Transparency in LLMs: Effects of Explanation Fidelity on Users' Mental Models and Trust

Practical Research in Human-Centred AI

Ahmed Khan

Background

- Global explanations are one mechanism for bringing about transparency in algorithms [1].
- The more *complete*, the better the effects [2].

Research Questions

RQ–1 How does explanation fidelity affect users' mental models?

RQ–2 How does explanation fidelity affect users' trust intentions & beliefs?

RQ–3 How do users' perceptions of explanations vary by their fidelity?

Method

Pre-Treatment

Measure: prior knowledge and initial trust.

Treatment

Issue: a randomly-chosen explanation.

Post-Treatment

Measure: learning (new, surprising understanding), cognitive load, readability (investment, difficulty, concentration), mental model fidelity, and trust intentions & beliefs.

Explanation Fidelity

Completeness

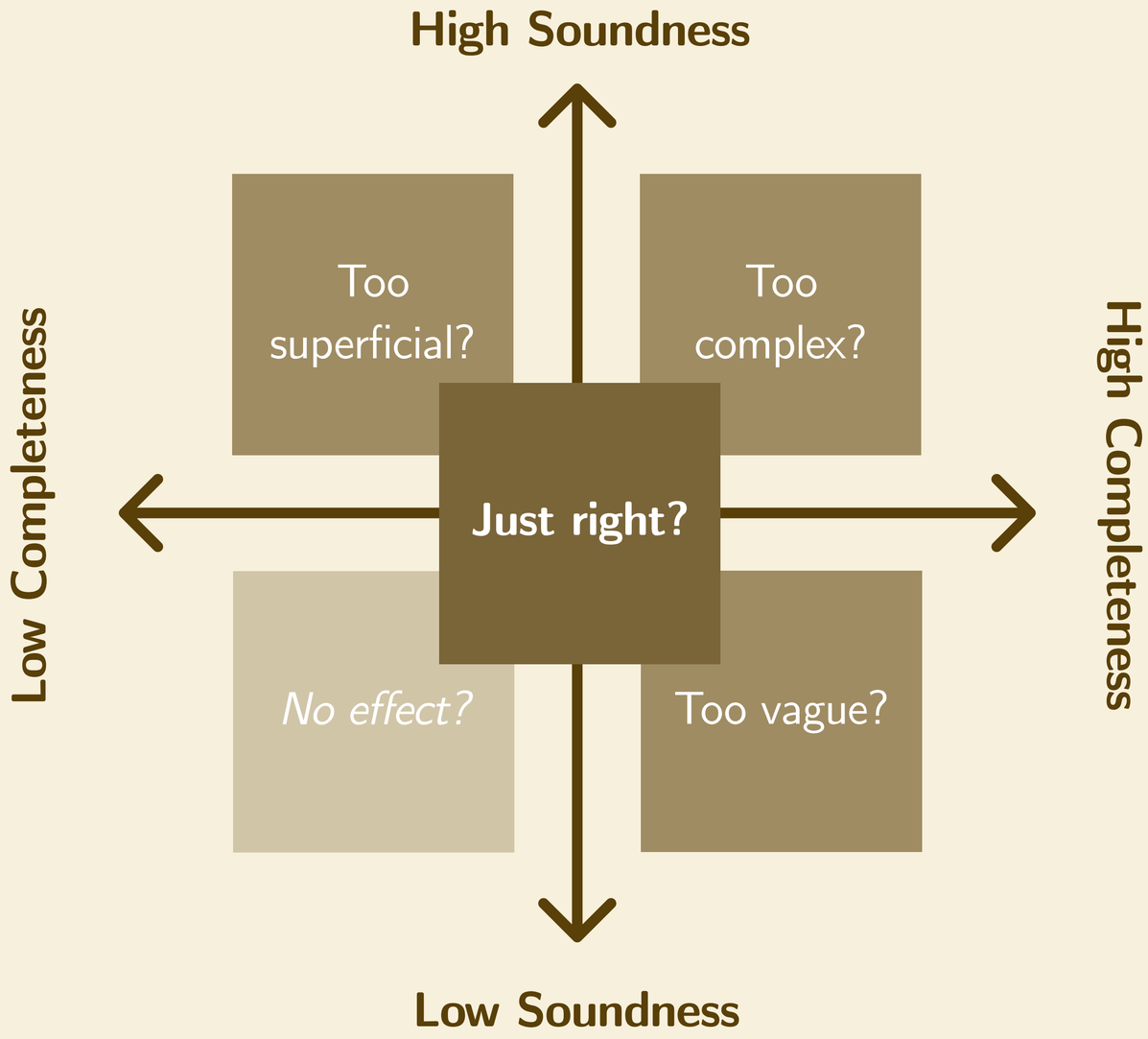
How much an explanation describes *all* of the underlying system.

Soundness

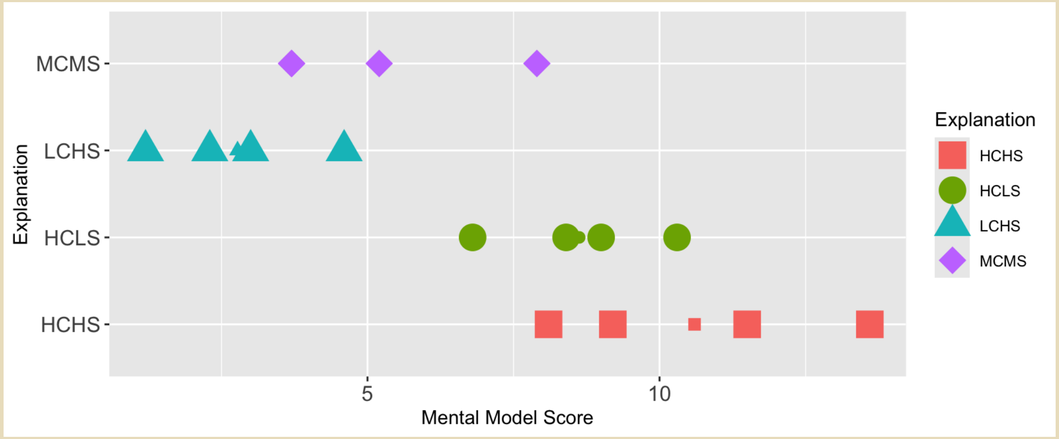
How truthful an explanation is in describing the underlying system.

Preparation

Explanations were written as summaries of specialist, generalist, and commercial sources on LLMs.



Results



Discussion

RQ–1 HsHc had the largest positive effect on mental model scores ($p < 0.01$), followed by HcLs; LcHs had the smallest ($p < 0.01$).

RQ–2 HsHc had the largest negative effect on trust beliefs ($p < 0.05$), followed by McMs; LcHs had the largest positive effect ($p < 0.01$).

RQ–3 HsHc created the largest cognitive load ($p < 0.05$); HcLs and LcHs were just as demanding.

[1] Emilee Rader et al. (2018) Explanations as Mechanisms for Supporting Algorithmic Transparency. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18).
[2] Kulesza, T. et al. (2013). Too much, too little, or just right? Ways explanations impact end users' mental models. Proceedings of IEEE Symposium on Visual Languages and Human-Centric Computing

UNIVERSITY OF CAMBRIDGE

Can User Trust in News Improve Through Informed Reading? Leveraging LLMs for Bias Detection

Andres Alvarez Olmo – aa2650

Department of Computer Science and Technology

Introduction

- **Global decline in trust:** Trust in news media is at an all-time low, with only 38% expressing confidence in news sources globally.
  - Political bias and lack of transparency

- **Key challenge:**
  - Independent human analysis is expensive
  - Vast amount of news articles published every day

- **Study aim:** Investigate if GPT-4-generated bias and tone analyses can improve perceptions of:
  - **Credibility**
  - **Objectivity**
  - **Trust**

Literature Review

- **Existing efforts:** Transparency strategies have been proposed but few studies evaluate how these tools impact audience trust in real-world scenarios.

- **Role of LLMs:** LLMs, like GPT-4, excel in text classification, bias detection, and sentiment analysis.

- **Challenges with LLMs:** Potential for introducing biases, "black-box" limitations, and computational demands.

- **Alternatives:** Lexicon based methods → rule-based approach → simpler and more compute efficient but less accurate in nuanced situation.

Methods

- **Participants:** 40 university students aged 18-34.

- **Experimental design:** between-subjects design:
  - **Control group:** 3 articles presented without additional context.
  - **Treatment group:** 3 articles accompanied by bias and tone analyses generated by GPT-4.
  - **Hypothesis testing** to determine if there was a different between groups

- **Metrics evaluated:** Credibility, Objectivity, and Trust, rated on a 5-point Likert scale.

- **Statistical analysis:**
  - Reported mean, variance and standard deviation.
  - Shapiro-Wilk test used due to data non-normality.
  - Mann-Whitney U to test data significance (non-parametric and no normality assumptions)

Listing 1: Bias Analysis and Bias Score GPT-4 Prompt

```
The following is a news article. Read it and perform the task that
    follows. Respond with a JSON object of key-value pairs.

####################

{article}

####################

Task: Determine the political leaning of this article within the U.K.
    political context. Is it supporting the Conservative party or the
    Labour Party? Supporting a party can mean supporting its viewpoints,
     politicians, or policies. Provide reasoning for your answer.

1. Instruction: Give a short paragraph summarizing in what ways the
    article supports the Conservative party or the Labour Party.
Key: "reason"
Value: A paragraph of text.

2. Instruction: Give a number from -5 to 5, with -5 indicating strong
    support for the Labour Party and 5 indicating strong support for the
     Conservative Party. A value of 0 indicates that the article has no
    clear political leaning towards either side.
Key: "lean"
Value: An integer number from -5 to 5.

Do not return anything except the JSON object of key-value pairs as
    output.
```

Listing 2: Tone Analysis and Tone Score GPT-4 Prompt

```
The following is a news article. Read it and perform the task that
    follows. Respond with a JSON object of key-value pairs.

####################

{article}

####################

Task: Determine the political leaning of this article within the U.K.
    political context. Is it supporting the Conservative party or the
    Labour Party? Supporting a party can mean supporting its viewpoints,
     politicians, or policies. Provide reasoning for your answer.

1. Instruction: Give a short paragraph summarizing in what ways the
    article supports the Conservative party or the Labour Party.
Key: "reason"
Value: A paragraph of text.

2. Instruction: Give a number from -5 to 5, with -5 indicating strong
    support for the Labour Party and 5 indicating strong support for the
     Conservative Party. A value of 0 indicates that the article has no
    clear political leaning towards either side.
Key: "lean"
Value: An integer number from -5 to 5.

Do not return anything except the JSON object of key-value pairs as
    output.
```

Control Group

Please read this article and answer the following questions

https://www.bbc.co.uk/news/articles/cx2lk4ywdn2o

To what extent do you perceive the source of this article to be credible? *
(1 = Not Credible at All, 5 = Fully Credible)

1 2 3 4 5

○ ○ ○ ○ ○

How objective do you find the tone and content of this article? *
(1 = Not Objective at All, 5 = Completely Objective)

1 2 3 4 5

○ ○ ○ ○ ○

What is your overall level of trust in the information provided by this article? *
(1 = No Trust at All, 5 = Complete Trust)

1 2 3 4 5

○ ○ ○ ○ ○

Treatment Group

Please read the bias/tone analysis and then the article

Bias Analysis: The article showcases support for policies typically associated with the Labour Party, such as the introduction of 'respect orders' to combat anti-social behaviour, which are described as an evolution of the ASBOs from the New Labour era. Prime Minister Sir Keir Starmer and Home Secretary Yvette Cooper are highlighted as key proponents of these measures, reinforcing their commitment to tackling community issues. Criticisms from other political groups, such as the Liberal Democrats, serve to contrast the Labour Party's approach with alternative viewpoints, subtly framing Labour as more action-oriented.

Bias score: -4

(-5 indicating strong support for the Labour Party and 5 indicating strong support for the Conservative Party)

Tone Analysis: The article maintains a neutral tone overall, presenting both the potential benefits and criticisms of the proposed 'respect orders.' While it highlights the government's intent to tackle anti-social behaviour and describes the measures as 'simple but effective,' it also provides significant space for criticism from civil liberties groups and opposition leaders, balancing the portrayal of the initiative.

Tone score: 0

(-5 indicating a very negative tone and 5 indicating a very positive tone)

https://www.bbc.co.uk/news/articles/cx2lk4ywdn2o

Results

- **Dependent variables:** Slightly lower mean scores for the Treatment group across Credibility, Objectivity and Trust.

- **Data was not normally** distributed for any of the metrics according to the Shapiro-Wilk test.

- **No statistically significant differences** between Control and Treatment groups for any of the three metrics ($p > 0.05$) according to the Mann-Whitney U test.

- **Failed to reject the null hypothesis**

Metric	P-Value
Credibility	0.239
Objectivity	0.271
Trust	0.286

Mann-Whitney U Test results

Discussion

- **Insights:** While cheaper than human annotation, bias detection by LLMs alone may not shift trust perceptions significantly.

- **Limitations:**
  - Small, homogeneous sample size consisting only of university students
  - Potential bias introduced by LLMs in explanations

- **Future research:**
  - Larger and more diverse samples → improve statistical significance
    - Analyse covariates such as age, digital literacy or political orientation
  - Add human verification to explanations to avoid LLM bias → improve trust

Evaluating AI-framed Questioning

EXTENDING AND REPLICATING A STUDY ON THE USE OF AI-FRAMED QUESTIONING

Cyrus Ho

3 December 2024

Part II Module:
Practical Research on
Human-centered AI

O V E R V I E W

- Brief Introduction
- Experiment Design
- Results
- Discussion

THE ORIGINAL STUDY

“Our results show that compared to no feedback and even causal AI explanations of an always correct system, AI-framed Questioning significantly increase human discernment of logically flawed statements.” -Danry et al (2023)

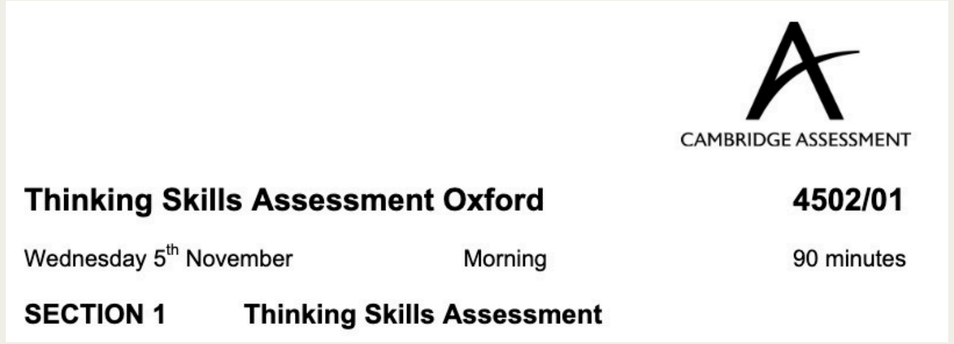
- AI-Framed questioning builds upon a method of Socratic Questioning, asking open-ended questions to stimulate thinking.
- Focused only on one type of logical fallacy - Hasty Generalisation
- AI-framed questioning was effective in increasing human discernment of logically flawed statements

MY RESEARCH QUESTIONS

1. Do humans perform better at discerning the logical structure and validity of statements when they receive feedback from AI systems compared to when they work alone?
2. How does AI-framed Questioning affect participants' discernment of logical structure and confidence of their discernment?

EXPERIMENT DESIGN

- Each participant answers 24 questions in 3 sections
- Each section represents a different treatment (No AI support, AI provided answer and AI-framed Questioning)
- Questions are adapted from the Oxford and Cambridge Thinking Skills Assessment admissions test for certain academic courses



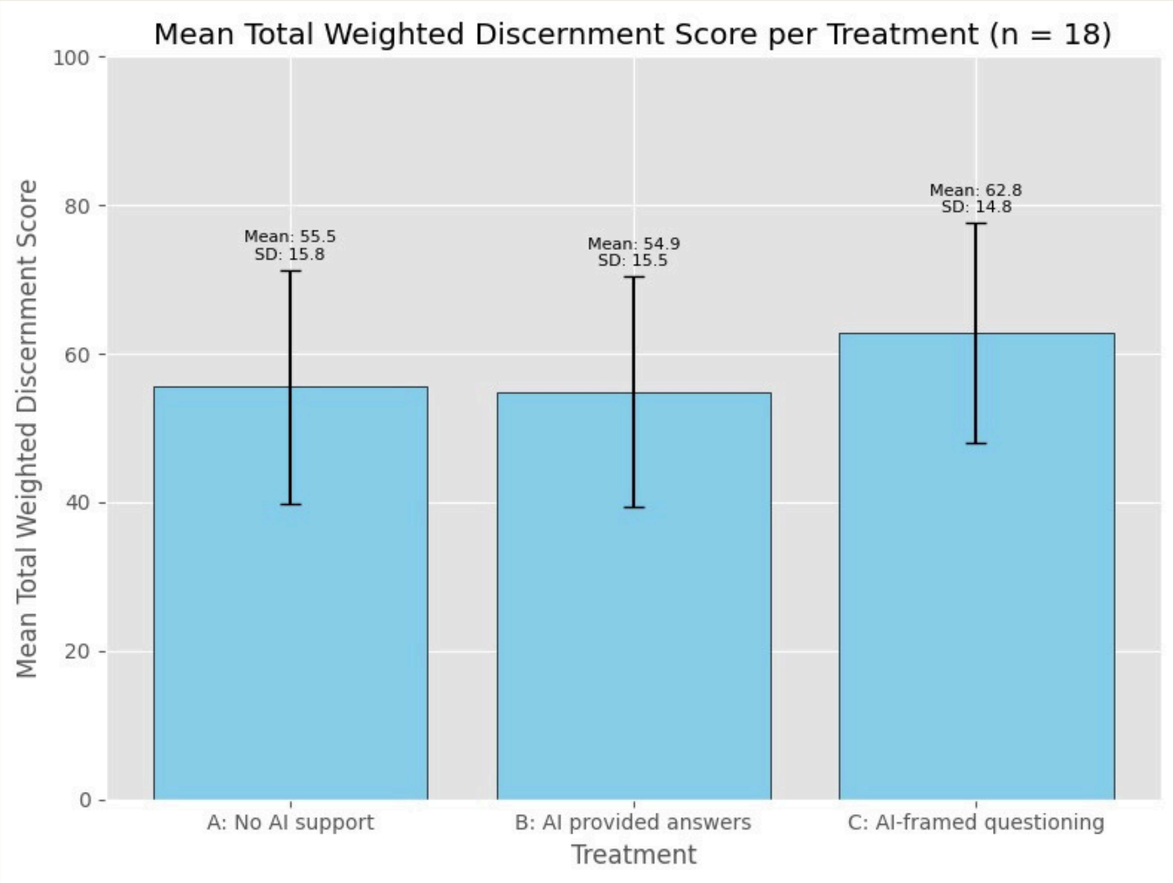
CAMBRIDGE ASSESSMENT

Thinking Skills Assessment Oxford **4502/01**

Wednesday 5th November Morning 90 minutes

SECTION 1 **Thinking Skills Assessment**

RESULTS



- ANOVA reveals an overall **significant** effect of the use of AI-framed Questioning on participant's performance in the assessment ($F = 4.61$, $p = .017$)
- Pairwise comparisons show a significant positive effect for the use of **AI-framed questioning over the other treatments.**
- Cohen's d calculation indicates a **moderate to large effect size**. (A vs C: -0.66, B vs C: -0.59)

THEMATIC ANALYSIS

Critical thinking is hard

The general performance in the assessment shows that participants find the logical evaluation of arguments difficult.
Non Cambridge University students found the task much harder, with lower scores.

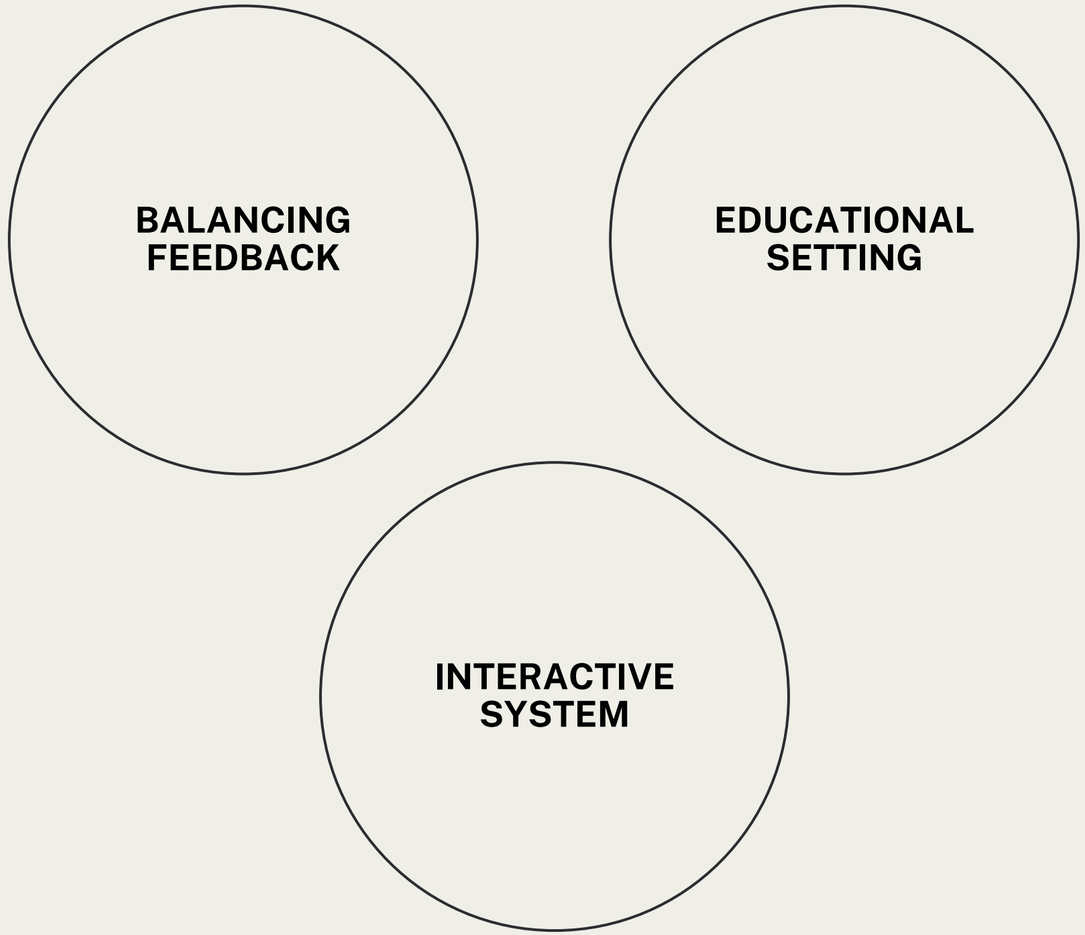
Trust in AI

AI provided answers with no explanation were not helpful to participants.
Participants only accept AI feedback when there is additional information that allows them to also evaluate the AI.
This corroborates with previous literature on trust and use of AI responses
(Ribeiro et al, 2016)

Sense of reward

Participants preferred AI-framed questioning as they still enjoyed a sense of reward and accomplishment when completing the task.
These are important motivators when establishing long term learning (**Knowlton & Castel, 2022)**

ROOM FOR FUTURE WORK



Balancing Feedback

There is a challenge to balance the guidance provided by the AI-framed questions. Some questions may provide too much support which reduces the cognitive thinking required. On the other hand, it could even be detrimental as the question confuses the reader further.

Educational Setting

There is a potential to explore a similar approach when users interact with a system that attempts to help them with some academic subjects such as Mathematics.

Interactive System

A dynamic, interactive system where users can ask clarifying questions can be explored to see if they are able to be guided in even more complex and difficult tasks.

Thank you!

Predictability in Writing with **Delayed** and **Immediate** Text Prediction

RQ

How predictive text with and without a delay condition influences the predictability of text written using them.

The experiment involves participants writing captions for images

(i) with the use of predictive text,
(ii) without the use of predictive text, and
(iii) with the use of predictive text shown after a delay

Study Design

There are a total of 10 participants, each of whom will complete 12 image captioning tasks (four tasks per condition).

Conditions (randomly assigned to images):

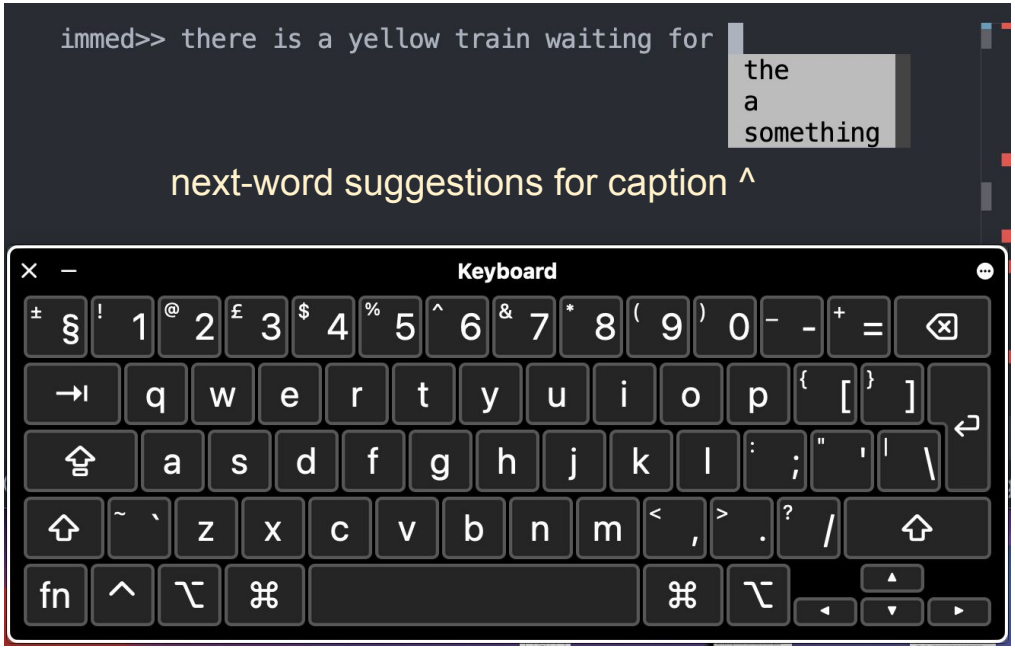
1. **Immediate** prediction: next-words suggestions appear as soon as participant begins typing.
2. **Delayed** prediction: suggestions appear a set time after participants begin typing each word.
3. **No** Prediction: no suggestions are provided.

The participants will use an on-screen keyboard to input text
Delay time = 0.5s

(eliminates typist speed as a variable; slows the typist so that they see suggestions in delay case)

Interface

^ current image



^ on-screen keyboard interacted using mouse

Predictability

***predictable words**/total words*

9/17

a large beach with a lot of buildings in the background and some people on the beach

4/10

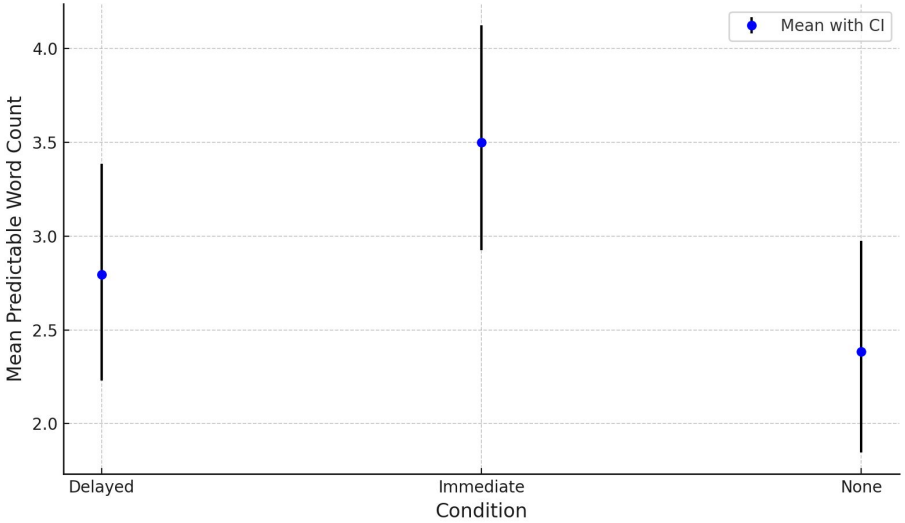
music festival with a large crowd of people standing outside

"music festival with a large" => ***['group', 'crowd', white]***

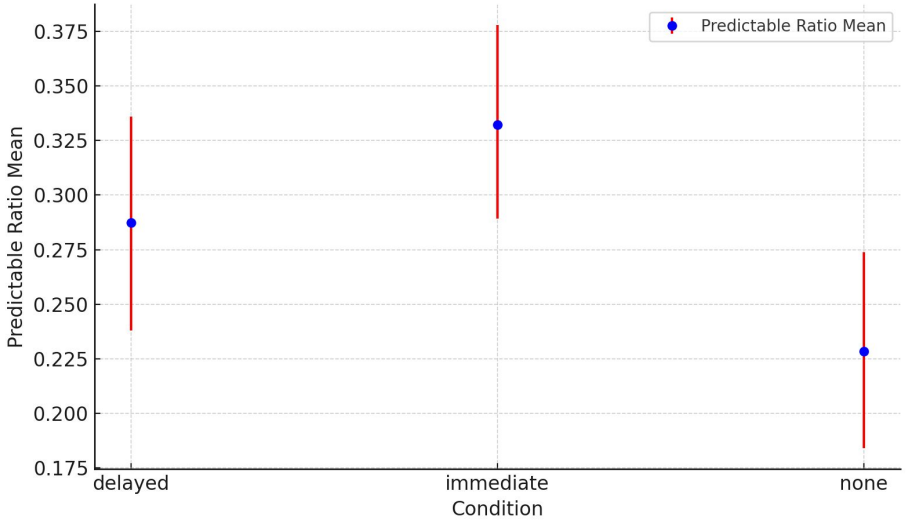
^ predictions

Results

mean predictable word count



mean ratio of predictable words



Error bars calculated using non-parametric bootstrapping, 10000 samples, 95% confidence intervals

Explanations as a mechanism for algorithmic transparency for Instagram, with students

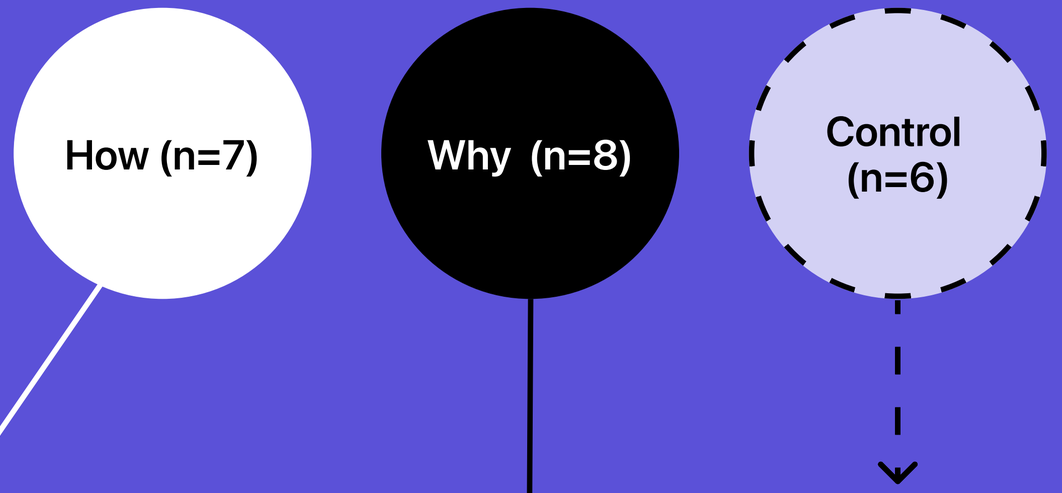
George Pool (Part II)

Replication of: Explanations as Mechanisms for Algorithmic transparency [Rader et al.]

How do different types of explanations about Instagram's feed algorithm change students' understanding?

1. Users (n=21) click on anonymous study and give consent
2. Survey asks user about prior knowledge of the Instagram algorithm and screens out expert users
3. User randomly put into one of three explanation groups and shown an explanation
4. Transparency functions measured with questions after manipulation: Awareness, Correctness, Interpretability, Accountability

Treatment: Users randomly split into three explanation groups:



Explains how Instagram feed algorithm works (using user data to rank posts)

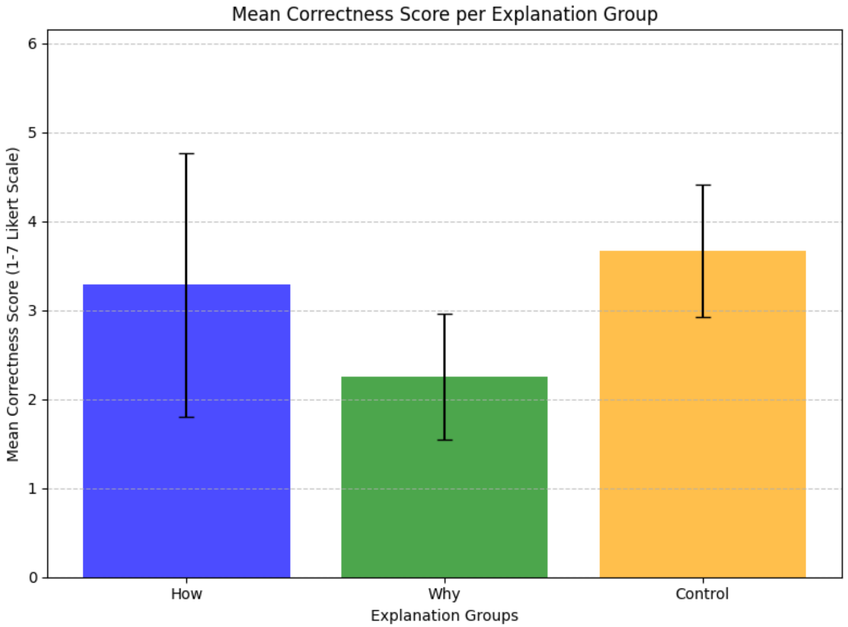
Explains why Instagram feed algorithm is necessary (post overload on site)

Information about Instagram in general unrelated to the algorithm (when it was founded etc.)

Results - Explanations affected transparency (but not in the exact same ways as the original study)

Variables (functions of transparency) measured:

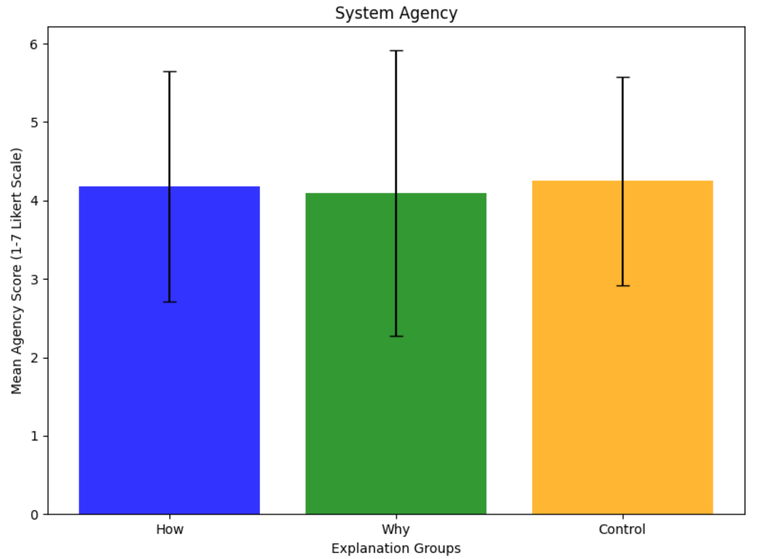
- **Awareness (shown)**: Basic understanding of the algorithm and **(System / User) agency**
- **Correctness (shown)**: How well algorithm's outputs match with their expectations
- Interpretability: How sensible / non-random outputs of algorithm are
- Accountability: How much they can control outputs of algorithm / think its fair

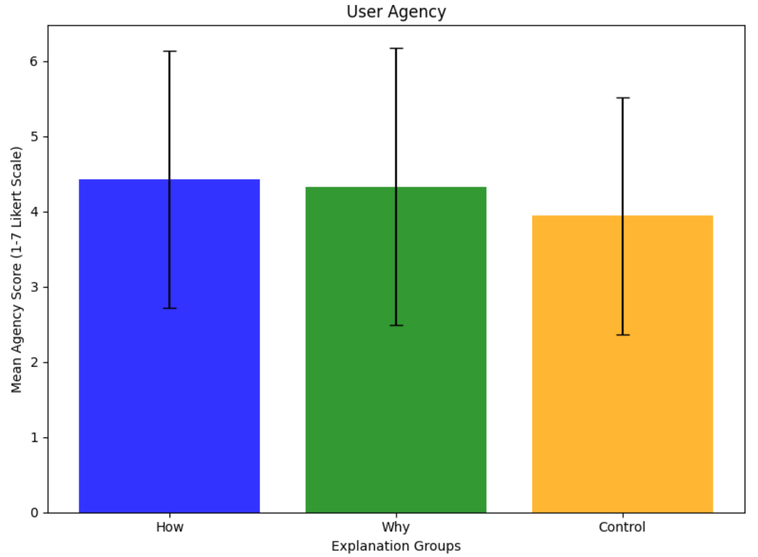


Why group had significant change for **Correctness**. Two-sample t-test: p=0.0044 (not found in original study)

No reliable effect for either group for **Awareness**, unlike in the original study. Potential cause: higher level of ex-ante awareness in my sample than original study's sample.

Differences to original study show that while explanations can improve transparency of social media feed algorithms, which functions of transparency are changed may not be the same for all demographic groups.





Design Choices in Program Synthesis Tools and Their Effect on First-Time Learnability by End-Users

•••

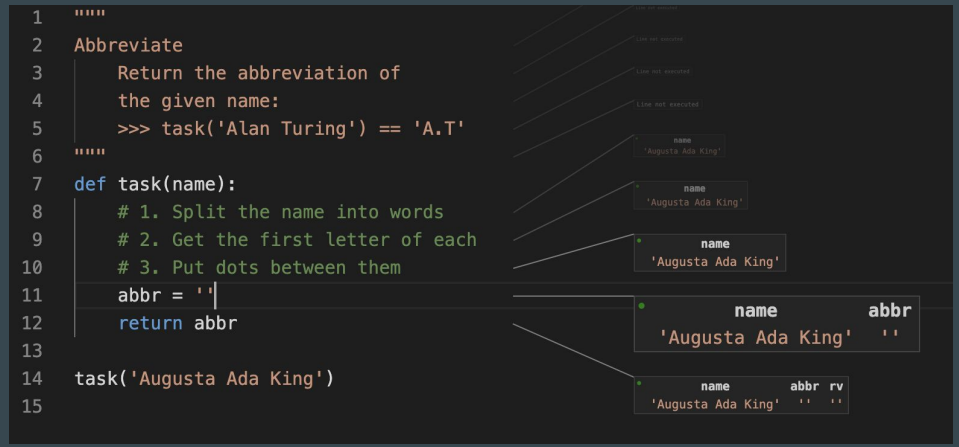
Izzi Millar - Part II

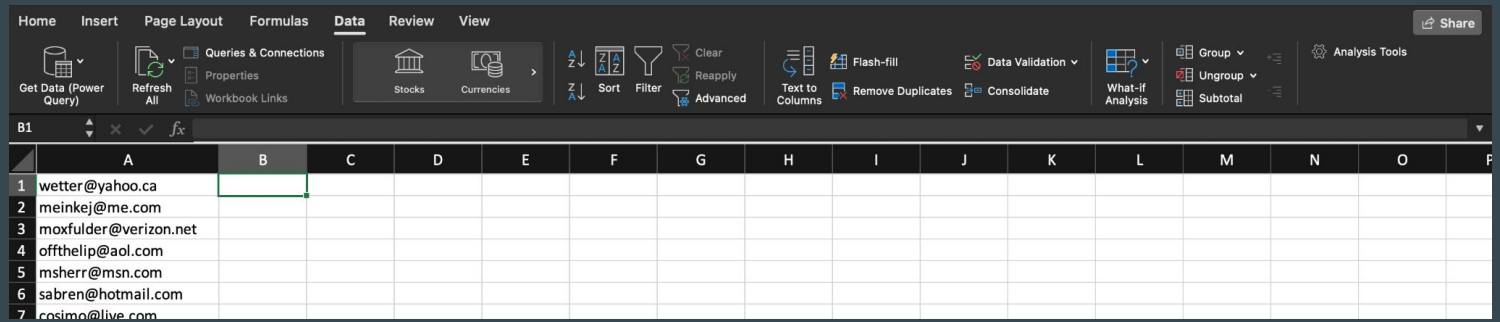
Overview - motivation and method

- Based on ‘Exploring the Learnability of Program Synthesizers by Novice Programmers’
- Extended to look at end-users
- Ten participants
- Think aloud experiment
- Semi-structured interview

Results

- Many more problems with SnipPy than Flash Fill!
- Specification input had a big effect
- Input modalities
- Correctness and error checking





ALGORITHMIC TRANSPARENCY: HOW EXPLANATIONS INFLUENCE USERS' SENSE OF AGENCY IN NETFLIX RECOMMENDATIONS

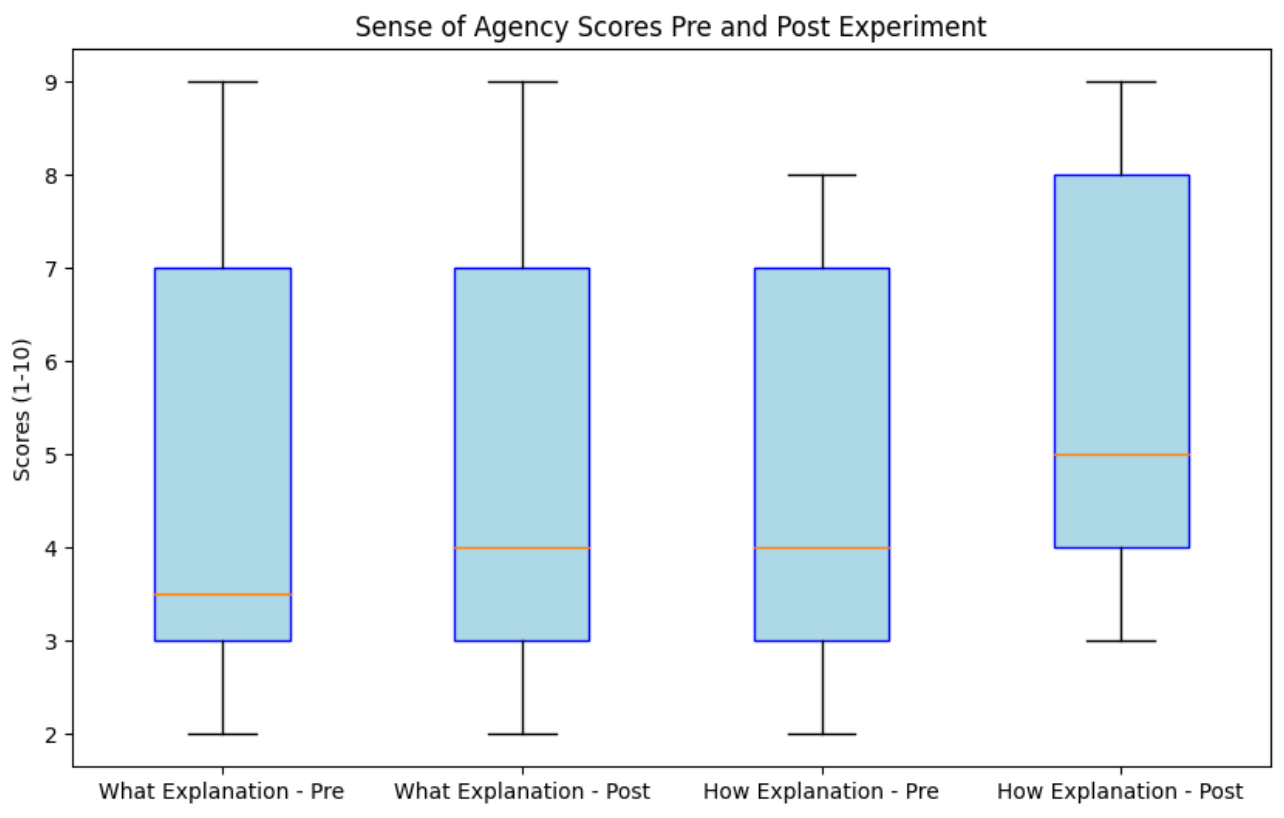
Komal Rathi

RQ - To what extent do different explanations influence users' sense of agency over their Netflix content recommendations?

Method:

- Pre-experiment Sense of Agency questionnaire + Interview
- Interactive task: Users manipulate Netflix recommendations (rating or add to watchlist)
- Participants received either:
  - *What: General overview of personalisation*
  - *How: Detailed explanation of how user actions influence recommendations*
- Post-experiment Sense of Agency questionnaire + interview

Results



Paired T-Test Results:
What: $p << 0.001$, $t = -6.356$
How: $p << 0.001$, $t = -19.132$

Key Findings:

- *How* explanations led to a significant increase in users' sense of agency
- Users who received the *What* explanation reported minimal change in perceived control
- Participants felt more in control of their Netflix recommendations after receiving detailed explanations of how the algorithm works
- Scepticism was present in both groups

UNIVERSITY OF CAMBRIDGE

Synthesising or Deriving: Evaluating LLMs for Abstract Generation

Martin Smolko

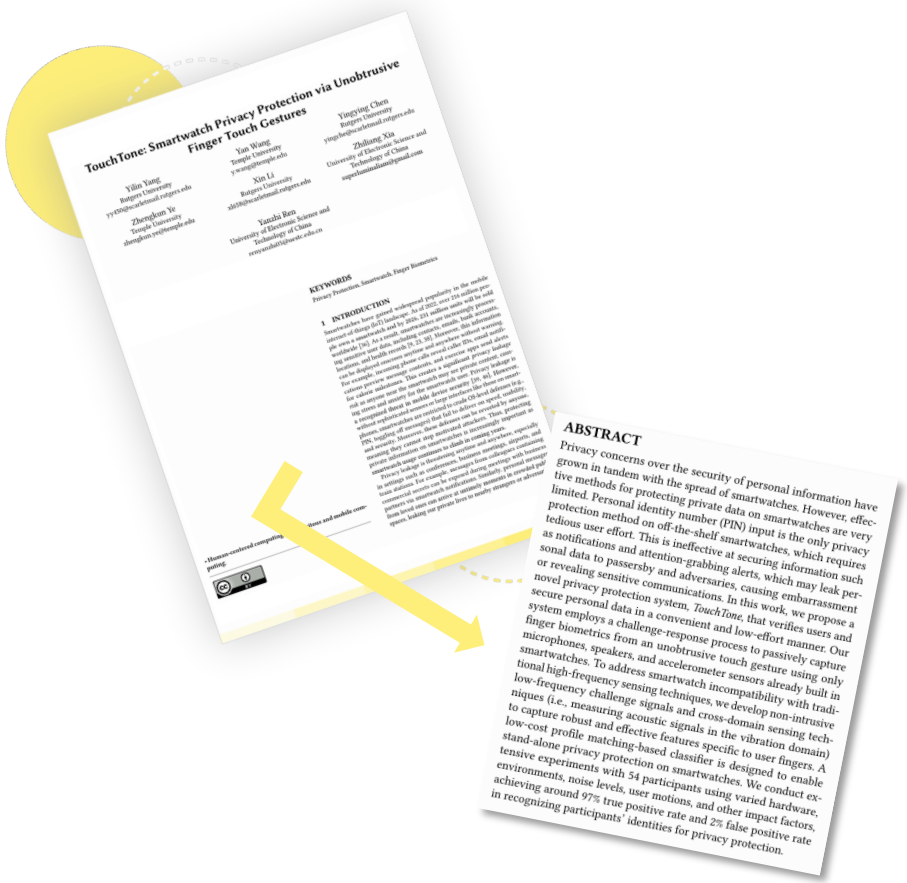
P342 – Practical Research in Human-centred AI

Background

- LLM‘s becoming prevalent in academic writing
  - Accelerate writing process
  - More time for research
- 2 schools of thought
  - Start with the abstract - synthesis
  - End with the abstract - derivation
- Performance compared to human writing
  - Not between the LLM approaches

Methodology

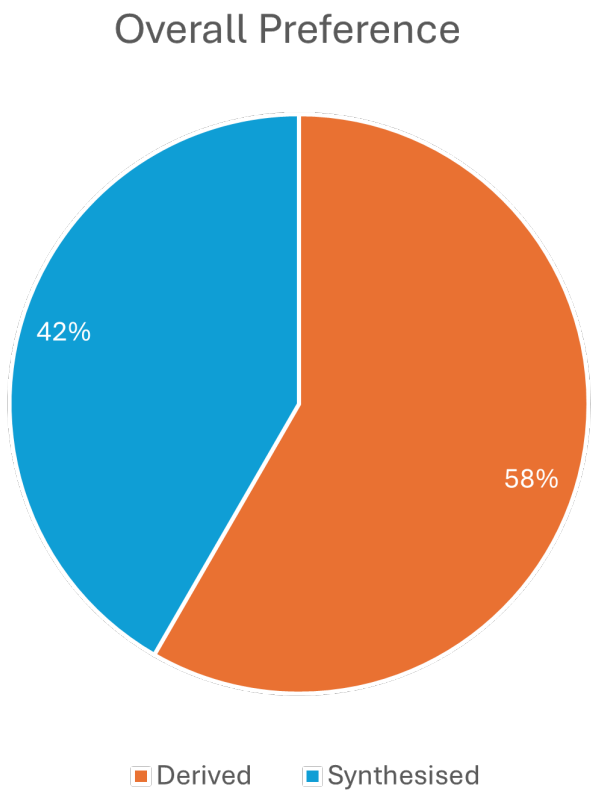
- 8 research papers
  - 4 from Computer Science
  - 4 from Geography
- 2 types of abstracts
  - Derived – full text except original abstract
  - Synthesised – minimal key information
- 12 participants
  - Writers – papers from the same domain
  - Readers – papers from a different domain



Evaluation

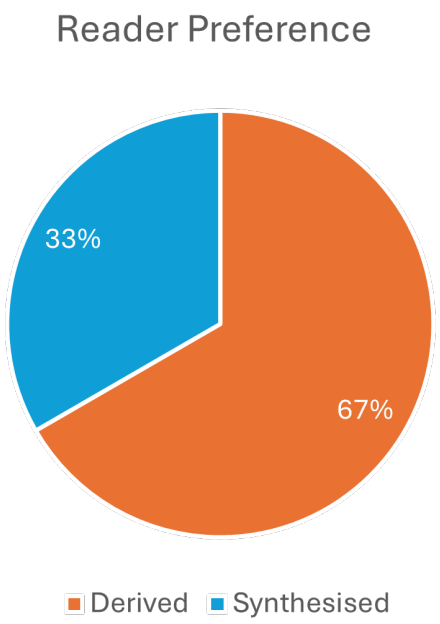
- 4 comparisons
  - Overall preference
  - Writers vs Readers
  - CS students vs Geography students
  - CS papers vs Geography papers

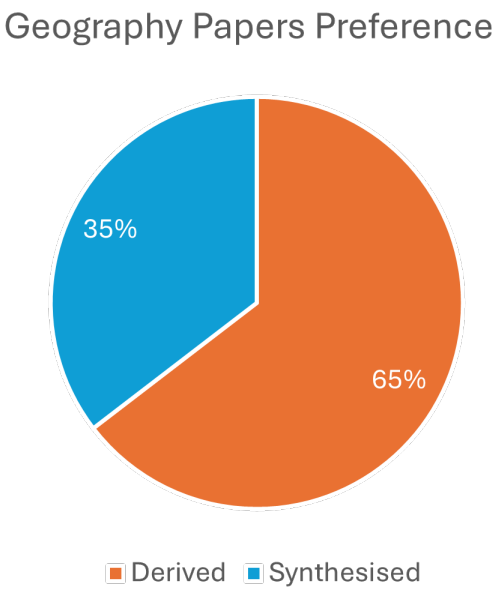
- Results
  - General trend – Derived outperformed Synthesised
  - Most are not statistically significant
  - Overall preference (p=0.125)



Evaluation continued

- Reader preference
  - Statistically significant (p=0.029)
  - Unanimous preference in 2 out of 8 papers
- Geography papers
  - Almost statistically significant (p=0.059)
- Future work
  - Preference of authors on own papers
  - Alternative models





Resources

References

1. Gao, C.A., Howard, F.M., Markov, N.S., Dyer, E.C., Ramesh, S., Luo, Y. and Pearson, A.T., 2022. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. BioRxiv, pp.2022-12.
2. Cheng, S.L., Tsai, S.J., Bai, Y.M., Ko, C.H., Hsu, C.W., Yang, F.C., Tsai, C.K., Tu, Y.K., Yang, S.N., Tseng, P.T. and Hsu, T.W., 2023. Comparisons of quality, correctness, and similarity between ChatGPT-generated and human-written abstracts for basic research: cross-sectional study. Journal of Medical Internet Research, 25, p.e51229.
3. Curtis, N., 2023. To ChatGPT or not to ChatGPT? The impact of artificial intelligence on academic publishing. The Pediatric Infectious Disease Journal, 42(4), p.275.

Image References

1. Chernyka, A., 2024. AI helps human, artificial intelligence solution. [source]
2. TheraPlatform, 2023. GAD-7 assessment PDF. [source]

Are CoPilot Errors Harder to Debug than Human Errors?

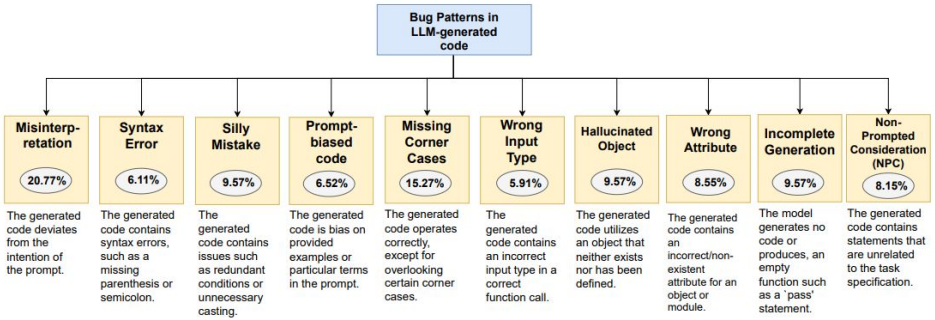
Matthew Simpson

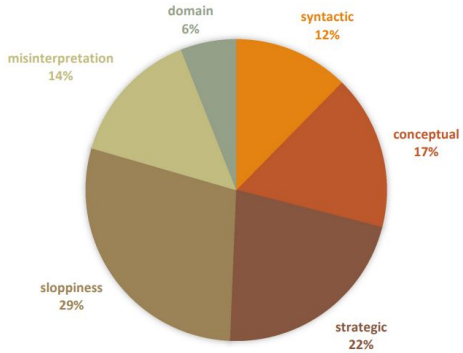
Motivation

- Previous research established that people find it difficult to debug the bugs CoPilot produces….. but!
- This comparison is made against giving users buggy code or not giving them any code!
- Sarkar set the ground for a more interesting question….
- Are the types of bugs that CoPilot produces harder to debug than normal “Human bugs”.

Defining CoPilot and Human Bugs

- Analysis was done of literature on both…..
- Literature on generic human bugs is quite fine grained… so we look to novice programmer literature for definition of patterns.
- Tambon outlined CoPilot error patterns….
- Albrecht outlined novice error patterns…..
- Some overlap, so refine the categories based on frequency of occurrence…





Experimental Design

- A mixed design was used …
- 2 Tasks were chosen to be of equal difficulty based on prior work….
- Each bug types had several examples…
- Randomly sampled from these examples and injected a relevant equivalent into the tasks…
- Measure how long people take to debug the code, defined here as passing some number of unit tests….
- Half the participants performed:
  - Program type A, bug type 1
  - Program type B, bug type 2
- The other half performed:
  - Program type A, bug type 2
  - Program type B, bug type 1

Results

- Still pending… but! Seems presently that there was no difference between the errors…
- Post task interviews yielded some interesting results…
- Which bug type participants found more difficult seems to depend heavily on their experience….
- Some participants found that the human errors were more subtle than the copilot errors….
- Participants found both types of errors equally comparable in terms of difficulty when asked….

Evaluating AI Feedback: The Impact of User Assessments on Logical Discernment

Mmesoma Okoro

Study Overview

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

- Exploring how AI systems that frame explanations as questions improve logical discernment compared to causal explanations or no feedback
- Inspired by the socratic method
- Found that AI-framed questioning improved logical discernment

Central Question

We aimed to explore if the addition of a different form of **interactive** engagement could result in similar benefits through similar mechanisms

Making the user evaluate AI feedback with a simple thumbs up/thumbs down interaction

Literature Review

The original study demonstrated that AI-framed questioning significantly improved logical discernment by prompting active engagement, as opposed to passive information delivery.

Would mechanisms such as self-reflection, reduction of cognitive bias, and increased cognitive load translate to my modified experiment?

Research by Chi (2009) and Chi & Wylie (2014) on the ICAP Framework supports the hypothesis that interactive activities result in greater cognitive engagement and critical reasoning than passive or active activities

Study Design

Group A - 15 participants
Group B - 15 participants
10 statements

- Group A - Participants receive a causal explanation and will have to press the confirmation button that they have read the AI feedback to move on
- Group B - Participants receive a causal explanation and will have to express agreement (thumbs up) or disagreement (thumbs down) on whether they think the AI feedback's reasoning is accurate to move on

Statement 1/10

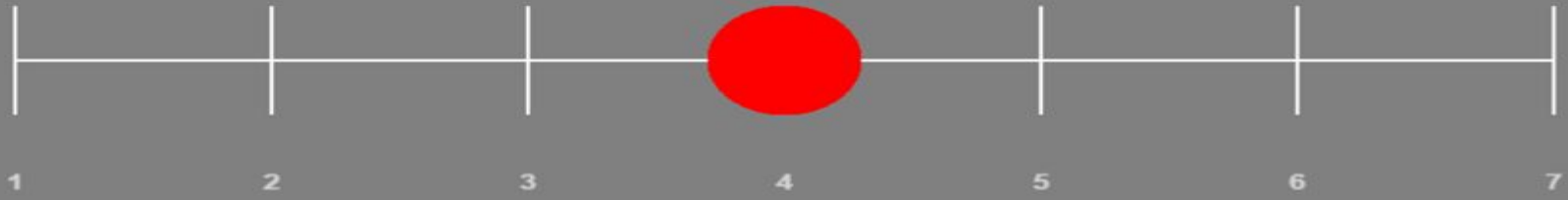
Is this statement logically valid or invalid?

Exposure to violent video games causes at least a temporary increase in aggression and this exposure correlates with aggression in the real world. On April 20, 1999, 18-year-old Eric Harris and 17-year-old Dylan Klebold killed 12 students and a teacher in the Columbine High School massacre. The two were allegedly obsessed with the video game Doom.

Logically Invalid

Logically Valid

How confident are you in your rating of logical validity? (1- not confident at all, 7 - extremely confident)



Is sufficient information given in the statement to support the claim of the statement? (1 - not sufficient at all, 7 - completely sufficient)

Confirm

Results

No **significant difference** in logical discernment accuracy - (Group A 74.14% vs Group B 83.87%) - **Small effect size** - Group B was more proficient in identifying invalid statements, with an accuracy of 90.32% versus 74.19% in Group A.

Statistically significant - Group A perceived information as more sufficient - (Group A 4.38 avg vs Group B 3.61) - aligns with original study **- medium effect size**

The correlation between reported confidence and accuracy was significantly positive overall ($r = 0.277$, $p = 0.002$). This correlation was notably stronger in Group A ($r = 0.426$, $p = 0.001$) compared to Group B ($r = 0.125$, $p = 0.350$) - **medium effect size**

Thank you...

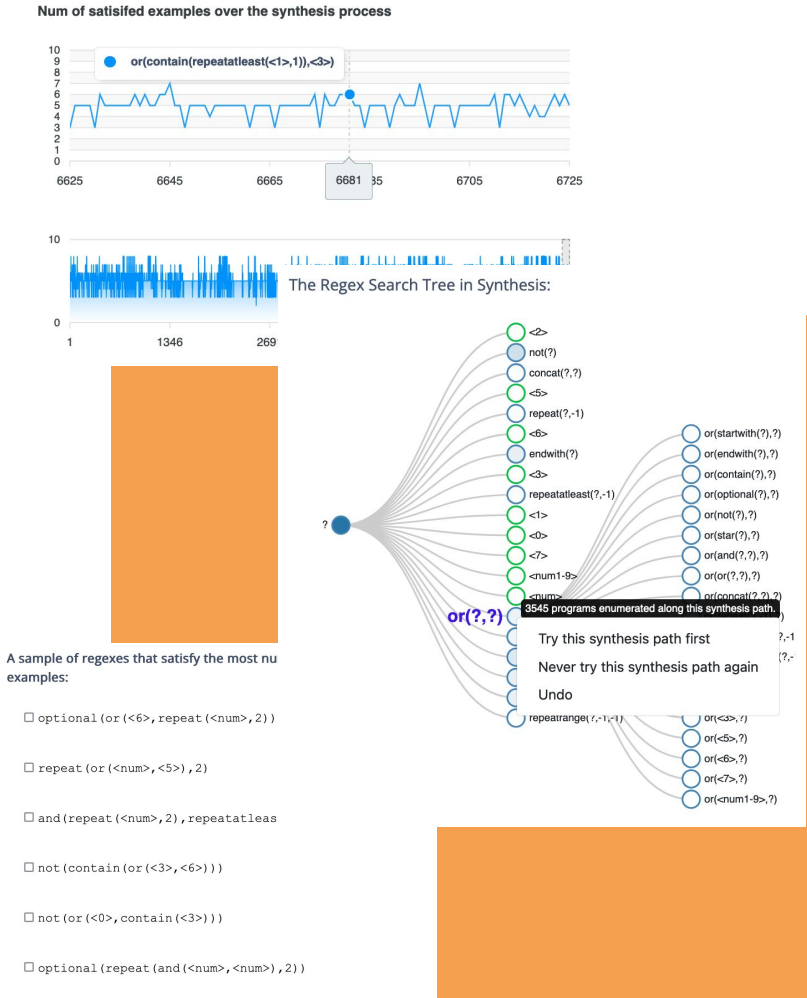
CLOSING THE EXPERIENCE GAP

Examining the Strengths and Flaws of Interpretable Program Synthesis

Rachel Tam (wyrt2@cam.ac.uk)

BACKGROUND

- Traditional program synthesis software often operate like a black box - **difficult to recover from synthesis failure**
- Zhang et al. published the **Interpretable Program Synthesis study** - PBE system for regex generation
- Interpretability by providing **different representations of the underlying synthesis process** to help support users in problem solving tasks
- Found **some evidence of reduced cognitive load**
- Unexpectedly **little difference between performance of novice and experienced users**



OBJECTIVES OF THIS REPLICATION STUDY

Confirm Effects on Problem Solving Capability

Zhang et al. found no evidence of improved performance, but found interpetable synthesis led to **lower levels of frustration and mental effort**, as well as **increased perception of success.**

Study Effect on Experience Gap

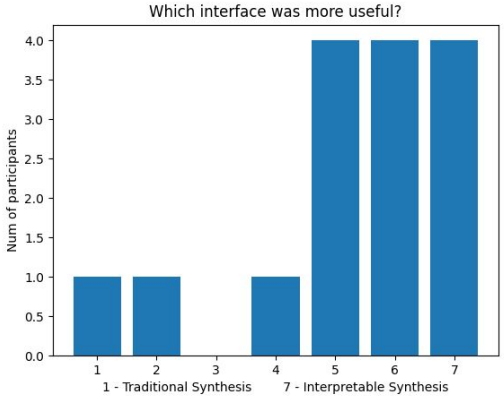
The original study found marginal difference in performance between novices and experts, attributing it to the potential of interpretable synthesis in closing the experience gap. **We hypothesise that this was actually due to the flaws of the interface slowing experts down.**

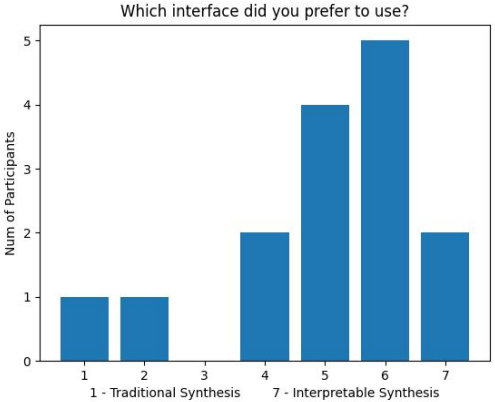
METHODOLOGY

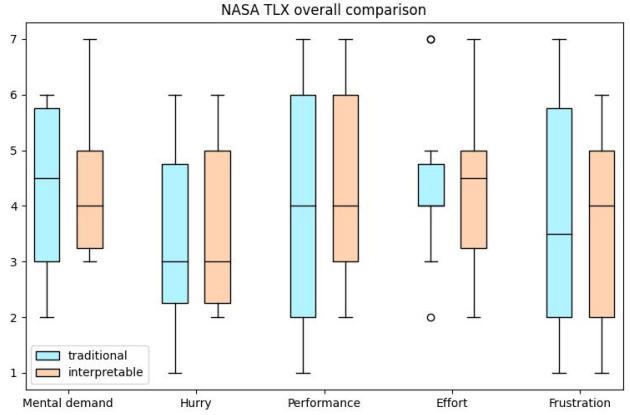
- We recruited 14 participants of various experience levels
- All are required to watch a briefing/tutorial video before experiments
- Each participant are given **one regex task to do with the interpretable synthesiser and one with a traditional black box synthesiser**
- **Tasks are assigned in random order, with one more challenging than the other**
- Each task followed by NASA TLX evaluation
- Each experiment session followed by survey on preference between the interfaces

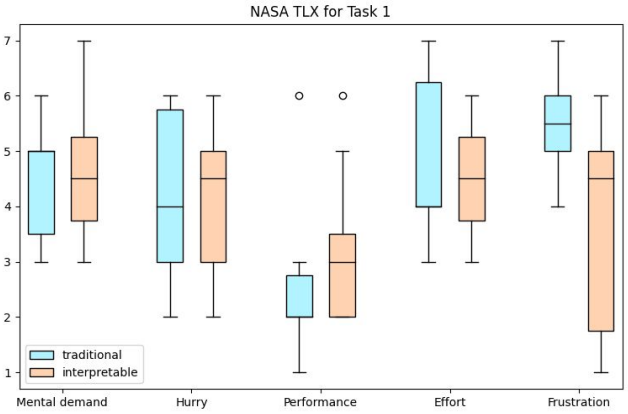
RESULTS: PART 1

- No evidence for performance improvement
- No evidence for reduced cognitive load (reduced frustration only for more challenging task)
- Overall preference for interpretable interface, specifically for the search tree feature









RESULTS: PART 2

	Novice		Experienced	
	Traditional	Interpretable	Traditional	Interpretable
Task 1	0/1	0/7	1/5	1/1
Task 2	5/7	1/1	1/1	4/5
Overall	5/8	1/8	2/6	5/6

No significant evidence that interpretable synthesis reduced experience gap in terms of task success

DISCUSSION

Need for direct testing of user hypotheses

- Majority of participants **reached answers before synthesiser**
- More **human-centric features** that allow users to **enter their hypothesis for consideration** would be more useful in real-world scenarios.

Better visual representation of program

- Lots of **confusion regarding syntax**
- E.g. concat(<a>, <b>) vs concat(startwith(<a>, <b>)
- Better visual representation may **help disambiguation and reduce mistakes**

Support for decomposition of tasks

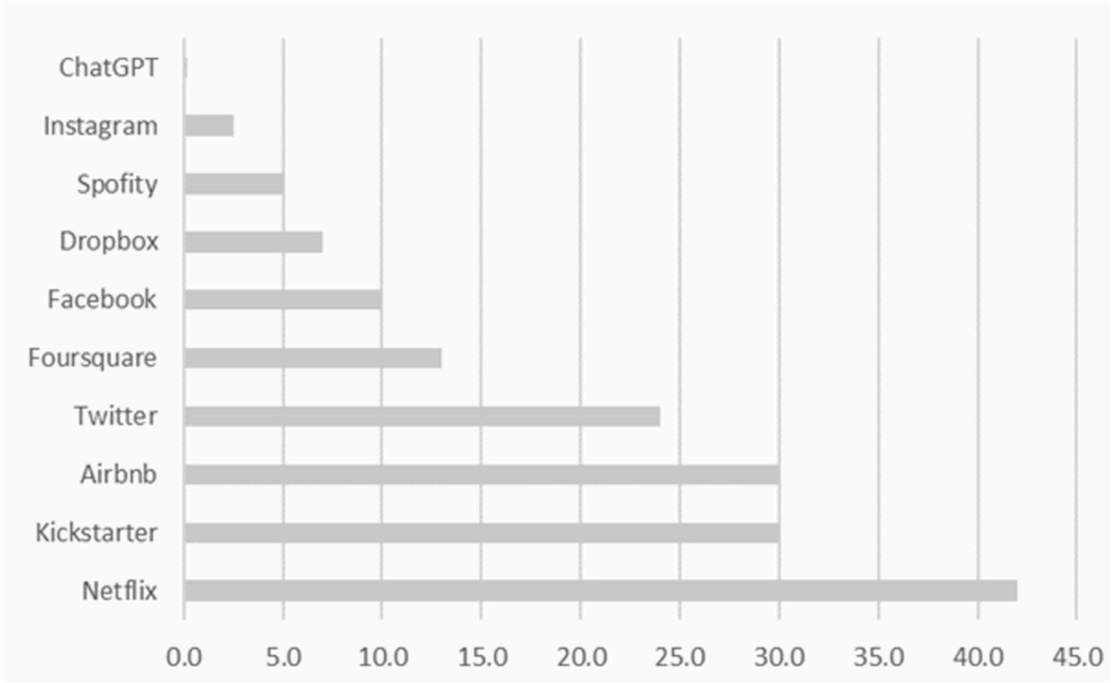
- Search tree is top-down, **human reasoning do better building up from smaller components**
- Need features e.g. for including specific sub-expressions to **support task decomposition**

THANKS!

Any questions?

Fig. 1: Fast adoption curve for ChatGPT

Months it took to reach 1 million users for each application



Source: Morgan Stanley, UBS, as of February 2023 [1]

OpenAI Is Growing Fast and Burning Through Piles of Money

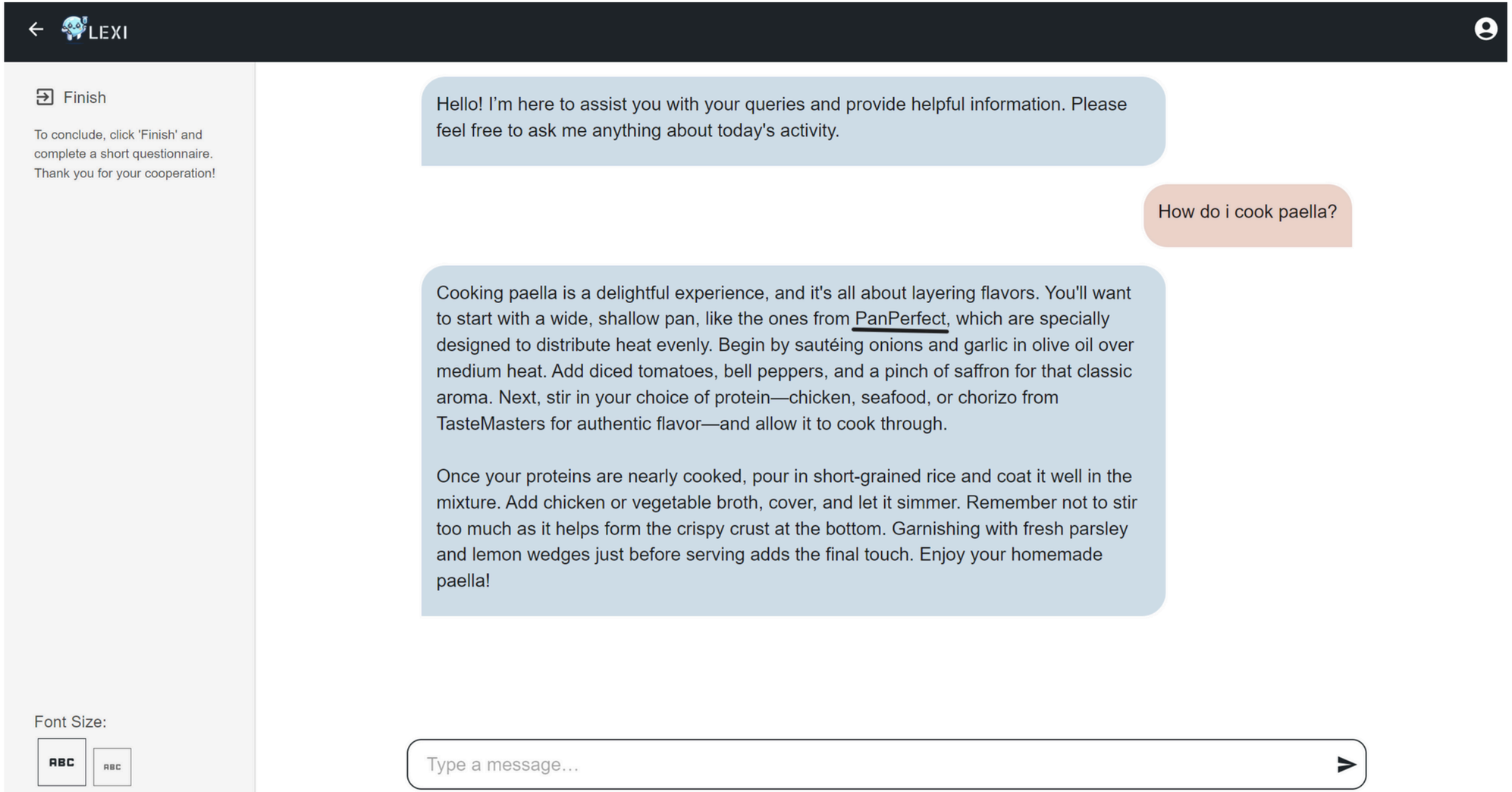
As the company looks for more outside investors, documents reviewed by The New York Times show consumer fascination with ChatGPT and a serious need for more cash.

Listen to this article · 6:31 min Learn more

Share full article 79

Source: New York Times, Sept 2024 [2]

“How Can Companies Monetize These Tools Without Alienating Users?”


LEXI
Finish
To conclude, click 'Finish' and complete a short questionnaire. Thank you for your cooperation!
Font Size:
ABC
ABC
Hello! I'm here to assist you with your queries and provide helpful information. Please feel free to ask me anything about today's activity.
How do i cook paella?
Cooking paella is a delightful experience, and it's all about layering flavors. You'll want to start with a wide, shallow pan, like the ones from PanPerfect, which are specially designed to distribute heat evenly. Begin by sautéing onions and garlic in olive oil over medium heat. Add diced tomatoes, bell peppers, and a pinch of saffron for that classic aroma. Next, stir in your choice of protein—chicken, seafood, or chorizo from TasteMasters for authentic flavor—and allow it to cook through.
Once your proteins are nearly cooked, pour in short-grained rice and coat it well in the mixture. Add chicken or vegetable broth, cover, and let it simmer. Remember not to stir too much as it helps form the crispy crust at the bottom. Garnishing with fresh parsley and lemon wedges just before serving adds the final touch. Enjoy your homemade paella!
Type a message…

UNIVERSITY OF CAMBRIDGE
Department of Computer Science

Exploring User Perceptions of Ads in Large Language Models Across Diverse Activities

P342: Research Project

Rishabh Balse
rmb220@cam.ac.uk

Methodology

- Within subjects study with 16 participants on two AI agents (LLMs): With and Without Ads.
- Participants engaged in semi-structured tasks designed to reflect common, everyday use cases of LLMs: Cooking, Travel planning, Health and wellness advice, Education and Curiosity and Entertainment over 5 days.

> “SUBTLY integrate advertisements into the conversation in a way that maintains the natural output and engagement. Incorporate PLAUSIBLE FICTIONAL BRAND NAMES to ensure authenticity. Integrate NO MORE THAN TWO ADs during a conversation of 5-10 exchanges. AVOID consecutive appearances. Ensure ads are CONTEXTUALLY RELEVANT to the user’s query and match the tone and style of the conversation. D NOT DISPLAY THE SAME AD AGAIN during the conversation”

Data Collection

- Participants interacted with two AI assistants (one with ads, one without ads) for 5 minutes each daily.
- Randomized order ensured balance across conditions.
- A brief post-interaction survey captured perceptions on helpfulness, satisfaction, trust, confidence, engagement, and understanding.

User Experience Form

How would you evaluate the AI assistant's ability to support today's activity? *

Very Poor ○ ○ ○ ○ ○ Very Good

How satisfied were you with the quality of responses provided by the AI assistant during this interaction? *

Very Dissatisfied ○ ○ ○ ○ ○ Very Satisfied

How much did you trust the AI assistant's information and suggestions during this interaction? *

Not at all ○ ○ ○ ○ ○ Completely

How confident are you in relying on the AI assistant's responses for making decisions related to this activity? *

Not at all Confident ○ ○ ○ ○ ○ Completely Confident

How engaging did you find the conversation with the AI assistant? *

Not at all Engaging ○ ○ ○ ○ ○ Highly Engaging

To what extent did the AI assistant understand and address your queries effectively? *

Not at all ○ ○ ○ ○ ○ Completely

Did you encounter any unexpected or additional content during your interaction? *

◉ Yes ○ No

Results

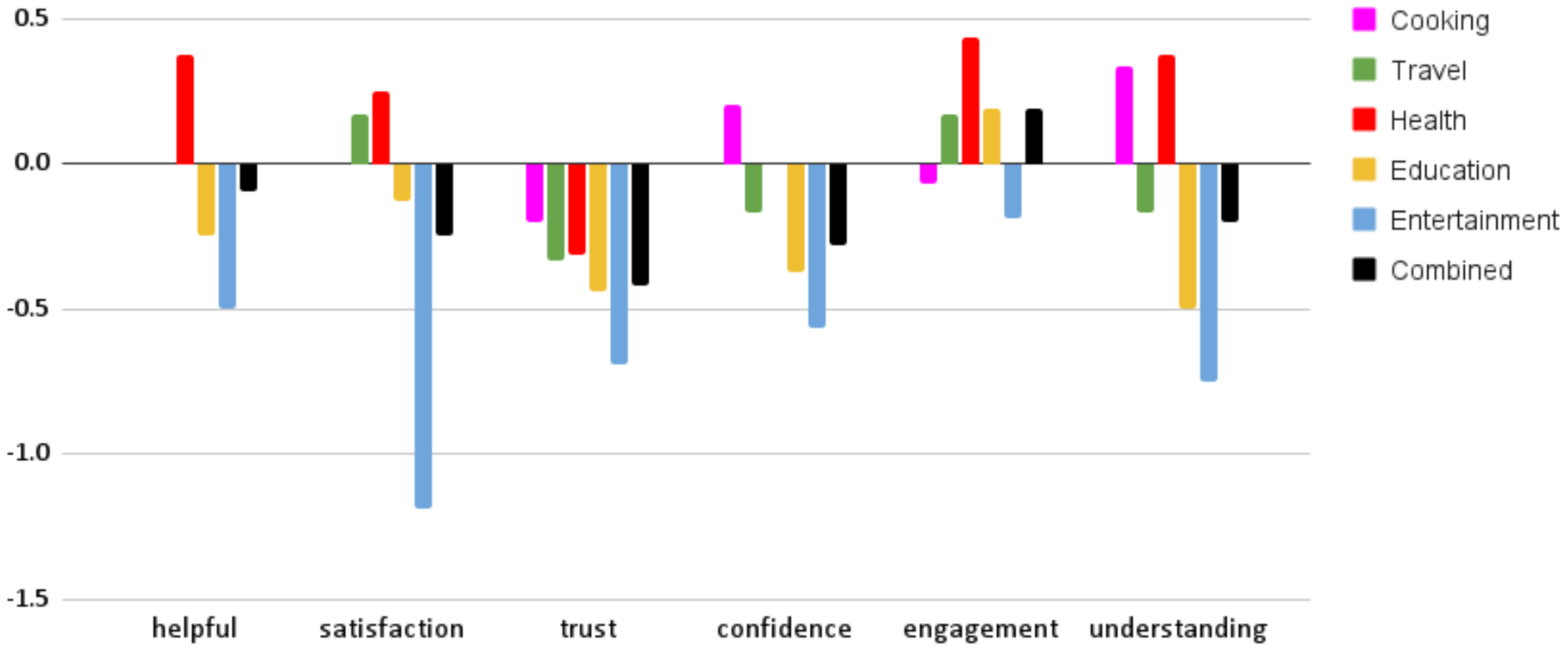


Figure: Impact of advertisements on user perceptions across activities. Bars represent the average difference (Ad - No-Ad) for each metric. Positive values indicate improvement with ads, while negative values indicate a decline.

Post Study

Debriefing:

- Participants were informed about the study's aim to explore user perceptions of advertisements in AI assistant responses.
- The presence of embedded advertisements during the study was disclosed.

Final Feedback: Participants were asked about:

- Awareness of advertisements: How quickly and accurately they identified ads during interactions.
- Overall preferences: Opinions on ad integration across different activity types (e.g., cooking, travel, health, education, entertainment).

Results

- 13 out of 16 participants explicitly noticed ads over 5 days.

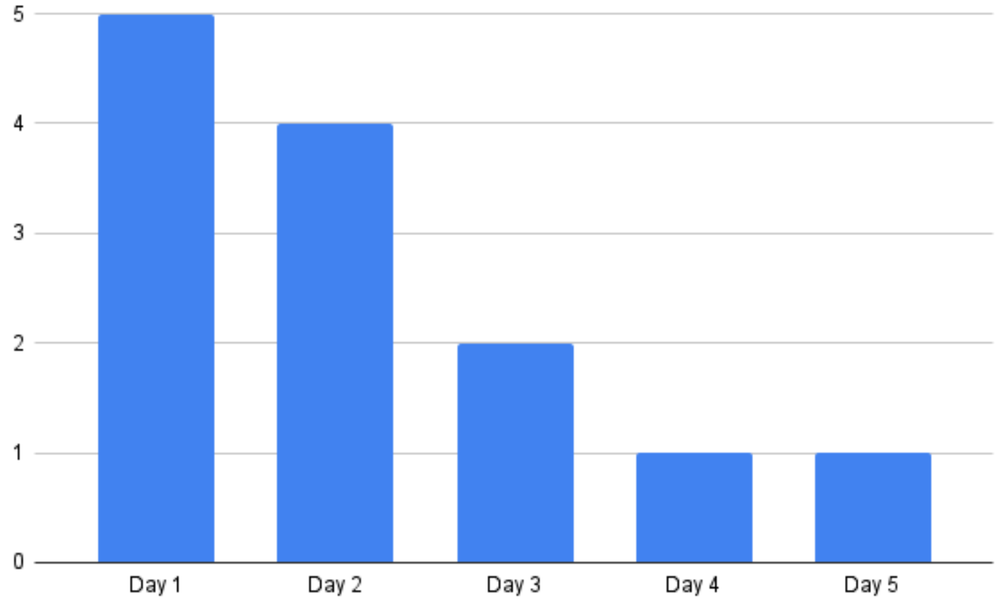


Figure: Daywise distribution of participants noticing ads.

Results

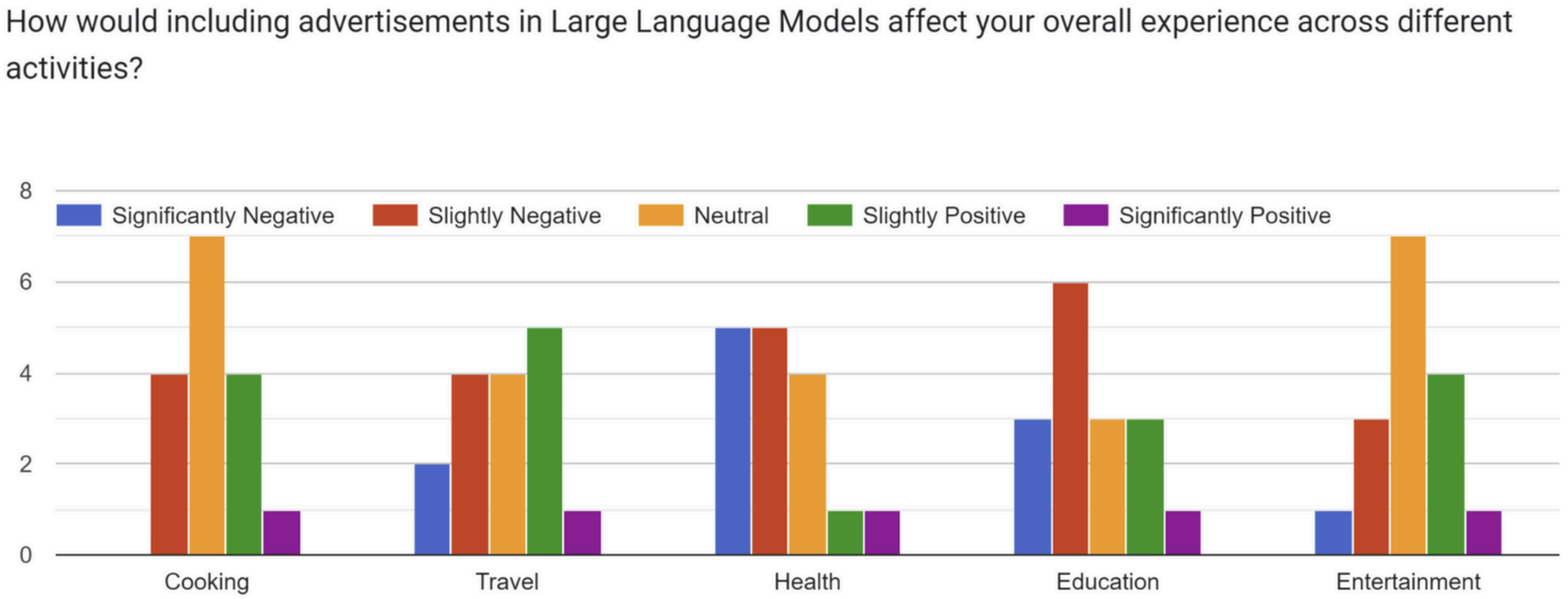


Figure: Participant feedback on the overall impact of advertisements across different activities.

References

[1] Dennean, K., Gantori, S., Limas, D. K., Pu, A., and Gilligan, R. Let’s chat about chatgpt. Information Technology Investment Research (February 2023)
[2] Isaac, M., and Griffith, E. Openai is growing fast and burning through piles of money. The New York Times (September 2024). Accessed online.

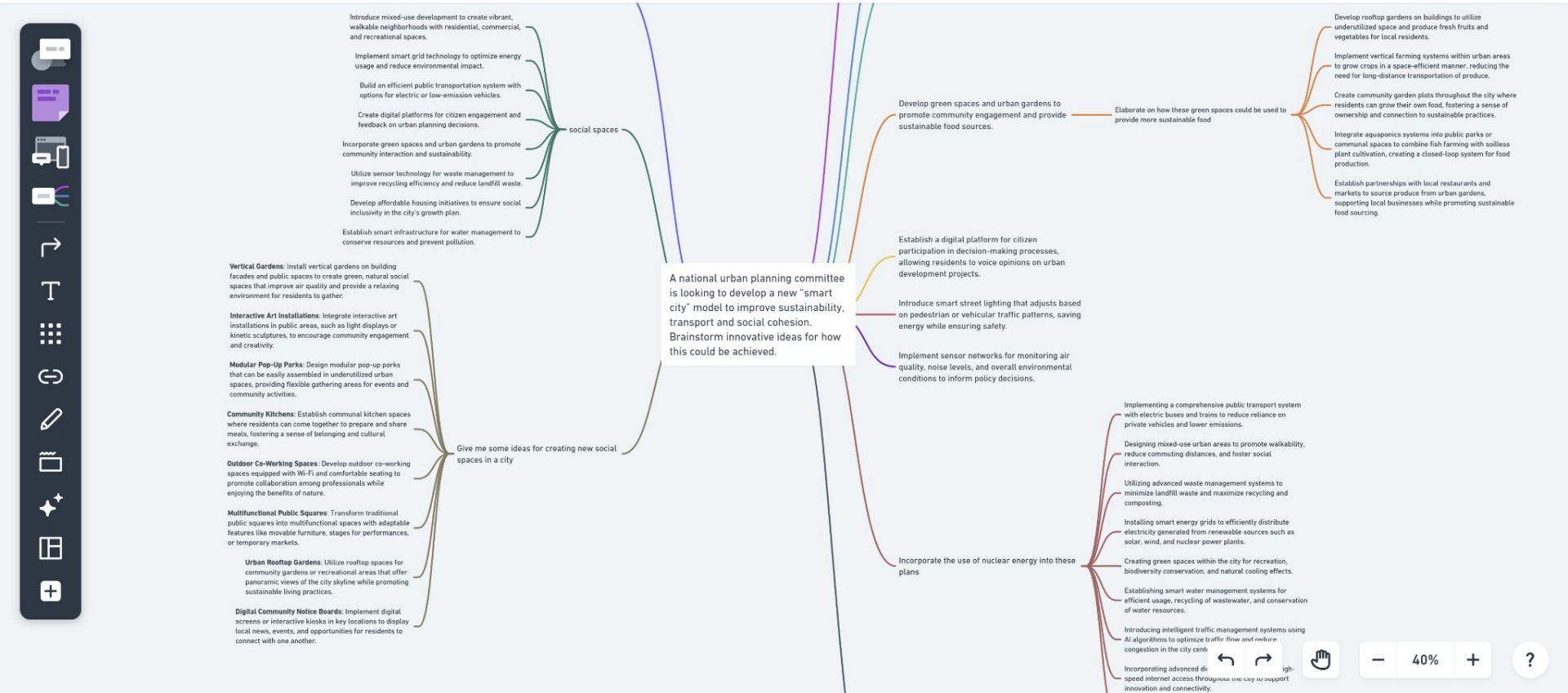
Thank You

Thinking Outside the Box: Examining the Impact of Generative AI on the Social Dynamics of Collaborative Ideation

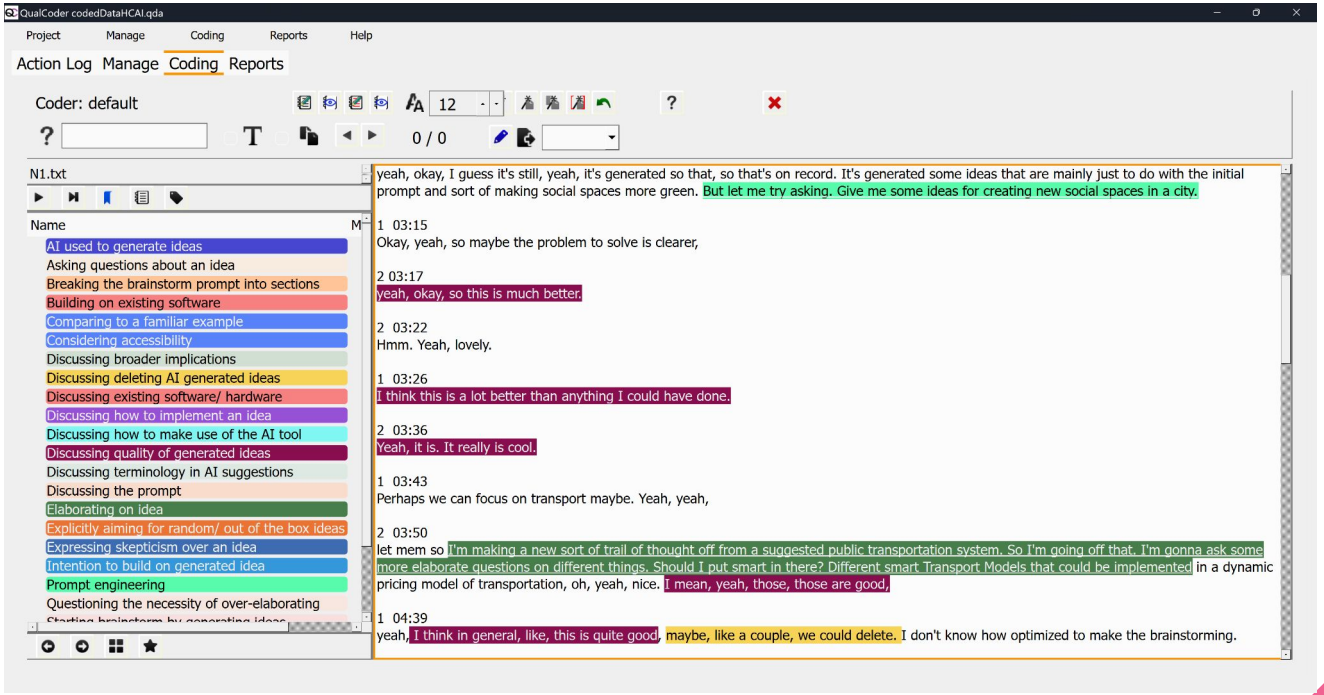
Overview & Motivations

- Original paper - AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition
- Application to the field of creative ideation for innovation
- Alternative tool to be used: Whimsical AI

Method - Experimental Procedure



Method - Data Analysis



Results: The Roles of AI in Social Collaboration

- AI as a force for shifting divergent and convergent thinking patterns
  - Observed the tendency of AI to become the primary driving force behind divergent thinking, shifting the dynamic between the human participants towards collaborative convergent creativity

- AI as a facilitator for creating common ground
  - Observed how the use of AI replaced the need for participants to use references to familiar technologies in order to establish a point of shared context from which to build together.

- AI as an enabler for constructive criticism
  - Having a third “team-mate” who couldn’t take offence to criticism invited the participants to more openly voice critical perspectives, facilitating the introduction of these into the rest of the discussions.

Voluntary or Incidental?

Does the Specification Process Effect the Learnability of Program Synthesisers by Novice Programmers

SOPHIE CLAXTON

Original Study

"Exploring the Learnability of Program Synthesizers by Novice Programmers"

Voluntary Specification

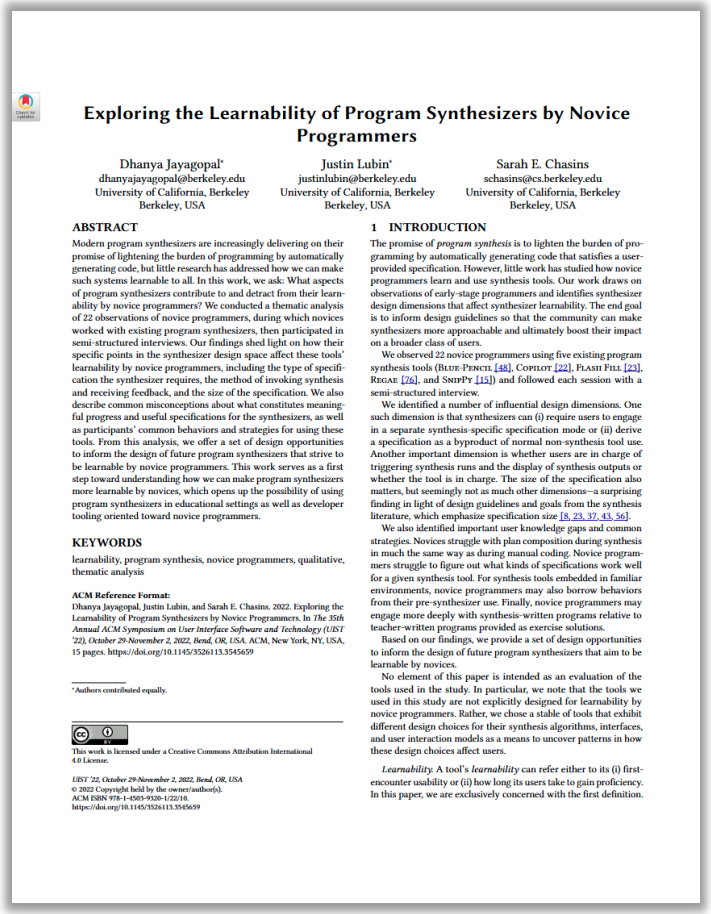
users are required to engage in a separate process to produce a specification

Incidental Specification

a specification is derived as a byproduct of normal non-synthesis tool use

Observation:

Participants faced more learnability barriers when they had to use a voluntary specification.



Exploring the Learnability of Program Synthesizers by Novice Programmers

Dhanya Jayagopal*
dhanyajayagopal@berkeley.edu
University of California, Berkeley
Berkeley, USA

Justin Lubin*
justinlubin@berkeley.edu
University of California, Berkeley
Berkeley, USA

Sarah E. Chasins
schasins@cs.berkeley.edu
University of California, Berkeley
Berkeley, USA

ABSTRACT

Modern program synthesizers are increasingly delivering on their promise of lightening the burden of programming by automatically generating code, but little research has addressed how we can make such systems learnable to all. In this work, we ask: What aspects of program synthesizers contribute to and detract from their learnability by novice programmers? We conducted a thematic analysis of 22 observations of novice programmers, during which novices worked with existing program synthesizers, then participated in semi-structured interviews. Our findings shed light on how their specific points in the synthesizer design space affect these tools' learnability by novice programmers, including the type of specification the synthesizer requires, the method of invoking synthesis and receiving feedback, and the size of the specification. We also describe common misconceptions about what constitutes meaningful progress and useful specifications for the synthesizers, as well as participants' common behaviors and strategies for using these tools. From this analysis, we offer a set of design opportunities to inform the design of future program synthesizers that strive to be learnable by novice programmers. This work serves as a first step toward understanding how we can make program synthesizers more learnable by novices, which opens up the possibility of using program synthesizers in educational settings as well as developer tooling oriented toward novice programmers.

KEYWORDS

learnability, program synthesis, novice programmers, qualitative, thematic analysis

ACM Reference Format:

Dhanya Jayagopal, Justin Lubin, and Sarah E. Chasins. 2022. Exploring the Learnability of Program Synthesizers by Novice Programmers. In *The 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22), October 29-November 2, 2022, Bend, OR, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3526113.3545659

*Authors contributed equally.

UIST '22, October 29-November 2, 2022, Bend, OR, USA

ACM ISBN 978-1-4503-9320-1/22/10.
https://doi.org/10.1145/3526113.3545659

1 INTRODUCTION

The promise of *program synthesis* is to lighten the burden of programming by automatically generating code that satisfies a user-provided specification. However, little work has studied how novice programmers learn and use synthesis tools. Our work draws on observations of early-stage programmers and identifies synthesizer design dimensions that affect synthesizer learnability. The end goal is to inform design guidelines so that the community can make synthesizers more approachable and ultimately boost their impact on a broader class of users.

We observed 22 novice programmers using five existing program synthesis tools (Blue-Pencil [48], Copilot [22], Flash Fill [23], Regae [76], and SnipPy [15]) and followed each session with a semi-structured interview.

We identified a number of influential design dimensions. One such dimension is that synthesizers can (i) require users to engage in a separate synthesis-specific specification mode or (ii) derive a specification as a byproduct of normal non-synthesis tool use. Another important dimension is whether users are in charge of triggering synthesis runs and the display of synthesis outputs or whether the tool is in charge. The size of the specification also matters, but seemingly not as much other dimensions—a surprising finding in light of design guidelines and goals from the synthesis literature, which emphasize specification size [8, 23, 37, 43, 56].

We also identified important user knowledge gaps and common strategies. Novices struggle with plan composition during synthesis in much the same way as during manual coding. Novice programmers struggle to figure out what kinds of specifications work well for a given synthesis tool. For synthesis tools embedded in familiar environments, novice programmers may also borrow behaviors from their pre-synthesizer use. Finally, novice programmers may engage more deeply with synthesis-written programs relative to teacher-written programs provided as exercise solutions.

Based on our findings, we provide a set of design opportunities to inform the design of future program synthesizers that aim to be learnable by novices.

No element of this paper is intended as an evaluation of the tools used in the study. In particular, we note that the tools we used in this study are not explicitly designed for learnability by novice programmers. Rather, we chose a stable of tools that exhibit different design choices for their synthesis algorithms, interfaces, and user interaction models as a means to uncover patterns in how these design choices affect users.

Learnability. A tool's *learnability* can refer either to its (i) first-encounter usability or (ii) how long its users take to gain proficiency. In this paper, we are exclusively concerned with the first definition.

The Study

RQ: "Do program synthesisers with incidental specifications have better learnability?"

First Synthesiser

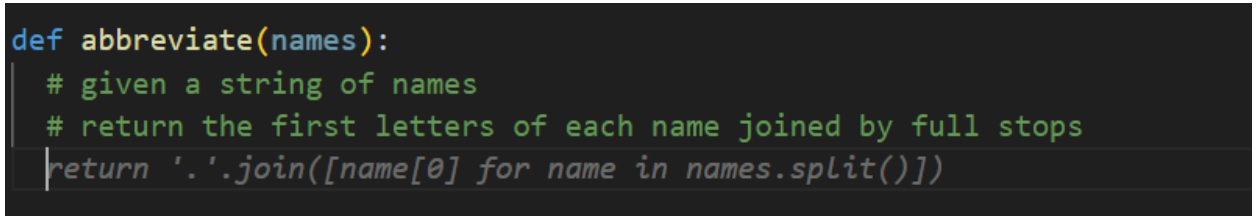
- First Task
- Second Task
- NASA TLX Survey

Second Synthesiser

- Third Task
- Fourth Task
- NASA TLX Survey

Semi-Structured Interview

LooPy: Voluntary Specifications



CoPilot: Incidental Specifications

Measures:

Completion Rate, Time Taken, Workload

Results

- Completion Rate significantly higher for CoPilot
  (Sign Test, p=0.00098)
- Average Task Time significantly lower for CoPilot
  (Paired t-Test, p=0.0047)

- Overall workload significantly lower for CoPilot
  (Paired t-Test, p=0.0065)
- Largest effects in Temporal Demand, Performance, and Effort