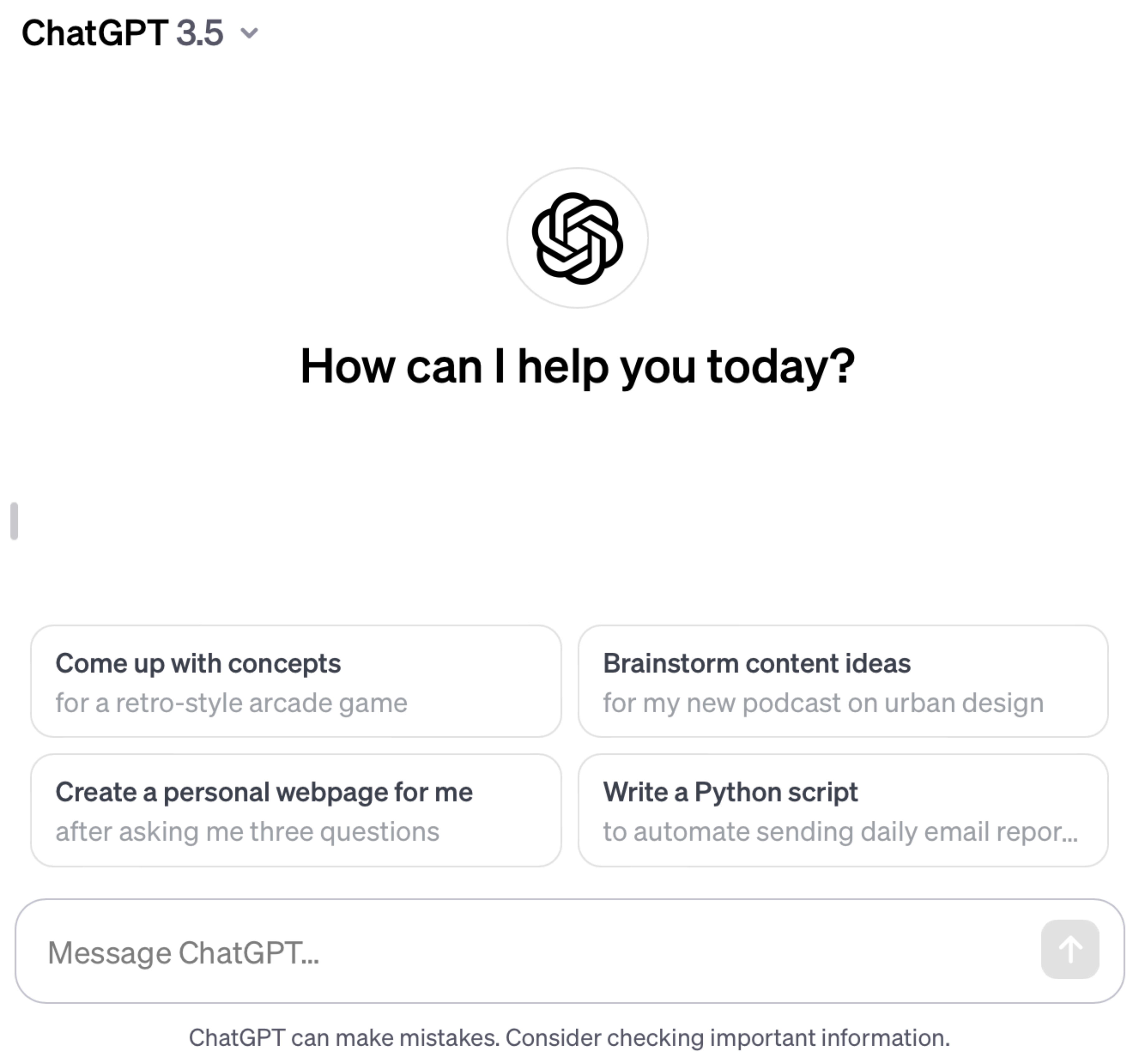
Wikipedia vs ChatGPT: A study on their Use as Tools for Topic Discovery and Learning

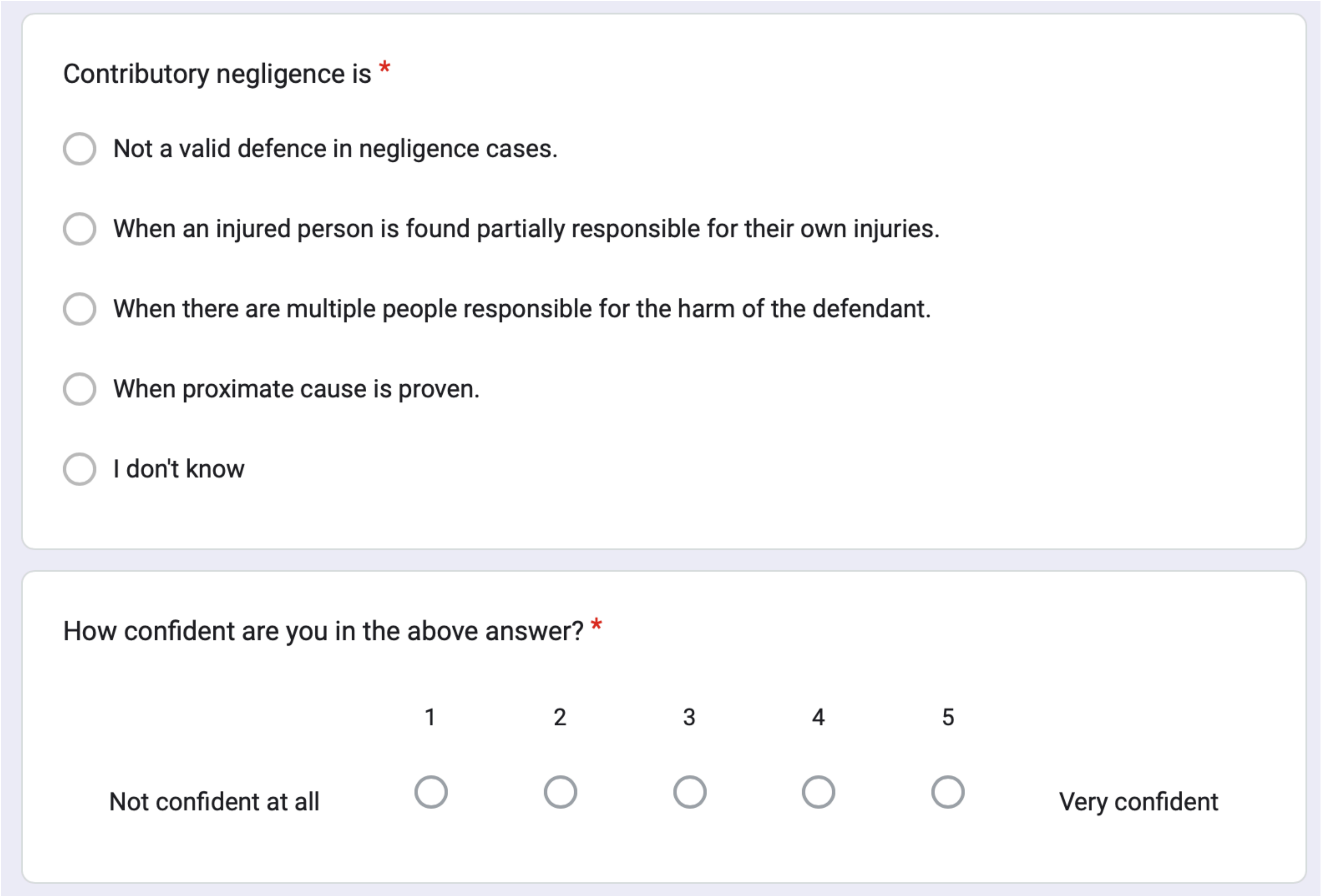
Alexandra Herghelegiu

Wikipedia vs ChatGPT

For learning



Knowledge test



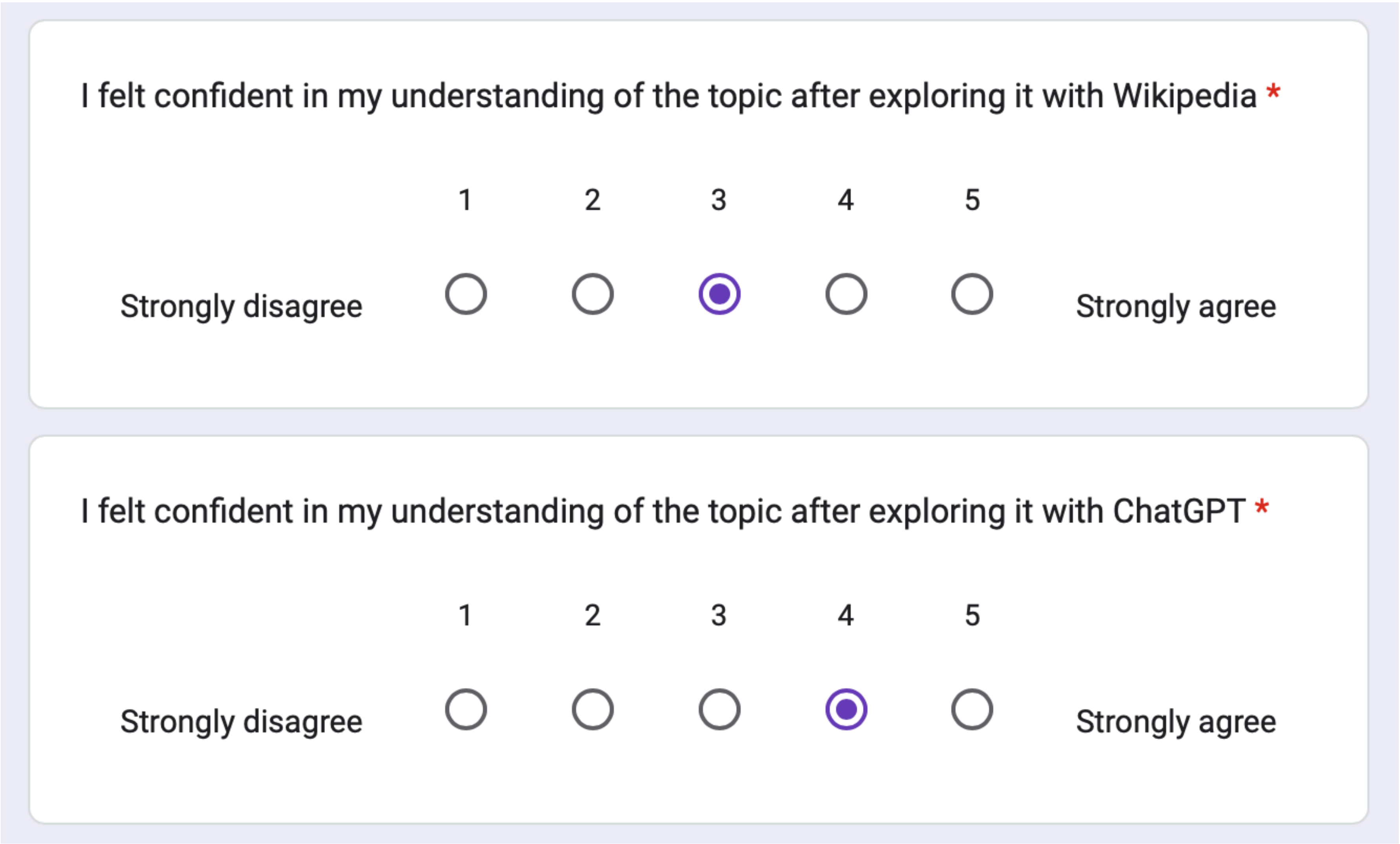
Contributory negligence is *

- Not a valid defence in negligence cases.
- When an injured person is found partially responsible for their own injuries.
- When there are multiple people responsible for the harm of the defendant.
- When proximate cause is proven.
- I don't know

How confident are you in the above answer? *

	1	2	3	4	5	
Not confident at all	○	○	○	○	○	Very confident

Post-Study Questionnaire



I felt confident in my understanding of the topic after exploring it with Wikipedia *

	1	2	3	4	5	
Strongly disagree	○	○	◉	○	○	Strongly agree

I felt confident in my understanding of the topic after exploring it with ChatGPT *

	1	2	3	4	5	
Strongly disagree	○	○	○	◉	○	Strongly agree

- 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

Symmetrical photo of subway platform. Predominantly red and dark colours, apart from platform floor which is illuminated by central lights.

We are in a long corridor, slate tile floors, metallic red panels line the walls. The ceiling is a mix of long white lights, red panels and two signs hang from the ceiling showing a pram, a wheel chair, and number 1.

An underground station. It is imposing and dystopian - red, glaring, neon, empty. It is very clean and unnatural, with disabled toilets shown on signs in the hallway.

eerie train platform that is very dark above and below the platform contrasted by bright lights at the centre of the platform. there are signs to show where the accessible carriges are and when the next train is coming

A symmetrical image of an open subway station without people, which has been photoshopped to be monochrome with the exception of red lights. The platform and lights are at the centre of the image, with spaces for tracks in either side

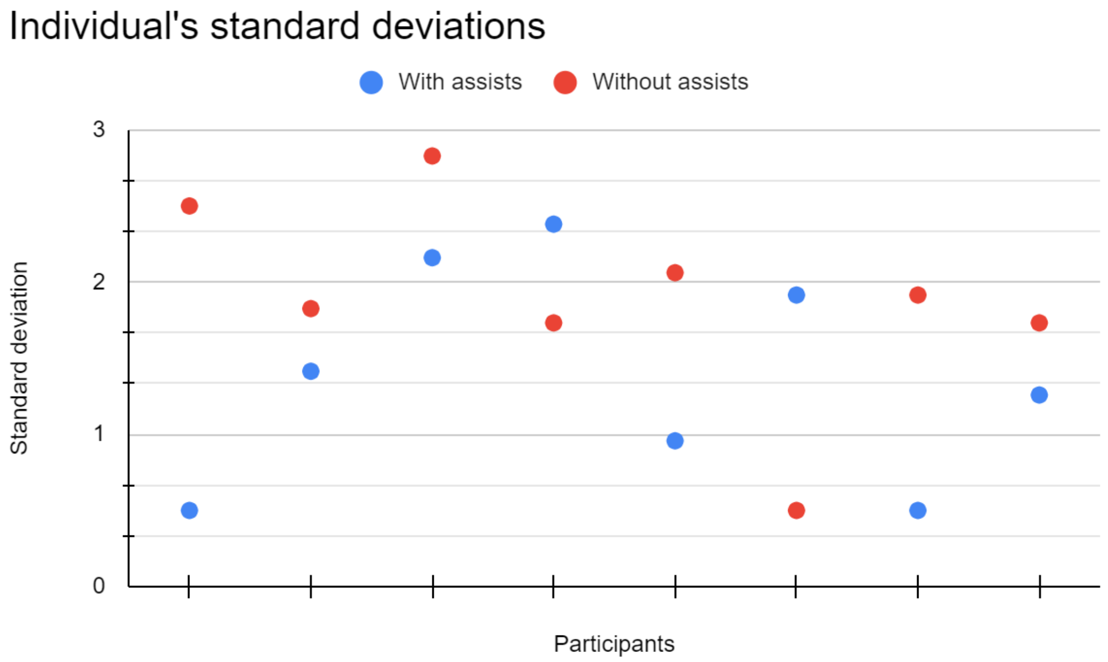
A view along long corridor with red tiled walls. There are lampposts down the centre of the platform.

An ominous hallway which appears to be an airport corridor, signs give directions.

middle of underground train with bright lights in centr and dark sides with train rails unvisible and signs haning down on each side with a red vibe

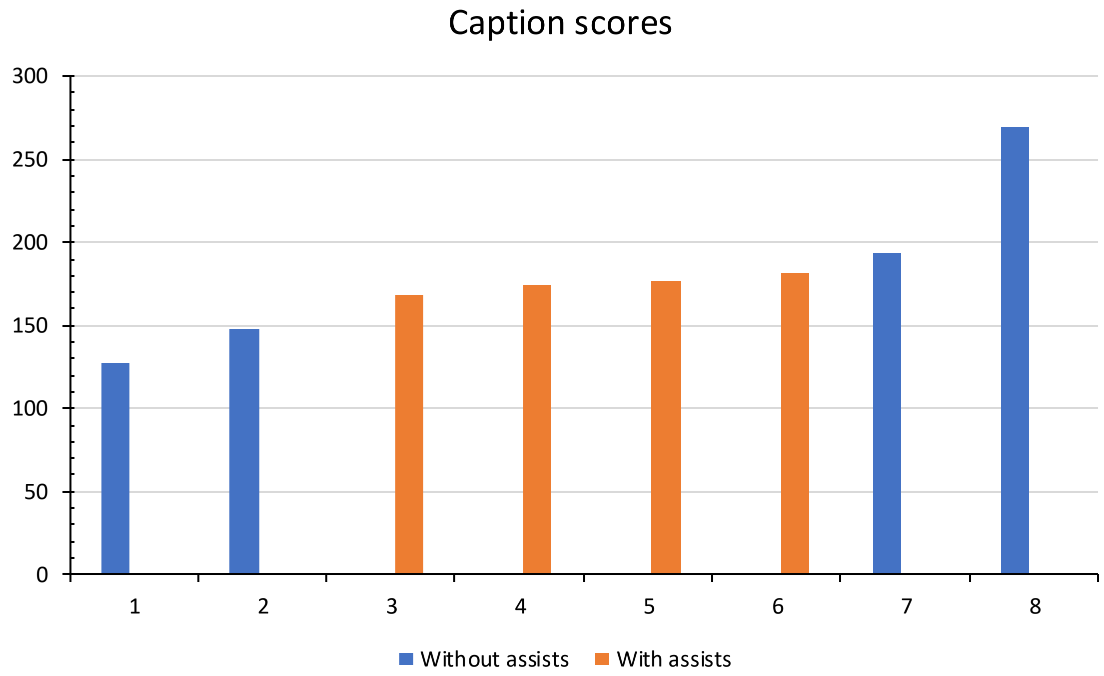
The results

Expert comparison



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

Wider comparison



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

Travel Advice

Attraction

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

activity

nature

history

cusine

culture

Reliability

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

Hallucination

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

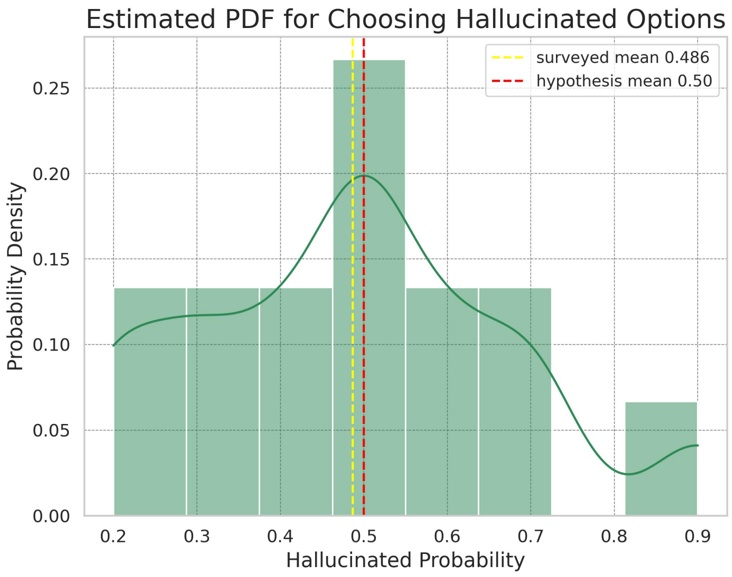
[1] Seyidov, Javid, and Roma Adomaitienė. "Factors influencing local tourists’ decision-making on choosing a destination: a case of Azerbaijan." Ekonomika 95.3 (2016).

[2] Mohd Shariff, Shafiza, Xiuzhen Zhang, and Mark Sanderson. "User perception of information credibility of news on twitter." ECIR 2014.

Results

Attraction - hallucinated probability

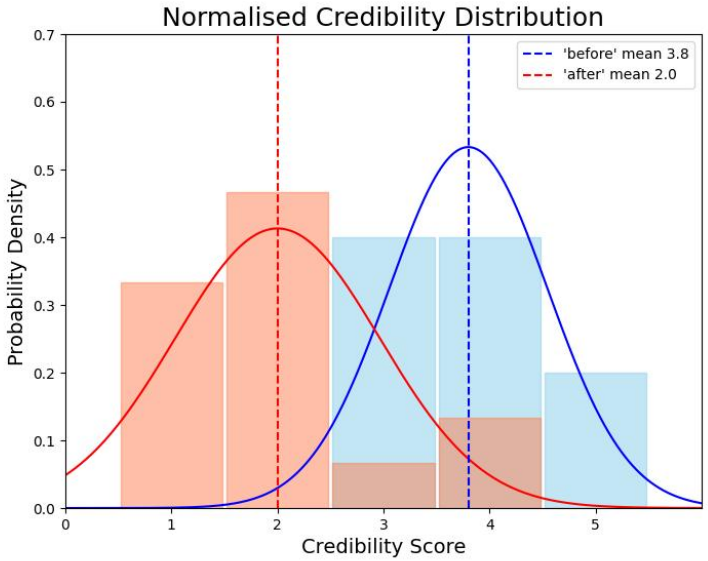
Null hypothesis:
the likelihood of choosing hallucinated and real travel advice is equal.



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.]

CHARLES YANG

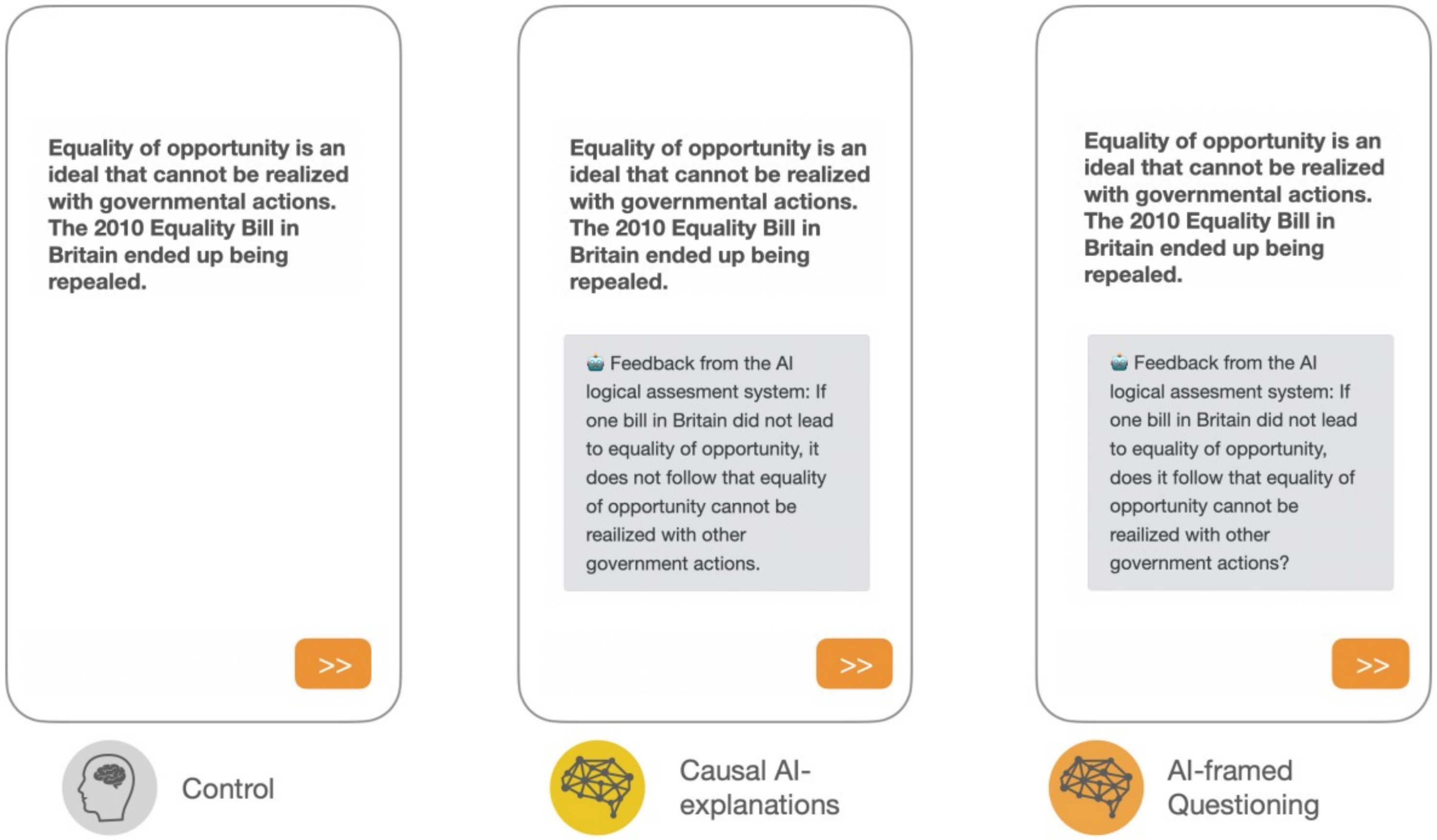
ABOUT

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?

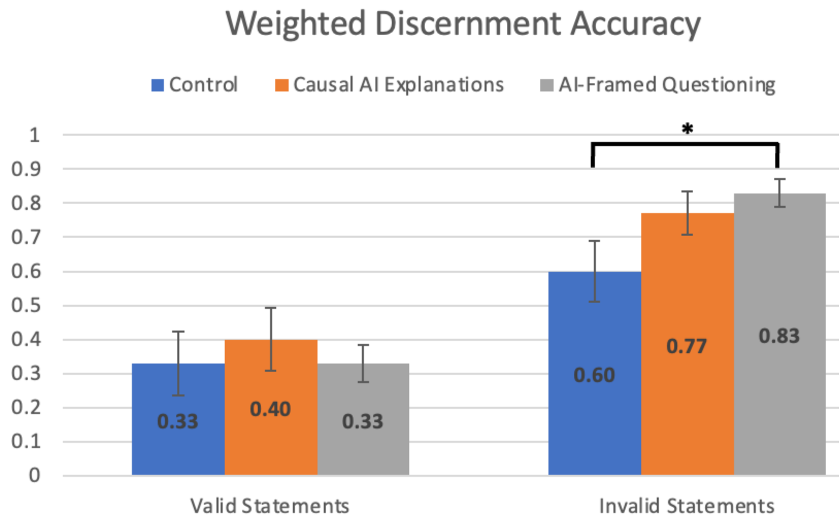
(3) How does the type of feedback impact the cognitive load imposed on the user?

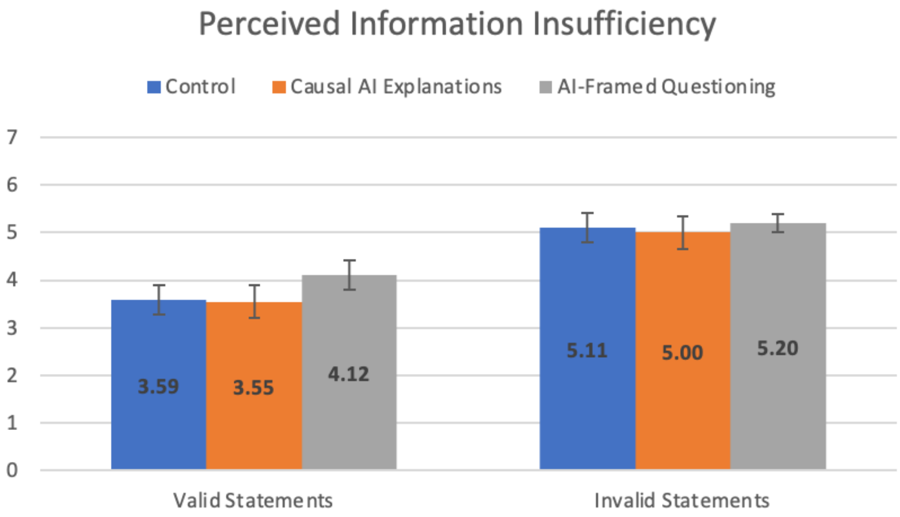


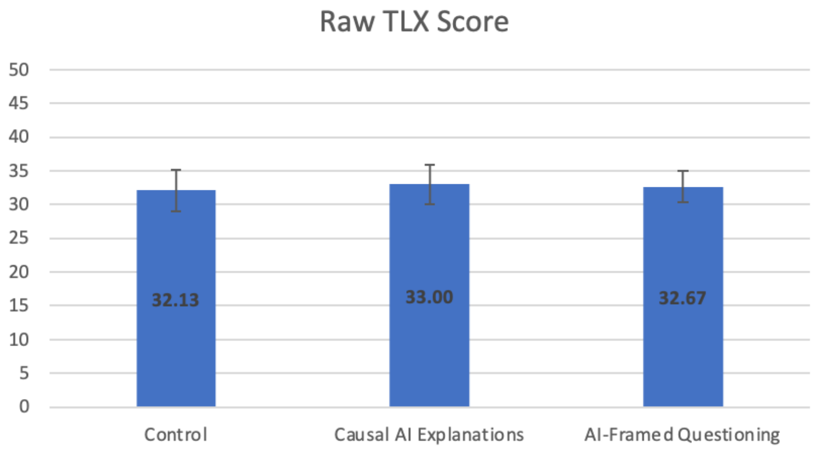
RESULTS

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)**







Impact of different factors on Weighted Accuracy and Perceived Information Insufficiency for Invalid statements

Factor	Dependent Variable	F	df	Sig.	Observed Power	Partial η squared
AI Feedback Group	Weighted Accuracy	2.414	2	0.096	0.473	0.060
	Information Insufficiency	0.226	2	0.798	0.084	0.006
Trust in AI	Weighted Accuracy	4.419	1	0.039*	0.546	0.056
	Information Insufficiency	0.092	1	0.763	0.060	0.001
Cognitive Reflection Level	Weighted Accuracy	3.323	1	0.072	0.436	0.042
	Information Insufficiency	3.122	1	0.081	0.415	0.040
Prior Knowledge	Weighted Accuracy	7.088	1	0.009**	0.748	0.086
	Information Insufficiency	5.832	1	0.018*	0.664	0.072
Prior Belief	Weighted Accuracy	0.997	1	0.321	0.167	0.013
	Information Insufficiency	1.527	1	0.220	0.230	0.020

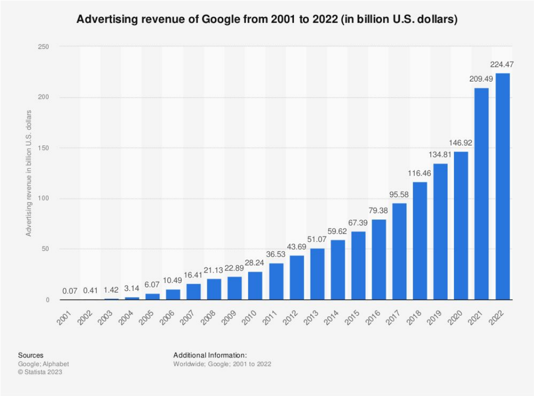
Learn Your Biases - Advertisers Already Exploit Them

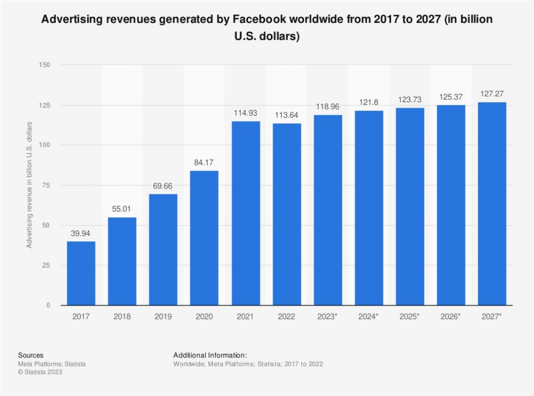
Cosmin Moroca

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 ~231 samples for α = 0.05, large effect size and power = 0.8. Only had 34.

	Novelty	Engagement	Pertinence	Purchase Intent	A/V Quality
Extraversion	0.4449(0.615)	0.4260(0.602)	0.1492(0.851)	1.4482(0.090)	0.0901(0.925)
Agreeableness	-1.1052(0.140)	-1.0987(0.114)	-0.4837(0.458)	-1.0235(0.136)	-0.7959(0.316)
Openness	-0.0200(0.965)	-0.2891(0.500)	-0.1272(0.759)	-0.2020(0.633)	-0.3201(0.525)
Conscientiousness	1.2059(0.183)	1.4034(0.100)	0.8927(0.269)	2.7335(0.004*)	0.1592(0.867)
Emotional Stability	0.5194(0.349)	-0.4495(0.379)	-0.4263(0.392)	-0.5458(0.285)	0.0594(0.920)
R^2	0.689	0.631	0.777	0.762	0.642

Table 4: β 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 $\alpha = 0.05$.

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

	Novelty	Engagement	Pertinence	Purchase Intent	A/V Quality
Extraversion	-0.1967(0.820)	0.1623(0.873)	-0.5253(0.594)	1.1012(0.346)	-1.1501(0.216)
Agreeableness	1.00882(0.135)	1.5222(0.060)	-0.1718(0.815)	-0.4681(0.589)	0.3821(0.575)
Openness	0.7246(0.101)	0.8560(0.099)	-0.1858(0.698)	-0.2970(0.597)	0.2446(0.580)
Conscientiousness	-0.1656(0.839)	0.7063(0.464)	0.5944(0.523)	2.1559(0.062)	-0.7343(0.395)
Emotional Stability	-0.2843(0.481)	-0.7779(0.113)	-1.2684(0.013*)	-1.2393(0.032*)	-0.2957(0.483)
R^2	0.698	0.668	0.653	0.649	0.574

Table 5: β 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 $\alpha = 0.05$.

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

28 NOVEMBER 2023

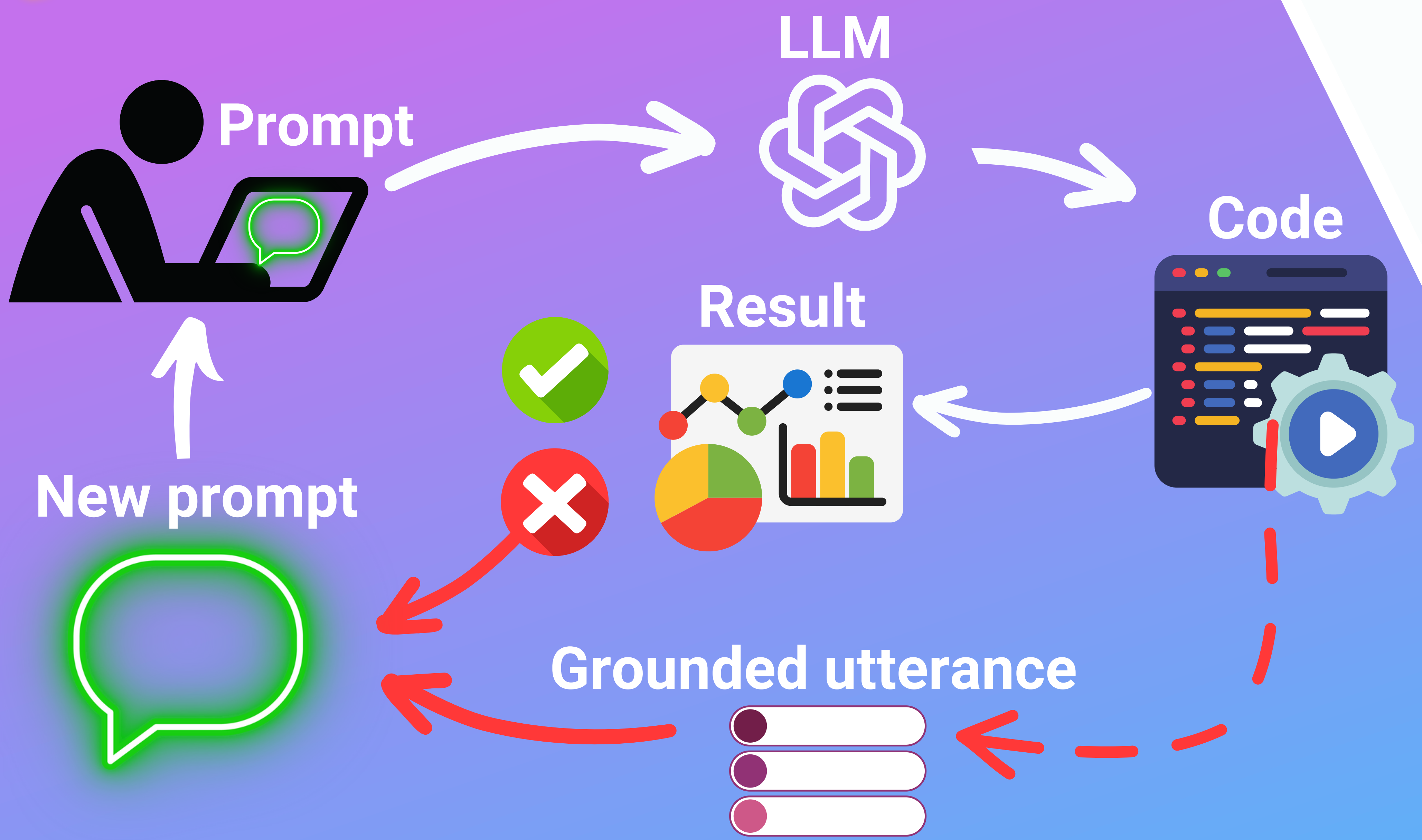
Grounded Abstraction Matching in interactions with code-generating LLMs

P342 Project

EMMA URQUHART (EU233)

Replication Study: "What It Wants Me To Say": Bridging the Abstraction Gap Between End-User Programmers and Code-Generating Large Language Models [1]

Grounded Abstraction Matching



Replication Method

- **Spreadsheet analysis in Excel**
  + API invocation in Python

- **Deterministic system + Non-deterministic system**

[1] Liu, M. X., Sarkar, A., Negreanu, C., Zorn, B., Williams, J., Toronto, N., & Gordon, A. D. (2023, April). "What It Wants Me To Say": Bridging the Abstraction Gap Between End-User Programmers and Code-Generating Large Language Models. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1-31).

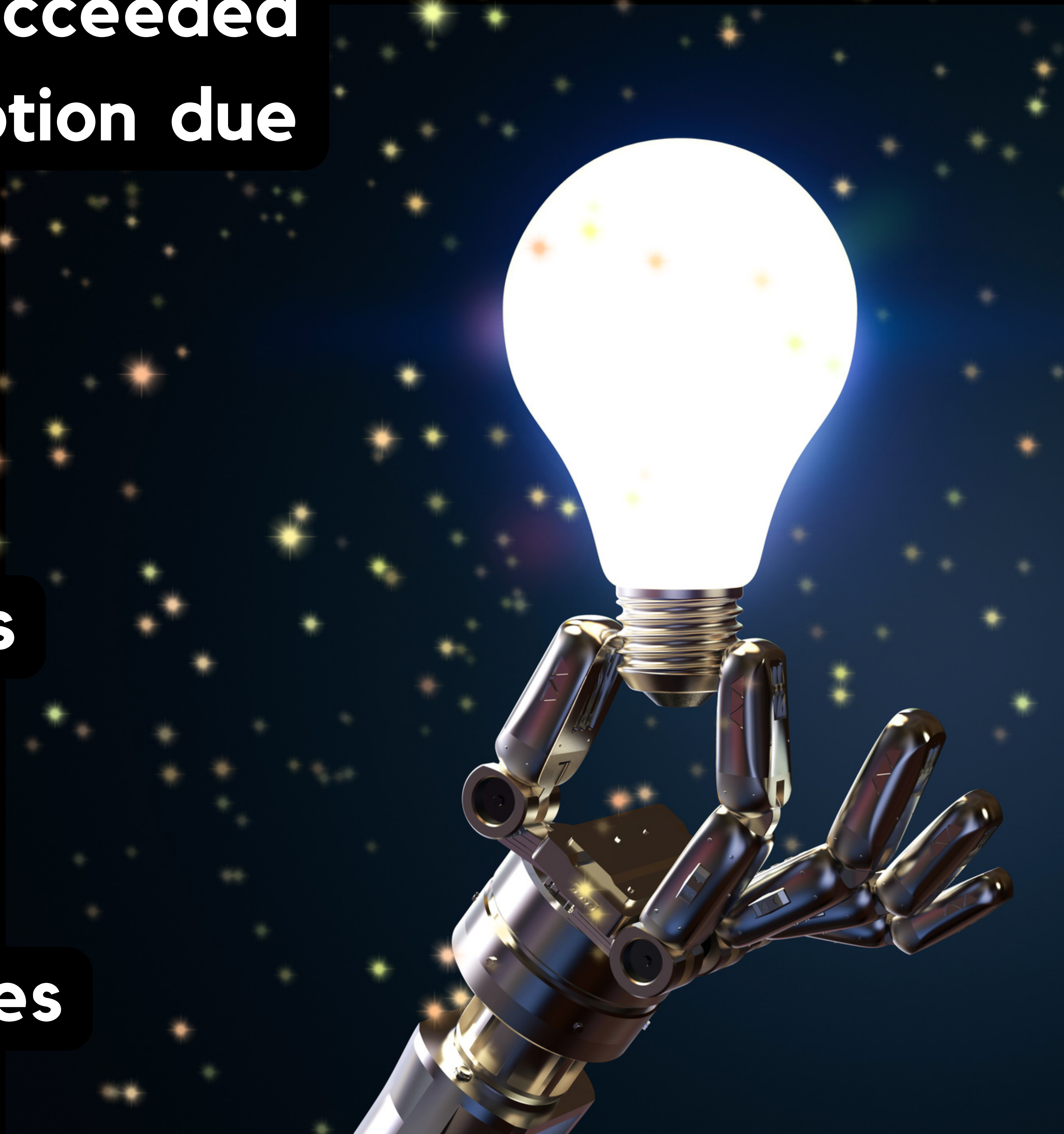
Category	Statement	Deterministic	Non-deterministic
Comprehensibility	I would consider my interactions with the tool to be understandable and clear.	3.0 (3.17 ± 0.69)	5.0 (4.67 ± 0.47)

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



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

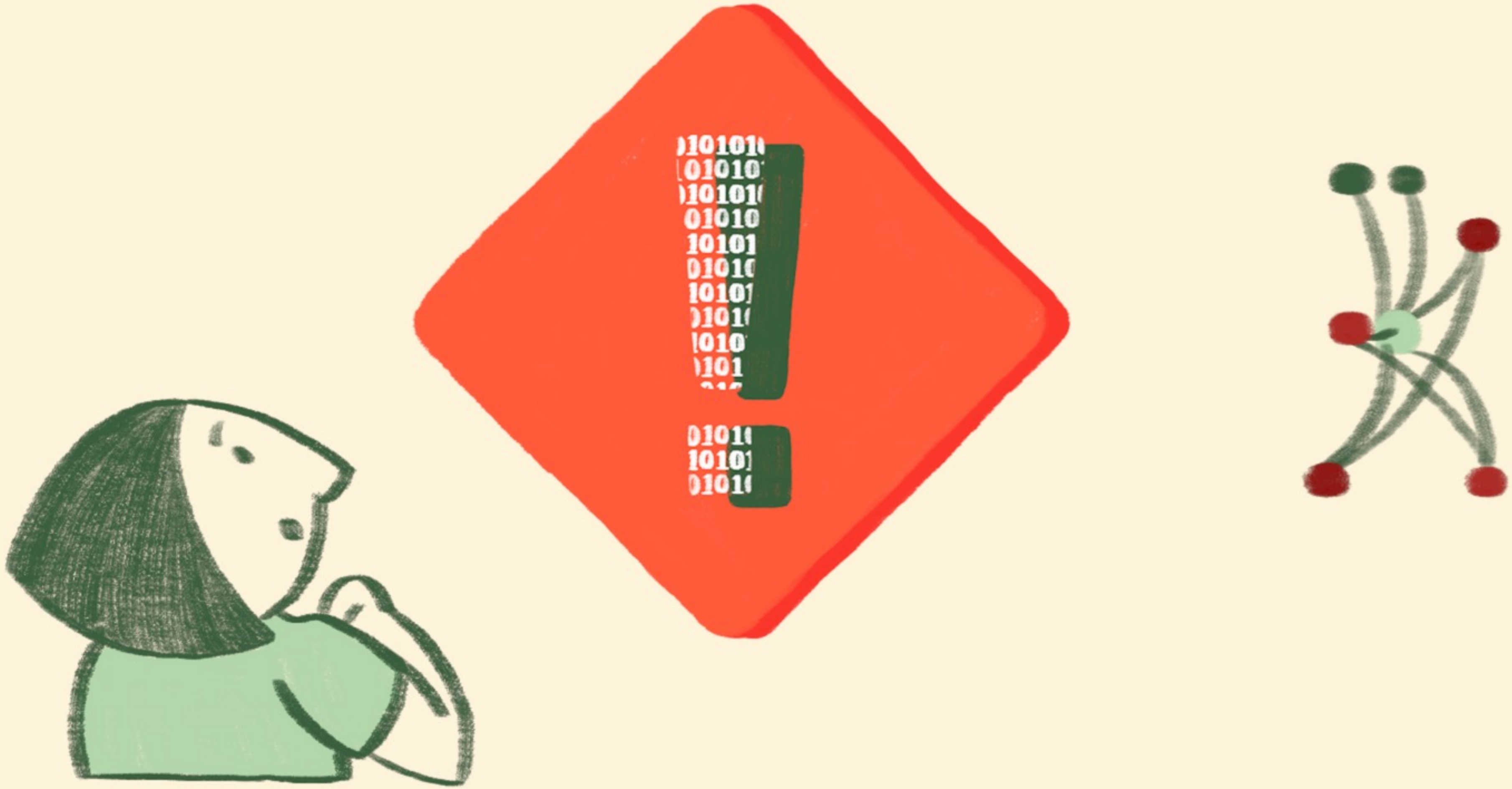
Results

- 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

Compared groups	Attitude	Clarity	Correctness
positive VS negative	0.68	0.24	0.45
neutral VS positive	0.84	0.67	0.57
neutral VS negative	0.03	0.13	0.74

Table 1: Change in attitude p-values

“Reducing Normative Dissociation And ‘The Thirty-Minute Ick’ On Instagram With BetterImagesOfAI “



Israel Mason-Williams

Amanda Baughan, Mingrui Ray Zhang, Raveena Rao, Kai Lukoff, Anastasia Schaadhardt, Lisa D Butler, and Alexis Hiniker. 2022. "I Don't Even Remember What I Read": How Design Influences Dissociation on Social Media. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 1–13.

Lisa Butler. 2011. Must dissociation be unusual? Journal of Trauma and Dissociation 12, 4 (2011), 454.

- 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.

"

-Minute Ick"


If Social Media Is Making You Sad, You're Not Alone

August 5, 2022 by Christian Zilles

JUNE 20, 2022 | 5 MIN READ

Why Social Media Makes People Unhappy—And Simple Ways to Fix It

Research suggests platform designs make us lose track of time spent on them and can heighten conflicts, and then we feel upset with ourselves

BY DAISY YUHAS

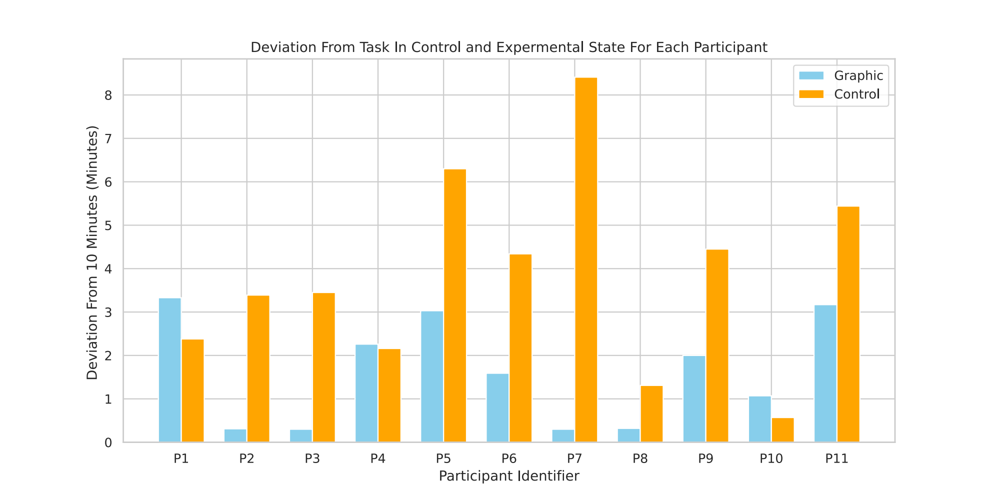
BEWARE OF SOCIAL MEDIA PLATFORM

Beware of Social Media Platform / 30 Minute Ick Factor / Dissociation / S Lakshmanan, Psychologist

Jonathan A. Tran, Katie S. Yang, Katie Davis, and Alexis Hiniker. 2019. Modelling the Engagement-Disengagement Cycle of Compulsive Phone Use. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/ 3290605.3300542

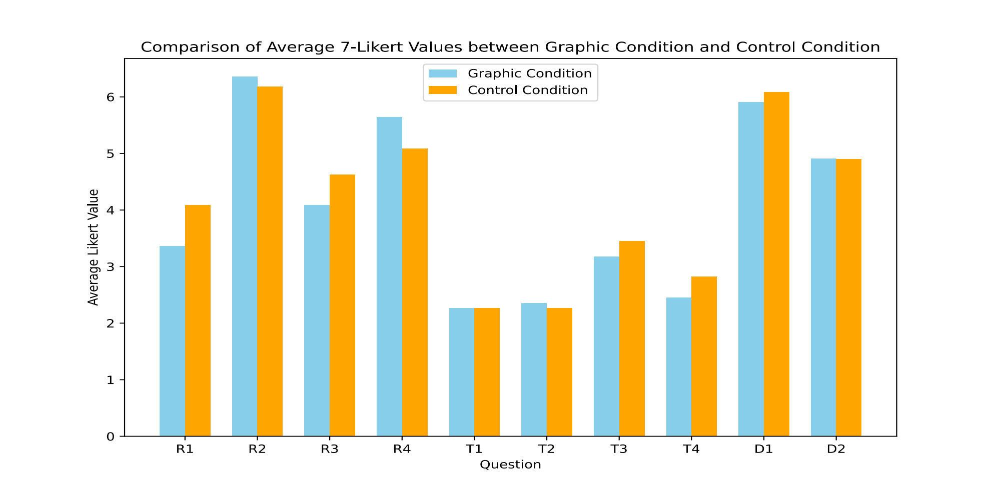
Measuring Dissociation

"Consume recommended content on Instagram for exactly 10 minutes"



T-statistic:-2.967 and Significance: 0.014

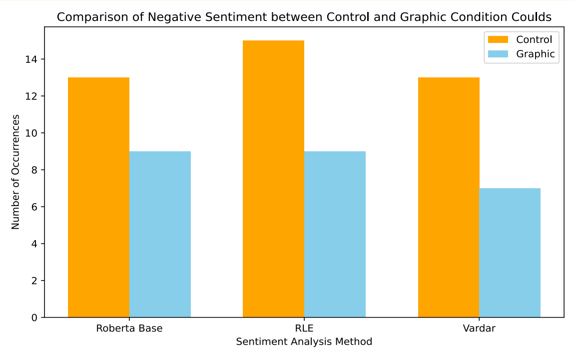
Measuring Perceptions of Social Media

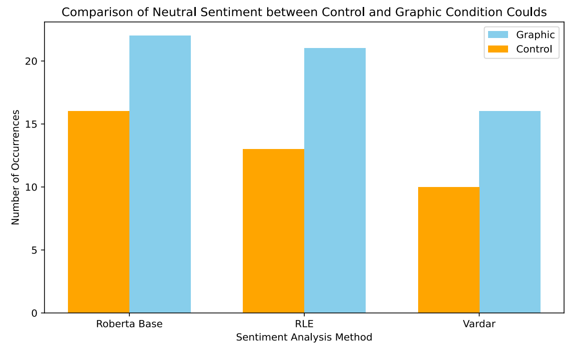


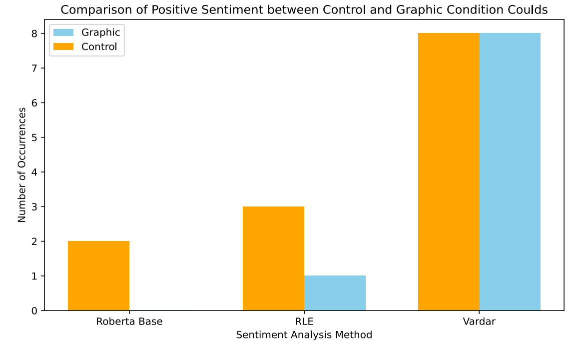
T-statistic:-2.967 and Significance: 0.014

Exit Point Sentiment Analysis

"As a result of using Instagram I feel?"

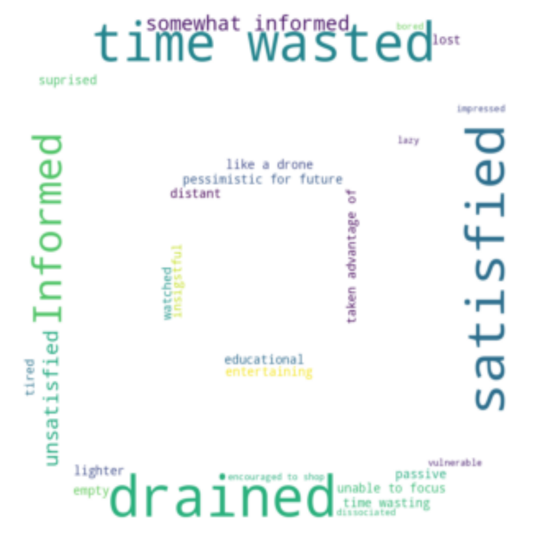




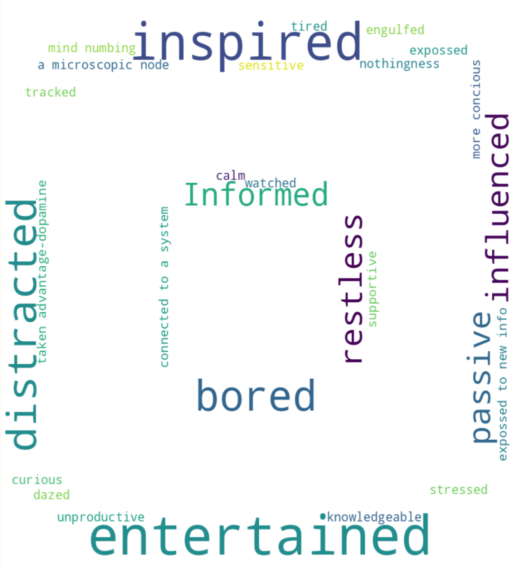


Sentiment	T-statistic	P-value
Negative	-8.000	0.015
Neutral	10.0	0.010
Positive	-2	0.184

Control Condition: Negative



Graphic Condition: Neutral



On the topic of Individualization

Adding people back into “When People and Algorithms Meet”

Jonathan Haley

About

When People and Algorithms meet: User-reported Problems in Intelligent Everyday Applications

Sets about discovering:

- Which problems do users encounter when using intelligent everyday applications?
  Based on the categories:
  Knowledge Base, Algorithm, User Choice, User Feedback
- What kind of support do users want for which problem?

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.

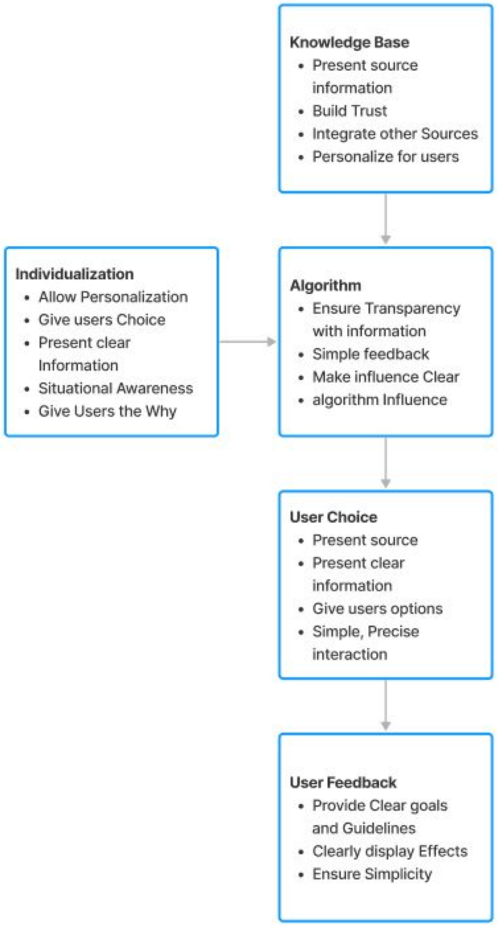
Methodology:

- 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.



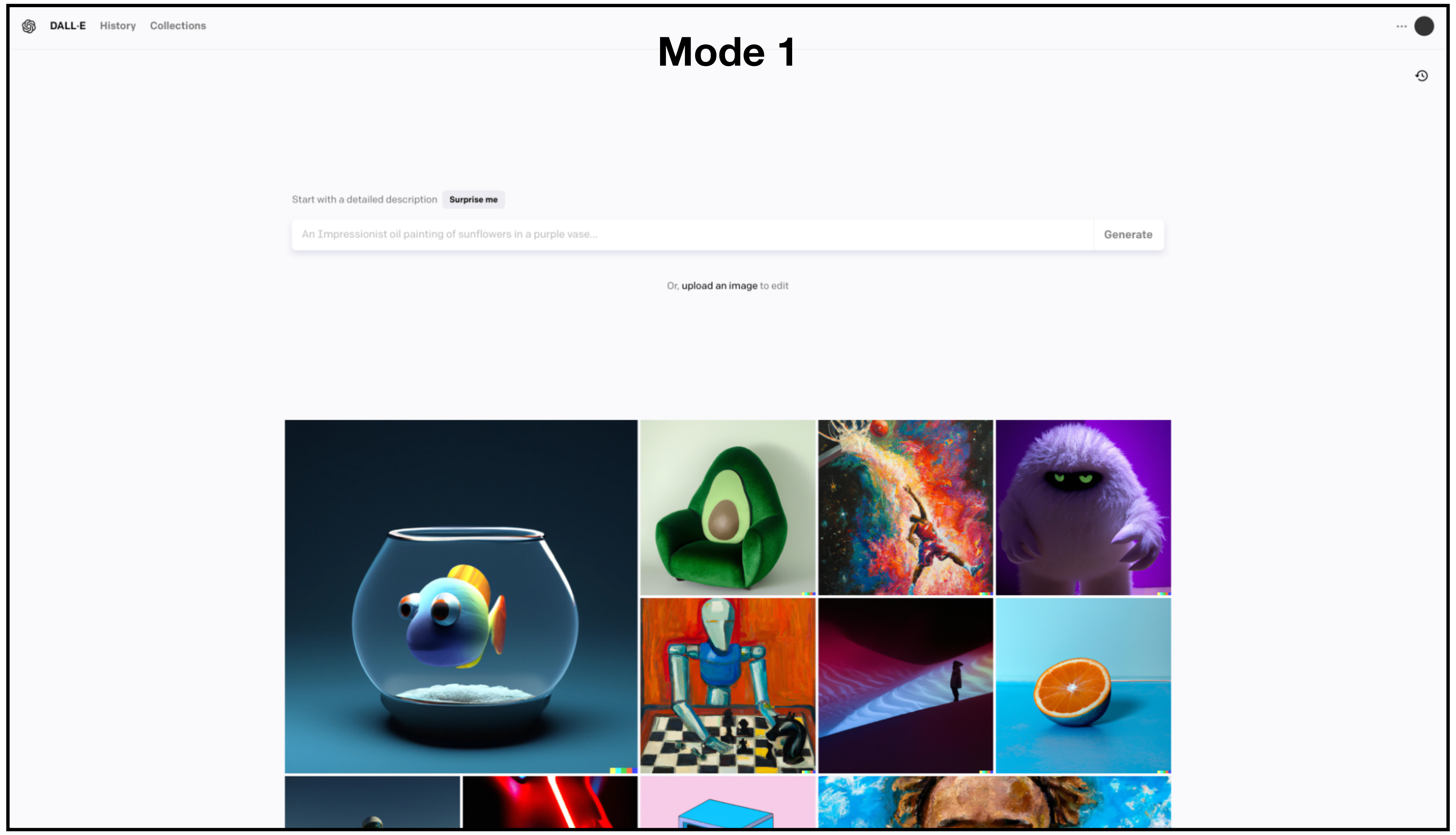
Investigating How Different Modes of Interaction Affect User Experience for Image Generation With DALL-E

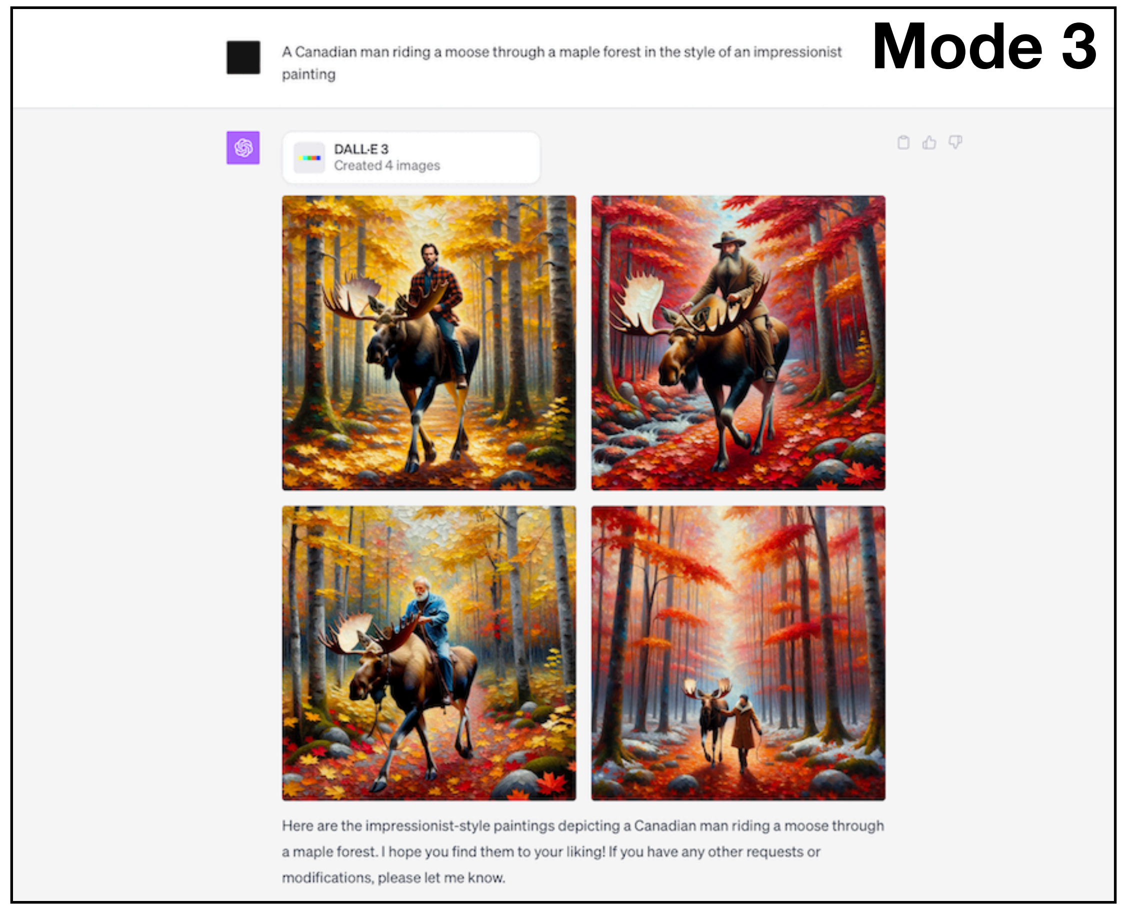
P342: Practical Research in Human-Centred AI

Joseph Cameron (jmc276) - 28/11/2023

Research Question

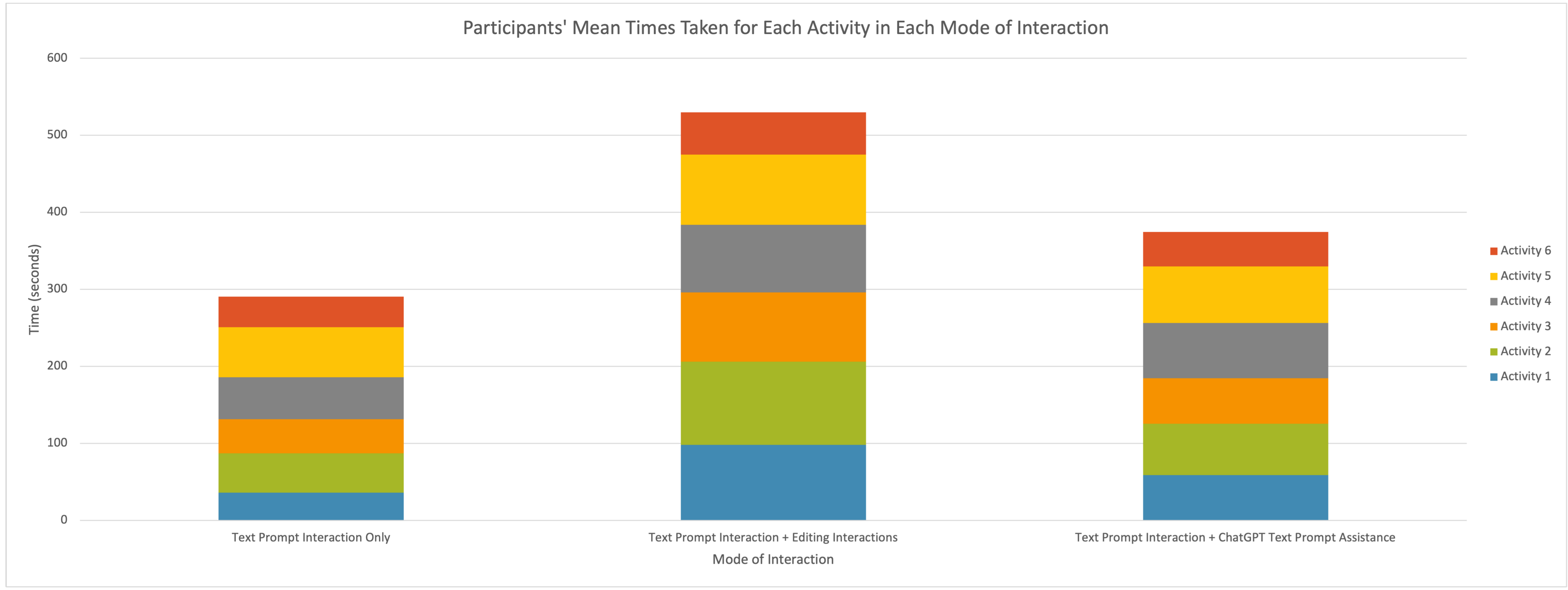
- What is the impact of different interaction modalities on user experience and its relevant time and error-rate usability metrics?
  - Mode 1: Default Text Prompts
  - Mode 2: Text Prompts + DALL-E's Editing Tools
  - Mode 3: Text Prompts + ChatGPT Prompt Assistance





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.



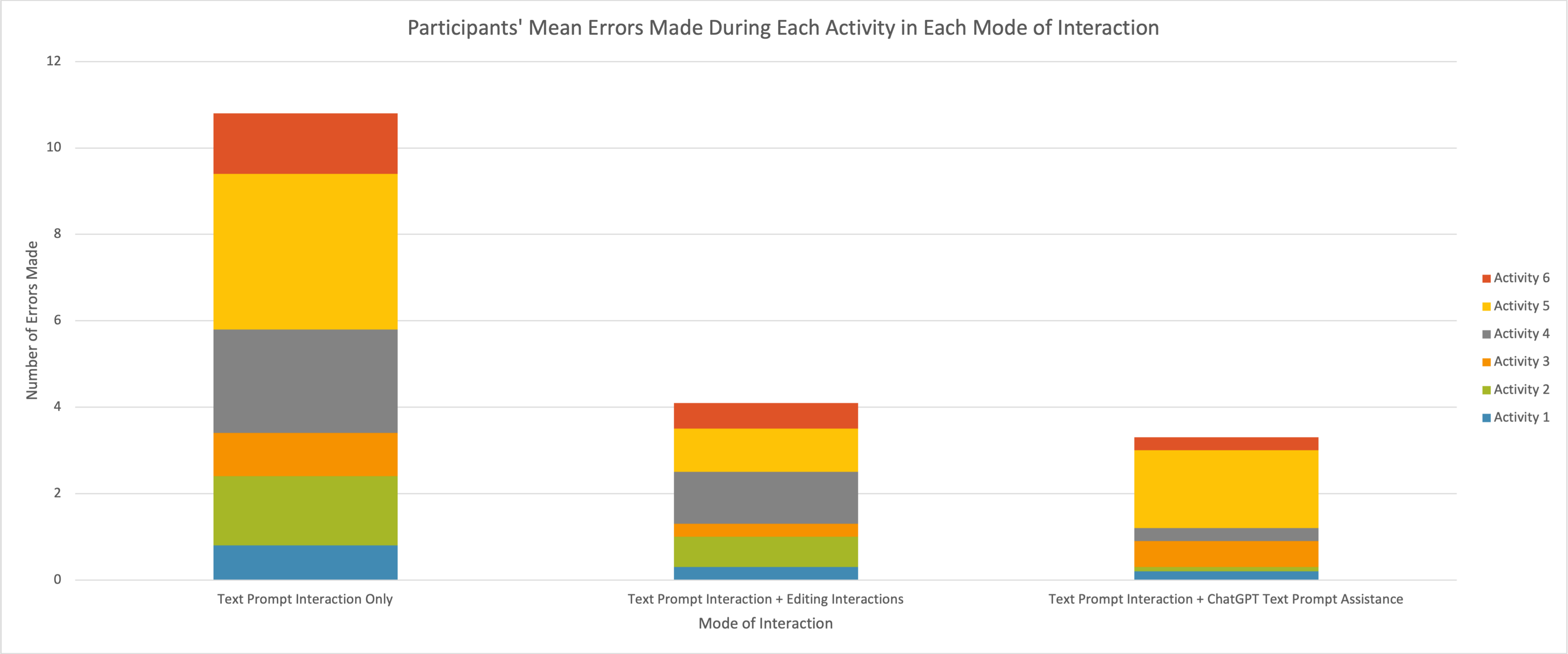


Table 1. Repeated Measures ANOVA Significance Tests for Participants' Mean Times Taken to Complete Each Activity

Activity	F(2, 18)	p–value
1	88.756	< 0.001
2	30.669	< 0.001
3	76.374	< 0.001
4	31.893	< 0.001
5	32.857	< 0.001
6	32.900	< 0.001

Table 2. Repeated Measures ANOVA Significance Tests for Participants' Mean Errors Made During Each Activity

Activity	F(2, 18)	p–value
1	3.532	0.051
2	10.329	0.001
3	5.286	0.016
4	28.009	< 0.001
5	34.696	< 0.001
6	5.356	0.015

Mwalimu Mbaya?

ON CHATGPT AS A SUPPORT TOOL FOR SWAHILI VOCABULARY ACQUISITION

JOSEPHINE REY

RESEARCH QUESTION & METHODS

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

Danry et al, 2023

1 exercise

Campos et al. 2004 (Latin)

2 assessments

Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 352, 1–13.

Alfredo Campos, María Pérez-Fabello and Rocío Gómez-Juncal. (2004). Gender and age differences in measured and self-perceived imaging capacity. Personality and Individual Differences. 37. 1383-1389. 10.1016/j.paid.2004.01.008.

FINDINGS & DISCUSSION

Are the differences between these means statistically significant between groups?

WHY?

ChatGPT Log Data...

- **Poor** explanations

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

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

One sunny morning (Asubuhi), the villagers gathered

YES

Average Proportion of Words Recalled

	After Exercise (T1)	*1 – 2 Days Later (T2)*
Group 1	96.25%	88.75%
Group 2	60.83%	55%
Group 3	74.17%	70%

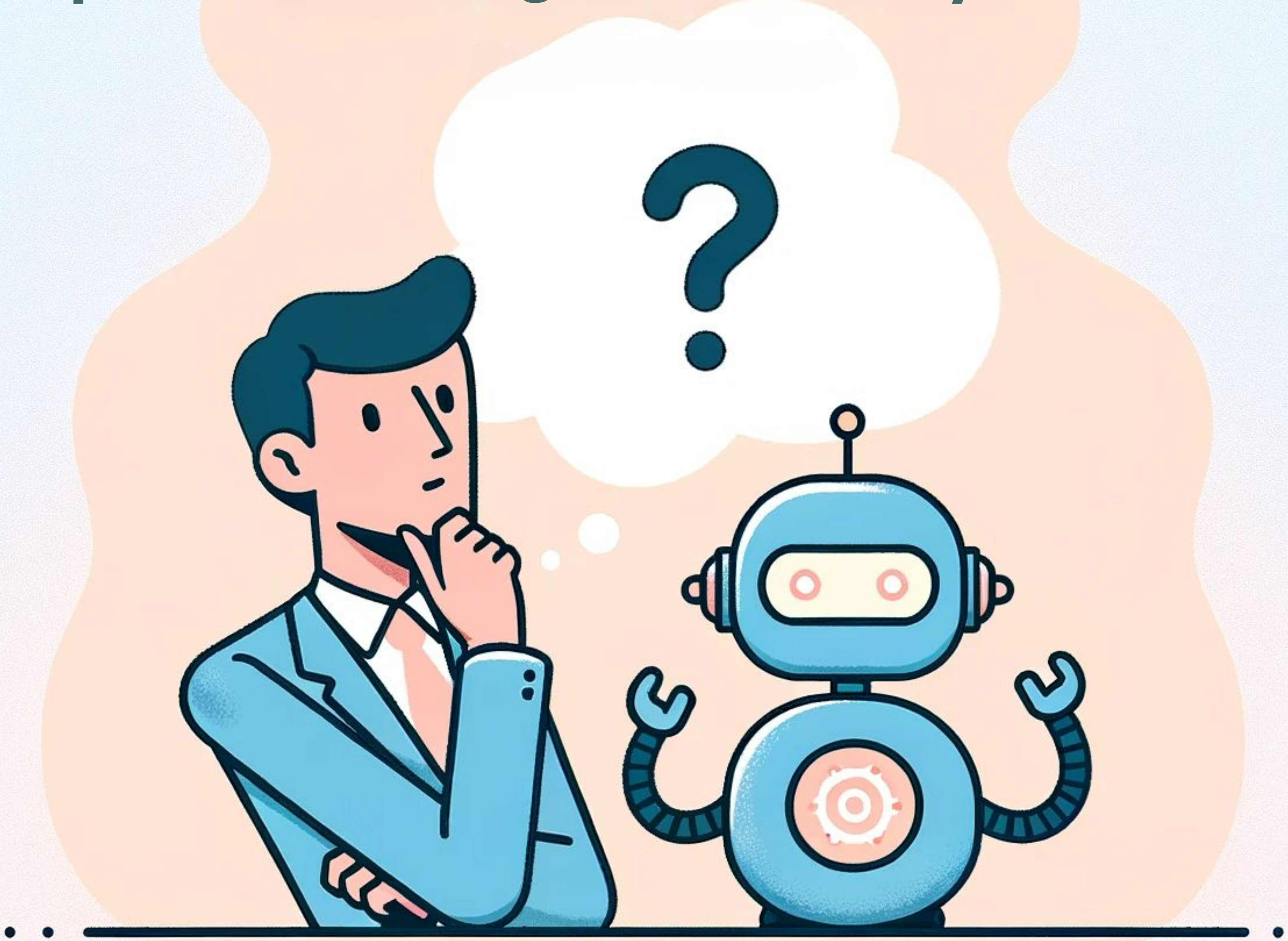
ANOVA TEST: Yes (p = 0) **SIGN TEST:** Yes (p = 0.001)

No: G2 & G3 (p = 0.9935)

Mwalimu Mbaya

"Bad Teacher"

Non-AI-Experts Predicting the Accuracy of LM on QA tasks



Lexin Zhou
(lz473@cam.ac.uk)

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?

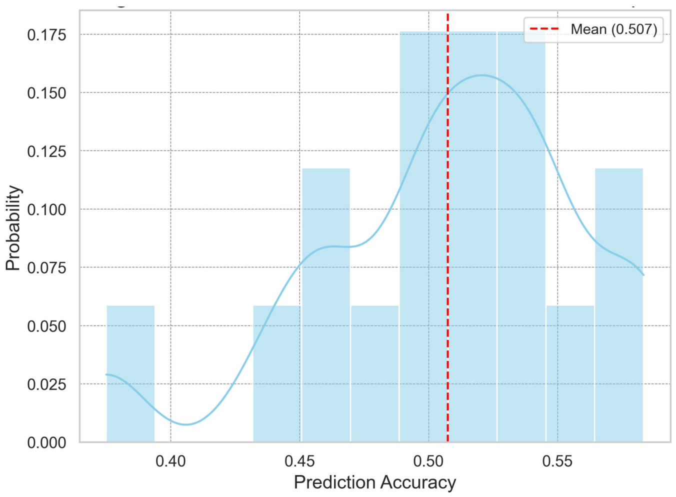
Q2. Can participants **improve** their **predictions** as they continuously **observe more examples** of successes and failures of the LM?

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?

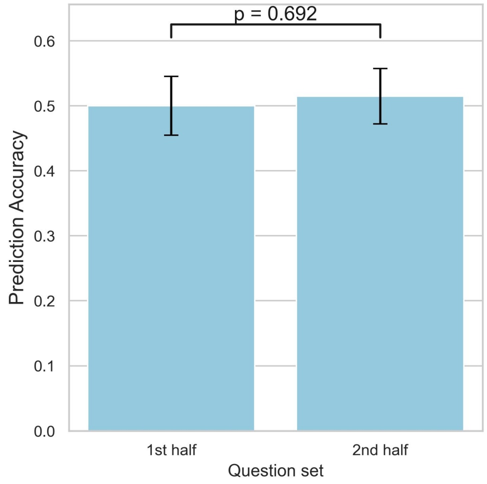
Experimental setup:

- 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.

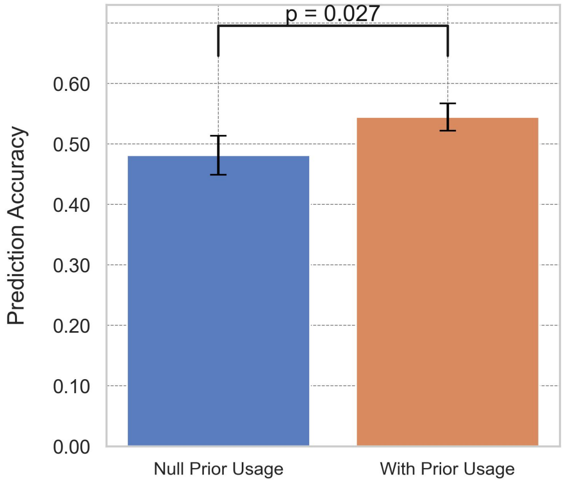
Q1. How predictable is the accuracy of an unfamiliar LM on QA tasks for non-AI-experts?

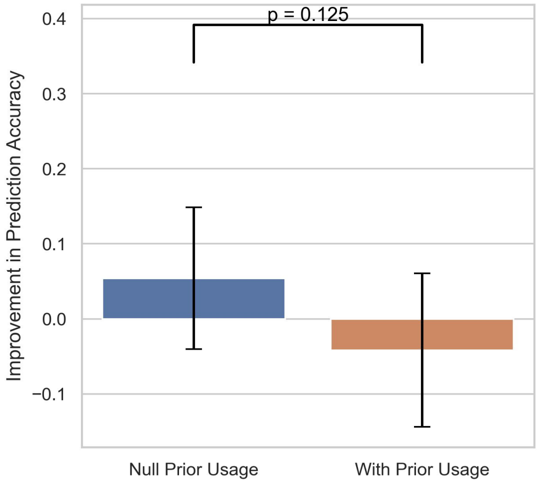


Q2. Can participants **improve** their **predictions** as they continuously **observe more examples** of successes and failures of the LM?



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

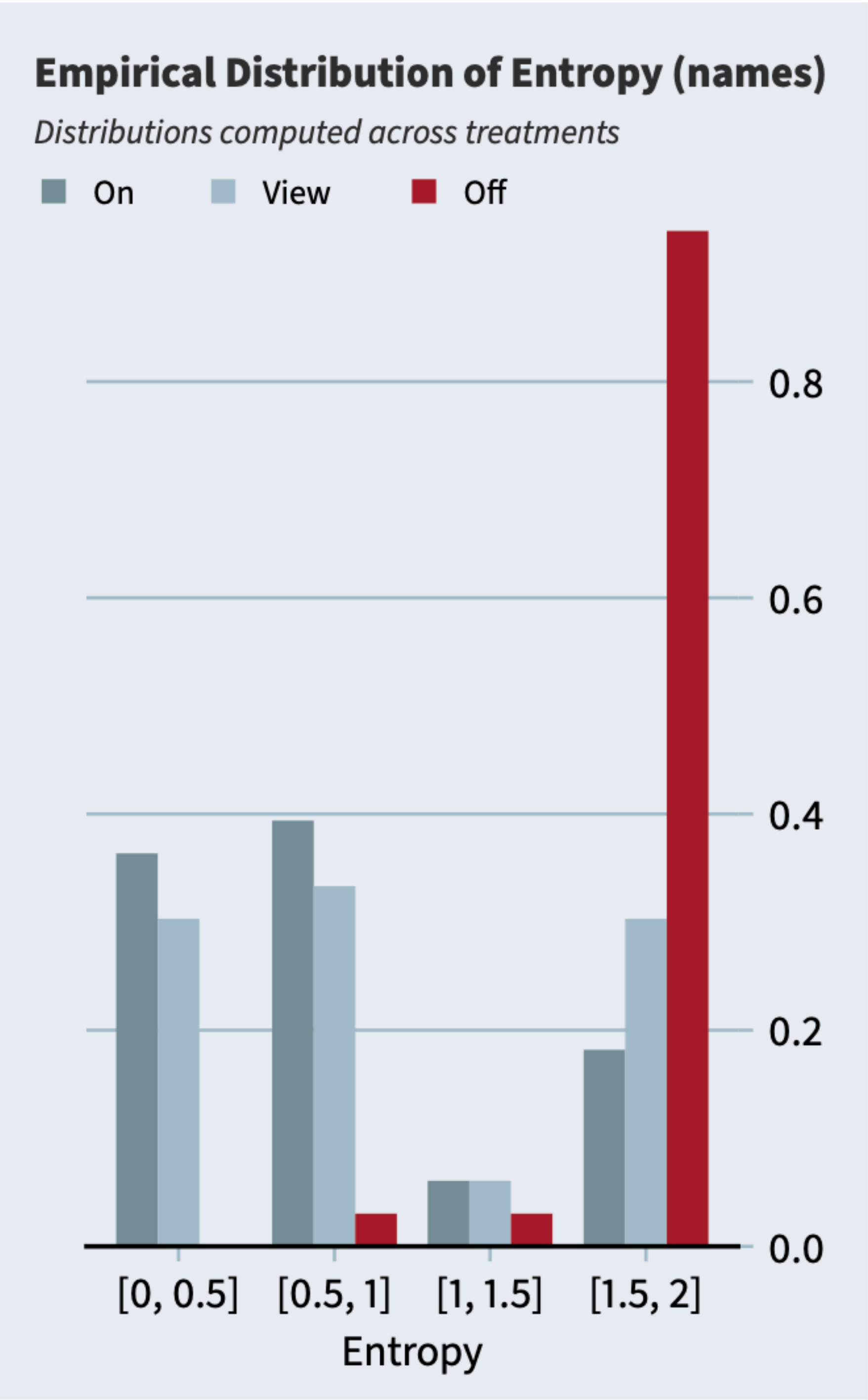
```
let ??? op base :
 fix (fun g ->
   base ++ option (op ++ g) ==> function
    | (e, None) -> e
    | (e, Some(f, e')) -> f e e')
```

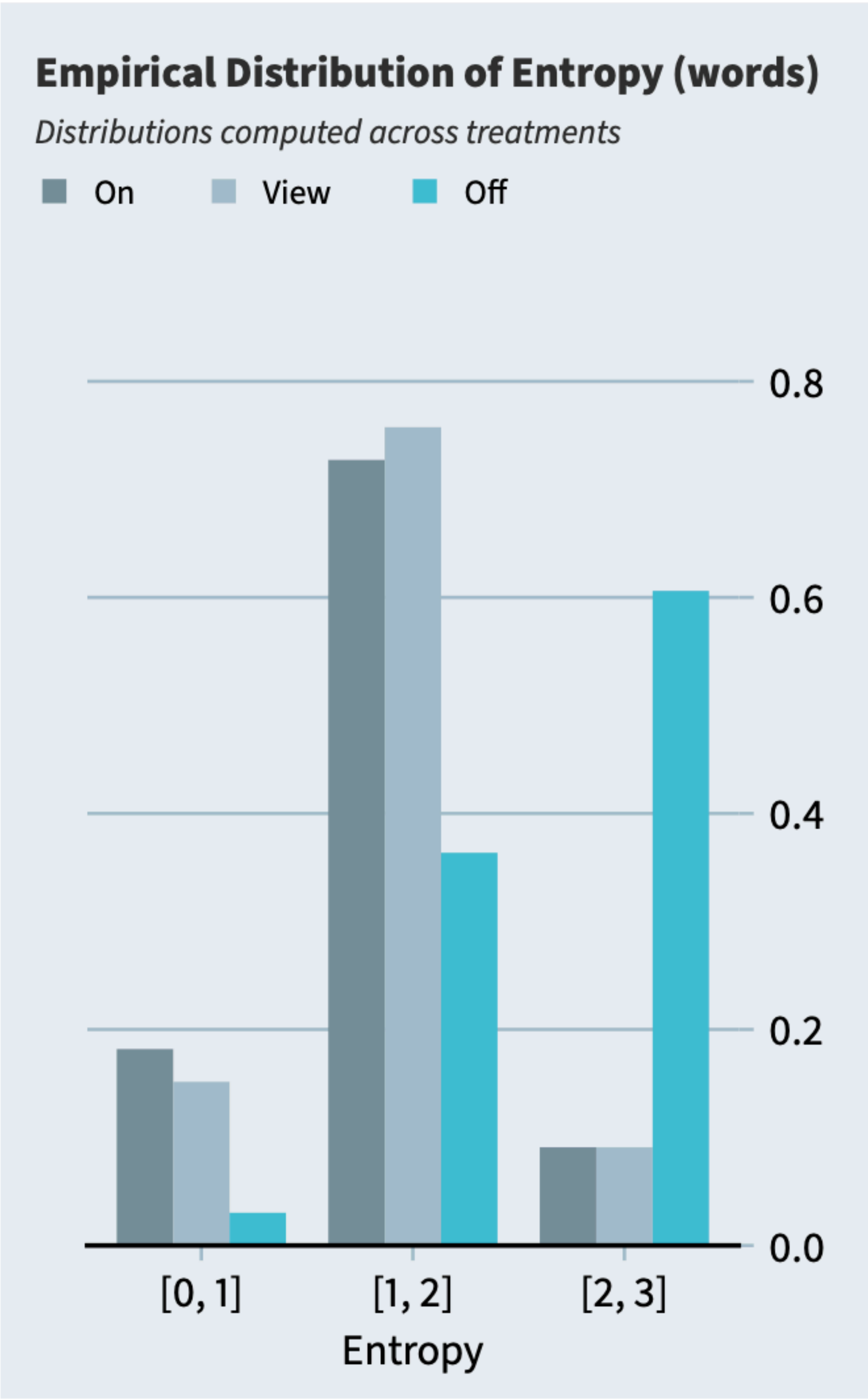
Krishnaswami and Yallop (2019),
A Typed Algebraic Approach to Parsing

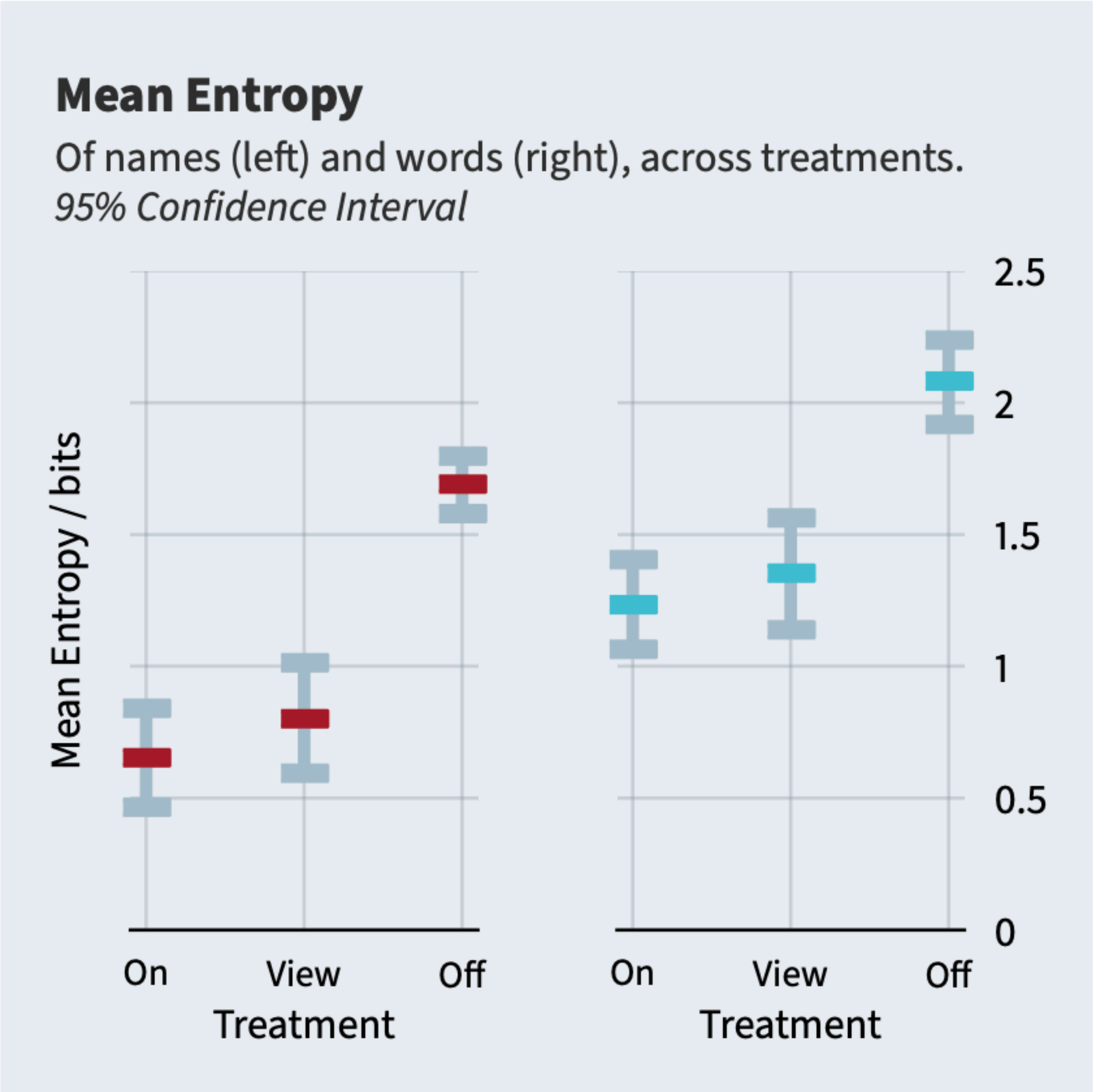
How would you caption this image?

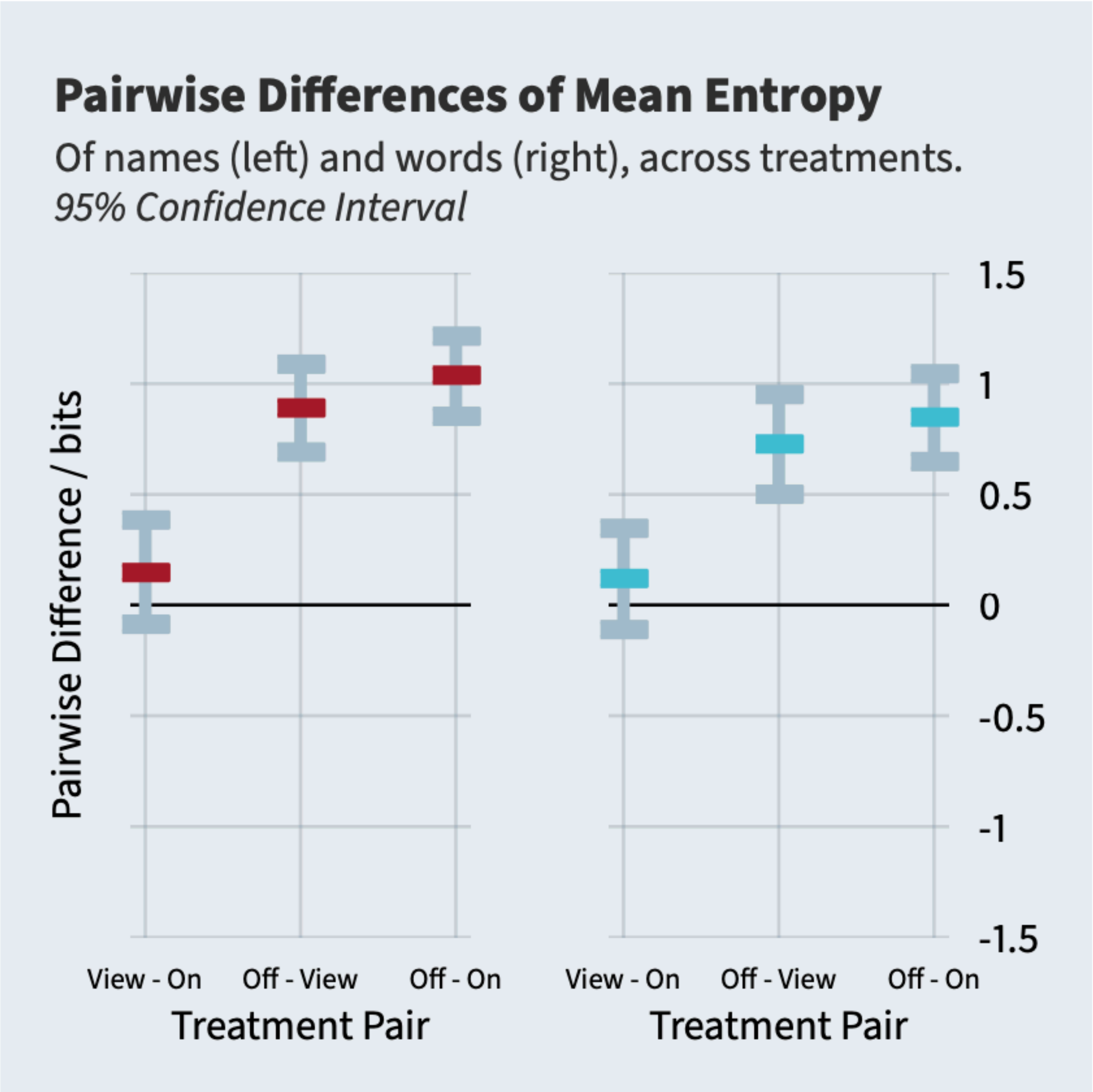
How would you name this abstract object?

Results









t	$P(H(c, \textsc{Off}) > H(c, t))$	95% CI
On	0.848	[0.727, 0.970]
View	0.909	[0.818, 1.000]

t	$P(\text{renamed} \mid t)$	95% CI
On	0.106	[0.061, 0.160]
View	0.197	[0.129, 0.267]
Off	0.258	[0.182, 0.333]

UNIVERSITY OF CAMBRIDGE

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

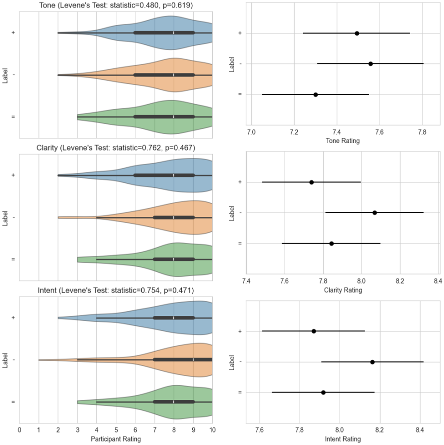
Text-based communication can be stressful or difficult

- 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

Label 1	Label 2	$\hat{y}_2 - \hat{y}_1$	Lower bound	Upper bound	p-value
+	-	0.0647	-0.4346	0.5640	0.9501
+	=	-0.1918	-0.6919	0.3084	0.6395
-	=	-0.2565	-0.7557	0.2428	0.4491

(b) Clarity

Label 1	Label 2	$\hat{y}_2 - \hat{y}_1$	Lower bound	Upper bound	p-value
+	-	0.3283	-0.1849	0.8415	0.2898
+	=	0.1027	-0.4113	0.6168	0.8854
-	=	-0.2256	-0.7388	0.2876	0.5560

(c) Intent

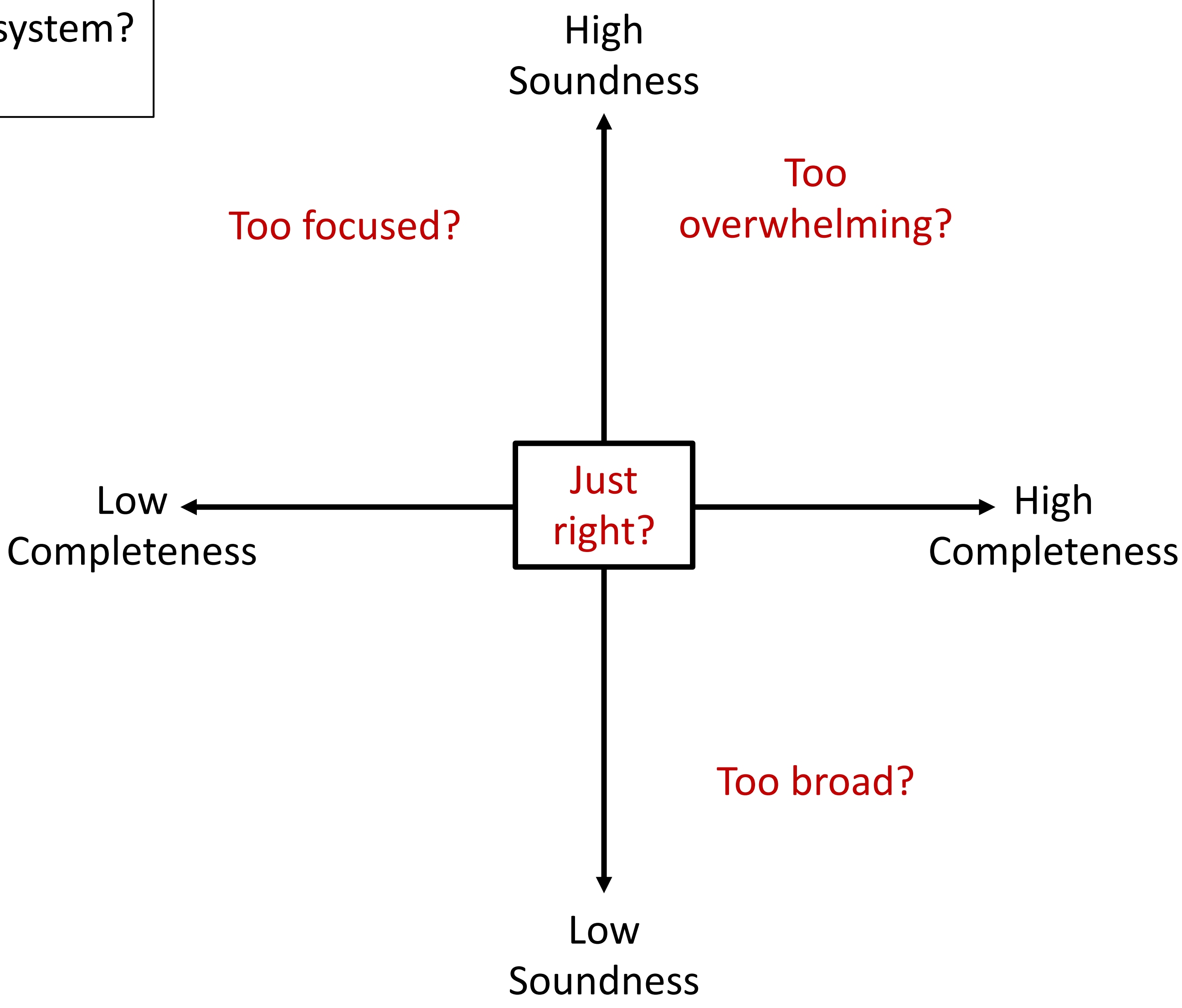
Label 1	Label 2	$\hat{y}_2 - \hat{y}_1$	Lower bound	Upper bound	p-value
+	-	0.2934	-0.2186	0.8054	0.3696
+	=	0.0479	-0.4649	0.5608	0.9737
-	=	-0.2455	-0.7574	0.2665	0.4976

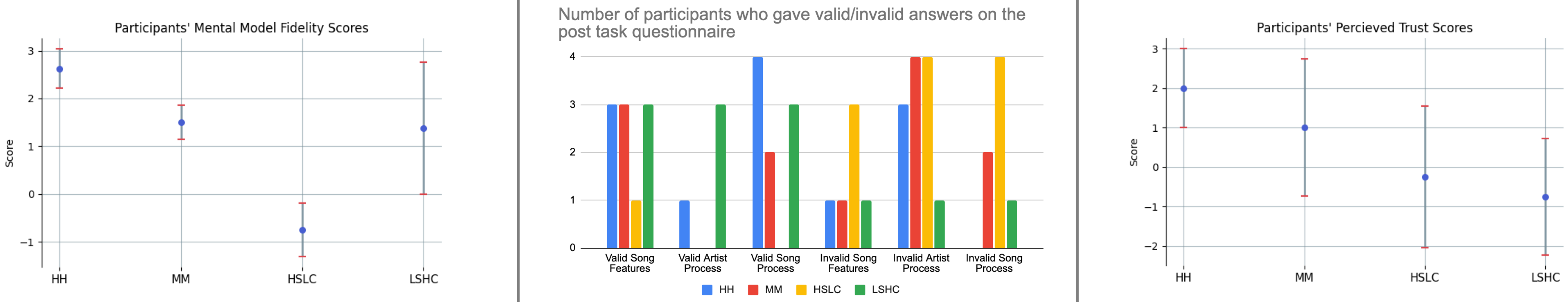
The Goldilocks Zone for Explanations: Finding the Sweet Spot in Recommender Systems

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?





Participants' Mental Model Fidelity Scores
Score
3
2
1
0
−1
HH
MM
HSLC
LSHC
Number of participants who gave valid/invalid answers on the post task questionnaire
4
3
2
1
0
Valid Song Features
Valid Artist Process
Valid Song Process
Invalid Song Features
Invalid Artist Process
Invalid Song Process
HH
MM
HSLC
LSHC
Participants' Percieved Trust Scores
Score
3
2
1
0
−1
−2
HH
MM
HSLC
LSHC

Exploring the Effect of Augmented Writing Systems on Creative Writing Processes and Outcomes

by Sol Dubock

The Premise

Basis Paper: **Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence**

Two text editors:

- Editor-Red (AI-Assisted) *[Google Docs for spellcheck/word completions & GPT3.5 extension for recommended story continuations]*
- Editor-Green (Unassisted) *[Windows Notepad]*

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

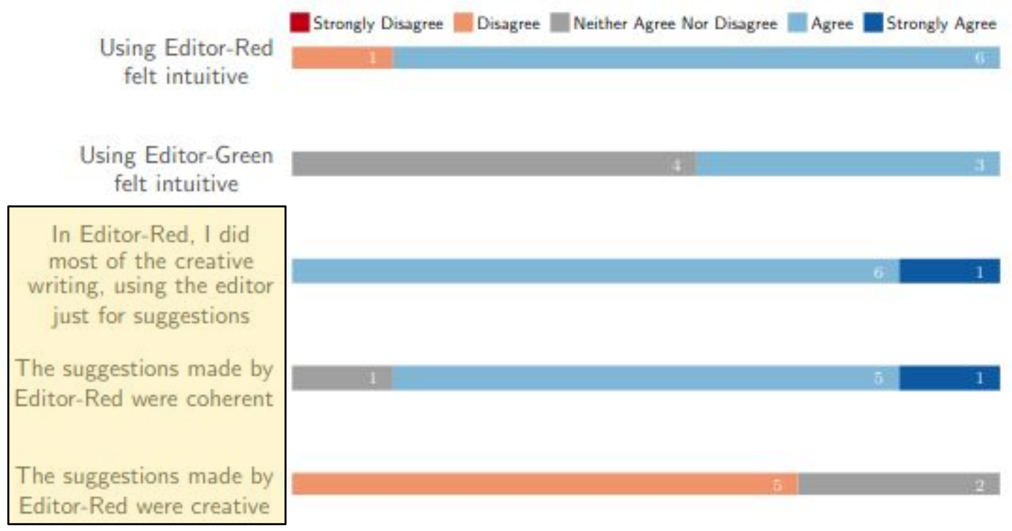
The Results

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”



Novice (e.g. GCSE/ Secondary School English Level)	Intermediate (e.g. Completes some independent creative writing)
2	2

Proficient (e.g. Creative writing is a significant pastime)	Expert (e.g. Published author)
4	1

Measure	Editor-Green (unassisted)	Editor-Red (AI-assisted)	t-value	$t_{14}(0.1)$
Avg. (mean) text length	398.4 words	455.1 words	1.005	1.345
Avg. spelling error count	2	0.14	1.462	1.345
Avg. spelling error rate (per 100 words)	0.597	0.034	1.505	1.345
Avg. grammar error count	1.857	1	0.684	1.345
Avg. grammar error rate (per 100 words)	0.640	0.218	0.8944	1.345
Avg. number of distinct AI-phrases*	N/A	2.714	N/A	N/A
Avg. AI-phrase* rate (per 100 words)	N/A	0.600	N/A	N/A

Impact of Conflict on User Perspectives and Problems with Intelligent Applications

Sophie Walker

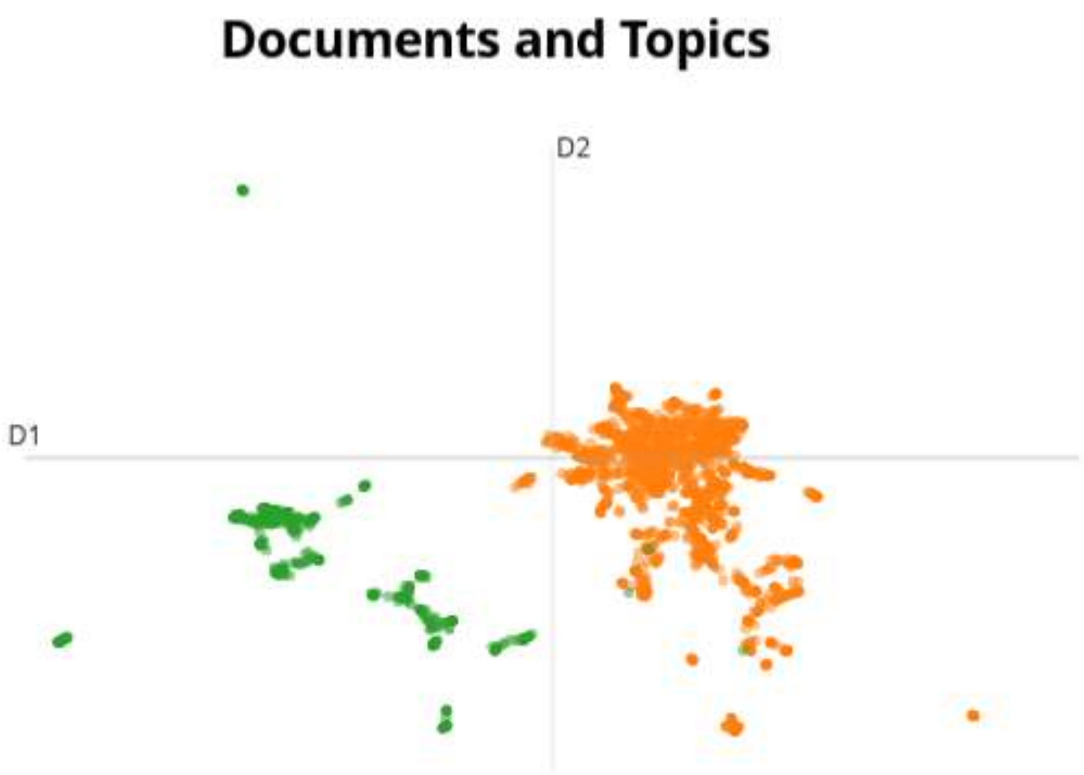
Research Questions

- 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

Eiband, M., Völkel, S. T., Buschek, D., Cook, S., & Hussmann, H. (2019, March). When people and algorithms meet: User-reported problems in intelligent everyday applications. In *Proceedings of the 24th international conference on intelligent user interfaces* (pp. 96-106).
Zhang, H., Wu, C., Xie, J., Kim, C., & Carroll, J. M. (2023). QualiGPT: GPT as an easy-to-use tool for qualitative coding. *arXiv preprint arXiv:2310.07061*.
Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Results



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

- Key Themes: ‘Political Bias’, ‘Removal of Sinai From Maps’, ‘Falsification’ and ‘Omission of Specific Locations’

https://misbar.com/en/editorial/2023/10/21/recent-claims-of-sinai-peninsula-name-removal-from-google-maps-are-inaccurate

Kazenwadel, Daniel, and Christoph V. Steinert. "How User Language Affects Conflict Fatality Estimates in ChatGPT." *arXiv preprint arXiv:2308.00072* (2023).

User reported problems in Spotify DJ

STEPHANIE CHO

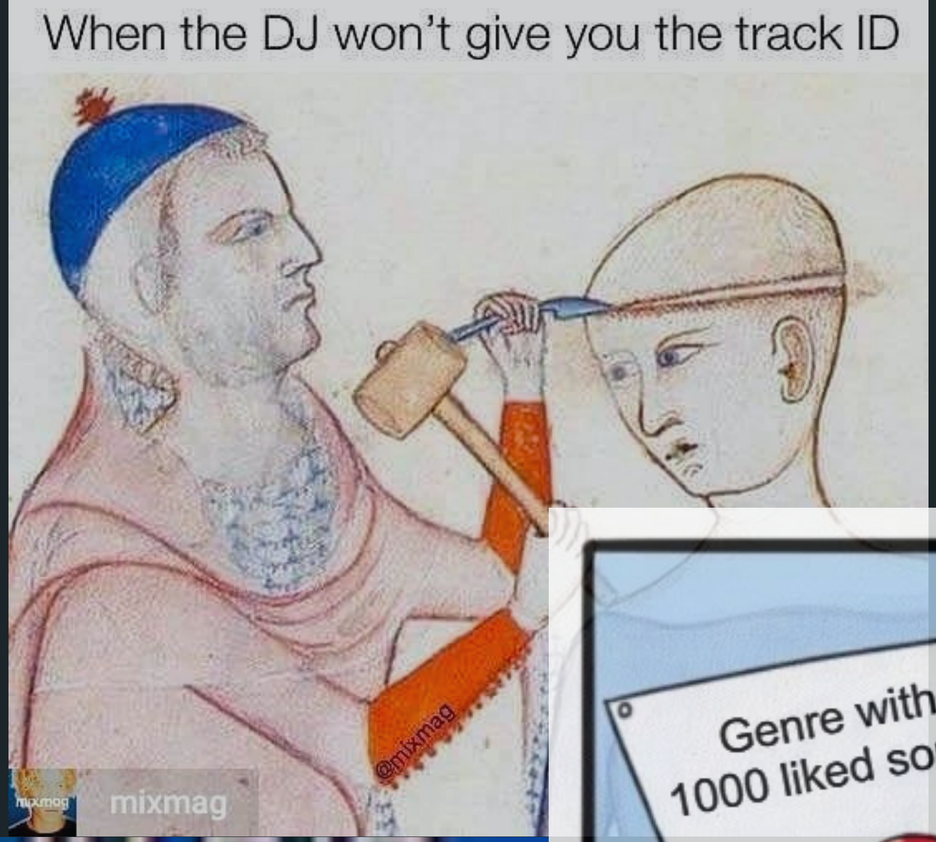


Image credits

HTTPS://WWW.GOOGLE.COM/URL?SA=I&URL=HTTPS%3A%2F%2FWWW.REDDIT.COM%2FR%2FMEMES%2FCOMMENTS%2F1794ZHX%2FDO_YALL_USE_THE_DJ_FEATURE%2F&PSIG=AOVVAW3G02Y-1A7DIARRB0NBBTHU&UST=1701250258592000&SOURCE=IMAGES&CD=VFE&OPI=89978449&VED=0CBQQJHXQFWOTCLDLQZYX5OIDFQAAAAADAAAAABBE
HTTPS://WWW.GOOGLE.COM/URL?SA=I&URL=HTTPS%3A%2F%2FWWW.TIKTOK.COM%2FDISCOVER%2FSPOTIFY-DJ-MEME&PSIG=AOVVAW3G02Y-1A7DIARRB0NBBTHU&UST=1701250258592000&SOURCE=IMAGES&CD=VFE&OPI=89978449&VED=0CBQQJHXQFWOTCLDLQZYX5OIDFQAAAAADAAAAABBM
HTTPS://WWW.GOOGLE.COM/URL?SA=I&URL=HTTPS%3A%2F%2FWWW.PINTEREST.FR%2FPIN%2F392094711310430697%2F&PSIG=AOVVAW3G02Y-1A7DIARRB0NBBTHU&UST=1701250258592000&SOURCE=IMAGES&CD=VFE&OPI=89978449&VED=0CBQQJHXQFWOTCLDLQZYX5OIDFQAAAAADAAAAABBU

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.

REDDIT AS SOURCE OF INFORMATION

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

DJ AS A CONCEPT

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

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?

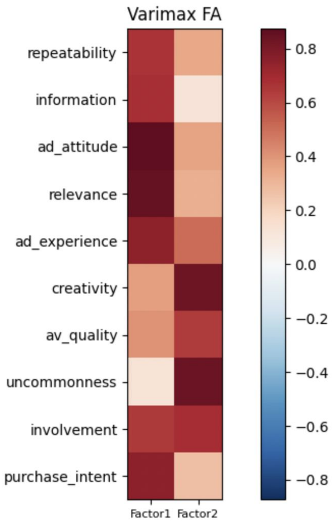
The Big Five personality traits

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

Sentiment of Video Ads

	Synonyms
Active	energetic, adventurous
Alert	attentive, curious
Amusing	humored, laughing
Calm	soothed, peaceful

Results



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

	Precision	Recall	Accuracy	F1
SVM	0.45	0.53	0.59	0.57
CART	0.4	0.44	0.47	0.44

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

	Overall Experience	Overall Quality
Extraversion	0.20 (0.23)	0.25 (0.06)
Agreeableness	0.26 (0.22)	0.22 (0.18)
Conscientiousness	-0.23 (0.15)	0.03 (0.79)
Neuroticism	0.16 (0.35)	-0.02 (0.87)
Openness	-0.07 (0.69)	-0.39 (<0.01)*
R2	0.23	0.52

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

	Overall Experience	Overall Quality
Extraversion	0.10 (0.28)	0.11 (0.38)
Agreeableness	0.41 (<0.01)*	0.31 (0.07)
Conscientiousness	-0.15 (0.12)	-0.25 (0.05)
Neuroticism	0.21 (0.05)	0.37 (<0.01)*
Openness	-0.09 (0.31)	0.18 (0.16)
R2	0.59	0.52

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

[1] Touvron, Hugo, Louis Martin, and Kevin Stone. “Llama 2: Open Foundation and Fine-Tuned Chat Models,”

[2] Anik, Ariful Islam, and Andrea Bunt. “Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency.” In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13. Yokohama Japan: ACM, 2021. https://doi.org/10.1145/3411764.3445736.

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