**XAI Greybox Ensemble with GPT4 for Story Analysis and Generation:**

**A Novel Framework for Diachronic Sentiment Analysis**

Jon Chun, Kenyon College, Gambier, OH

ORCID ID 0000-0002-5315-6784

Katherine Elkins, Kenyon College, Gambier OH

ORCID ID 0000-0001-9887-4854

**Keywords**: GPT4, LLM, XAI, greybox XAI, ensemble learning, GPT4 function calling, NLP, sentiment analysis, story generation

**Abstract**

The recent development of Transformers and large language models (LLMs) offer unique opportunities to work with natural language. They bring a degree of understanding and fluidity far surpassing previous language models, and they are rapidly progressing. They excel at representing and interpreting ideas and experiences that involve complex and subtle language and are therefore ideal for Computational Digital Humanities research. This paper briefly surveys how XAI can be used to augment two Computational Digital Humanities research areas relying on LLMs: (a) diachronic text sentiment analysis and (b) narrative generation. We also introduce a novel XAI greybox ensemble for diachronic sentiment analysis generalizable to any AI classification data points within a structured time series. Under human-in-the-loop supervision (HITL), this greybox ensemble combines the high performance of SOTA blackbox models like gpt-4-0613 with the interpretability, efficiency, and privacy-preserving nature of whitebox models. Two new local (EPC) and global (ECC) metrics enable multi-scale XAI at both the local and global levels. This greybox ensemble framework extends the SentimentArcs framework with OpenAI’s latest GPT models, new metrics and a modified supervisory HITL workflow released as open source software on github.com.

**Statements and Declarations**

No funds, grants, or other support was received.

**1. XAI FOR DH**

Digital Humanities has traditionally aligned with cultural thought leaders (O’Neil, 2017) who focus on the broader questions of Fairness, Accuracy, Transparency and Explainability (FATE) rather than the more technical aspects of XAI (Hussain, 2022). In particular, the culture of critique in the humanities (Latour, 2004) positions the digital humanist to identify and critique bias in data and models and explore definitions of fairness that affect both Digital Humanities scholarship and society at large. This paper, by contrast, will highlight the distinctive nature of XAI as applied to Computational Digital Humanities applications for both a discriminative task (sentiment analysis) and a generative task (narrative generation) (Offert, 2020).

Aside from the very recent rise of XAI Rationalization (Gurrapu, 2023), most XAI techniques require some level of statistical or technical expertise that represents a barrier for engagement with XAI technologies (Molnar, 2022). Instead, there can be a tendency, even in Computational Digital Humanities, to rely upon models without technical introspection, especially in cases in which the results align with intuition and theory. For white box GOFAI models like word frequency clouds, simple decision trees and even probabilistic topic modeling, trust can rest on the narrow gap between human understanding and model explainability (Jacovi, 2020). As the performance of massive black box models progress to offer human-level performance on more and more complex tasks like language and vision (EFF, 2023), however, these black box models will be increasingly incorporated into Computational Digital Humanities. For this reason, XAI must become part of the process.

The field of XAI is rapidly evolving (Jacovi, 2023), a development that bodes well for Computational Digital Humanists. Both AI and XAI have matured well beyond the confines of lab research (Liu H., 2021). However, broad adoption hinges upon trust, verification and accountability by experts beyond the AI research community. This includes non-technical regulators (Sovrano, 2022), users, and the public at large (Weidinger, 2022). Alongside progress in low/no-code AI tools like RapidMiner and Keras, XAI has evolved from a focus on largely technical considerations to include more human-centered issues (Nauta, 2022).

The humanization of XAI can be seen in several distinct trends related to XAI Rationalization (Gurrapu, 2023). One trend is towards providing non-technical natural language explanations. Another trend involves formulating explanations based on a deep understanding of what constitutes clear, concise and intuitive explanations for most humans (Paleja, 2022). To bridge the gap between complex black box models and broad non-technical understanding, XAI is becoming more interdisciplinary by incorporating principles from psychology, philosophy and Human-Computer Interface (HCI), a field itself at the intersection of computer science, behavioral science, cognitive science, design, media studies, and sociology (Heimerl, 2021).

The humanization of XAI is also visible in the shifting focus from models and data towards humans and interpretability (Langer, 2021). For example, some have even created a unified framework for multidisciplinary teams based upon a meta-analysis of survey papers from the social sciences, Human-Computer Interactions, visual analytics and traditional machine learning (Mohseni, 2018). This framework partitions the human XAI audience by expertise (novices, data and AI experts) each with distinct design goals and evaluation metrics. The AI model’s effectiveness is measured by both traditional machine metrics like accuracy and recall, but also newer human metrics, in other words by metrics that are both objective and subjective (e.g. perceived competence and understandability). Like Computational Digital Humanities itself, XAI is inherently collaborative and interdisciplinary.

A number of excellent free and open-source software (FOSS) tools, libraries and texts also support the Computational Digital Humanist in understanding, interrogating and quantifying XAI metrics like accuracy and fairness. These include frameworks and libraries like Quantus (Hedström, 2023), Google’s What-If-Tool (Wexler, 2019), IBM AI Fairness 360 (Bellamy, 2018), with many more resources compiled in github awesome repositories (Wang, 2023). These provide access to a broad range of XAI techniques from no-code visualizations to API calls and embedded code.

**2. XAI: Sentiment Analysis for Understanding Narrative**

Sentiment analysis is a popular task within the field of NLP that attempts to identify latent overall emotions, sentiments and opinions in data like text, images and sound (Poria, 2020). Sentiment analysis models generally assign a single unit of text (a) a number for emotional polarity (e.g. positivity or negativity) and sometimes (b) a label for emotion type (e.g. happiness, sadness, anger) (Zhang, 2022). Compared to other modalities like image and sound, text-based sentiment analysis is more popular (Github, 2023) and accurate (Paperswithcode, 2023).

The simplest sentiment models are based upon the compilation of thousands of semantically emotional words into lexicons based upon: positive/negative polarity (Xiaowen, 2008) and emotion type or even subjectivity, as in NRC (Mohammad, 2013). Some lexical models like VADER (Hutto, 2014) are augmented with heuristic rules to identify negations (not happy), intensifiers (very angry, VERY angry, very angry!!!), and non-text symbols (emojis and emoticons) that are missed by simple lexicons. These white box models are easy to interpret and often provide intuitive explainable visualizations at the cost of lower accuracy. Although lexical models have trouble detecting and accurately assessing sentiment in humor, irony and sarcasm, statistical patterns of natural language usage in corpora like novels show these errors are often symmetrically distributed around the average and often cancel each other out (Elkins, 2019).

Of the two sentiment analysis tasks, measuring emotional polarity is much simpler and more reliable than identifying emotional type. The emotional polarity task involves simply assigning a discrete label (e.g. positive/negative, [-1,0,1] or [1-5 stars]) or continuous value (-1.0 through +1.0) to each piece of data. In contrast, categorizing emotion types is usually challenging. Often emotional types overlap and are oversimplifications distilled from much more complex taxonomies like Plutchik’s Wheel of Emotions (Zeng, 2021). For example, the image-based facial recognition model Deepface (Serengil, 2021) bins all emotions within 6 mutually exclusive types (labels: angry, disgust, fear, happy, sad, surprise, neutral) while the text-based NRC Word-Emotion Association Lexicon (Mohammad, 2013) uses 8 categories (anger, anticipation, disgust, fear, joy, sadness, surprise and trust). Our preliminary research into facial recognition with DeepFace finds that while the emotional category is sometimes wrong, the sentiment is often correct (i.e. the label “fear” may not appear accurate but negative sentiment appears correct).

Despite popular critiques, most notably in *Atlas of AI* by Kate Crawford (Crawford, 2020), emotional analysis has proven to be one of the most common and effective NLP techniques for many narrow applications like movie or restaurant reviews. In contrast, Computational Digital Humanities approaches analyze more general and complex texts toward different ends. This requires a modified approach: exploratory data analysis (EDA) vs hypothesis-driven analysis, SOTA LLMs vs fast GOFAI models, and human-in-the-loop XAI vs fully-automated metrics. One should be careful to dismiss an entire field based upon standard sentiment analysis models and workflows from industry. These approaches are often inappropriate for more complex and nuanced texts. Diachronic sentiment analysis can detect the evolution of emotions in literary narratives, transformation of popular opinions on social media, or shifting attitudes in financial news. In addition, key features like peaks and valleys often correspond to narrative crux points or to financial forensic signatures like pump and dump schemes.

Diachronic sentiment analysis is particularly suited for Computational Digital Humanities research for several reasons. First, there are few computational approaches that can reveal as much about plot and story evolution as the emotional trajectory reflected in language. Second, time series signal processing techniques like simple moving average and LOWESS can effectively filter out random variations and noise to surface a coherent sentiment arc that often agrees with the reader’s experience. Finally, specific characteristics of narrative sentiment arcs like rapidly rising/falling trajectories or inflection points can offer new insights or confirm/challenge interpretations on key narrative elements and crux points.

Diachronic sentiment analysis works by segmenting a large corpus into an ordered sequence of semantically meaningful units (e.g. sentences, paragraphs, tweets, news articles), measuring the discrete sentiment value of each unit, joining these sentiment values into an ordered time series, and applying smoothing techniques. We have explored all of these segmentation methods and more, finding sentence level segmentation provides an optimal trade-off between granular and meaningful semantic units and the higher signal/noise ratio needed for extracting latent emotional arcs. Paragraph segmentation loses too much information (e.g. longer paragraphs can contain complex emotional structure) while extracting smaller coherent semantic units below the sentence level becomes error-prone, lowering signal/noise levels (e.g. ambiguous parse trees)

Popular applications for detecting and measuring discrete sentiment or opinions in social media and reviews (e.g. movies, restaurants, amazon products) are reflected in specialized resources (lexicons, training datasets and fine-tuned models) (Wankhade, 2022). These applications give a single sentiment value. In very short text, one can imagine that an error in a word or two might drastically change interpretation. In contrast, tools for diachronic sentiment analysis produce a sequence of sentiment values showing the changing evolution of sentiment over time. The effect of individual errors is moderated by averaging sentiment over sentences and then applying time-series smoothing techniques. Global underlying sentiment arcs are then surfaced. Many key narrative features hidden within massive corpora can be quickly identified and selectively prioritized for close readings. Diachronic sentiment analysis tools include simple lexical+heuristic models available as libraries in R (Jockers, 2020). More sophisticated ensembles of models exist in Python using Jupyter Colab notebooks (Chun, 2023).

By applying common signal processing techniques to raw sentiment time series values, diachronic sentiment analysis transforms thousands of individual sentiment point values into a smooth narrative arc that intuitively reveals key narrative features. Signal processing includes a variety of smoothing, peak detection and warping algorithms. Nonetheless, issues have arisen in the Digital Humanities research concerning sentiment analysis, particularly when it comes to over-reliance on trusting fully automated diachronic sentiment analysis (Swafford, 2015).

Although SentimentArcs (Chun, 2021a) comes with a large ensemble of SOTA models and sensible defaults for tuning signal processing, it emphasizes the human-in-the-loop expert overseeing the entire process. When the sentiment plots of dozens of diverse models in the ensemble are in relative agreement, this strongly suggests these narrative periods have a strong emotional coherence; the human-in-the-loop expert can have confidence in the surfaced emotional arc. Periods where the sentiment plots diverge and/or disagree alert the expert to disambiguate conflicts with closer readings. This enables the expert to judiciously allocate research time to areas that can provide new insights as well as challenge or reinforce existing theories.

Unlike previous tools, SentimentArcs (Chun 2021) emphasizes a multi-model approach, highly tunable signal processing, and feature extraction. All of these model elements remain under the supervision of a human-in-the-loop expert. This ensemble of dozens of models listed in Figure 1 is classified in order of increasing complexity from simple lexical models to SOTA LLMs based on BERT (Devlin, 2019). The text processing pipeline in Figure 2 has been used to surface story arcs in a wide variety of corpora ranging from more traditional narratives to compilations of social media posts and financial news articles (Kenyon, 2023). *The Shapes of Stories* (Elkins 2022) elaborates upon a methodology to use SentimentArcs for analyzing literary narratives.

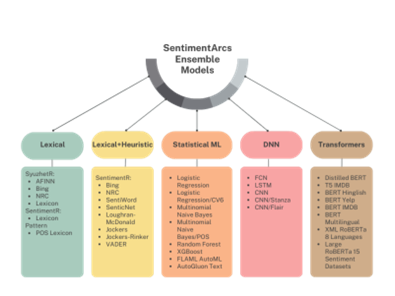


Figure 1. SentimentArcs Ensemble Models

A picture containing text, circle, compact disk, graphics

Description automatically generated

Figure 2. SentimentArcs Workflow Pipeline

Computational Digital Humanities scholars often work with more open-ended and sophisticated uses of language than industry applications in which the style and vocabulary tend to be straightforward (e.g. tweets or movie reviews). SentimentArcs offers a good example of applying a human-in-the-loop solution to address these unusual challenges. Simpler white box models like lexical and statistical models provide intuitive whitebox baselines to build trust, while more performant black box LLMs like BERT and GPT can deal with the edge cases of complex language constructions. BERT variants like RoBERTa (Liu Y., 2019) sometimes even appear to mimic the conceptual structure humans use to represent emotions (Li, 2023).

Explainability in a more traditional humanities context often means interrogating an argument through analysis of close readings. In this context, one often places more trust in the larger argument when the scholar’s attention to particular words or phrases seem accurate. Explainable AI in the sentiment analysis realm looks very different. The sentiment of single words can be misclassified even as the larger statistical model remains relatively stable and accurate. A good way to assess credibility is to assess crux points to determine whether key points in the arc seem to comport with readerly experience. Nonetheless, it’s important to stress that simpler models remain explainable in a fundamental sense: the lexicons can easily be inspected, and it’s easy to understand exactly how the model works.

Sometimes these simpler models struggle to surface arcs in narratives that are unusually challenging. Figure 3 shows how Toni Morrison’s *Beloved* (Elkins, 2022) is relatively less coherent than Virginia Woolf’s *To the Lighthouse* in Figure 4. In that case, both simpler lexical models and statistical machine learning models failed to clearly surface a narrative that identified crux points in the peaks and valleys that aligned with the typical reader’s perspective. In cases like these, LLMs do sometimes surface the most accurate sentiment arc from a scholarly perspective. They are, however, black boxes in the sense that while we can interrogate them for the sentiment value of a particular word or sentence, they do not provide a clear picture of how they come to those findings.

These two novels provide an ideal case study since, out of the entire SentimentArcs corpus, they offer two of the most diametrically opposite case studies. In the case of Woolf, the coherence of the black box and white box models gives us some measure of confidence in the black box. Morrison’s novel is the least coherent with the ensemble model, and LLMs are far more performant in this instance. Here, we require a novel greybox method to assure explainability of these more performant black box models.

A picture containing diagram, line, plot, text

Description automatically generated

Figure 3. SentimentArcs Incoherent Ensemble Plots

Toni Morrison’s *Beloved*

A picture containing plot, line, diagram

Description automatically generated

Figure 4. SentimentArcs Coherent Ensemble Plots

Woolf’s *To the Lighthouse*

 In general, Deep Learning LLMs offer better performance in exchange for weaker explainability (Bell, 2022). One surprising finding of working with LLMs in surfacing emotional arcs is that sometimes these large language models struggle or even perform worse than simpler models. In the case of sentiment analysis for Digital Humanities, the best performance is achieved by jointly optimizing the combination of both a particular corpus and a specific model (along with a particular set of hyperparameters and training regime). Overall, LLMs still provide good performance but are also the most opaque, difficult to explain and to trust.

Analyzing natural language in some situations when simpler models fail using the high-performance of large language models like BERT often makes sense or GPT. The larger lesson is that, unlike Yelp or Amazon reviews, for example, narratives vary tremendously in both style and content. There is simply no single model that works best for every novel. Yet another error that is often made in Digital Humanities research using sentiment analysis is the conclusion that it does not work in a particular case because a single model fails. In fact, sentiment analysis for such challenging cases requires optimizing the model to the particularities of the narrative.

SentimentArcs comes with an initial reference corpus of 25 carefully curated novels selected for familiarity as well as diversity across time, geography, cultures, styles and perspectives. These provide informed baselines to guide exploration of new works and to establish best practice guidelines. In the complex interplay between models, corpora, hyperparameter selection and training settings there is sometimes no universal best setting, as seen in Figure 3. Results can vary in agreement with the gold standard of human experts. Default initial settings and similarities with works in the reference corpus can provide initial guidance on model selection. Default settings also allow the exploration of new works according to coherence metrics. SentimentArcs workflow pipeline empowers human experts to efficiently filter through an ensemble of dozens of models over millions of individual data points.

**3. XAI and ChatGPT**

ChatGPT can be seen as (a) a conversational web interface backed by (b) a GPT-3.5/4.0+ large language model fine-tuned for dialogue (OpenAI, 2022). Although trained on dialogue, it can produce nearly anything that can be described in text form from essays to code, music to poetry (DAIR-AI, 2023). Part of ChatGPT’s exceptional performance is due to an effective method for fine-tuning that included both supervised learning and reinforcement learning from human feedback (RLHF) (Ouyang, 2022). ChatGPT is considered a text-to-text generative LLM that can respond to novel questions posed as natural language via prompts (PromptsLab, 2023). It can answer simple questions in response to a single prompt (zero-shot) or answer more complex problems when given more elaborate prompts that include several examples through few-shot/multi-shot prompts. In addition to simply generating text, it can perform arithmetic, solve puzzles and conduct multi-step reasoning with suitable prompt engineering. Periodically ChatGPT can suffer from non-response, omissions, tangents and even errors sustained with unwarranted confidence. These generally increase in frequency with task difficulty (Borji, 2023).

LLMs like ChatGPT are an invaluable tool for Computational Digital Humanities because they are unusually adept at understanding, analyzing and generating natural language (Zhou, 2023). LLMs can solve a wide range of NLP tasks like summarization, Q&A and sentiment analysis in ways that far surpass previous-generation, narrow, task-specific NLP models (Chee, 2023). At times they can even outperform humans (Zhang, 2022). NLP analysis of complex tasks like the detection of humor, sarcasm and irony have really only become practical with the development of these models. This also holds for NLP generative tasks like creating relatively coherent essays, writing creatively, and editing and critiquing existing prose.

There is widespread misunderstanding about how ChatGPT works from an end user perspective. It’s difficult to appreciate its capabilities from only casual interactions. Using simple zero-shot prompts like “write an essay on how the Smoot-Hawley Tariff Act played a role in the origins of the Great Depression” typically results in a fairly simple Wikipedia-level overview of the topic. The process of writing both fiction and non-fiction prose depends upon skills beyond good writing, involving not only a fluidity with natural language and interview skills (Chun, 2022), but also an accurate theory of mind for how ChatGPT processes language (Chun, 2021b). It also requires both simple and sophisticated prompt engineering strategies (Lester, 2021).

LLMs like ChatGPT represent not only a change in degree but also a change in kind from previous AI model types. The most concrete example of “more is different” is seen in the emergence of unpredictable capabilities observed as LLMs grow (Wei, 2022). With few-shot prompting and little to no task-specific training, LLMs beyond various size thresholds have been seen to acquire new abilities like adding/subtracting three-digit numbers (13B), estimating the probability of answers being true (52B), performing multi-step reasoning tasks (100B) and achieving word sense disambiguation (540B) where model parameter size is measured as x billion (xB).

LLMs are approaching a trillion parameters and could become a million times larger within the decade (Laird, 2023); we have not reached a ceiling on performance. Beyond just model parameter count, other factors like training dataset size (Hashimoto, 2021) and dataset composition (Ivgi, 2022) can also improve performance. Better training regime design (e.g. RLHF and sample efficient learning), training hyperparameters (e.g. iterations and learning rates) and fundamental compute and architecture advances (e.g. RLHF used in InstructGPT and ChatGPT) (Bai, 2022) can also enhance performance.

Aside from the multiple avenues for increased performance on core functionality like language generation, we have to consider possible unanticipated emergent capabilities. The nature of large DNN black box models like LLMs is fundamentally different from previous models and hence requires a different approach to XAI. Traditional AI models are generally viewed as deterministic transformations from fixed inputs to fixed outputs. LLMs like ChatGPT are interactive, open-ended, highly variable and dependent upon internal states that evolve over time. Within the narrow domain of Computational Digital Humanities, LLMs like ChatGPT offer both traditional and prompt-based approaches to XAI. Traditional XAI at the local level offers insights both practical (e.g. attention head visualizations) (Rush, 2018) and more theoretical (e.g. localizing knowledge in DNN) (Meng, 2022) insights. Traditional XAI at the global level, especially related to Computational Digital Humanities tasks like text generation, is largely exploratory and assessed through careful and thoughtful prompt engineering (Microsoft, 2022).

Since the release of ChatGPT in 2022 there has been a focus on trying to understand and explain ChatGPT on a functional level through prompt templates (Liu P., 2021), taxonomies (White, 2023), and compilations (Akin, 2023). This type of functional XAI is driven by prompt engineering, which is a dynamic, evolving and often unpredictable interactive conversation with humans (Reynolds, 2021). Sometimes called software 3.0 (Beurer-Kellner, 2022), prompt engineering is a skill that involves defining goals clearly, decomposing a path into discrete steps, encoding each step unambiguously, and adjusting and adapting to unexpected detours. Although there are attempts at automation (e.g. Google’s Automatic Prompt Engineer) (Zhou, 2022), these are guided by domain expertise and a strong theory of mind about how LLMs think, respond and can be guided.

For these reasons, prompt engineering does not fit into the traditional XAI framework. Given the scale and highly variable behavior of LLMs, typical black box global XAI approaches fail. The same slightly perturbed input prompt can result in highly variable outputs. LLMs typically transform input prompts into very open-ended outputs that resist characterization by common global XAI perturbation methodologies like SHAP and LIME described in Section 4. Surrogate models or distilled models would only mimic this same behavior.

ChatGPT prompt engineering for story generation faces several additional limitations. From a technical standpoint, ChatGPT currently has a working memory of about 4000 subword tokens (equivalent to 3000 English words) to remember facts and context around any given dialogue (as of June 2023). Depending on the length of the conversation and how it evolves over time, early statements, facts or other context may be lost, misremembered or hallucinated (Sallam, 2023). Finally, from a content standpoint, standalone ChatGPT only knows facts and information contained in its training dataset through September 2021. It can also slip into toxic or undesirable forms of behavior (Kumar, 2022). These limitations are partially addressed with heuristic safety rules as well as augmentation with other systems (Schick, 2023) and tooling like LangChain (Langchain, 2023).

Most ChatGPT prompts only generate about 400-600 words per turn, though some variation of a ‘continue’ prompt may cause ChatGPT to generate text as part of a longer request. From a stylistic standpoint, and with minimal prompt guidance, ChatGPT tends to generate stories that are short, simple, similarly structured, and filled with common tropes (Kallio, 2023). This is due to the fact that ChatGPT generates stories based upon similar text in its training dataset. Familiar, overused and trite phrasing and structures are statistically more likely to be recreated.

Where traditional analytical and statistical XAI methods fail, more informal conversational XAI approaches show promise. Through simple question and answer dialogue, ChatGPT has been given the same battery of psychological profiling used to gauge the beliefs, biases, and knowledge in humans. Although it can demonstrate limited knowledge, a lack of self-awareness, and a confidence in occasional hallucinations--all these much like humans--ChatGPT has demonstrated a distinctive persona (Kocoń, 2023; Li X, 2022). Across a dozen surveys for political opinions, it consistently favors liberal-libertarian positions over conservative-authoritarian ones (Hartmann, 2023). It exhibits a passing understanding of material on professional exams in medicine (Kung, 2022), law (Choi, 2023) and business school exams (Terwiesch, 2023). As a conversation partner, ChatGPT has demonstrated a theory of mind at the level of a nine year-old (Kosinski, 2023) that has been critiqued (Smith, 2023) and counter-critiqued (Boland, 2023).

In these cases, ChatGPT exhibits a unique pattern of strengths and shortcomings that lie closer to human cognition than GOFAI model computation (Mahowald, 2023). Although ChatGPT is trained to generate the next word (actually subword token) with varying levels of certainty set by the underlying model hyperparameters (OpenAI, 2023), the next word choice is not simply conditioned on linguistic statistical patterns (Benzon, 2023). Instead, ChatGPT text generation is conditioned on language as well as cultural, psychological, and physical aspects of reality based on a human perspective (Reynolds, 2021). For example, ChatGPT can be told to create stories in a given period, genre, author, style, or tone that contain particular characters, places, objects or events. Any story element can be added, expanded, modified, or deleted within a conversational feedback loop. While the human acts as the head writer, domain expert or creative genius, their ChatGPT partner plays the role of writing staff, ghostwriter, or editor.

To use ChatGPT to generate narratives that are longer, more complex and intentional, a number of approaches are promising (Mori, 2022 Wang, 2022). Controllable text generation (CTG) is adapted to storytelling via prompt engineering (Zhang H., 2022; Yang, 2022) and tight LLM revision loops (Du, 2022). Narrative coherence can be addressed with various structured strategies including slot, goal directed, analogy-based pipelines (Alhussain, 2022). Other top-down hierarchical approaches like TaleBrush use simple visual story arcs to guide the generation process (Chung, 2022). These allow for scoring, ranking, and selecting ensemble story elements (Goldfarb-Tarrant, 2020). LLMs can also provide more immediate control in terms of Human-Computer Interaction (HCI) collaborative writing using systems like Cue-Me-In (Brahman, 2020), Wordcraft (Yuan, 2022) (Ippolito, 2022), Loose Ends (Kreminski, 2022), and Dramatron (Mirowski, 2022).

**4. Local and Global XAI with SentimentArcs-Greybox Ensemble**

A fundamental trade-off with respect to XAI is the high performance of opaque blackbox models like GPT4 vs the more explainable and interpretable nature of whitebox models like decision trees, regression models or sentiment lexicons. The terms ‘explainable’ and ‘interpretable’ are overloaded with different meanings in different contexts (Marcinkevičs, 2023), but are used synonymously here.

ML/AI whitebox models like VADER are relatively transparent. Decisions can be interrogated by examining internal mechanisms (e.g. VADER unambiguously assigns a numeric emotional polarity to each word in its lexicon). Many mission critical applications like medical diagnosis and bank lending require whitebox models whose decisions can be interpreted and explained to non-specialists patients, customers and regulators. Whitebox models are generally much smaller and are easier to customize, fine-tune and roll out to devices on the edge of the cloud, such as smartphones and camera sensors.

In contrast, blackbox models like LLMs are orders of magnitude larger and require massive compute, data and operations support. While whitebox regression models frequently have under 10 to 100 parameters, blackbox models can range in scale from 66M (DistilBERT) to the rumored 8x220B = 1.76T parameters for GPT4 (The Decoder, 2023). It’s been estimated ChatGPT costs $700k/day to operate, but the performance, flexibility and emergent capabilities of LLMs have justified the rapidly-growing research and investment.

XAI can be divided into two categories: intrinsic and post hoc. A whitebox model is an intrinsically explainable model, simple and transparent enough to be understood without additional tools. Blackbox models are post hoc models that require additional tools to obtain explanations. XAI can also be divided into local and global explanations, where local explanations focus on explaining individual predictions like the peaks/valleys of a time series, while global explanations focus on explaining the overall behavior of the model, such as the general shapes of diachronic sentiment plots.

LIME (Ribeiro, 2016) and SHAP (Lundberg, 2017) are widely-used local post hoc methods in Explainable AI (XAI) for blackbox models. LIME generates local explanations by perturbing input instances and approximating the model's behavior, providing insights into the important features that drive individual predictions. SHAP, based on cooperative game theory, assigns importance values to features by quantifying their contribution to the prediction. Both methods offer transparent explanations for individual model predictions, enhancing interpretability and understanding of complex machine learning models.

Unfortunately, popular local post hoc methods like LIME or SHAP, are inappropriate for diachronic sentiment analysis both in scale and nature. In terms of scale, there are 3,700 individual data points for the case studies in this paper. In terms of nature, these open-ended, blackbox LLMs can produce widely varying outputs in response to small perturbations, thereby failing to yield reliable insights. Instead, we utilize a novel greybox ensemble of models from simple lexical (VADER) to SOTA LLMs (GTP4 0613) with a supervisory, human-in-the-loop guide to arbitrate both local points of disagreement and variations in global plots.

In general, a greybox model is one that combines the unparalleled performance of opaque blackbox models with the explainability of whitebox models. Teacher-student greybox transfer learning is a technique where a knowledgeable "teacher" model shares its expertise with a less-capable "student" model and the student model leverages this transferred knowledge to improve its performance on the target task. The performance of the student model on narrow tasks can sometimes approach that of the blackbox model, but with the explainability of whitebox models.

The github repo accompanying this paper contains a Jupyter Colab notebook featuring a greybox ensemble that combines the performance of LLMs with the explainability of whitebox models. Here we feature it for diachronic sentiment analysis, but it could be used for any time series that is based on classification decisions. To our knowledge, this is the first greybox ensemble for diachronic sentiment analysis.

On the global level, it surfaces an XAI metric that allows comparison of whitebox and blackbox models (i.e., a greybox model) to quickly optimize the whitebox model for a given corpus. This is because we can compare the SOTA blackbox LLM (which is most performant) with the explainable models that come closest to this performance for the particular case at hand.

The result is that we can improve whitebox performance by comparing it with the SOTA blackbox model, while simultaneously achieving some level of explainability of the blackbox model by using the whitebox model explainability. A novel global XAI metric, Ensemble Curve Cohesiveness (ECC), allows measurement of small, whitebox model alignment with large, high-performance LLMs. Clear heat map visualizations of ECC for every model pair allows for quick human-in-the-loop arbitration to identify the best whitebox and blackbox model alignment for both whitebox explainability and blackbox performance

A novel local incoherence XAI metric within the greybox ensemble model further allows selective investigation of point classification disagreement among ensemble models. Here we can peek inside the hood to ascertain which classifications--according to the human-in-the-loop--capture the emotional arc best. We return to both the global and local aspects of explainability. First, we turn to the new function calling feature of 4.0, which allows new capabilities in terms of sentiment analysis with LLMs, and the greybox ensemble workflow.

To our knowledge, this is the first published use of the new function calling feature in the newest GPT4/3.5 version 0613 for sentiment analysis. The SentimentArcs-Greybox ensemble is a simplified version of SentimentArcs (Chun, 2020) augmented with (a) novel local and global XAI metrics in the context of diachronic sentiment analysis, (b) the latest SOTA GPT4/3.5 models exploiting the newest function calling feature of version 0613 for more accurate and structured replies, and (c) new visualizations for exploratory data analysis (EDA) and local/global explainability.

SentimentArcs-Greybox simplifies the original SentimentArcs ensemble from nearly three dozen models across five families down to six models across three families. These include: (a) Lexical/GOFAI: VADER, TextBlob, (b) BERT Transformers: DistilBERT, NLPTown, RoBERTa Large and (c) SOTA LLM: GPT4, GPT3.5 (versions 0613). The Lexical/GOFAI models provide the advantages of whitebox explainability while fine-tuned BERT models vie for top spots on the SOTA leaderboards outperforming older versions of GPT (PapersWithCode, 2023). We compare several popular BERT models with the newest SOTA GPT models using naïve zero-shot prompting and new structured function API interface (OpenAI, 2023).

The new workflow for SentimentArcs-Greybox is shown in Figure 5. For our purposes the two case studies were chosen because they contrast a highly-coherent ensemble model (Woolf) and a highly-incoherent ensemble model (Morrison). In the case of Morrison, we ascertained that Transformer models (RoBERTa) were best able to capture the nuance of the emotional arc (Elkins, 2022). These Transformer models sacrificed explainability, however. Now using a grey-box methodology, we are able to use even more performant LLMs alongside coherence and incoherence metrics that allow more focused exploration for explainability.

Text is first read and segmented into semantic units of sentences using this functionality in either the NLTK (Bird, 2009) or PySBD (Sadvilkar, 2020) Python libraries. For the two novels in this paper, the more sophisticated PySBD segmentation of Woolf’s *To the Lighthouse* resulted in exactly 3,700 sentences. Next, Morrison’s *Beloved* was first segmented into 8,057 sentences. Access to the new GPT4 API is in private beta at this time and several automated runs to process 3,700 calls took an average of 3.5 hours with several catastrophic failures. Since each sentence requires a separate API call with a relatively constant overhead, it would require an estimated 7.6 hours to process 8,057 sentences. So starting with the shortest, each sentence of *Beloved* was agglomerated with the adjacent shorter sentence until the 8,057 sentences were reduced to exactly 3,700 sentence(s). This drastically reduced the estimated API function call overhead time to 3.5 hours and makes future comparisons across works simpler.

The ensemble sentiment plots in Figure 3 (‘Beloved’ by Toni Morrison) and Figure 4 (‘To the Lighthouse’ by Virginia Woolf) highlight the contrast between text that just happen to be highly coherent (‘To the Lighthouse’) with text that are relatively incoherent (‘Beloved’) given the our initial minimal ensemble of only half a dozen models. The rest of this paper will explore the more coherent novel ‘To the Lighthouse’ due to limited space and to avoid complications that would distract from explaining core aspects of this novel greybox ensemble approach. The initial coherence of the ensemble plots depend upon many factors, principally the text in question, the extensiveness/diversity of the models in the ensemble and how well these two happen to be matched. For curious readers, the next steps for analyzing the more challenging language and style of texts like ‘Beloved’ would be to add more of the 22 whitebox models in the original SentimentArcs ensemble, many of which can have multiple variants based upon different hyperparameter settings (e.g. GOFAI classifier models like logistic regression, naive bayes, random forests, XGBoost, etc). Using our greybox ensemble methodology to find the best whitebox model is not a deterministic process with guaranteed results, although casting a wider net via larger model ensembles increases the probability of success.

A diagram of a workflow

Description automatically generated with low confidence

Figure 5. SentimentArcs-Greybox Workflow

Next, the raw text of each sentence(s) segment was directly processed by the GPT4/3.5 models with API calls using the new function object feature in the latest 0613 version. This function object defines a structured sentiment analysis interface in JSON, which is the same format used to fine-tune version 0613 models. This new function object drastically reduces the problem of guiding GPT to perform narrow tasks and return data in consistent format. The function object can be viewed as a very structured form of zero-shot prompting that leads to better and more consistent results. The function object provides a function “description” (“Finds sentiment polarity and emotion”) along with a function signature of all expected return variables along with their types/values (“polarity”: string in ["positive", "negative", "neutral"] and “emotion”: string in [“happiness”,”sadness”,”anger","fear","disgust","surprise"]). Using this new API function object resulted in less than 1/3,700 malformed sentiment labels. This obviated much of the need for advanced prompt engineering strategies to perform sentiment classification. Additional minimal text cleaning was done to avoid the most serious potential errors. Details of the models within the simplified ensemble are given in Table 1.

| Novel | Model | Type | Size | Returns | Notes: |
| --- | --- | --- | --- | --- | --- |
| B/TTL | VADER | Lexicon/GOFAI | 7,500 words | -1.0/+1.0 | Plus heuristic rules |
| B/TTL | TextBlob | Lexicon/GOFAI | 2,917 words | -1.0/+1.0 | Plus heuristic rules |
| TTL | DistilBERT | BERT | 66M | Pos/Neg | Uncased, trained on SST2 |
| B | NLPTown | BERT | 110M-340M? | 1-5 stars | Uncased, product reviews/6 languages |
| B/TTL | RoBERTa Lg | BERT | 355M | Pos/Neg | 15 diverse datasets |
| B/TTL | gpt-3.5-turbo-0613 | GPT | >> 175B (GPT3) | Pos/Neu/Neg | Released June 27, 2023 |
| B/TTL | gpt-4-0613 | GPT | 1.7T (8\*220B) | Pos/Neu/Neg | Limited Beta as of June 27, 2023 |

Table 1: SentimentArcs-GreyBox ensemble model descriptions

1. https://github.com/cjhutto/vaderSentiment

2. https://github.com/sloria/TextBlob/blob/dev/textblob/en/en-sentiment.xml

3. https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english

4. https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

5. https://huggingface.co/siebert/sentiment-roberta-large-english

6. https://platform.openai.com/docs/models

7. https://the-decoder.com/gpt-4-is-1-76-trillion-parameters-in-size-and-relies-on-30-year-old-technology/

At inference time, each of the 3,700 sentence(s) segments are sequentially passed to each model to get a sentiment value. Models output as sentiment real numbers (floating point number between -1.0 to 1.0) or a label (positive/negative, positive/negative/neutral or 1-5 stars) as shown in histograms Figure 6 and described in Table 1. Labels are converted to floating point numbers. Then all values are normalized to the range -1.0 to 1.0 and concatenated to form a time series representing the evolution of emotional polarity over narrative time (Fig 3, Fig 4). Two exceptions to this are (a) to remove DistillBERT from the ensemble for *Beloved* and (b) to remove NLPTown from the ensemble for *To the Lighthouse*. In each case this was due to relatively poor performance on the respective novel.

A picture containing text, diagram, screenshot, plot

Description automatically generated

Figure 6. Sample distribution of sentiment values by model

Raw sentiment time series are then smoothed using a simple moving average (SMA) with a 10% window size. All curves are normalized to fall within the same range. At this point, curves can be compared, and features extracted: story beginnings/endings, crux point peaks/valleys, and rapidly rising/falling slopes. Several customizable algorithms attempting automated peak/valley detection and visualization make it easy for a human-in-the-loop expert to adjust and analyze thousands of data points quickly (Fig.7).

A picture containing diagram, line, text, screenshot

Description automatically generated

Figure 7. Sentiment plot with automatic peak/valley detection

Before the release of ChatGPT on November 30, 2022, after analyzing over a hundred diverse works from literature, finance, medicine, social media, etc., using the old SentimentArcs ensemble, we determined that no one model worked best for every corpus. Prior to the newest OpenAI GPT4 model release, the main goal of SentimentArcs was to identify which model resulted in maximum performance for any given corpus or corpora.

A picture containing plot, line, diagram, text

Description automatically generated

Figure 8. Ensemble of sentiment plots with local coherence XAI metric below

Now, however, we are able to use the state-of-the-art (SOTA) LLMs, which are ideal for Computational Digital Humanities projects due to their ability to capture both nuance and subtlety. Their lack of explainability poses real issues however, which the grey-box method is designed to address. The top plot in Fig.8 shows ensemble sentiment curves for the relatively coherent *To the Lighthouse*. We calculate a local XAI metric called Ensemble Point Coherence (EPC) by measuring the Euclidean distance between the maximum and minimum sentiment value at each data point. The lower plot in Fig.7 uses EPC values to plot in red how ensemble agreement varies over time, highlighting potential segments of interest and enabling closer inspection of areas of both coherence (model agreement) and incoherence.

The recent restricted beta release of gpt-4-0613 fine-tuned for structured function calling plus the rapid progress in LLMs more generally may soon provide a single best model for every situation. However, such SOTA LLMs are relatively costly at scale, complex to engineer/maintain, and too large to run/fine-tune locally. Our greybox method uses massive blackbox LLMs to identify the best aligned (most performant) whitebox model for a narrow task (sentiment analysis) on a particular corpus. When one cannot be easily identified using standard hyperparameters, this allows for human-in-the-loop adjustment of parameters to bring white-box models more in line with the performance of LLMs. Such aligned whitebox models often demonstrate the performance of a LLM with all the benefits of small whitebox models.

SentimentArcs-Greybox introduces another XAI metric, Ensemble Curve Coherence (ECC) that helps identify when smaller, whitebox models are aligned with, and can be good proxies for, high-performing SOTA LLMs. Unlike Ensemble Point Coherences that use local pointwise Euclidean distance metrics to inspect details of individual sentence coherence, Ensemble Curve Coherence uses global curve Euclidean distance metrics as shown in fig.9. A second round of smoothing using the Savitzky-Golay filter further filters noise and better visualizes the fundamental model curves before calculating the ECC. SG-smoothing was chosen because it tends to preserve the area, position and width of peaks and valleys.

A picture containing diagram, plot, text, line

Description automatically generated

Figure 9. A second Savitzky-Golay smoothing enables

To identify model alignment with our SOTA LLMs (GPT4/3.5), we compute the Euclidean distance or area between these SG-smoothed curves. The heatmap in fig.10 shows in lighter colors those models best aligned using the global distance metric ECC for the *To the Lighthouse* corpus. Here, RoBERTa is best aligned with both GPT4 and GPT35 while TextBlob better aligned with only GPT35.

A blue and white squares

Description automatically generated with low confidence

Figure 10. LLM Alignment using global distance metric ECC

Moving from the global coherence metrics (model alignment) to the local point metrics (local explainability), several preliminary conclusions about GPT 3.5 and 4 can be drawn from our two case studies, although further study is warranted. First, points of maximum incoherence or disagreement were relatively rare (i.e. 47/3700 for Woolf and 78/3700 for Morrison). In most cases, disagreement can be attributed to several factors. Unlike the whitebox models, GPT3.5 and 4 were more biased towards choice (i.e. negative or positive) in cases in which there was clear emotional valence that was subtle. In contrast, the whitebox models were centered around 0.0 (Figure 6) with a granularity that allowed for slightly more nuanced results, downplaying extreme error (i.e. a score of 0.2 as opposed to “positive”). For a large number of the discrepancies between models, there was clear ambivalence as judged by the human-in-the-loop expert. For example, there is one passage in Woolf in which a character has a positive estimation of himself but the sentiment is filtered through the mind of another, whose judgment is negative. In cases like these, the incoherence points the human-in-the-loop towards the nuance and complexity of these points in the narrative.

Other points of incoherence are due to error, for example the LLMs occasionally classified the emotion correctly (i.e. fear) but classified the sentiment value incorrectly (i.e. neutral). While this seemed to happen more with GPT 3.5 than with 4, these errors occurred with both models, though we can expect improvement of newer models with time. By contrast, and far improved (in our opinion), is the ability of the LLMs to correctly identify emotion. This contrasts with our findings in earlier studies of older models, in which sentiment is often accurate but emotion classification is not.

More broadly, and given given that GPT4 and future SOTA LLMs will likely serve as a human-level sentiment classifier in the near future, the goal of SentimentArcs model ensembles has shifted from finding the best model among many to finding the best whitebox model that approximates the new gold standard of SOTA LLMs on the narrow task of sentiment classification. In a greybox ensemble configuration, SentimentArcs can identify the best small, whitebox models with guidance from SOTA LLMs. These smaller models can bring enhanced performance for specific corpora without the limitations of LLMs that include cost, complexity and lack of privacy.

**5. Conclusion**

The performance of LLMs continues to improve along a logarithmic curve as both training datasets and model parameter size grows (Tay, 2021; Maloney, 2022). As LLMs grow, they exhibit new, uncanny, human-like emergent functionality (Wei, 2022). With this new functionality, XAI becomes even more challenging. In the end, of course, these are still just models. If all models are wrong by design--merely distillations or approximations of a higher dimensional reality –-LLM explanations are even more wrong, since they are even greater simplifications. Nonetheless, LLMs like BERT and ChatGPT open exciting new opportunities that require careful interpretations. Human-in-the-loop XAI can help elucidate these.

**Reference List**

Akin, Fatih Kadir. (2022, February 25). ChatGPT Prompt Curation. Github.com. Retrieved February 26, 2023, from https://github.com/f/awesome-chatgpt-prompts?ref=FutureTools.io

Alhussain, A.I., & Azmi, A.M. (2021). Automatic Story Generation. *ACM Computing Surveys (CSUR), 54*, 1 - 38.

Anderson, P. W. (1972). More is different: broken symmetry and the nature of the hierarchical structure of science. *Science*, *177*(4047), 393-396.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T.J., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T.B., Clark, J., McCandlish, S., Olah, C., Mann, B., & Kaplan, J. (2022). Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *ArXiv, abs/2204.05862*.

Bell, A., Solano-Kamaiko, I., Nov, O., & Stoyanovich, J. (2022). It’s Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy. *2022 ACM Conference on Fairness, Accountability, and Transparency*.

Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P.K., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J.T., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., & Zhang, Y. (2018). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. ArXiv, abs/1810.01943.

Benzon, Bill. (2023, February 20). The idea that ChatGPT is simply "predicting" the next word is, at best, misleading. LessWrong.com. <https://www.lesswrong.com/posts/sbaQv8zmRncpmLNKv/the-idea-that-chatgpt-is-simply-predicting-the-next-word-is>

Beurer-Kellner, L., Fischer, M., & Vechev, M.T. (2022). Prompting Is Programming: A Query Language For Large Language Models. *ArXiv, abs/2212.06094*.

Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Toolkit. NLTK 3.5. Available at <https://www.nltk.org/>.

Boland, Joseph. "Theory of Mind in Large Language Models: A Response to Whang, Sap, and Ullman." Accessed June 28, 2023. https://www.linkedin.com/pulse/theory-mind-large-language-models-response-whang-sap-ullman-boland. Written April 3, 2023.

Borji, A. (2023). A Categorical Archive of ChatGPT Failures. *ArXiv, abs/2302.03494*.

Brahman, F., Petrusca, A., & Chaturvedi, S. (2020). Cue Me In: Content-Inducing Approaches to Interactive Story Generation. *AACL*.

Chee, Cedrick. (2023, February 26). Awesome Transformers. Github.com. Retrieved February 26, 2023, from https://github.com/cedrickchee/awesome-transformer-nlp

Choi, J.H., Hickman, K.E., Monahan, A.B., & Schwarcz, D.B. (2023). ChatGPT Goes to Law School. *SSRN Electronic Journal*.

Chun, J. (2021a). SentimentArcs: A Novel Method for Self-Supervised Sentiment Analysis of Time Series Shows SOTA Transformers Can Struggle Finding Narrative Arcs. *ArXiv, abs/2110.09454*.

Chun, J., Katz, Lauren, Elkins, Katherine and Dennen, Jim (2021b, January 25). Can a machine play? AI, Theatre, and Improvisation. Arts at Denison. Retrieved February 26, 2023, from https://www.youtube.com/watch?v=Ldb0vlCQXtQ&t=90s

Chun, J., & Elkins, K. (2022). What the Rise of AI Means for Narrative Studies: A Response to “Why Computers Will Never Read (or Write) Literature” by Angus Fletcher. *Narrative, 30*, 104 - 113.

Chun, J. (2023). SentimentArcs Jupyter Notebooks Repository. Github.com. Retrieved February 26, 2023, from <https://github.com/jon-chun/sentimentarcs_notebooks>

Chun, Jon. "SentimentArcs-Greybox." 2023. GitHub repository. Accessed June 28, 2023. URL: <https://github.com/jon-chun/SentimentArcs-Greybox>.

Chung, J.J., Kim, W., Yoo, K., Lee, H., Adar, E., & Chang, M. (2022). TaleBrush: Sketching Stories with Generative Pretrained Language Models. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.

Crawford, 1974, Kate. (2020). *Atlas of AI*. Yale University Press.

DAIR-AI. (2023, February 26). Prompt Engineering Guide. DAIR-AI. Retrieved February 26, 2023, from https://github.com/dair-ai/Prompt-Engineering-Guide

The Decoder. “GPT-4 Is 1.76 Trillion Parameters in Size and Relies on 30-Year-Old Technology.” The Decoder. Accessed June 28, 2023. <https://the-decoder.com/gpt-4-is-1-76-trillion-parameters-in-size-and-relies-on-30-year-old-technology/>.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv, abs/1810.04805*.

Du, W., Kim, Z., Raheja, V., Kumar, D., & Kang, D. (2022). Read, Revise, Repeat: A System Demonstration for Human-in-the-loop Iterative Text Revision. *ArXiv, abs/2204.03685*.

Electronic Freedom Frontier (EFF) (2023). Measuring the Progress of AI Research. Retrieved February 20, 2023 from <https://www.eff.org/ai/metrics>

Elkins, Katherine. *The Shapes of Stories*: Sentiment Analysis for Narrative. Cambridge: Cambridge UP, 2022.

Elkins, K., & Chun, J. (2019). Can Sentiment Analysis Reveal Structure in a Plotless Novel? *ArXiv, abs/1910.01441*.

Github.com. (2019, March 25). Topic: Sentiment Analysis. Github.com. Retrieved February 26, 2023, from <https://github.com/search?o=desc&q=sentiment+analysis&s=stars&type=Repositories>

Goldfarb-Tarrant, S., Chakrabarty, T., Weischedel, R.M., & Peng, N. (2020). Content Planning for Neural Story Generation with Aristotelian Rescoring. *Conference on Empirical Methods in Natural Language Processing*.

Gurrapu, S., Kulkarni, A., Huang, L., Lourentzou, I., Freeman, L.J., & Batarseh, F.A. (2023). Rationalization for Explainable NLP: A Survey. *ArXiv, abs/2301.08912*.

Hartmann, J., Schwenzow, J., & Witte, M. (2023). The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *ArXiv, abs/2301.01768*.

Hashimoto, T. (2021). Model Performance Scaling with Multiple Data Sources. *International Conference on Machine Learning*.

Hedström, A., Weber, L., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., & Höhne, M.M. (2022). Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations. ArXiv, abs/2202.06861.

Heimerl, A., Weitz, K., Baur, T., & André, E. (2022). Unraveling ML Models of Emotion With NOVA: Multi-Level Explainable AI for Non-Experts. IEEE Transactions on Affective Computing, 13, 1155-1167.

Hussain, F., Hussain, R., & Hossain, E. (2021). Explainable Artificial Intelligence (XAI): An Engineering Perspective. *ArXiv, abs/2101.03613*.

Hutto, C.J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*.

Ippolito, D., Yuan, A., Coenen, A., & Burnam, S. (2022). Creative Writing with an AI-Powered Writing Assistant: Perspectives from Professional Writers. *ArXiv, abs/2211.05030*.

Ivanovs, Maksims, Roberts Kadikis and Kaspars Ozols. “Perturbation-based methods for explaining deep neural networks: A survey.” *Pattern Recognit. Lett.* 150 (2021): 228-234.

Ivgi, M., Carmon, Y., & Berant, J. (2022). Scaling Laws Under the Microscope: Predicting Transformer Performance from Small Scale Experiments. *Conference on Empirical Methods in Natural Language Processing*.

Jacovi, Alon. “Trends in Explainable AI (XAI) Literature.” *ArXiv* abs/2301.05433 (2023): n. Pag.

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2020). Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Jockers, M. (2020, November 24). syuzhet: Extracts Sentiment and Sentiment-Derived Plot Arcs from Text. CRAN-R Repository. Retrieved February 26, 2023, from https://cran.r-project.org/web/packages/syuzhet/index.html

Kallio, Celeste and Kurenkov, Andrey. (2023, February 24). Would Chat GPT3 Get a Wharton MBA? Wharton Business School. Retrieved February 26, 2023, from Fagan, J. (2019, March 25). ChatGPT for Creative Writing - the Good and the Bad. Storiesby.ai. Retrieved February 26, 2023, from https://storiesby.ai/p/chatgpt-for-creative-writing-the

Kenyon College (2023). Digital Humanities Research on Sentiment Analysis. Integrated Program for Humane Studies. Retrieved February 26, 2023, from <https://digital.kenyon.edu/do/search/?q=sentiment&start=0&context=13717103&facet=>

Köbis, N.C., & Mossink, L. (2020). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Comput. Hum. Behav., 114*, 106553.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydlo, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Milkowski, P., Oleksy, M., Piasecki, M., Radlinski, L., Wojtasik, K., Woźniak, S., & Kazienko, P. (2023). ChatGPT: Jack of all trades, master of none. *ArXiv, abs/2302.10724*.

Kosinski, M. (2023). Theory of Mind May Have Spontaneously Emerged in Large Language Models. *ArXiv, abs/2302.02083*.

Kreminski, M., Dickinson, M., Wardrip-Fruin, N., & Mateas, M. (2022). Loose Ends: A Mixed-Initiative Creative Interface for Playful Storytelling. *Artificial Intelligence and Interactive Digital Entertainment Conference*.

Kumar, S., Balachandran, V., Njoo, L., Anastasopoulos, A., & Tsvetkov, Y. (2022). Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey. *ArXiv, abs/2210.07700*.

Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2022). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health, 2*.

Laird, Jeremy. (2023, February 23). NVIDIA predicts explosive future for AI even as gaming graphics revenue plummet by 46%. Retrieved February 28, 2023, from <https://www.pcgamer.com/nvidia-predicts-explosive-future-for-ai-even-as-gaming-graphics-revenues-plummet-by-46/>

LangChain. (2023). LangChain: A framework for developing applications powered by language models. Retrieved from https://docs.langchain.com/docs/

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What Do We Want From Explainable Artificial Intelligence (XAI)? - A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research. *Artif. Intell., 296*, 103473

Latour, B. (2004). Why has critique run out of steam? From matters of fact to matters of concern. *Critical inquiry*, *30*(2), 225-248.

Learn Prompting. (2023, February 26). Prompt Engineering Course. Learn Prompting. Retrieved February 26, 2023, from https://learnprompting.org/docs/intro

Lester, B., Al-Rfou, R., & Constant, N. (2021). The Power of Scale for Parameter-Efficient Prompt Tuning. *ArXiv, abs/2104.08691*.

Li, M., Su, Y., Huang, H., Cheng, J., Hu, X., Zhang, X., Wang, H., Qin, Y., Wang, X., Liu, Z., & Zhang, D. (2023). Human Emotion Knowledge Representation Emerges in Large Language Model and Supports Discrete Emotion Inference. *ArXiv, abs/2302.09582*.

Li, X., Li, Y., Liu, L., Bing, L., & Joty, S.R. (2022). Is GPT-3 a Psychopath? Evaluating Large Language Models from a Psychological Perspective. *ArXiv, abs/2212.10529*.

Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Jain, A.K., & Tang, J. (2021). Trustworthy AI: A Computational Perspective. *ACM Transactions on Intelligent Systems and Technology (TIST)*.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys, 55*, 1 - 35.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv, abs/1907.11692*.

Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." In Advances in neural information processing systems, 4765-4774. 2017.

Mahowald, K., Ivanova, A.A., Blank, I.A., Kanwisher, N.G., Tenenbaum, J.B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: a cognitive perspective. *ArXiv, abs/2301.06627*.

Maloney, A., Roberts, D.A., & Sully, J. (2022). A Solvable Model of Neural Scaling Laws. ArXiv, abs/2210.16859.

Marcinkevičs, R., & Vogt, J. E. "Interpretable and explainable machine learning: A methods-centric overview with concrete examples." WIREs Data Mining and Knowledge Discovery 13, no. 3 (2023): e1493. Accessed June 28, 2023. <https://doi.org/10.1002/widm.1493>.

Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and Editing Factual Knowledge in GPT. *ArXiv, abs/2202.05262*.

Microsoft. (2022, October 12). Prompt Engine. Microsoft. Retrieved February 26, 2023, from https://github.com/microsoft/prompt-engine

Mirowski, P.W., Mathewson, K.W., Pittman, J.J., & Evans, R. (2022). Co-Writing Screenplays and Theatre Scripts with Language Models: An Evaluation by Industry Professionals. *ArXiv, abs/2209.14958*.

Mohammad, S.M., Kiritchenko, S., & Zhu, X. (2013). NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. *International Workshop on Semantic Evaluation*.

Mohseni, S., Zarei, N., & Ragan, E.D. (2018). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Trans. Interact. Intell. Syst., 11*, 24:1-24:45.

Molnar, Christoph (2022 December 14). *Interpretable Machine Learning*. Retrieved February 20, 2023 from <https://christophm.github.io/interpretable-ml-book/>

Mori, Y., Yamane, H., Mukuta, Y., & Harada, T. (2022). Computational Storytelling and Emotions: A Survey. *ArXiv, abs/2205.10967*.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., Keulen, M.V., & Seifert, C. (2022). From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. ACM Computing Surveys.

Offert, F., & Bell, P. (2020). Generative Digital Humanities. *Workshop on Computational Humanities Research*.

O’Neil, C. (2017). *Weapons of math destruction*. Penguin Books.

OpenAI (2022, November 30). ChatGPT: Optimizing Language Models for Dialogue. Openai.com. Retrieved February 26, 2023, from <https://openai.com/blog/chatgpt/>

OpenAI (2023). OpenAI API Documentation. Openai.com. Retrieved February 26, 2023, from <https://platform.openai.com/docs/introduction/overview>

OpenAI. "Function Calling and Other API Updates." OpenAI Blog, OpenAI, June 13, 2023. Accessed June 28, 2023. <https://openai.com/blog/function-calling-and-other-api-updates?ref=upstract.com>.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L.E., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., & Lowe, R.J. (2022). Training language models to follow instructions with human feedback. *ArXiv, abs/2203.02155*.

Paleja, R.R., Ghuy, M., Arachchige, N.R., Jensen, R., & Gombolay, M.C. (2022). The Utility of Explainable AI in Ad Hoc Human-Machine Teaming. ArXiv, abs/2209.03943.

PapersWithCode.com. (2023). NLP: Sentiment Analysis. Paperswithcode.com. Retrieved February 26, 2023, from <https://paperswithcode.com/area/natural-language-processing/sentiment-analysis>

Patel, Dylan. "How Much Does ChatGPT Cost to Run? $700K/day, Per Analyst." Business Insider, April 20, 2023. Accessed June 28, 2023. https://www.businessinsider.com/how-much-chatgpt-costs-openai-to-run-estimate-report-2023-4.

Pintelas, Emmanuel, Ioannis E. Livieris, and Panagiotis Pintelas. 2020. "A Grey-Box Ensemble Model Exploiting Black-Box Accuracy and White-Box Intrinsic Interpretability" Algorithms 13, no. 1: 17. <https://doi.org/10.3390/a13010017>

Poria, S., Hazarika, D., Majumder, N., & Mihalcea, R. (2020). Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research. *ArXiv, abs/2005.00357*.

PromptsLab. (2023, February 20). Awesome Prompt Engineering. Github.com. Retrieved February 26, 2023, from https://github.com/promptslab/Awesome-Prompt-Engineering

Reynolds, L., & McDonell, K. (2021). Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?: Explaining the predictions of any classifier." In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 1135-1144. 2016.

Rozado, David. (2023, January 20). ChatGPT Prompt Curation. Rozado's Visual Analytics. Retrieved February 26, 2023, from https://davidrozado.substack.com/p/political-bias-chatgpt

Rush, A.M., & Biderman, S.R. (2018). The Annotated Transformer.

Sadvilkar, N. and Neumann, M., "PySBD: Pragmatic Sentence Boundary Disambiguation," in Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS) (Online: Association for Computational Linguistics, 2020), 110-114, <https://www.aclweb.org/anthology/2020.nlposs-1.15>.

Sallam, M. (2023). The Utility of ChatGPT as an Example of Large Language Models in Healthcare Education, Research and Practice: Systematic Review on the Future Perspectives and Potential Limitations. *medRxiv*.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language Models Can Teach Themselves to Use Tools. *ArXiv, abs/2302.04761*.

Serengil, S.I., & Ozpinar, A. (2021). HyperExtended LightFace: A Facial Attribute Analysis Framework. *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, 1-4.

Smith, John. 2023. “A Critical Review of Kosinski’s Theory of Mind in Large Language Models.” Journal of Artificial Intelligence Research 54, no. 2: 123-145.

Sovrano F, Sapienza S, Palmirani M, Vitali F. Metrics, Explainability and the European AI Act Proposal. J. 2022; 5(1):126-138. <https://doi.org/10.3390/j5010010>

Swafford, Annie. (2015, March 30). Why Syuzhet Doesn’t Work and How We Know. Blog. Retrieved February 26, 2023, from https://annieswafford.wordpress.com/2015/03/30/why-syuzhet-doesnt-work-and-how-we-know/

Tay, Y., Dehghani, M., Rao, J., Fedus, W., Abnar, S., Chung, H.W., Narang, S., Yogatama, D., Vaswani, A., & Metzler, D. (2021). Scale Efficiently: Insights from Pre-training and Fine-tuning Transformers. ArXiv, abs/2109.10686.

Terwiesch, Christian. (2023, January). Would Chat GPT3 Get a Wharton MBA? Wharton Business School. Retrieved February 26, 2023, from https://mackinstitute.wharton.upenn.edu/wp-content/uploads/2023/01/Christian-Terwiesch-Chat-GTP.pdf

Universität Bern Digital Humanities. (2022 October 31). *Biased Data, Models, and Algorithms*. Retrieved February 20, 2023, from <https://www.dh.unibe.ch/e330319/e336052/e1282030/Biased_Data_Models_2022-final_ger.pdf>

Wankhade, M., Rao, A.C., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review, 55*, 5731 - 5780.

Wang, Y., Lin, J., Yu, Z., Hu, W., & Karlsson, B.F. (2022). Open-world Story Generation with Structured Knowledge Enhancement: A Comprehensive Survey. *ArXiv, abs/2212.04634*.

Wang, Yongjie (2023). *Awesome Explainable AI*. Github.com, Retrieved February 20, 2020, from <https://github.com/wangyongjie-ntu/Awesome-explainable-AI>

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent Abilities of Large Language Models. ArXiv, abs/2206.07682.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P., Mellor, J.F., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S.M., Kenton, Z., Hawkins, W.T., Stepleton, T., Birhane, A., Hendricks, L.A., Rimell, L., Isaac, W.S., Haas, J., Legassick, S., Irving, G., & Gabriel, I. (2022). Taxonomy of Risks posed by Language Models. *2022 ACM Conference on Fairness, Accountability, and Transparency*.

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F.B., & Wilson, J. (2019). The What-If Tool: Interactive Probing of Machine Learning Models. IEEE Transactions on Visualization and Computer Graphics, 26, 56-65.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D.C. (2023). A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *ArXiv, abs/2302.11382*.

Xiaowen Ding, Bing Liu and Philip S. Yu. "A Holistic Lexicon-Based Approach to Opinion Mining." Proceedings of First ACM International Conference on Web Search and Data Mining (WSDM-2008), Feb 11-12, 2008, Stanford University, Stanford, California, USA.

Yang, K., Liu, D., Lei, W., Yang, B., Xue, M., Chen, B., & Xie, J. (2022). Tailor: A Prompt-Based Approach to Attribute-Based Controlled Text Generation. *ArXiv, abs/2204.13362*.

Yuan, A., Coenen, A., Reif, E., & Ippolito, D. (2022). Wordcraft: Story Writing With Large Language Models. *27th International Conference on Intelligent User Interfaces*.

Zeng, X., Chen, Q., Chen, S., & Zuo, J. (2021). Emotion Label Enhancement via Emotion Wheel and Lexicon. *Mathematical Problems in Engineering, 2021*, 1-11.

Zhang, D.T., Maslej, N., Brynjolfsson, E., Etchemendy, J., Lyons, T., Manyika, J., Ngo, H., Niebles, J., Sellitto, M., Sakhaee, E., Shoham, Y., Clark, J., & Perrault, R. (2022). The AI Index 2022 Annual Report. *ArXiv, abs/2205.03468*.

Zhang, H., Song, H., Li, S., Zhou, M., & Song, D. (2022). A Survey of Controllable Text Generation using Transformer-based Pre-trained Language Models. *ArXiv, abs/2201.05337*.

Zhang, W., Li, X., Deng, Y., Bing, L., & Lam, W. (2022). A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges. *ArXiv, abs/2203.01054*.

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., Peng, H., Li, J., Wu, J., Liu, Z., Xie, P., Xiong, C., Pei, J., Yu, P., University, L.S., University, B., University, L., University, M., University, N.T., Diego, U.O., University, D., Chicago, U.O., & Research, S.A. (2023). A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. *ArXiv, abs/2302.09419*.

Zhou, Y., Muresanu, A.I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2022). Large Language Models Are Human-Level Prompt Engineers. *ArXiv, abs/2211.01910*.