

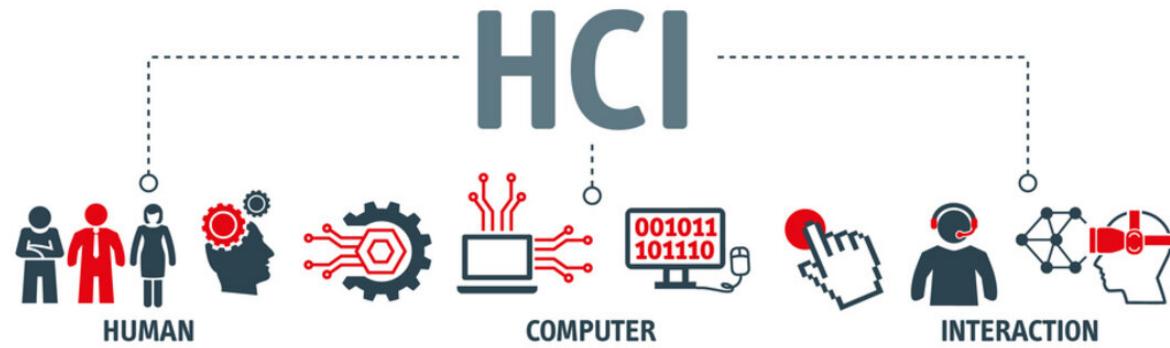
بِسْمِ اللّٰهِ الرَّحْمٰنِ الرَّحِيْمِ

# Information Technology Fundamentals

Mohammad Hossein Manshaei

[manshaei@gmail.com](mailto:manshaei@gmail.com)





# Human Computer Interaction

## Modul 3: Part 2

# Contents

## ➤ Human-Computer Integration

- ✓ The Eras of Computing
- ✓ Main Idea of HInt

## ➤ Low Vision

- ✓ Designing AR Visualization
- ✓ Visualizations For Projection-Based AR
- ✓ Visualizations for Smartglasses

## ➤ EarBuddy

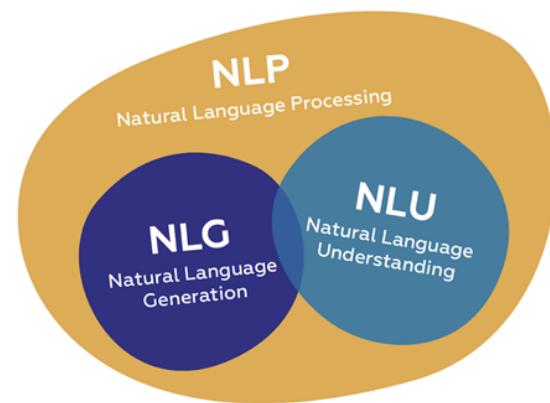
- ✓ Gesture Selection, Data collection, Usability Evaluation, Potential Application

## ➤ LLM & HCI

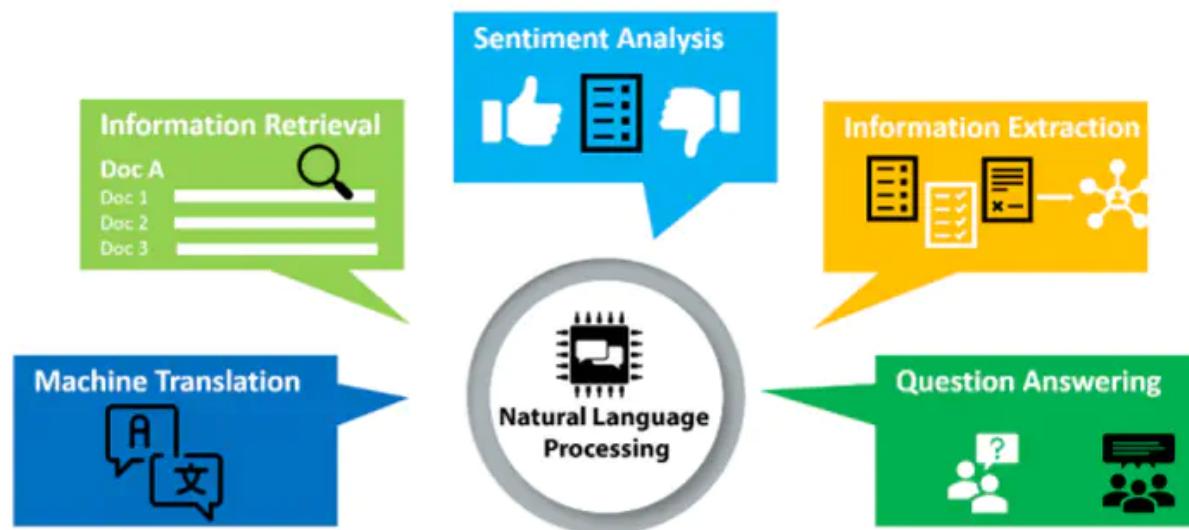
- ✓ Large Language Model
- ✓ LLM in Generating Synthetic HCI Data
- ✓ LLM-Driven Analysis in HCI Research

# What is NLP?

- Natural-language processing – computer activity in which computers are entailed to analyze, understand, alter, or generate natural language.
- Components of NLP: NLP is divided into two components.
  - Natural Language Understanding
  - Natural Language Generation



# Applications of NLP



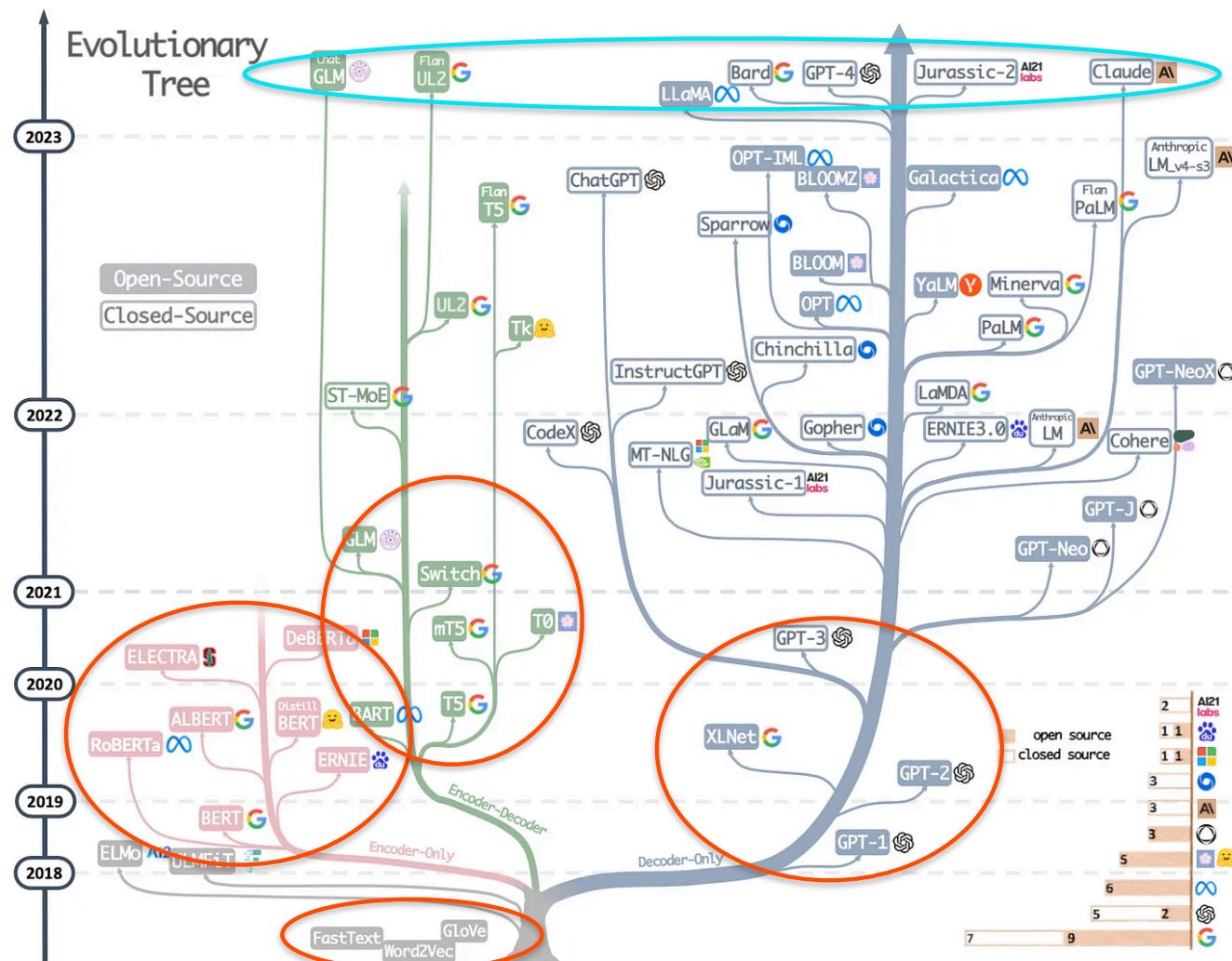
# Language models

- A language model in NLP is a probabilistic statistical model that determines the probability of a given sequence of words occurring in a sentence based on the previous words. It helps to predict which word is more likely to appear next in the sentence
- A large language model is an advanced type of language model that is trained using deep learning techniques on massive amounts of text data.
  - These models are capable of generating human-like text and performing various natural language processing tasks.

input/feature #1      input/feature #2      output/label  
Thou      shalt      \_\_\_\_\_

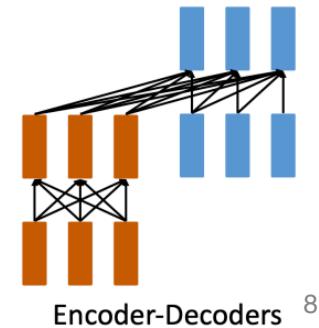
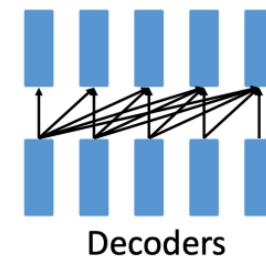
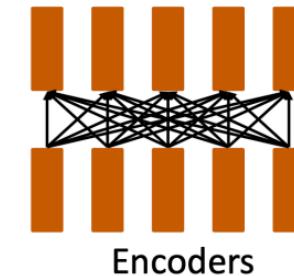


# Language models



# Transformer Architectures

- Encoder-only (e.g., BERT): bidirectional contextual embeddings
- Decoder-only (e.g., GPT-x): unidirectional contextual embeddings, generate one token at a time
- Encoder-decoder (e.g., T5): encode input, decode output



Encoder-Decoders 8

# Contents

## ➤ Human-Computer Integration

- ✓ The Eras of Computing
- ✓ Main Idea of HInt

## ➤ Low Vision

- ✓ Designing AR Visualization
- ✓ Visualizations For Projection-Based AR
- ✓ Visualizations for Smartglasses

## ➤ EarBuddy

- ✓ Gesture Selection, Data collection, Usability Evaluation, Potential Application

## ➤ LLM & HCI

- ✓ Large Language Model
- ✓ LLM in Generating Synthetic HCI Data
- ✓ LLM-Driven Analysis in HCI Research

# Reference

- Hämäläinen, P., Tavast, M. and Kunnari, A., 2023, April. **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* (pp. 1-19).

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

Perttu Hämäläinen\*  
perttu.hämäläinen@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 questions about HCI research data, ranging from games as art, a topic not treatable with traditional computational approaches. 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. This 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, 39 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?

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



This work is licensed under a Creative Commons Attribution International 4.0 License.

CHI '23, April 23–28, 2023, Hamburg, Germany  
© 2023 Copyright held by the owner/author(s).  
ACM ISBN 978-1-4503-9321-5/23/04.  
<https://doi.org/10.1145/3544548.3580688>

AI: My favorite writing tool would have to be Microsoft Word. I find it is 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<sup>1</sup>, 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. The result is characteristic of the phenomenon we investigate in this paper: Through learning to model and predict the kinds of natural language texts ranging from formal documents to online discussions, large 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 neuropsychological 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 CommonCrawl [27], which means they also learn to predict the words 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 corpus. Unlike traditional search engines, LLMs can be queried for the form of a narrative or as a fictional image. Furthermore, LLMs exhibit at least some generalization capability to new tasks and data (e.g., [45, 51, 81]). This presents an untapped opportunity for counterfactual *What if?* exploration, e.g., allowing

<sup>1</sup><https://beta.openai.com/playground>

## The prompts used to generate the GPT-3 data

### 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:

## Do LLMs present a new tool for HCI research?

- Three experiments have been proposed to assess the human-likeness of data generated by LLMs, 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)

## Condition of Generating the Synthetic Data

- 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.

## Condition of Generating the Synthetic Data (cont.)

- 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 used in Experiment 3, we used temperature=1.0 instead of 0.7,
- With temperature=1.0, the token sampling probabilities directly correspond to those learned from the training data.
- **Dataset:** Bopp et al. dataset is about real human participant responses to questions about art experiences in video games

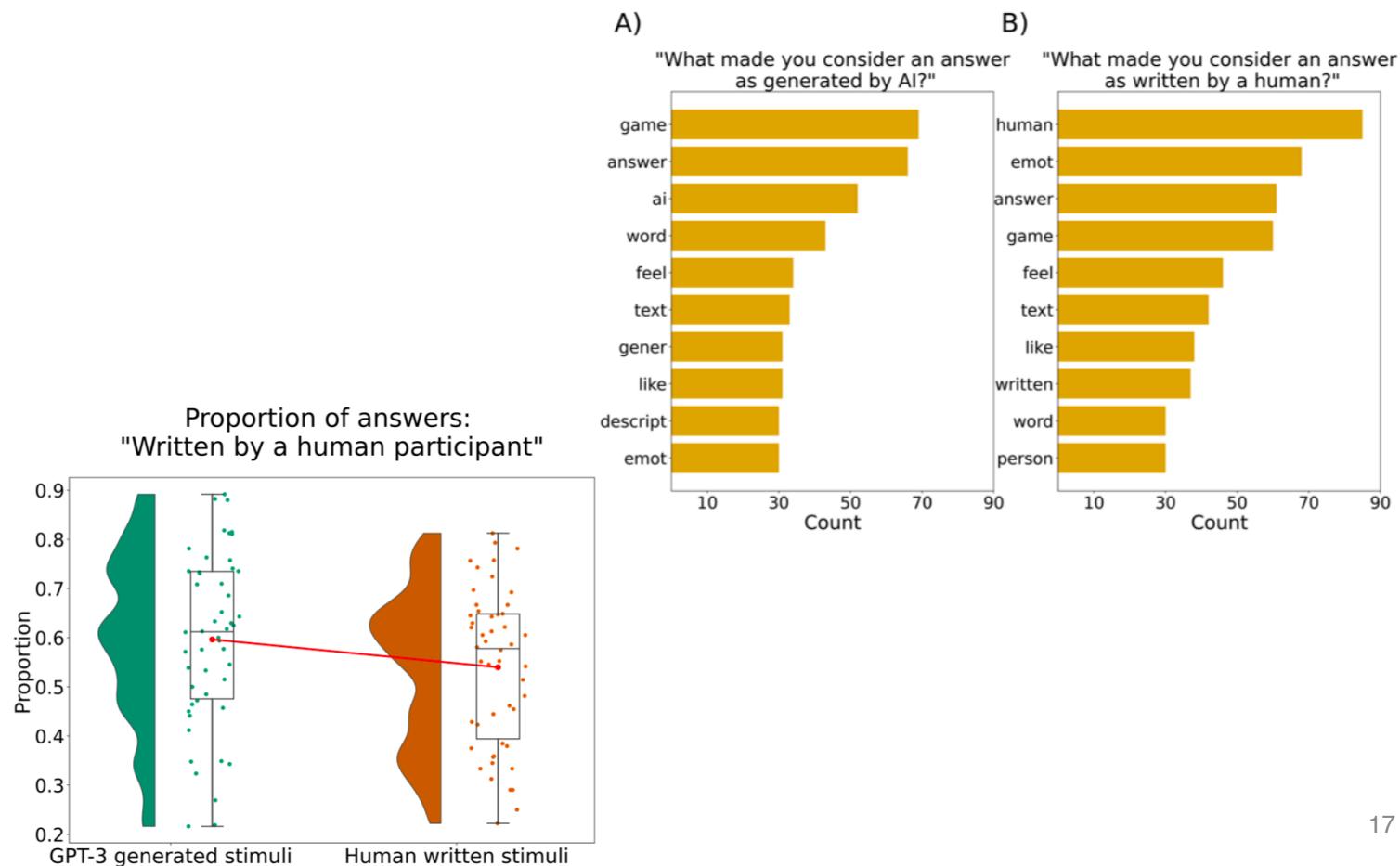
# Participants and Stimuli

- **175 (155) adult participants:** 55.48% men, 43.23% women (under 35 years old)
- 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.
- Another a scale from 1 (I have never heard the NLP term before) to 5 (I am an AI 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
- 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 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<sup>15</sup> words (SD: 60.68, median: 64.0).

## **EI: DISTINGUISHING BETWEEN GPT-3 AND REAL DATA**

- A first experiment provides a quantitative study of how distinguishable GPT-3 are from real human responses.
- 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.
- After the 20 evaluations, the participants were asked to answer two open questions regarding their decision process:
  - "What made you consider an answer as written by a human?"
  - "What made you consider an answer as generated by AI?".

# EI: DISTINGUISHING BETWEEN GPT-3 AND REAL DATA



## EI: DISTINGUISHING BETWEEN GPT-3 AND REAL DATA

- On aggregate, human written texts were correctly recognized 54.45% of the time, with 95% confidence interval (95% CI).
- The average accuracy of recognizing GPT-3 generated texts as AI-written was below chance level 40.45% (95% CI).
- 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".

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

## EI: DISTINGUISHING BETWEEN GPT-3 AND REAL DATA

- 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.
- 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.

## **E2: WHAT KINDS OF ERRORS DOES GPT-3 MAKE?**

- 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.

## **E2: WHAT KINDS OF ERRORS DOES GPT-3 MAKE?**

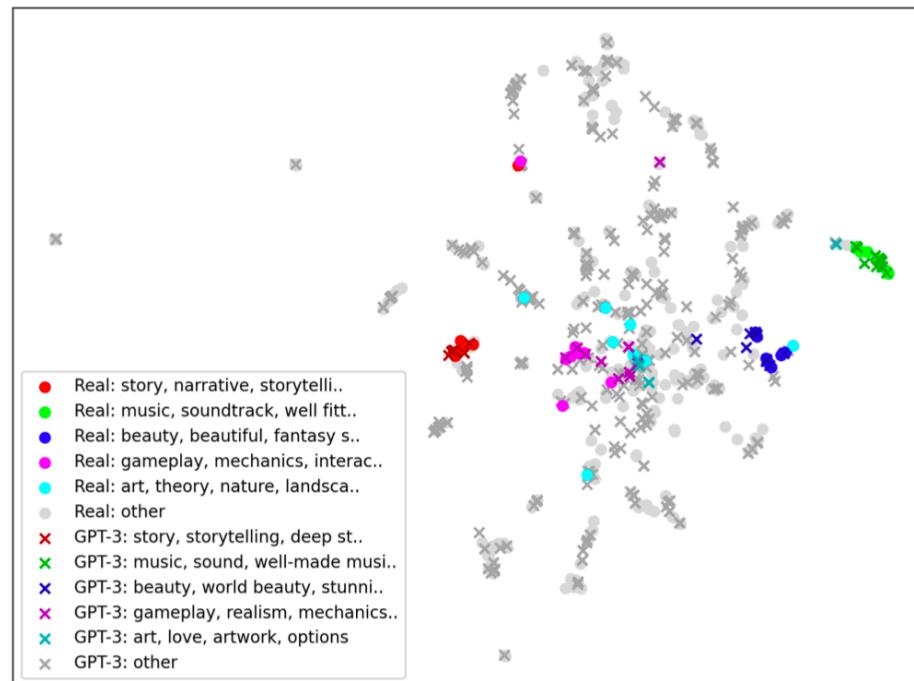
- We identified 8 distinct types of anomalies plus an "other" category.
  - Too long or rambling incoherently (27.5%).
  - Not answering the question or dodging the question (20.5%).
  - Incomplete: answer ends before getting to the point (15.2%).
  - Contradicts the positive answer in the prompt by claiming that has not experienced the game as art (8.2%).
  - Imaginary game or gameplay (7.6%)
  - Talks about art more generally, not about a game (4.1%).
  - Sentences that clearly contradict each other (3.5%).
  - Names a film or other type of media instead of a game (1.8%).

## **E3: DIFFERENCES BETWEEN REAL AND GPT-3 CONTENT**

- We 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:
- **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?")

## E3: DIFFERENCES BETWEEN REAL AND GPT-3 CONTENT

- A scatterplot of 2D dimensionality-reduced code embeddings of both real and GPT-3 data.



## E3: DIFFERENCES BETWEEN REAL AND GPT-3 CONTENT

- 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.

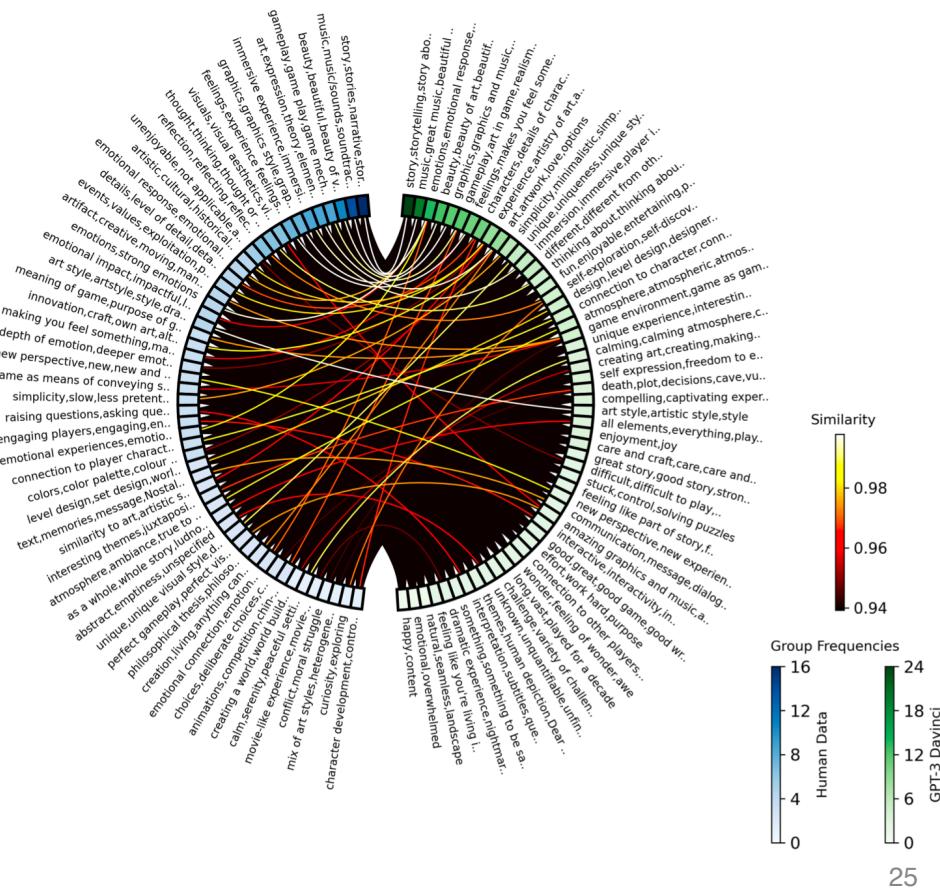
Rank	Human data		GPT-3 davinci		GPT-3 text-davinci-002	
1.	<i>The Legend of Zelda: BOTW</i>	<b>10</b>	<i>Journey</i>	<b>44</b>	<i>Journey</i>	<b>151</b>
2.	<i>Journey</i>	7	<i>The Last of Us</i>	<b>12</b>	<i>Flower</i>	5
3.	<i>Nier: Automata</i>	7	<i>Dear Esther</i>	<b>8</b>	<i>That Dragon, Cancer</i>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

# A circular graph presenting the human and GPT-3 davinci data

Human Data

GPT-3 Davinci

- 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.



# Contents

## ➤ Human-Computer Integration

- ✓ The Eras of Computing
- ✓ Main Idea of HInt

## ➤ Low Vision

- ✓ Designing AR Visualization
- ✓ Visualizations For Projection-Based AR
- ✓ Visualizations for Smartglasses

## ➤ EarBuddy

- ✓ Gesture Selection, Data collection, Usability Evaluation, Potential Application

## ➤ LLM & HCI

- ✓ Large Language Model
- ✓ LLM in Generating Synthetic HCI Data
- ✓ **LLM-Driven Analysis in HCI Research**

# Reference

- Torii, M.G., Murakami, T. and Ochiai, Y., 2024. Expanding Horizons in HCI Research Through LLM-Driven Qualitative Analysis. *arXiv preprint arXiv:2401.04138*.

arXiv:2401.04138v1 [cs.HC] 7 Jan 2024

## Expanding Horizons in HCI Research Through LLM-Driven Qualitative Analysis

MAYA GRACE TORII,

R&D Center for Digital Nature, University of Tsukuba, Japan

TAKAHITO MURAKAMI,

R&D Center for Digital Nature, University of Tsukuba, Japan

YOICHI OCHIAI, R&D Center for Digital Nature, University of Tsukuba, Japan

How would research be like if we still needed to "send" papers typed with a typewriter? Our life and research environment have continually evolved, often accompanied by controversial opinions about new methodologies. In this paper, we embrace this change by introducing a new approach to qualitative analysis in HCI using Large Language Models (LLMs). We detail a method that uses LLMs for qualitative data analysis and present a quantitative framework using SBART cosine similarity for performance evaluation. Our findings indicate that LLMs not only match the efficacy of traditional analysis methods but also offer unique insights. Through a novel dataset and benchmark, we explore LLMs' characteristics in HCI research, suggesting potential avenues for further exploration and application in the field.

CCS Concepts: • Human-centered computing → Human computer interaction (HCI); HCI design and evaluation methods.

Additional Key Words and Phrases: Large Language Models (LLMs), qualitative data analysis, LLM in HCI research, methodology

### 1 INTRODUCTION

As HCI researchers, our expertise lies in exploring human-computer interactions, not in laborious qualitative data coding. Qualitative research in Human-Computer Interaction (HCI) has traditionally been constrained by the extensive time, effort, and financial resources required for coding, categorizing, and interpreting large volumes of data. Traditional methods, despite their rigor, often face challenges in scaling and maintaining consistency across diverse contexts. The advent of Natural Language Processing (NLP) technologies has been offering some relief by automating aspects of these processes assisted in creating codebooks or categorizing data [1].

A significant shift towards leveraging Large Language Models (LLMs) has been observed since the introduction of ChatGPT in 2022. The inherent strength of LLMs in processing and providing insights on substantial language data suggests a promising avenue for HCI research [11]. Studies have begun exploring the potential of LLMs in aiding qualitative analysis, particularly in tasks like codebook and categorizing [5, 23]. However, these approaches often see NLP and LLMs as mere supplements to human effort, focusing on data compilation rather than interpretation, indicating an opportunity for further automation. As an exception, Byun et al.(2023) have made initial forays into using LLMs for generating discussions from raw qualitative data, although their approach lacks quantitative evaluation [6].

In this paper, we propose a method for conducting qualitative analysis in HCI research using LLMs and evaluate the output by comparing it with the published paper. We aim to not only process and analyse datasets but also to enhance the reproducibility and scalability of qualitative research. This approach overcome the limitations of human-intensive methods, offering a pathway to handle vast amounts of information more efficiently. Specifically, we explored

Authors' addresses: Maya Grace Torii,  
R&D Center for Digital Nature, University of Tsukuba, 1-2 Kasuga, Tsukuba, Ibaraki, Japan, 305-0018, toriparu@diginnature.slis.tsukuba.ac.jp; Takahito  
Murakami, takahito@diginnature.slis.tsukuba.ac.jp.  
R&D Center for Digital Nature, University of Tsukuba, 1-2 Kasuga, Tsukuba, Ibaraki, Japan, 305-0821; Yoichi Ochiai, R&D Center for Digital Nature,  
University of Tsukuba, 1-2 Kasuga, Tsukuba, Ibaraki, Japan, 305-0821, wizard@slis.tsukuba.ac.jp.

Manuscript submitted to ACM

1

# **Exploring Human-Computer Interactions**

- We propose a method for conducting qualitative analysis in HCI research using LLMs and evaluate the output by comparing it with the published paper.
- To execute the experiment, three prompt system were prepared,
  - analysis prompt,
  - summarise for QA prompt,
  - question generator prompt

## **Exploring Human-Computer Interactions (cont.)**

- This experimental setup includes the use of two different LLMs:
  - GPT-4 of openAI and Llama 13B from meta.
  - To assess the capabilities of Large Language Models (LLMs) in qualitative data analysis, we have chosen the Sentence- BART(SBART) as our primary tool.

# **Dataset for Quality Analysis**

- Our study created a dataset comprising three CHI papers. These papers were selected based on the presence of qualitative evaluation related to open questions. We specifically focus on sections of these papers that describe the findings of open-question evaluations. The dataset we created includes:
  - Summary: formatted summaries of the selected CHI papers.
  - Raw data: data pertaining to the open-ended questions and their responses.
  - Chosen passages from papers: set of passages that contains qualitative analysis findings and the preceding paragraph.

# Qualitative Analysis LLMs System

- Three prompts used in the qualitative analysis LLMs system and the evaluation system.

