Wikipedia vs ChatGPT: A study on their Use as Tools for Topic Discovery and Learning

Alexandra Herghelegiu
Wikipedia vs ChatGPT
For learning
• No statistically significant difference in the students’ performance in the knowledge tests

• ChatGPT:
  • Better user engagement
  • More personalised content
  • Tended to repeat itself

• Wikipedia:
  • Immediate overview of the concepts (table of content)
  • Filtering content was difficult

• participants perceived their confidence of topic understanding to be higher after using ChatGPT

• however less confident answers when faced with the knowledge test after using ChatGPT
Does Predictive Text Affect the Quality of Writing

Cameron Round

Part II

cr667@cam.ac.uk
What I did

- Image captioning task
- Setwise comparison
The results

Expert comparison

Individual's standard deviations

Mean standard deviations TTest: $p = 0.087$
Mean with assists: 3.94  Mean without assists: 4.97
Scores TTest: $p = 0.03$

Wider comparison

Caption scores

Levene test: $p = 0.048$
Mean with assists: 175  Mean without assists: 185
Impact of LLM Hallucinations on Travel Advice: Entertaining and Less Reliable

Chang Liu
Research Question

Attraction
motivates participants’ desire to visit a place. [1]

Reliability
key component building trust between users and a service. [2]

Hallucination
Salty Ice Cream turns taste buds into playful seals dancing on icebergs!

Travel Advice

Results

Attraction - hallucinated probability
Null hypothesis: the likelihood of choosing hallucinated and real travel advice is equal.

Fail to reject the null hypothesis, as difference is not statistically significant (one sample t-test p = 0.80).

Reliability - credibility score
2 rounds of participant ratings of perceived reliability of travel advice: ‘before’ and ‘after’ recognising hallucinations

Significant decrease from initial 3.8 to 2.0. (p = 6.87e-6)
Sometimes Tell Me, Sometimes Ask Me: Comparing Logical Discernment using AI Systems that Intelligently Frame Explanations and AI Systems with Causal Explanations

A Replication Study of:

Don’t Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations

[Danry et al.]
Research Questions:

(1) Do AI interventions influence the discernment accuracy and perceived information insufficiency (including when controlling for personal factors)?

(2) Do personal factors impact discernment accuracy or perceived information insufficiency?

(3) How does the type of feedback impact the cognitive load imposed on the user?
RESULTS

Key Findings:

1. Participants were much better at identifying logical fallacies with AI interventions (statistically significant in the case of AI-framed questioning) (Similar to original study)

2. AI-framed questioning wasn’t as successful as causal AI explanations at persuading participants a statement was logically valid even when it was (Similar to original study)

3. Participants generally preferred the causal AI explanations as opposed to the AI-framed questioning (Contrasts original study)

4. The cognitive load in all 3 conditions is very similar (New to this study)
Learn Your Biases - Advertisers Already Exploit Them

Cosmin Moroica

Replicating “Recommendation for Video Advertisements based on Personality Traits and Companion Content” by Dey, S. et al
How does a person’s personality impact the kind of video advertisement they prefer?

- Custom Audiences & Lookalike Audiences
- Google Ad Rank
Results

Power Analysis: Needed \( \sim 231 \) samples for \( \alpha = 0.05 \), large effect size and power = 0.8. Only had 34.

Conscientious participants were found to have a higher purchase intent for alert advertisements. Many alert advertisements were for financial products like insurance.

Table 4: \( \beta \) coefficients of the linear regression analysis performed on the participants’ opinion about alert video ads. In the brackets is the probability of a Type I error. Asterisks denote statistical significance with \( \alpha = 0.05 \).

<table>
<thead>
<tr>
<th>Extraversion</th>
<th>Novelty</th>
<th>Engagement</th>
<th>Pertinence</th>
<th>Purchase Intent</th>
<th>A/V Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2059(0.183)</td>
<td>0.5194(0.349)</td>
<td>0.4495(0.379)</td>
<td>0.4263(0.392)</td>
<td>0.5458(0.283)</td>
<td>0.0594(0.920)</td>
</tr>
</tbody>
</table>

\( R^2 = 0.689 \)  

Participants with high neuroticism were found to have higher purchase intent and pertinence for amusing advertisements. High neuroticism -> mood swings.

Table 5: \( \beta \) coefficients of the linear regression analysis performed on the participants’ opinion about amusing video ads. In the brackets is the probability of a Type I error. Asterisks denote statistical significance with \( \alpha = 0.05 \).

<table>
<thead>
<tr>
<th>Extraversion</th>
<th>Novelty</th>
<th>Engagement</th>
<th>Pertinence</th>
<th>Purchase Intent</th>
<th>A/V Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.1650(0.839)</td>
<td>-0.2843(0.481)</td>
<td>-0.7779(0.113)</td>
<td>-1.2684(0.013*)</td>
<td>-1.2993(0.032*)</td>
<td>-0.2957(0.483)</td>
</tr>
</tbody>
</table>
Grounded Abstraction Matching in interactions with code-generating LLMs

P342 Project

EMMA URQUHART (EU233)
"What It Wants Me To Say": Bridging the Abstraction Gap Between End-User Programmers and Code-Generating Large Language Models [1]

**Replication Study:**

**Grounded Abstraction Matching**

**Replication Method**

- Spreadsheet analysis in Excel + API invocation in Python
- Deterministic system + Non-deterministic system

Results:

**Diminished abstraction gap:** All users succeeded on their first attempt (with one exception due to misinterpretation)

**Deterministic system:**
- concise and technical
- less accessible to non-technical users

**Non-deterministic system:**
- verbose & contextualized
- more flexible to different query types

<table>
<thead>
<tr>
<th>Category</th>
<th>Statement</th>
<th>Deterministic</th>
<th>Non-deterministic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comprehensibility</td>
<td>I would consider my interactions with the tool to be understandable and clear.</td>
<td>3.0 (3.17 ± 0.69)</td>
<td>5.0 (4.67 ± 0.47)</td>
</tr>
</tbody>
</table>
Transforming Textual Discourse: Evaluating ChatGPT’s Influence on Attitude and Discussion Dynamics Among Cambridge’s Postgraduate Students

BY HANNA FOERSTER (MPHIL ACS)
Research Question

- How do attitude, sentiment, and phrasing choices in text discussions change after exposure to a biased LLM?

- Experiment setup:
  - Discussion topics (Cambridge’s dining hall food, bicycle infrastructure, …)
  - ChatGPT produced reference text (pos.-/neg.-/non-biased)
  - Produce own text discussion
Results

- **Non-biased:**
  - Tendency to include discussion of diverse opinions
  - More diplomatic phrasing

- **Biased:**
  - Tendency for one-sided discussions
    - High attitude clarity: Participants echoed own views
    - Lower attitude clarity: Participants echoed more of reference text views
  - More subjective and extreme phrasing

- **Implications:**
  - ChatGPT transforming textual discourse: Diversifying or polarizing views
  - Need for AI literacy of students & Research on bias in LLMs

<table>
<thead>
<tr>
<th>Compared groups</th>
<th>Attitude</th>
<th>Clarity</th>
<th>Correctness</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive VS negative</td>
<td>0.68</td>
<td>0.24</td>
<td>0.45</td>
</tr>
<tr>
<td>neutral VS positive</td>
<td>0.84</td>
<td>0.67</td>
<td>0.57</td>
</tr>
<tr>
<td>neutral VS negative</td>
<td>0.03</td>
<td>0.13</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Table 1: Change in attitude p-values
“Reducing Normative Dissociation And ‘The Thirty-Minute Ick’ On Instagram With BetterImagesOfAI “

Yasmin Dwiputri & Data Hazards Project / Better Images of AI / Managing Data Hazards / CC-BY 4.0
Reduced Sense of Agency
Reduced Self Awareness
Reduced Sense of Time
Reduced Memory

Flow States
Meaningful and creative endeavours: reading and socialising.

Zone States
Meaningless activities: gambling and other addictive activities.

"30:00 -Minute Ick"

Measuring Dissociation
“Consume recommended content on Instagram for exactly 10 minutes”

Measuring Perceptions of Social Media

T-statistic: -2.967 and Significance: 0.014

Exit Point Sentiment Analysis
“As a result of using Instagram I feel?”

Control Condition: Negative

Graphic Condition: Neutral

T-statistic: -2.967 and Significance: 0.014

<table>
<thead>
<tr>
<th>Sentiment</th>
<th>T-statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>-8.000</td>
<td>0.015</td>
</tr>
<tr>
<td>Neutral</td>
<td>10.0</td>
<td>0.010</td>
</tr>
<tr>
<td>Positive</td>
<td>-2</td>
<td>0.184</td>
</tr>
</tbody>
</table>
On the topic of Individualization

Adding people back into “When People and Algorithms Meet”

Jonathan Haley
My paper aims to replicate this study:

- To discover: Have the user issues or solutions proposed with intelligent systems changed since 2019?
- It also: Extends thematic analysis to include Individualistic Issues.

Methodology:

- Scrape 10,000 user reviews each from Netflix and Google Maps.
- Use GPT 3.5 to pick out key quotes relating to HCAI issues.
- Perform 8 Interviews to investigate possible solutions to these HCAI Issues.
Results

Main Takeaways:

- **Explain Why** – Users had, been failed by intelligent systems in the past. They were therefore wary of any and all data so wanted to understand the values and information provided.

- **Give Users Information** - This helps to build user trust and allows them to make more informed decisions.

- **Give Users Choice and Options** - Users want fine-grained control options both for practical reasons and also to best account for users individual situations.
Investigating How Different Modes of Interaction Affect User Experience for Image Generation With DALL-E

P342: Practical Research in Human-Centred AI

Joseph Cameron (jmc276) - 28/11/2023
Research Question

- What is the impact of different interaction modalities on user experience and its relevant time and error-rate usability metrics?

- Mode 1: Default Text Prompts
- Mode 2: Text Prompts + DALL-E’s Editing Tools
- Mode 3: Text Prompts + ChatGPT Prompt Assistance
Results

- DALL-E’s Editing Tools and ChatGPT Prompt Assistance Increase Time, but also Decrease Errors.

- Participants felt more comfortable to explore when feedback from DALL-E or ChatGPT is available. Sole Text Prompting Stifles Interaction and Connection.

- Participants felt more agency with ChatGPT and DALL-E’s editing assistance.
Mwalimu Mbaya?

ON CHATGPT AS A SUPPORT TOOL FOR SWAHILI VOCABULARY ACQUISITION

JOSEPHINE REY
How might the use of ChatGPT improve acquisition & retention of Swahili vocabulary?

Motivations

- Elevate an **Africa-inclusive context** in AI for education
- Investigate **adaptability** of AI systems to African languages
- Assess one aspect of ChatGPT as a learning tool: **Learning new vocabulary**

Methods

- 16 participants
- 3 groups
- 1 exercise
- 2 assessments

*Danry et al, 2023*

*Campos et al. 2004 (Latin)*
FINDINGS & DISCUSSION

Are the differences between these means statistically significant between groups?

ANOVA TEST: Yes (p = 0)  
SIGN TEST: Yes (p = 0.001)

Group 1: 96.25%  88.75%  
Group 2: 60.83%  55%  
Group 3: 74.17%  70%

Why?

ChatGPT Log Data...

- Poor explanations
- Forgetting rules of engagement (out of scope vocabulary use)
- Implicit stereotypes

3. Asante - This word resembles the English word "asante,"

One sunny morning (Asubuhi), the villagers gathered

Mwalimu Mbaya
"Bad Teacher"
Non-AI-Experts Predicting the Accuracy of LM on QA tasks
Why predict AI’s accuracy?

● Human’s understanding—mental model—of the system’s error boundaries.
● To foresee potential errors and decide when to bypass the system and when to delegate.
● Prevent disappointment, time wastage and inefficient use of computational resources.

Research Questions:

Q1. How predictable is the accuracy of an unfamiliar LM on QA tasks for non-AI-experts?
Q2. Can participants improve their predictions as they continuously observe more examples of successes and failures of the LM?
Q3. What is the effect of prior familiarity with generative AI on the two questions above, after controlling the effects of personal age and sex?

Experimental setup:

● A pilot study of 6 Cambridge graduate students and 2 crowdsourced workers.
● A final sample of 17 UK participants (sex-balanced distribution and fluent in English) passed the quality check.
● Predicting Falcon-7B-instruct's accuracy on 48 questions from the TruthfulQA benchmark.
● Statistical Analysis: T-test and ANCOVA, as Shapiro-Wilk test does not reject the normality assumption.
Q1. How predictable is the accuracy of an unfamiliar LM on QA tasks for non-AI-experts?

Q2. Can participants improve their predictions as they continuously observe more examples of successes and failures of the LM?

Q3. What is the effect of prior familiarity with generative AI on the two questions above, after controlling the effects of personal age and sex?

Take Home Messages

- Non-AI-experts showed random performance in anticipating LM accuracy, although there is a marginal advantage of prior experience.
- No evidence supporting that participants could adjust their expectations (or mental models) regarding the LM’s error boundaries over more interaction, regardless of participants’ prior familiarity.
- These show a concerning trend, implying that users may frequently encounter disappointment and resource wastage, while unable to significantly improve their expectations on LM’s error boundaries.
An Industrial Devolution: Naming Under the Influence of Copilot.

Michael Lee
How would you caption this image?

How would you name this abstract object?

```ml
let ??? op base :
  fix (fun g ->
    base ++ option (op ++ g) ==> function
    | (e, None) -> e
    | (e, Some(f, e')) -> f e e')

Krishnaswami and Yallop (2019),
A Typed Algebraic Approach to Parsing
```
Results

Empirical Distribution of Entropy (names)  
Distributions computed across treatments
- On
- View
- Off

Empirical Distribution of Entropy (words)  
Distributions computed across treatments
- On
- View
- Off

Mean Entropy
Of names (left) and words (right), across treatments.  
95% Confidence Interval

Pairwise Differences of Mean Entropy
Of names (left) and words (right), across treatments.  
95% Confidence Interval

<table>
<thead>
<tr>
<th></th>
<th>( P(H(c, \text{OFF}) &gt; H(c, t)) )</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>On</td>
<td>0.848</td>
<td>[0.727, 0.970]</td>
</tr>
<tr>
<td>View</td>
<td>0.909</td>
<td>[0.818, 1.000]</td>
</tr>
</tbody>
</table>

|   | \( P(\text{renamed}|t) \) | 95% CI        |
|---|------------------------------|-------------|
| On | 0.106                        | [0.061, 0.160] |
| View | 0.197                       | [0.129, 0.267] |
| Off | 0.258                        | [0.182, 0.333] |
Guidance for AI-Mediated Communication: AI Does Not Alter Perceptions of Text Messages

N’yoma Diamond

Department of Computer Science and Technology, University of Cambridge, UK
### Problem, Motivation

**Text-based communication can be stressful or difficult**

- Emotion, sarcasm, social nuance are difficult to convey via text
- Anxiety, depression, other conditions can exacerbate stress
- Text messaging can be difficult for neurodivergent people

**Generative AI has the potential to assist**

- AI-MC has been shown to improve user speed and confidence
- Generative AI (e.g., ChatGPT) are useful text composition tools
- Does the belief of AI usage alter perceptions? *(Results say no)*
Results (Levene & Tukey Tests)

(a) Tone

<table>
<thead>
<tr>
<th>Label 1</th>
<th>Label 2</th>
<th>$\hat{y}_2 - \hat{y}_1$</th>
<th>Lower bound</th>
<th>Upper bound</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>-</td>
<td>0.0647</td>
<td>-0.4346</td>
<td>0.5640</td>
<td>0.9501</td>
</tr>
<tr>
<td>+</td>
<td>=</td>
<td>-0.1918</td>
<td>-0.6919</td>
<td>0.3084</td>
<td>0.6395</td>
</tr>
<tr>
<td>-</td>
<td>=</td>
<td>-0.2565</td>
<td>-0.7557</td>
<td>0.2428</td>
<td>0.4491</td>
</tr>
</tbody>
</table>

(b) Clarity

<table>
<thead>
<tr>
<th>Label 1</th>
<th>Label 2</th>
<th>$\hat{y}_2 - \hat{y}_1$</th>
<th>Lower bound</th>
<th>Upper bound</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>-</td>
<td>0.3283</td>
<td>-0.1849</td>
<td>0.8415</td>
<td>0.2898</td>
</tr>
<tr>
<td>+</td>
<td>=</td>
<td>0.1027</td>
<td>-0.4113</td>
<td>0.6168</td>
<td>0.8854</td>
</tr>
<tr>
<td>-</td>
<td>=</td>
<td>-0.2256</td>
<td>-0.7388</td>
<td>0.2876</td>
<td>0.5560</td>
</tr>
</tbody>
</table>

(c) Intent

<table>
<thead>
<tr>
<th>Label 1</th>
<th>Label 2</th>
<th>$\hat{y}_2 - \hat{y}_1$</th>
<th>Lower bound</th>
<th>Upper bound</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>-</td>
<td>0.2934</td>
<td>-0.2186</td>
<td>0.8054</td>
<td>0.3696</td>
</tr>
<tr>
<td>+</td>
<td>=</td>
<td>0.0479</td>
<td>-0.4649</td>
<td>0.5608</td>
<td>0.9737</td>
</tr>
<tr>
<td>-</td>
<td>=</td>
<td>-0.2455</td>
<td>-0.7574</td>
<td>0.2665</td>
<td>0.4976</td>
</tr>
</tbody>
</table>
The Goldilocks Zone for Explanations: Finding the Sweet Spot in Recommender Systems

Ria Mundhra

Replicating Kulesza et al’s “Too much, too little, or just right? Ways explanations impact end users' mental models”
How does changing the completeness and soundness of explanations affect end users mental models of the system? What about trust?
Exploring the Effect of Augmented Writing Systems on Creative Writing Processes and Outcomes

by Sol Dubock
The study consisted of an introductory survey, two 20 minute writing tasks (one in each editor), a conclusion survey, and a lightly structured discussion.
The Results

Participant Quotes:
P1 “I was specifically writing something sad, and the AI kept wanting to make it positive again”
P5 “It limited my sense of expression when I used it”
P2 “I was finding that I could use it to suggest something and then if it was inspiring I could go back and change a few words and make it fit”

<table>
<thead>
<tr>
<th>Measure</th>
<th>Editor-Green (unassisted)</th>
<th>Editor-Red (AI-assisted)</th>
<th>t-value</th>
<th>( t_{14}(0.1) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. (mean) text length</td>
<td>398.4 words</td>
<td>455.1 words</td>
<td>1.005</td>
<td>1.345</td>
</tr>
<tr>
<td>Avg. spelling error count</td>
<td>2</td>
<td>0.14</td>
<td>1.462</td>
<td>1.345</td>
</tr>
<tr>
<td>Avg. spelling error rate (per 100 words)</td>
<td>0.597</td>
<td>0.034</td>
<td>1.505</td>
<td>1.345</td>
</tr>
<tr>
<td>Avg. grammar error count</td>
<td>1.857</td>
<td>1</td>
<td>0.684</td>
<td>1.345</td>
</tr>
<tr>
<td>Avg. grammar error rate (per 100 words)</td>
<td>0.640</td>
<td>0.218</td>
<td>0.8944</td>
<td>1.345</td>
</tr>
<tr>
<td>Avg. number of distinct AI-phrases*</td>
<td>N/A</td>
<td>2.714</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Avg. AI-phrase* rate (per 100 words)</td>
<td>N/A</td>
<td>0.600</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Impact of Conflict on User Perspectives and Problems with Intelligent Applications

Sophie Walker
Research Questions

- **RQ1**: How have current events affected user problems with intelligent navigation applications?
- **RQ2**: Is there an impact current events and conflicts have on implicit user trust in intelligent systems?

- Web scraping and Sentiment Analysis
- BERTopic Topic Analysis
- QualiGPT and ChatGPT Thematic Analysis

---


Results

- 619 of 2497 reviews (24.8%) were relevant to the conflicts
- At least 225 reviews referenced Sinai

User reported problems in Spotify DJ

STEPHANIE CHO
RESEARCH QUESTIONS X2

1. Which problems do users encounter when using Spotify DJ?
2. What kind of features or improvements do users want from Spotify DJ?

METHODOLOGY

Reddit and articles as source of data.
Topic modelling 1) using LDA, interpreting results manually or 2) using ChatGPT, or 3) using only ChatGPT to produce cumulative summaries.
**USER VARIABILITY**

To have new songs or not to have? Variability within and between users.

**DJ AS A CONCEPT**

Overwhelmingly negative comments on DJ voice. Why is Spotify doing this anyways? Is this a step forward or backwards?

**REDDIT AS SOURCE OF INFORMATION**

Discussions going off topic. Many layers of comments. How much is relevant?

**CHATGPT FOR SUMMARIES**

Interpretable and accurate by-topic summaries over small input size, with some difficulty in prompting.
The Impact of Personality Traits on the Sentiment of People’s Preferred Video Ads

Tamisa Ketmalasiri
Research Question

RQ: How does personality traits affect the sentiment of people’s preferred video ads?

RQ: To what degree can personality traits be used to predict the sentiment of people’s preferred video ads?

The Big Five personality traits

<table>
<thead>
<tr>
<th>Traits</th>
<th>Dominant Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraversion</td>
<td>excitability, sociability, talkativeness, assertiveness</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>trust, altruism, kindness, affection</td>
</tr>
<tr>
<td>Openness</td>
<td>creativity, openness to trying new things, focus on tackling new challenges</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>thoughtfulness, good impulse control, goal-directed behaviors, organized</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>sadness, moodiness, emotional instability</td>
</tr>
</tbody>
</table>

Sentiment of Video Ads

<table>
<thead>
<tr>
<th>Synonyms</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>energetic, adventurous</td>
</tr>
<tr>
<td>Alert</td>
<td>attentive, curious</td>
</tr>
<tr>
<td>Amusing</td>
<td>humored, laughing</td>
</tr>
<tr>
<td>Calm</td>
<td>soothed, peaceful</td>
</tr>
</tbody>
</table>
Results

Coefficients of multiple regression analysis on participants’ opinion on active video ads. Asterisk(*) denotes statistical significance (p<0.01).

<table>
<thead>
<tr>
<th></th>
<th>Overall Experience</th>
<th>Overall Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraversion</td>
<td>0.20 (0.23)</td>
<td>0.25 (0.06)</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>0.26 (0.22)</td>
<td>0.22 (0.18)</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>-0.23 (0.15)</td>
<td>0.03 (0.79)</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>0.16 (0.35)</td>
<td>-0.02 (0.87)</td>
</tr>
<tr>
<td>Openness</td>
<td>-0.07 (0.69)</td>
<td>-0.39 (&lt;0.01)*</td>
</tr>
<tr>
<td>R²</td>
<td>0.23</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Coefficients of multiple regression analysis on participants’ opinion on alert video ads. Asterisk(*) denotes statistical significance (p<0.01).

<table>
<thead>
<tr>
<th></th>
<th>Overall Experience</th>
<th>Overall Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraversion</td>
<td>0.10 (0.28)</td>
<td>0.11 (0.38)</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>0.41 (&lt;0.01)*</td>
<td>0.31 (0.07)</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>-0.15 (0.12)</td>
<td>-0.25 (0.05)</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>0.21 (0.05)</td>
<td>0.37 (&lt;0.01)*</td>
</tr>
<tr>
<td>Openness</td>
<td>-0.09 (0.31)</td>
<td>0.18 (0.16)</td>
</tr>
<tr>
<td>R²</td>
<td>0.59</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Precision, Recall, Accuracy and F1 measures of SVM and CART classifiers.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.45</td>
<td>0.53</td>
<td>0.59</td>
<td>0.57</td>
</tr>
<tr>
<td>CART</td>
<td>0.4</td>
<td>0.44</td>
<td>0.47</td>
<td>0.44</td>
</tr>
</tbody>
</table>
Data-centric explanations affect trust in LLM output

Zeno Kujawa
Research question

- Llama-2 models were trained on 89.7% English data (German: 0.17%) [1]
- Previous research indicates that data-centric explanations affect trust [2]
- How does trust change when users are informed of language imbalance?
- Trust and trustworthiness matter in both ethical and economic sense


The study

- 10 English and 8 German speakers, all aged 18-29, majority used to LLMs
- Show 3 LLM-generated instructions, measure trust, inform about data
- Small drop in trust across all participants ($p \sim 0.03$)
- No statistically significant difference between language groups