OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Machine Versus Human: A Comparative Study on Satire Annotation

Tailor Ray Corley and Avashna Govender
tailorcorley@gmail.com and agovender@theinnoverse.co.za

ABSTRACT

Satire is a notorious blind spot for Natural Language Processing (NLP) systems due to its embedded sarcasm, parody, and contextual clues; this pragmatic language often leaves the intended meaning separate from the literal one. In this study, we ask if the newest frontier of NLP, Large Language Models (LLMs), can compare to, or surpass, the abilities of human annotators in identifying satirical language. We source our assessed excerpts from theatrical texts, since they are considered the original form of satire, and because their longevity allows us to utilize texts from across three historical eras. We find that the LLMs significantly outperform the human annotators, indicating a great increase in the ability of NLP systems to identify satirical language. Our findings raise questions about why these models have gained this ability, and possible implications for their use.

INTRODUCTION

Satire represents one of the most enduring and challenging forms of human expression for computational analysis. Among the many complexities of human language, satire poses a particular difficulty for Natural Language Processing (NLP) and Machine Learning (ML) systems because its meaning frequently diverges from its literal form. From its origins in the comedies of Aristophanes in Ancient Greece, satire has functioned as a tool of social commentary, employing irony, sarcasm, and ridicule to critique individuals, institutions, and cultural practices. Over centuries, from medieval morality plays to Shakespearean drama and into contemporary media, satire has relied on a reader's or audience's ability to recognize contextual cues, shared cultural references, and rhetorical strategies such as hyperbole, parody, and inversion. These features render satire especially resistant to surface-level text analysis, since the intended meaning is often the opposite of what is explicitly stated.

Prior research in NLP has long acknowledged the challenge of satire detection. Satire detection is often grouped with irony and humor recognition, tasks that require pragmatic reasoning, world knowledge, and cultural competence. Early approaches, such as rule-based systems or feature engineering using lexical and syntactic markers, achieved only limited success [1]. Even supervised models trained on annotated corpora frequently struggled to generalize across genres and contexts. This reflects that satire's interpretive demands exceed the scope of traditional text classification pipelines [2]. Humans, by contrast, can draw upon a vast reservoir of experiential, social, and cultural knowledge when making judgments about whether a passage is satirical. This raises the central question of our work: can machines, with their

extensive training on vast corpora, detect satire with greater consistency than humans?

The rise of Large Language Models (LLMs) such as ChatGPT, Claude, and Gemini introduces a new moment in this field. OpenAI's GPT-1 was first developed in 2018, based on the Tranformer model, with a relatively small 117 million parameters trained on BooksCorpus, a dataset of 7000 unpublished books [3][4]. Continued work yielded GPT-3, with 175 billion parameters trained on a massive corpus of text, taken from websites, books, and other available written language [3]. Its release to the public began the AI boom, leading to the releases of Claude and Gemini 1.0. Since then, each company has iterated on its models in a proprietary manner, not releasing datasets or parameter counts.

Trained on unprecedented volumes of diverse textual data, these models demonstrate emergent capabilities in pragmatic reasoning, metaphor interpretation, and humor generation. The question remains, however, whether such advances extend to the more elusive task of satire recognition, and if so, whether they can approximate or even surpass human performance in annotation tasks. Given that satire is inherently context-dependent, evaluating the performance of LLMs on this dimension provides both a test of their current ability to interpret literature and a window into the broader issue of whether AI can engage with culturally embedded forms of meaning.

This paper addresses the central question through a comparative study of satire annotation across literary texts spanning three distinct eras. We employ a dataset of dramatic excerpts. These texts are presented both to human annotators and to several state-of-the-art LLMs. By systematically comparing annotation accuracy between humans and machines, we assess the extent to which current AI systems can emulate human interpretive judgments of satire. In doing so, we contribute to ongoing debates in computational linguistics and AI ethics concerning the limits of machine understanding of language, culture, and humor.

The remainder of this paper is organized as follows: Section 2 reviews related work on satire detection and figurative language in NLP. Section 3 outlines our methodology, including text selection, annotation protocols, and evaluation metrics. Section 4 presents the results of our comparative analysis. Section 5 discusses the implications of these findings. Finally, Section 6 concludes with reflections on the challenges that remain for AI in understanding culturally embedded forms of human communication.

LITERATURE REVIEW

Natural Language Processing and Its Limits

NLP is a subfield of artificial intelligence focused on enabling computers to understand and generate human language. It is what powers things like chatbots, translation apps, and voice assistants such as Siri and Alexa. Over the past few decades, NLP has made huge progress, but it still struggles with one of the hardest things about language: *context*.

Words and phrases often change interpretation depending on the situation, tone, or shared cultural background. For example, the phrase *"That's just great"* can express genuine approval or sarcasm depending on context. This is a distinction that remains difficult for machines to detect.

Mallery (1994) described this issue as "AI-complete," meaning that solving it would require the full range of human-like intelligence, including linguistic, semantic, and cultural reasoning, things that humans do naturally but machines find difficult [5].

Wu et al. (2024) argued that because pragmatic language does not have clear-cut right or wrong answers, models must develop deeper, more flexible internal representations to approximate human understanding [6]. Despite improvements in computational power and model design, Hu et al. (2023) demonstrated that state-of-the-art NLP systems still perform worse than human annotators when analyzing context-dependent language such as irony or indirect speech [7]. Similarly, Treviso et al. (2023) noted that such tasks require considerable computational resources, highlighting the practical limits of even today's largest models [8]. NLP has come far, but it still tends to struggle with the subtle, flexible nature of human communication.

The Impact of Transformers on NLP

The introduction of the Transformer architecture by Vaswani et al. (2017) marked a turning point in NLP research [9]. Unlike earlier models that processed text sequentially, transformers use an *attention mechanism* that allows them to consider multiple parts of a sentence simultaneously. This innovation enabled models to capture long-range dependencies and contextual relationships more effectively, dramatically improving performance across a range of NLP tasks.

Building on this architecture, Devlin et al. (2018) introduced *BERT*, which leveraged bidirectional language modeling to achieve state-of-the-art results in eleven benchmarks [10]. Subsequently, Brown et al. (2020) presented *GPT-3*, a large language model with 175 billion parameters capable of performing new tasks through *in-context learning* which is capable of learning from examples rather than explicit retraining [11]. Achiam et al (2023) showed scaling trends have continued, with **GPT-4** reaching performance levels comparable to top human test-takers on standardized assessments [12].

However, researchers caution that scaling alone has not resolved the issue of contextual understanding. Wei et al. (2022) observed that models sometimes display "emergent abilities" that appear unpredictably at certain scales, while Chakrabarty et al. (2022) found that apparent successes in figurative language tasks often rely on superficial patterns rather than genuine comprehension [13,14]. Thus, even as models grow more powerful, their understanding of meaning may remain partial and inconsistent.

Sentiment Analysis

Sentiment analysis refers to the process of classifying a piece of text according to the emotion or opinion it expresses. The field has progressed steadily, moving from rule-based systems to modern neural networks, yet it continues to face challenges when dealing with non-literal or context-dependent language.

Early work by Pang, Lee, and Vaithyanathan (2002) demonstrated that machine-learning techniques could outperform lexicon-based approaches, achieving around 90% accuracy on topic classification [15]. However, their sentiment classification performance was lower, at 83%, even after feature optimization, indicating the inherent complexity of interpreting emotional tone. The introduction of deep learning

brought notable improvements. Socher et al. (2013) developed a Recursive Neural Tensor Network that yielded 85.4% accuracy on binary sentiment tasks, an increase over prior methods, which had not previously reached 80% [16].

More recently, transformer-based models such as *BERT* have advanced sentiment classification further. Zheng (2025) reported accuracies between 80% and 90% when applying BERT to film dialogue. However, performance declined for movies with subtle or culturally specific emotions [17]. Farías et al. (2015) found similar difficulties in the *SemEval-2015 Task 11*, where the top systems achieved less than 76% accuracy when analyzing tweets containing figurative language [18]. Their work highlighted that words used sarcastically, such as positive words expressing negative sentiment, remained a major obstacle.

Dia and Pettersson (2024) later compared several modern sentiment models including BERT, VADER, Flair, and TextBlob and observed accuracies between 48% and 60% on social-media data [19]. These results highlight that despite architectural progress, sentiment analysis continues to perform poorly on text that depends heavily on context, irony, or cultural nuance.

Satire Detection

Machine Learning Models

Satire detection extends the challenges of sentiment analysis, as it requires identifying deliberate exaggeration, irony, and subversion of expectations. Early computational approaches relied on classical machine-learning methods. Burfoot and Baldwin (2009) employed support-vector machines (SVMs) using lexical and semantic features such as profanity, slang, and semantic inconsistencies, achieving an F1 score of 0.8 [20]. While promising, these models struggled to generalize because satire often breaks linguistic norms intentionally. Stöckl (2018) achieved similar results (F1 = 0.76) when testing on publishers outside the training set, reinforcing the difficulty of adapting to new writing styles or contexts [21].

Neural-network architectures introduced measurable gains. Yang, Mukherjee, and Dragut (2017) applied hierarchical neural networks with attention mechanisms to detect satirical news, improving F1 scores to 0.9 [22]. However, such systems required large, well-annotated datasets that were difficult to obtain. Khodak, Saunshi, and Vodrahalli (2018) addressed this limitation by compiling the **SARC** dataset comprising 1.3 million Reddit comments annotated for sarcasm using the "/s" marker which provides an extensive resource for future work [23].

Recent LLMs have produced mixed outcomes. Gole, Nwadiugwu, and Miranskyy (2024) found that a fine-tuned GPT-3 model achieved an F1 score of 0.81 on the SARC dataset, while GPT-4 reached 0.75 in zero-shot settings [24]. Dobre and Gross (2025) observed that LLMs tended to rate AI-generated satire as more satirical than human-written examples, suggesting bias toward familiar patterns in training data [25]. Similarly, Niu et al. (2024) reported that GPT-4's satire annotations were judged as more consistent but

not necessarily more accurate than human raters [26]. These studies indicate that while LLMs show progress, their comprehension of satirical intent remains superficial.

Human Annotators

Understanding satire poses difficulties even for humans. Akimoto et al. (2013) used neuroimaging to show that irony comprehension engages multiple brain regions, including the medial prefrontal cortex, superior temporal gyrus, and amygdala [27]. Each is responsible for social and emotional processing. This cognitive complexity explains why individuals vary widely in their ability to detect satire.

Olkoniemi, Ranta, and Kaakinen (2016) found that sarcasm was harder to interpret than literal or metaphorical language, with accuracy linked to working-memory capacity and emotional reasoning [28]. Bojić et al. (2025) compared 33 human annotators and 8 LLMs, finding both groups achieved low inter-rater reliability (Krippendorff's $\alpha = 0.25$), meaning that even humans often disagree on whether a passage contains satire [29].

Theatrical Texts

Theatre offers a distinctive and relatively unexplored domain for computational satire analysis. Andresen and Reiter (2024) noted that dramatic texts such as acts, scenes, and dialogues have explicit structural boundaries that make them suitable for computational modeling while preserving rich variation in tone and expression [30]. These features can both aid and challenge satire detection. Clear structure simplifies parsing, but diverse literary techniques introduce subtler forms of irony.

Historically, theatrical works provide some of the earliest examples of satire, dating back to ancient Greek plays. Piotrowski (2012) observed that linguistic changes over time, such as shifts in spelling, vocabulary, and grammar, add another layer of difficulty for NLP models trained on modern data [31]. Rosen (2012) further argued that satire's themes evolve with society's moral and political context, meaning that detecting satire across historical eras requires sensitivity to both language and culture [32].

Research Question

Across all these domains, a consistent pattern emerges: Both NLP systems and humans struggle with context-dependent language. Sentiment analysis remains unreliable for figurative text, satire detection models perform inconsistently across domains, and LLMs show variable results depending on the data source. Meanwhile, theatrical texts, which embody some of the oldest and most sophisticated uses of satire, remain under-examined.

This study therefore aims to address a critical and novel question: How accurately can state-of-the-art large language models detect satire in plays from different historical eras, compared with human

annotators?

METHODOLOGY

Text Selection

The dataset used in this work consisted of 300 excerpts, evenly distributed across three historical eras: modern (late 1800s - 1930s), Elizabethan (1560s - mid 1600s), and ancient (classical Greek). Each era contained 50 satirical and 50 non-satirical samples, resulting in a balanced design. This sample size was chosen to balance scale with feasibility, as too many samples would have required exponentially more human annotators.

Texts were obtained from publicly available digital archives, such as Project Gutenberg, Internet Archive, Folger Shakespeare Library, and Tufts' Perseus Digital Collection [33,34,35,36]. Satirical and non-satirical works were selected only if there was a broad scholarly consensus on their genre classification, established by consulting multiple independent sources, such as scholarly analyses, authoritative reference works, script forwards, and library catalog metadata. Any plays subject to sustained historical debate regarding their satirical intent were excluded. These consensus labels served as the gold-standard reference for evaluating model accuracy. For example, for the ancient corpus, excerpts were primarily drawn from Aristophanes, whose works Silk (2002) indicates are widely recognized as foundational to theatrical satire [37]. Non-satirical plays, by contrast, represented a range of genres including naturalist tragedies, historical dramas, and thrillers but were consistently serious in tone and non-ironic in intent.

Each excerpt consisted of five consecutive lines of dialogue, chosen to provide sufficient context for satire identification while remaining concise to reduce annotator fatigue. Although excerpts were initially selected through random sampling, a manual review process was subsequently applied to ensure genre representativeness. Many plays, regardless of category, contain extended passages of exposition, transitional dialogue, or episodic humor unrelated to the central thematic structure. Simple random sampling risked producing excerpts that did not adequately reflect the satirical or non-satirical nature of the work, thereby introducing potential confounds.

To mitigate this, excerpts were retained only if they met specific criteria. For satirical texts, the excerpt was required to reference the social, political, or cultural issue being satirized, and to fall within an average range of 5 to 100 words per line across the passage. Non-satirical excerpts were required to reflect the dramatic or thematic conflict central to the play, and to fall within the same words per line average. This procedure aimed to ensure that each excerpt was both representative and interpretable.

The reliance on manual review inevitably introduces an element of researcher judgment and potential bias. However, the use of consistent selection criteria and a balanced design across eras and categories was intended to reduce this risk. Readers should nonetheless consider the interpretive role of the researchers when evaluating the findings presented in this study.

Preprocessing

All excerpts were presented in English. For the modern and Elizabethan corpora, plays were included in their original published form without modernization of language. The ancient corpus consisted of established modern English translations of the original Greek texts, drawn from Tufts' Perseus Collection's standardized editions [36]. This ensured that linguistic accessibility was balanced with fidelity to the source material.

To avoid inadvertently revealing authorial intent, stage directions and paratextual commentary were excluded from all excerpts. Only spoken dialogue was retained. Line breaks and speaker attributions were preserved to maintain the natural structure of the plays and to reflect contextual cues that could influence satire recognition. In some cases, character names themselves carried satirical significance (e.g., "Mr. Zero" in Elmer Rice's *The Adding Machine*), and their inclusion was deemed essential for preserving interpretive integrity.

This pre-processing protocol was designed to present all excerpts in a consistent format while minimizing the risk of biasing annotators toward particular interpretations through extraneous information.

Models Selection

Three state-of-the-art LLMs were selected to automatically annotate the excerpts: OpenAI's GPT-5, Google's Gemini 2.5 Pro, and Anthropic's Claude Opus 4.1. These models represent the most advanced commercially available LLMs at the time of the study and were chosen for three reasons. First, they are widely used within both research and applied settings, ensuring that their performance has broad implications. Second, they originate from distinct research organizations with differing design philosophies and training pipelines, allowing for a comparative analysis across architectures. Third, preliminary benchmarks published by their developers and independent evaluations identify these models as leading performers in natural language understanding and reasoning tasks [38,39,40].

While full details of each model's training data, parameter counts, and fine-tuning processes are proprietary, existing documentation indicates substantial variation in scale, alignment methods, and safety tuning between the three systems. This diversity supports the aim of assessing whether differences in training and architecture influence performance in satire annotation.

All three models were accessed through their official APIs under comparable usage conditions. Default temperature settings and deterministic decoding strategies (temperature = 0 where applicable) were used to minimize variance across repeated trials. This deterministic approach promoted consistency and reproducibility of results across the three LLMs. However, as LLMs are proprietary and have inherent randomness, this cannot be controlled for absolutely. Results will likely vary with replication. Model updates may also alter results

Experimental Setup

AI Annotation

Each of the 300 excerpts was submitted to the three LLMs via their official Software Development Kits (SDKs). This predeveloped code allowed interaction with the LLMs using a custom Python annotation script. To minimize variability in responses, Claude and Gemini were run with temperature set to 0, ensuring deterministic outputs. GPT-5, which does not allow temperature adjustment, was run with its default settings. No maximum token limit was imposed, and internet access was disabled for all runs. Each model call was independent, with no reference to previous inputs.

The inputs provided to the LLMs were identical to those given to the human annotators, consisting solely of the excerpted dialogue without metadata such as play title or author. Alongside each excerpt, the models were supplied with the same definitions and examples of satire and non-satire that were presented to human participants. The task was to classify the excerpt with a single-word label: "satirical" or "non-satirical." Because excerpts were evaluated in isolation, randomization of order was not required.

Outputs were returned as plain text and automatically checked for compliance with the one-word format. Responses containing additional words, punctuation, or capitalization were standardized where possible (e.g., "Satirical." → "satirical"). If a response could not be standardized, the excerpt and instructions were resent until a valid one-word output was obtained. This procedure was repeated until all 300 excerpts had been annotated by each of the three LLMs.

Human Annotation

Twenty-five human annotators participated in the study. All were fluent speakers of English and followed identical annotation procedures. Participants were not required to have prior experience with satire analysis, ensuring that judgments reflected general interpretive ability rather than specialized expertise. Annotators ranged in age from 18 to 75 and ranged in education level from high-school graduates to earning doctorates.

Annotations were conducted through a customized online survey interface. To balance workload while maintaining reliability, excerpts were distributed using a balanced incomplete block design: each annotator evaluated 36 excerpts randomly drawn from the full dataset of 300, and each excerpt was independently reviewed by three different annotators.

This randomization process inherently led to variability in the excerpts each annotator received. The number of satirical samples an annotator evaluated had a mean of 18 and ranged between 11 and 26, with non-satirical samples likewise ranging from 10 to 25. The standard deviation was 4.14 samples, so no set of excerpts exceeded two standard deviations away from the mean, indicating no outsized outliers.

The number of ancient samples an annotator evaluated had a mean of 12, ranged between 5 and 19, and had a standard deviation of 3.96. The number of Elizabethan samples an annotator evaluated had a mean of 12, ranged between 7 and 17, and had a standard deviation of 2.68. The number of modern samples an

annotator evaluated had a mean of 12, ranged between 4 and 19, and had a standard deviation of 3.92. There was one outlier, an annotator who received 4 modern samples, more than two standard deviations beyond the mean of 12.

The mean word count per excerpt across all annotators was 122.54 words, with a range between 101.81 words per excerpt and 139.06 words per excerpt, with a standard deviation of 10.14 words per excerpt, and therefore no outlier annotators.

This data on the content of each annotator's sample set demonstrates that while randomization effectively distributed the samples, the variation led to little overloading of particular sample types for any annotators.

Prior to beginning, participants completed a short pre-task self-assessment, rating their perceived ability to identify satire on a ten-point scale. They were then provided with:

1. A brief description of the task and the survey interface,
2. Clear definitions of satire and non-satire,
3. One illustrative example of each type, and
4. Instructions to carefully read each excerpt in full, to refrain from using external sources, and to complete the task in a single sitting.

Excerpts were presented to annotators one at a time in randomized order. For each excerpt, participants classified the text as either *satirical* or *non-satirical* by selecting from two mutually exclusive options. The response format was strictly binary, and no additional responses were permitted. All annotators were exposed to identical instructions and task conditions to maximize consistency.

Inter-rater reliability across annotators was assessed to evaluate consistency of judgments, with details reported in the Results section.

RESULTS

Overall Accuracy

Accuracy was defined as the proportion of excerpts correctly classified as satirical or non-satirical relative to the literary consensus gold-standard labels, as established in 3.1. As shown in Table 1, LLMs substantially outperformed human annotators. Across all excerpts, the LLMs achieved an average accuracy of 85.4%, compared to 61.0% for the human participants.

Performance, however, varied notably across models. GPT-5 and Gemini 2.5 Pro achieved near-identical accuracies of 93.0% and 93.3%, respectively, while Claude Opus 4.1 performed considerably lower at 70.0%. This reduction in mean accuracy reflects the impact of including Claude in the overall LLM average.

In addition to individual performance, accuracy was also evaluated under a consensus framework. Each excerpt was independently annotated by three LLMs and three human participants, with the majority label within each group taken as the consensus decision. Under this method, the LLM consensus accuracy reached 93.3%, whereas the human consensus accuracy was 65.0%. These findings further underscore the relative advantage of LLMs in satire annotation tasks.

Table 1: Accuracy across LLM and Human Annotators

Annotator	Accuracy (%)
GPT 5	93.0%
Gemini	93.3%
Claude	70.0%
LLM Mean	85.4%
LLM Range	70.0%-93.3%
Human Mean	61.0%
Human Range	44.4%-75.0%
LLM Consensus	93.3%
Human Consensus	65.0%

Satirical vs. Non-Satirical Classification

Figure 1 presents the breakdown in performance by genre and highlights systematic differences across annotators. Claude Opus 4.1 exhibited a strong bias toward satire: while it achieved perfect recall (100%) on satirical excerpts, it misclassified 90 of 150 non-satirical excerpts as satirical, yielding a comparatively low precision of 62.5% and an F1-score of 0.769.

Gemini 2.5 Pro achieved a more balanced performance, with precision of 90.6% and recall of 96.7%, leading to the highest F1-score among the models (0.935). GPT-5 showed a slightly different pattern, achieving precision of 96.4% but somewhat lower recall (89.3%), producing an F1-score of 0.927.

Human consensus judgments revealed both lower overall performance and systematic bias. Annotators correctly identified 74.7% of satirical excerpts (recall) but only 55.3% of non-satirical excerpts, yielding precision of 62.6% and an F1-score of 0.681. This indicates a tendency to over-identify satire, though less pronounced than Claude's.

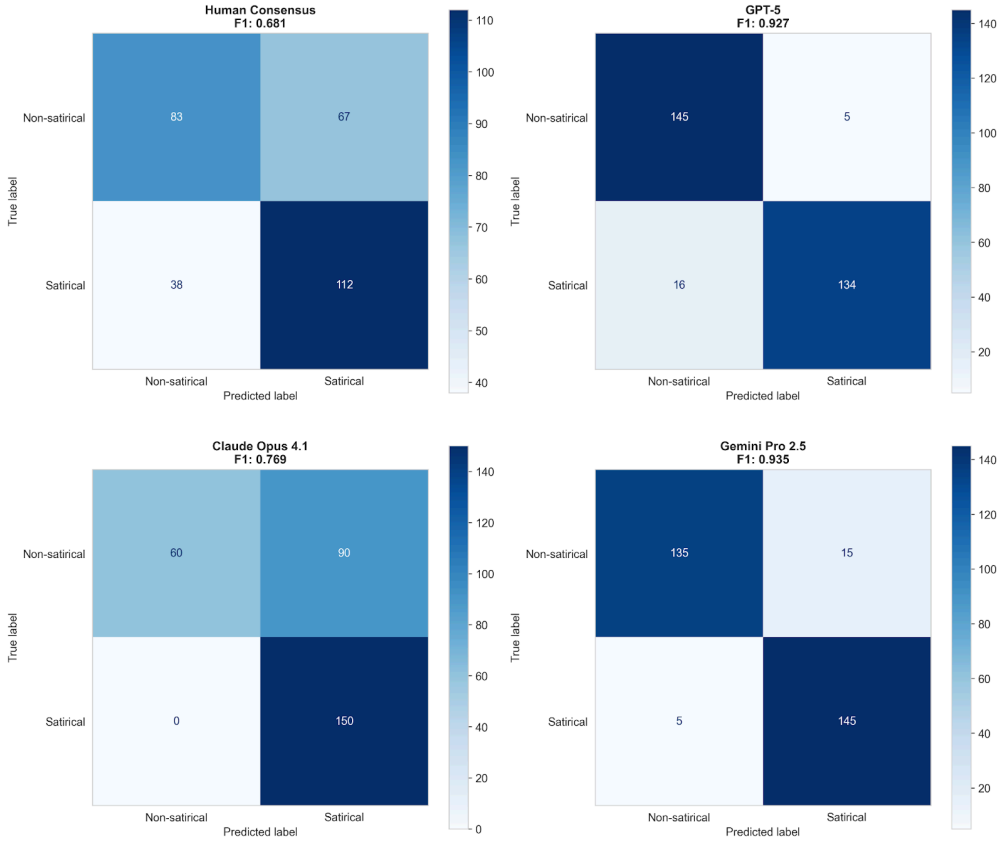

Figure 1: Confusion matrices for satire vs. non-satire classification across annotator groups. Results are shown for (a) human consensus, (b) GPT-5, (c) Claude Opus 4.1, and (d) Gemini 2.5 Pro. Each matrix reports the number of correctly and incorrectly classified excerpts, with rows indicating true labels and columns predicted labels. F1-scores are displayed for each group, reflecting the balance between precision and recall.

Table 2: Results for satire vs. non-satire classification across annotator groups. Results are shown for (a) human consensus, (b) GPT-5, (c) Claude Opus 4.1, and (d) Gemini 2.5 Pro. Each row reports the number

of correctly and incorrectly classified excerpts, as well as the F1-score, Precision, and Recall for each group.

Annotator	Precision	Recall	F1 Score	True Pos	False Pos	True Neg	False Neg
Human Consensus	0.626	0.747	0.681	112	67	83	38
GPT-5	0.964	0.893	0.927	134	5	145	16
Claude	0.625	1.000	0.769	150	90	60	0
Gemini	0.906	0.967	0.935	145	15	135	5

Accuracy by Historical Era

Figure 2 illustrates that further differences emerge when performance is analyzed by historical era. Both GPT-5 and Gemini achieved their highest accuracy on the ancient corpus, with 97% and 99%, respectively. Their performance declined as texts became more recent, falling to 95% and 91% on Elizabethan plays, and further to 87% and 90% on modern plays. Claude exhibited a different pattern: accuracy dropped from 71% on ancient texts to 64% on Elizabethan texts, but then increased to 75% on modern texts.

The LLM consensus followed a similar downward trajectory across eras, decreasing from 98% on ancient texts to 92% on Elizabethan and 90% on modern excerpts. Human consensus showed more consistency, declining from 67% on ancient samples to 63% on Elizabethan plays, before rising slightly to 65% on modern texts.

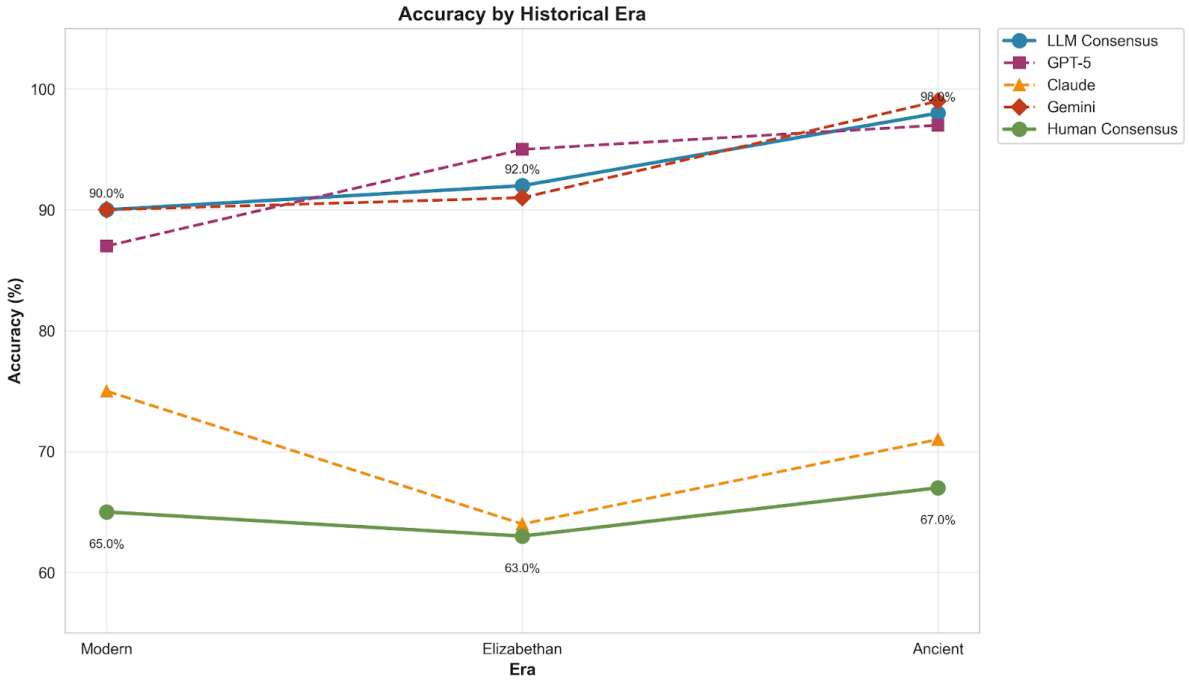

Figure 2: Accuracy of GPT-5, Claude Opus 4.1, Gemini 2.5 Pro, LLM consensus, and human consensus across three historical eras (ancient, Elizabethan, modern). The plot illustrates declining performance of GPT-5 and Gemini on more recent texts, partial recovery of Claude in the modern corpus, and relative instability in human consensus accuracy.

Inter-Rater Agreement

Inter-rater agreement was assessed by calculating the proportion of excerpts for which all three annotators within a group produced identical labels, with results displayed in Figure 3. The three LLMs reached unanimous agreement on 65.7% of excerpts (197/300), while human annotators achieved unanimous consensus on only 33.3% of excerpts (100/300).

For both groups, accuracy increased substantially under conditions of full agreement compared to cases with partial disagreement. Among the LLMs, accuracy rose from 84.6% in partially agreed cases to 98.0% in cases with unanimous agreement. A similar effect was observed for human annotators: accuracy improved from 57.0% in cases of partial disagreement to 81.0% in cases with complete agreement.

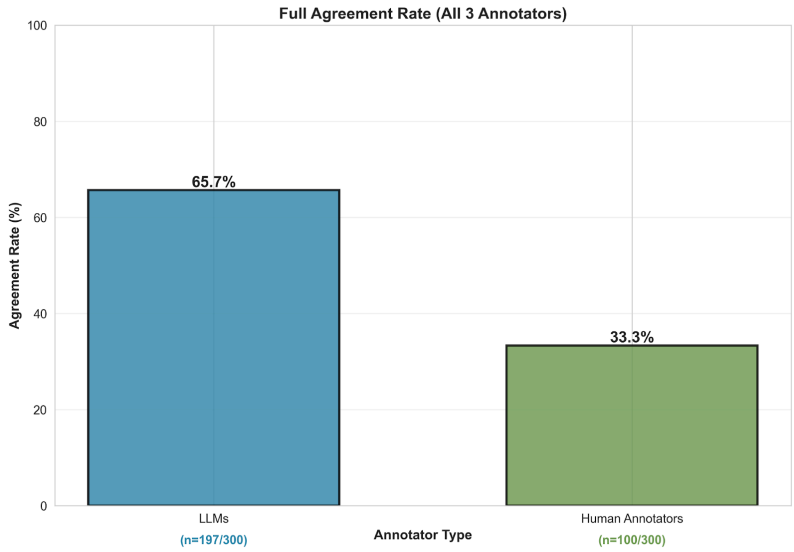

Figure 3: Bar charts reflecting the percentage full agreement rate of the (a) LLM Consensus group and the (b) Human Consensus group. Below the chart, the fractional accuracies are also given.

ANALYSIS

Null Hypothesis

The primary null hypothesis for this study is as follows: *There is no difference in satirical theatrical text annotation accuracy between Large Language Models and human annotators.* This hypothesis was tested through multiple statistical approaches to investigate the overall performance differences, pairwise comparisons, inter-rater reliability, and error distribution patterns of our data. For our analysis, an α of 0.05 was used to determine if a result was statistically significant from our null hypothesis.

Cochran's Q Test for Overall Performance Differences

Cochran's Q Test was employed to determine whether there was a statistically significant difference in overall accuracy between the 3 LLMs and the human annotators. When comparing all four annotating groups (human consensus, GPT, Gemini, and Claude), the test produced a Q statistic of 141.0584 ($p < 0.001$), leading to a rejection of the null hypothesis. This test indicates that at least one classifier differed significantly from the others in performance.

Cochran's Q Test was also employed for a second analysis on the three LLMs alone, without the Human consensus. This secondary test yielded a Q statistic of 93.8058 ($p<0.001$), indicating that at least one of the 3 LLMs' performances differed significantly.

McNemar's Test for Pairwise Comparison

McNemar's Test with continuity correction was performed to determine if there was a statistically significant difference in the proportion of correct and incorrect classifications between pairings of the four classifiers.

Human Consensus vs. LLM Comparisons

Human consensus differed significantly (refer to Table 3) from GPT-5 (χ^2=71.27, p<0.001), Gemini (χ^2=72.74, p<0.001), and the LLM consensus (χ^2=71.27, p<0.001).

Notably, human consensus did not differ significantly from Claude (χ^2=1.867, p=0.172). This unexpected result suggests that the human annotators and Claude may exhibit similar error patterns. The contingency table for this pairing reveals that both groups classified 45 identical samples incorrectly.

LLM vs. LLM comparison

For this test, each LLM was not tested against the consensus LLM data, as they were not independent of each other.

GPT-5 and Gemini showed no significant difference (χ^2=0.000, p=1.000), which is consistent with their nearly identical accuracies (93.0% vs. 93.3%)

Claude differed significantly from both GPT-5 (χ^2=45.78, p<0.001) and Gemini (χ^2=59.51, p<0.001), which is also consistent with its comparably lower accuracy of 70%.

Table 3: A pairwise comparison matrix for McNemar's test statistic (χ^2), as well as each pair's p-value. LLMs were not tested against consensus LLM data, as they were not independent of each other.

Groups	Human Consensus	LLM Consensus	GPT-5	Gemini	Claude
Human Consensus					
LLM Consensus	χ^2=71.27 p<0.001				
GPT-5	χ^2=64.99 p<0.001				

Gemini	χ^2=72.74 p<0.001		χ^2=0.000 p=1.000		
Claude	χ^2=1.867 p=0.172		χ^2=45.78 p<0.001	χ^2=59.51 p<0.001	

Fleiss' Kappa Test for Inter-Rater Agreement

We assessed inter-rater reliability using Fleiss' kappa, which quantifies the agreement beyond chance among multiple raters. It is on a scale of -1 to 1, with -1 being absolute disagreement, 0 being chance-level agreement, and 1 being absolute agreement. The human raters received a kappa of 0.0945, indicating slight agreement. This low value indicates human annotators' substantial inconsistency in how they interpret theatrical satire.

In contrast, the three LLMs received a kappa of 0.5236, indicating moderate agreement. This represents a large improvement over human accuracy and indicates a greater reproducibility of LLM satirical annotations compared to humans, even when LLMs are trained using different approaches and employ separate architectures.

This difference could be reflective of the greater number of human annotators over LLM annotators. Another possible reasoning is an increased difference in human context compared to the LLMs. While it is impossible to know the exact datasets these LLMs are trained on, it is likely that a large amounts of their training data is from similar or shared open-source datasets, creating similar learning patterns and biases. The lives of human annotators, on the other hand, will vary greatly depending on age, race, education, sex, and other personal factors. This could lead to a more randomized response distribution.

Chi-Square Test for Similarity in Misclassification Patterns.

A chi-square test of independence was employed to determine if there were significant differences in the error distributions of the human and LLM consensus. The test yielded χ^2=7.12 (df = 1, p = 0.008), indicating that the error distributions differ significantly between the two groups. This difference in misclassification patterns could be reflective of a difference in decision processes between Humans and LMMs. Since LLMs are trained to find similarities and predict outcomes in an automatic process of trial and error, it is difficult to know why they reach their decisions. Analyzing where LLMs differ could allow future research to uncover aspects of this process.

DISCUSSION

Our study highlights the improving capabilities of state-of-the-art Large Language Models to recognize satirical content through their analysis of theatrical excerpts. Our findings suggest they have advanced

past the detection capabilities of humans, possibly allowing for the development and use of many new NLP tools, if our findings apply across many fields.

LLM vs. Human Accuracy

We found that the LLMs statistically significantly outperformed the human annotators in identifying the genre of plays across all categories, with an overall consensus accuracy of 93.3% compared to 65.0%. This unexpected finding suggests that the capability of Machine Learning models to detect satirical language has increased greatly. Whereas before, Srivastava and Siddiqui (2004) found that NLP models often struggled with the non-literal language of satire, likely due to an inability to understand the societal context in which it is written, as they had little reasoning capabilities and a relatively small set of training data, transformer-based context-aware LLMs represent a new generation of NLP [41]. The immense amount of data these models are trained on has been shown to give them reasoning abilities beyond basic pattern recognition, and recent studies like one authored by Yax, Anlló and Palminteri (2024) found that this reasoning can extend beyond the capabilities of some humans in certain categories [42].

Due to the proprietary nature of LLM development, it is impossible to identify the exact updates or internal parameters that have led to this advancement past human capabilities. One possible explanation is that LLMs advanced reasoning capability has greatly elevated their abilities to analyze and understand the text. This may allow the LLMs to detect satirical content at a much higher level than humans. Future research should compare the accuracy of past models, noting the model updates that increased accuracy. By analyzing publicly released information on that models training, possible reasons could be identified.

Another possible explanation is that the model's training data provides significant context to the topics the plays discuss, which many modern humans do not have. As text modernized, we observed the difference between LLM consensus accuracy and human-annotator accuracy continually decreased; a difference of 31% for the ancient texts, 29% for the Elizabethan texts, and 25% for the modern texts. As an example of a possible difference in contextual understanding, many of the Elizabethan era plays satirized the courts of England. This topic holds nearly no importance for modern humans living in democracies, so they do not have the required context for the situations. Li et al. (2025) indicated that LLMs can exhibit a wide understanding of historical events and trends [43]. This information would prove invaluable for understanding the topics these plays discussed and would provide ample context for an informed decision on their satirical content.

A third explanation surrounds the language these plays were written in. Our dataset contained historical plays from the public domain; many annotators reported difficulty in understanding the original language and grammar contained in the excerpts. Alrefaie et al (2025) found that LLMs, conversely, understand historical vocabularies and grammar through their training data, allowing them to get a much better grasp of what is occurring in each excerpt [44]. This may have contributed to the LLMs outperforming the human annotators on the sample excerpts. More research is needed to identify if these models continue to outperform on more modern, unseen data.

The findings raise the question: Is there a difference between *recognizing* satire and *understanding* it? LLMs have no lived experience. They have not built the ability to understand emotional intent or shared cultural context through conversation the way humans do. However, their performance at an above-human level intuitively suggests there is a possibility that the LLMs do not need this "living" aspect of existence to produce the same results. Extrapolated beyond satire, this mimicry could eventually allow for the presentation of emotional understanding or logical reasoning at a level indistinguishable from a true human, or perhaps far beyond.

Genre-Specific Performance

We also found that the human annotators and Claude showed a heavy bias towards identifying excerpts as satirical. Humans correctly identified 74.7% of satirical excerpts, but only 55.3% of non-satirical excerpts, while Claude correctly identified 100% of satirical excerpts, but only 40% of non-satirical excerpts.

LLM's Performance

As shown in Figure 1, Claude demonstrated an extreme satirical bias. GPT-5 and Gemini did not show the same bias, with GPT-5 showing slightly greater accuracy with non-satirical excerpts (96.7% non-satirical vs. 89.3% satirical), and Gemini a slightly greater accuracy with satirical excerpts (90% non-satirical vs. 96.7% satirical). Compared to other models, Claude's unexpected tendency towards satire is evidence of how differences in ML training and architecture can impact results. Due to the proprietary nature of this training, it is difficult to ascertain a definite reason for this. With greater transparency from Anthropic, and indeed all other LLM designers, a more certain explanation could be given.

Using some of the limited training principles Anthropic has published, we can make informed guesses about why the model demonstrates a preference for satire. Unlike OpenAI or Google, Athropic uses a training concept called "Constitutional AI," in which it has written a "Constitution" of values it wants its LLM to possess. The training then involves the model evaluating several responses to a prompt and selecting the one that aligns with the Constitution's principles most closely. Principles from the Constitution include avoiding harmful, dishonest, or offensive language towards any group. Satire often relies on the depiction of others in a negative light and the use of dishonest or sarcastic language as a form of mockery. In training the model this way, a feedback loop could persist in which the model excessively identifies these aspects in text. This could then lead to the model over-selecting for satire, especially in situations where a person is portrayed negatively, but in the realistic style of a drama, or when tragedies overindulge in a way that can read as disingenuous.

Another possibility is that Anthropic had noted the difficulties LLMs were having in detecting irony and sarcasm and purposefully trained its models to detect these types of language. This could have had the inverse impact of the model, overemphasizing these types of language prevalent in satire, and led to its bias towards a satirical annotation.

Human Performance

Like Claude, the Human consensus accuracy for satire, at 74.7%, outperformed their accuracy for non-satire, at 55.4%. This difference of 19.3% was, however, less than that of Claude, with a difference of 60%.

The explanation behind this is ultimately a question of psychology, but a possible reason is that the human annotators were similarly predisposed to look for satire. Because most writing humans read is non-satirical, the annotators may have relied on finding signs of satire rather than signs of non-satire. This could lead to an over-inclusivity bias, where small parts of a non-satirical excerpt may have been received as sarcastic or ironic by the annotator, who would then label an otherwise non-satirical sample as satirical.

Another possible explanation is that, due to the public domain nature of the selected plays, the grammar and language employed in them may have seemed frivolous or overdone to some modern readers, elements that can be part of satire. This may have also contributed to the over-selection of pieces as satirical.

IMPACT

If LLMs have advanced past human satirical-analysis capabilities, the impact would be large. A practical application would be for online content moderation at scale. As social media apps have increased in size, the requirement to police the content being posted to their feeds has also increased. A historic difficulty in this moderation has been detecting when content is satiric or serious, which often requires manual review of edge cases, or risks either over- or under-aggressive moderation. If our findings can be reproduced in the setting of online content, the implementation of LLMs for content moderation could greatly reduce these demands and risks by more accurately assessing content without the need for manual review.

Another use case is for detecting satirical or "fake" news, as studied by Stöckl.[19] "Fake" news, or disinformation, spread under the guise of being true articles or headlines, is a significant and growing issue in our modern online and polarized world. Large swaths of people have been deceived into believing this incorrect information, and Tomassi Falegnami and Romano found that it has had real-world consequences [45]. Future research should explore whether LLM capabilities to detect satire are effective on completely unseen data and could be used in disinformation detection tools to find and restrict fake news before it can spread.

A third possible use is for interpreting unknown historical texts at mass. If, for example, a library or museum gained access to a large array of theatrical texts and wished to assess their satirical intent for labeling, an LLM-automated annotating system could be used. This would remove the burden from staff, as well as likely be more accurate, as shown by our results.

LIMITATIONS

As noted, the manual review process by the researchers inserts bias into our study; however, we believed this review was necessary to make sure our excerpts were representative of their source material and did not contain noise that could impact our results. To mitigate the impact of this bias, consistent rules were used in the review process of our excerpts.

The translation of ancient Greek pieces also introduces another layer of human input. However, modern English-speaking annotators would not be able to read the excerpts in their original language, so the translation was necessary. To mitigate this translation's impact as much as possible, text was selected from standardized editions.

Our study was limited to plays in the public domain. This may have decreased the accuracy of human annotators, as the older language and historical context could have been lost on modern readers, while being more easily picked up on by the LLMs.

The data the models are trained on is proprietary, and as such, it is impossible to know if the plays used in our study were included. Because of this knowledge, the LLMs may have identified the play the excerpt was from and then used that knowledge to decide if the text was satirical or not. Steps were taken to evaluate this hypothesis, such as using one LLM to generate satire and another to evaluate it, but these tests showed no significant difference from our original results. However, there is always a risk in using public domain or publicly available text for LLM research.

Ethically, LLMs replicating human judgment could also raise questions about replacing human interpretation in education or journalism. Taken alone, these results do not necessarily indicate that this would be successful, and further research, as well as detailed discussions and considerations, are required before LLMs can actively supplement human work in these fields.

Our data was limited to theatrical dialogue. As the first form of satire, this writing form is an excellent source for literary satire; however, for wider impact, data from other forms of media would be required.

CONCLUSION AND FURTHER RESEARCH

This study examined the ability of LLMs and human annotators to detect satire in theatrical texts drawn from different historical periods. The results demonstrate that modern LLMs have made substantial progress in interpreting complex, context-dependent language, achieving accuracy levels that surpass those of human evaluators. This suggests that LLMs have developed a form of contextual sensitivity that, while not equivalent to true understanding, allows them to recognize patterns and linguistic cues indicative of satire with remarkable consistency.

One possible reason for this advantage lies in the extensive training data of LLMs, which include

exposure to centuries of literary and cultural material. In contrast, human readers often have limited familiarity with older texts, linguistic styles, or historical references, making interpretation more challenging. The model's broad textual knowledge base therefore offers an edge in recognizing satirical intent that draws on diverse eras and writing traditions.

However, these findings should not be interpreted as evidence of genuine comprehension. LLMs excel at detecting patterns but may still rely on statistical correlations rather than an appreciation of irony or social critique - the core features of satire.

Future research should explore whether similar results are observed across other forms of satire, including contemporary plays, prose, and digital media. Such work will help clarify whether the observed performance reflects robust linguistic understanding or task-specific optimization.

Future Research could also explore samples where the satirical intent is unknown. By adding LLM and Human answer confidence as an additional variable, the differences between these two groups decision making processes could be further analyzed.

Future work should additionally explore the relative similarities of GPT-5 and Gemini, as well as between Claude and human annotators, as noted in this paper. Understanding where they differ could be key to understanding what aspect of LLMs allow them to accurately detect theatrical satire.

In summary, while state-of-the-art LLMs outperform human annotators in identifying satire within theatrical texts, their success highlights both the progress and the remaining limitations of machine understanding. These results invite deeper reflection on what it truly means to "understand" language.

REFERENCES

1. Reyes, A., Rosso, P. and Buscaldi, D. (2012). From Humor Recognition to Irony detection: the Figurative Language of Social Media. *Data & Knowledge Engineering*, [online] 74, pp.1–12. doi:https://doi.org/10.1016/j.datak.2012.02.005.

2. Jang, H. and Frassinelli, D. (2024). Generalizable Sarcasm Detection is Just Around the corner, of Course! In: K. Duh, H. Gomez and S. Bethard, eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. [online] Mexico City, Mexico: Association for Computational Linguistics, pp.4238–4249. doi:https://doi.org/10.18653/v1/2024.naacl-long.238.

3. Ray, P.P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and CyberPhysical Systems*, [online] 3, pp.121–154. doi:https://doi.org/10.1016/j.iotcps.2023.04.003.

4. Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*. [online] OpenAI. Available at: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

5. Mallery, J.C. (1987). *Thinking about foreign Policy–finding an appropriate role for artificially intelligent computers*.

6. Wu, S., Yang, S., Chen, Z. and Su, Q. (2024). Rethinking Pragmatics in Large Language models: Towards open-ended Evaluation and Preference Tuning. In: Y. Al-Onaizan, M. Bansal and Y.-N. Chen, eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. [online] Miami, Florida, USA: Association for Computational Linguistics, pp.22583–22599. doi:https://doi.org/10.18653/v1/2024.emnlp-main.1258.

7. Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E. and Gibson, E. (2023). A fine-grained Comparison of Pragmatic Language Understanding in Humans and Language Models. In: A. Rogers, J. Boyd-Graber and N. Okazaki, eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. [online] Toronto, Canada: Association for Computational Linguistics, pp.4194–4213. doi:https://doi.org/10.18653/v1/2023.acl-long.230.

8. Treviso, M., Lee, J.-U., Ji, T., Aken, B. van, Cao, Q., Ciosici, M.R., Hassid, M., Heafield, K., Hooker, S., Raffel, C., Martins, P.H., Martins, Forde, J.Z., Milder, P., Simpson, E., Slonim, N., Dodge, J., Strubell, E., Balasubramanian, N. and Derczynski, L. (2023). Efficient Methods for Natural Language processing: a Survey. *Transactions of the Association for Computational Linguistics*, [online] 11, pp.826–860. doi:https://doi.org/10.1162/tacl_a_00577.

9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017). Attention Is All You Need. In: I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds., *Advances in Neural Information Processing Systems*. [online] Curran Associates, Inc. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

10. Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational linguistics: Human Language Technologies*. [online] Minneapolis, Minnesota: Association for Computational Linguistics, pp.4171–4186. doi:https://doi.org/10.18653/v1/N19-1423.

11. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C. and Hesse, C. (2020). Language Models Are few-shot Learners. In: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan and H. Lin, eds.,

Advances in Neural Information Processing Systems. [online] Curran Associates, Inc., pp.1877–1901. Available at: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

12. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. and OpenAI (2023). *GPT-4 Technical Report*. [online] *arXiv*. Available at: https://arxiv.org/pdf/1810.04805.

13. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D. and Metzler, D. (2022). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*. [online] doi:https://doi.org/10.48550/arXiv.2206.07682.

14. Chakrabarty, T., Saakyan, A., Ghosh, D. and Muresan, S. (2022). FLUTE: Figurative Language Understanding through Textual Explanations. In: Y. Goldberg, Z. Kozareva and Y. Zhang, eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. [online] Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp.7139–7159. doi:https://doi.org/10.18653/v1/2022.emnlp-main.481.

15. Pang, B., Lee, L. and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification Using Machine Learning Techniques. In: *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*. [online] Association for Computational Linguistics, pp.79–86. doi:https://doi.org/10.3115/1118693.1118704.

16. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A. and Potts, C. (2013). Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank. In: D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu and S. Bethard, eds., *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. [online] Seattle, Washington, USA: Association for Computational Linguistics, pp.1631–1642. Available at: https://aclanthology.org/D13-1170.pdf.

17. Zheng, H. (2025). Artificial intelligence-driven sentiment analysis and optimization of movie scripts. *Discover Artificial Intelligence*, 5(1). doi:https://doi.org/10.1007/s44163-025-00361-2.

18. Farías, D., Sulis, E., Patti, V., Ruffo, G. and Bosco, C. (2015). ValenTo: Sentiment Analysis of Figurative Language Tweets with Irony and Sarcasm. In: P. Nakov, T. Zesch, D. Cer and D. Jurgens, eds., *Proceedings of the 9th International Workshop on Semantic Evaluation*. [online] Association for Computational Linguistics, pp.694–698. doi:https://doi.org/10.18653/v1/S15-2117.

19. Dia, H. and Pettersson, N. (2024). *Evaluating the Accuracy of Sentiment Analysis Models When Applied to Social Media Texts*. [online] Available at: https://kth.diva-portal.org/smash/get/diva2:1890072/FULLTEXT02.pdf.

20. Burfoot, C. and Baldwin, T. (2009). Automatic Satire detection: Are You Having a laugh? In: K.-Y. Su, J. Su, J. Wiebe and H. Li, eds., *Proceedings of the Association for Computational Linguistics-International Joint Conference on Natural Language Processing*. [online] Association for Computational Linguistics, pp.161–164. Available at: https://aclanthology.org/P09-2041/.

21. Stöckl, A. (2018). Detecting Satire in the News with Machine Learning. [online] Available at: http://arxiv.org/abs/1810.00593.

22. Yang, F., Mukherjee, A. and Dragut, E. (2017). Satirical News Detection and Analysis Using Attention Mechanism and Linguistic Features. In: M. Palmer, R. Hwa and S. Riedel, eds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. [online] Copenhagen, Denmark: Association for Computational Linguistics, pp.1979–1989. doi:https://doi.org/10.18653/v1/D17-1211.

23. Khodak, M., Saunshi, N. and Vodrahalli, K. (2018). A Large self-annotated Corpus for Sarcasm. In: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis and T. Tokunaga, eds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. [online] Miyazaki, Japan: European Language Resources Association (ELRA). Available at: https://aclanthology.org/L18-1102.pdf.

24. Gole, M., Nwadiugwu, W.-P. and Miranskyy, A. (2024). On Sarcasm Detection with OpenAI GPT-Based Models. In: *Proceedings of the 34th International Conference on Collaborative Advances in Software and COmputiNg (CASCON)*. [online] pp.1–6. doi:https://doi.org/10.1109/cascon62161.2024.10837875.

25. Dobre, A.-S. and Gross, E.-C. (2025). Evaluating AI-Generated Satire against human-written content: a Comparative Analysis. *Bulletin of the Transilvania University of Braşov Series VII Social Sciences • Law*, 18(67), pp.157–166. doi:https://doi.org/10.31926/but.ssl.2025.18.67.1.17.

26. Niu, M., El-Tawil, Y., Romana, A. and Provost, E.M. (2025). Rethinking Emotion Annotations in the Era of Large Language Models. *IEEE Transactions on Affective Computing*, pp.1–12. doi:https://doi.org/10.1109/TAFFC.2025.3584775.

27. Akimoto, Y., Sugiura, M., Yomogida, Y., Miyauchi, C.M., Miyazawa, S. and Kawashima, R. (2014). Irony comprehension: Social Conceptual Knowledge and Emotional Response. *Human Brain Mapping*, [online] 35, pp.1167–1178. doi:https://doi.org/10.1002/hbm.22242.

28. Olkoniemi, H., Ranta, H. and Kaakinen, J. (2016). Individual Differences in the Processing of Written Sarcasm and metaphor: Evidence from Eye Movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(3), pp.433–450. doi:https://doi.org/10.1037/xlm0000176.

29. Bojić, L., Zagovora, O., Zelenkauskaite, A., Vuković, V., Čabarkapa, M., Jerković, S. and

Jovančević, A. (2025). Comparing Large Language Models and Human Annotators in Latent Content Analysis of sentiment, Political leaning, Emotional Intensity and Sarcasm. *Scientific Reports*, 15(11477). doi:https://doi.org/10.1038/s41598-025-96508-3.

30. Andresen, M. and Reiter, N. (2024). *Computational Drama Analysis: Reflecting on Methods and Interpretations*. Walter de Gruyter GmbH & Co KG, pp.1–3.

31. Piotrowski, M. (2012). *Natural Language Processing for Historical Texts*. San Rafael: Morgan & Claypool Publishers, pp.1–12.

32. Rosen, R.M. (2012). Efficacy and meaning in ancient and modern political satire: Aristophanes, lenny bruce, and jon stewart. *Social Research*, [online] 79(1), pp.1–32. Available at: http://www.jstor.org/stable/23350296.

33. Project Gutenberg. (2025). *Project Gutenberg Library*. [online] Available at: https://www.gutenberg.org.

34. Internet Archive. (2025). *Internet Archive eBooks and Texts*. [online] Available at: https://archive.org/details/texts.

35. Folger Shakespeare Library. (2025). *All Shakespeare's works*. [online] Available at: https://www.folger.edu/explore/shakespeares-works/all-works.

36. Tufts Perseus Collection. (2025). *Greek and Roman Materials*. [online] Available at: https://www.perseus.tufts.edu/hopper/collection?collection=Perseus:collection:Greco-Roman.

37. Silk, M.S. (2002). *Aristophanes and the Definition of Comedy*. [online] Oxford University Press. doi:https://doi.org/10.1093/acprof:oso/9780199253821.001.0001.

38. OpenAI. (2025). *Introducing GPT-5.* [online] Available at: https://openai.com/index/introducing-gpt-5.

39. Anthropic. (2025). *Claude Opus 4.1.* [online] Available at: https://www.anthropic.com/news/claude-opus-4-1.

40. Google. (2025). *Gemini 2.5: Our most intelligent AI model.* [online] Available at: https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#enhanced-reasoning.

41. Srivastava, A. and Siddiqui, S. (2024). Sarcasm Unveiled: a Comprehensive Systematic Review of Transformer-Based Detection Models. *Journal of Electrical Systems*, [online] 20(3), pp.6151–6164. Available at: https://journal.esrgroups.org/jes/article/view/6679.

42. Yax, N., Anlló, H. and Palminteri, S. (2024). Studying and Improving Reasoning in Humans and

Machines. *Communications Psychology*, 2(51), pp.1–16. doi:https://doi.org/10.1038/s44271-024-00091-8.

43. Li, N., Yuan, S., Chen, J., Liang, J., Wei, F., Liang, Z., Yang, D. and Xiao, Y. (2025). Past Meets present: Creating Historical Analogy with Large Language Models. In: W. Che, J. Nabende, E. Shutova and M.T. Pilehvar, eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. [online] Association for Computational Linguistics, pp.3942–3957. doi:https://doi.org/10.18653/v1/2025.acl-long.200.

44. Alrefaie, M.T., Salem, F., Morsy, N.E., Samir, N. and Gaber, M.M. (2025). The Dynamics of Meaning through time: Assessment of Large Language Models. doi:https://doi.org/10.48550/arXiv.2501.05552.

45. Tomassi, A., Falegnami, A. and Romano, E. (2025). Disinformation in the Digital age: Climate change, Media dynamics, and Strategies for Resilience. *Publications*, 13(2). doi:https://doi.org/10.3390/publications13020024.