Large Language Models for Sentiment Analysis on Political Tweets: A Case Study on Environmental Issues

DRAFT

Brandon Derr

PS 491

Spring 2024

Introduction

Sentiment analysis allows researchers to gain insight into corpora of text by assigning documents within those corpora certain values. Automated sentiment analysis is an essential tool for studying large corpora of text data. Two methods of sentiment analysis (manual assignment and large language models) are tested on the same corpus of Tweets. I manually assign sentiment values based on the endorsement or disapproval of renewable energy sources. I use 5 large language models of varying sizes to automatically assign values to the same corpus based on the same instructions that I used to manually assign values. The runtime and scores of each large language model are recorded and compared to manual assignment. Differences between the large language model value assignments and manual assignments are noted. These results have implications for those using automated sentiment analysis and large language models in social science research, particularly with large corpora of short documents like those in Twitter analysis.

Literature Review

Sentiment analysis is the process of assigning values to opinions based on some established criteria (Liu 2022). These values can then be used to gain insight into opinion trends within a single actor or groups of actors. The average opinion of a group of actors, change over time for a single actor's opinion, and differences between groups of actors are all examples of insights that can be gained from sentiment analysis. This process has seen an increase in use with the advent of social media and online review forums. Marketing professionals can utilize sentiment analysis to determine the effectiveness of their marketing efforts (Micu et al. 2017). Institutions of higher learning can use sentiment analysis to gain a better understanding of

DRAFT

student sentiment towards any number of institutional factors including courses, instructors, and student services (Dalipi, Zdravkova, and Ahlgren 2021).

In common use and in the literature, sentiment analysis usually refers to automated processes for applying opinion values to a large corpus of text. Manual assignment by one researcher is both extremely inefficient and presents questions of bias. Before the advent of large language models, common methods for sentiment analysis included crow-coding, dictionary word-embeddings, and untrained machine learning algorithms. In 2021, van Attenveldt et al. compared these methods on a common corpus of Dutch news headlines (Van Atteveldt, Van Der Velden, and Boukes 2021). They found high validity with manual and crowd-coding values, poor validity with dictionary methods, and promising results with machine learning approaches. The machine learning approaches tested in this study are early versions of the technology upon which current large language models are built. Comparisons between manual sentiment analysis and automated methods are not novel questions. The previously cited van Atterveldt et al. study is one such example.

The use of social media as a medium for political communication by both the public and political elites has dramatically increased since the late aughts. Although political speech on social media significantly differs from political speech in other mediums, it has become a crucial part of personal political networks (Kruse, Norris, and Flinchum 2018).

This research seeks to add to the growing literature surrounding the use of pre-trained generative large language models in social science research. Preliminary work has been produced on its potential as a source of zero-shot classification (Ziems et al. 2024). This particular work finds sufficient parity between human coded values and values resulting from large language

DRAFT

models to be used in social science research. The research recommends the use of large language models in tandem with manual coding.

Large language models are notoriously biased in their current implementation. Harmful gender, religious, and ethnic stereotypes have been found in some form in several models (Abid, Farooqi, and Zou 2021; Kotek, Dockum, and Sun 2023; Navigli, Conia, and Ross 2023). These stereotypes are most likely unavoidable due to the nature of how large language models are created and curated for end users. This produces a major impediment to their use in scientific research.

OpenAI is the organization behind the GPT models used in this research. These models are the same models used by the popular “ChatGPT” service, also run by OpenAI. With the release of ChatGPT, OpenAI became the most well-known artificial intelligence company in the world (Kay 2024). The organization has been under consistent scrutiny since its founding due to questions surrounding their hybrid public/private business structure and closed nature of their models.

The Mistral team began working in 2023 with the stated purpose of creating open access models as opposed to the restricted models coming from private companies like Google and OpenAI. In late February 2024, they received a substantial investment from Microsoft (also an investor in OpenAI) and announced a new partnership with them (Malik and Hu 2024; Mistral AI team 2024).

Claude 3 Sonnet is the newest model of those included in this research. Anthropic is the organization that releases the Claude series of models. They present themselves as a company focused first and foremost on “AI safety and research”, dedicated to “building systems that

DRAFT

people can rely on and generating research about the opportunities and risks of AI" (Anthropic PBC n.d.). Claude 3 Sonnet is the most recent of the models tested in this research but testing suggests that the Claude 3 family of model are extremely competitive with and sometimes outright win against large language model offerings (Sun 2024).

Along with the explosion of large language models for personal and professional use, academics have begun to see their potential as research tools. From deciphering financial sentiment to assisting with statistical questions, the use of the tools in meaningful research is increasing and shows no signs of slowing down (De Kok 2023; Sufi 2024).

Methodology

I analyze a corpus of microblogging posts made on X, formerly Twitter, using four methods of sentiment analysis to assess their efficiency and accuracy. They have not been available long enough yet to find common use in social science research. However, pre-trained large language models have the potential to be extremely useful to social scientists. Models which are pre-trained can be implemented into a project faster than a model which needs to be trained using a large corpus of training data. Cheap cloud computing options like Microsoft Azure and Google Colab are becoming increasingly available alongside the deployment of large language models, opening the door for more social scientists to answer important questions. Comparisons of efficiency and accuracy between emerging methods is crucial information to researchers deciding on methodologies for their specific use cases.

The text corpus consists of posts made by Senators on Twitter/X during the 116th Congress. Data are gathered from Tweets of Congress, a repository which contains automatically recorded posts from thousands of accounts relevant to Congress (Litel 2023). These data are

DRAFT

available until July 11, 2023, when Twitter changed their API access policies. All data for the 116th Congress (January 3, 2019 – January 3, 2021) is available. I only use posts made from a single account for each Senator. I consider only the accounts used the most by their respective Senator and avoid campaign accounts. In total, there are 102 accounts represented in the corpus after two seats changed during the Congressional term. Martha McSally (R-AZ) was replaced by Mark Kelly (D-AZ) in a special election, and Johnny Isakson (R-GA) was replaced by Kelly Loeffler (R-GA) due to health concerns.

I then take 2000 random statements from the total corpus to create the dataset against which I test each method of sentiment analysis. This number is large enough to capture a wide range of topics and sentiment while also being small enough to manually code each statement individually. This step would not be necessary if we had no need for manually assigned, baseline values. The methods tested in this case study enable social scientists to stretch their resources further to answer broader questions than they would otherwise be able to. In research dedicated to the totality renewable energy sentiment from Senators, all 222,367 statements could be used. However, I am more interested in how the methods compare to each other in practice than I am in capturing the sentiment expressed in the entire population of Senatorial tweets in this period.

Only two methods are being tested in this research. The first is the manual assignment method traditionally used by social scientists before modern computing. The second method utilizes large language models to automatically detect topics and assign sentiment scores. I test this method using the small and large versions of 3 distinct large language models on a single topic. The most current, publicly available models from leading developers are the ones tested.

DRAFT

Manual Assignment

I manually assign values to each of the 2000 Twitter statements, timing myself while doing so. The process happens inside of an R script that loops through each row in a data frame, prints the row's associated text data, asks for a user input, and saves that input to a separate column. The loop does not stop until every row in the data frame has been assigned a value. For manual assignment, I loop through all 2000 statements and assign statements values based on sentiment towards renewable energy. Statements which approve of renewable energy or disapprove of fossil fuel usage are assigned "pr" or pro-renewable. Statements which approve of fossil fuels or disapprove of renewable energy proposals are assigned "pff" or pro-fossil fuels. Statements that have neutral, no, or ambiguous sentiment towards energy sources are assigned "0". This process produces the list of "true" values against which the models' results can be analyzed.

I arrive at my own set of instructions throughout the course of several preliminary rounds of manual assignment. Initially, I only followed the following set of instructions: "Assign pro-renewable statements as 'pr'. Assign pro-fossil fuel statements as 'pff'. Assign statements that are neither as '0'". I quickly realized that these instructions would not sufficiently capture the full range of sentiment expressed towards energy generation by Senators. The next evolution of my manual instructions added negative sentiment, statements that were anti-fossil fuel were assigned as pro-renewable and vice versa. Statements with neutral or no sentiment towards fuel sources were still assigned "0". This set of instructions got me further along in the manual coding process than before. However, edge cases with respect to this coding scheme started to become apparent. These edge cases involved nuclear energy, ethanol, and the Nord Stream Pipeline. Nuclear energy statements were assigned "0" as nuclear energy is technically non-

DRAFT

renewable but is certainly not a fossil fuel. Ethanol and related products are considered renewable energy, and statements mentioning them were labeled as such. Many statements about the Nord Stream pipeline were tied to larger conversations about Russian influence in Western Europe. These statements were assessed on a case-by-case basis and only labeled as “pr” or “pff” if they contained broader endorsements or condemnation of fossil fuels. The prompt sent to each large language model is a copy of my own coding instructions with the exception of the first and last sentences.

Large Language Models

Tests involving large language models (GPT and Mistral) are conducted using their current large models (GPT 4 Turbo and Mistral Large) and small models (GPT 3.5 Turbo and Mistral Small). All four tests use the same prompt for each variable tested. These prompts include the guidelines that I use for manual assignments as well as instructions to ensure clean formatting. The following is the user prompt passed to each model:

> “Given the following criteria for assessing the sentiment of statements about energy sources: pr: pro renewables and/or clean energy and/or anti fossil fuels. pff: pro fossil fuels and/or anti renewables and/or pro coal. 0: no stance on energy source Nuclear energy statements are assigned 0 because they are nonrenewable. Ethanol is considered clean energy. Statements that espouse both pr and pff views are assigned 0. Many tweets reference the Nord Stream pipeline. These are given 0 unless they connect to a larger narrative about fossil fuel usage in the US. Statements referencing specific legislation are only coded if the purpose of the legislation can be assessed within the Tweet's text. Classify the following statement as 'pr', 'pff', or '0'. Include nothing else in your response:”

DRAFT

The large language model tests are run inside of a Python script. Each large language model has an API, or Application Programming Interface. This allows programmers to communicate with web services. Python is the primary coding language used to interact with large language model APIs. My scripts repeatedly send instructions to the large language model asking for a coding response for each Tweet in an Excel file. Each model is fed the prompt along with the data contained within a single Twitter/X statement. The output of the model is recorded in a new column on the same row as the statement it was given. Each row of the corpus is fed to the model individually until all statements have been given values. Additionally, the start time and end time of the script are recorded and printed once the script is complete, giving the script's runtime and thus the model's efficiency.

I find instances where the results from automatic methods differ from manual assignment values particularly useful. These instances can provide valuable information about the specificity of coding guidelines and biases within models on the subject domain. For methods utilizing pre-trained large language models, researchers can use these differences to create prompts that produce more reliable sentiment values. See *Prompt Engineering* in the Discussion section for more information.

Some models produced values which were not exactly "pr", "pff", or "0". These values sometimes had quotes around them. Others included explanations about why the model chose a particular value. In instances of non-standard value assignments, I clean the data to reflect what the data would be if returned correctly. This process was not timed and is not included in the model runtime. However, specific instances of this are noted in Discussion section and should be factored into one's choice of model.

DRAFT

Model Specifics

OpenAI’s GPT 4 Turbo model is the flagship model from the leading artificial intelligence organization. This model powers the paid tier of ChatGPT. The model used in this research is “gpt-4-0125-preview”. OpenAI’s GPT 3.5 Turbo currently powers the free tier of ChatGPT. The specific model used in these tests is gpt-3.5-turbo-0125.

Mistral Large and Mistral Small are the names of the Mistral models used in this research. At the time of these tests, “mistral-large-2402” and “mistral-small-2402” were the specific model numbers used as they were the most currently released versions of each model respectively.

Claude’s equivalent to GPT 3.5 Turbo and Mistral Small is called Claude 3 Haiku. However, this model was not public at the time of this research. Instead, their medium-sized current model, Claude Sonnet, is used. Claude 3 Opus is their large model, comparable to GPT 4 Turbo and Mistral Large. I initially planned to test this model as well. However, Claude places a limit on the number of API requests that each account is allowed to use in a given time frame. I was able to test Claude 3 Sonnet with no issues in this area. However, while testing Claude 3 Opus, I encountered a 7-day restriction. Results from the Sonnet model are included in the results of this report, while the half-completed Opus results are not. Find more on the Claude rate limit in the Discussion section.

The distinction between “small” and “large” models refers to the size of the dataset a model is trained upon.

Inter-coder Reliability Scores

DRAFT

To compare results from each model and the manually assigned values, I treat the models as individual coders and use established methods to compute inter-coder reliability. Values need to be converted into numeric factors to make these computations. “Pr” assignments are converted to 1. “Pff” assignments are converted to -1. Assignments of neutral or no sentiment remain 0. I use the “icr” R package to compute Krippendorf’s Alpha for sets of 5 sets of two models, each set containing the manually assigned scores and the scores from a large language model.

DRAFT

Results

Table 1: *Runtimes*

Method/Model	*Time to Complete*
Manual Assignment	270 minutes
GPT 3.5 Turbo	14 minutes
GPT 4 Turbo	22 minutes
Mistral Small	32 minutes
Mistral Large	32 minutes
Claude 3 Sonnet	60 minutes

Source: Author

Table 2: *Costs*

Model	*Input Token Cost (per million)*	*Output Token Cost (per million)*	*Total Run Cost*
*GPT 3.5 Turbo**	$0.5	$1.50	$0.57
*GPT 4 Turbo**	$10	$30	$5.54
*Mistral Small***	$2	$6	$1.47
*Mistral Large***	$8	$24	$5.10
*Claude 3 Sonnet****	$3	$15	$3.30

Sources: *(OpenAI 2024) **(Mistral AI 2024) ***(Anthropic PBC 2024)

Table 3: *Sentiment Scores Frequency*

Model	*0*	*PFF*	*PR*
Manual Assignment	1960	12	28
GPT 3.5 Turbo	1830	36	134

GPT 4 Turbo	1904	17	79
*Mistral Small**	1937	6	57
Mistral Large	1885	18	97
Claude 3 Sonnet	1902	23	75

Source: Author

Table 4: *Krippendorff's Alpha (Model + Manual Assignment)*

Method/Model	*Krippendorff's Alpha*
Manual Assignment	1
GPT 3.5 Turbo	0.303
GPT 4 Turbo	0.508
Mistral Small	0.553
Mistral Large	0.467
Claude 3 Sonnet	0.486

Source: Author

Comparison Summary

A total of 196 statements (9.8%) were assigned a value by at least one model. Of these, 170 statements were non-consensus. In other words, 170 statements had different values between manual assignment and the five models. Manual assignment values have agreement with 4 models in 81 cases, agreement with 3 models in 24 cases, agreement with 2 models in 18 cases, agreement with only one model in 19 cases, and no agreement with any model in 28 cases.

The fastest method of those tested was automatic assignment with GPT 3.5 Turbo. It finished outputting values in 5% of the time that it took to manually code the corpus and was 33% faster than the next fastest model. These differences are small in absolute values for this corpus. However, when scaled to larger corpora, these differences can result in much longer computing times. Claude 3 Sonnet was the slowest model at just under an hour runtime.

DRAFT

To add another feather to the hat of GPT 3.5 Turbo, it was by far the cheapest model tested. The entire run cost was nearly an order of magnitude less than the other models. Mistral Small also stood out as notably cheap.

In terms of score accuracy when compared to manually assigned sentiment scores, GPT 3.5 performed terribly. It assigned over 100 more “pr” statements than I manually assigned. Krippendorf’s Alpha is a measure of inter coder agreement. The values in Table 4 come from computing this value for each model as it agrees with the manually assigned scores. An Alpha value of 1 shows complete agreement. An Alpha value of 0 shows complete disagreement. Mistral Small performed the best of the model in terms of accuracy with a Krippendorf’s Alpha of 0.553.

Discussion

Incorrect Responses

The models were given prompts which specified to only respond with one of three values: “0”, “pff”, or “pr”. Three models returned at least one value assignment that was outside of these values: GPT 4 Turbo, Mistral Large, and Mistral Small. GPT 4 Turbo only gave one incorrect response. It was assigning sentiment to a Twitter statement which only contained a link to a YouTube video. The model said that it could not access links or view online videos. Mistral Large contained 4 incorrect responses. One contained a valid response in quotes. The other 3 incorrect responses were assigned “0” with an additional explanation as to why this might be the case. Mistral Small only produced 1034 valid responses out of 2000 requests. Of these responses, 960 were “0” assignments with explanations like in the case of Mistral Large. The other 6 incorrect responses were valid responses contained within single quotes. This problem may be

DRAFT

able to be fixed with more specific prompts. It also may just be a limitation of the model's training. For this sample size, 1 or 4 incorrect responses are easy to fix.

Limited Sample

My work only represents a single attempt by a single researcher on a single corpus of text. This sample size is dangerously small to make any substantial wide-reaching conclusions. Instead, this research should be a contributing factor towards a methodological decision that requires many more data points. Methods may differ in their efficiency and accuracy with different topics or different users. Corpora with larger documents such as blog posts, legal documents, or books are radically different from microblogging posts. This research should not factor into any decision regarding corpora that differ in this way.

Time to Launch

This research does not consider the time each method takes to learn, or any bugs/difficulties encountered with each method. Every researcher will have different experiences with each method, learning some API's quicker or encountering fewer bugs on their systems than others. Personally, I experienced no difficulty using the API's from OpenAI and Mistral, but Anthropic's API (Claude 3) took about an hour to learn to use. I did not include these setup times in my analysis because they are a one-time cost and are not affected by the size of a corpus. Once one invests in learning how to use a particular model, they can continue to use that knowledge on any number of corpora of any length.

Constantly Changing Landscape

AI is changing extremely quickly, and this rate does not seem to be slowing down. Changes have even been made within released models that alter their functionality in ways that

DRAFT

the end user can measure. Results from these tests and similar ones should be considered as temporary knowledge until AI development slows or AI developers put more care into maintaining models in the status that they released.

New models are being released weekly by the largest software developers in the world. These models will ideally outperform older models more efficiently. While the models tested in this research are the most well-known at the time of writing, newer models are likely to be released and usurp their older counterparts.

Scalability

Large language models reveal their strength with large corpora and corpora with poorly engaging content. Manually coding corpora of sufficient size becomes increasingly unrealistic as the time commitment scales positively. Poorly engaging content is more demanding on human coders than engaging content, decreasing the rate at which human coders manually code each statement. While large language models are well suited to large scale projects. Small percentage differences in the performance between models can result in large absolute differences in large projects. API rate limits like the one in

Mixed Assignment is Still Helpful

Manual assignment is beneficial to automated methods of sentiment analysis. It provides values against which a researcher can test the results from large language models. The process performed in this research of manually coding a set of documents and comparing these manual values with the automatic values can and should be done with subsamples in projects using this technology.

DRAFT

Prompt Engineering

Costs can be reduced by reducing the number of tokens sent to the API and received from the API. Efforts in similar research should be made to reduce the length of the coding instructions sent to the large language model API. The smallest number of tokens that still result in accurate values is the most ideal set of instructions as the input token cost will be smallest. Better prompts can also reduce the rate of incorrect responses in models like Mistral Small, reducing output token costs.

Active Coding or Passive Automation

The times recorded for each model are their runtimes with no external input required. Manual assignment requires quite a lot of focus for extended periods of time. Even an activity like listening to an engaging podcast can break this concentration enough to dramatically slow down manual coding. However, for automated sentiment analysis methods using large language models, the code runs until completion without any effort required from the researcher. *Even if the time required from an automated method were several times the time required by manual assignment, it would remain beneficial to use the automated method on a corpus of any significant size.*

Costs

The current cost structure of all 5 models scales linearly with the size of the corpus. The models bill based on the number of tokens included in the prompt and returned in the response. Testing larger corpora will be linearly more expensive than smaller ones. Corpora with fewer, larger statements are billed equally to corpora with more, smaller statements as the quantity of text sent to the model does not change.

DRAFT

Claude Rate Limits

As mentioned earlier, the quickly evolving nature of artificial intelligence availability and technology may make some of this report irrelevant at any point in the next few years. However, at the time of writing this report, the rate limits implemented by Claude are unacceptable for an enterprise-grade artificial intelligence service. This is that much truer due to limits being placed on even those who pay to use the service. The rate limits are restrictive enough to make any large corpus of data impossible to analyze with their model. As such no recommendation can be made to choose Claude over alternatives for large-scale sentiment analysis until these limits are removed. Similar limits exist with industry leader OpenAI's models for ChatGPT's paid tier, varying with the organization's cloud hardware availability. If Anthropic continue to maintain Claude models and scale their hardware solutions, they may become a true competitor in the sentiment analysis field later.

Bias

Biases are inherent in large language models. "Garbage in, garbage out" is an adage that reflects biased training data resulting in biased results. Large language models also result from biases after they are trained. The organizations that produce these models tweak them to bring them into "alignment" with user desires. This process also results in biases. Human coders are not immune from biases, and this area could be a potential strength for large language models.

Diverse medium inputs

Pre-trained large language models are beginning to offer image and video inputs into their prompts. This offers new automated analysis methods which were unavailable with previous methods of automated analysis. Tweets and other forms of social media posts often

DRAFT

contain images and videos which enhance or modify the meaning of the associated text. In some cases, these media are the only content included in a post. In the past, these posts would simply be removed from the corpus. In cases with both text and media, the text would be included without the added context of the media. However, new tools will allow for a more comprehensive analysis of social media posts.

Conclusion and Recommendations

In this research, I performed a novel evaluation of the capacity of 5 large language models to assign sentiment values to a corpus of Tweets. I evaluate each model based on its completion time, cost, and accuracy compared to manually assigned values. These evaluations resulted in a few key findings. The first is that GPT 3.5 model is the most cost-effective and fastest model of those tested. The second is this model's advantages end there as its accuracy is much worse than any other model. The third is that while being competitive in all measurements, Mistral Small has an almost 50% incorrect response rate. With these results in mind, if one is primarily concerned with efficiency, cost or time, GPT 3.5 Turbo is the go-to model for sentiment assignment. For those who do not care about price or time and do not want to clean their results as long as they are mostly accurate, GPT 4 Turbo is the model of choice. However, if one can put some time into cleaning the results that they are given or find a prompt that makes incorrect responses rare, they should use Mistral Small. It is relatively cheap, fast, and accurate with the large caveat that, currently, its results must be cleaned.

DRAFT

Bibliography

Abid, Abubakar, Maheen Farooqi, and James Zou. 2021. “Persistent Anti-Muslim Bias in Large Language Models.” In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, Virtual Event USA: ACM, 298–306. doi:10.1145/3461702.3462624.

Anthropic PBC. 2024. “Pricing.” *Claude API*. https://www.anthropic.com/api (March 15, 2024).

Anthropic PBC. “Our Purpose.” *Company*. https://www.anthropic.com/company (March 15, 2024).

Dalipi, Fisnik, Katerina Zdravkova, and Fredrik Ahlgren. 2021. “Sentiment Analysis of Students’ Feedback in MOOCs: A Systematic Literature Review.” *Frontiers in artificial intelligence* 4: 728708.

De Kok, Ties. 2023. “Generative LLMs and Textual Analysis in Accounting: (Chat) GPT as Research Assistant?” *SSRN Electronic Journal*. doi:10.2139/ssrn.4429658.

Kay, Grace. 2024. “The History of ChatGPT Creator OpenAI, Which Elon Musk Helped Found before Parting Ways and Criticizing.” *Business Insider*. https://www.businessinsider.com/history-of-openai-company-chatgpt-elon-musk-founded-2022-12 (March 7, 2024).

Kotek, Hadas, Rikker Dockum, and David Sun. 2023. “Gender Bias and Stereotypes in Large Language Models.” In *Proceedings of The ACM Collective Intelligence Conference*, Delft Netherlands: ACM, 12–24. doi:10.1145/3582269.3615599.

Kruse, Lisa M., Dawn R. Norris, and Jonathan R. Flinchum. 2018. “Social Media as a Public Sphere? Politics on Social Media.” *The Sociological Quarterly* 59(1): 62–84. doi:10.1080/00380253.2017.1383143.

Litel, Alex. 2023. “Tweets of Congress.” https://github.com/alexlitel/congresstweets (March 4, 2023).

Liu, Bing. 2022. *Sentiment Analysis and Opinion Mining*. Springer Nature.

Malik, Yuvraj, and Krystal Hu. 2024. “Microsoft Partners with OpenAI’s French Rival Mistral.” *Reuters*. https://www.reuters.com/technology/microsoft-partners-with-openais-french-rival-mistral-2024-02-26/.

Micu, Adrian, Angela Eliza Micu, Marius Geru, and Radu Constantin Lixandroiu. 2017. “Analyzing User Sentiment in Social Media: Implications for Online Marketing Strategy.” *Psychology & Marketing* 34(12): 1094–1100.

Mistral AI. 2024. “Pricing.” *Mistral technology*. https://mistral.ai/technology/#pricing (March 15, 2024).

Mistral AI team. 2024. “Au Large.” https://mistral.ai/news/mistral-large/ (March 7, 2024).

DRAFT

Navigli, Roberto, Simone Conia, and Björn Ross. 2023. "Biases in Large Language Models: Origins, Inventory, and Discussion." *Journal of Data and Information Quality* 15(2): 1–21. doi:10.1145/3597307.

OpenAI. 2024. "Pricing." https://openai.com/pricing#language-models (March 15, 2024).

Sufi, Fahim. 2024. "Generative Pre-Trained Transformer (GPT) in Research: A Systematic Review on Data Augmentation." *Information* 15(2): 99. doi:10.3390/info15020099.

Sun, Wei. 2024. "Claude 3 Dethrones GPT-4 to Mark Phase Two in LLM Competition." *Counterpoint*. https://www.counterpointresearch.com/insights/claude-3-dethrones-gpt-4-phase-two-llm-competition/.

Van Atteveldt, Wouter, Mariken A. C. G. Van Der Velden, and Mark Boukes. 2021. "The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms." *Communication Methods and Measures* 15(2): 121–40. doi:10.1080/19312458.2020.1869198.

Ziems, Caleb, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. "Can Large Language Models Transform Computational Social Science?" *Computational Linguistics*: 1–55. doi:10.1162/coli_a_00502.

DRAFT