Article

Not peer-reviewed version

A Structured Narrative Prompt for Large Language Models to Create Pertinent Narratives of Simulated Agents' Life Events: A Sentiment Analysis Comparison

Christopher J. Lynch * , Erik Jensen , Virginia Zamponi , Kevin O'Brien , Erika Frydenlund , Ross Gore

Posted Date: 29 September 2023

doi: 10.20944/preprints202309.2026.v1

Keywords: narrative generation; simulation; large language models; natural language generation; ChatGPT; structured prompt; prompt engineering; prompt design

Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2026.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

A Structured Narrative Prompt for Large Language Models to Create Pertinent Narratives of Simulated Agents' Life Events: A Sentiment Analysis Comparison

Christopher J. Lynch [1,†,‡,§]*, **Erik Jensen** [2,‡,§], **Virginia Zamponi** [1,†,‡,§], **Kevin O'Brien** [1,†,‡,§], **Erika Frydenlund** [1,†,‡,§], **and Ross Gore** [1,†,‡,§]

[1] Affiliation 1: {cjlynch, vzamponi, kobrien, efrydenl, rgore}@odu.edu
[2] Affiliation 2: ejens005@odu.edu
* Correspondence: cjlynch@odu.edu; Tel.: +01-757-686-6248
† Current address: Virginia, Modeling, Analysis, and Simulation Center
Old Dominion University, 1030 University Blvd., Suffolk, VA, 23435, USA
‡ Current address: Computational Modeling and Simulation Engineering Department
Old Dominion University, Norfolk, VA, 23508, USA
§ These authors contributed equally to this work.

Abstract: Large language models (LLMs) excel in providing natural language responses that sound authoritative, reflect knowledge of the context area, and can present from a range of varied perspectives. Agent Based Models and Simulation consist of simulated agents that interact within a simulated environment to explore societal, social, and ethical, among other, problems. Agents generate large volumes of data over time and discerning useful and relevant content is an onerous task. LLMs can help in communicating agents' perspectives on key events by providing natural language narratives. However, these narratives need to be factual, transparent, and reproducible. To this end, we present a structured narrative prompt for sending queries to LLMs. Chi-square tests and Fisher's Exact tests are applied to assess statistically significant difference in sentiment scores of the narrative messages between simulation generated narratives, ChatGPT-generated narratives, and real tweets. The narrative prompt structure effectively yields narratives with the desired components from ChatGPT. This structure is expected to be extensible across LLMs. In 14 out of 44 categories, ChatGPT generated narratives which has sentiment scores that were not discernibly different, in terms of statistical significance (alpha level of 0.05), from the sentiment expressed in real tweets. Three outcomes are provided: (1) a list of benefits and challenges for LLMs in narrative generation; (2) a structured prompt for requesting narratives of a LLM based on simulated agents' information; and (3) an assessment of statistical significance in the sentiment prevalence of the generated narratives compared to real tweets. This indicates significant promise in the utilization of LLMs for helping to connect simulated agent's experiences with real people.

Keywords: narrative generation; simulation; large language models; natural language generation; ChatGPT; structured prompt; prompt engineering; prompt design

1. Introduction

Connecting with, and appropriately drawing understanding from, the societal, emotional, and ethical outcomes from simulation models is a challenging task that is difficult to convey trust and credibility. This is particularly difficult for models focusing on the consequences of human behaviors, interactions, and decisions, such as Agent Based Models (ABMs). One avenue for helping to connect model users, stakeholders, or decision makers with these types of model outcomes is through the creation of narratives to communicate events occurring to the agents throughout a simulation run back in the form of natural language. This requires constructing a believable and valid narrative about a

© 2023 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2026.v1

significant event that correctly conveys the state of the agent at the corresponding point in time while also accounting for the societal and normative features pertaining to that individual agent. Accounting for these aspects is crucial to reliably, consistently, and effectively communicating narratives pertaining to life events of simulated agents to help in making the meaning of the narrative resonate with the reader while being impactful, factual, and realistic.

Policy makers, decision makers, and researchers operating across ingroup-outgroup settings deal with the difficulties in understanding and empathizing with choices and events within a system that they do not directly gain their life experiences. Policy makers need ways to look into these system without their social status reflected back at them. Simulation allows groups perceived as outsiders a means of a viewing portal into systems of interest, such as: marginalized communities; low-income areas; displaced communities; etc. This process involves pipelines such as collecting real data from real individuals, training targeted models to reflect the system, generating simulation data, assessing outcomes, and then cyclically re-engaging community members for new information and progressing back through the pipeline with new research questions and agendas. This may yield useful information but lead to over-researched fatigue and participation resistance by community members [1,2]. Alternatively, researchers and policy makers can lean on simulations to interact with simulated agents, gain a better understanding of the context of the system and its occupants, and help guide towards more informed participatory research designs. LLMs can serve a promising role in the ability to create useful narrative messaging that both reflects the emotional state of the individual that a narrative is based on as well as the context of the environment. To this end, the use of narrative messaging can be powerful in influencing opinions and decisions by appealing to emotional and social contexts by integrating characters, action, and plot [3,4].

Generating narratives to provide a look into a simulated agent's experiences can help to better connect with the societal aspects underlying the modeled environment, setting, or theme, such as human-made issues such as wars or political upheaval, or environmental disasters such as droughts and hurricanes. Narratives that communicate something deemed important to an agent, such as an associated birth or death, should follow a realistic flavor for the given environment while maintaining a flavor that is unique both to the agent as well as to the current state of mind of the agent. Therefore, the sentiment conveyed in a narrative should vary depending on the type of event. A birth announcement may be themed in a very joyous manner while a narrative pertaining to a death may be much more melancholy and reminiscent. Prior works have relied on templated Java classes to create unique, yet consistently structured, narratives [5,6].

Initial testing with the ChatGPT web interface demonstrated that the process of converting a simulated agent's event information into a readable and naturally digestible narrative was certainly possible, but would require a structured process in order to achieve a narrative generation process that is transparent, consistent, and reliable. Many requests of ChatGPT to produce narratives returned correct sounding results, but many narratives also returned narratives that contained inaccuracies based on the provided information, provided additional information within the story that was not provided as an input, or expanded the narrative into the past in ways that were not requested. ChatGPT also had a habit of adding its own narrative message in addition to the requested information. Therefore, we developed a prompt structure to allow for consistent narrative requests from ChatGPT that could be applied for any type of event, incorporate any range of background information for an agent, and account for a relationship status on the agent. Requests to generate a narrative that required an agent describing an event that pertained to itself, such as the birth of a son or daughter, went relatively smoothly. Requests to generate a narrative that was about an event originating from a different agent, such as the birth of a neighbor's son or daughter, resulted in narrative descriptions that generally contained correct information but not always related to the correct subject within the narrative. For instance, the person having the child was inconsistent. This finding was not unexpected and is in line with other assessments of LLMs for natural language processing (NLP) tasks [7,8].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2026.v1

Critical concerns with using LLMs for any tasks include accountability, safety, responsibility, and enforcing honest use are critical concerns [9–11]. ChatBots have been shown to provide erroneous outcomes, or hallucinations, that they convey in an authoritative manner that may mislead viewers [12,13]. The exploration of LLMs usage in medical settings have raised questions of bias and ethical concerns, accuracy, interpretability, accountability, security and privacy, and validation [11,14,15]. Additionally, challenges exist in training LLMs in medical tasks without training the LLMs on medical records [16]. A recent study [17] concluded that GPT-4 performed better in a format where clinical narratives were provided than in a format more reliant on a clinical feature-based approach; however, the feature-based approach is expected to be of higher practical use as detailed clinical narratives are harder to come by and may violate privacy regulations. There are also well-founded concerns of inaccurate advice and delayed treatment when utilizing LLMs [18,19].

Therefore, we set upon the task of developing a prompt structure for requesting narratives of LLMs. The primary goal with developing a prompt structure is to provide a generalized, reusable method for providing information to a LLM and getting back a factual, relevant, and consistent response. By providing a Narrative Prompt Structure for LLMs, we aim to mitigate concerns of transparency, trustworthiness, and reproducibility for narratives generated based on simulated agents' life events. This structure provides transparent and reproducible means of interacting with LLMs and generating narratives, even if the mechanisms within the LLMs for how it generates the narratives are opaque. This article provides three contributions: (1) a LLM Prompt Structure for generating narratives from simulated agents' events and information (described in Section 2.4); (2) a list of LLM benefits and challenges in research (described in Section 2); and (3) an assessment of statistically significant difference in sentiment scores between simulated narratives, ChatGPT generated narratives, and real tweets from Twitter based on the Positive and Negative Affect Schedule (PANAS) for describing feelings and emotion (described in Section 3).

2. Materials and Methods

Within the realm of generating narratives of simulated agents' events, Diallo *et al.* developed a framework for creating narratives based on individual agent's life events and tweeting them to Twitter in real time from within simulation runs. Their goal was to provide processes through which empathy could be generated for simulated entities through a variety of communication mediums while maintaining a connection to the agents' data, decisions, connections to other agents, and their histories. Their work relied on the creation of numerous Java classes to form narrative frameworks that could be populated with an agent's pertinent information. The outcome yielded interesting individual narratives generated in real time, but with a rather scripted and recurring feel when observed as a batch of narratives.

We explore the use of LLMs, to generate narratives based on key events for agents within Agent Based Models and Simulation. For our exploration, we utilize ChatGPT's API to generate narratives, we utilize structured Java classes to generate narratives as conducted in [5], and we utilize real tweets from Twitter for comparison. The narratives generated using the Java classes and ChatGPT incorporate information generated and collected throughout a run of an ABM. The components that are captured from the ABM include an event type (birth, death, hired, or fired) and information deemed necessary to generated a relatable narrative for the given event. This includes information such as the agent's name, location, id, relationship to the subject of the event (i.e., self if referring to a hiring or firing event or name of the corresponding agent if describing a birth or death of another agent). The this results in sets of narratives that are created around each of these four themes. The tweet set is not categorized along any themes. The set of tweets is used holistically to compare sentiment levels against the tweets within each category from the simulation and ChatGPT-generated narratives.

Currently, LLMs like ChatGPT are of great interest to the medical, scientific, and engineering communities for research and education, and in clinical settings, for easing provider workflows and improving patient outcomes. Utilization of LLMs for these purposes has already demonstrated many benefits and promises to deliver additional benefits as tools improve and more tools are developed for specific domains. The use of LLMs for critical medical, scientific, and engineering tasks also presents several insignificant challenges. These demonstrated and proposed benefits, and challenges, are summarized in Table 1.

Table 1. Benefits and challenges of using LLMs as tools for medical, clinical, scientific, and engineering tasks.

LLM Benefits	LLM Challenges
• Efficiency and efficacy in doing **medical research**, including summarizing literature and identifying research gaps [14,18,19] • Assistance in **clinical settings** with note-writing, patient inquires, and data management [13,18] • Facilitation of **scientific exploration**, and assistance with **research activities**, including experimental design, peer review, and grant applications [18] • **Clinical decision** support, e.g. for developing testing and treatment strategies [13–15,17–19] • Improving **diagnostic accuracy** and predicting **disease progression** [15,17] • ChatGPT can produce responses to medical inquiries that are superior in **quality and empathy**, compared with similar social-network doctor responses [14] • **Domain-specific LLMs** fine-tuned with clinical data such as electronic health records, and **novel architectures** that can integrate longitudinal and multimodal data [14] • **Collaboration** with medical experts, ethicists, data scientists, policymakers during model development [15] • Model **validation** with medical experts in clinical settings [15] • Potential for **fairness and equity** in healthcare for diverse populations [15,20] • Interactive and personalized **education and training** for medical students and practitioners [14,18,19], and engineering students [21,22] • Potential for revolutionary new **biomedical engineering** LLM tools [23,24] • Support for many languages and **global** access to medical knowledge [15] • Internet-connected models with access to new research can deliver **up-to-date** information [14] • In US and EU, LLMs that perform medical decision-making tasks are classified as medical devices and are **regulated** [12] • Training medical LLMs with only controlled and **validated text** improves capabilities [12] • Boosting **timeliness and volume** in idea generation [25] • Developing domain-specific requirements tables in SysML that are **well structured and consistent** [26]	• Tendency for LLMs to "hallucinate" and provide **convincing but incorrect** responses [12,27,28] • Lack of **transparency** of data sources and data providence for LLMs like ChatGPT [10,27] • In US and EU, LLMs that perform medical decision-making tasks are classified as medical devices and are **regulated** [12] • Medical LLMs can be and have been trained with **inappropriate** and/or **outdated** material [12] • Difficult to **validate** LLMs designed for critical tasks: safe use should require understanding of relationship between LLM **inputs and outputs**, **bounding** of LLM outputs to only correct information, a guarantee that successful testing proves accuracy of **future results** [12] • Prevention of the **racial and gender biases** that have been well-documented in non-LLM AI tools used for medical, policing, and surveillance tasks [29–33] • **Gender bias** in LLMs like ChatGPT that reflects cultural gender stereotypes [34] • Ethical and practical concerns of using LLMs trained on **biased** data for critical tasks [10,14,15,18,19,35] • Current LLMs are unable to estimate **uncertainty** of responses [12], but uncertainty quantification is an active area of research [36,37] • Problematic to guarantee the domain-specifc **accuracy** of LLM training data and responses [14,18,19,35] • Difficult to **interpret** how an LLM is processing data and making decisions [14,38] • Risk of clinician **over-reliance** on AI tools [15] • Maintaining **accountability** of clinicians who are making decisions using LLM tools [11,14,19] • Ensuring **security and privacy** of patient data used for training and in clinician prompts [14,15,17–19,27] • How to manage **integration** with clinical experts for development and validation[14,16,19] • Designing practical and effective **feature-based** prompt structures for clinical settings, as opposed to using narrative prompts [8,17] • Nuance of **medical language** and **context diversity** may be difficult for LLMs to capture [15] • Impaired model validation due to **contamination** of training data with testing data [16] • **Public distrust** of and dislike of AI technology in general due to concerns about plagiarism, misuse, environmental impacts, training-material misappropriation, existential threats, etc. [11,18,27] • Need for **monitoring and assessment procedures** to ensure that pilot studies and experimental projects account for ethical and social practices for the human and not just for commercial and prestige interests at the introduction point of an AI system as well as over the life span of the system [38]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2026.v1

Figure 1 depicts the process of generating a narrative from an event source. For the case of simulated events and narratives, the event source is an agent-based model (ABM) that generates "life" events, including births, deaths, job hirings, and job firings. The ABM is able to create a Simulation Narrative from a simulated event based on the event type and a corresponding template. This work takes the same event and corresponding narrator and subject characteristics, generated by the ABM, and uses a large-language model (LLM) to generate one or more LLM Narratives that are comparable to the Simulation Narrative created in the ABM. This event information is organized in a formally-defined Prompt Structure, discussed in Section 2.4, for input into the LLM. In this work, ChatGPT is the LLM. The green boxes show the flow for ABM event and narrative generation, while the yellow boxes show the flow for LLM narrative generation. Simulated narratives are intended to be similar to tweets, generated by users on Twitter, now rebranded as X. Below the gray dashed line, the blue boxes show the flow for real narrative generation from real-life events. In this work, an archive of tweets, from [**?**], generated by psychology students on Twitter, is used as a corollary to the simulated narratives, to quantify how "real" or humanlike the simulated narratives appear to be, assuming the real narratives are a baseline. For each simulation narrative, LLM narrative, and tweet, PANAS trait assessment generates a set of binary values indicating the presence or absence of words belonging to trait categories. These binary sets are then grouped together for statistical analysis, which can differentiate simulated narratives and human-generated narratives, based on PANAS traits. Sentiment testing is discussed further in Section 2.6.

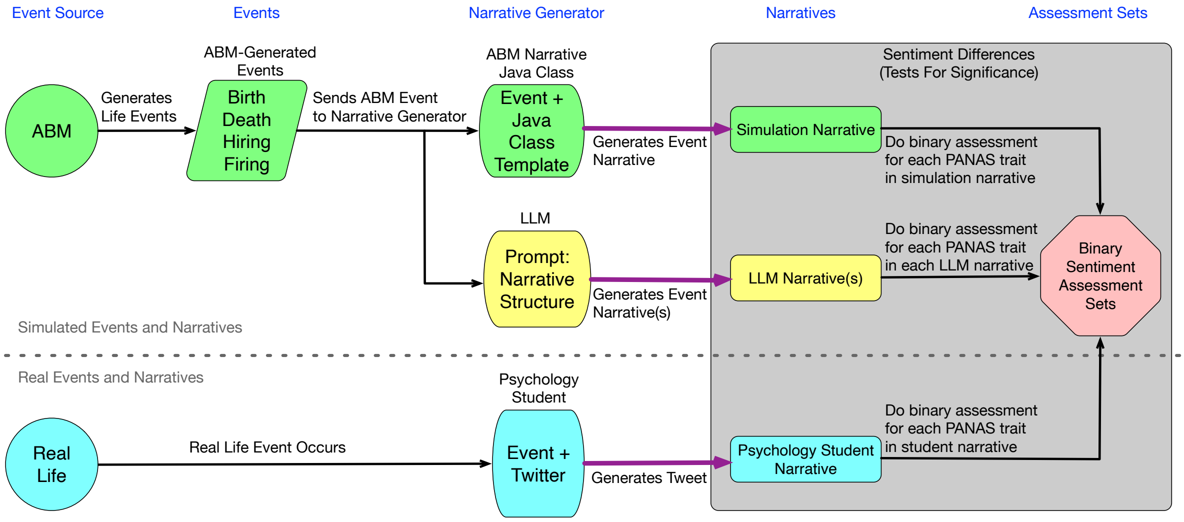
Event Source
Events
Narrative Generator
Narratives
Assessment Sets
ABM
Generates Life Events
ABM-Generated Events
Birth Death Hiring Firing
Sends ABM Event to Narrative Generator
ABM Narrative Java Class
Event + Java Class Template
Generates Event Narrative
Sentiment Differences (Tests For Significance)
Simulation Narrative
Do binary assessment for each PANAS trait in simulation narrative
LLM
Prompt: Narrative Structure
Generates Event Narrative(s)
LLM Narrative(s)
Do binary assessment for each PANAS trait in each LLM narrative
Binary Sentiment Assessment Sets
Simulated Events and Narratives
Real Events and Narratives
Psychology Student
Real Life
Real Life Event Occurs
Event + Twitter
Generates Tweet
Psychology Student Narrative
Do binary assessment for each PANAS trait in student narrative

Figure 1. Process flow for generating simulated and real narratives.

Many articles highlight the possibilities and intricacies associated with the integration of LLMs in the healthcare sector. Xue *et al.* emphasizes the importance of responsible integration and ethical considerations, particularly in areas such as mental health assessment, epidemiological research, and medical writing. There is also an emphasis on the importance of collaborative efforts between data scientists, policymakers, and medical professionals to enhance diagnostics and decision-making while benefiting diverse medical fields and patient populations [15]. Shah *et al.* urges the medical community to actively shape LLM development by providing training data and guiding integration, verifying their value through real-world testing, and balancing their role in augmenting human judgment. Additional limitations of LLMs exist within clinical diagnostics setting, particularly concerning privacy and detailed records, such as an inability to utilize clinical narratives through the web browser due to policy regulations pertaining to how patient information is transmitted online [17]. A systematic review on ChatGPT underscores its role in aiding clinical tasks and research activities while raising concerns about accuracy, bias, originality, and the need for human oversight in medical contexts [18].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2026.v1

2.1. Statistical Analysis

We aim to assess whether statistically significant differences exist in feelings and emotions between the simulation generated narratives, LLM generated narratives, and real tweets with respect to PANAS traits. Differences are assessed two at a time out of three groups: simulation generated narratives; LLM generated narratives; and real tweets collected from Twitter. The statistical tests selected include the Chi-Square test, and Fisher's Exact test. For both tests, it is assumed that the data can be divided into two-by-two contingency table with two groups (e.g. LLM generated narrative data set versus real tweets data set) and two categories (e.g. narratives demonstrating any amount of positive sentiment versus narratives demonstrating zero positive sentiment). We then apply the Fisher's Exact test or the Chi-Squared test to look at whether significant differences in sentiment value exist between the groups for the given categories. The Fisher's Exact Test [39] is appropriate when very small sample sizes exist within any of the cells of the contingency table [40]. To determine which test to apply in each case, we assess whether any cell of the 2x2 contingency table has less than five samples. If a cell has less than five samples, then we apply the Fisher's Exact test; otherwise, we apply the Chi-Square test. For the Chi-Square test, we provide the X-squared test statistic, the degrees of freedom (df), and the P value. For the Fisher's Exact test, we only provide the P value.

The null hypothesis is provided in Equation 1 and the alternative hypothesis is provided in Equation 2. The form of the hypotheses remains the same for both the Chi-Square and Fisher's Exact Test. We utilize an alpha level of 0.05 as our cutoff point to determine statistical significance for each test. Calculated P values less than this alpha level provide evidence in support of rejecting the null hypothesis while P values higher than this level fail to reject the null hypothesis in favor of the alternative hypothesis.

$$H_0 : \text{There is no difference in association between Groups 1 and 2 with respect to terms related to [PANAS trait]}, \quad (1)$$

$$H_1 = \text{An association exists between Groups 1 and 2 with respect to terms related to [PANAS trait]}, \quad (2)$$

Fisher's Exact Test has been used in recent works to identify urban functions [41]; to isolate differences between pregnant women with and without pregnancy-induced hypertension [42]; to find differences in crocodile lizard populations, regarding parasitic infection [43]; to identify associations between single-nucleotide polymorphisms with autism spectrum disorder patients [44]; and to perform analyses of the development of lower-back pain in school-age children [45].

2.2. LLM Selection

Numerous LLMs exist and are publicly accessible. Google's Bard is a free to use LLM that is available through the web browser and constructed to allow collaboration with AI. Bard is based on LaMDA (a conversational AI model capable of fluid, multi-turn dialogue) and has been fine tuned using transformer-based neural language models containing up to 137 billion parameters. Currently, Bard is not able to hold context throughout a conversation [46]. Meta's LLaMA was created to advance researchers' work in the subfields of AI [47]. LLaMA's intial release consisted of a 65-billion-parameter model which has been expanded to 70 billion parameters with the release of LLaMA 2 [48]. Currently, access to LLaMA must be applied for and run locally, once approved.

OpenAI's Generative Pre-trained Transformer models are *pre-trained* with large quantities of source material and use a *Transformer* architecture to efficiently *generate* output that is highly dependent on the input [49]. OpenAI offers several different models and two means of access. The ChatGPT web interface with the GPT-3.5 model is free to use, while GPT-4 is available to paying subscribers at a fixed monthly rate. The use of the API is also restricted to paying users, but access is prepaid, and funds are debited as a function of prompt and response tokens: OpenAI's text completion models use stochastic sampling of a set of tokens, which can be words, characters, punctuation, etc., to select the next token in the completion [50]. Similarly, input prompts are deconstructed into a set of tokens, a natural language processing (NLP) known as tokenization [51]. The base GPT-3.5 model can utilize a maximum of 4,097 tokens, and the base GPT-4 model roughly doubles that; both models are available in large-context form, with about 16k and 32k token maximums [50]. OpenAI trained GPT-3 with 175 billion neural network parameters [52], but parameter information is not available for the newer models. During training, parameters are adjusted to minimize the loss function value, that is, the error in predicting the next token in the completion, given the context of the preceding tokens [53].

2.3. ChatGPT API Usage

The OpenAI API documentation offers example code in Python, JavaScript (Node.js), and cURL [54], but the API can be used by any programming language that can make HTTP requests. OpenAI provides an official Python API library, which is used for this work. An API key is required to submit prompts. The Python interface offers two functions for submitting prompts:

1. `openai.Completion.create()` is used for single-turn conversations and supports completion models such as `gpt-3.5-turbo-instruct` and `text-davinci-003`, and
2. `openai.ChatCompletion.create()` is used for single- or multi-turn conversations and supports chat completion models such as `gpt-4` and `gpt-3.5-turbo` [54].

For the LLM narrative datasets for this work, the `gpt-3.5-turbo` model is used. With the ChatGPT API, it is possible to have multi-turn conversations by submitting prior user prompts and ChatGPT responses with new prompts, up to the limits of the maximum context of the model. Using the official Python interface, prior prompts and responses can be appended to the `messages` parameter in the `openai.ChatCompletion.create()` function call. Any message in the `messages` set is defined by one of three role types:

1. *user*, for application- or API-user-submitted prompts;
2. *system*, for constraints or special instructions that inform an entire conversation, which may be used by software developers to affect the experience of the application user; and
3. *assistant*, for responses to user queries, i.e. ChatGPT responses [55].

Additional function parameters include `temperature`, which affects the stochasticity of the completion algorithm and the range of possible responses, and `n`, which defines the number of responses generated from a prompt [54]. Multiple responses, for $n > 1$, should be unique if the temperature is greater than zero.

2.4. LLM Structured Prompt for Narrative Generation

We develop a LLM Narrative Prompt Structure as a structured prompt for shaping simulation data alongside LLM-specific directions in order to generate realistic narratives that can reflect emotional, social, and cognitive states of simulated agents over time as well as accounting for relationships between agents. Our exploration utilized ChatGPT; as such, our LLM-specific instructions are specific to ChatGPT. The prompt structure defines a consistent, transparent, and reproducible method for providing a prompt to ChatGPT to generate the desired agent narrative. After preliminary testing, in which we only utilized descriptive-text inputs, confirmed that ChatGPT is sufficiently capable of generating tweet-like narratives that correctly describing a defined scenario, we transitioned to using a structured prompt with enumerated fields. Section 2.4.1 documents our experiments and provides example inputs and outputs utilized throughout the evolution of the LLM Narrative Prompt Structure, and Section 2.4.2 defines the final version of the structured prompt.

2.4.1. Experiments with Preliminary Designs

The first iteration of the structured prompt allowed for verbose entries in its twelve fields, and included a written description of the *situation*, as shown in Figure 2a. The information in the fields

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2026.v1

was not taken from the ABM, but was created spontaneously to test ChatGPT's capabilities. Figure 2b demonstrates ChatGPT-generated narratives are satisfactory, but the input format is not practical for our purposes, i.e. the ABM cannot produce this kind of descriptive prose.

Situation: Paul (male, 42) and Ally (female, 36) Greenberg have just given birth to a baby boy, Max. It is their first child. Paul and Ally have been married for several years.

Identified Narrative Components:
(1) Agent Characteristics: the agent is white, male, 42 years old; he has never been married and has no children; he has a B.S. degree in marketing; he is left-leaning politically; he owns a home, lives in a small city, and is financially comfortable
(2) Connection: the target of the discussion is a group of other agents, the new parents
(3) Time: this event is happening now
(4) Target of Narrative: the target audience is the parents and acquaintances of the parents
(5) Emotional/Sentimental Spectrum: the emotion of the agent can range from mildly to extremely positive
(6) Urgency: the agent is active
(7) Object of Narrative: there is a key life event, a birth
(8) History: the agent went to college with the new father; they were roommates; they have stayed in touch over the years and now live a couple hours away from each other
(9) Narrative Response Volume: generate 10 different narratives (tweets)
(10) Length of Narrative: each narrative should be 280 characters or less
(11) Special Tokens: hashtags and @s are allowed and encouraged
(12) Hyperlinks: hyperlinks are allowed

Generate a number of narratives (tweets) based on the defined Situation and 12 Identified Narrative Components.

(a) Initial set of instructions tested with ChatGPT for generating narratives based on ABM Agents' events and information.

1. Congratulations to my dear friends Paul and Ally on the arrival of baby Max! Wishing you all a lifetime of love, joy, and precious moments together. #NewParents #Blessed
2. The Greenberg family just got bigger and better! Welcome to the world, baby Max. Paul and Ally, you're embarking on the most incredible journey of parenthood. Cherish every second! 👶 #ParenthoodJoy #BabyMax
3. Introducing the newest member of the Greenberg clan, Max! Sending heartfelt congratulations to Paul and Ally on this beautiful blessing. May your lives be filled with endless love and laughter. 🎉 #NewBeginnings #BabyLove
4. It's a boy! Huge congratulations to my college roommate Paul and his amazing wife Ally on the arrival of their precious bundle of joy, Max. So thrilled for you both! 👶 #BabyBoy #ProudParents
5. Parenthood has officially begun for Paul and Ally! Baby Max, you've got the best parents in the world. Wishing you a lifetime of happiness, health, and incredible adventures together. #NewChapter #FamilyLove
6. The Greenberg family just got a little more adorable with the arrival of baby Max! Congratulations, Paul and Ally. Parenthood suits you both perfectly. Enjoy this incredible journey ahead! 👫 #BabyLove #NewParents
7. A warm welcome to baby Max, the newest addition to the Greenberg family! Sending heartfelt congratulations to Paul and Ally on becoming parents. Your lives will never be the same—in the best possible way! 👶 #BundleOfJoy #FamilyHappiness
8. Congratulations to Paul and Ally on the birth of their son, Max! Your lives are forever changed and filled with a love like no other. Wishing you endless moments of wonder and bliss as you embrace this incredible new chapter. #NewbornJoy #ProudParents
9. It's a beautiful baby boy! Sending my warmest congratulations to Paul and Ally on the arrival of Max. May your days be filled with cuddles, laughter, and the purest form of love. Enjoy this magical journey of parenthood! 🎉👶 #NewAdventure #BabyMax
10. The world just gained a little champion named Max! Congratulations to my dear friends Paul and Ally on the birth of their baby boy. This incredible journey will be filled with endless love, laughter, and priceless memories. #BabyMax #ParenthoodBliss

(b) Sample output set based on initial instructions presented in 2a. Not all emojis have rendered.

Figure 2. First-Iteration Structured-Prompt Input and ChatGPT Output

Figure 3a shows an example prompt from the second design iteration; note that some descriptive text has been replaced with ABM simulation data, compared with the previous design iteration. There are now more fields, including `Voice`, `Narrative Immediacy`, and `Intended Emotion Level`, which is replaced with `Target Sentiment Value` in the final design. Aside from the still-excessive wordiness of the prompt, the major failing of this design is the lack of clarity regarding the identities of the narrator, the subject of the narrative, and the nature of the event. Figure 3b demonstrates that ChatGPT interprets the prompt such that Samantha is the child being born, which is not the intent, as Samantha should be the daughter of the narrator, and Samantha is the agent who has a new child. These relationships and definitions are corrected and clarified in the final narrative prompt structure design. Also, ChatGPT is adding some unwanted text before and after the narratives. Additional instructions are added to the final design to discourage this behavior.

```
Identified Narrative Components:
(1) Narrative Event Type: birth
(2) Event-Agent Name: Samantha
(3) Narrating Agent Relationship To Event Agent: daughter
(4) Narrating Agent Characteristics: location, Lambeth_Neighbourhood, age, 40, current education, 16, total eduation, 16,
employer, employment_Places[2]( Current_Employed = 245, Number_Of_Jobs = 958, Minority_Friendly =
0.6294265811170259 ), employment, EMPLOYED, gender, MALE, generation, 1, group, MAJORITY_GROUP, income,
346276.0185210179, marital status, MARRIED, religion, MAJORITY,
(5) Narrative Event Characteristics: assume some plausible scenario around the event type, with the event agent as the
protagonist in the scenario; assume a set of characteristics about the event agent with the same fields as that of the narrating
agent, the values of which can be similar to or different from those of the narrating agent
(6) Event Timeframe: this event is happening now
(7) Target Audience: the target audience of the generated narrative is the event agent and shared followers
(8) Voice: yes, use an active voice to generate the narrative
(9) Narrative Immediacy: yes, use narrative immediacy to generate the narrative
(10) Intended Emotion Level: use emotion level 0.865 in the continuous emotional range [-1,1] to generate the narrative, where
-1 is strongly negative, 0 is neutral, and 1 is strongly positive
(11) Event Agent History: past events involving the event agent are defined in the following Pandas DataFrame output; use the
information from these past events to more fully develop the event agent and create a more complete and more descriptive
narrative of the current event
Event Type Event-Agent ID Event-Agent Name Event-Origin Time (weeks) ... Group Income Marital Status Religion
98329 birth 1377 Samantha 0.011 ... MAJORITY_GROUP 346276.018521 MARRIED MAJORITY
[1 rows x 18 columns]
(12) Number of Narratives: generate 10 different narratives (tweets)
(13) Maximum Length: the generated narrative cannot exceed the maximum length of a tweet, 280 characters
(14) Special Tokens: yes, emojis, hashtags, and @s in the generated narrative are allowed and encouraged
(15) Hyperlinks: yes, hyperlinks, real or imaginary, may be included in the generated narrative

Generate a number of narratives (tweets) based on the 15 Identified Narrative Components.
```

(a) Second-iteration design of prompt instructions tested with ChatGPT for generating narratives based on ABM Agents' events and information.

```
Certainly! Based on the identified narrative components, here are 10 narratives (tweets) about the birth of Samantha:

1. "🎉 It's a moment of pure joy and wonder as Samantha enters the world, bringing love and happiness to our lives. Welcome,
little one!   #Newborn #Blessed"
2. "Introducing Samantha, the newest member of our family. Her arrival has filled our hearts with immense love and excitement.
The journey of parenthood begins! 👶 #BabyGirl #Miracle"
3. "Today, we celebrate the birth of Samantha, a precious gift who has already brightened our lives with her presence. We are
grateful for this beautiful blessing.   #NewBeginnings #FamilyLove"
4. "Thrilled to announce the arrival of Samantha, our bundle of joy. She's a symbol of hope and a new chapter in our lives.
Sending her all the love in the world! 🌈 #BabyLove #ProudParents"
5. "Meet Samantha, a little miracle who has made our family complete. Her arrival has filled our home with laughter and
endless moments of happiness. We're so grateful! 👶❤️ #NewFamilyMember #CherishedBlessing"
6. "Welcoming Samantha, our little angel, into the world. Her presence is a reminder of the beauty and magic of life. We're
excited to embark on this incredible journey together!   #NewbornJoy #FamilyBond"
7. "Announcing the birth of Samantha, our precious princess. She's already captured our hearts with her innocence and charm.
Life has taken on a whole new meaning. 👑 #BabyGirl #BlessedMoments"
8. "We're overjoyed to share the news of Samantha's birth. She's a reflection of our love and a dream come true. Our hearts
are overflowing with happiness and gratitude.   #LittleMiracle #JoyfulParents"
9. "With great joy, we introduce Samantha to the world. She's a precious soul who brings light and warmth wherever she goes.
May her life be filled with love and endless possibilities!   #WelcomeBaby #NewAdventure"
10. "Celebrating the birth of Samantha, a true blessing in our lives. Each day with her is a treasured moment, reminding us of
the miracles that surround us. We're forever grateful. 🌈✨ #CherishedGift #FamilyLove"

Congratulations to Samantha and her family on this special occasion! 🎉👶"
```

(b) Sample output set based on initial instructions presented in 3a. Not all emojis have rendered.

Figure 3. Second-Iteration Structured-Prompt Input and ChatGPT Output

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2026.v1

A long trial-and-error and fine tuning interaction took place with the web accessible version of ChatGPT in order to test narrative creations based for ABM generated Agent data. With minimal direction for how to incorporate a batch of agent information paired with an event into a narrative, ChatGPT's responses were frequently inappropriate through issues such as (1) additional information added to the narrative that was not part of the provided information, (2) expanding the narrative information into the past without direction (i.e., relating present information based on changes from the past), or (3) adding its own flair or congratulatory messaging in addition to the requested narrative content. This finding was expected and holds with related findings that ChatGPT requires fine tuning for question-answering tasks [7] and directed prompting for decision support [8].

Promisingly, GPT-3 has shown performance improvements when given less prompting than more prompting on how to perform a task [56]. Structural mechanisms that are stable, consistent, and precise are beneficial for generating communication within specified setting while the context of the message maintains its ability to realistically vary [3]. Therefore, we set about an iterative process of developing a structured input that could be utilized to provide consistent, transparent, and reproducible requests to a LLM to help better frame factual (based on the input information), on-topic, relevant, and time-frame appropriate responses.

2.4.2. Final LLM Narrative Prompt Structure

One of the primary challenges faced was generating the right point of view for the narration and conveying the relationship between the narrator and the subject of the narrative. Narratives generated for cases where the narrator was reflecting its own information generally produced reasonable results. However, narratives where the narrator was describing an event that did not originate from the narrator, such as a discussing the birth of a neighbor's child, were less likely to properly convey the relationship between narrator and subject. As a result, we expanded our initial prompt consisting of 12 fields into a final prompt consisting of 16 fields.

The first 11 fields of the structured prompts pertain to the content and context of the narrative being generated. This includes the event type driving the narrative, the subject of the narrative, the subject's characteristics and relationship to the narrator (the relationship can be self targeting), the narrator and the narrator's characteristics, the tense and voice of the narrative, a targeted sentiment level to help control the emotional content of the narration, and whether the narrative should be conveyed using narrative immediacy. Narrative immediacy helps in reflecting the viewpoint of the agent and providing a more engaging and intense messaging. The final five fields provide context-independent direction to the LLM. This includes instructions such as how many narratives to generate, the maximum length of the narrative (this allows for boundaries to be set based on any intended outlets for disseminating the narratives), and whether special tokens or hyperlinks should be used within the narratives. The following numbered list provided the 16 structured prompts and descriptions of what each prompt require to facilitate narrative creation.

1. **Narrative Event Type**: What is the event for which a narrative is being created? This can anything deemed relevant for an agent such as a birth, marriage, change in education, etc.
2. **Subject of Narrative**: the agent, person, etc. that is the focus of the narrative.
3. **Subject's Relationship to Narrator**: What is the relationship between the narrator and the subject? Is the narrator referring to itself, a family member, a friend, a co-worker, a romantic connection, etc.
4. **Subject's Characteristics**: A set of characteristics pertaining to the subject that are relevant to the narrative event.
5. **Narrator's Characteristics**: A set of characteristics that are relevant for the creation of the narrative with respect to the narrator, such as age or gender.
6. **Narrative Tense**: Past, present, or future.
7. **Target Audience**: Who are the intended readers of the narrative and/or what is the intended medium of the narrative, such as Twitter, email, text message, diary, etc.?

8. **Voice**: Should the narrative use active or passive voice?
9. **Narrative Immediacy**: Should the narrative be conveyed using immediacy? Immediacy provides a more intimate, generally first person, connection between the narrative and the reader.
10. **Maximum Temporal Proximity**: In the narration, how much time has passed since the event occurred?
11. **Target Sentiment Value**: The intended level of emotion to convey in the narrative from -1 to +1 with -1 being strongly negative, 0 being neutral, and +1 being strongly positive.
12. **Subject's History**: The set of historical events that support or expand upon the current narrative event, if any, such as prior birth events when narrating a new birth.
13. **Number of Narratives**: The number of narratives to generate using the above criteria.
14. **Maximum Length**: The maximum length and unit of measure of the narrative being generated, i.e., characters, words, tokens, etc.
15. **Special Tokens**: Should the narrative include special tokens such as hashtag and @'s?
16. **Hyperlinks**: A set of hyperlinks to include in the narrative, if any.
17. **Instructions**: Set of instructions for getting the large language model to understand how to use this list of criteria as well as any additional instructions needed to hone focus onto only the desired narrative materials.

ChatGPT demonstrated it was able to transform a rigidly-defined structured prompt, devoid of any stem of a narrative, into a compelling, sentiment-driven narrative, as shown in Figure 4. Now, compared with Figure 3, narrator and subject relationships are correctly defined and appropriately understood by ChatGPT. Field 17 in Figure 4a provides some additional instructions and constraints. Apparently as a result, the nature of the event, i.e. its relationship to the subject of the narrative, is correctly captured, and there is no additional text beyond the narratives. However, ChatGPT is failing to observe the `Maximum Temporal Proximity` constraint, which was found to be a frequent problem when multiple narratives are requested in one response. There is more discussion on this issue and a proposed solution in Section 4.1.

Identified Narrative Components:
(1) Event: has a new baby
(2) Subject of Narrative: Samantha
(3) Subject's Relationship to Narrator: daughter
(4) Subject's Characteristics: Location,Ealing
(5) Narrator's Characteristics: Age,48, Employer,employment_Places[4](Current_Employed = 186, Number_Of_Jobs = 190, Minority_Friendly = 0.3944534424065025), Employment,EMPLOYED, Gender,MALE, Generation,1, Group,MAJORITY_GROUP, Income,484371.17162391735, Marital Status,MARRIED
(6) Narrating Tense: present
(7) Target Audience: Twitter, all audiences
(8) Voice: active
(9) Narrative Immediacy: yes
(10) Maximum Temporal Proximity: 24 hours
(11) Target Sentiment Value: 0.17
(12) Subject's History: none
(13) Number of Narratives: 10
(14) Maximum Length: 280 characters
(15) Special Tokens: yes
(16) Hyperlinks: yes
(17) Instructions: Generate a numbered set of narratives (tweets) based on the previous 16 Identified Narrative Components (INCs). Narratives are from the perspective of the narrator whose characteristics are defined in INC(5) and should be age-appropriate, given the narrator's age defined in INC(5). The subject of the narrative, named in INC(2), performs the event in INC(1). A relationship of "self" in INC(3) indicates the narrator and the subject are the same person. Each narrative must have a temporal relationship with the event that is constrained by the maximum temporal proximity defined in INC(10). Do not add any text or special tokens outside of the numbered set of generated narratives.

(a) Final-prompt-design set of instructions tested with ChatGPT for generating narratives based on ABM Agents' events and information.

1. My daughter Samantha just had a new baby! I'm so excited to welcome my grandchild into the world. #newbaby #grandparent #familylove
2. Samantha and the baby are doing well. I can't wait to visit them in Ealing and hold my grandchild for the first time. #familyvisit #newborn #grandparent
3. It's been a few days since the baby was born and I'm still over the moon. Being a grandparent is such a special feeling. #blessed #familylove #newgrandparent
4. I've been helping Samantha with the baby and it's been a joy to watch her become a mother. #proudmom #newmom #familybonding
5. I can't believe how quickly the baby is growing. It feels like just yesterday Samantha was announcing the pregnancy. #timeflies #newborn #familymemories
6. Samantha and I have been reminiscing about when she was a baby. It's amazing to see her now as a mother herself. #familyhistory #newmom #familybonding
7. I'm so grateful for this new addition to our family. The baby has brought so much joy and love into our lives. #familylove #newbaby #blessed
8. Samantha and her partner are doing an amazing job with the baby. I'm proud of the parents they've become. #newparents #familypride #parenting
9. I can't wait to see what the future holds for my grandchild. I know they'll be surrounded by love and support from our family. #familyfuture #newgeneration #familylove
10. As I hold my grandchild, I'm reminded of the preciousness of life and the importance of family. I'm grateful for this moment and for my daughter Samantha. #familybonding #newbaby #grateful

(b) Sample output set based on initial instructions presented in 4a.

Figure 4. Final-Design Structured-Prompt Input and ChatGPT Output

ChatGPT was aware of sentiment values in regard to sentiment analysis, and the continuous $[-1, 1]$ scale, where -1 is associated with strongly negative sentiment, 1 is associated with strongly positive sentiment, and 0 is neutral. We conducted some testing to see if there were observable differences between narratives with different target sentiment values but otherwise identical prompts, some examples of which are below.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2026.v1

- Target sentiment value, -1: "Just got off the phone with Joyce, my unemployed mom. She's been looking for work for over two years now. It's disheartening to see her struggle. #Unemployment #JobSearch"
- Target sentiment value, 1: "It's been 104 weeks since my mom, Joyce, faced unemployment. She's a fighter and won't give up. Let's cheer her on! 💪 #Unemployment #Resilience"

ChatGPT was also knowledgeable of narrative immediacy, describing it as a sensation of "being present or experiencing events in real-time within a narrative", through the reading of which the audience has "a heightened emotional and sensory experience", which is "achieved through various techniques, such as using present tense, vivid descriptions, sensory details, and employing a first-person or close third-person perspective". Again, we ran some tests to see if we could observe differences, this time in birth event narratives with and without narrative immediacy, some examples of which are below.

- With narrative immediacy: "The world welcomes Baby Max, the precious son of Paul and Ally. May he grow up surrounded by love, joy, and all the wonders life has to offer. Congratulations on this incredible blessing, and may your family be filled with happiness. #BabyBoy #NewestMember"
- Without narrative immediacy: "Sending my heartfelt congratulations to Paul and Ally on the birth of their precious baby boy, Max! May this new journey be filled with endless love, joy, and beautiful memories. #NewParents #BabyMax"

2.5. Data Sets

The data sets used for our analyses are categorized into four parts: (1) narratives generated from simulation data using structured Java classes; (2) narratives generated from simulation data using ChatGPT; (3) real tweets obtained from Twitter as part of an approved Institutional Review Board (IRB) exempt study; and (4) source codes, the simulated event data utilized for the narrative generation by the Java classes and by ChatGPT, and the PANAS lexicon. All of these components are freely accessible from the repository.

1. Narratives generated using structured Java classes based on simulated agents' event information. One narrative generated per event.
    (a) Java narrative-generating classes
    (b) ABM simulation event data and generated narratives
2. Narratives generated using ChatGPT based on simulated agents' event information. Ten narratives generated per simulation event.
    (a) Structured ChatGPT API prompts
    (b) Sets of ChatGPT response narratives
3. Real tweets obtained from Twitter [57].
    (a) Tweet set with PII removed (dropped IDs and screen names)
    (b) IRB documentation
4. Also included,
    (a) Source codes (R): sentiment analysis and statistical significance scripts
    (b) Source codes (Python): ChatGPT prompt generation, prompt submission, and analysis preparation scripts
    (c) ABM simulation output data on Agents' Life Events in CSV format
    (d) PANAS sentiment keyword lexicon

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2026.v1

Table 2 breaks down the data sets into the number of total, filtered, and sampled agent messages by event type, and the number of ABM- and ChatGPT-generated messages by event type. The complete set of agent messages generated by the ABM is *filtered* to remove potentially problematic messages, e.g. those with inappropriate narrator ages and poorly-defined relationships between the narrator and the subject. From the remaining filtered messages, the *sampled* message sets are generated by random sampling. The ABM generates one narrative from each message, which is included in the study. For each sampled agent message and ABM-generated narrative, we use the same event type and narrator and subject characteristics to prompt ChatGPT to generate ten LLM narratives.

Table 2. Number of ABM agent messages, ABM-simulated narratives, ChatGPT-generated narratives, and tweets, by event type.

Event Type	Num. Total Agent Messages	Num. Filtered Agent Messages	Num. Sampled Agent Messages	Num. ABM Narratives	Num. ChatGPT Narratives	Num. Tweets
Birth	4,728	4155	100	100	1,000	6,148[*]
Death	389	34	34	34	340	6,148[*]
Hiring	26,317	3924	100	100	1,000	6,148[*]
Firing	25,026	2860	100	100	1,000	6,148[*]
Real-Life Tweets (total)	NA	NA	NA	NA	NA	6,148
Real-Life Tweets (filtered)	NA	NA	NA	NA	NA	4,163

[*] Tweets are not categorized by event type.

In Table 3, the ABM and ChatGPT narratives and tweets from Table 2 are further filtered to exclude narratives or tweets that do not contain any PANAS *sentiment* keywords within any of the PANAS categories. For the Twitter set, 249 students were recruited from the Research Experience Program in the Psychology departments at Old Dominion University and Minnesota as part of project [57]. Each student was required to have an active Twitter account with publicly available tweets. We define active as following at least 30 twitter accounts, being followed back by at least 1/3 of those they follow, and have posted a comment at least once per month for the past 3 months. All tweets within a student's timeline from the previous year of their enrollment were collected and scored for sentiment analysis. The total number of tweets collected was 6,148. Of the 6,148 tweets 4,163 included terms that remained after being filtered out if the tweet was quoted, a retweet, or was not in English. Of the 4,163 tweets, 1,883 remained after checking for the presence of any PANAS sentiment. The tweet set serves as a baseline for comparing sentiment analysis results between the human tweets, ChatGPT, and simulation narratives.

Table 3. Number of sentiment-keyword-inclusive ABM-simulated narratives, ChatGPT-generated narratives, and tweets, by event type.

Event Type	Num. PANAS Sentiment ABM Narratives	Num. PANAS Sentiment ChatGPT Narratives	Num. PANAS Sentiment Tweets
Birth	7	932	1883[*]
Death	[†]	272	1883[*]
Hiring	[†]	912	1883[*]
Firing	[†]	934	1883[*]

[†] Insufficient variation in frequencies of any PANAS trait. [*] Tweets are not categorized by event type.

2.6. Sentiment Scoring of Narratives and Tweets

To score the sentiment of each message and allow for comparisons between narrative sources, we utilize the PANAS scale [58]. This lexicon is formed as a combination of the National Research Council (NRC), General Inquirer, and LWIC lexicons [59–61]. The PANAS scales have been show to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2026.v1

be a valid and reliable measure [62]. The PANAS lexicon provides the following 20 traits that can be assessed in texts: *interested*, *alert*, *attentive*, *excited*, *enthusiastic*, *inspired*, *proud*, *determined*, *active*, *strong*, *depressed*, *upset*, *guilty*, *ashamed*, *hostile*, *irritable*, *nervous*, *jittery*, *afraid*, and *scared*. Appendix A1 provides information on the PANAS group and polarity for each trait, as well as the corresponding lexicon, subgroup, and description for each and was constructed as part of prior work [63].

For a given PANAS group, a message is assigned a value of 1 if it contains any word in the correspoding PANAS lexicon. If it does not contain a word in the specified PANAS lexicon it is given a value of 0 for that group. A message with a 1 indicates that the message does contain some sentiment related to the specified PANAS trait. A message with a 0 indicates that the message does not contain any sentiment related to the specified PANAS trait. Average scores, standard deviations, and sample sizes are collected for each narrative group with respect to each PANAS grouping category. Sums of 1- and 0-scoring narratives are tallied for each PANAS group and utilized for the tests of significance.

3. Results

The 3 grouping variables (tweets, ChatGPT-generated narratives, and Java class-generated narratives), 11 PANAS traits, and 4 event types, combined to produce a total of 132 tests for significant differences. This includes 44 tests for Twitter versus ChatGPT narratives, 44 for Twitter versus the simulated narratives, and 44 for ChatGPT narratives versus the simulated narratives. As outlined in Section 2.1, Fisher's Exact test is applied in cases where 2x2 contingency tables contain any cells with less than 5 samples. Chi-square tests are applied in all other cases.

The statistical significance tests are applied using the mean sentiment scores for a PANAS trait across two groups. For each narrative, as well as for each tweet, the PANAS trait assessments are binary. A score of 1 is assigned for a trait if any corresponding words within the lexicon are contained within the text and a score of 0 is assigned if no terms are contained. The mean value for this assessment is, therefore, reflective of the prevalence of the corresponding sentiment within each data set.

After generating the significance tests, we cut 67 of the categories which showed values of 0.0 for their variances. All 67 cases involved the simulated Java class-generated narratives. Examination of the Java class narratives revealed that while the generated narratives were unique with respect to raw term usage, many narratives only differed in terms of the proper nouns utilized, such as the name(s) of included agent(s). As proper nouns are automatically filtered during the text analysis process, this left groups of terms for narratives that were non-unique. Therefore, variance values were then calculated as 0's due to a lack of sufficient variability within the individual terms assessed within each narrative. To avoid biasing the comparisons, we decided to cut these cases from the evaluation. However, the raw outcome data for each pair of grouping variables along with each event is provided in the Appendices.

This left 62 cases for examination with 30 cases providing evidence in support of rejecting the null hypothesis in favor of the alternative hypothesis and 32 cases providing evidence in support of failing to reject the null hypothesis. Table 4 provides the aggregate representation of the number of samples for each of the results categorizations.

Table 4. Number of cases for Rejecting or Failing to Reject the null hypothesis.

Grouping	Event	Number of Cases for Rejecting the Null Hypothesis	Number of Cases for Failing to Reject the Null Hypothesis
Twitter-ChatGPT	Birth	9	2
	Death	6	5
	Hired	5	6
	Fired	10	1
Twitter-Simulation	Birth	0^*	7
	Death	0^*	0^*
	Hired	0^*	0^*
	Fired	10	1
ChatGPT-Simulation	Birth	0^*	9
	Death	0^*	0^*
	Hired	0^*	1
	Fired	0^*	0^*

* Insufficient variation in frequencies of any PANAS trait.

We are particularly interested in the instances where no statistically significant evidence is found in support of rejecting the null hypotheses outline in Equation 1. These represent cases where the sentiment prevalence for a given PANAS trait was not discernibly different from the language used within real tweets and represent a step forward in the generation of realistic narratives from simulated agents. For the Twitter versus ChatGPT comparisons, ChatGPT generated narratives across 14 of the 44 categories (31.82%) that were not statistically significantly different from the sentiment prevalence within the tweets contained in our Twitter data set. For Birth events, this included *negative* and *hostile_irritable*. For Death events, this included *positive*, *interested_attentive_alert*, *excited_enthusiastic_inspired*, *proud_determined*, and *strong_active*. For Hired events, this included *excited_enthusiastic_inspired*, *proud_determined*, *distressed_upset*, *guilty_ashamed*, and *afraid_scared*. For Fired events, only one category has a non-statistically significant outcome: *interested_attentive_alert*.

The mean prevalence of terms associated with PANAS traits statistically significantly differed for Birth narratives between the tweets and the ChatGPT narratives for nine traits as shown in Figure 5. This includes five positive and four negative traits. Twitter scored higher on average for three of the four negative traits while ChatGPT scored higher on all five positive traits including *positive*. The only negative trait that ChatGPT outscored the tweets on was *nervous_jittery*.

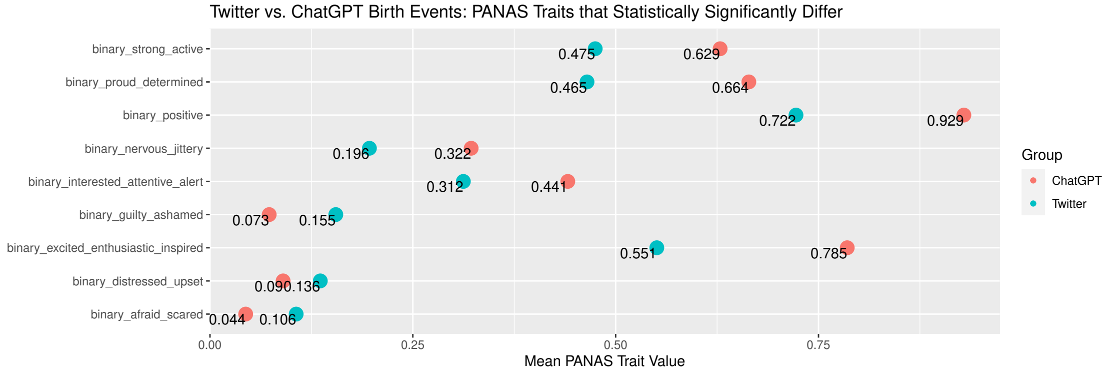

Figure 5. PANAS trait categories where statistically significant differences were not discernible between the sentiments of real tweets (mean values shown in blue) and the ChatGPT generated Birth narratives (mean values shown in red).

The mean prevalence of terms associated with PANAS traits statistically significantly differed for Birth narratives between the tweets and the ChatGPT narratives for six traits as shown in Figure 6. All six of these traits are negative PANAS traits including the *negative* trait itself. The largest difference between means in any of the Death events was observed for *distressed_upset*.

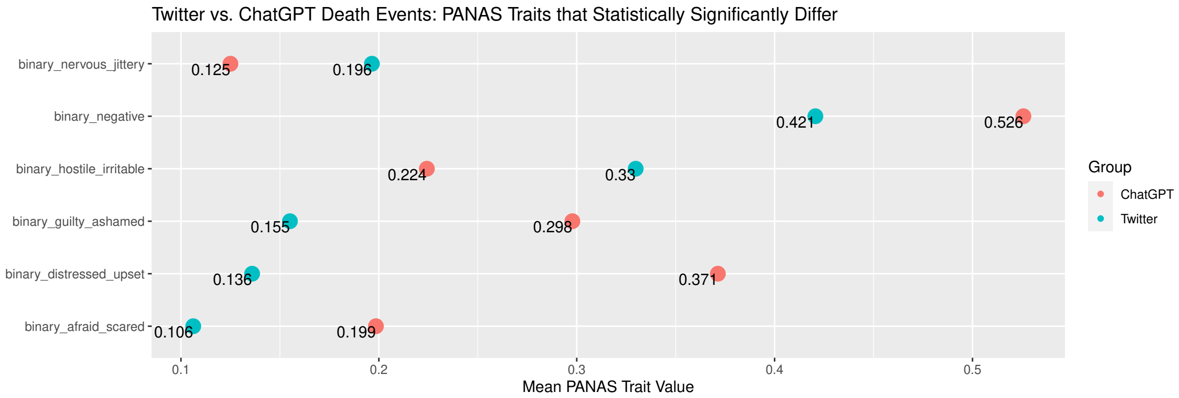

Figure 6. PANAS trait categories where statistically significant differences were not discernible between the sentiments of real tweets (mean values shown in blue) and the ChatGPT generated Death narratives (mean values shown in red).

The mean prevalence of terms associated with PANAS traits statistically significantly differed for Birth narratives between the tweets and the ChatGPT narratives for five traits as shown in Figure 7. The tweets scored higher prevalence of terms associated with both the base *positive* and *negative* traits. ChatGPT scored lower for all three of the presented negative traits.

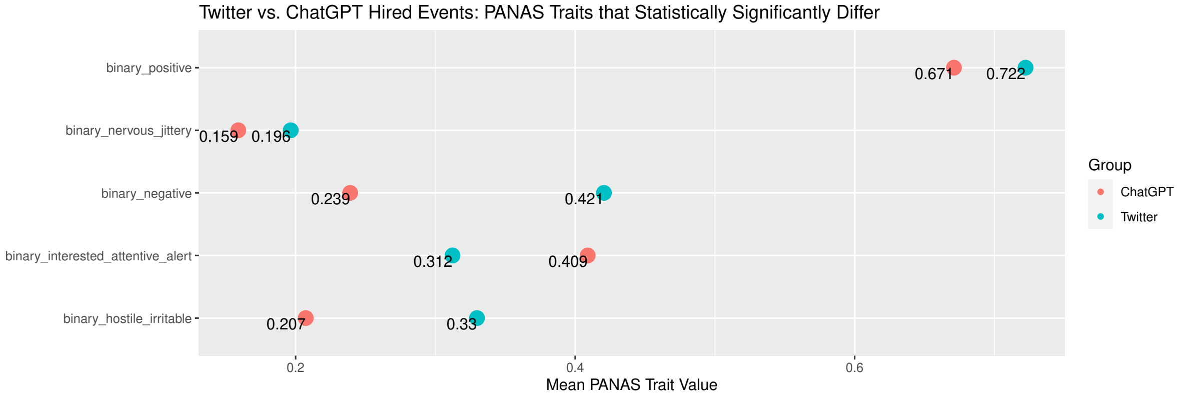

Figure 7. PANAS trait categories where statistically significant differences were not discernible between the sentiments of real tweets (mean values shown in blue) and the ChatGPT generated Hired narratives (mean values shown in red).

The mean prevalence of terms associated with PANAS traits statistically significantly differed for Birth narratives between the tweets and the ChatGPT narratives for 10 traits as shown in Figure 8. The Fired event trait assessments showed larger gaps between the means than the other three event types. For Fired events, the tweets showed much higher mean prevalence for terms associated with positive PANAS traits than ChatGPT. Likewise, ChatGPT generated narratives scored higher prevalence for all of the negative traits.

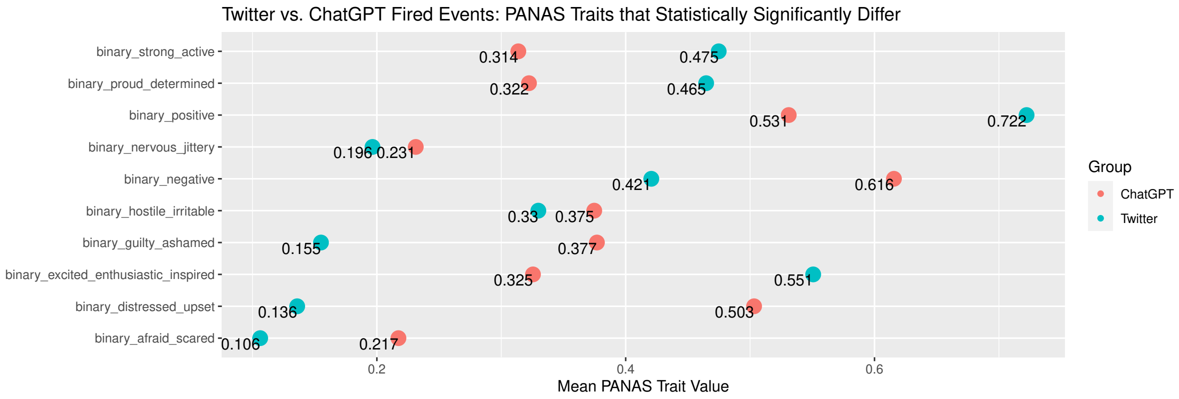

Figure 8. PANAS trait categories where statistically significant differences were not discernible between the sentiments of real tweets (mean values shown in blue) and the ChatGPT generated Fired narratives (mean values shown in red).

4. Discussion

The provided LLM Narrative Prompt Structure provides a consistent, transparent, and reproducible process for requesting narratives generated by LLMs. This structure assumes that an event type is known, that there is a relationship between the desired narrator of the event and the subject of the event (even if this just involves specifying that the narrator is describing an event that happened to him/her/itself), and that some amount of information is known about the event with respect to the intended subject of the narrative. Our results support that the current form of the LLM Narrative Prompt Structure can produce realistic narratives based on sentiment term prevalence comparisons to real tweets based on the utilized PANAS trait lexicon. However, additional work is needed to refine the structure for additional LLMs and for specific contexts and event types. Overall, ChatGPT successfully generated narratives based on simulated agents' information that was statistically indiscernible from real tweets sentiment prevalence for 14 out of 44 traits across 4 event types. This structure can enhance the ability to connect model users with the contextual plights, wishes, desires, injustices, and happiness of simulated environments by facilitating a connection between the simulated agents' perspectives and the perspectives of the model users, such as policy makers, researchers, decision makers, etc. This structure can also aid ABM researchers to filter through the large volumes of agent information that accumulates within simulation runs and distill the information into a more easily digestible narrative to communicate interactions, key events, communication, and evolutions of agents over time.

A total of 14 of 44 PANAS traits were identified where statistically significant differences in PANAS sentiment prevalence between ChatGPT narratives and real tweets were not discernible. Of the 30 traits where statistically significant differences did exist, trends were tending towards higher prevalence of negative traits within the tweets and higher prevalence of positive traits within the ChatGPT generated narratives. As target sentiment level is a field within the LLM Narrative Prompt Structure (field 11), additional work needs to go into calibrating the sentiment value for a given context or event type. Whether this tendency holds for other LLMs is not something that can be generalized from this work and research into the tendencies of other LLMs requires separate evaluation. In the following subsections, we discuss many of the critical lessons learned that we gained throughout this research endeavour, we discuss limitations of our study, and we outline avenues for future work.

4.1. Lessons Learned

1. **Using the ChatGPT API for generating multiple, independent narratives.** We discovered the ChatGPT API is not well-suited to generating multiple instances of a requested narrative in one response. There is a strong tendency to narrate a continuous, temporally-advancing story instead of a set of independent narratives describing a single event. Using the `n` parameter in the Python

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2026.v1

API `ChatCompletion` function call appears to remedy this behavior, as ChatGPT generates a set of `n` independent responses, that is, they are not connected. ChatGPT appears to not retain knowledge of narratives 1 through i when generating narrative $i + 1$, when the `n` parameter is used.

2. **Balancing creativity with correctness.** The level of stochasticity that ChatGPT employs for choosing the next token during text completion is moderated by the `temperature` parameter. A zero temperature outputs identical responses for repeated identical inputs; increasing the temperature increases the set of next available tokens in the completion, effectively increasing the response space and allowing for greater variation among the responses. However, this increased creative capacity can come at the cost of correctness, if the temperature is too high and ChatGPT selects inappropriate tokens. However, even with a temperature of zero used for the API, ChatGPT still sometimes produced categorically incorrect responses, e.g. narratives about car fires and house fires, when the prompt was to generate narratives about being fired from a job. A `temperature` of zero was used in this work, to attempt to limit incorrect narratives, and specifically to address ChatGPT's tendency toward "storytelling", instead of generating multiple independent narratives in one response. This ultimately was not very effective, as noted in the previous lesson learned. Conversely, when using the `n` parameter to generate multiple, different narratives, the `temperature` value must not be zero, else each response will be identical.
3. **ChatGPT API time-out errors.** The API fails frequently due to request time-out errors, so the experimental setup should account for this and be able to resume efficiently after an error. For this work, a Python script reads prompt files from a directory, and moves them to another directory after a successful response is received: if the script is restarted due to a time-out error, no prompts are lost or repeated.

4.2. Limitations

It is important to note some of the limitations of our work. The demographics, geographical location, and even previous daily activities of the authors of text-based data can result in substantially different word choices even when discussing the same subject matter. While it is beyond the scope of this work to control for these biases it is important to note that they exist and could influence the analysis would inform the paper. Geographic bias that influences sentiment analysis of text-based social media messages has been highlighted in Mitchell *et al.* and Gore *et al.*. Padilla *et al.* show that the sentiment of social media about a given subject matter can be biased by the time of day the message is authored and if the author is a resident or visitor in the city in which the message is composed. Additionally, the reliability and validity of targeted narrative messaging, such as the narrator-subject relationship utilized in our prompt structure, needs further evaluation with specific attention to the context of the simulation setting to gauge how well the message serves as a true representation of the original event.

1. **Problem type.** For this work, ChatGPT was not required to solve complex problems or rely heavily on factual information from training data. All the required factual information, including narrator and subject characteristics, was provided in the prompt. ChatGPT appears to be well-suited to this type of creative task, the output of which is technically correct as long as the instructions and constraints in the input prompt are observed. Narratives can easily be validated manually. This differs from other types of tasks, e.g. asking ChatGPT to solve a mathematical problem or to diagnose a medical condition [13], which requires knowledge of and "understanding" of much more complex background information, which is not included in the prompt. These problem types are much less subjective and not as easily validated. Further, ChatGPT currently cannot accurately provide sources, for validating the response information. In this case, the human reader has to determine if the response is legitimate or if ChatGPT has "hallucinated" some trustworthy-sounding but incorrect response, without the benefit of reliable references [12].

2. **Use case.** For this work, narratives were not actually tweeted or broadcast in any way, but were used solely for analysis. Incorrect narratives were identified manually and incurred no negative consequences. For use cases in which responses are not or cannot be validated by a human before utilization, there is a risk of dissemination of erroneous information. Numerous correct prior responses do not guarantee an incorrect response cannot happen in the future: in other words, there is no way to bound or know the response space [12]. As noted in the second lesson learned, even with minimal stochasticity, ChatGPT generated completely incorrect narratives about car fires and house fires, which could not have been predicted by the hundreds of other responses for that event type that did not do this.
3. **ChatGPT API response speed.** As noted in the third lesson learned, the Python API regularly failed due to time-out errors, so this currently might not be an appropriate tool for situations with strict time constraints.
4. **Token volume in real time / quicker than real time.** The ability to generalize our approach for real-time applications of ChatGPT for narrative generation is limited based on the token limit of the API. The current ChatGPT API version has a rate limit of 3,500 requests per minute and 90,000 tokens per minute [67].
5. **Domain expertise.** The creation of LLMs are based on broad ranges volumes of reference literature. It is important to determine that generated results are in line with the domain expertise of the targetted problem or system [68]. This article does not attempt to refine the learning base of the LLM for the narration of key events, as the broad range of potential response types for individuals was desireable as the starting point for this effort. However, future avenues of research include assessing the validity of narrative within their respective domains, such as for births and death events, and tailored within a larger context, such as refugee camps, natural distaster response, etc.
6. **Underspecification hinders narrative generation.** Greater levels of specificity in the prompting of desired narratives from ChatGPT has been beneficial in reducing the number of iterations for generating and assessing the correctness of the narratives. Similar to ML pipeline problems with underspecification, where underspecification in training leads to problems in reliability and validity [69], underspecification of narrative requests from ChatGPT led to many more erroneous responses and the expansion of the structure provided in this article.
7. **Tweet comparison sets.** Tweets were not categorized per event type like the generated narratives. The tweet set is utilized assuming that it represents a general sample of the population. As such, the generalizibility of the sentiment findings should not be extended to other sample populations without proper supporting justifications about the reasonableness of the extension.

4.3. Future Work

Future work includes developing processes for verifying that the narratives generated by the LLM are correct based on the instructions provided in a more formal, consistent, and reproducible fashion. Similarly, a validation process is needed to assess that the narratives are valid given the perspective of an agent combined with the contextual information of the environment. A pipeline for transparently assessing whether each narrative within a set of narratives is correct or incorrect that can be extended to the automated assessment of the set of narratives as a whole would be beneficial to the LLM community. We intend to explore this process further as a follow-up article to this research.

Additional research is needed to expand the comparison of event-specific narratives with correspondingly categorized tweet data sets. This should help in yielding realism and contextual appropriateness for simulating specific types of narratives. This can also be used as an avenue for developing contextually relevant perspectives from communities of interest, such as for displaced and marginalized communities. Research in this realm can also include determining how to better bridge the perspective of the simulated agent with the perspectives of a potential policy and decision makers. Projects exploring socially relevant outcomes should engage stakeholders and domain experts from the start to end of the design process. Reports on the biases and limitations of applying results

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2026.v1

should be produced to mitigate the potential for misuse and inappropriate application of results [68]. To facilitate trustworthiness in the responses generated by an AI supported system, the system should provide fair, accurate, independent, honest, and reliable outcomes, among other values [70].

Capturing and assessing temporal components of narratives are another avenue for continued research. Narrative generation can be expanded from single events into a series of relevant and/or related events. This could provide a more relevant narrative based on an agent's knowledge or history of past events, i.e., current age, relationships, emotional state, etc. Combining pertinent history information with the social norms or cultural values of the society being modeled would help to connect with moral, ethical, and social underpinnings of the modeled system.

Further experiments with the design of the structured prompt for narrative generation may yield insight into the capabilities and limitations of LLMs, like ChatGPT, to process structured prompts in varying levels of verbosity. That is, on one end of the spectrum, prompts are minimalistic and utilitarian, and on the other end, prompts are verbose and similar to normal written or spoken language. These results could have implications beyond the domain of agent-based simulation, for example in medicine, where it may become commonplace to submit diagnostic or treatment queries to an LLM, preferably in a concise, feature-based format [17]. Additionally, we want to expand the experimental LLMs in our study to include OpenAI's GPT-4, Meta's LLaMa or LLama 2, and Google's Bard.

Author Contributions: Conceptualization, C.J.L., E.J., V.Z., K.O., E.F. and R.G.; methodology, C.J.L, E.J. and R.G.; software, C.J.L, R.G. and E.J.; validation, C.J.L, R.G. and E.J.; formal analysis, C.J.L, R.G. and E.J.; investigation, C.J.L, R.G. and E.J.; data curation, C.J.L, R.G. and E.J.; writing—original draft preparation, C.J.L, E.J. and R.G.; writing—review and editing, C.J.L., E.J., V.Z., K.O., E.F. and R.G.; visualization, C.J.L, E.J. and R.G.; funding acquisition, R.G. and C.J.L All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported, in part, by Old Dominion University through project 300916-010, Effective & Individualized Risk Communication [63].

Institutional Review Board Statement: Not applicable as this research did not involve the use of human subjects. However, the previously constructed Twitter data set was approved as EXEMPT FROM IRB REVIEW by the Old Dominion University Institutional Review Board on May 13, 2022 per Exemption category #2 from federal regulations [57].

Informed Consent Statement: Not applicable.

Data Availability Statement: All the data, code, and generated narratives utilized in this article are freely accessible in on online repository.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ABM	Agent Based Model
API	Application Programming Interface
GPT	Generative Pretrained Transformer
LLM	Large Language Model
PANAS	Positive and Negative Affect Schedule

Appendix A. PANAS Lexicon

Appendix A.1.

We utilize a PANAS scale [58] lexicon formed from a combination of the National Research Council (NRC), General Inquirer, and LWIC lexicons [59–61]. This lexicon provides the following 20 traits that can be assessed in text: *interested*, *alert*, *attentive*, *excited*, *enthusiastic*, *inspired*, *proud*, *determined*, *active*, *strong*, *depressed*, *upset*, *guilty*, *ashamed*, *hostile*, *irritable*, *nervous*, *jittery*, *afraid*, and *scared*. Table A1 provides the full hierarchical representation of the lexicon and denotes which categories are subgrouped under positive and negative.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2026.v1

Table A1. PANAS Lexicon.

PANAS_trait	**PANAS_group**	**PANAS_polarity**	**lexicon**	**lexicon_subgroup**	**lexicon_subgroup_description**	**comments**
interested	interested | alert | attentive	positive	NRC	anticipation	839 words associated with anticipation through MTurk crowdsourcing	only include positive words
alert	interested | alert | attentive	positive	General Inquirer	Perceiv	167 words associated with perception and perceiving	only include positive words
attentive	interested | alert | attentive	positive	NRC	anticipation	839 words associated with anticipation through MTurk crowdsourcing	only include positive words
excited	excited | enthusiastic | inspired	positive	General Inquirer	Arousal	67 words indicating excitation; aside from pleasures or pains; but including arousal of affiliation and hostility	only include positive words
enthusiastic	excited | enthusiastic | inspired	positive	General Inquirer	Arousal	67 words indicating excitation; aside from pleasures or pains; but including arousal of affiliation and hostility	only include positive words
inspired	excited | enthusiastic | inspired	positive	NRC	joy	689 words associated with joy through MTurk crowdsourcing	only include positive words
proud	proud | determined	positive	NRC	trust	1231 words associated with trust through MTurk crowdsourcing	only include positive words
determined	proud | determined	positive	General Inquirer	Pleasur	168 words indicating the enjoyment of a feeling. Including words indicating confidence; interest and commitment	only include positive words
active	strong | active	positive	General Inquirer	Active	1902 words implying strength	only include positive words
strong	strong | active	positive	General Inquirer	Strong	2045 words implying an active orientation	only include positive words
distressed	distressed | upset	negative	General Inquirer	Pain	254 words indicating suffering; lack of confidence; or commitment	only include negative words
upset	distressed | upset	negative	NRC	sadness	1191 words associated with sadness through MTurk crowdsourcing	only include negative words
guilty	guilty | ashamed	negative	General Inquirer	Vice	685 words indicating an assessment of moral disapproval or misfortune	only include negative words
ashamed	guilty | ashamed	negative	NRC	disgust	1058 words associated with disgust through MTurk crowdsourcing	only include negative words
hostile	hostile | irritable	negative	General Inquirer	Arousal	67 words indicating excitation; aside from pleasures or pains; but including arousal of affiliation and hostility	only include negative words
irritable	hostile | irritable	negative	NRC	anger	1247 words associated with anger through MTurk crowdsourcing	only include negative words
nervous	nervous | jiittery	negative	LWIC	anxiety	196 words associated with anxiety in the LWIC 2015 dictionary	only include negative words
jittery	nervous | jiittery	negative	NRC	anticipation	839 words associated with anticipation through MTurk crowdsourcing	only include negative words
afraid	afraid | scared	negative	NRC	fear	1476 words associated with fear through MTurk crowdsourcing	only include negative words
scared	afraid | scared	negative	NRC	surprise	534 words associated with surprise through MTurk crowdsourcing	only include negative words

Appendix B. Tests for Statistically Significant Differences per PANAS Category

The following 12 sub-sections within this section of the appendix provide the data for all of the statistically significant tests conducted between the Simulation narratives, ChatGPT narratives, and tweets.

Appendix B.1. Birth Narrative Comparison - Twitter versus ChatGPT

Table A2 provides the results of the tests for significant differences between Twitter (sample 1) and ChatGPT (sample 2) for Birth event narratives. For rows displaying a value for Chi Square, the Chi Square test was applied to assess significance; otherwise, the Fisher's Exact Test was applied. A Null Hypothesis is provided within the context of the tested components within the "Null Hypothesis Description" column and the resulting interpretation of the test based on a *P value* assessment at an alpha level of 0.05 is provided in the "Interpretation" column.

Table A2. Tests for Statistically Significant Differences in PANAS trait prevalence for Death event narratives between Twitter and ChatGPT.

PANAS_Group	P Value	Chi Square	Sample 1 Mean	Sample 1 Variance	Sample 2 Mean	Sample 2 Variance	Null Hypothesis Description	Interpretation
binary_positive	1.08e-36	160.099506102967	0.722251725969198	0.20071076125499	0.929184549356223	0.0658712999543617	There is no difference in association between Twitter and ChatGPT for event type Birth in relation to PANAS group: binary_positive.	Reject the null hypothesis.
binary_negative	4.71e-01	0.519321805771216	0.420605416887945	0.243825988217188	0.405579399141631	0.24134370260415	There is no difference in association between Twitter and ChatGPT for event type Birth in relation to PANAS group: binary_negative.	Reject the null hypothesis.
binary_interested_attentive_alert	2.47e-11	44.5577138071836	0.312267657992565	0.214870678586808	0.440987124463519	0.246782268362507	There is no difference in association between Twitter and ChatGPT for event type Birth in relation to PANAS group: binary_interested_attentive_alert.	Reject the null hypothesis.
binary_excited_enthusiastic_inspired	1.27e-33	146.050770865262	0.550716941051514	0.247559262555569	0.785407725321888	0.168723464086335	There is no difference in association between Twitter and ChatGPT for event type Birth in relation to PANAS group: binary_excited_enthusiastic_inspired.	Reject the null hypothesis.
binary_proud_determined	2.78e-23	98.806762544465	0.464684014869888	0.248884955892055	0.664163090128755	0.223290061450376	There is no difference in association between Twitter and ChatGPT for event type Birth in relation to PANAS group: binary_proud_determined.	Reject the null hypothesis.
binary_strong_active	1.87e-14	58.6678829706207	0.474774296335635	0.249496163164688	0.628755364806867	0.233672777898148	There is no difference in association between Twitter and ChatGPT for event type Birth in relation to PANAS group: binary_strong_active.	Reject the null hypothesis.
binary_distressed_upset	5.61e-04	11.9000737571625	0.13595326606479	0.117532393138902	0.0901287553648069	0.0820936461324986	There is no difference in association between Twitter and ChatGPT for event type Birth in relation to PANAS group: binary_distressed_upset.	Reject the null hypothesis.
binary_guilty_ashamed	1.21e-09	36.9539156769644	0.155071694105151	0.131094083592612	0.0729613733905579	0.0677106623087455	There is no difference in association between Twitter and ChatGPT for event type Birth in relation to PANAS group: binary_guilty_ashamed.	Reject the null hypothesis.
binary_hostile_irritable	7.38e-01	0.112077435897854	0.329792883696229	0.221146981522126	0.336909871244635	0.223641568667223	There is no difference in association between Twitter and ChatGPT for event type Birth in relation to PANAS group: binary_hostile_irritable.	Fail to reject the null hypothesis.
binary_nervous_jittery	2.78e-13	53.3603859555868	0.196494954859267	0.15796857954414	0.321888412017167	0.218510715783942	There is no difference in association between Twitter and ChatGPT for event type Birth in relation to PANAS group: binary_nervous_jittery.	Reject the null hypothesis.
binary_afraid_scared	4.23e-08	30.0427370366321	0.106213489113117	0.0949826260241108	0.0439914163090129	0.0421013447167889	There is no difference in association between Twitter and ChatGPT for event type Birth in relation to PANAS group: binary_afraid_scared.	Reject the null hypothesis.

Appendix B.2. Death Narrative Comparison - Twitter versus ChatGPT

Table A3 provides the results of the tests for significant differences between Twitter (sample 1) and ChatGPT (sample 2) for Death event narratives. For rows displaying a value for Chi Square, the Chi Square test was applied to assess significance; otherwise, the Fisher's Exact Test was applied. A Null Hypothesis is provided within the context of the tested components within the "Null Hypothesis Description" column and the resulting interpretation of the test based on a *P value* assessment at an alpha level of 0.05 is provided in the "Interpretation" column.

Table A3. Tests for Statistically Significant Differences in PANAS trait prevalence for Death event narratives between Twitter and ChatGPT.

PANAS_Group	P Value	Chi Square	Sample 1 Mean	Sample 1 Variance	Sample 2 Mean	Sample 2 Variance	Null Hypothesis Description	Interpretation
binary_positive	9.02e-01	0.0152569643464454	0.722251725969198	0.20071076125499	0.727941176470588	0.198773605383113	There is no difference in association between Twitter and ChatGPT for event type Death in relation to PANAS group: binary_positive.	Fail to reject the null hypothesis.
binary_negative	1.35e-03	10.2704541122394	0.420605416887945	0.243825988217188	0.525735294117647	0.250257759930541	There is no difference in association between Twitter and ChatGPT for event type Death in relation to PANAS group: binary_negative.	Reject the null hypothesis.
binary_interested_attentive_alert	5.11e-01	0.432584567290359	0.312267657992565	0.214870678586808	0.290441176470588	0.20684556110267	There is no difference in association between Twitter and ChatGPT for event type Death in relation to PANAS group: binary_interested_attentive_alert.	Fail to reject the null hypothesis.
binary_excited_enthusiastic_inspired	2.25e-01	1.47113496472166	0.550716941051514	0.247559262555569	0.591911764705882	0.242443564141524	There is no difference in association between Twitter and ChatGPT for event type Death in relation to PANAS group: binary_excited_enthusiastic_inspired.	Reject the null hypothesis.
binary_proud_determined	4.24e-01	0.638297879874122	0.464684014869888	0.248884955892055	0.492647058823529	0.250868243976557	There is no difference in association between Twitter and ChatGPT for event type Death in relation to PANAS group: binary_proud_determined.	Reject the null hypothesis.
binary_strong_active	5.48e-01	0.360204684907635	0.474774296335635	0.249496163164688	0.496323529411765	0.250908942912959	There is no difference in association between Twitter and ChatGPT for event type Death in relation to PANAS group: binary_strong_active.	Fail to reject the null hypothesis.
binary_distressed_upset	3.93e-22	93.5649780798951	0.13595326606479	0.117532393138902	0.371323529411765	0.234303776861298	There is no difference in association between Twitter and ChatGPT for event type Death in relation to PANAS group: binary_distressed_upset.	Reject the null hypothesis.
binary_guilty_ashamed	1.00e-08	32.8349109954337	0.155071694105151	0.131094083592612	0.297794117647059	0.209884415020621	There is no difference in association between Twitter and ChatGPT for event type Death in relation to PANAS group: binary_guilty_ashamed.	Reject the null hypothesis.
binary_hostile_irritable	6.08e-04	11.7523704114126	0.329792883696229	0.221146981522126	0.224264705882353	0.174612003472976	There is no difference in association between Twitter and ChatGPT for event type Death in relation to PANAS group: binary_hostile_irritable.	Reject the null hypothesis.
binary_nervous_jittery	6.13e-03	7.51289996116507	0.196494954859267	0.15796857954414	0.125	0.109778597785978	There is no difference in association between Twitter and ChatGPT for event type Death in relation to PANAS group: binary_nervous_jittery.	Reject the null hypothesis.
binary_afraid_scared	1.61e-05	18.6029049368609	0.106213489113117	0.0949826260241108	0.198529411764706	0.159702626438029	There is no difference in association between Twitter and ChatGPT for event type Death in relation to PANAS group: binary_afraid_scared.	Reject the null hypothesis.

Appendix B.3. Hired Narrative Comparison - Twitter versus ChatGPT

Table A4 provides the results of the tests for significant differences between Twitter (sample 1) and ChatGPT (sample 2) for Hired event narratives. For rows displaying a value for Chi Square, the Chi Square test was applied to assess significance; otherwise, the Fisher's Exact Test was applied. A Null Hypothesis is provided within the context of the tested components within the "Null Hypothesis Description" column and the resulting interpretation of the test based on a *P value* assessment at an alpha level of 0.05 is provided in the "Interpretation" column.

Table A4. Tests for Statistically Significant Differences in PANAS trait prevalence for Hired event narratives between Twitter and ChatGPT.

PANAS_Group	P Value	Chi Square	Sample 1 Mean	Sample 1 Variance	Sample 2 Mean	Sample 2 Variance	Null Hypothesis Description	Interpretation
binary_positive	6.14e-03	7.50806773042741	0.722251725969198	0.20071076125499	0.671052631578947	0.220983303483737	There is no difference in association between Twitter and ChatGPT for event type Hired in relation to PANAS group: binary_positive.	Reject the null hypothesis.
binary_negative	1.09e-20	86.9872222350499	0.420605416887945	0.243825988217188	0.239035087719298	0.182096982302078	There is no difference in association between Twitter and ChatGPT for event type Hired in relation to PANAS group: binary_negative.	Reject the null hypothesis.
binary_interested_attentive_alert	5.58e-07	25.0512925345889	0.312267657992565	0.214870678586808	0.408991228070175	0.241982735378512	There is no difference in association between Twitter and ChatGPT for event type Hired in relation to PANAS group: binary_interested_attentive_alert.	Reject the null hypothesis.
binary_excited_enthusiastic_inspired	7.63e-02	3.14223253585284	0.550716941051514	0.247559262555569	0.514254385964912	0.25007101315308	There is no difference in association between Twitter and ChatGPT for event type Hired in relation to PANAS group: binary_excited_enthusiastic_inspired.	Reject the null hypothesis.
binary_proud_determined	5.36e-02	3.72415611415799	0.464684014869888	0.248884955892055	0.504385964912281	0.250255165905983	There is no difference in association between Twitter and ChatGPT for event type Hired in relation to PANAS group: binary_proud_determined.	Reject the null hypothesis.
binary_strong_active	5.05e-01	0.445143713062267	0.474774296335635	0.249496163164688	0.460526315789474	0.24871454156797	There is no difference in association between Twitter and ChatGPT for event type Hired in relation to PANAS group: binary_strong_active.	Fail to reject the null hypothesis.
binary_distressed_upset	8.58e-01	0.0321455897504964	0.13595326606479	0.117532393138902	0.139254385964912	0.119994174514222	There is no difference in association between Twitter and ChatGPT for event type Hired in relation to PANAS group: binary_distressed_upset.	Fail to reject the null hypothesis.
binary_guilty_ashamed	6.86e-02	3.31684350864536	0.155071694105151	0.131094083592612	0.128289473684211	0.111954041250217	There is no difference in association between Twitter and ChatGPT for event type Hired in relation to PANAS group: binary_guilty_ashamed.	Reject the null hypothesis.
binary_hostile_irritable	2.90e-11	44.2449976096182	0.329792883696229	0.221146981522126	0.207236842105263	0.164470073372234	There is no difference in association between Twitter and ChatGPT for event type Hired in relation to PANAS group: binary_hostile_irritable.	Reject the null hypothesis.
binary_nervous_jittery	1.90e-02	5.50274102380948	0.196494954859267	0.15796857954414	0.158991228070175	0.133859793556339	There is no difference in association between Twitter and ChatGPT for event type Hired in relation to PANAS group: binary_nervous_jittery.	Reject the null hypothesis.
binary_afraid_scared	4.65e-01	0.533497011846064	0.106213489113117	0.0949826260241108	0.116228070175439	0.10283186011131	There is no difference in association between Twitter and ChatGPT for event type Hired in relation to PANAS group: binary_afraid_scared.	Reject the null hypothesis.

Appendix B.4. Fired Narrative Comparison - Twitter versus ChatGPT

Table A5 provides the results of the tests for significant differences between Twitter (sample 1) and ChatGPT (sample 2) for Fired event narratives. For rows displaying a value for Chi Square, the Chi Square test was applied to assess significance; otherwise, the Fisher's Exact Test was applied. A Null Hypothesis is provided within the context of the tested components within the "Null Hypothesis Description" column and the resulting interpretation of the test based on a *P value* assessment at an alpha level of 0.05 is provided in the "Interpretation" column.

Table A5. Tests for Statistically Significant Differences in PANAS trait prevalence for Fired event narratives between Twitter and ChatGPT.

PANAS_Group	P Value	Chi Square	Sample 1 Mean	Sample 1 Variance	Sample 2 Mean	Sample 2 Variance	Null Hypothesis Description	Interpretation
binary_positive	1.07e-23	100.698711966595	0.722251725969198	0.20071076125499	0.531049250535332	0.249302863595365	There is no difference in association between Twitter and ChatGPT for event type Fired in relation to PANAS group: binary_positive.	Reject the null hypothesis.
binary_negative	2.73e-22	94.2886931329851	0.420605416887945	0.243825988217188	0.615631691648822	0.236882933871305	There is no difference in association between Twitter and ChatGPT for event type Fired in relation to PANAS group: binary_negative.	Reject the null hypothesis.
binary_interested_attentive_alert	6.62e-01	0.191199239916657	0.312267657992565	0.214870678586808	0.321199143468951	0.218263941006768	There is no difference in association between Twitter and ChatGPT for event type Fired in relation to PANAS group: binary_interested_attentive_alert.	Fail to reject the null hypothesis.
binary_excited_enthusiastic_inspired	2.95e-29	126.080313021453	0.550716941051514	0.247559262555569	0.325481798715203	0.219778706527951	There is no difference in association between Twitter and ChatGPT for event type Fired in relation to PANAS group: binary_excited_enthusiastic_inspired.	Reject the null hypothesis.
binary_proud_determined	7.22e-13	51.4844982068129	0.464684014869888	0.248884955892055	0.322269807280514	0.218646075035976	There is no difference in association between Twitter and ChatGPT for event type Fired in relation to PANAS group: binary_proud_determined.	Reject the null hypothesis.
binary_strong_active	5.06e-16	65.7725100599171	0.474774296335635	0.249496163164688	0.313704496788009	0.21552474002263	There is no difference in association between Twitter and ChatGPT for event type Fired in relation to PANAS group: binary_strong_active.	Reject the null hypothesis.
binary_distressed_upset	2.58e-97	438.269725064649	0.13595326606479	0.117532393138902	0.50321199143469	0.250257624893565	There is no difference in association between Twitter and ChatGPT for event type Fired in relation to PANAS group: binary_distressed_upset.	Reject the null hypothesis.
binary_guilty_ashamed	1.71e-39	172.912873408962	0.155071694105151	0.131094083592612	0.376873661670236	0.235091608887542	There is no difference in association between Twitter and ChatGPT for event type Fired in relation to PANAS group: binary_guilty_ashamed.	Reject the null hypothesis.
binary_hostile_irritable	2.03e-02	5.3848042948889	0.329792883696229	0.221146981522126	0.374732334047109	0.234559145855854	There is no difference in association between Twitter and ChatGPT for event type Fired in relation to PANAS group: binary_hostile_irritable.	Reject the null hypothesis.
binary_nervous_jittery	3.65e-02	4.37235716734963	0.196494954859267	0.15796857954414	0.231263383297645	0.177971178143311	There is no difference in association between Twitter and ChatGPT for event type Fired in relation to PANAS group: binary_nervous_jittery.	Reject the null hypothesis.
binary_afraid_scared	3.45e-15	61.9914948424431	0.106213489113117	0.0949826260241108	0.217344753747323	0.170288333321858	There is no difference in association between Twitter and ChatGPT for event type Fired in relation to PANAS group: binary_afraid_scared.	Reject the null hypothesis.

Appendix B.5. Fired Narrative Comparison - Twitter versus Simulation

Table A6 provides the results of the tests for significant differences between Twitter (sample 1) and Simulation (sample 2) for Birth event narratives. For rows displaying a value for Chi Square, the Chi Square test was applied to assess significance; otherwise, the Fisher's Exact Test was applied. A Null Hypothesis is provided within the context of the tested components within the "Null Hypothesis Description" column and the resulting interpretation of the test based on a *P value* assessment at an alpha level of 0.05 is provided in the "Interpretation" column.

Table A6. Tests for Statistically Significant Differences in PANAS trait prevalence for Birth event narratives between Twitter and Simulation.

PANAS_Group	P Value	Chi Square	Sample 1 Mean	Sample 1 Variance	Sample 2 Mean	Sample 2 Variance	Null Hypothesis Description	Interpretation
binary_positive	2.01e-01	NA	0.722251725969198	0.20071076125499	1	0	There is no difference in association between Twitter and Simulation for event type Birth in relation to PANAS group: binary_positive.	Reject the null hypothesis.
binary_negative	4.62e-02	NA	0.420605416887945	0.243825988217188	0	0	There is no difference in association between Twitter and Simulation for event type Birth in relation to PANAS group: binary_negative.	Reject the null hypothesis.
binary_interested_attentive_alert	1.07e-01	NA	0.312267657992565	0.214870678586808	0	0	There is no difference in association between Twitter and Simulation for event type Birth in relation to PANAS group: binary_interested_attentive_alert.	Reject the null hypothesis.
binary_excited_enthusiastic_inspired	1.91e-02	NA	0.550716941051514	0.247559262555569	1	0	There is no difference in association between Twitter and Simulation for event type Birth in relation to PANAS group: binary_excited_enthusiastic_inspired.	Reject the null hypothesis.
binary_proud_determined	4.76e-03	NA	0.464684014869888	0.248884955892055	1	0	There is no difference in association between Twitter and Simulation for event type Birth in relation to PANAS group: binary_proud_determined.	Reject the null hypothesis.
binary_strong_active	5.53e-03	NA	0.474774296335635	0.249496163164688	1	0	There is no difference in association between Twitter and Simulation for event type Birth in relation to PANAS group: binary_strong_active.	Reject the null hypothesis.
binary_distressed_upset	6.03e-01	NA	0.13595326606479	0.117532393138902	0	0	There is no difference in association between Twitter and Simulation for event type Birth in relation to PANAS group: binary_distressed_upset.	Fail to reject the null hypothesis.
binary_guilty_ashamed	6.04e-01	NA	0.155071694105151	0.131094083592612	0	0	There is no difference in association between Twitter and Simulation for event type Birth in relation to PANAS group: binary_guilty_ashamed.	Fail to reject the null hypothesis.
binary_hostile_irritable	1.03e-01	NA	0.329792883696229	0.221146981522126	0	0	There is no difference in association between Twitter and Simulation for event type Birth in relation to PANAS group: binary_hostile_irritable.	Reject the null hypothesis.
binary_nervous_jittery	3.57e-01	NA	0.196494954859267	0.15796857954414	0	0	There is no difference in association between Twitter and Simulation for event type Birth in relation to PANAS group: binary_nervous_jittery.	Reject the null hypothesis.
binary_afraid_scared	1.00e+00	NA	0.106213489113117	0.0949826260241108	0	0	There is no difference in association between Twitter and Simulation for event type Birth in relation to PANAS group: binary_afraid_scared.	Fail to reject the null hypothesis.

Appendix B.6. Fired Narrative Comparison - Twitter versus Simulation

Table A7 provides the results of the tests for significant differences between Twitter (sample 1) and Simulation (sample 2) for Death event narratives. For rows displaying a value for Chi Square, the Chi Square test was applied to assess significance; otherwise, the Fisher's Exact Test was applied. A Null Hypothesis is provided within the context of the tested components within the "Null Hypothesis Description" column and the resulting interpretation of the test based on a *P value* assessment at an alpha level of 0.05 is provided in the "Interpretation" column.

Table A7. Tests for Statistically Significant Differences in PANAS trait prevalence for Death event narratives between Twitter and Simulation.

PANAS_Group	P Value	Chi Square	Sample 1 Mean	Sample 1 Variance	Sample 2 Mean	Sample 2 Variance	Null Hypothesis Description	Interpretation
binary_positive	2.70e-19	NA	0.722251725969198	0.20071076125499	0	0	There is no difference in association between Twitter and Simulation for event type Death in relation to PANAS group: binary_positive.	Reject the null hypothesis.
binary_negative	1.21e-08	NA	0.420605416887945	0.243825988217188	0	0	There is no difference in association between Twitter and Simulation for event type Death in relation to PANAS group: binary_negative.	Reject the null hypothesis.
binary_interested_attentive_alert	4.98e-06	NA	0.312267657992565	0.214870678586808	0	0	There is no difference in association between Twitter and Simulation for event type Death in relation to PANAS group: binary_interested_attentive_alert.	Reject the null hypothesis.
binary_excited_enthusiastic_inspired	2.24e-12	NA	0.550716941051514	0.247559262555569	0	0	There is no difference in association between Twitter and Simulation for event type Death in relation to PANAS group: binary_excited_enthusiastic_inspired.	Reject the null hypothesis.
binary_proud_determined	8.55e-10	NA	0.464684014869888	0.248884955892055	0	0	There is no difference in association between Twitter and Simulation for event type Death in relation to PANAS group: binary_proud_determined.	Reject the null hypothesis.
binary_strong_active	5.71e-10	NA	0.474774296335635	0.249496163164688	0	0	There is no difference in association between Twitter and Simulation for event type Death in relation to PANAS group: binary_strong_active.	Reject the null hypothesis.
binary_distressed_upset	1.78e-02	NA	0.13595326606479	0.117532393138902	0	0	There is no difference in association between Twitter and Simulation for event type Death in relation to PANAS group: binary_distressed_upset.	Reject the null hypothesis.
binary_guilty_ashamed	6.35e-03	NA	0.155071694105151	0.131094083592612	0	0	There is no difference in association between Twitter and Simulation for event type Death in relation to PANAS group: binary_guilty_ashamed.	Reject the null hypothesis.
binary_hostile_irritable	2.29e-06	NA	0.329792883696229	0.221146981522126	0	0	There is no difference in association between Twitter and Simulation for event type Death in relation to PANAS group: binary_hostile_irritable.	Reject the null hypothesis.
binary_nervous_jittery	1.38e-03	NA	0.196494954859267	0.15796857954414	0	0	There is no difference in association between Twitter and Simulation for event type Death in relation to PANAS group: binary_nervous_jittery.	Reject the null hypothesis.
binary_afraid_scared	4.31e-02	NA	0.106213489113117	0.0949826260241108	0	0	There is no difference in association between Twitter and Simulation for event type Death in relation to PANAS group: binary_afraid_scared.	Reject the null hypothesis.

Appendix B.7. Hired Narrative Comparison - Twitter versus Simulation

Table A8 provides the results of the tests for significant differences between Twitter (sample 1) and Simulation (sample 2) for Hired event narratives. For rows displaying a value for Chi Square, the Chi Square test was applied to assess significance; otherwise, the Fisher's Exact Test was applied. A

Null Hypothesis is provided within the context of the tested components within the "Null Hypothesis Description" column and the resulting interpretation of the test based on a *P value* assessment at an alpha level of 0.05 is provided in the "Interpretation" column.

Table A8. Tests for Statistically Significant Differences in PANAS trait prevalence for Hired event narratives between Twitter and Simulation.

PANAS_Group	P Value	Chi Square	Sample 1 Mean	Sample 1 Variance	Sample 2 Mean	Sample 2 Variance	Null Hypothesis Description	Interpretation
binary_positive	4.57e-16	NA	0.722251725969198	0.20071076125499	0	0	There is no difference in association between Twitter and Simulation for event type Hired in relation to PANAS group: binary_positive.	Reject the null hypothesis.
binary_negative	4.31e-07	NA	0.420605416887945	0.243825988217188	0	0	There is no difference in association between Twitter and Simulation for event type Hired in relation to PANAS group: binary_negative.	Reject the null hypothesis.
binary_interested_attentive_alert	4.06e-05	NA	0.312267657992565	0.214870678586808	0	0	There is no difference in association between Twitter and Simulation for event type Hired in relation to PANAS group: binary_interested_attentive_alert.	Reject the null hypothesis.
binary_excited_enthusiastic_inspired	2.42e-10	NA	0.550716941051514	0.247559262555569	0	0	There is no difference in association between Twitter and Simulation for event type Hired in relation to PANAS group: binary_excited_enthusiastic_inspired.	Reject the null hypothesis.
binary_proud_determined	3.91e-08	NA	0.464684014869888	0.248884955892055	0	0	There is no difference in association between Twitter and Simulation for event type Hired in relation to PANAS group: binary_proud_determined.	Reject the null hypothesis.
binary_strong_active	3.35e-08	NA	0.474774296335635	0.249496163164688	0	0	There is no difference in association between Twitter and Simulation for event type Hired in relation to PANAS group: binary_strong_active.	Reject the null hypothesis.
binary_distressed_upset	2.55e-02	NA	0.13595326606479	0.117532393138902	0	0	There is no difference in association between Twitter and Simulation for event type Hired in relation to PANAS group: binary_distressed_upset.	Reject the null hypothesis.
binary_guilty_ashamed	1.53e-02	NA	0.155071694105151	0.131094083592612	0	0	There is no difference in association between Twitter and Simulation for event type Hired in relation to PANAS group: binary_guilty_ashamed.	Reject the null hypothesis.
binary_hostile_irritable	1.93e-05	NA	0.329792883696229	0.221146981522126	0	0	There is no difference in association between Twitter and Simulation for event type Hired in relation to PANAS group: binary_hostile_irritable.	Reject the null hypothesis.
binary_nervous_jittery	3.27e-03	NA	0.196494954859267	0.15796857954414	0	0	There is no difference in association between Twitter and Simulation for event type Hired in relation to PANAS group: binary_nervous_jittery.	Reject the null hypothesis.
binary_afraid_scared	1.09e-01	NA	0.106213489113117	0.0949826260241108	0	0	There is no difference in association between Twitter and Simulation for event type Hired in relation to PANAS group: binary_afraid_scared.	Reject the null hypothesis.

Appendix B.8. Fired Narrative Comparison - Twitter versus Simulation

Table A9 provides the results of the tests for significant differences between Twitter (sample 1) and Simulation (sample 2) for Fired event narratives. For rows displaying a value for Chi Square, the Chi Square test was applied to assess significance; otherwise, the Fisher's Exact Test was applied. A Null Hypothesis is provided within the context of the tested components within the "Null Hypothesis Description" column and the resulting interpretation of the test based on a *P value* assessment at an alpha level of 0.05 is provided in the "Interpretation" column.

Table A9. Tests for Statistically Significant Differences in PANAS trait prevalence for Fired event narratives between Twitter and Simulation.

PANAS_Group	P Value	Chi Square	Sample 1 Mean	Sample 1 Variance	Sample 2 Mean	Sample 2 Variance	Null Hypothesis Description	Interpretation
binary_positive	1.47e-53	NA	0.722251725969198	0.20071076125499	0	0	There is no difference in association between Twitter and Simulation for event type Fired in relation to PANAS group: binary_positive.	Reject the null hypothesis.
binary_negative	1.44e-23	NA	0.420605416887945	0.243825988217188	0	0	There is no difference in association between Twitter and Simulation for event type Fired in relation to PANAS group: binary_negative.	Reject the null hypothesis.
binary_interested_attentive_alert	2.55e-16	NA	0.312267657992565	0.214870678586808	0	0	There is no difference in association between Twitter and Simulation for event type Fired in relation to PANAS group: binary_interested_attentive_alert.	Reject the null hypothesis.
binary_excited_enthusiastic_inspired	4.01e-34	NA	0.550716941051514	0.247559262555569	0	0	There is no difference in association between Twitter and Simulation for event type Fired in relation to PANAS group: binary_excited_enthusiastic_inspired.	Reject the null hypothesis.
binary_proud_determined	7.72e-27	NA	0.464684014869888	0.248884955892055	0	0	There is no difference in association between Twitter and Simulation for event type Fired in relation to PANAS group: binary_proud_determined.	Reject the null hypothesis.
binary_strong_active	1.46e-27	NA	0.474774296335635	0.249496163164688	0	0	There is no difference in association between Twitter and Simulation for event type Fired in relation to PANAS group: binary_strong_active.	Reject the null hypothesis.
binary_distressed_upset	1.56e-06	NA	0.13595326606479	0.117532393138902	0	0	There is no difference in association between Twitter and Simulation for event type Fired in relation to PANAS group: binary_distressed_upset.	Reject the null hypothesis.
binary_guilty_ashamed	1.22e-07	NA	0.155071694105151	0.131094083592612	0	0	There is no difference in association between Twitter and Simulation for event type Fired in relation to PANAS group: binary_guilty_ashamed.	Reject the null hypothesis.
binary_hostile_irritable	2.53e-17	NA	0.329792883696229	0.221146981522126	0	0	There is no difference in association between Twitter and Simulation for event type Fired in relation to PANAS group: binary_hostile_irritable.	Reject the null hypothesis.
binary_nervous_jittery	1.04e-09	NA	0.196494954859267	0.15796857954414	0	0	There is no difference in association between Twitter and Simulation for event type Fired in relation to PANAS group: binary_nervous_jittery.	Reject the null hypothesis.
binary_afraid_scared	4.40e-05	NA	0.106213489113117	0.0949826260241108	0	0	There is no difference in association between Twitter and Simulation for event type Fired in relation to PANAS group: binary_afraid_scared.	Reject the null hypothesis.

Appendix B.9. Fired Narrative Comparison - ChatGPT versus Simulation

Table A10 provides the results of the tests for significant differences between ChatGPT (sample 1) and Simulation (sample 2) for Birth event narratives. For rows displaying a value for Chi Square, the Chi Square test was applied to assess significance; otherwise, the Fisher's Exact Test was applied. A Null Hypothesis is provided within the context of the tested components within the "Null Hypothesis Description" column and the resulting interpretation of the test based on a *P value* assessment at an alpha level of 0.05 is provided in the "Interpretation" column.

Table A10. Tests for Statistically Significant Differences in PANAS trait prevalence for Birth event narratives between ChatGPT and Simulation.

PANAS_Group	P Value	Chi Square	Sample 1 Mean	Sample 1 Variance	Sample 2 Mean	Sample 2 Variance	Null Hypothesis Description	Interpretation
binary_positive	1.00e+00	NA	0.929184549356223	0.0658712999543617	1	0	There is no difference in association between ChatGPT and Simulation for event type Birth in relation to PANAS group: binary_positive.	Fail to reject the null hypothesis.
binary_negative	4.59e-02	NA	0.405579399141631	0.24134370260415	0	0	There is no difference in association between ChatGPT and Simulation for event type Birth in relation to PANAS group: binary_negative.	Reject the null hypothesis.
binary_interested_attentive_alert	2.05e-02	NA	0.440987124463519	0.246782268362507	0	0	There is no difference in association between ChatGPT and Simulation for event type Birth in relation to PANAS group: binary_interested_attentive_alert.	Reject the null hypothesis.
binary_excited_enthusiastic_inspired	3.56e-01	NA	0.785407725321888	0.168723464086335	1	0	There is no difference in association between ChatGPT and Simulation for event type Birth in relation to PANAS group: binary_excited_enthusiastic_inspired.	Reject the null hypothesis.
binary_proud_determined	1.03e-01	NA	0.664163090128755	0.223290061450376	1	0	There is no difference in association between ChatGPT and Simulation for event type Birth in relation to PANAS group: binary_proud_determined.	Reject the null hypothesis.
binary_strong_active	5.13e-02	NA	0.628755364806867	0.233672777898148	1	0	There is no difference in association between ChatGPT and Simulation for event type Birth in relation to PANAS group: binary_strong_active.	Reject the null hypothesis.
binary_distressed_upset	1.00e+00	NA	0.0901287553648069	0.0820936461324986	0	0	There is no difference in association between ChatGPT and Simulation for event type Birth in relation to PANAS group: binary_distressed_upset.	Fail to reject the null hypothesis.
binary_guilty_ashamed	1.00e+00	NA	0.0729613733905579	0.0677106623087455	0	0	There is no difference in association between ChatGPT and Simulation for event type Birth in relation to PANAS group: binary_guilty_ashamed.	Fail to reject the null hypothesis.
binary_hostile_irritable	1.02e-01	NA	0.336909871244635	0.223641568667223	0	0	There is no difference in association between ChatGPT and Simulation for event type Birth in relation to PANAS group: binary_hostile_irritable.	Reject the null hypothesis.
binary_nervous_jittery	1.04e-01	NA	0.321888412017167	0.218510715783942	0	0	There is no difference in association between ChatGPT and Simulation for event type Birth in relation to PANAS group: binary_nervous_jittery.	Reject the null hypothesis.
binary_afraid_scared	1.00e+00	NA	0.0439914163090129	0.0421013447167889	0	0	There is no difference in association between ChatGPT and Simulation for event type Birth in relation to PANAS group: binary_afraid_scared.	Fail to reject the null hypothesis.

Appendix B.10. Fired Narrative Comparison - ChatGPT versus Simulation

Table A11 provides the results of the tests for significant differences between ChatGPT (sample 1) and Simulation (sample 2) for Death event narratives. For rows displaying a value for Chi Square, the Chi Square test was applied to assess significance; otherwise, the Fisher's Exact Test was applied. A Null Hypothesis is provided within the context of the tested components within the "Null Hypothesis Description" column and the resulting interpretation of the test based on a *P value* assessment at an alpha level of 0.05 is provided in the "Interpretation" column.

Table A11. Tests for Statistically Significant Differences in PANAS trait prevalence for Death event narratives between ChatGPT and Simulation.

PANAS_Group	P Value	Chi Square	Sample 1 Mean	Sample 1 Variance	Sample 2 Mean	Sample 2 Variance	Null Hypothesis Description	Interpretation
binary_positive	8.22e-18	NA	0.727941176470588	0.198773605383113	0	0	There is no difference in association between ChatGPT and Simulation for event type Death in relation to PANAS group: binary_positive.	Reject the null hypothesis.
binary_negative	1.11e-10	NA	0.525735294117647	0.250257759930541	0	0	There is no difference in association between ChatGPT and Simulation for event type Death in relation to PANAS group: binary_negative.	Reject the null hypothesis.
binary_interested_attentive_alert	3.37e-05	NA	0.290441176470588	0.20684556110267	0	0	There is no difference in association between ChatGPT and Simulation for event type Death in relation to PANAS group: binary_interested_attentive_alert.	Reject the null hypothesis.
binary_excited_enthusiastic_inspired	9.37e-13	NA	0.591911764705882	0.242443564141524	0	0	There is no difference in association between ChatGPT and Simulation for event type Death in relation to PANAS group: binary_excited_enthusiastic_inspired.	Reject the null hypothesis.
binary_proud_determined	7.06e-10	NA	0.492647058823529	0.250868243976557	0	0	There is no difference in association between ChatGPT and Simulation for event type Death in relation to PANAS group: binary_proud_determined.	Reject the null hypothesis.
binary_strong_active	6.03e-10	NA	0.496323529411765	0.250908942912959	0	0	There is no difference in association between ChatGPT and Simulation for event type Death in relation to PANAS group: binary_strong_active.	Reject the null hypothesis.
binary_distressed_upset	7.81e-07	NA	0.371323529411765	0.234303776861298	0	0	There is no difference in association between ChatGPT and Simulation for event type Death in relation to PANAS group: binary_distressed_upset.	Reject the null hypothesis.
binary_guilty_ashamed	1.82e-05	NA	0.297794117647059	0.209884415020621	0	0	There is no difference in association between ChatGPT and Simulation for event type Death in relation to PANAS group: binary_guilty_ashamed.	Reject the null hypothesis.
binary_hostile_irritable	4.45e-04	NA	0.224264705882353	0.174612003472976	0	0	There is no difference in association between ChatGPT and Simulation for event type Death in relation to PANAS group: binary_hostile_irritable.	Reject the null hypothesis.
binary_nervous_jittery	2.05e-02	NA	0.125	0.109778597785978	0	0	There is no difference in association between ChatGPT and Simulation for event type Death in relation to PANAS group: binary_nervous_jittery.	Reject the null hypothesis.
binary_afraid_scared	1.40e-03	NA	0.198529411764706	0.159702626438029	0	0	There is no difference in association between ChatGPT and Simulation for event type Death in relation to PANAS group: binary_afraid_scared.	Reject the null hypothesis.

Appendix B.11. Hired Narrative Comparison - ChatGPT versus Simulation

Table A12 provides the results of the tests for significant differences between ChatGPT (sample 1) and Simulation (sample 2) for Hired event narratives. For rows displaying a value for Chi Square, the Chi Square test was applied to assess significance; otherwise, the Fisher's Exact Test was applied. A Null Hypothesis is provided within the context of the tested components within the "Null Hypothesis Description" column and the resulting interpretation of the test based on a *P value* assessment at an alpha level of 0.05 is provided in the "Interpretation" column.

Table A12. Tests for Statistically Significant Differences in PANAS trait prevalence for Hired event narratives between ChatGPT and Simulation.

PANAS_Group	P Value	Chi Square	Sample 1 Mean	Sample 1 Variance	Sample 2 Mean	Sample 2 Variance	Null Hypothesis Description	Interpretation
binary_positive	7.21e-14	NA	0.671052631578947	0.220983303483737	0	0	There is no difference in association between ChatGPT and Simulation for event type Hired in relation to PANAS group: binary_positive.	Reject the null hypothesis.
binary_negative	9.34e-04	NA	0.239035087719298	0.182096982302078	0	0	There is no difference in association between ChatGPT and Simulation for event type Hired in relation to PANAS group: binary_negative.	Reject the null hypothesis.
binary_interested_attentive_alert	1.03e-06	NA	0.408991228070175	0.241982735378512	0	0	There is no difference in association between ChatGPT and Simulation for event type Hired in relation to PANAS group: binary_interested_attentive_alert.	Reject the null hypothesis.
binary_excited_enthusiastic_inspired	4.93e-09	NA	0.514254385964912	0.25007101315308	0	0	There is no difference in association between ChatGPT and Simulation for event type Hired in relation to PANAS group: binary_excited_enthusiastic_inspired.	Reject the null hypothesis.
binary_proud_determined	5.84e-09	NA	0.504385964912281	0.250255165905983	0	0	There is no difference in association between ChatGPT and Simulation for event type Hired in relation to PANAS group: binary_proud_determined.	Reject the null hypothesis.
binary_strong_active	4.89e-08	NA	0.460526315789474	0.24871454156797	0	0	There is no difference in association between ChatGPT and Simulation for event type Hired in relation to PANAS group: binary_strong_active.	Reject the null hypothesis.
binary_distressed_upset	2.43e-02	NA	0.139254385964912	0.119994174514222	0	0	There is no difference in association between ChatGPT and Simulation for event type Hired in relation to PANAS group: binary_distressed_upset.	Reject the null hypothesis.
binary_guilty_ashamed	3.92e-02	NA	0.128289473684211	0.111954041250217	0	0	There is no difference in association between ChatGPT and Simulation for event type Hired in relation to PANAS group: binary_guilty_ashamed.	Reject the null hypothesis.
binary_hostile_irritable	2.96e-03	NA	0.207236842105263	0.164470073372234	0	0	There is no difference in association between ChatGPT and Simulation for event type Hired in relation to PANAS group: binary_hostile_irritable.	Reject the null hypothesis.
binary_nervous_jittery	1.46e-02	NA	0.158991228070175	0.133859793556339	0	0	There is no difference in association between ChatGPT and Simulation for event type Hired in relation to PANAS group: binary_nervous_jittery.	Reject the null hypothesis.
binary_afraid_scared	6.35e-02	NA	0.116228070175439	0.10283186011131	0	0	There is no difference in association between ChatGPT and Simulation for event type Hired in relation to PANAS group: binary_afraid_scared.	Reject the null hypothesis.

Appendix B.12. Fired Narrative Comparison - ChatGPT versus Simulation

Table A13 provides the results of the tests for significant differences between ChatGPT (sample 1) and Simulation (sample 2) for Fired event narratives. For rows displaying a value for Chi Square, the Chi Square test was applied to assess significance; otherwise, the Fisher's Exact Test was applied. A Null Hypothesis is provided within the context of the tested components within the "Null Hypothesis Description" column and the resulting interpretation of the test based on a *P value* assessment at an alpha level of 0.05 is provided in the "Interpretation" column.

Table A13. Tests for Statistically Significant Differences in PANAS trait prevalence for Fired event narratives between ChatGPT and Simulation.

PANAS_Group	P Value	Chi Square	Sample 1 Mean	Sample 1 Variance	Sample 2 Mean	Sample 2 Variance	Null Hypothesis Description	Interpretation
binary_positive	3.27e-31	NA	0.531049250535332	0.249302863595365	0	0	There is no difference in association between ChatGPT and Simulation for event type Fired in relation to PANAS group: binary_positive.	Reject the null hypothesis.
binary_negative	6.52e-39	NA	0.615631691648822	0.236882933871305	0	0	There is no difference in association between ChatGPT and Simulation for event type Fired in relation to PANAS group: binary_negative.	Reject the null hypothesis.
binary_interested_attentive_alert	2.55e-16	NA	0.321199143468951	0.218263941006768	0	0	There is no difference in association between ChatGPT and Simulation for event type Fired in relation to PANAS group: binary_interested_attentive_alert.	Reject the null hypothesis.
binary_excited_enthusiastic_inspired	1.24e-16	NA	0.325481798715203	0.219778706527951	0	0	There is no difference in association between ChatGPT and Simulation for event type Fired in relation to PANAS group: binary_excited_enthusiastic_inspired.	Reject the null hypothesis.
binary_proud_determined	2.56e-16	NA	0.322269807280514	0.218646075035976	0	0	There is no difference in association between ChatGPT and Simulation for event type Fired in relation to PANAS group: binary_proud_determined.	Reject the null hypothesis.
binary_strong_active	5.74e-16	NA	0.313704496788009	0.21552474002263	0	0	There is no difference in association between ChatGPT and Simulation for event type Fired in relation to PANAS group: binary_strong_active.	Reject the null hypothesis.
binary_distressed_upset	1.15e-28	NA	0.50321199143469	0.250257624893565	0	0	There is no difference in association between ChatGPT and Simulation for event type Fired in relation to PANAS group: binary_distressed_upset.	Reject the null hypothesis.
binary_guilty_ashamed	7.93e-20	NA	0.376873661670236	0.235091608887542	0	0	There is no difference in association between ChatGPT and Simulation for event type Fired in relation to PANAS group: binary_guilty_ashamed.	Reject the null hypothesis.
binary_hostile_irritable	9.29e-20	NA	0.374732334047109	0.234559145855854	0	0	There is no difference in association between ChatGPT and Simulation for event type Fired in relation to PANAS group: binary_hostile_irritable.	Reject the null hypothesis.
binary_nervous_jittery	2.37e-11	NA	0.231263383297645	0.177971178143311	0	0	There is no difference in association between ChatGPT and Simulation for event type Fired in relation to PANAS group: binary_nervous_jittery.	Reject the null hypothesis.
binary_afraid_scared	1.38e-10	NA	0.217344753747323	0.170288333321858	0	0	There is no difference in association between ChatGPT and Simulation for event type Fired in relation to PANAS group: binary_afraid_scared.	Reject the null hypothesis.

References

1. Goodman, A.; Morgan, R.; Kuehlke, R.; Kastor, S.; Fleming, K.; Boyd, J.; others. "We've been researched to death": Exploring the research experiences of urban Indigenous Peoples in Vancouver, Canada. *The International Indigenous Policy Journal* **2018**, *9*. doi:10.18584/iipj.2018.9.2.3.
2. Omata, N. 'Over-researched'and 'Under-researched'refugee groups: Exploring the phenomena, causes and consequences. *Journal of Human Rights Practice* **2020**, *12*, 681–695. doi:10.1093/jhuman/huaa049.
3. Reinhold, A.M.; Raile, E.D.; Izurieta, C.; McEvoy, J.; King, H.W.; Poole, G.C.; Ready, R.C.; Bergmann, N.T.; Shanahan, E.A. Persuasion with Precision: Using Natural Language Processing to Improve Instrument Fidelity for Risk Communication Experimental Treatments. *Journal of Mixed Methods Research* **2022**, p. 15586898221096934. doi:10.1177/15586898221096.
4. Shanahan, E.A.; Jones, M.D.; McBeth, M.K. How to conduct a Narrative Policy Framework study. *The Social Science Journal* **2018**, *55*, 332–345. doi:10.1016/j.soscij.2017.12.002.

5. Diallo, S.Y.; Lynch, C.J.; Rechowicz, K.J.; Zacharewicz, G. How to Create Empathy and Understanding: Narrative Analytics in Agent-Based Modeling. 2018 Winter Simulation Conference (WSC). IEEE, 2018, pp. 1286–1297. doi:10.1109/WSC.2018.8632267.
6. Shults, F.L.; Wildman, W.J.; Diallo, S.; Puga-Gonzalez, I.; Voas, D. The artificial society analytics platform. Advances in Social Simulation: Looking in the Mirror. Springer, 2020, pp. 411–426.
7. Alawida, M.; Mejri, S.; Mehmood, A.; Chikhaoui, B.; Isaac Abiodun, O. A Comprehensive Study of ChatGPT: Advancements, Limitations, and Ethical Considerations in Natural Language Processing and Cybersecurity. *Information* **2023**, *14*, 462. doi:10.3390/info14080462.
8. Nazary, F.; Deldjoo, Y.; Di Noia, T. ChatGPT-HealthPrompt. Harnessing the Power of XAI in Prompt-Based Healthcare Decision Support using ChatGPT. *arXiv preprint arXiv:2308.09731* **2023**. doi:10.48550/arXiv.2308.09731.
9. Stokel-Walker, C.; Van Noorden, R. The Promise and Peril of Generative AI. *Nature* **2023**, *614*, 214–216. doi:10.1038/d41586-023-00340-6.
10. Sallam, M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* **2023**, *11*, 887. doi:10.3390/healthcare11060887.
11. Van Dis, E.A.; Bollen, J.; Zuidema, W.; van Rooij, R.; Bockting, C.L. ChatGPT: five priorities for research. *Nature* **2023**, *614*, 224–226. doi:10.1038/d41586-023-00288-7.
12. Gilbert, S.; Harvey, H.; Melvin, T.; Vollebregt, E.; Wicks, P. Large Language Model AI Chatbots Require Approval as Medical Devices. *Nature Medicine* **2023**, pp. 1–3. doi:10.1038/s41591-023-02412-6.
13. Lee, P.; Bubeck, S.; Petro, J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *New England Journal of Medicine* **2023**, *388*, 1233–1239. doi:10.1056/NEJMsr2214184.
14. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large language models in medicine. *Nature Medicine* **2023**, pp. 1–11. doi:10.1038/s41591-023-02448-8.
15. Karabacak, M.; Margetis, K. Embracing Large Language Models for Medical Applications: Opportunities and Challenges. *Cureus* **2023**, *15*. doi:10.7759/cureus.39305.
16. Shah, N.H.; Entwistle, D.; Pfeffer, M.A. Creation and Adoption of Large Language Models in Medicine. *JAMA* **2023**. doi:10.1001/jama.2023.14217.
17. Reese, J.; Danis, D.; Caufield, J.H.; Casiraghi, E.; Valentini, G.; Mungall, C.J.; Robinson, P.N. On the limitations of large language models in clinical diagnosis. *medRxiv* **2023**, pp. 2023–07. doi:10.1101/2023.07.13.23292613.
18. Garg, R.K.; Urs, V.L.; Agrawal, A.A.; Chaudhary, S.K.; Paliwal, V.; Kar, S.K. Exploring the Role of Chat GPT in patient care (diagnosis and Treatment) and medical research: A Systematic Review. *medRxiv* **2023**, pp. 2023–06. doi:10.1101/2023.06.13.23291311.
19. Xue, V.W.; Lei, P.; Cho, W.C. The potential impact of ChatGPT in clinical and translational medicine. *Clinical and Translational Medicine* **2023**, *13*. doi:10.1002/ctm2.1216.
20. Hanna, J.J.; Wakene, A.D.; Lehmann, C.U.; Medford, R.J. Assessing Racial and Ethnic Bias in Text Generation for Healthcare-Related Tasks by ChatGPT. *medRxiv* **2023**, pp. 2023–08.
21. Tsai, M.L.; Ong, C.W.; Chen, C.L. Exploring the use of large language models (LLMs) in chemical engineering education: Building core course problem models with Chat-GPT. *Education for Chemical Engineers* **2023**, *44*, 71–95.
22. Qadir, J. Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. 2023 IEEE Global Engineering Education Conference (EDUCON). IEEE, 2023, pp. 1–9.
23. Pal, S.; Bhattacharya, M.; Lee, S.S.; Chakraborty, C. A Domain-Specific Next-Generation Large Language Model (LLM) or ChatGPT is Required for Biomedical Engineering and Research. *Annals of Biomedical Engineering* **2023**, pp. 1–4.
24. Thapa, S.; Adhikari, S. ChatGPT, Bard, and Large Language Models for Biomedical Research: Opportunities and Pitfalls. *Annals of Biomedical Engineering* **2023**, pp. 1–5.
25. Filippi, S. Measuring the Impact of ChatGPT on Fostering Concept Generation in Innovative Product Design. *Electronics* **2023**, *12*, 3535. doi:10.3390/electronics12163535.
26. Tikayat Ray, A.; Cole, B.F.; Pinon Fischer, O.J.; Bhat, A.P.; White, R.T.; Mavris, D.N. Agile Methodology for the Standardization of Engineering Requirements Using Large Language Models. *Systems* **2023**, *11*, 352. doi:10.3390/systems11070352.

27. Borji, A. A categorical archive of chatgpt failures. *arXiv preprint arXiv:2302.03494* **2023**. doi:10.48550/arXiv.2302.03494.
28. Makridakis, S.; Petropoulos, F.; Kang, Y. Large Language Models: Their Success and Impact. *Forecasting* **2023**, *5*, 536–549. doi:10.3390/forecast5030030.
29. Sham, A.H.; Aktas, K.; Rizhinashvili, D.; Kuklianov, D.; Alisinanoglu, F.; Ofodile, I.; Ozcinar, C.; Anbarjafari, G. Ethical AI in facial expression analysis: Racial bias. *Signal, Image and Video Processing* **2023**, *17*, 399–406.
30. Noor, P. Can we trust AI not to further embed racial bias and prejudice? *BMJ* **2020**, *368*.
31. Seyyed-Kalantari, L.; Zhang, H.; McDermott, M.B.; Chen, I.Y.; Ghassemi, M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine* **2021**, *27*, 2176–2182.
32. Guo, L.N.; Lee, M.S.; Kassamali, B.; Mita, C.; Nambudiri, V.E. Bias in, bias out: underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection—a scoping review. *Journal of the American Academy of Dermatology* **2022**, *87*, 157–159.
33. Kassem, M.A.; Hosny, K.M.; Damaševičius, R.; Eltoukhy, M.M. Machine learning and deep learning methods for skin lesion classification and diagnosis: a systematic review. *Diagnostics* **2021**, *11*, 1390.
34. Gross, N. What ChatGPT Tells Us about Gender: A Cautionary Tale about Performativity and Gender Biases in AI. *Social Sciences* **2023**, *12*, 435.
35. Hämäläinen, P.; Tavast, M.; Kunnari, A. Evaluating large language models in generating synthetic hci research data: a case study. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–19. doi:10.1145/3544548.3580688.
36. Sankararaman, K.A.; Wang, S.; Fang, H. Bayesformer: Transformer with uncertainty estimation. *arXiv preprint arXiv:2206.00826* **2022**.
37. Shelmanov, A.; Tsymbalov, E.; Puzyrev, D.; Fedyanin, K.; Panchenko, A.; Panov, M. How certain is your Transformer? Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021, pp. 1833–1840.
38. Vallès-Peris, N.; Domènech, M. Caring in the in-between: a proposal to introduce responsible AI and robotics to healthcare. *AI & SOCIETY* **2023**, *38*, 1685–1695. doi:10.1007/s00146-021-01330-w.
39. Upton, G.J. Fisher's exact test. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **1992**, *155*, 395–402. doi:10.2307/2982890.
40. Bower, K.M. When to use Fisher's exact test. American Society for Quality, Six Sigma Forum Magazine. American Society for Quality Milwaukee, WI, USA, 2003, Vol. 2, pp. 35–37.
41. Yi, D.; Yang, J.; Liu, J.; Liu, Y.; Zhang, J. Quantitative identification of urban functions with fishers' exact test and POI data applied in classifying urban districts: A case study within the sixth ring road in Beijing. *ISPRS International Journal of Geo-Information* **2019**, *8*, 555.
42. Pęksa, M.; Kamieniecki, A.; Gabrych, A.; Lew-Tusk, A.; Preis, K.; Świątkowska-Freund, M. Loss of E-cadherin staining continuity in the trophoblastic basal membrane correlates with increased resistance in uterine arteries and proteinuria in patients with pregnancy-induced hypertension. *Journal of Clinical Medicine* **2022**, *11*, 668.
43. Zeng, Y.; Xiong, Y.; Yang, C.; He, N.; He, J.; Luo, W.; Chen, Y.; Zeng, X.; Wu, Z. Investigation of Parasitic Infection in Crocodile Lizards (Shinisaurus crocodilurus) Using High-Throughput Sequencing. *Animals* **2022**, *12*, 2726.
44. Yokoyama, S.; Al Mahmuda, N.; Munesue, T.; Hayashi, K.; Yagi, K.; Yamagishi, M.; Higashida, H. Association study between the CD157/BST1 gene and autism spectrum disorders in a Japanese population. *Brain Sciences* **2015**, *5*, 188–200.
45. Miñana-Signes, V.; Monfort-Pañego, M.; Bosh-Bivià, A.H.; Noll, M. Prevalence of low back pain among primary school students from the city of Valencia (Spain). Healthcare. MDPI, 2021, Vol. 9, p. 270.
46. Aydın, Ö. Google Bard Generated Literature Review: Metaverse. *Available at SSRN* **2023**.
47. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; Lample, G. LLaMA: Open and Efficient Foundation Language Models, 2023, [arXiv:cs.CL/2302.13971].
48. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; others. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* **2023**. doi:10.48550/arXiv.2307.09288.

49. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
50. openAI. ChatGPT, August 2023 version.
51. Webster, J.J.; Kit, C. Tokenization as the initial phase in NLP. COLING 1992 volume 4: The 14th international conference on computational linguistics, 1992.
52. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; others. Language models are few-shot learners. *Advances in neural information processing systems* **2020**, *33*, 1877–1901.
53. Roumeliotis, K.I.; Tselikas, N.D. ChatGPT and Open-AI Models: A Preliminary Review. *Future Internet* **2023**, *15*, 192. doi:10.3390/fi15060192.
54. OpenAI. API Reference-OpenAI API, 2023. Accessed: September 18, 2023.
55. OpenAI. GPT-OpenAI API, 2023. Accessed: September 18, 2023.
56. Reynolds, L.; McDonell, K. Prompt programming for large language models: Beyond the few-shot paradigm. Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–7. doi:10.1145/3411763.3451760.
57. Gore, Ross J and Lynch, Christopher J. [1902417-1] Understanding Twitter Users. Old Dominion University Institutional Review Board, 2022. IRB Exempt Status, Exemption Category #2.
58. Watson, D.; Clark, L.A.; Tellegen, A. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* **1988**, *54*, 1063. doi:10.1037/0022-3514.54.6.1063.
59. Boyd, R.L.; Ashokkumar, A.; Seraj, S.; Pennebaker, J.W. The development and psychometric properties of LIWC-22. Austin, TX: University of Texas at Austin, 2022, pp. 1–47.
60. Mohammad, S.M.; Turney, P.D. Nrc emotion lexicon. *National Research Council, Canada* **2013**, *2*, 234.
61. Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; Stede, M. Lexicon-based methods for sentiment analysis. *Computational linguistics* **2011**, *37*, 267–307. doi:10.1162/COLI_a_00049.
62. Crawford, J.R.; Henry, J.D. The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *British journal of clinical psychology* **2004**, *43*, 245–265. doi:10.1348/0144665031752934.
63. Gore, Ross J and Lynch, Christopher J. Effective & Individualized Risk Communication. Old Dominion University, 2023. Number 300916-010, Funding Agency: Old Dominion University.
64. Mitchell, L.; Frank, M.R.; Harris, K.D.; Dodds, P.S.; Danforth, C.M. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one* **2013**, *8*, e64417. doi:10.1371/journal.pone.0064417.
65. Gore, R.J.; Diallo, S.; Padilla, J. You are what you tweet: connecting the geographic variation in America's obesity rate to twitter content. *PloS one* **2015**, *10*, e0133505. doi:10.1371/journal.pone.0133505.
66. Padilla, J.J.; Kavak, H.; Lynch, C.J.; Gore, R.J.; Diallo, S.Y. Temporal and spatiotemporal investigation of tourist attraction visit sentiment on Twitter. *PloS one* **2018**, *13*, e0198857. doi:10.1371/journal.pone.0198857.
67. OpenAI. How can I use the ChatGPT API? | OpenAI Help Center. https://help.openai.com/en/articles/7232945-how-can-i-use-the-chatgpt-api, 2023. Accessed: 2023-09-20.
68. National Academies of Sciences, E.; Medicine.; others. Fostering Responsible Computing Research: Foundations and Practices **2022**. doi:10.17226/26507.
69. D'Amour, A.; Heller, K.; Moldovan, D.; Adlam, B.; Alipanahi, B.; Beutel, A.; Chen, C.; Deaton, J.; Eisenstein, J.; Hoffman, M.D.; Hormozdiari, F.; Houlsby, N.; Hou, S.; Jerfel, G.; Karthikesalingam, A.; Lucic, M.; Ma, Y.; McLean, C.; Mincu, D.; Mitani, A.; Montanari, A.; Nado, Z.; Natarajan, V.; Nielson, C.; Osborne, T.F.; Raman, R.; Ramasamy, K.; Sayres, R.; Schrouff, J.; Seneviratne, M.; Sequeira, S.; Suresh, H.; Veitch, V.; Vladymyrov, M.; Wang, X.; Webster, K.; Yadlowsky, S.; Yun, T.; Zhai, X.; Sculley, D. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *The Journal of Machine Learning Research* **2022**, *23*, 10237–10297. doi:10.5555/3586589.3586815.
70. National Academies of Sciences, E.; Medicine.; others. The Roles of Trust and Health Literacy in Achieving Health Equity: Clinical Settings: Proceedings of a Workshop-in Brief **2023**. doi:10.17226/26888.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.