.

Figure 7 specifically examine the output of ChatGPT in repsonse to the 'Upworthy-style' instruction, demonstrating that even specific prompts do not alleviate this trend.

In the context of Upworthy.com's goals, namely to produce headlines that can be easily read and understood while scrolling through webpages, ease-of-reading is a highly prized quality of a good headline. Our analysis suggests that human writers are better able to navigate the trade-off of using more words in their headlines in order to produce sentences that can be digested quickly.

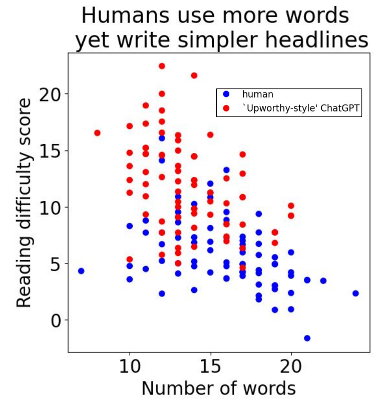

Fig. 7: Comparing human headlines (blue) to 'Upworthy-style' headlines (red) by word count and reading level. Human-written headlines use more words but are linguistically simpler.

Conclusions and Future Directions

Limitations and Future Directions

A major limitation of our analysis is the comparatively small datasets of ChatGPT outputs we were able to produce. Future studies, either with in-house gpt models or with better access to OpenAI services, would benefit from generating much larger datasets and repeating the above analysis. Another limitation is the comparatively low-dimension feature embedding used. While this was necessitated in part by the aforementioned limited dataset, higher-dimensional embeddings may better capture headline attributes not considered in this analysis, including named-entity recognition.

Another avenue for further study is how the above results may vary with time. We compared ChatGPT headlines to headlines from Q3 of 2014. Between 2013 and 2015, the internal goals and ideal style of Upworthy.com's content shifted dramatically, including a change away from "clickbait-like" content [2]. If the above analysis is repeated for articles written in other quarters, would the findings change?

Recommendations to Media Firms

As large language models continue to advance, we expect the output of services like ChatGPT to only continue to improve. The cost-effectiveness of employing automation in media production will place pressure on firms to make principled decisions about the role AI will play in their workflow. Given our findings, we highlight to such firms the role of linguistic complexity. In particular, we emphasize that employing human writers–who are experts at crafting readable language–remains a vital necessity, given the current state of large language models.

References

[1] IBM.com. “Augmented work for an automated, AI-driven world.” https://www.ibm.com/downloads/cas/NGAWMXAK. Accessed 10 Nov. 2023.

[2] David Karpf. “Analytic Activism: Digital Listening and the New Political Strategy.” Oxford University Press 2016. ISBN 9780190266141.

[3] Upworthy Research Archive. “About the Archive.” https://upworthy.natematias.com/about-the-archive. Accessed 7 November 2023.

[4] Matias, J., Munger, K., Le Quere, M.A., Ebersole, C. “ The Upworthy Research Archive, a time series of 32,487 experiments in U.S. media.” Nature Scientific Data. https://www.nature.com/articles/s41597-021-00934-7. Accessed 7 November 2023.

[5] Krystal Hu. “ChatGPT sets record for fastest-growing user base.” Reuters.com. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/. Accessed 9 November 2023.

[6] Ken Rice, Thomas Lumley. “Permutation tests,” University of Washington. https://faculty.washington.edu/kenrice/sisg/SISG-08-06.pdf. Accessed 9 November 2023.