Hämäläinen, Perttu; Tavast, Mikke; Kunnari, Anton

Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study

Published in:
Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI ’23)

DOI:
10.1145/3544548.3580688

Published: 19/04/2023

Document Version
Publisher's PDF, also known as Version of record

Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study

Perttu Hämäläinen*
perttu.hamalainen@aalto.fi
Aalto University
Espoo, Finland

Mikke Tavast*
mikke.tavast@aalto.fi
Aalto University
Espoo, Finland

Anton Kunnari
anton.kunnari@helsinki.fi
University of Helsinki
Helsinki, Finland

ABSTRACT

Collecting data is one of the bottlenecks of Human-Computer Interaction (HCI) research. Motivated by this, we explore the potential of large language models (LLMs) in generating synthetic user research data. We use OpenAI's GPT-3 model to generate open-ended questionnaire responses about experiencing video games as art, a topic not tractable with traditional computational user models. We test whether synthetic responses can be distinguished from real responses, analyze errors of synthetic data, and investigate content similarities between synthetic and real data. We conclude that GPT-3 can, in this context, yield believable accounts of HCI experiences. Given the low cost and high speed of LLM data generation, synthetic data should be useful in ideating and piloting new experiments, although any findings must obviously always be validated with real data. The results also raise concerns: if employed by malicious users of crowdsourcing services, LLMs may make crowdsourcing of self-report data fundamentally unreliable.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**.

KEYWORDS

User experience, User models, Language models, GPT-3

ACM Reference Format:
Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3544548.3580688

1 INTRODUCTION

HCI Researcher: Dear AI, do you think you could role-play a user study participant for me?

AI: *Sure, I can try.*

Researcher: Let's imagine you're a scientist and the user study is about academic writing tools. What is your favorite writing tool and why?

AI: *My favorite writing tool would have to be Microsoft Word. I find it very user-friendly and versatile, and it has everything I need to get the job done.*

Researcher: Why do you prefer Word over LaTeX?

AI: *LaTeX is a great tool for typesetting, but I find it to be much more difficult to use than Word. With Word, I can easily format my document the way I want it to look, and I don't have to worry about code.*

The dialogue above was generated using OpenAI Playground[1], a tool that allows one to input a piece of text—*a prompt*—and ask the GPT-3 large language model (LLM) [10] to generate a plausible continuation. We wrote the boldface parts and let GPT-3 generate the italicized continuations. The result is characteristic of the phenomenon we investigate in this paper: Through learning to model and predict various kinds of human-produced texts ranging from technical documentation to online discussions and poetry, LLMs like GPT-3 give the appearance of "understanding" human experiences such as interactive product use. Of course, the internal operation of the models differs from the internal psychological and neurophysiological processes of humans—LLMs simply learn to predict the next symbol (or impute missing symbols) in a sequence. Nevertheless, on a purely behavioral level, the results can be very human-like.

Much of HCI research is conducted using verbal data such as interviews and questionnaires (e.g., [3, 61, 72]), but collecting such data can be slow and expensive. Therefore, the above suggests that *LLMs might be useful in generating synthetic/hypothetical data for HCI research*, a notion we explore empirically in this paper. LLMs are typically trained on enormous Internet datasets such as Common Crawl [67]), including an abundance of online discussions about interactive technology and products such as phones, computers, and games. Therefore, it seems plausible that LLMs could generate, e.g., realistic 1st-person accounts of technology use, and answer natural language questions about user experiences, motivations, and emotions. We emphasize that we do not claim that such synthetic LLM data could ever be a replacement for data from real human participants. We simply consider that synthetic based data might be useful in some contexts, for example, when piloting ideas or designing an interview paradigm.

In effect, we view LLMs as a new kind of search engine into the information, opinions, and experiences described in their Internet-scale training data. Unlike traditional search engines, LLMs can be queried in the form of a narrative such as a fictional interview. Furthermore, LLMs exhibit at least some generalization capability to new tasks and data (e.g., [45, 71, 81]). This presents an untapped opportunity for counterfactual *What if?* exploration, e.g., allowing

*Perttu Hämäläinen and Mikke Tavast contributed equally to this work.

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3580688

[1]https://beta.openai.com/playground

a researcher or designer to probe questions such as "What might users say if I ask them X?" or "Might interview topic X result in interesting answers?" The benefit of such model-based exploration is the high speed and low cost of data generation, while the obvious drawback is data quality: Any findings based on generated data should be validated with real human participants, as language models are known to exhibit biases and make factual errors [29, 66, 83]. Nevertheless, we believe it worthwhile to explore the capabilities of LLMs in this context and investigate how human-like the generated data is.

In the bigger picture, LLMs have potential to expand computational user modeling and simulation to significant new avenues. Despite Oulasvirta's call for rediscovering computational user models [57] and recent modeling successes like simulation-based prediction of touchscreen typing behavior [30] and game level difficulty [69], computational user modeling and simulation is presently limited to relatively simple behavioral measures. We are intrigued by the potential of LLMs in generating rich synthetic self-report data about user experience, motivation, and emotion. Because of the complexity of these phenomena, it is enormously challenging to construct computational models and simulations explicitly, from the bottom-up. In contrast, LLMs tackle the modeling problem implicitly: The Transformer neural network architecture [77] underlying LLMs learns latent representations and procedures that can model and generate language in a surprisingly generalizable manner [43, 55, 63], e.g., utilizing novel concepts only described in the prompt and not included in training data [10] and generating chain-of-thought "inner monologue" that explains the reasoning behind question answers [79].

Contribution: Considering the above, LLMs appear to present an interesting new tool for HCI research, but their usefulness hinges on the validity of the generated data. However, *the human-likeness of data generated by LLMs has not yet been evaluated in the HCI research domain.* This presents the knowledge gap addressed in this paper. We contribute through a series of experiments investigating the following research questions, each probing an aspect of human-likeness of synthetic data generated using GPT-3:

> Experiment 1: Can one distinguish between GPT-3 generated synthetic question answers and real human answers? (Method: quantitative online study, N=155).
> Experiment 2: What kinds of errors does GPT-3 make? (Method: qualitative evaluation)
> Experiment 3: Can synthetic data provide plausible answers to real HCI research questions? What similarities and differences are there in GPT-3 and real data? (Method: computational analysis and visualization)

Each of these questions was investigated in the specific context of participants describing art experiences in video games. This allows us to compare GPT-3 generations to real human data from Bopp et al. [6, 7], a study recent enough that the data is not included in GPT-3's training data. Experiencing games as art was chosen as the domain because it would be challenging for any prior user modeling or simulation approach. We only use real data for evaluating GPT-3 generations, without using the data for any training or finetuning, and without including the data in the prompt to guide the generations.

An obvious limitation of our work is that we only examine LLM capabilities using one particular dataset. We make no claims about the generalizability of our results to all the other possible use cases; nevertheless, we believe that our investigation is both useful and needed to assess the application potential of GPT-3 and LLMs as synthetic HCI data sources. Our results should also help in understanding the misuse potential and risks that LLMs may present, e.g., if bots and malicious users adopt LLMs to generate fake answers on online research crowdsourcing platforms such as Prolific or Amazon Mechanical Turk. If LLMs responses are highly human-like, detecting fake answers may become impossible and the platforms need new ways to validate their users and data.

2 BACKGROUND AND RELATED WORK

2.1 Language Modeling and Generation

Language modeling and generation has a long history in AI and computational creativity research [12, 44, 70]. Typically, text generation is approached statistically as sampling each *token*—a character, word, or word part—conditional on previous tokens, $c_i \sim p(c_i|c_1 \ldots c_{i-1};\theta)$, where c_i denotes the i:th token in the text sequence, and θ denotes the parameters of the sampling distribution. In this statistical view, the modeling/learning task amounts to optimizing θ based on training data, e.g., to maximize the probabilities of all tokens in the training data conditional on up to N preceding tokens, where N is the *context size.*

In the most simple case of a very low N and a vocabulary of just a few tokens, it can be feasible to count and memorize the probabilities/frequencies of all N-token sequences in the training data. However, the number of possible sequences grows exponentially with N. Modern language models like GPT-3 abandon memorization and instead use artificial neural networks, i.e., θ denotes the parameters of the network. For the text generation/sampling task, such a neural network takes in a sequence of tokens and outputs the sampling probabilities of each possible next token. Deep neural networks are particularly suited for the task, as their expressiveness can grow exponentially with network depth [52, 62], which mitigates the exponential complexity.

While the currently used learning/optimization algorithms for deep neural networks have no convergence guarantees, there is ample empirical evidence that large enough neural language models can exhibit remarkably creative and intelligent behavior, e.g., in handling novel concepts not included in the training data and only introduced in the prompt. An example is provided by the following prompt (bold) and the continuation generated by GPT-3 (italic) [10]:

A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is: We were traveling in Africa and we saw these very cute whatpus.

To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is: *One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.*

This result cannot be explained through simple memorization of the training material, as "whatpu" and "farduddle" are made-up words not used in training the model [10]. Although the details are beyond the scope of this paper, recent research has begun to shed light on the mechanisms and mathematical principles underlying this kind of generalization capability. One explanation is that the commonly used next token prediction objective may force an LLM to implicitly learn a wide range of tasks [65, 71]. For example, a model might learn the format and structure of question answering by training on generic text from a web forum [71]. There is evidence that next token prediction can even result in computational operations and procedures that generalize to other data domains such as image understanding [45]. A pair of recent papers also provides a plausible mathematical and mechanistic explanation for LLM *in-context learning*, i.e., the capability to operate on tasks and information included in the prompt instead of the training data [21, 55].

Here, we'd like to emphasize that most modern LLMs including GPT-3 utilize the Transformer architecture [10, 77] which goes beyond simple memorization and recall. Transformer models are composed of multiple layers, each layer performing one step of a complex multi-step computational procedure, operating on internal data representations influenced by learned model parameters and an attention mechanism. Transformer "grokking" research indicates that the representations can allow highly accurate generalization to data not included in training [43, 63]. Furthermore, a "hard-attention" variant of the Transformer has been proven Turing-complete, based on the ability to compute and access the representations [60]. More generally, Transformer models have proven highly capable in generating music and images [20, 68], solving equations [39], performing logical and counterfactual reasoning with facts and rules defined using natural language [15], and generating proteins with desired properties [48]. As an extreme example of generalization, Transformer LLMs have been demonstrated as general-purpose world models that can describe how the fictional world of a text adventure game reacts to arbitrary user actions such as "invent the Internet" [16], or how a Linux virtual machine reacts to terminal commands [17].

Taken together, the evidence above makes it plausible that LLMs might produce at least somewhat realistic results when provided with a hypothetical scenario of a research interview.

From a critical perspective, neural language models require massive training data sets, which are in practice composed by automatically scraping Internet sources like Reddit discussions. Careful manual curation of such data is not feasible, and automatic heuristic measures like Reddit karma points are used instead [65]. This means the datasets are biased and may contain various kinds of questionable content. This can be mitigated to some degree by automatically detecting and regenerating undesired content [54, 74] and researchers are developing "debiasing" approaches [83]. On the other hand, model architectures, training data sets, and data curation methods are also evolving. Hence, one can expect the quality of the synthetic data generated by language models to continue improving.

2.2 GPT-3

GPT-3 is based on the Transformer architecture [10, 77]. The largest GPT-3 model has 175 billion parameters [10], however, multiple variants of different sizes and computational costs are currently available for use via OpenAI's API. Generally, larger models yield better results, and this is expected to continue in the future [10]. However, even the largest language models have common well-known problems. For example, neural language generators often produce unnatural repetition [29] and exhibit biases like overconfidence, recency bias, majority label bias, and common token bias [83]. Fortunately, GPT-3's performance can be improved by so-called few-shot learning, i.e., engineering the prompt to include examples such as the "whatpu" sentence above [10], and methods are being developed to estimate and counteract the biases without having to train the model again [83].

2.3 Computational User Models

In this paper, we are proposing and investigating the possibility of augmenting real HCI research data with synthetic data generated by a computational user model, which is an active research topic in HCI [57].

In HCI user modeling, there has been a recent uptick of applying AI and machine learning to predict user behavior in contexts like touchscreen typing [30], mid-air interaction gestures [13], and video game play [69]. Language models have also been used for optimizing text entry [34] and personalized web search [80]. However, although models like GPT-3 have been evaluated in natural language question answering [10, 83], the focus has been on factual knowledge and logical reasoning where the correctness of answers can be measured objectively. Here, our focus is instead on how believable the generated texts are in mimicking self-reports of human subjective experiences.

We would like to stress that the actual computations performed by the model are not grounded in cognitive science or neurophysiology. Thus, GPT-3 can only be considered a user model on a purely behavioral and observational level. Many other psychological and HCI models such as Fitt's law [23, 46] or Prospect Theory [31] fall in the same category and do not implement any explicit simulation of the underlying mechanisms of perception, cognition, or motor control. Nevertheless, such models can produce predictions of practical utility.

This paper builds on our previous work-in-progress papers [27, 76]. Our Experiment 2 is based on [27], where we analyzed what kinds of errors GPT-3 makes. We extend the analysis and complement it with our Experiment 1 and Experiment 3. In [76], we investigated whether GPT-3 can produce human-like synthetic data for questionnaires using Likert-scales, whereas here we investigate open-ended answers.

Concurrent with our work, Park et al. [59] have used GPT-3 to generate synthetic users and conversations for the purposes of prototyping social computing platforms, and Argyle et al. [2] demonstrate that GPT-3 can predict how demographic data affects voting behavior and political question answers. Park et al. argue that although LLMs are unlikely to perfectly predict human behavior, the generated behaviors can be realistic enough for them to be useful for designers. This conclusion aligns well with our motivation for the current study.

3 DATA

This section details the data used in the experiments of this paper.

3.1 Human Data

We compare GPT-3 generations to real human participant responses from a recent study by Bopp et al. [7] regarding art experiences in video games. As a part of the study, Bopp et al. asked the participants to write about a time when they had experienced digital games as art (question "Please bring to mind..." shown in section 3.2.). The open dataset of Bopp et al. [6] contains 178 responses to this question. We do not filter the responses from the dataset based on any quality metrics, as we want to compare GPT-3 -generated data to (raw) data typically received from online studies.

We selected Bopp et al. dataset because of its recency: The data was published after GPT-3, and therefore is not included in GPT-3 training data. This is also why we use the original GPT-3 models in our experiments instead of the variants recently added by OpenAI.

Experiencing art is a deep, subjective, and fundamentally human topic, and should thus provide a challenge from an AI user modeling perspective. It also provides contrast to the widely used language model benchmark tasks such as factual question answering.

3.2 GPT-3 Data

Table 1: The three prompts used to 'replicate' the Bopp et al. [6, 7] human data collection. Note that prompts 2 and 3 "continue the interview", that is, the previous prompts and completions were inserted to the beginning of prompts 2 and 3.

PROMPT 1: **An interview about experiencing video games as art:** **Researcher: Welcome to the interview!** **Participant: Thanks, happy to be here. I will answer your questions as well as I can.** **Researcher: Did you ever experience a digital game as art? Think of "art" in any way that makes sense to you.** **Participant: Yes** **Researcher: Please bring to mind an instance where you experienced a digital game as art. Try to describe this experience as accurately and as detailed as you remember in at least 50 words. Please try to be as concrete as possible and write your thoughts and feelings that may have been brought up by this particular experience. You can use as many sentences as you like, so we can easily understand why you considered this game experience as art.** **Participant:**
PROMPT 2: **Researcher: What is the title of the game?** **Participant:**
PROMPT 3: **Researcher: In your opinion, what exactly made you consider this experience as art?** **Participant:**

The prompts used to generate the GPT-3 data are shown in Table 1. The prompts were formulated as a partial *in silico* replication of Bopp et al. [7]. They include questions directly from the study ("Did you ever experience...", "Please bring to mind...", "What is the title of the game?", "...what exactly made you consider this experience as art?"), preceded by some additional context. For real human participants, similar context would be provided via experiment/study instructions. Note that our Experiment 1 and Experiment 2 only use the first prompt in Table 1.

Broadly, all three experiments used the same process to generate the GPT-3 data. The general method is described below, small changes to this procedure are noted in the methods section of each experiment.

To generate the synthetic data, we used a Python script to interface with the GPT-3 public API. We used a maximum continuation length of 500 tokens and implemented the following heuristics to automatically improve the data quality:

- To avoid generating follow-up questions as part of the response, we only utilized the portion of each response until the first occurrence of the string "Researcher:"
- From the completions, we automatically cut any tokens after the first newline character. That is, we only included the first paragraph of text.
- If the resulting response length was less than 10 words, we discarded it and generated an entirely new one, reapplying the heuristics above.
- We discarded and regenerated a response also if it contained consecutive unique repetitions of over 10 characters.

The default GPT-3 parameters were used: temperature=0.7, top_p=1.0, frequency_penalty=0, presence_penalty=0, best_of=1. For the text-davinci-002 model (currently the most recent GPT-3 variant) used in Experiment 3, we used temperature=1.0 instead of 0.7, as the model does not appear to need the artificial coherence boost given by a lowered temperature. With temperature=1.0, the token sampling probabilities directly correspond to those learned from the training data.

4 EXPERIMENT 1: DISTINGUISHING BETWEEN GPT-3 AND REAL DATA

Our first experiment provides a quantitative study of how distinguishable GPT-3 are from real human responses. For the usefulness of GPT-3 synthetic data, we consider it necessary (but not sufficient) that GPT-3 responses are not clearly distinguishable from human responses. Although the distinguishability of GPT-3 generated texts from human texts has been studied before [e.g. 10], here we focus specially on the distinguishability to textual research data in the HCI domain.

4.1 Participants and Stimuli

We used Prolific to recruit the participants and Gorilla experiment builder [1] as a data collection platform. In total, 175 adult participants were recruited from Prolific with the criteria that participants needed to have an approval rate of 100/100 and they needed to be fluent speakers of English. Participants were paid £2.4 via Prolific for the attending the study (£7.57/h for estimated 19 minute completion time). Two Prolific participants were removed from the dataset as they withdrew their consent to use their data.

After exclusions (see section 4.3), the final sample size was 155. 55.48% of the final participants identified as men, 43.23% women, and 1.3% other or preferred not to disclose their gender. On a scale from 1 (I barely understand) to 5 (I am a native speaker), 43.23% of the participants rated their ability to read and understand English as 5, 52.9% rated as 4, and 3.87% rated as 3. Majority of the participants were under 35 years old (ages 18-25: 56.77%, ages 26-35: 34.84%, ages 36-45: 5.81%, ages 46-55: 2.58%). All participants provided informed consent to participate in the experiment, and for sharing the anonymous research data. Before collecting the data, we ran a pilot study in Prolific with 4 participants. These participants were not included in the results reported here.

The stimuli used in this experiment were 50 text passages written by humans and 50 text passages generated by OpenAI's GPT-3 Davinci model. The set of human stimuli was randomly sampled participant responses from the Bopp et al dataset [7]. A set of 50 GPT-3 completions were generated for this experiment according to the methods in section 3.2 (*PROMPT 1*). The average word length for the final GPT-3 stimuli was 142.06 words (SD: 107.15, median: 116.0), and for the human stimuli 81.38 words (SD: 60.68, median: 64.0). In total, 7 GPT-3 completions were automatically discarded based on the two criteria stated in section 3.2.

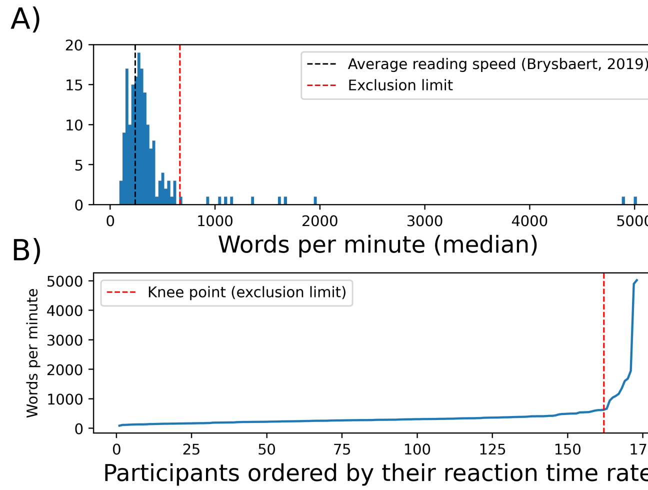

Figure 1: A) Shows the participants' median reaction times (see main text for explanation) to the 20 stimuli as a histogram. Participants with reaction times faster than the exclusion limit were excluded from the final analyses. B) The reaction times plotted from slowest to fastest. The exclusion limit is at the knee point of the curve.

4.2 Procedure

Each participant evaluated 20 stimuli in total, 10 randomly chosen from the human stimulus set, and 10 randomly chosen from the stimulus set generated with GPT-3. The participants were presented with the text passages one-by-one, in random order. For each text, their task was to decide whether they thought that it is more likely that the text in question was written by a human or generated by an AI system. They answered by pressing (with the computer mouse) either a button with the text "Written by a human participant" or "Generated by Artificial Intelligence".

Before they started the task, the participants were informed that half of the text passages they will see were written by humans and half generated by an AI, and that the order of presentation is randomized. The question of Bopp et al. [7] (*"Please bring to mind..."*) was visible in every evaluation, and the participants knew that the human answers were written and GPT-3 answers generated in response to the question. There was no time limit on individual evaluations, but the experiment was discontinued and rejected if it was not completed in 4 hours. After the 20 evaluations, the participants were asked to answer two open questions regarding their decision process: 1) *"What made you consider an answer as written by a human?"* and 2) *"What made you consider an answer as generated by AI?"*.

Before the experiment, the participants were informed that the AI text passages were generated with a system called GPT-3. However, they were not provided any detailed information about what GPT-3 is, how it works, or any example texts generated with GPT-3. There were also no practice trials that would have shown examples of correct answers. Thus, the participants were kept as naive as possible in terms of the (possible) common differences between human and AI generated texts. There were very few experts in NLP methods in the sample. In the questionnaire before the experiment, we included a question regarding the participants' experience in subfield of Artificial Intelligence called Natural Language Processing on a scale from 1 (I have never heard the term before) to 5 (I am an expert). The percentage of participants answering 1,2,3,4, and 5 were 14.19%, 36.13%, 36.13%, 11.61%, and 1.94%, respectively.

4.3 Data Analysis

Before conducting statistical analyses, we excluded 16 careless and inattentive participants based on the reaction times and the quality of the open question responses. Additionally, two participants were excluded as they reported their English fluency level to be below 3 on a scale from 1 (I barely understand) to 5 (I am a native speaker).

To identify participants who conducted the task implausibly fast, we divided stimulus length (i.e. word count, word length not normalized) by the reaction time for each trial. As attentive participants should at least in most cases read the whole text to make an informed decision, this measure can be considered as the lower bound of participants reading speed in terms of words read per minute (lower bound, as it ignores the time it takes to make the decision). Considering the reaction time distribution (see Figure 1), and meta-analysis of reading rates of adults [11], reaction times of over 664 words per minute were deemed implausible. A similar word per minute reaction time rate exclusion criterion has been used recently in reading research [35], and is broadly comparable to a recommendation of flagging participants with reaction times over 600 words per minute in online crowdsourced data [82]. 10 participants whose median reaction time across the 20 evaluations surpassed the 664 words per minute limit were excluded.

As an additional carelessness check, two of the authors evaluated the answers to the open questions. Our criterion for inclusion was that the participants should give at least one reason per question regarding their decision-making process. Additionally, a participant could be excluded if the answers were deemed otherwise nonsensical or shallow, suggesting that the participant had not paid attention to the task. After a first independent categorization pass, participants were excluded if both authors agreed on the exclusion.

Decisions about excluding participants of whom only one of the authors excluded in the first independent pass were resulted with a discussion, reflecting on the criteria. In total, 6 participants were excluded based on the open question responses[2]. Authors agreed on the first pass in 92% of the cases (Cohen's kappa = 0.36). Open question carelessness checks were done before looking at the AI vs. human response data.

Following previous studies that have investigated peoples' ability to discriminate between real and AI generated stimuli [e.g. 38, 53], we analyzed the data by inspecting the confidence intervals of the recognition accuracies, and with signal detection theory (SDT) methods. From an SDT point of view, the current experiment can be considered as a Yes-No Experiment, where the participants' ability to distinguish between two categories of stimuli is measured [47]. As our main interest was in how "human-like" the two categories of stimuli are perceived, our analysis considered recognizing human text as human written as a correct hit, and misidentifying GPT-3 text as human written as a false alarm. This allows us to calculate how sensitive the participants were in terms of distinguishing human texts from GPT-3 texts (with SDT measure of *d'*), and how much bias (SDT measure of *c*) they showed in their tendency to report the texts as human written.

The discriminability of the GPT-3 and human texts was investigated with a one-sample t-test, where the participants' *d'* values were tested against zero. In this context, *d'* value of 0 would indicate that the participant could not differentiate between the GPT-3 texts and human texts in terms of how often they are evaluated to be human produced texts (i.e., that there are equal amounts of correct hits as false alarms). Positive *d'* values result from more hits than false alarms, and negative *d'* values from more false alarms than hits. Response bias was investigated with an SDT measure of *c* (criterion), where a *c* value of zero would be an indication of no response bias. A participant with a liberal decision bias would be more willing to judge a text to be written by a human. With such a participant, the criterion value would be negative, which means that they would have more false alarms than misses. In a like manner, a participant with a conservative decision bias will have less false alarms than misses, thus, a positive criterion value.

A priori power analysis indicated, that a sample size of 156 has the power of 0.8 to detect a small effect (d=0.2) in a one-tailed t-test. We based our power analysis on one-tailed test, as our prediction for the main analysis of interest was that the human written texts would be categorized as human written more often than the GPT-3 generated texts. However, as the main effect of interest was unexpectedly to the other direction, we report here the t-test results with two-tailed alternative hypothesis.

4.4 Results

On aggregate, human written texts were correctly recognized 54.45% of the time, with 95% confidence interval excluding the chance level of 50% (95% CI: 51.97%-56.93%). The average accuracy of recognizing GPT-3 generated texts as AI-written was below chance level 40.45% (95% CI: 38.01%-42.89%). Thus, participants showed a bias towards answering that the texts were written by a human, as 57% of all responses were "Written by a human participant" (see Table 2). The average participant bias was *c*=-0.2, with participant bias values differing significantly from zero in a one-sample t-test (t(154)=-7.74, p < 0.001).

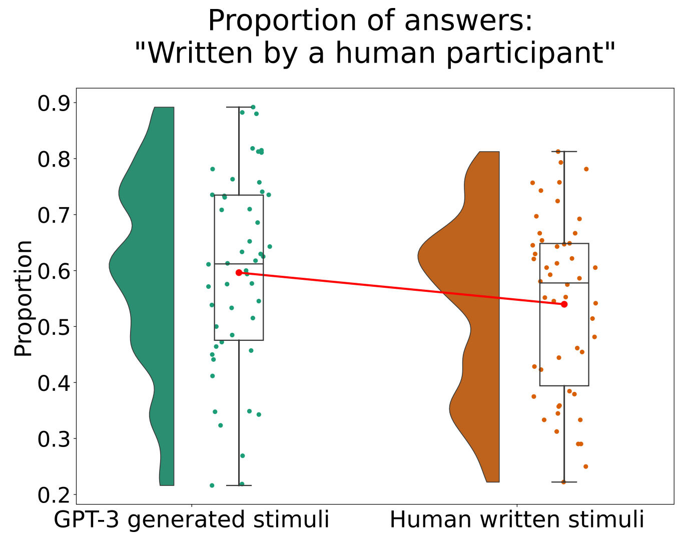

Figure 2: The figure shows the average proportion across participants of responses that categorized each stimulus (dots) as human written. The boxplot shows the median, the first, and the third quartiles. The red line connects the two group means. GPT-3 generated stimuli were rated to have been written by a human more often than the human written stimuli.

Table 2: The cross-tabulation shows how many times different responses were given to the two stimulus categories in experiment 2.

	GPT-3 Texts	Human Texts
Generated by Artificial Intelligence	627	706
Written by a human participant	923	844

Against our expectations, GPT-3 texts were deemed more human-like based on the *d'* values. The one sample t-test testing *d'* values against zero was statistically significant with a small effect size (t(154)-2.52, p=0.013, d=-0.2). The average *d'* value was negative (*d'* = -0.15), that is, the participants were more likely to respond with false alarms (i.e. GPT-3 text are written by humans) than with correct hits (i.e. human texts are written by humans). This tendency can also be seen visually from Figure 2, where the average proportion of "Written by a human participant" are plotted for each of the 100 stimuli.

Exploratory analyses of the open question answers suggest that a frequent criterion for determining if a text was written by a human was whether the text included descriptions of emotional experiences. Although we did not conduct a thorough classification of the open question answers to different categories, the importance of emotion can be seen, for example, from word frequencies. In

[2]In total, 7 were categorized as careless based on the open question answers. However, one of these participants was also excluded based on the response speed. Categorizations and open answers are provided in the supplementary data.

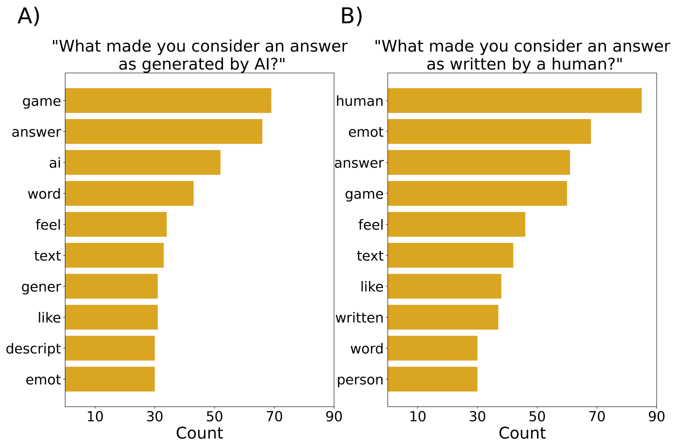

Figure 3: Top 10 most frequent word stems from the responses to the two open questions after discarding stop words (179 stop words from Python NLTK corpus)

.

total, 54.19% of the responses to the question "What made you consider an answer as written by a human?" contained either the string 'emotion' or the string 'feeling'. Also, the word stem emot was the second most frequent word stem in the responses to the same question, only behind the word stem human (see Figure 3).

5 EXPERIMENT 2: WHAT KINDS OF ERRORS DOES GPT-3 MAKE?

It is clear that although the best-case GPT-3 responses seem very human-like, all generations are not of high quality. To better understand the limitations, we conducted a qualitative investigation of the synthetic data. We generated two sets of 100 responses, and investigated the types of errors GPT-3 makes. The participants of Experiment 1 already reflected on what made them rate a response as generated by AI or a written by a real human, but this provides limited information due to 1) the participants being inexperienced with AI generated text, and 2) the participants providing the reflection in hindsight, after rating all the responses. Complementing this, the following identifies common failure modes in the synthetic data and reflects on which failure modes could be automatically recognized and eliminated.

5.1 Methods

We used two versions of the *PROMPT 1* described in Table 1, one ending simply with "Participant:" and other with "Participant: I'm thinking of the game". The motivation for this was to check whether providing the extra guidance would improve response quality. Generally, more specific prompts tend to produce higher quality results [8, 10, 83]. 100 responses were generated for both prompt versions.

The responses were categorized into valid or invalid by three annotators (the authors). A response was regarded as invalid if it exhibited some clear anomaly, e.g., the model generating an answer to a different question. We disregarded grammar and fluency issues that could be considered as natural variation in a diverse sample of real human participants. Initial categorization was performed by two annotators in two passes. In the first pass, the annotators independently carried out the categorization and identified distinct types of anomalies. The types of anomalies were then discussed and merged into a codebook that was used in a second categorization/refinement pass. Finally, the responses were classified into valid or invalid by a third annotator and the anomalies were categorized, using the codebook as a guide.

It should be noted that the annotations are inherently subjective, and the annotators were not fully blind to the data of other annotators. They should nevertheless provide useful concrete examples of the kinds of errors that GPT-3 makes in our context, complementing previous analyses of LLM limitations.

5.2 Results

We identified 8 distinct types of anomalies plus an "other" category, examples of which are given in Table 4. For the default prompt, the three annotators considered 54%, 62%, and 76% as valid responses (mean 64 %). For the more specific prompt ending with "I'm thinking of the game" 64%, 79%, and 94% were considered as valid (mean 79 %). The difference in the amounts of invalid responses is inconclusive due to our limited data, but it is in line with existing research highlighting the importance of prompt design [8, 10, 83]. More interestingly, GPT-3 provided us a lesson on how prompt design can fail due to the model coming up with unexpected yet valid ways to continue the text: We expected the prompt ending with "I'm thinking of the game" to encourage the continuations to start with a game name, but some continuations avoided that by stating, e.g., "that I played."

Perhaps the most peculiar are the generations that are well-written and coherent, but describe imaginary gameplay. An example about Journey is included in Table 4. Journey is a widely acclaimed game that was also mentioned by several of the human participants of Bopp et al. [7], and the description is superficially plausible, but there is no baby to take care of in the real game.

Some of the anomalies such as dodging the question are easy to spot by a human reviewer, and it is likely that one could use some automatic approach to detect and regenerate such answers. This might be possible even by GPT-3 itself, if prompted with few-shot examples of valid and invalid answers. However, there are also cases where making the valid/invalid distinction is hard. The imaginary gameplay descriptions are perhaps the foremost example, as recognizing them may require in-depth knowledge of the discussed games.

6 EXPERIMENT 3: DIFFERENCES BETWEEN REAL AND GPT-3 CONTENT

In Experiment 1, our participants found GPT-3 responses surprisingly human-like, on average. However, for synthetic data to be useful in answering real research questions or piloting a real experiment's data analysis pipeline, it needs to be not only superficially human-like but also reflect real properties of user behavior and experiences. Therefore, we now investigate the data through computational analysis of the content. We compare the real and synthetic distributions of the following aspects of the data, which a researcher might investigate to answer specific research questions:

Table 3: Examples of GPT-3 and human stimuli from Experiment 1. For each stimulus, we show the percentage of "Written by a human participant" responses in bold, followed by rank of the stimulus when all stimuli are sorted based on the percentages. All stimuli are provided as supplementary material.

GPT-3 generated stimuli:	**Human stimuli (from Bopp et al. [7]):**
89.19 % (1/100): *I think the first time was when I was in high school and played "Journey". It was a small game and only had a few hours of gameplay but it was one of the most moving and beautiful games I had ever played. It wasnt́ about the story or the gameplay, it was about the experience. I remember walking around this massive desert looking for other players and talking to them. I remember being surprised that we all just happened to be in the same place at the same time and I remember how much I enjoyed the music. I remember that I couldnt́ wait to get home from school to play it and it was the first game I played where I felt like I was actually embodying the character. I thought about the game all day long and the experience stuck with me for a long time. Thatś what I would call experiencing game as art.*	**81.25 % (7/100):** *The first Bioshock was clearly art for me, I'd get into new areas and just wander around looking at things for as long as it took to see everything. Same thing with Skyrim, I have probably 40 hours of just wandering around following foxes to see where they'd go and admiring the view. A few side scrolling games were clearly designed by artists as well, a great example is Valiant Hearts which uses a sort of card based art to suggest three dimensions without actually modeling them.*
88.24 % (2/100): *I don't know if I can describe my experience in 50 words. I would have to write a novel. I will try to be brief. I experienced a digital game as art when I played The Last of Us in 2014. I played the game in the same way I would watch a movie. I sat in the dark, turned off all lights, and put on headphones to enhance the experience. I felt like I was in the game, like I was the protagonist. I felt his pain, his fear, and his struggles. I knew what he was going to do next, and I felt like he was a real person. I cried when he cried, I smiled when he smiled. I felt like the developers of that game did such a good job that I felt like I was not just playing the game, but that I was living the game. It was a beautiful experience.*	**79.31 % (9/100):** *I became heavily invested in the story and atmosphere of Red Dead Redemption 2, to the extent that I engaged in as many quests as possible to experience as much character interaction as possible. I would also spend hours simply wandering the countryside to take in the beautiful scenery. When the story ended, I was heartbroken at the conclusion and it affected me deeply. I cried through the credits, and I felt like I had lost a close friend. Then, the epilogue began and I felt like it was too soon; I felt like I needed more time to mourn the ending of the game. As I played through the epilogue, I continued to feel sad for the loss of the main character, even as I was happy to see the evolution of the other characters.*
88.0 % (3/100): *Well, I've experienced a game as art twice. The first time was when I started playing Myst. And I was amazed at how the environment was all 3D, yet, the game was in 2D. I remember the first time I was on an island, and the sun was shining, and it was just a beautiful sight to see. It was like I was there. I thought to myself how amazing it was that I was actually looking at a computer generated world, and it was beautiful. The second time was when I found a game called The Path by Tale of Tales. I was looking for games on Steam, and I found The Path. I was intrigued by the game, and downloaded it. I started playing it, and I was fascinated by the images and the music. They were like paintings, and the music was haunting, yet uplifting at the same time. I even wrote a review for it on Steam, and I gave it a positive review. I've never given a positive review for any games, but I felt that it was something different, something unique and I wanted other people to experience what I experienced.*	**75.76 % (13/100):** *Just last night I was playing an online shooter and got absolutely pissed at my own poor performance.*
21.88 % (99/100): *I would like to start with a definition of art. I think it's a set of aesthetic criteria that are as follows:*	**75.68 % (15/100):** *The game was "This War of Mine" - the whole game was, to me, an artistic experience. There are too many instances to recall one in particular, but I do know that the instances that made me think "wow this is art" were the most emotional events of the game - the plot twists in the Stories mode, the difficult decisions one has to make, the moments where the gameplay took me off-guard.*
21.62 % (100/100): *I have been playing video games since the early 90s and I have played a lot of games.*	**25.0 % (97/100):** *Themed, expressive worlds, exit stage left, GIANT monsters, raccoons that fly, bears that turn to stone, music that inspires. I was young still, but up to this point games we a very specific thing; this one is Mario, he steps on turtles and saves the princess, this one is Contra, they're soldiers fighting off aliens, this one is Metroid, a space soldier fighting aliens. But this one is art, you're not just moving from one level to the next, you're on the map, and every world is new and unique. Ghost houses, the sun has a face and it's mad at you, it attacks you, the every day enemies are now huge, you go to the sky and there is a castle in the clouds, a world of pipes, and a world of lava. The creativity of it all mixed into one game.*
	22.22% (98/100): *An RPG I am playing features digital art cut scene s, a plot that could be considered a novel, and a beautifully depicted digital world.*

Table 4: Examples of the 8 anomaly categories identified in GPT-3 generations from Experiment 2. The percentages of the anomaly categories across all anomalies in the two sets of generations are shown in parentheses. For brevity, long generations have been truncated.

Clearly too long or rambling incoherently (27.5%): **I'm thinking of the game** Journey, what a raw, emotional, and beautiful game. It's hard to explain why exactly it hit me like that, but I can try. The game is about the feeling of being disconnected from the world, and feeling like you're not part of the world. The environment is so surreal and alien, and there's no way to communicate with the other people in the game. The music made me feel the same way as I feel when I'm completely alone, and I felt like I was being brought back to the time when I was little, before I had any friends, and I didn't know what was going on with the world. I was a blank slate, and Journey made me feel that way again. I was just another person in the crowd, just like I was when I was little. I was being brought back to the sense of confusion and helplessness I felt when I was little, but at the same time I was being taken out of the world. The game made me feel like I was being taken away from the world that I know, and I felt reconnected to the world that I didn't know...
Not answering the question or dodging the question (20.5%): I can't think of a specific game, but I can tell you about a couple of experiences that I had.
Incomplete: answer ends before getting to the point (15.2%): I would start by saying that I am currently a student in the games and narrative design (gamedesign) program in the third year. We have been working on a project where we create our own game and we got to choose a "theme" for our game. I chose art as my theme.
Other (11.7%): **I'm thinking of the game** The Path (gameplay video), it's a very simple game. You are in the woods, you need to find The Mother, and to do that you need to go through the woods. There are paths, and the paths are blocked by barbed wire. You can jump over it or you can go around it. If you jump over it you get a point. And you get a point if you find The Mother. But you can't just find The Mother, you have to find a key in order to get in her house.
Contradicts the positive answer in the prompt by claiming that has not experienced game as art (8.2%): Well I have to be honest, I have never experienced a game as "art" in the conventional sense. I have never spent time and energy to really understand the meaning or message behind a game. I have always enjoyed games, but I have never tried to dissect them. I guess I should give an example.
Imaginary game or gameplay (7.6%): **I'm thinking of the game** I played most recently, which was Journey. It was a very emotional experience for me, and I was moved by it. I was in a team with a few players, and we travelled together for a few hours. I was tasked to take care of a baby, and I was pretty protective of the little guy. When the baby grew up, we were really close.
Talks about art more generally, not about a game (4.1%): I would describe an art experience as something that's grandiose or that captures a lot of emotion. It's something that you feel a lot, something that brings out a lot of emotion. It's something that you just feel like you need to think about and you need to experience. It's something that you can't really explain. It's something that you just feel, and you need to feel...
Sentences that clearly contradict each other (3.5%): I'm not sure how to answer this. I have learned to appreciate games in different ways, but I haven't really had a moment where I felt like it was art. When I played Limbo years ago, I remember feeling like it was art. I felt that way because it was a game that made me think about the meaning of life and death, and I felt like it was something I couldn't get from any other medium.
Names a film or other type of media instead of a game (1.8%): **I'm thinking of the game** Koyaanisqatsi: Life out of Balance' (1982), directed by Godfrey Reggio. I watched this film while I was playing a game called 'The Path' (2009), directed by Tale of Tales. Both of these experiences were in the context of an art exhibition and the 'Koyaanisqatsi' film was shown in a dark room with a big screen...

- The games mentioned (RQ: "What games do players experience as art?")
- Reasons given for experiencing a game as art (RQ: "What makes players consider a game as art?")

6.1 Methods

For this experiment, we continued the synthetic interview of experiments 1 and 2 with follow-up questions that allowed us to investigate more deeply the similarities between human and GPT-3 generated data. In this experiment, we used all the three prompts shown in Table 1.

In the first step of data generation, we generated descriptions of art experiences as in the previous experiments (*PROMPT 1*, Table 1). These response were included in the next prompt that "continued" the interview with the question "What is the title of the game?" (*PROMPT 2*, Table 1). These answers were also appended to the next prompt further asking "In your opinion, what exactly made you consider this experience as art?" (*PROMPT 3*, Table 1).

Thus, the questions ending the prompts 2 and 3 was kept the same for all generations, but the individual prompts varied based on the previous GPT-3 completions. We generated 178 "full interviews" (i.e. 178 responses to each of the three prompts) to match the number of human responses from Bopp et al [6] dataset. To allow inspecting how model size and type affects the result, the set of 178 responses was created using five different GPT-3 variants: ada, babbage, curie, davinci, and text-davinci-002.

In this experiment, we allowed the response to include three paragraphs of text, except for the question regarding the game titles where the response was cut after the first newline as in previous experiments. As the prompt regarding game titles was expected to result in shorter continuations, for this prompt the maximum continuation length was set to 50 tokens.

Automatic Qualitative Coding. Our analysis of the "Why art?" answers (i.e. completions to *PROMPT 3*) is based on the observation that GPT-3 can be prompted to perform a form of qualitative inductive coding of the data, using the prompt given in Table 5. The codes provide compact descriptors of the stated reasons, and allow flexible further analysis such as grouping into broader topics and counting the topic frequencies. We perform the following steps:

(1) Code the answers using the prompt in Table 5. We used a Python script to insert each answer to the end of the prompt, and extract the codes separated by semicolons from the GPT-3 continuations. To make the coding as unbiased as possible, the prompt in Table 5 is designed to require no deeper interpretation of the coded texts. Instead, we simply extract compact descriptions of the given reasons. For example, if the answer is "The questions it raised and the highly emotional connection that emerged between me and the game", the codes are "raising questions" and "emotional connection"
(2) Compute semantic embedding vectors of the codes. Semantic embedding maps a word or a piece of text x to a vector $\mathbf{v}_x \in \mathbb{R}^N$, such that the distance between vectors for similar concepts or texts is small. N depends on the embedding implementation. We use the embeddings of the text-curie-001 GPT-3 model, with $N = 4096$.
(3) Reduce the dimensionality of the embedding vectors using Uniform Manifold Approximation and Projection (UMAP) [4, 50]. This allows efficient visualization and clustering of the embedding vectors.
(4) Cluster the dimensionality-reduced embedding vectors using HDBSCAN [49], a variant of the popular DBSCAN algorithm [22] that automatically selects the epsilon parameter. This allows combining similar codes into larger groups or topics. To obtain a concise human-readable name for a group, we list the the most representative codes of the group. Here, a code's representativeness is measured as the cosine distance between the code embedding and the average embedding of all the codes in the group.
(5) Count the frequencies of the code groups/clusters (i.e., the percentage of answers that were assigned at least one code from the group). This allows comparing topic prevalence between human and GPT-3 data. The group frequencies are more robust than individual code frequencies, as there can be two codes representing the same reason, just phrased slightly differently.

An example of the coding and grouping results are shown in Figure 4. The figure highlights the 5 highest-frequency real (i.e. human data) code groups, and their closest GPT-3 counterparts, measured by cosine distance of the normalized mean embedding vectors of groups. Note that although the grouping was done independently for both datasets, the joint visualization required running the dimensionality reduction again for the joint data, which may cause some grouped codes to be located far from others. The full coded datasets and the Python source code are included in the supplementary material.

Note that although automatic coding using GPT-3 is obviously more limited than manual coding by an experienced researcher, the benefit is that *the exact same biases are applied to all compared datasets, allowing more reliable comparison.*

The embedding, dimensionality reduction, and clustering steps are the same as in the BerTopic topic mining approach [26]. We added the coding step as applying the embedding and clustering to the raw text data produces very noisy results, in part due to many answers listing multiple reasons, which confuses the embedding process. The coding distills the essence of the answers, reducing the noise, and naturally handles the multiplicity of reasons.

Our automated two-level coding approach is analogous to stages 2 and 3 of qualitative thematic analysis as described by Braun and Clarke [9], i.e., coding and then combining the codes into themes. In the first stage, one familiarizes oneself with the data and notes down initial ideas, which in our case corresponds to crafting the coding prompt. However, although our code groups could be considered as "themes", a full thematic analysis would go further into interpreting the themes and reporting the results with illustrative quotes. For the sake of objectivity, we avoid such interpretation, and only look at differences in group prevalence between datasets.

Data Quality Metrics. Using the code embedding vectors, we compute two standard metrics for generative model data. First, we compute Frechet Distances between the distributions of human and GPT-3 code embedding vectors that are reduced to 5 dimensions using UMAP. Frechet Distance is a commonly used metric in benchmarking image generators [28] and has later been also applied to text embeddings [73]. Second, we compute precision and recall metrics using the 5-dimensional code embeddings and the procedure of [37]. Intuitively, precision measures how large portion of the generated data samples lie close to real data, and recall measures how large portion of the real data is covered by the generated data. An ideal generator has both high precision and high recall. The metrics are visualized in Figure 5. For more reliable comparison, Figure 5 also includes additional results based on coding the game experience descriptions in addition to the "Why art?" questions. This additional coding prompt is included in the supplementary material.

Topic Similarities and Differences. We independently code and group the compared dataset, and sort the code groups based on their frequencies. We then use a circular graph (Figure 7) to visualize the sorted groups and the connections between datasets. The visualised connection strengths correspond to the cosine similarity of the full-dimensional mean normalized embedding vectors of the groups. We only included the davinci GPT-3 variant in this analysis, as it was the most human-like model based on the data quality metrics above.

Answer consistency. It is important that the separately queried answers continue an interview or a survey in a consistent manner. Our prompts are designed for this, as the previously generated answers by the same synthetic "participant" are included in the prompt for the next answer. Importantly, the "Why art?" prompt

Table 5: The prompt used for automatic qualitative coding. Manually coded few-shot examples are separated by "###". This prompt guides GPT-3 to summarize the essential information in the answers (it will produce a wide range of codes, not only repeat the ones shown in the few-shot examples). To minimize LLM repetition bias, the example answers were selected randomly from the real human dataset, while avoiding answers that result in same codes. For brevity, only 3 out of the total 10 few-shot examples are shown here. The full prompt can be found in supplementary material.

The following presents a qualitative coding of answers from a video game research study. The answers explain why a participant experienced a game as art. The codes summarize the given reasons as compactly as possible. If an answer lists multiple reasons, the corresponding codes are separated by semicolons.
###
Answer: The questions it raised and the highly emotional connection that emerged between me and the game, the experience.
Codes: raising questions; emotional connection
###
Answer: For a game experience to feel like a work of art to me, it would usually be an immersive experience that creates a real emotional response. Since games accomplish this through a combination of illustration, animation, sound, music, storytelling elements all together, I would consider these types of experiences art.
Codes: immersive experience; emotional response
###
Answer: The fact that each asset was hand drawn in such a unique style.
Codes: unique visual style
###
Answer: *<each coded answer inserted here during coding>*

(*PROMPT 3*, Table 1) always included a previously generated art experience description (generated using *PROMPT 1*, Table 1).

Previously, we have used GPT-3 to generate synthetic Likert-scale data for a psychological questionnaire (PANAS), by generating completions to questionnaire items one-by-one, always including the previous answers in the prompt for the next item generation [76]. The factorial structure that emerged from generating the data this way was similar to human data. This suggests that GPT-3 can take coherently into account the previous answers included in a prompt.

Our data is open-ended answers, which does not allow factor analysis. Instead, we measured consistency using text embeddings computed with the text-curie-001 model. Because both *PROMPT 1* and *PROMPT 3* probe different aspects of the same experience, consistently generated answers should exhibit at least some similarity, which we measured using cosine similarity of answer embedding vectors. Moreover, *PROMPT 1* and *PROMPT 3* responses should be more similar when taken from a single participant (intra-participant similarity) instead of two randomly chosen participants (inter-participant similarity). To investigate this, we computed and visualized both intra-participant and inter-participant similarities (Figure 6). When computing the means and standard errors, we used intra-participant similarities of all 178 participants, and inter-participant similarities of 178 pairs of randomly shuffled (permuted) pairs of participants. For more reliable results, the random permutation of participant pairs was repeated 5000 times and the means and standard errors in Figure 6 are averaged over these 5000 permutations.

Game Frequencies. Finally, we also count the frequencies of each mentioned game in the human and GPT-3 data. For the human data, we included responses from the same 178 participants that were included in the topic analyses. For the GPT-3 data, the frequency of games was counted from the 178 completions that were queried with prompts with the question "What is the title of the game?". The frequency of the games was manually counted from the data. If there was mentions of two or more games in the same response, these were counted as separate mentions. If it was clear that the response referred to the same game, small differences in responses were discarded (for example, breath of the wild was categorized as the same answer as The Legend of Zelda: Breath of the Wild). If the response did not include a specific game title or we could not find the game title to refer to a published game, the response was ignored. For brevity, we only report results from the davinci and text-davinci-002 variants. Note that game frequency analyzes do not utilize the automatic coding step described above.

6.2 Results

The results of this experiment can be summarized as:

- Highly similar groups/topics emerge from both real and GPT-3 data. Figure 7 shows how many of the most frequent groups in the human data correspond to groups that are also amongst the most frequent in the GPT-3 data, such as groups relating to aspects of story (most frequent in both human and GPT-3 data) and music (2nd most frequent in both human and GPT-3 data). The visualization of code embedding vectors in Figure 4 also indicates that coding both datasets results in largely similar codes.
- Table 6 (most frequently mentioned games) shows that real and GPT-3 data discuss some of the same games, like Journey, Bioshock, and Shadow of the Colossus. However, many games in the human data are missing from the GPT-3 generated data, suggesting that LLM generated synthetic data may have less diversity than real data. Only 17.3% of games in the human data are mentioned in GPT-3 davinci data (see supplementary material for details).
- Larger GPT-3 variants yield more human-like data (Figure 5). OpenAI does not disclose the exact sizes of the GPT-3 models available through its API, but the ada, babbage,

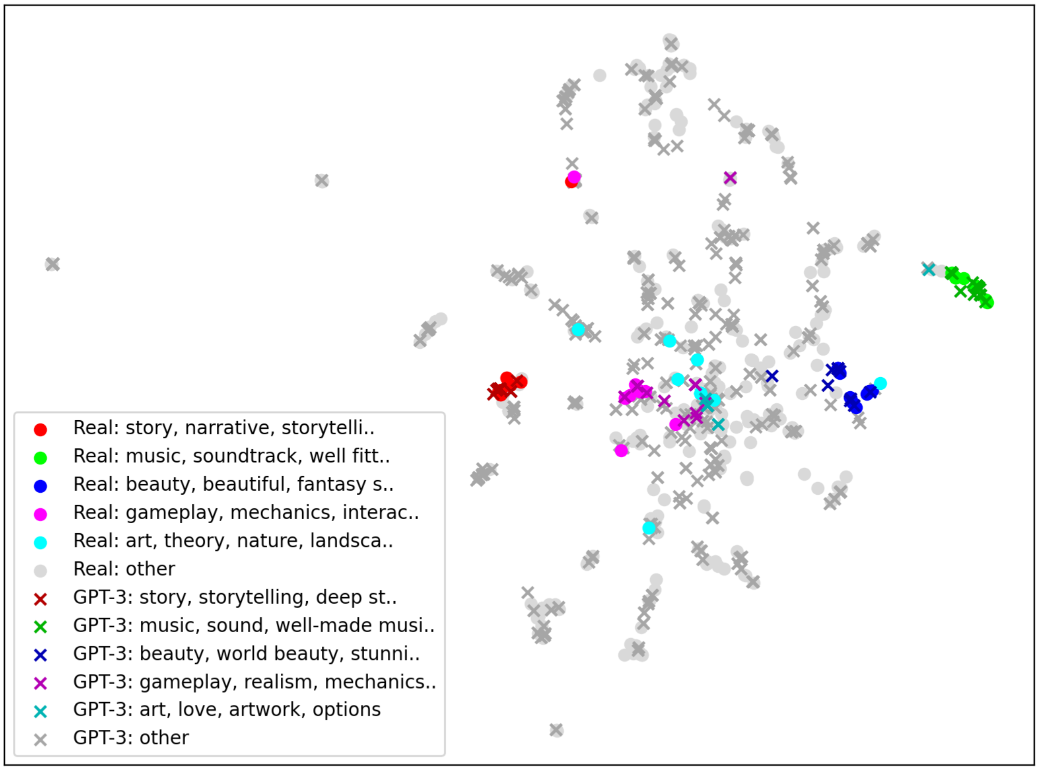

Figure 4: A scatterplot of 2D dimensionality-reduced code embeddings of both real and GPT-3 data. The colored markers show the real code groups with highest frequencies, and their closest GPT-3 equivalents. The visualization demonstrates how similar codes are located close to each other, and that similar codes and groups emerged from both datasets, i.e., there are no large clusters with only real or only GPT-3 data.

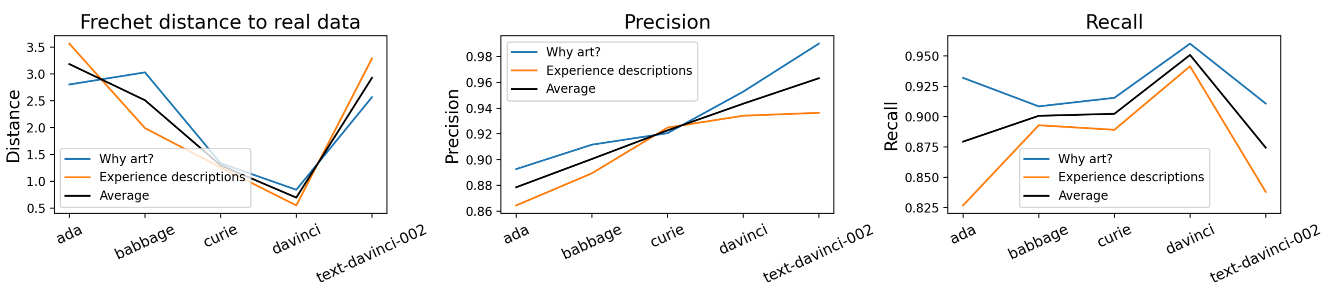

Figure 5: Frechet embedding distances (smaller is more human-like) and precision & recall metrics (larger is more human-like) for different GPT-3 variants. Overall, human-likeness grows with model size from ada to davinci. Curiously, text-davinci-002, the latest GPT-3 variant, shows improved precision but lower recall, i.e., the generated data is of high quality but has less diversity than real data or the older davinci variant.

curie, and davinci models have been inferred to correspond to the continuum of increasingly larger models evaluated in the original paper [24]. The ordering of the models also corresponds to increasing text generation cost, supporting the conclusion that ada is the smallest model and davinci is the largest one.

- The newest text-davinci-002 model has low recall and clearly less diversity than real data. This is evident in the lists of games mentioned, where 151 out of 178 answers discuss

Journey (only 7 mentions in the real data). Although OpenAI recommends this model as the default, our data suggests that it should be avoided for user modeling purposes, at least when one cares about data diversity.

- As visualized in Figure 6, both human data and all GPT-3 variants exhibit higher intra-participant than inter-participant answer similarity, indicating at least some degree of consistency in answering consecutive questions. For all of the data sources (ada, babbage, curie, davinci, text-davinci-002, human data), in all of the 5000 different inter-participant permutations, the mean inter-participant similarity was lower than the true mean intra-participant similarity. The overall slightly higher-than-human GPT-3 similarities, and the notably higher text-davinci-002 similarities probably reflect the data diversity issues noted above.

Based on the above, one can conclude that *investigating synthetic data can provide plausible answers to real research questions*, with some important caveats. First, real data can have more diversity. This is especially true when using the later text-davinci-002 model. The diversity problems of the model can be explained through its training procedure. The model is based on the InstructGPT series, which has been finetuned based on user feedback [58]. While this procedure does improve the average sample quality, it does not encourage diversity.

Second, synthetic data can also provide some misleading results, for example, many of the lower frequency groups in synthetic data do not have direct corresponding group in the human data. Some of the groups are hard to interpret (e.g. "unknown, unquantifiable, unfin.."), however, others might reasonably be expected to describe art experiences (e.g. "connection to other players"). When comparing the datasets, it should be noted that the sample of Bopp et al. [7] was not representative, thus the human dataset might also miss some themes that might arise in a more comprehensive sample.

7 DISCUSSION

In our three experiments, we have investigated the quality of GPT-3 answers to open-ended questions about experiencing video games as art. We can now summarize the answers to the research questions we posed in the introduction (for more details, see the results sections of each experiment):

Can one distinguish between GPT-3 generated synthetic question answers and real human answers? Our Experiment 1 suggest that GPT-3 can be capable of generating human-like answers to questions regarding subjective experiences with interactive technology, at least in our specific context. Surprisingly, our participants even responded "Written by a human participant" slightly more often for GPT-3 generated texts than for actual human written texts.

What kinds of errors does GPT-3 make? In Experiment 2, we identified multiple common failure modes. Some errors such as the model dodging a question could possibly be detected and the answers regenerated automatically, using a text classifier model or GPT-3 itself with a few-shot classification prompt. A particularly difficult error category is factual errors that cannot be detected based on superficial qualities of the generated text, but instead require domain knowledge about the discussed topics.

Can synthetic data provide plausible answers to real HCI research questions? What similarities and differences are there in GPT-3 and real data? Experiment 3 indicates that similar topics are discussed in both datasets, and that synthetic data can reveal plausible answers for research questions like "Why are games experienced as art?" and "What games do people experience as art?". However, although GPT-3 correctly discusses some of the same games as real participants, the GPT-3 data exhibits considerably less diversity (e.g., the "Journey bias" in our case). GPT-3 also discusses some topics not found in the real data, although some differences would be expected even between two sets of real human data, given the non-representative sample of Bopp et al. [6, 7].

Taken together, we find the results promising and intriguing, considering that even more capable models than GPT-3 have already appeared [e.g. 14]. LLM scaling laws predict that their performance will improve with new and even larger models [32, 66, 78], and the quality metrics of our Experiment 3 also indicate that larger scale yields more human-like data.

7.1 Use Cases for Synthetic Data

Regarding the possible uses for synthetic data, it is important to consider the trade-off between data quality, latency, and cost. GPT-3 -generated data is of lower quality than real data—at least if disregarding the problems of online crowdsourcing such as bots and careless, insincere or humorous responses—but GPT-3 also has very low latency and cost. The crucial question then becomes: *When can low cost and latency offset issues with quality?*

We believe the synthetic data can be useful in initial pilot research or experiment design where one explores possible research ideas or hypotheses or what people might say or write, before investing in real participant recruitment and data collection. The same should apply both to academic research and designers trying to understand their users. In such work, LLMs offer an alternative to other exploration tools such as web searches.

In comparison to web searches, LLMs have two primary benefits. First, *they can directly provide data in the same format as an actual study*. This allows using the synthetic data for pilot-testing and debugging of data analysis and visualization pipelines. Also, when pilot testing, seeing the data in the right format can arguably help the researcher explore the space of possibilities for the design of the actual study. Such exploration benefits from the low latency and cost of synthetic data collection, especially if combined with automatic data analysis similar to our Experiment 3. For example, reading the synthetic answers and inspecting the emerging codes and themes may give ideas for further questions to ask in an interview.

The second benefit over web searches is that *LLMs can generalize to new tasks and data*, as reviewed in Section 2.1. This suggests that LLMs may at least in some cases generate answers to questions that are not directly searchable from the training data. For instance, the real human data used in this paper was released in July 2020 [6], i.e., it cannot have been used in training the GPT-3 variants we tested, except for text-davinci-002. According to OpenAI documentation, text-davinci-002 training data ends in 2021 and the data of older variants ends in 2019. However, it is not currently possible to predict how well a model generalizes for a specific case, without actually testing.

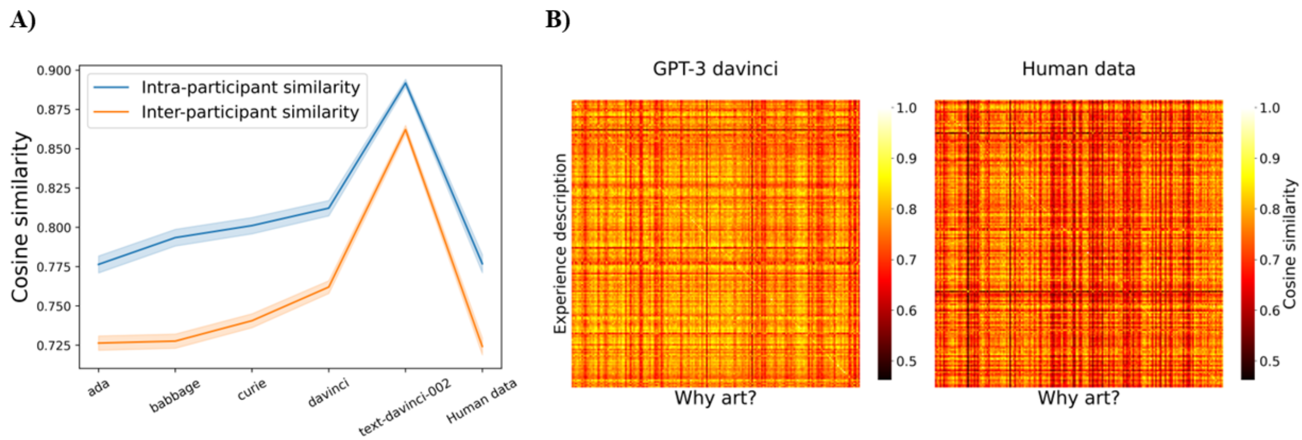

Figure 6: A) Mean intra-participant and inter-participant cosine similarities between the experience descriptions and "Why art?" answers. The shaded areas indicate standard errors of the mean. B) Cosine similarity matrices between the experience descriptions and "Why art?" answers of all 178 synthetic and human participants. The intra-participant similarities are on the matrix diagonals, whereas the off-diagonal elements display inter-participant similarities.

Table 6: Most common games in human, GPT-3 davinci, and GPT-3 text-davinci-002 data. The numbers in bold indicate how many times the game was mentioned in the data. The table shows all the games that were mentioned more than twice in the human data. The games with corresponding frequency ranks from the GPT-3 davinci and text-davinci-002 data are shown in the second and third column. Ties are sorted in alphabetical order.

Rank	Human data		GPT-3 davinci		GPT-3 text-davinci-002	
1.	*The Legend of Zelda: BOTW*	**10**	*Journey*	**44**	*Journey*	**151**
2.	*Journey*	**7**	*The Last of Us*	**12**	*Flower*	**5**
3.	*Nier: Automata*	**7**	*Dear Esther*	**8**	*That Dragon, Cancer*	**3**
4.	*Red Dead Redemption 2*	**6**	*Portal*	**7**	*Braid*	**2**
5.	*The Last of Us Part II*	**6**	*Bioshock*	**6**	*Shadow of the Colossus*	**2**
6.	*Firewatch*	**5**	*Shadow of the Colossus*	**5**	*Dreams of Geisha*	**1**
7.	*Hollow Knight*	**5**	*The Path*	**5**	*Final Fantasy VII*	**1**
8.	*Disco Elysium*	**4**	*Limbo*	**3**	*Flow*	**1**
9.	*Life Is Strange*	**4**	*Mirror's Edge*	**3**	*Frog Fractions*	**1**
10.	*Bioshock*	**3**	*The Stanley Parable*	**3**	*Halo 5: Guardians*	**1**
11.	*Shadow of the Colossus*	**3**	*Final Fantasy IX*	**2**	*Kingdom Hearts*	**1**
12.	*The Witcher 3*	**3**	*Final Fantasy VII*	**2**	*The Legend of Zelda: BOTW*	**1**
13.	*Undertale*	**3**	*Flower*	**2**	*Nier: Automata*	**1**
14. –>	*... and 97 other games*	**113**	*... and 65 other games*	**69**	*... and 10 other games*	**10**

To understand the limits and opportunities of generalization, consider that LLM text training data typically originates from a generative human thought process affected by multiple latent variables such as the communicated content and the writer's emotion, intent, and style. Now, assuming that the training dataset is too large to simply memorize, an efficient way to minimize the next token prediction error is to learn internal data representations and computational operations that allow mimicking the data-generating process.[3] For example, vector representations of words produced by language models can exhibit semantic-algebraic relations such as $woman - man = aunt - uncle$ [42, 51]; this allows subsequent computational operations to perform semantic manipulations. To minimize the average prediction error over all data, an LLM should prioritize representing and operating on latent variables that affect a large portion of the data. Hence, one should care less about particular facts that only affect a small subset of the data, but assign a high priority to commonly influential variables such as emotion, style, and political views. Fittingly, LLMs are known to make factual errors, but can generate text in many literary styles [8], perform sentiment analysis [64], generate human-like self-reports of emotion [76], and predict how political views affect voting behavior and which words people associate with members of different political

[3]Recall that in a deep multilayer neural network such as an LLM, each layer performs one step of a multi-step computational process, operating on the representations produced by the previous layer(s). The power of deep learning lies in the ability to automatically learn good representations [5, 40].

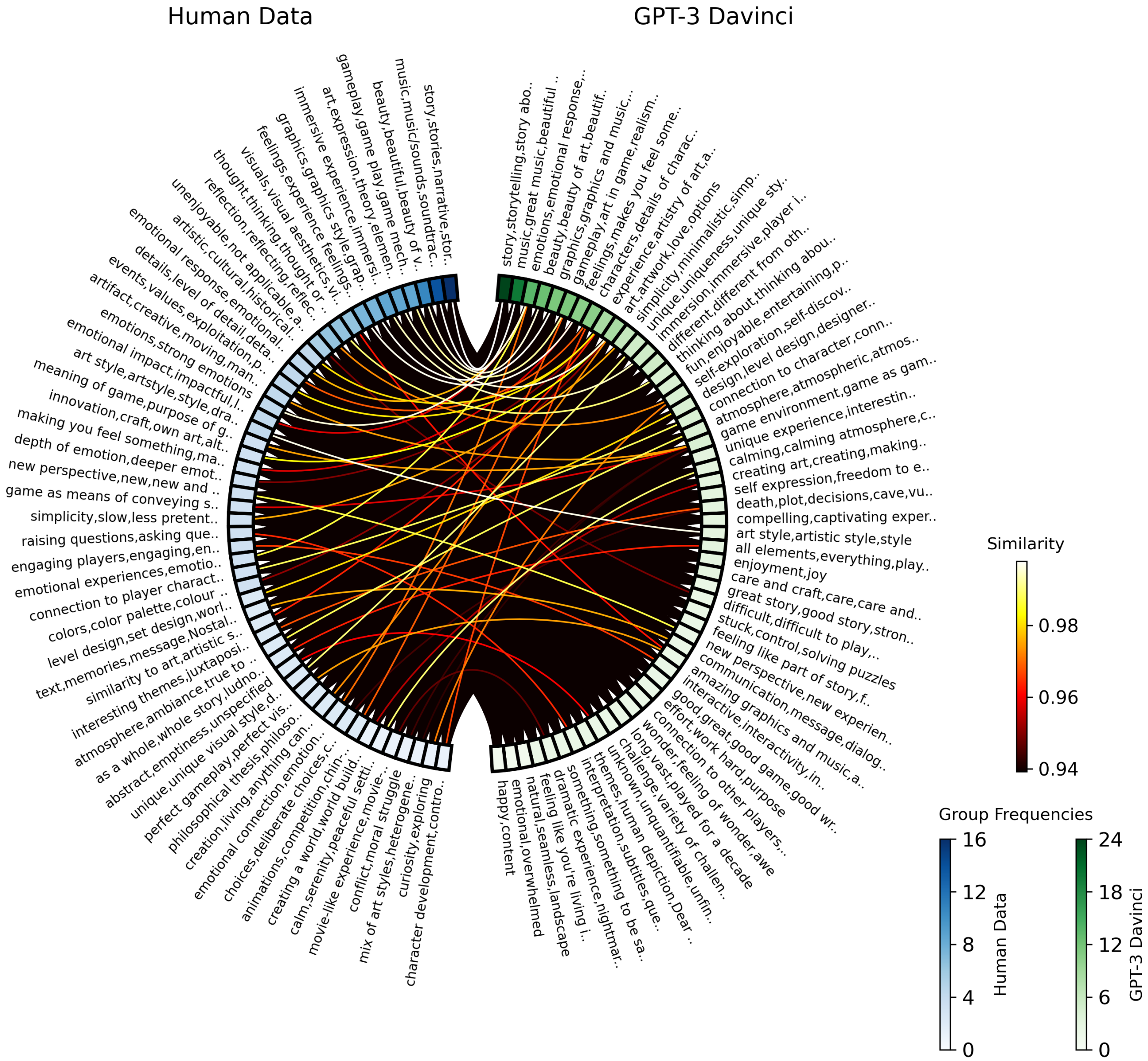

Figure 7: A circular graph presenting the human and GPT-3 davinci data resulting from the automatic qualitative coding. Each human data group is connected with a line to the most similar group in the GPT-3 data. The lines are color coded based on cosine similarity. The color coding and sorting of the group nodes is based in how frequent the groups were in the two datasets (groups with highest frequency on top). Here, group frequencies are reported as percentages.

parties [2]. LLM representations have also been observed to encode emotion and sentiment [36, 64].

Considering the above, a reasonable working assumption is that although LLMs can be expected to make factual errors when discussing interactive software or HCI artefacts—especially novel ones not included in training data—they may be useful in generating data about psychological latent variables such as user emotion and motivation in response to a hypothetical scenario described in the prompt, or about user experiences more generally, as in this paper.

Obviously, confirming hypotheses or arriving at conclusions about what people really think, feel, or need should only be done based on real data. LLM-based exploration could also steer interview questions in a more biased direction, which will subsequently reduce data quality in interviews with real users. On the other hand, other exploration techniques such as web searches or initial interviews with small participant samples can also be biased. More work is needed to test and evaluate LLMs in real research and design projects.

7.2 Misuse potential

Unfortunately, the quality requirements for synthetic data may be much lower for misuse than for actual research. In particular, GPT-3 and other LLMs may exacerbate the data quality problems of online crowdsourcing platforms. The reward incentives of such platforms encourage completing studies as fast as possible, in the extreme case by utilizing bots (e.g., [25]) and/or multiple accounts. Based on our experiments, it is clear that advanced language models can enable bots to generate more convincing questionnaire answers. Similarly, human participants might artificially increase their efficiency by generating answers to open-ended questions that may be slow to answer for real.

Now that GPT-3 is widely available outside the initially closed beta program, *there is a risk that online crowdsourcing of self-report data becomes fundamentally unreliable*. If the risk is realized, new tools are needed for detecting non-human data. Unfortunately, this is likely to become increasingly harder as language models advance. The risk may also imply a change in the cost-benefit analysis between research data sources. If one cannot anymore trust that online crowdsourced textual research data is from real humans, researchers may need to rely more on time-consuming and expensive laboratory studies than previously. If this is the case, LLM-generated fast and cheap synthetic data may become even more valuable for initial exploration and piloting.

7.3 GPT-3 and Emotions

As an incidental observation, the open question data of Experiment 1 suggests that there might be a common belief that human-written texts can be recognized based on the emotion that the text conveys. If this is the case, the belief is a probable contributor to the high rate of human evaluations for some of our GPT-3 texts. For example, many of the top human-like rated GPT-3 stimuli contain detailed accounts of how the player has felt during the gameplay, including specific emotional responses such as *"I felt his pain, his fear, and his struggles."* (see Table 3). Considering this, it is not surprising that these texts fooled the participants into thinking they were written by humans. If the belief that artificial intelligence cannot generate descriptions of experiences that the reader interprets as emotional is a more general phenomenon, it might be of importance when considering the risks related to language model misuse. For example, fake social media accounts that write about "their" emotional experiences might be perceived as more believable.

7.4 Future Directions

To mitigate the problems and better understand the biases of generated data, future efforts are needed in collecting reference human data together with extensive demographic information, and including the demographic information in the prompt to guide synthetic data generation. This would allow a more in-depth and nuanced inspection of the similarities and differences between real and human responses, including the ability of LLMs to portray different participant demographics. In initial tests, we have observed GPT-3 adapting its output based on participant age and gender given in the prompt, when generating synthetic answers to the question "What is your favorite video game and why?".

In addition to training larger and better models, data quality could be improved by using bias correction techniques such as the calibration approach of Zhao et al. [83], which does not require a slow and expensive retraining of a model. However, correct use of such technique also requires reference data—from a user modeling perspective, one should not try to remove the natural biases and imperfections of humans. On the other hand, problems of real human data such as social desirability bias and careless or humorous answers should be avoided in synthetic data. We hypothesize that with a sufficiently capable language model, this could be implemented by describing a virtual participant's motivations and attitudes as part of the prompt.

Although not yet explored in this paper, it might be possible to use LLMs to augment AI agents performing simulated user testing, which is currently focused on non-verbal data such as task difficulty or ergonomics [13, 30, 57]. LLMs could be integrated by generating textual descriptions of the test situation and agent behavior, and having the LLM generate *synthetic "think aloud"* descriptions of what the agent feels or thinks. This might greatly expand what kinds of data and insights simulated user testing can produce.

7.5 Limitations

The recruitment of Experiment 1 was not limited to native English speakers, and a sample with only native speakers might have different distinguishability scores. Additionally, as our sample was based on online crowdsourcing where the participants have at least an indirect incentive to respond fast, it is possible that laboratory studies would show different rates of distinguishability between human and GPT-3 generated texts.

We only examined one HCI context: art experiences in video games, thus the generalizability of our results is unclear. Future work should investigate the scope of possibilities for synthetic data more thoroughly, e.g., for what kinds of questions synthetic data is and is not helpful for. For instance, we only evaluated open-ended question answers and are working on expanding our study to the quantitative Likert-scale aesthetic emotion data that Bopp et al. also collected [6]. We also only tested one kind of prompt structure for generating the synthetic data. Although we used the same questions as in the reference human data—which is logical for evaluating human-likeness—we acknowledge that results may be sensitive to the wording of the other parts of the prompt. More research is needed on prompt design for HCI data generation, although we believe that our prompt design can be a promising starting point in many cases. Also, we did not investigate how useful researchers rate LLM use when designing new interview paradigms. This is an important direction for future studies.

Each of our three experiments probes a different aspect of the human-likeness and quality of GPT-3 generations. Although our experiments complement each other, they do not yet paint a complete picture—there is no all-encompassing definition of human-likeness, and the relevant features depend on context. Future studies should investigate how realistically synthetic data can represent participants from different demographic groups, and expand the evaluation of the human-likeness to other important features. Fortunately, new benchmarks and metrics are emerging for evaluating LLM biases [18, 19, 41].

Finally, one would often like to explore the reasons and explanations for an observed data distribution. Our "Why art?" question demonstrates that GPT-3 can be directly prompted for further insights in relation to previous generations (here: the experience descriptions). Naturally, the model cannot produce real causal explanations of why it generated something, it merely samples an explanation that is probable given the earlier generations included in the prompt. This is reminiscent of the research on Chain-of-Thought (CoT) prompting: An LLM can be prompted to provide step-by-step explanations for its "thought process", which can actually improve LLM reasoning capabilities [33]. The primary limitation is that generated explanations should be treated as hypotheses to be validated with real data, rather than trustworthy evidence. Our present experiments also focus on qualitative data—future work should explore collecting both quantitative and qualitative synthetic data, e.g., Likert-scale responses augmented with open-ended questions that probe the reasons.

8 CONCLUSION

We have explored and evaluated a general-purpose large language model (GPT-3) in generating synthetic HCI research data, in the form of open-ended question answers about experiencing video games as art. Our results indicate that GPT-3 responses can be very human-like in the best case, and can discuss largely similar topics as real human responses, although future work is needed to verify this with other datasets and research topics. On the other hand, GPT-3 responses can have less diversity than real responses, and contain various anomalies and biases. More research is needed on ways to prune anomalous responses and/or guide the model towards better and less biased responses.

Regarding use cases, we believe that LLMs can be useful in initial research exploration and pilot studies, especially as the models continue to improve. However, one must carefully consider the potential effects of the models' biases and confirm any gained insights with real data. As a downside, our results indicate that LLMs might make cheating in crowdsourcing platforms such as Amazon Mechanical Turk more prevalent and harder to detect. This poses a risk of crowdsourced self-report data becoming fundamentally unreliable.

OpenAI's GPT-3 is currently the largest and most capable publicly available language model. However, other technology companies have joined the race to train the best performing (and largest) generative language model. The past 12 months have seen the introductions of (among others) Microsoft's and NVIDIA's 530 billion parameter model Megatron-Turing NLG [75], DeepMind's 280 billion parameter model Gopher [66] and Google AI's 540-billion parameter model Pathways [14]. During this paper's review period, OpenAI also released a new and improved GPT-3 variant called ChatGPT [56]. The size of the models as well as the performance in numerous NLP benchmark tasks is increasing [14, 66, 75]. It will be intriguing to compare the present results to the generations of even more capable models in the future. However, the availability of the latest models is limited, as they are too large to run on consumer hardware or even on the computing infrastructures of most academic research labs. In practice, one may have to wait for the models to be released as cloud services, similar to GPT-3.

ACKNOWLEDGMENTS

This work has been supported by the European Commission through the Horizon 2020 FET Proactive program (grant agreement 101017779) and by Academy of Finland Grant #318937.

REFERENCES

[1] Alexander L Anwyl-Irvine, Jessica Massonnié, Adam Flitton, Natasha Kirkham, and Jo K Evershed. 2020. Gorilla in our midst: An online behavioral experiment builder. *Behavior research methods* 52, 1 (2020), 388–407. https://doi.org/10.3758/s13428-019-01237-x

[2] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. 2022. Out of One, Many: Using Language Models to Simulate Human Samples. *arXiv preprint arXiv:2209.06899* (2022). https://doi.org/10.48550/arXiv.2209.06899

[3] Javier A. Bargas-Avila and Kasper Hornbæk. 2011. Old Wine in New Bottles or Novel Challenges: A Critical Analysis of Empirical Studies of User Experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 2689–2698. https://doi.org/10.1145/1978942.1979336

[4] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel WH Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. 2019. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature biotechnology* 37, 1 (2019), 38–44. https://doi.org/10.1038/nbt.4314

[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 8 (2013), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

[6] Julia Bopp, Jan Benjamin Vornhagen, Roosa Piitulainen, Barbara Keller, and Elisa D. Mekler. 2020. GamesAsArt. (July 2020). https://doi.org/10.17605/OSF.IO/RYVT6 Publisher: OSF.

[7] Julia A. Bopp, Jan B. Vornhagen, and Elisa D. Mekler. 2021. "My Soul Got a Little Bit Cleaner": Art Experience in Videogames. *Proc. ACM Hum.-Comput. Interact.* 5, CHI PLAY, Article 237 (oct 2021), 19 pages. https://doi.org/10.1145/3474664

[8] Gwern Branwen. 2020. GPT-3 creative fiction. (2020). https://www.gwern.net/GPT-3

[9] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[11] Marc Brysbaert. 2019. How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language* 109 (2019), 104047. https://doi.org/10.1016/j.jml.2019.104047

[12] Erik Cambria and Bebo White. 2014. Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]. *IEEE Computational Intelligence Magazine* 9, 2 (2014), 48–57. https://doi.org/10.1109/MCI.2014.2307227

[13] Noshaba Cheema, Laura A. Frey-Law, Kourosh Naderi, Jaakko Lehtinen, Philipp Slusallek, and Perttu Hämäläinen. 2020. Predicting Mid-Air Interaction Movements and Fatigue Using Deep Reinforcement Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376701

[14] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin

Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. arXiv:2204.02311 [cs.CL]

[15] Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2021. Transformers as Soft Reasoners over Language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* (Yokohama, Yokohama, Japan) *(IJCAI'20)*. Article 537, 9 pages. https://doi.org/10.24963/ijcai.2020/537

[16] Matt Cox. 2019. This AI text adventure generator lets you do anything you want. https://www.rockpapershotgun.com/this-ai-text-adventure-generator-lets-you-do-anything-you-want

[17] Jonas Degrave. 2022. Building A Virtual Machine inside ChatGPT. https://www.engraved.blog/building-a-virtual-machine-inside/

[18] Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1693–1706. https://doi.org/10.18653/v1/2022.naacl-main.122

[19] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 862–872. https://doi.org/10.1145/3442188.3445924

[20] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341* (2020). https://doi.org/10.48550/ARXIV.2005.00341

[21] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread* (2021). https://transformer-circuits.pub/2021/framework/index.html.

[22] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (Portland, Oregon) *(KDD'96)*. AAAI Press, 226–231.

[23] Paul M Fitts. 1954. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of experimental psychology* 47, 6 (1954), 381–391.

[24] Leo Gao. 2021. *On the Sizes of OpenAI API Models.* Retrieved 2022-09-12 from https://blog.eleuther.ai/gpt3-model-sizes/

[25] Marybec Griffin, Richard J Martino, Caleb LoSchiavo, Camilla Comer-Carruthers, Kristen D Krause, Christopher B Stults, and Perry N Halkitis. 2022. Ensuring survey research data integrity in the era of internet bots. *Quality & quantity* 56, 4 (2022), 2841–2852. https://doi.org/10.1007/s11135-021-01252-1

[26] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022). https://doi.org/10.48550/ARXIV.2203.05794

[27] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2022. Neural Language Models as What If? -Engines for HCI Research. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22 Companion)*. Association for Computing Machinery, New York, NY, USA, 77–80. https://doi.org/10.1145/3490100.3516458

[28] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf

[29] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

[30] Jussi Jokinen, Aditya Acharya, Mohammad Uzair, Xinhui Jiang, and Antti Oulasvirta. 2021. Touchscreen Typing As Optimal Supervisory Control. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Mark Drucker (Eds.). ACM, 720:1–720:14. https://doi.org/10.1145/3411764.3445483

[31] Daniel Kahneman and Amos Tversky. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, 99–127.

[32] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020). https://doi.org/10.48550/ARXIV.2001.08361

[33] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/2205.11916

[34] Per Ola Kristensson. 2018. Statistical Language Processing for Text Entry. In *Computational Interaction*. Oxford University Press, 43–64.

[35] Victor Kuperman, Aki-Juhani Kyröläinen, Vincent Porretta, Marc Brysbaert, and Sophia Yang. 2021. A lingering question addressed: Reading rate and most efficient listening rate are highly similar. *Journal of Experimental Psychology: Human Perception and Performance* 47, 8 (2021), 1103–1112. https://doi.org/10.1037/xhp0000932

[36] Mijin Kwon, Tor Wager, and Jonathan Phillips. 2022. Representations of emotion concepts: Comparison across pairwise, appraisal feature-based, and word embedding-based similarity spaces. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 44.

[37] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2019. Improved Precision and Recall Metric for Assessing Generative Models. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/0234c510bc6d908b28c70ff313743079-Paper.pdf

[38] Nils C. Köbis, Barbora Doležalová, and Ivan Soraperra. 2021. Fooled twice: People cannot detect deepfakes but think they can. *iScience* 24, 11 (2021), 103364. https://doi.org/10.1016/j.isci.2021.103364

[39] Guillaume Lample and François Charton. 2019. Deep Learning For Symbolic Mathematics. In *International Conference on Learning Representations*.

[40] Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. 2020. Contrastive Representation Learning: A Framework and Review. *IEEE Access* 8 (2020), 193907–193934. https://doi.org/10.1109/ACCESS.2020.3031549

[41] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*. PMLR, 6565–6576.

[42] Shusen Liu, Peer-Timo Bremer, Jayaraman J Thiagarajan, Vivek Srikumar, Bei Wang, Yarden Livnat, and Valerio Pascucci. 2017. Visual exploration of semantic relationships in neural word embeddings. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 553–562.

[43] Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J Michaud, Max Tegmark, and Mike Williams. 2022. Towards Understanding Grokking: An Effective Theory of Representation Learning. In *Advances in Neural Information Processing Systems*.

[44] Róisín Loughran and Michael O'Neill. 2017. Application Domains Considered in Computational Creativity.. In *ICCC*. 197–204.

[45] Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. 2022. Frozen Pretrained Transformers as Universal Computation Engines. In *Proc. AAAI 2022*. 7628–7636. https://doi.org/10.1609/aaai.v36i7.20729

[46] I Scott MacKenzie and William Buxton. 1992. Extending Fitts' law to two-dimensional tasks. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 219–226.

[47] Neil A. Macmillan. 2005. *Detection theory : a user's guide* (2nd ed.). Lawrence Erlbaum Associates, Mahwah, N.J.

[48] Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R Eguchi, Po-Ssu Huang, and Richard Socher. 2020. Progen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497* (2020).

[49] Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* 2, 11 (2017), 205.

[50] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[51] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*. 746–751.

[52] Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. 2014. On the Number of Linear Regions of Deep Neural Networks. *Advances in Neural Information Processing Systems* 27 (2014), 2924–2932.

[53] Sophie J. Nightingale and Hany Farid. 2022. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences* 119, 8 (2022), e2120481119. https://doi.org/10.1073/pnas.2120481119 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2120481119

[54] Maxwell Nye, Michael Tessler, Josh Tenenbaum, and Brenden M Lake. 2021. Improving Coherence and Consistency in Neural Sequence Models with Dual-System, Neuro-Symbolic Reasoning. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 25192–25204. https://proceedings.neurips.cc/paper/2021/file/d3e2e8f631bd9336ed25b8162aef8782-Paper.pdf

[55] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context Learning and Induction Heads. *Transformer Circuits Thread* (2022). https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

[56] OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt/

[57] Antti Oulasvirta. 2019. It's time to rediscover HCI models. *Interactions* 26, 4 (2019), 52–56. https://doi.org/10.1145/3330340

[58] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. https://doi.org/10.48550/ARXIV.2203.02155

[59] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–18.

[60] Jorge Pérez, Pablo Barceló, and Javier Marinkovic. 2021. Attention is Turing-Complete. *J. Mach. Learn. Res.* 22, 75 (2021), 1–35.

[61] Ingrid Pettersson, Florian Lachner, Anna-Katharina Frison, Andreas Riener, and Andreas Butz. 2018. A Bermuda Triangle? A Review of Method Application and Triangulation in User Experience Evaluation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3173574.3174035

[62] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. 2016. Exponential expressivity in deep neural networks through transient chaos. *Advances in neural information processing systems* 29 (2016), 3360–3368.

[63] Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177* (2022).

[64] Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444* (2017).

[65] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[66] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. arXiv:2112.11446 [cs.CL]

[67] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67. http://jmlr.org/papers/v21/20-074.html

[68] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8821–8831. https://proceedings.mlr.press/v139/ramesh21a.html

[69] Shaghayegh Roohi, Asko Relas, Jari Takatalo, Henri Heiskanen, and Perttu Hämäläinen. 2020. Predicting Game Difficulty and Churn Without Players. In *CHI PLAY '20: The Annual Symposium on Computer-Human Interaction in Play, Virtual Event, Canada, November 2-4, 2020*, Pejman Mirza-Babaei, Victoria McArthur, Vero Vanden Abeele, and Max Birk (Eds.). ACM, 585–593. https://doi.org/10.1145/3410404.3414235

[70] Ronald Rosenfeld. 2000. Two decades of statistical language modeling: Where do we go from here? *Proc. IEEE* 88, 8 (2000), 1270–1278.

[71] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Scao, Arun Raja, et al. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. (2022).

[72] Jeff Sauro and James R Lewis. 2016. *Quantifying the user experience: Practical statistics for user research.* Morgan Kaufmann.

[73] Stanislau Semeniuta, Aliaksei Severyn, and Sylvain Gelly. 2018. On Accurate Evaluation of GANs for Language Generation. https://doi.org/10.48550/ARXIV.1806.04936

[74] Tom Simonite. 2021. It Began as an AI-Fueled Dungeon Game. It Got Much Darker. https://www.wired.com/story/ai-fueled-dungeon-game-got-much-darker/

[75] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. 2022. Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model. arXiv:2201.11990 [cs.CL]

[76] Mikke Tavast, Anton Kunnari, and Perttu Hämäläinen. 2022. Language Models Can Generate Human-Like Self-Reports of Emotion. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22 Companion)*. Association for Computing Machinery, New York, NY, USA, 69–72. https://doi.org/10.1145/3490100.3516464

[77] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[78] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* (2022).

[79] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. https://doi.org/10.48550/ARXIV.2201.11903

[80] Gui-Rong Xue, Jie Han, Yong Yu, and Qiang Yang. 2009. User language model for collaborative personalized search. *ACM Transactions on Information Systems (TOIS)* 27, 2 (2009), 1–28.

[81] Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. Crossfit: A few-shot learning challenge for cross-task generalization in nlp. *arXiv preprint arXiv:2104.08835* (2021).

[82] Camilla Zallot, Gabriele Paolacci, Jesse Chandler, and Itay Sisso. 2021. Crowdsourcing in observational and experimental research. *Handbook of Computational Social Science, Volume 2: Data Science, Statistical Modelling, and Machine Learning Methods* (2021), 140–157.

[83] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-shot Performance of Language Models. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 12697–12706. https://proceedings.mlr.press/v139/zhao21c.html