Performance, Workload, Emotion, and Self-Efficacy of Novice Programmers Using AI Code Generation

Nicholas Gardella, Raymond Pettit, and Sara L. Riggs

Introduction

AI is great at simple programming tasks.

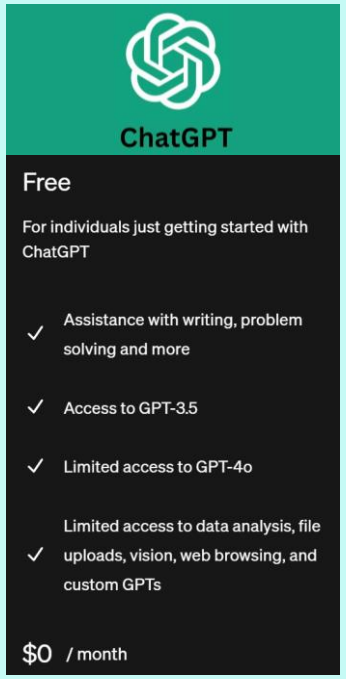
Default (GPT-3.5)

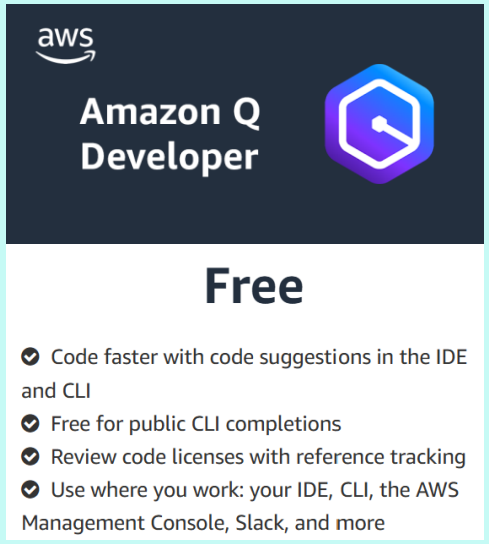
N

write a python script to read in a csv file that contains columns a,b,c and finds a grand sum of all the values of a,b, and c added together by row and then adding up all the rows

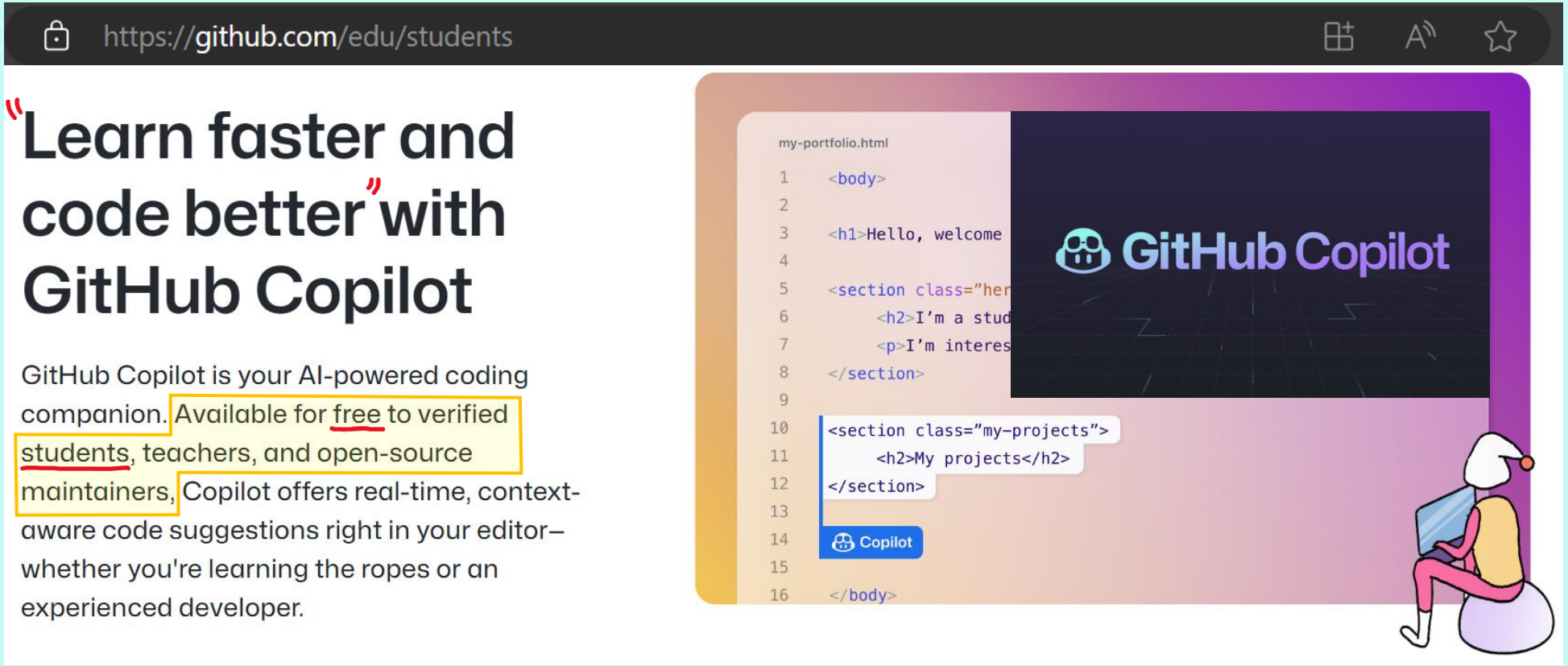

```python
import csv
# Initialize the variables to store the grand sum and row sums
grand_sum = 0
row_sums = []
# Specify the path to your CSV file
csv_file_path = 'your_csv_file.csv'
# Open the CSV file and read its contents
with open(csv_file_path, 'r') as csv_file:
    csv_reader = csv.reader(csv_file)
    # Skip the header row if it exists
    next(csv_reader, None)
    for row in csv_reader:
        try:
            # Assuming that columns a, b, and c are at index 0, 1, and 2 respectively
            a, b, c = map(float, row[0:3])
            # Calculate the sum of a, b, and c for this row
            row_sum = a + b + c
            # Add the row sum to the list of row sums
            row_sums.append(row_sum)
            # Add the row sum to the grand sum
            grand_sum += row_sum
        except ValueError:
            # Handle cases where the CSV contains non-numeric values
            print(f"Skipping invalid row: {row}")
# Calculate the grand total of all row sums
grand_total = sum(row_sums)
# Print the results
print(f"Grand Total of all rows: {grand_total}")
```

Novices can use high quality AI for free…



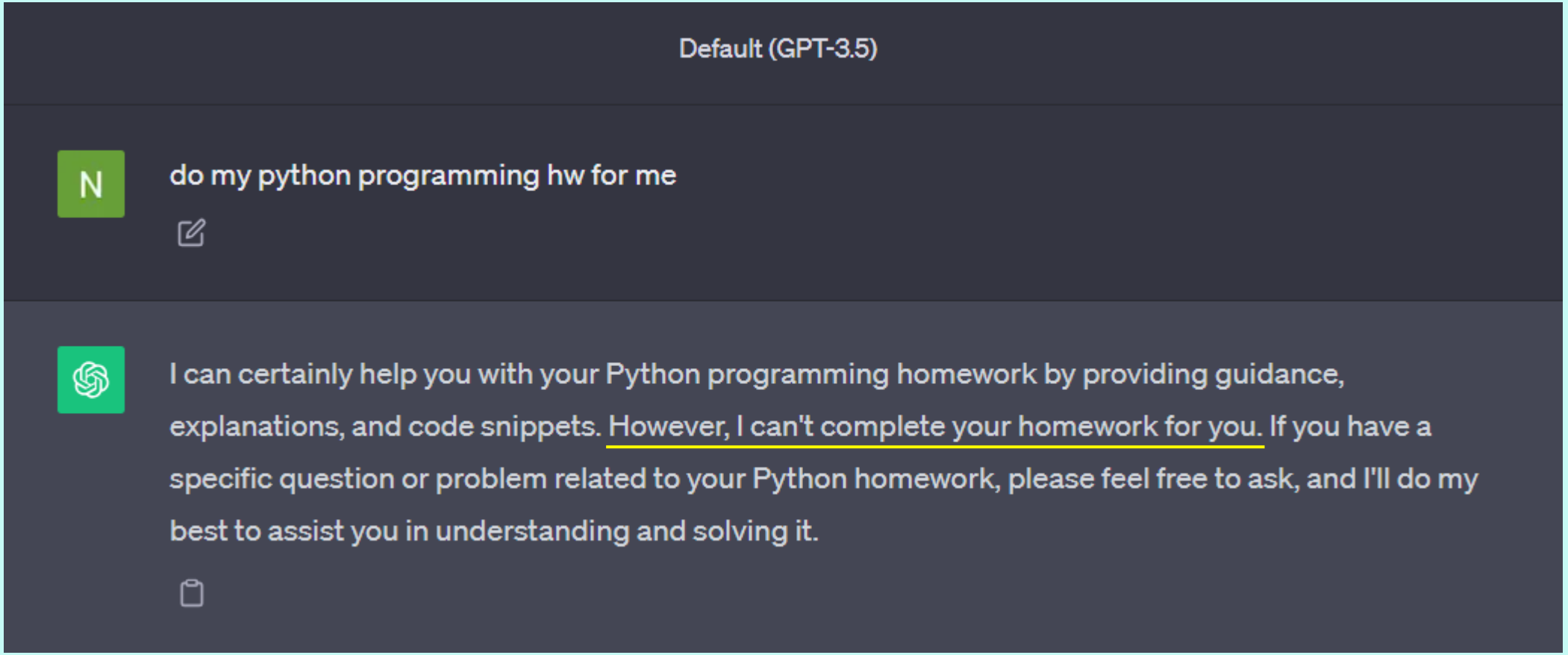


including the powerful GitHub Copilot AIDE



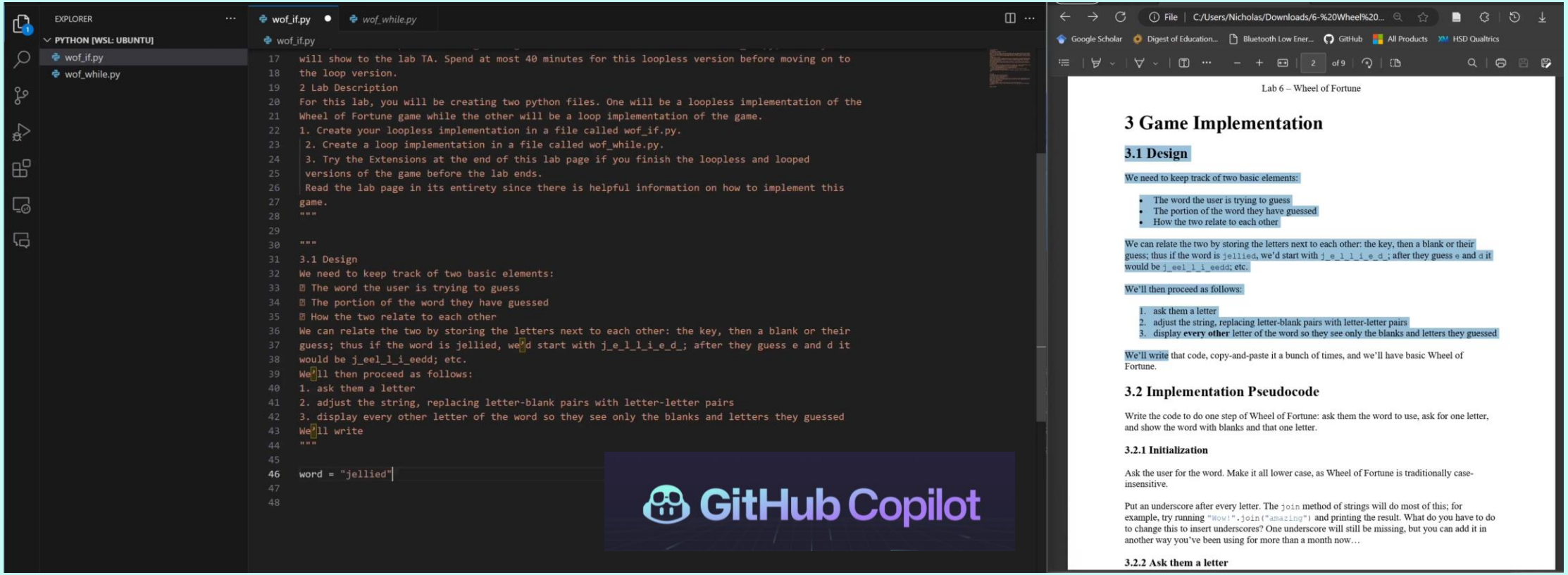
AIDE = Artificial Intelligence-driven Development Environment

But AI won’t do students’ work, right?



Unfortunately, it will…

see Chen et al. [8], Finnie-Ansley et al. [12], Kazemitabaar et al. [21]



So what is the impact of AI on a novice?

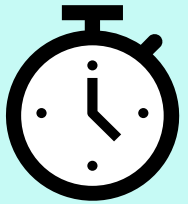
Student

\+

AI

\+

Time Pressure



= ?

Literature at a glance: Performance

Objectively Measuring Performance

Vaithilingam & Glassman [35], Xu et al. [37]

Significant Performance *Benefit*

Kazemitabaar et al. [21]

Literature at a glance: Well-Being

Kazemitabaar et al. [21]
“slightly **less stressed**” (p=.06)
“**more eager** […] to continue learning” (p = **.025**)

Vaithilingam & Glassman [35]
“Participants found code generated by Copilot **more helpful** than code generated by Intellisense” (p < **.001**)

Xu et al. [37]
“participants report having a **neutral** (15/31; 48.4%) **or** at least somewhat **positive** (15/31; 48.4%) experience”

But what measurable effects can we show?

Student

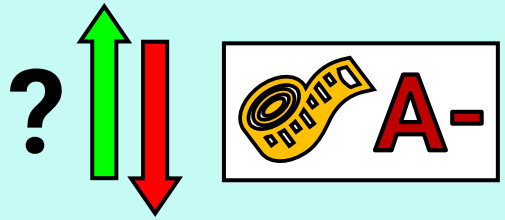
+

AI

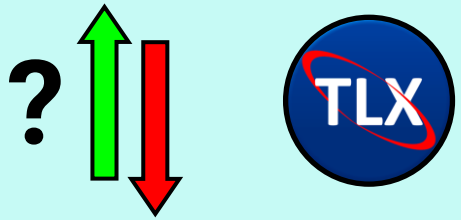
+

Time Pressure

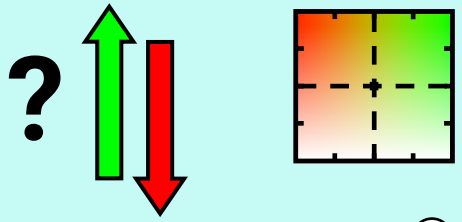
=



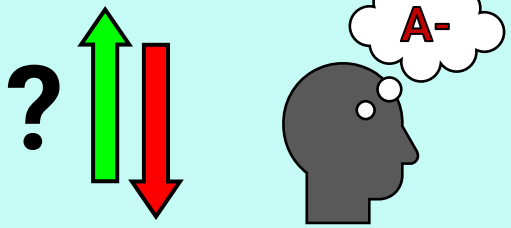
Performance



Workload



Emotion



Self-Efficacy

Research Questions: How do AIDEs…

- **RQ1:** affect performance, workload, emotion, & self-efficacy of novice programmers under time pressure?

- **RQ2:** influence the effects of additional time spent programming on novices' performance and self-efficacy

Method

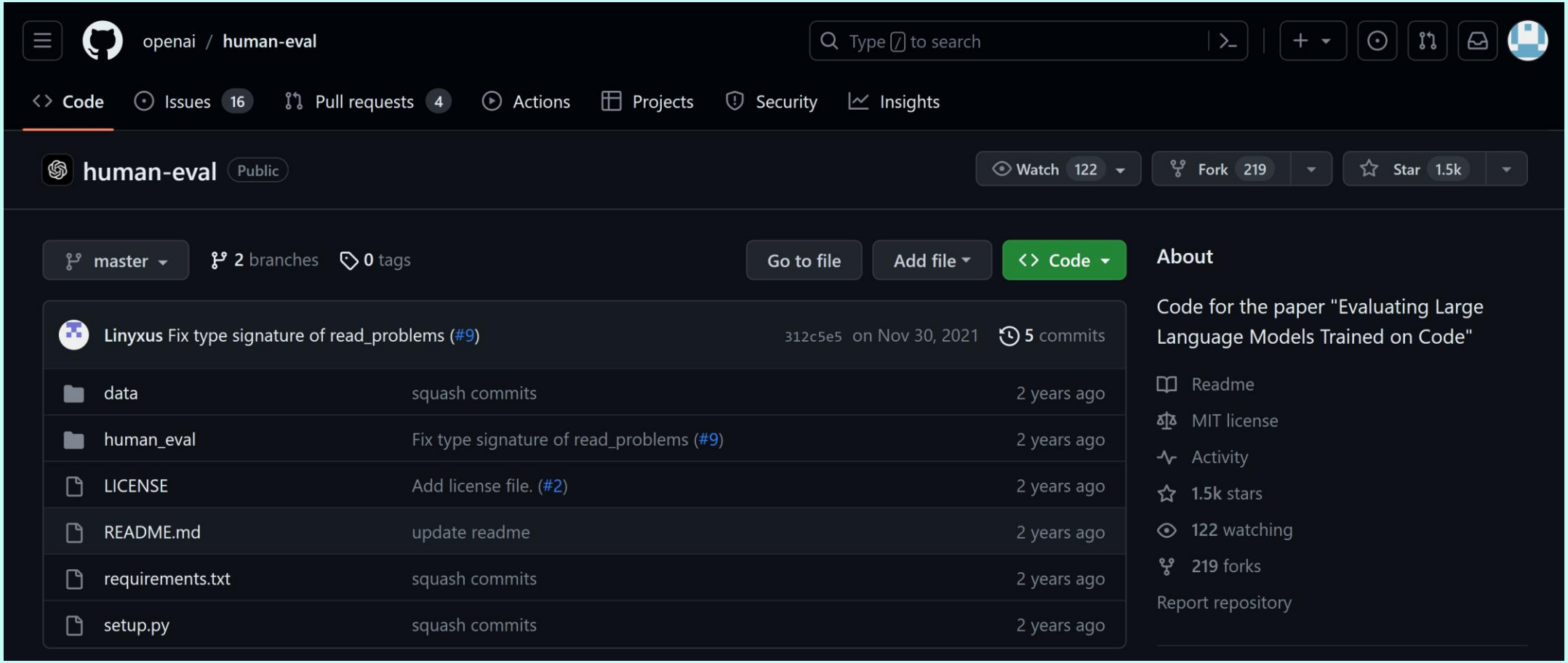
We recruited 17 students from a CS1 course.

- Ability reports (day of participation):

- Gender reports:

- Racial reports:
  - Asian- 11
  - White- 8
  - Hispanic/Latinx- 2
  - Black- 1

Tasks came from the HumanEval dataset.

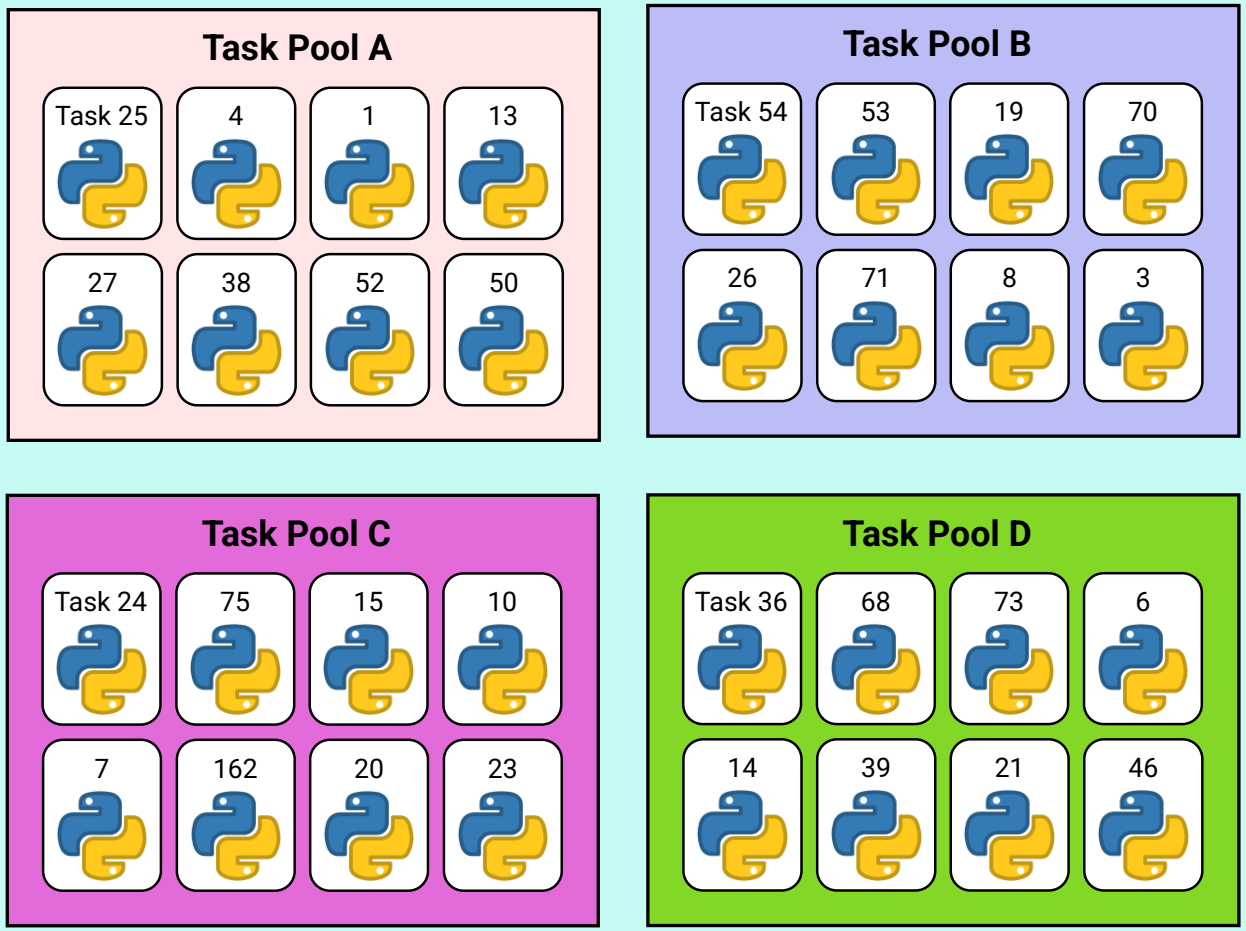


Chen et al. [8]

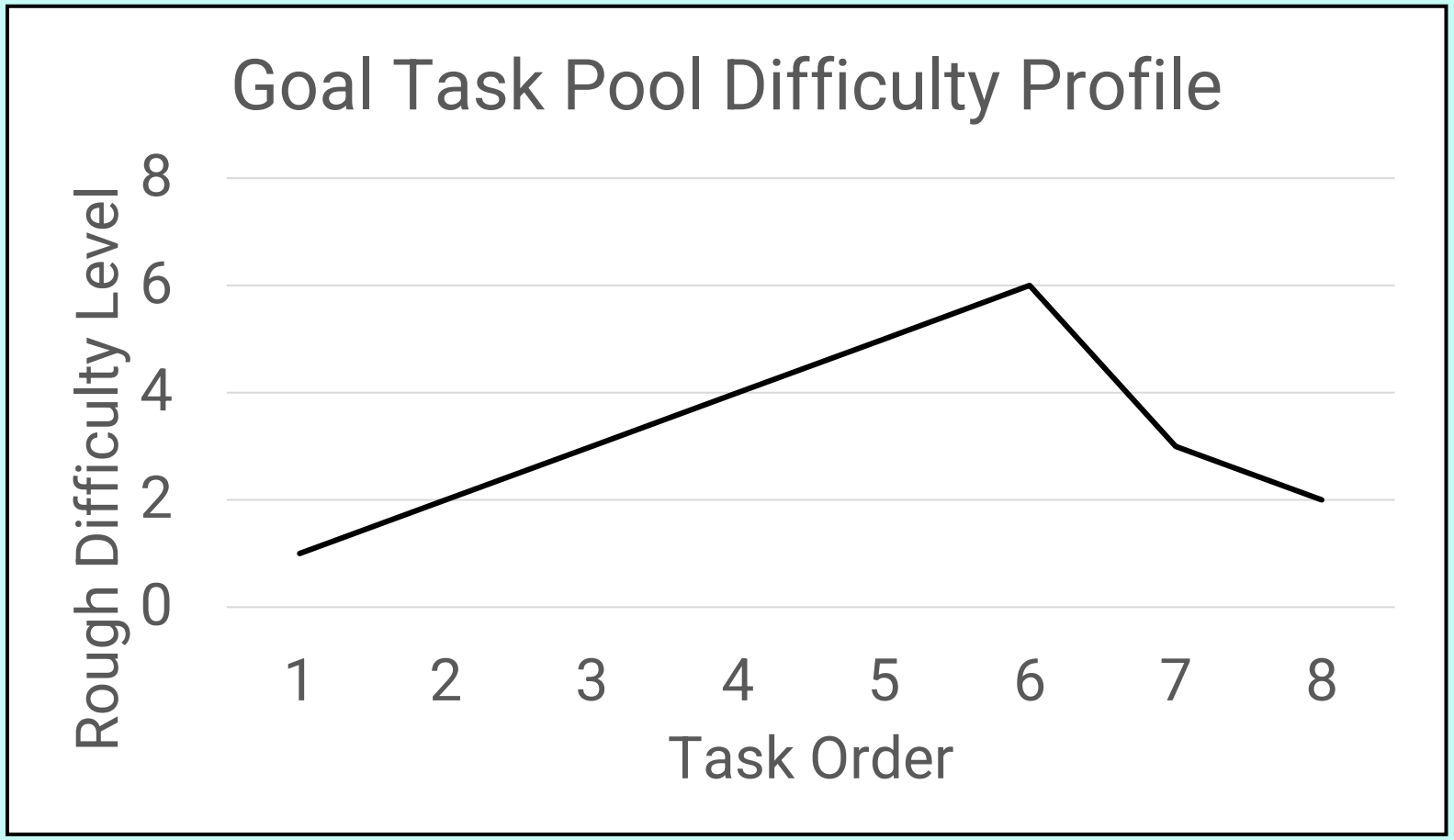
We created four pools of tasks.

164 Tasks

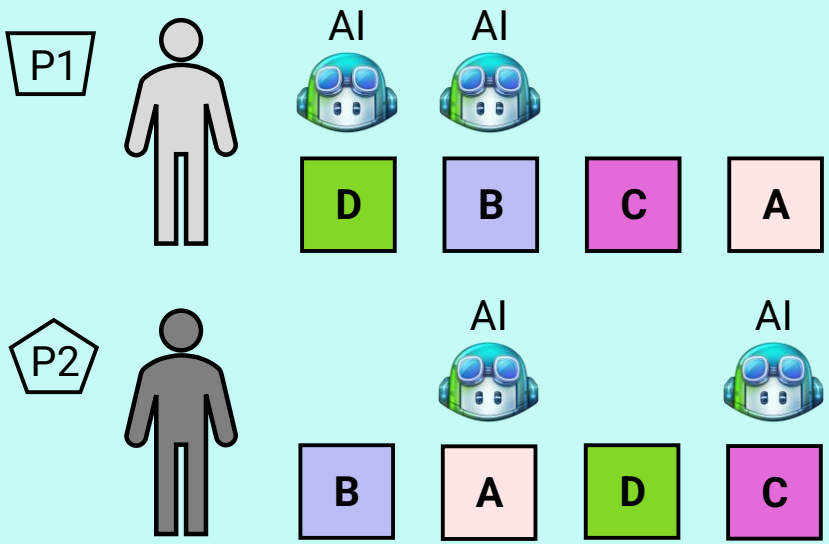
9 x

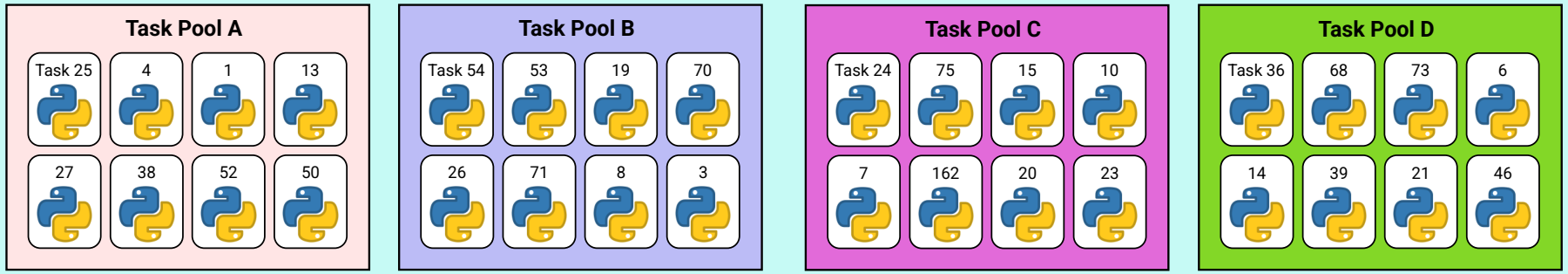


We placed easier tasks early in each pool.

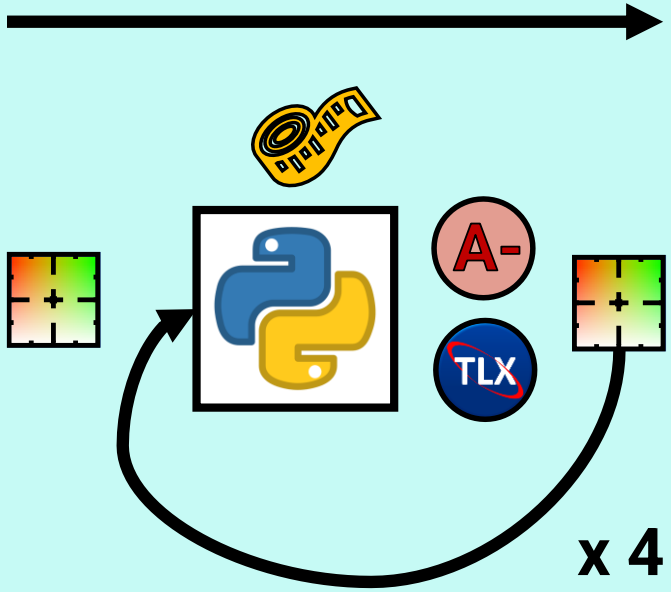


We used a within-subjects design.





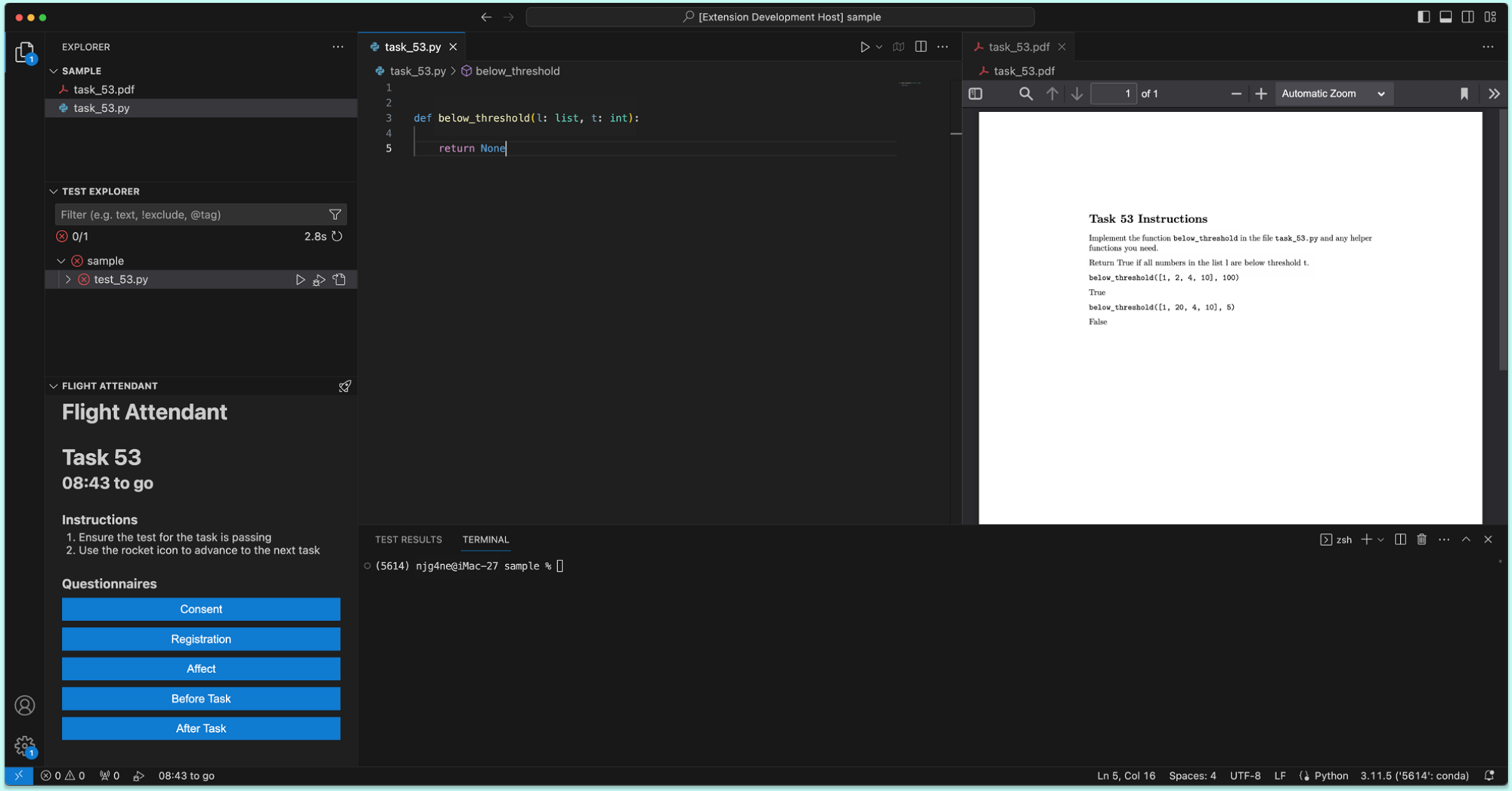
We took repeated measures four times.



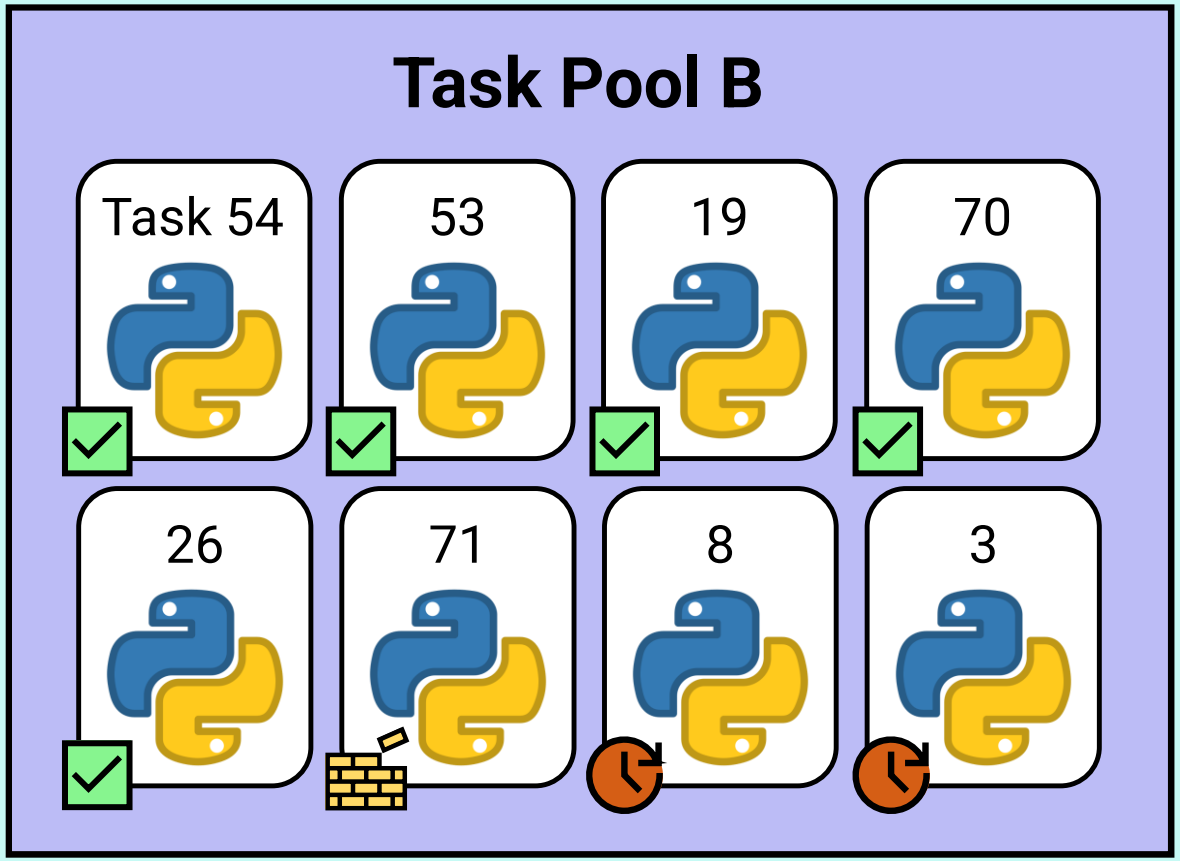
Legend

- Emotional affect
- 20-minute trial
- Ordinal Score
- Self Grade
- NASA-TLX

Participants worked in Visual Studio Code.



We scored task pools ordinally.

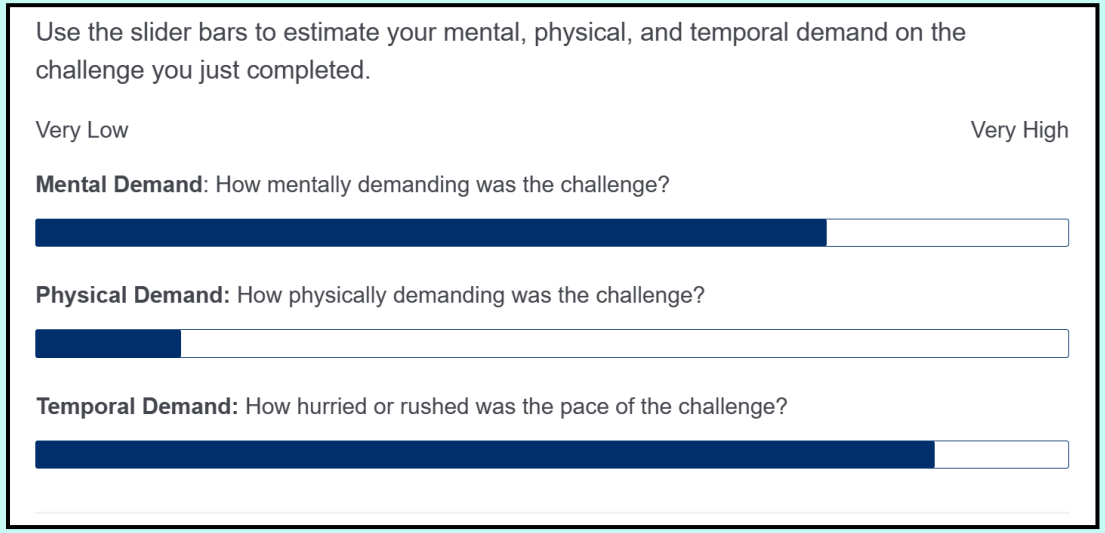


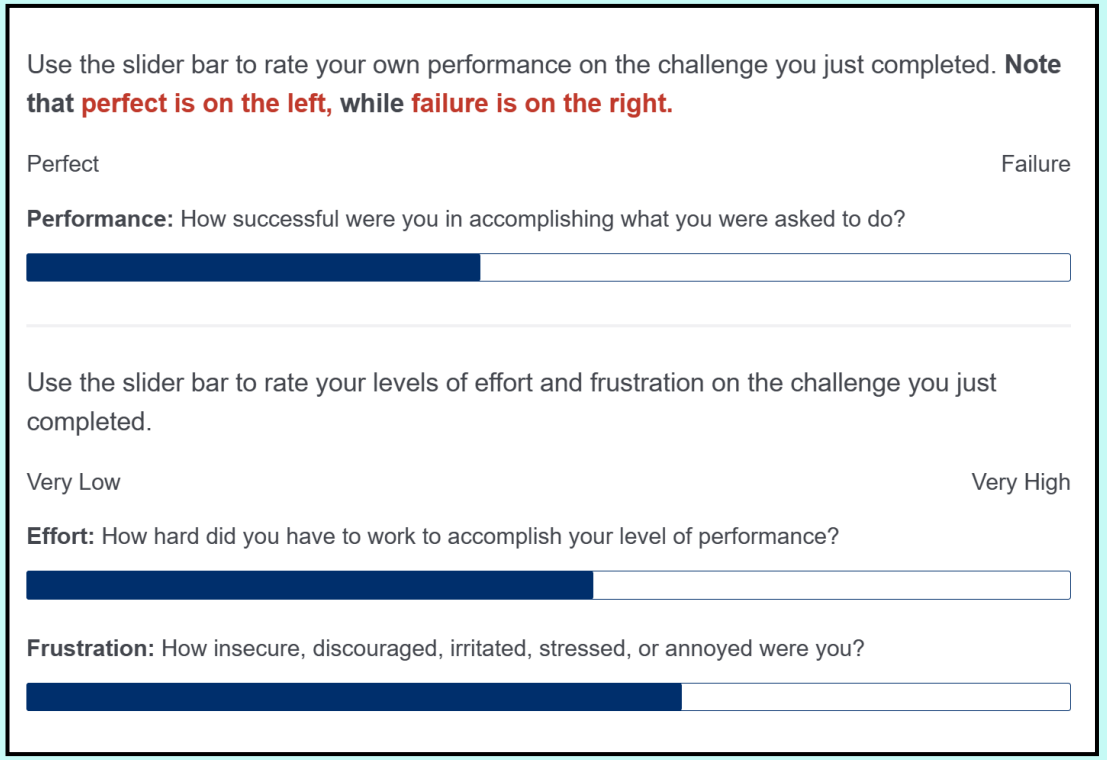
Score = 5

Legend

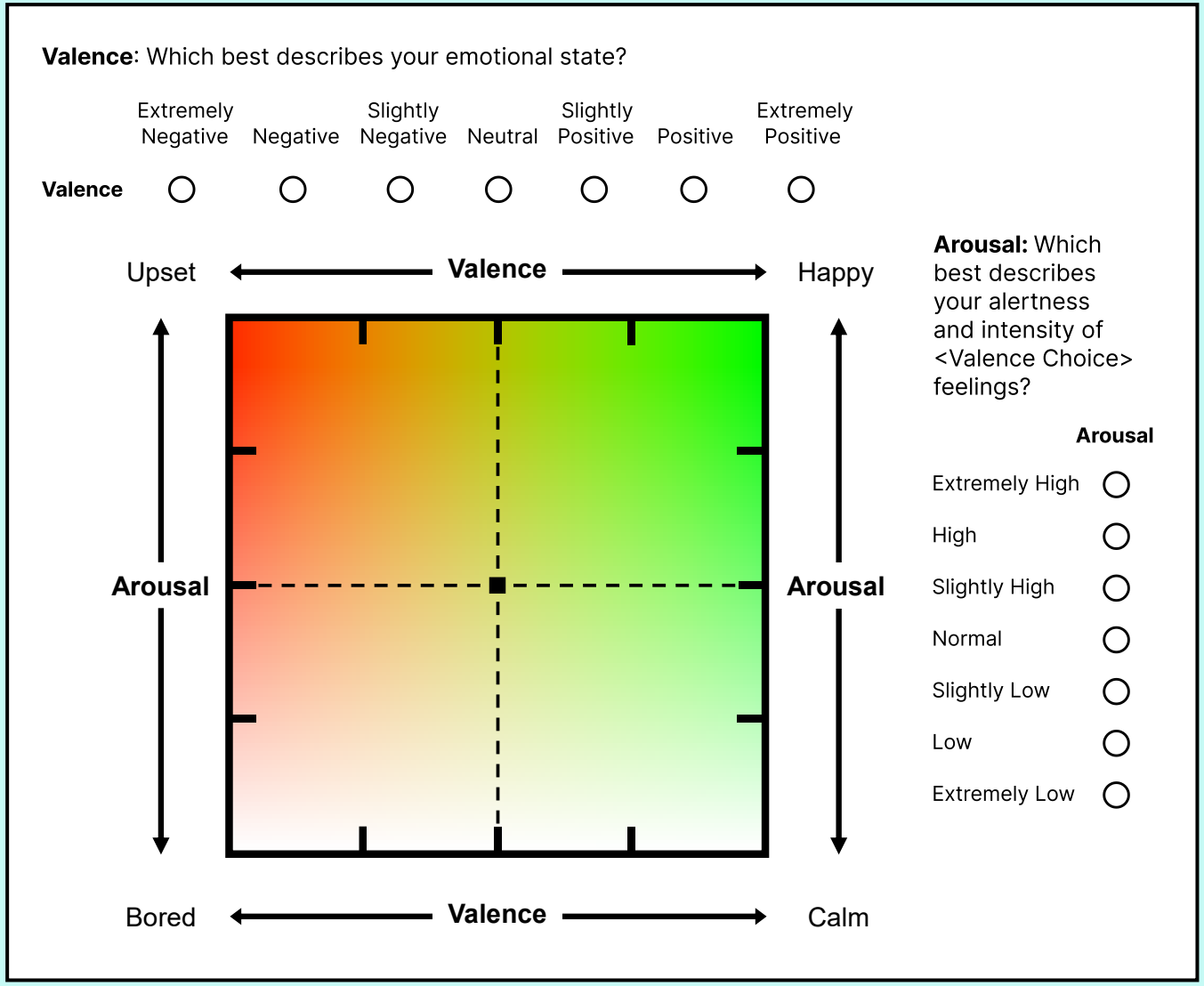
- Passed
- In-progress
- Not started

We measured workload with NASA-TLX.

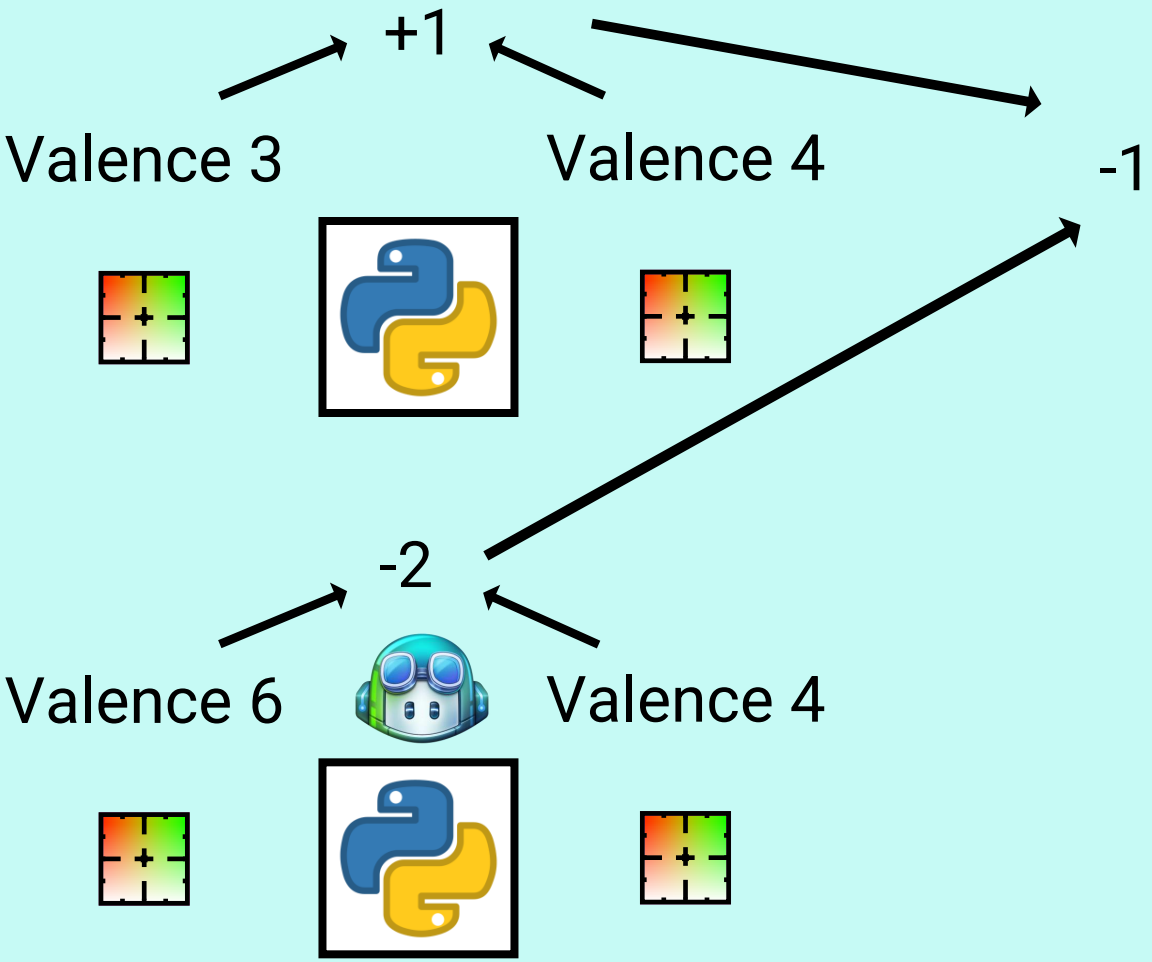




We measured emotion on two dimensions.



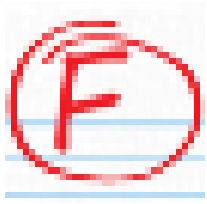
We measured **changes** in emotional state.



We measured self-efficacy with a letter grade.

If you had to give **yourself** a grade for your work on the last challenge, what would it be?

If you had to give **GitHub Copilot** a grade for its contribution to the last challenge, what would it be?



What percentage of the work for the challenge do you think **GitHub Copilot** was responsible for?

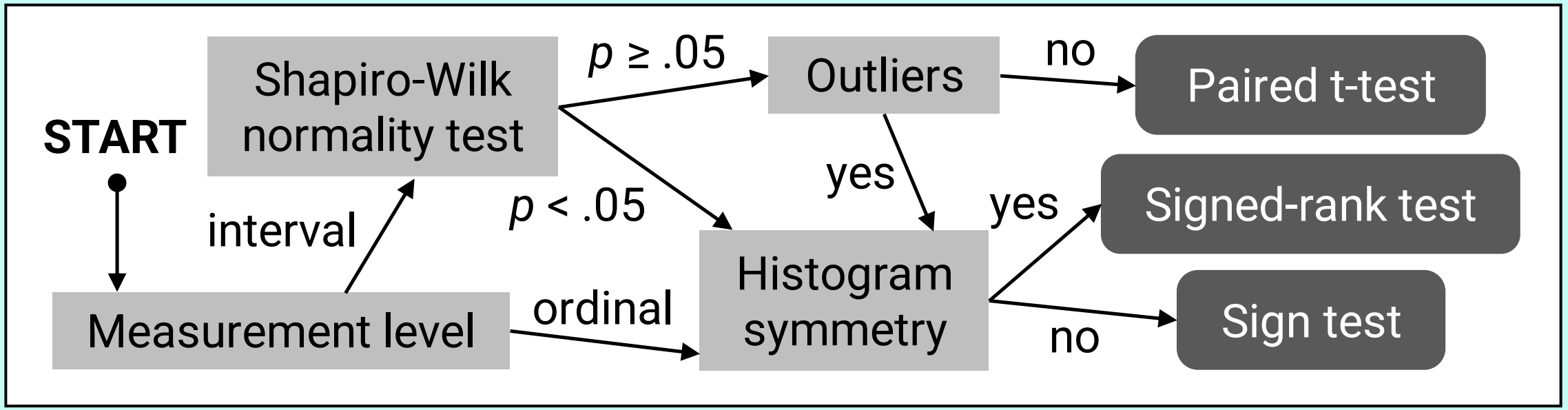
0% 100%

75

% of Work done by Copilot

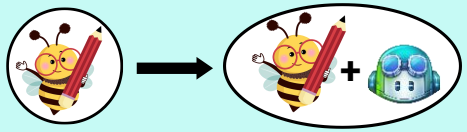
Grade	Value
A+	13
A	12
A-	11
B+	10
...	...
D	3
D-	2
F	1

We took a conservative statistical approach.



Results

RQ1 Results



	Change	p	
Score	$\widetilde{\Delta x} = 1$	$\mathbf{.001}*$	**Be more successful**
Δ Valence	$\overline{\Delta x} = 0.206$	.436	Feel better
Δ Arousal	$\widetilde{\Delta x} = -.5$	.191	Feel more relaxed
TLX Mental	$\overline{\Delta x} = -8.971$	$\mathbf{.024}*$	**Think less hard**
TLX Physical	$\widetilde{\Delta x} = 0$	.773	Exert no differently
TLX Temporal	$\overline{\Delta x} = -5.294$	.200	Feel less rushed
TLX Performance	$\overline{\Delta x} = -6.029$	.230	Feel more successful
TLX Effort	$\widetilde{\Delta x} = -5$	$\mathbf{.043}*$	**Try less hard**
TLX Frustration	$\overline{\Delta x} = -2.206$	.581	Feel less frustrated
Self-Grade	$\overline{\Delta x} = .265$	.455	Feel more successful

RQ2 Results

		1st Trial → 2nd Trial	Change	*p*	
Solo 	 Score	—	$\widetilde{\Delta x} = 0$	1.000	Plateau over time
	 Self-Grade	—	$\widetilde{\Delta x} = 0$	.594	Feel the plateau
with AI 	 Score	↗	$\widetilde{\Delta x} = 1$	**0.046***	**Improve over time**
	 Self-Grade	↗	$\overline{\Delta x} = 1.588$	**0.021 ***	**Feel the improvement**

Discussion

Internet may explain performance differences.

Objectively Measuring Performance

Vaithilingam & Glassman [35], Xu et al. [37]

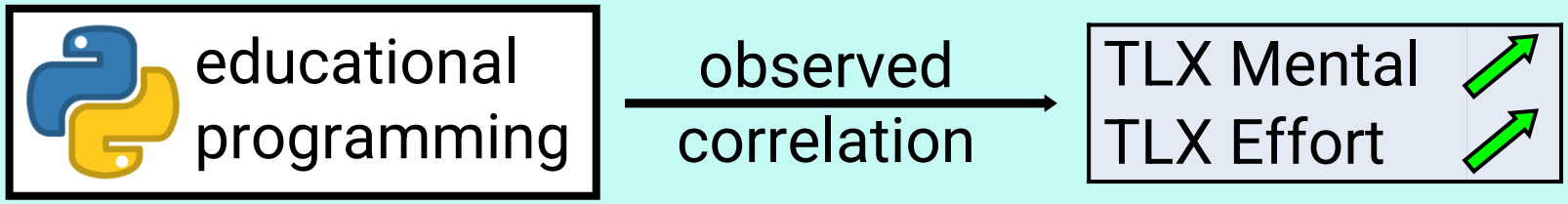
Significant Performance *Benefit*

Kazemitabaar et al. [21]

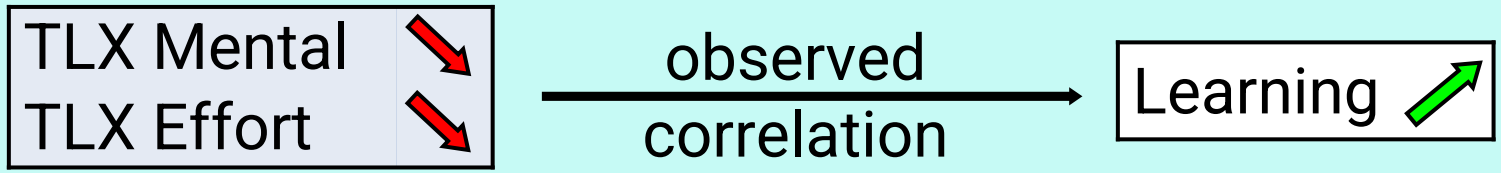
Our study

Workload decreases could aid learning.

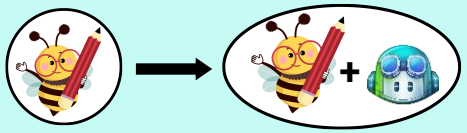
Al Madi et al. [2]



Gerjets et al. [13]

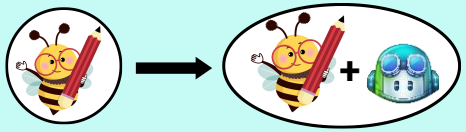


AI may not make you feel much better.



		Change	*p*	
Δ Valence	↗	$\overline{\Delta x} = 0.206$	.436	Feel better
Δ Arousal	↘	$\widetilde{\Delta x} = -.5$	.191	Feel more relaxed
TLX Frustration	↘	$\overline{\Delta x} = -2.206$	.581	Feel less frustrated

Novices don’t take credit for AI’s work.



		Change	*p*
TLX Performance	↘	$\overline{\Delta x} = -6.029$	.230
Self-Grade	↗	$\overline{\Delta x} = .265$	.455

Feel more successful

Novices take credit when they use AI better.

1st Trial → 2nd Trial

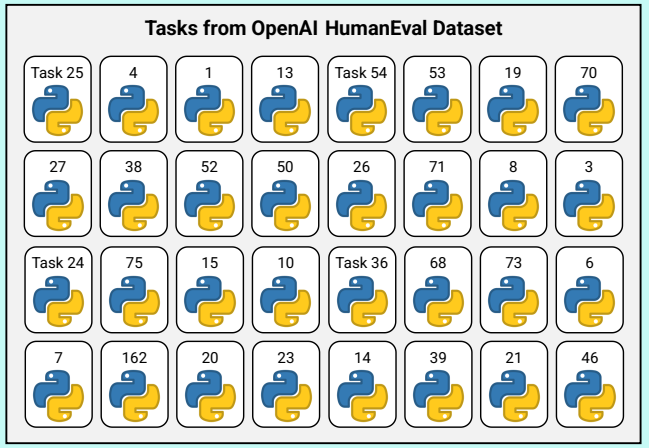
		Change	*p*	
with AI 	 Score	↗ $\widetilde{\Delta x} = 1$	**0.046***	**Improve over time**
	 Self-Grade	↗ $\overline{\Delta x} = 1.588$	**0.021 ***	**Feel the improvement**

Conclusion

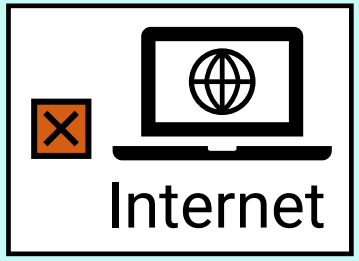
Ecological and external validity are concerns.

Training Data

contained?



n = 17 all from UNIVERSITY of VIRGINIA



Key Takeaways

- Novice programmers…
  - can **be more successful** with AI
  - can **think and try less hard** with AI
  - can **improve** at using AI **over time** and **take credit for the improvement**
- But…
  - **don’t take credit** for AI’s work
  - **don’t feel much better** using AI

Nicholas Gardella, Raymond Pettit and Sara L. Riggs. 2024. Performance, Workload, Emotion, and Self-Efficacy of Novice Programmers Using AI Code Generation. In Proceedings of the 2024 Conference on Innovation and Technology in Computer Science Education V. 1 (ITiCSE 2024), July 8- 10,2024, Milan, Italy. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3649217.3653615

Contact: njg4ne@virginia.edu

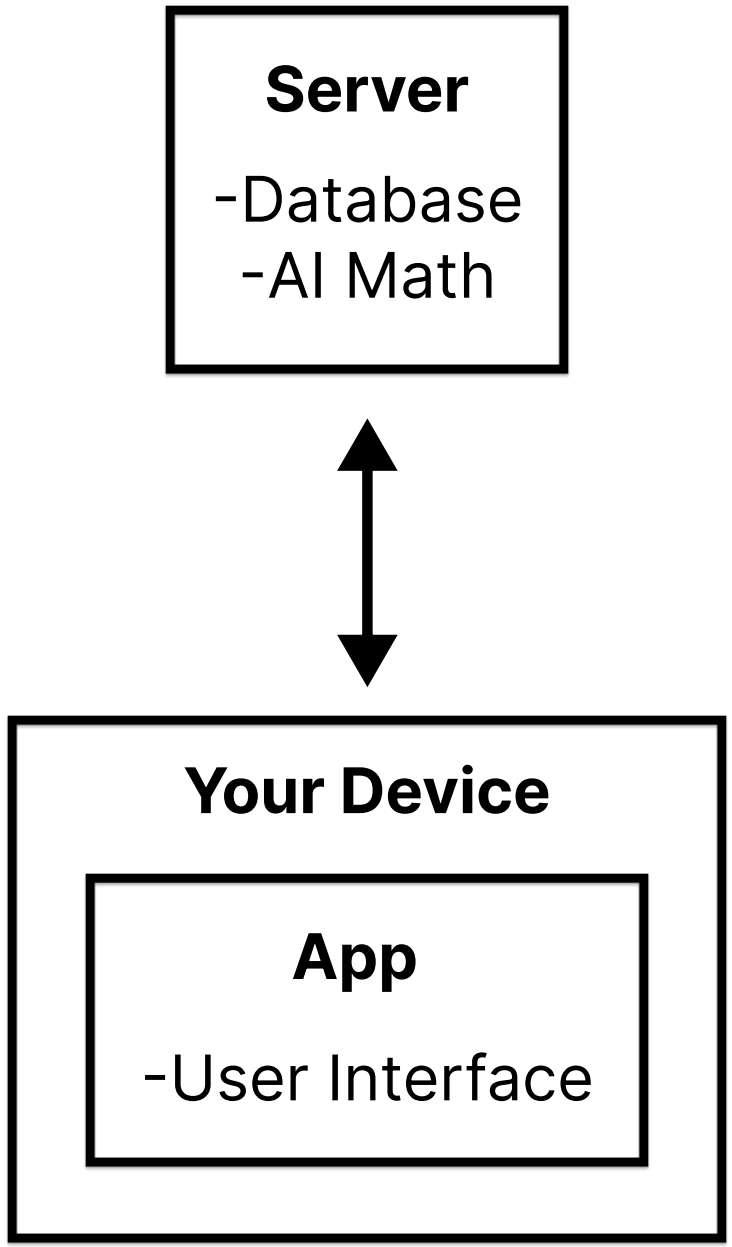
Website: https://n.gardella.cc

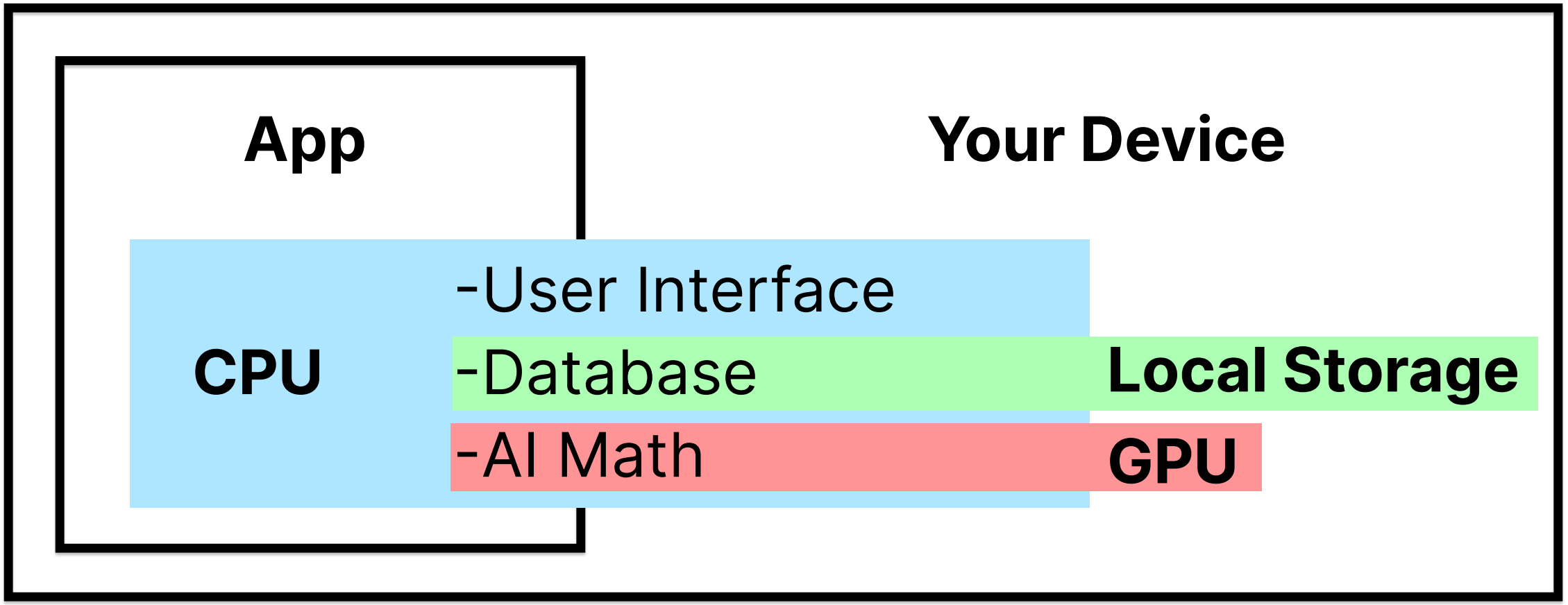
Funding support:

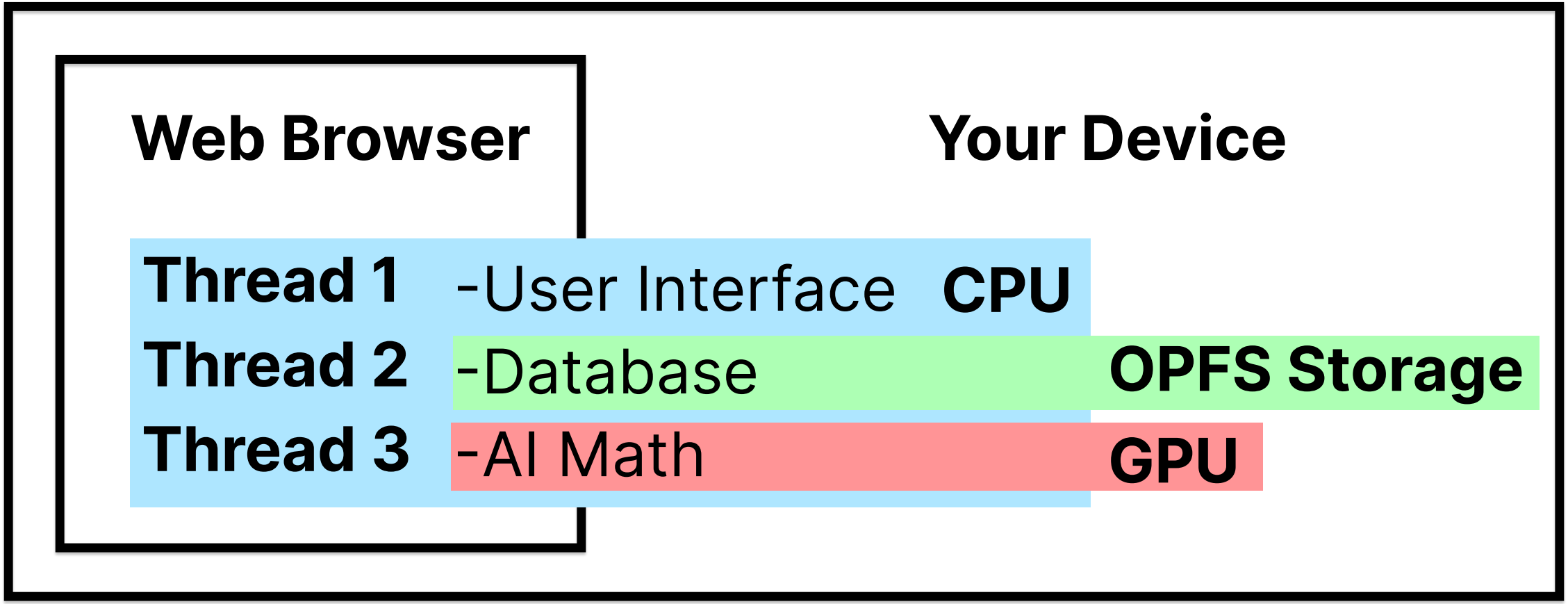
- UVA Distinguished Fellowship
- NSF NRT 1842490
- NSF GRFP 1829004
- Virginia CCI CVN GR102900

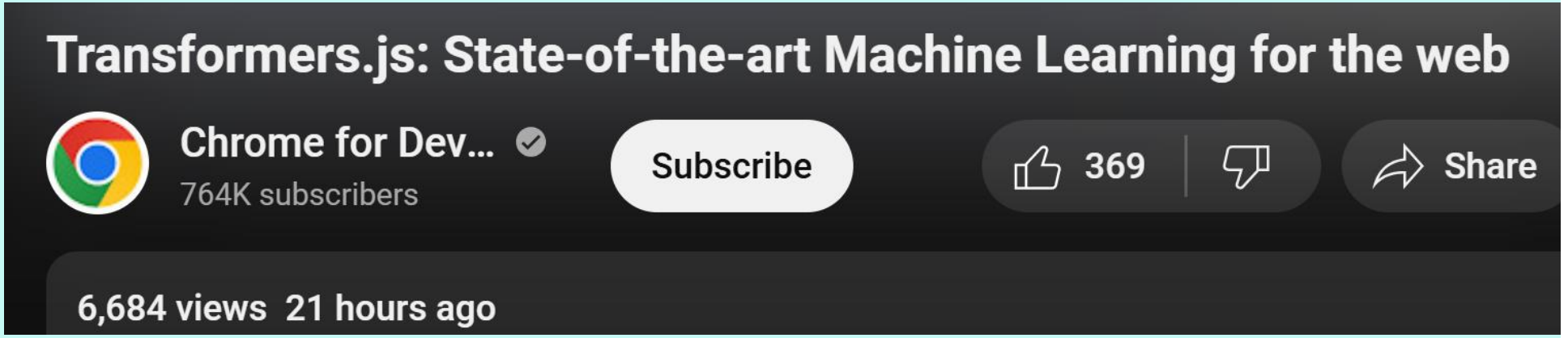
We thank Dylan Do, Zack Landsman, Michael Lin, Alex Myers, Eric Nguyen, Jacob Rice, Allison Sawyer, Olivia Seto, and Alice Warner for assistance in selecting and arranging HumanEval tasks.

Appendix


Server
-Database
-AI Math
Your Device
App
-User Interface


App
Your Device
CPU
-User Interface
-Database
-AI Math
Local Storage
GPU


Web Browser
Your Device
Thread 1
-User Interface
CPU
Thread 2
-Database
OPFS Storage
Thread 3
-AI Math
GPU



https://n.gardella.cc/rag/

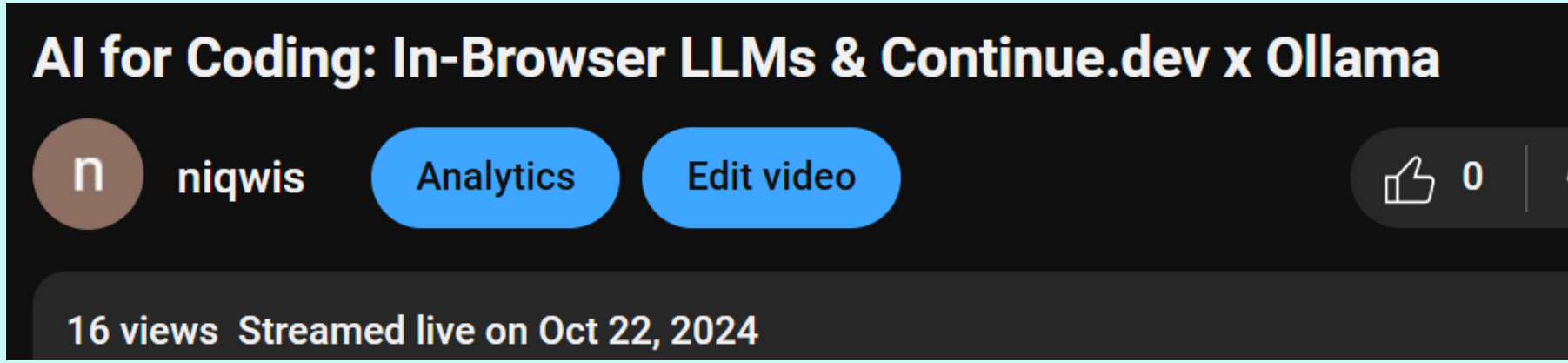


Table 1

	Solo	with AI	Change
Score	$\tilde{x} = 2$	$\tilde{x} = 2.5$	$\widetilde{\Delta x} = 1$
Δ Valence	$\bar{x} = -.088$	$\bar{x} = .118$	$\overline{\Delta x} = 0.206$
Δ Arousal	$\tilde{x} = 0$	$\tilde{x} = -.5$	$\widetilde{\Delta x} = -.5$
TLX Mental	$\bar{x} = 58.382$	$\bar{x} = 49.412$	$\overline{\Delta x} = -8.971$
TLX Physical	$\tilde{x} = 5$	$\tilde{x} = 2.5$	$\widetilde{\Delta x} = 0$
TLX Temporal	$\bar{x} = 54.559$	$\bar{x} = 49.265$	$\overline{\Delta x} = -5.294$
TLX Performance	$\bar{x} = 61.912$	$\bar{x} = 55.882$	$\overline{\Delta x} = -6.029$
TLX Effort	$\tilde{x} = 62.5$	$\tilde{x} = 55$	$\widetilde{\Delta x} = -5$
TLX Frustration	$\bar{x} = 47.059$	$\bar{x} = 44.853$	$\overline{\Delta x} = -2.206$
Self-Grade	$\bar{x} = 7.647$	$\bar{x} = 7.912$	$\overline{\Delta x} = .265$

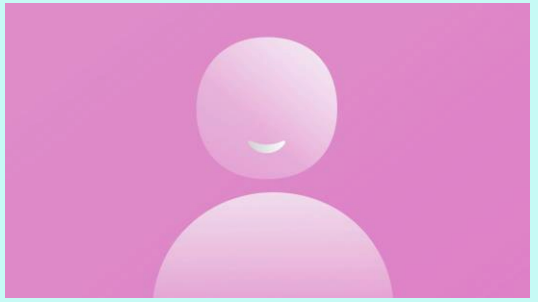
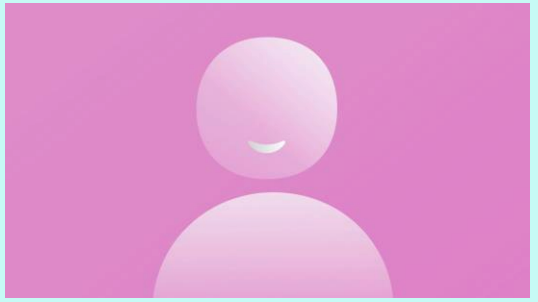


Table 2

	Statistic	*p*	d
Score	$\mathbf{z_{Wilcoxon} = 3.012}$	$\mathbf{.001*}$	$\mathbf{.368}$
Δ Valence	$t(16) = .800$	$.436$	$.098$
Δ Arousal	$z_{Wilcoxon} = -1.329$	$.191$	$-.162$
TLX Mental	$\mathbf{t(16) = -2.487}$	$\mathbf{.024*}$	$\mathbf{-.304}$
TLX Physical	$z_{Sign\ Test} = -.289$	$.773$	$-.035$
TLX Temporal	$t(16) = -1.338$	$.200$	$-.163$
TLX Performance	$t(16) = -1.248$	$.230$	$-.152$
TLX Effort	$\mathbf{z_{Wilcoxon} = -2.023}$	$\mathbf{.043*}$	$\mathbf{-.247}$
TLX Frustration	$t(16) = -.563$	$.581$	$-.069$
Self-Grade	$t(16) = .765$	$.455$	$.093$



Tables 3 & 4

		1st Trial	2nd Trial	Change
Solo	Score	$\tilde{x} = 2$	$\tilde{x} = 2$	$\widetilde{\Delta x} = 0$
	Self-Grade	$\tilde{x} = 8$	$\tilde{x} = 7.25$	$\widetilde{\Delta x} = 0$
with AI	Score	$\tilde{x} = 2$	$\tilde{x} = 3$	$\widetilde{\Delta x} = 1$
	Self-Grade	$\bar{x} = 7.118$	$\bar{x} = 8.706$	$\overline{\Delta x} = 1.588$

		Statistic	*p*	d
Solo	Score	$z_{Wilcoxon} = .090$	1.000	.016
	Self-Grade	$z_{Wilcoxon} = -.517$	.594	−.091
with AI	**Score**	$z_{Wilcoxon} = 2.029$	**0.046***	.348
	Self-Grade	$t(16) = 2.567$	**0.021***	.440