Temporal Memory

Exploring ChatGPTs Predictive Abilities in Human Location Memory

Simon Dennis
Supervising Professor

Yu Pin Gan
Summer Research Intern

Where were you yesterday at 7pm?

Are you able to recall this?

The Case of Ronald Cotton

- Cotton (left) sentenced to life & 50 years in 1984
- Two counts of rape and two counts of burglary
- Cotton was exonerated in 1995, after spending over 10 years in prison when DNA evidence demonstrated his innocence.

Ronald Cotton (left) and Bobby Poole (Right)

The Case of Ronald Cotton

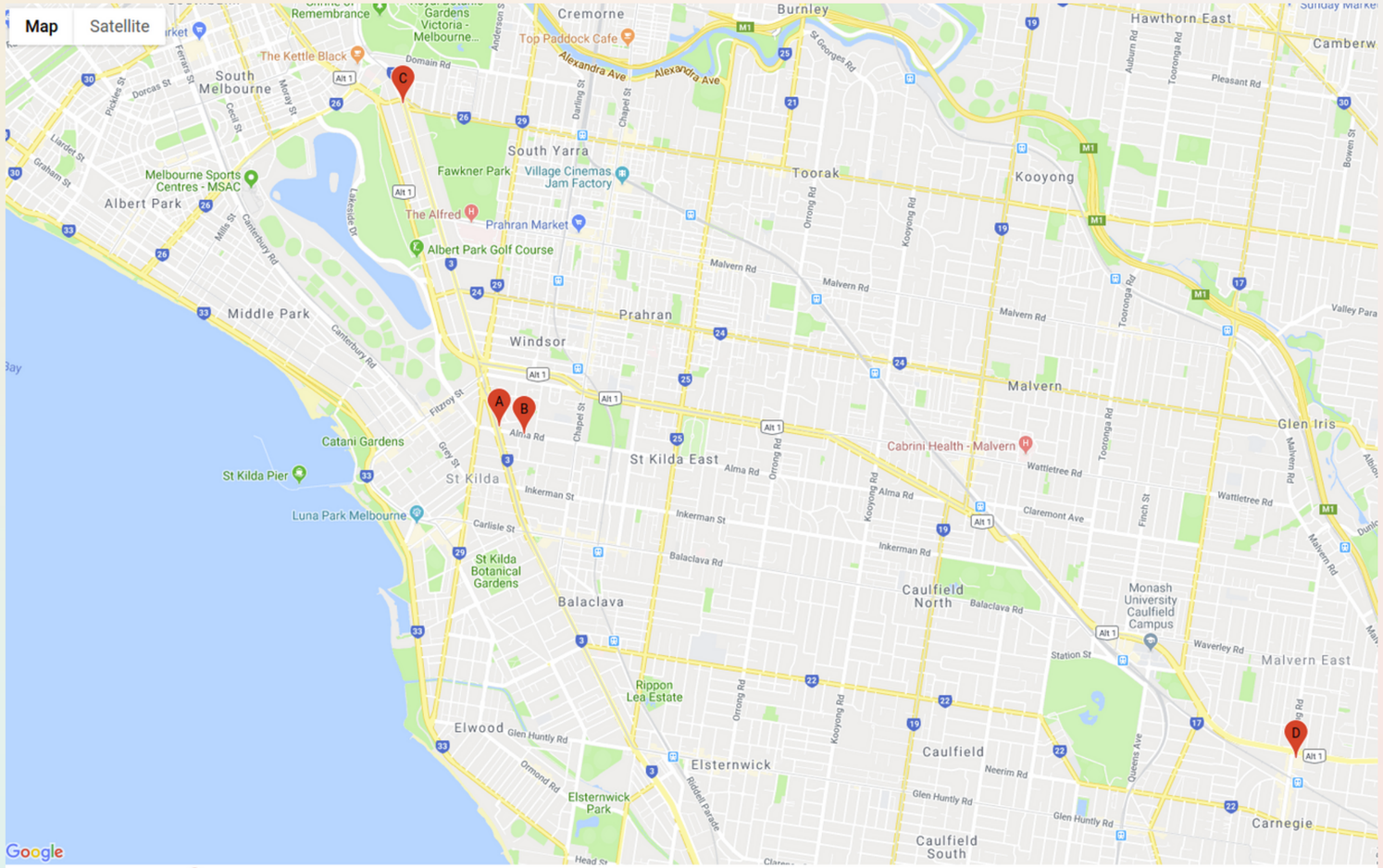
Ronald Cotton (left) and Jennifer Cannino Thompson (right)

- Jennifer Thompson (victim) misidentified Cotton 3 times
- **Cotton misremembered where he was at the time of the crime (a week ago)**
- Alibi could not be corroborated, leading jurors to conclude Cotton was lying

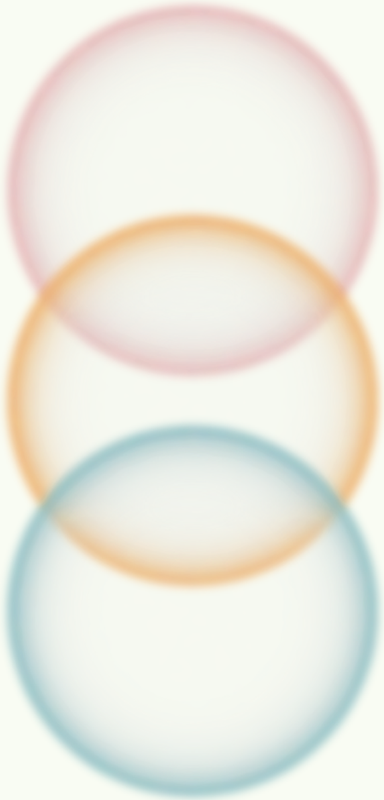
Data (Events)

- N=66
- Collected across 2 weeks
- Accelerometry Data 10 times a second
- GPS Coordinates once every 10 minutes
- 3 seconds of Audio segments every 10 minutes
- Emotion (Discrete): Angry, Anxious, Bored, Disappointed, Irritable, Sad, Confident, Content, Excited, Happy, Relaxed

Data (Experiment)



- After a one week retention interval participants were asked where they were at different times
- Choose 1 from 4 options
- **67% Correct**



How do LLMs Compare? Are the Mistakes Similar?

Translate the task such that LLMs can process it and then perform some analysis on the results

Methodology

Production of Results
• Pick an LLM (GPT 3.5 Turbo 1106) • Instruct CGPT (ChatGPT) on how to respond (encoded as a ‘system’ role) • Pull each entry (for each participant) from the events file and extract only the **datetime** and **GPS cluster**, and encode each event (as an entry from a user) • Pull each entry (for each participant) from the experiments file and extract only the **datetime**, the **4 options** and the **correct answer**, and encode each event (as an entry from a user) • Give the instructions and event trials to CGPT (API Calls), run it against each experiment trial for every participant and save the output as a .csv file

Methodology

Methodology

Result Analysis
• Run the CGPT output against the correct answers to generate the accuracy per model • Check the match between humans and CGPT when they are both incorrect • Generate confusion matrices

Prompt Wording

Base Case

- Instruction
  - ‘Today is {**new day**}, {**new date**}. Your task is to judge your location based on the given time. You'll have 4 options to choose from. You must always choose the most suitable letter and return only the letter.
  - Example:
    - Sunday, August 11 at 5PM, you were at location 3.
    - Friday, August 16 at 12AM, you were at location 0.
  - Question:
    - Where were you on Friday, 16 August 12AM? A. 2 B. 3 C. 0 D. 30
      - Response: C
    - Where were you on Sunday, 11 August 5AM? A. 2 B. 3 C. 9 D. 30
      - Response:B’

Prompt Wording

Base Case

- Event Trials
  - Format: {**day**}, {**date**} {**time**} you were at location {**gps cluster**}
  - ‘Tuesday, 20 August 8AM you were at location 0'

- Experiment Trials
  - Format: Where were you on {**day**}, {**date**} {**time**}?\nA {**loc1**}\B {**loc2**}\nC {**loc3**}\nD {**loc4**}
  - 'Where were you on Sunday, 11 August 3PM?\nA 23\nB 37\nC 0\nD 26'

Prompt Wording

Dependent Case

- Every last question is added into the conversation

Independent Case

- Every question is independently submitted to CGPT without the previous questions

Independent with Spacing (0 Shot) Case

- Independent case
- Added spacing between the time
  - instead of 8AM -> 8 AM
- Removed examples from the instructions (0 shot)

Prompt Wording

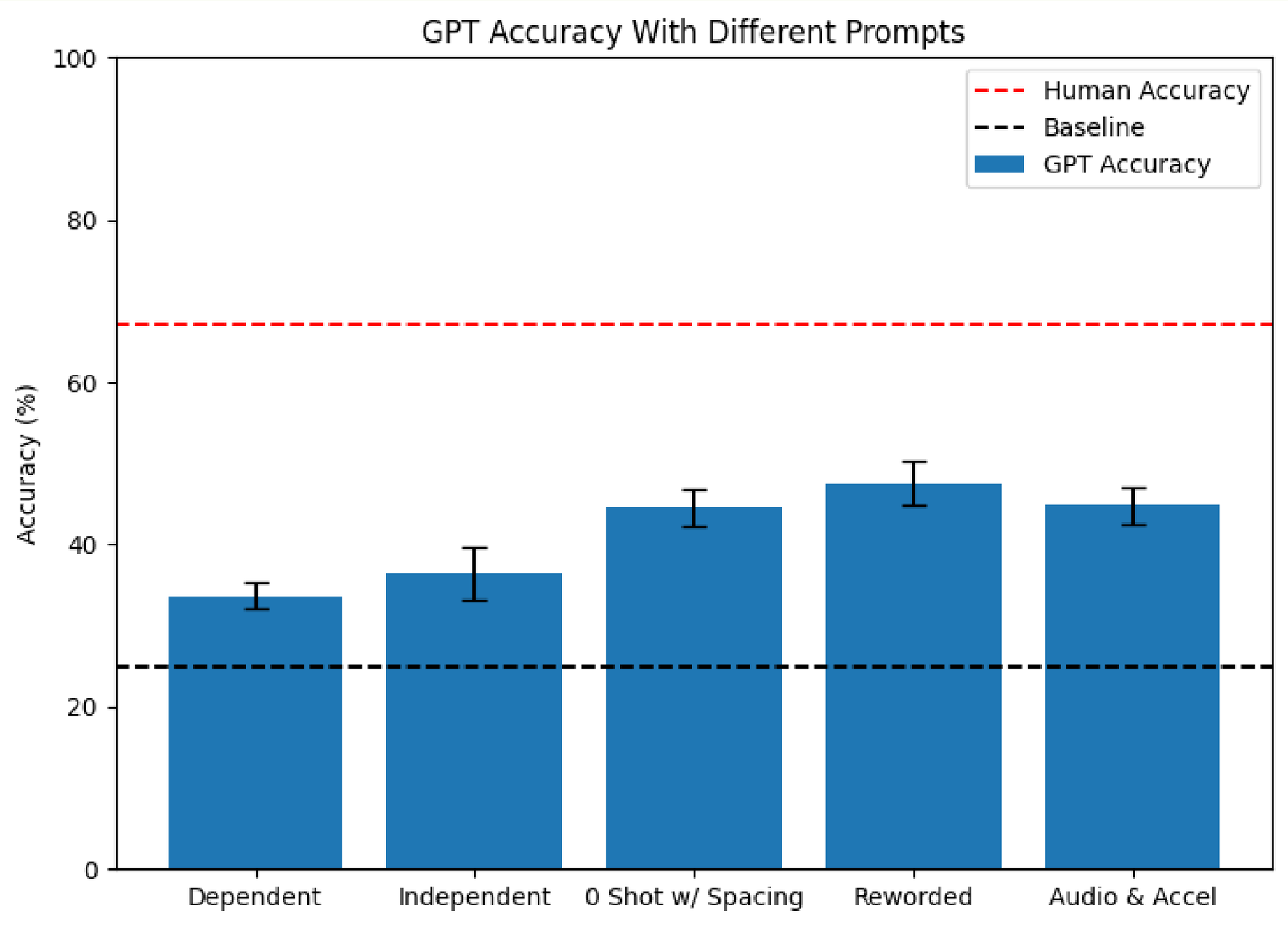
Independent with Rewording (0 Shot) Case

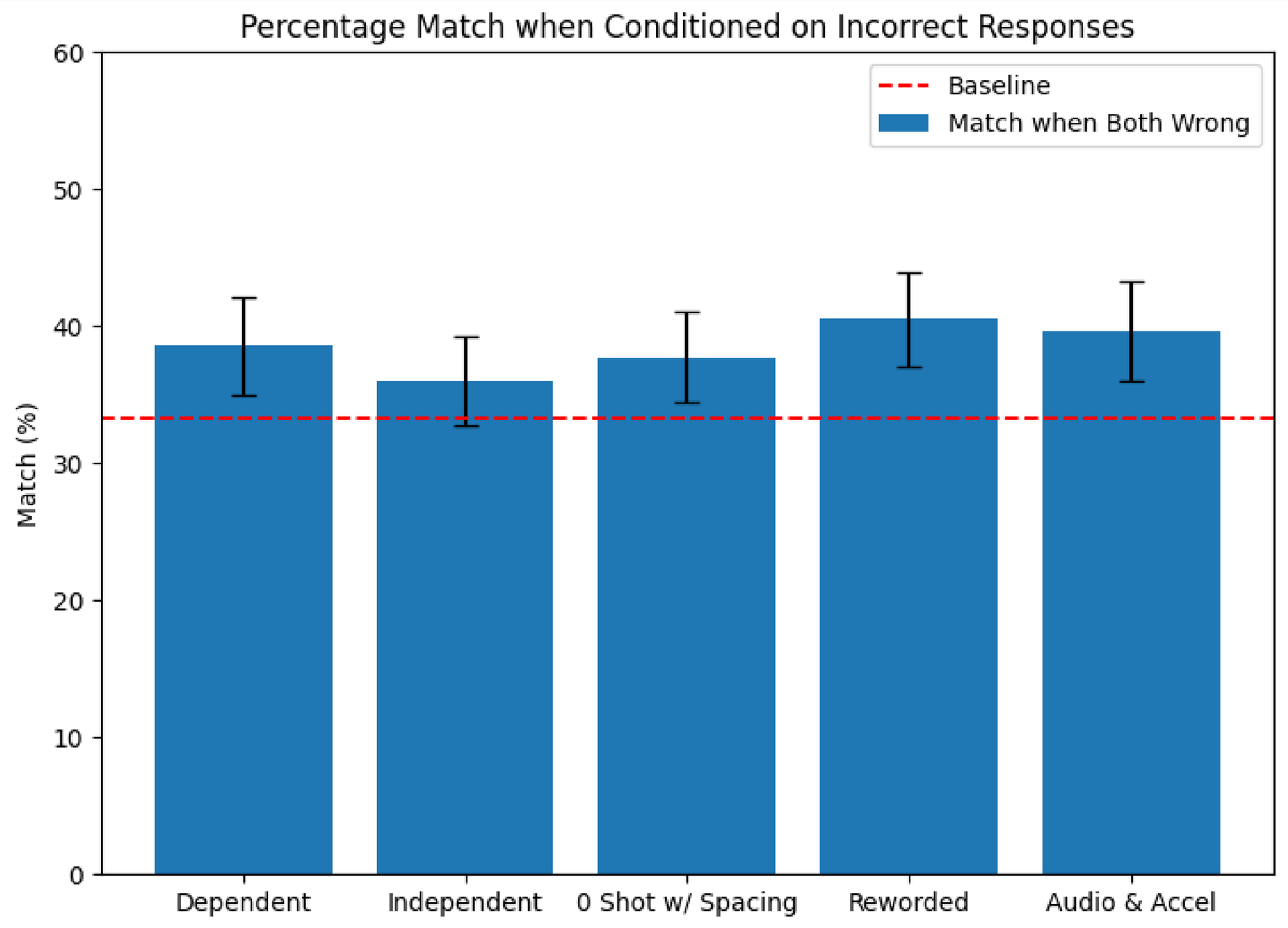
- Independent Case
- The time is brought forward in all prompts, followed by day, then date
- 0 Shot

Independent with Rewording, Audio and Accelerometry (0 Shot) Case

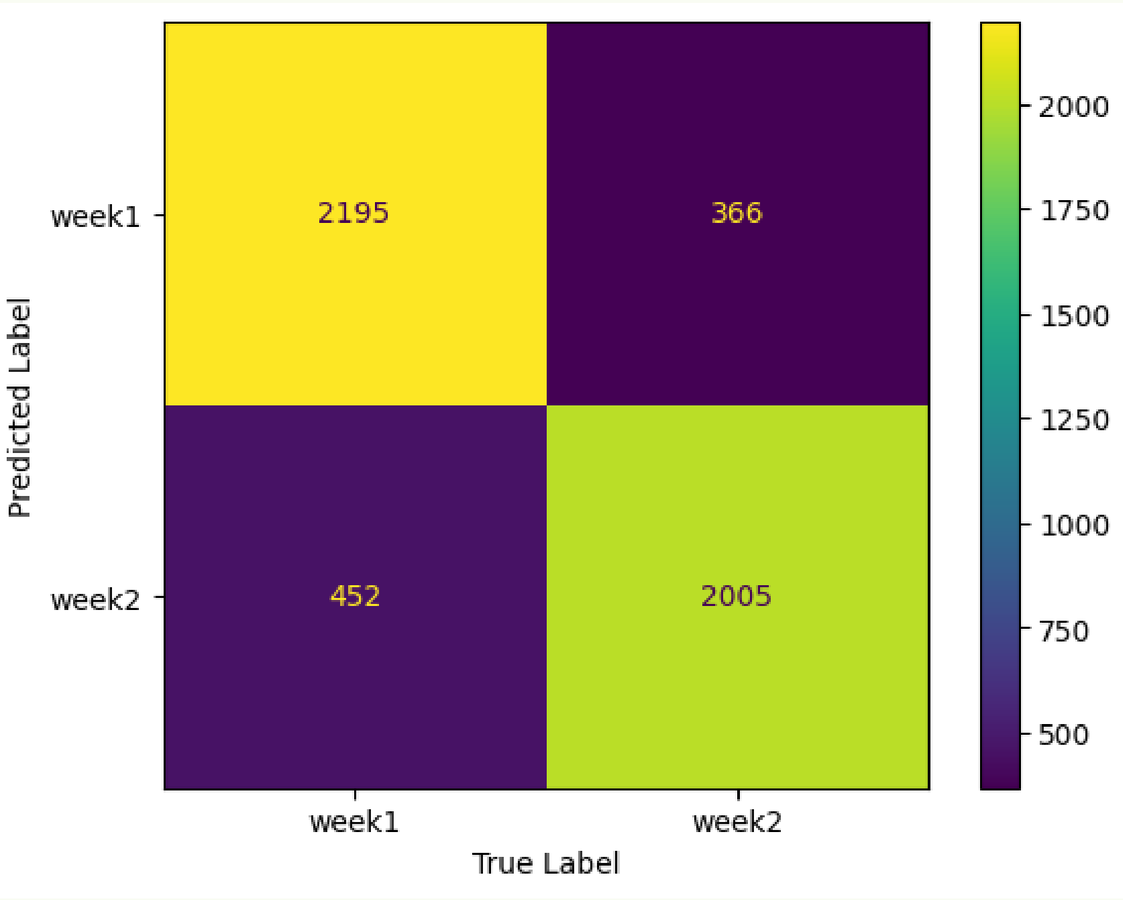
- Same as the previous case
- Added the audio and accelerometry data into the prompt
  - … you heard sound {**sound cluster**} and you were doing movement {**accel cluster**}

Results


GPT Accuracy With Different Prompts
Human Accuracy
Baseline
GPT Accuracy
Accuracy (%)
100
80
60
40
20
0
Dependent
Independent
0 Shot w/ Spacing
Reworded
Audio & Accel


Percentage Match when Conditioned on Incorrect Responses
Baseline
Match when Both Wrong
Match (%)
60
50
40
30
20
10
0
Dependent
Independent
0 Shot w/ Spacing
Reworded
Audio & Accel

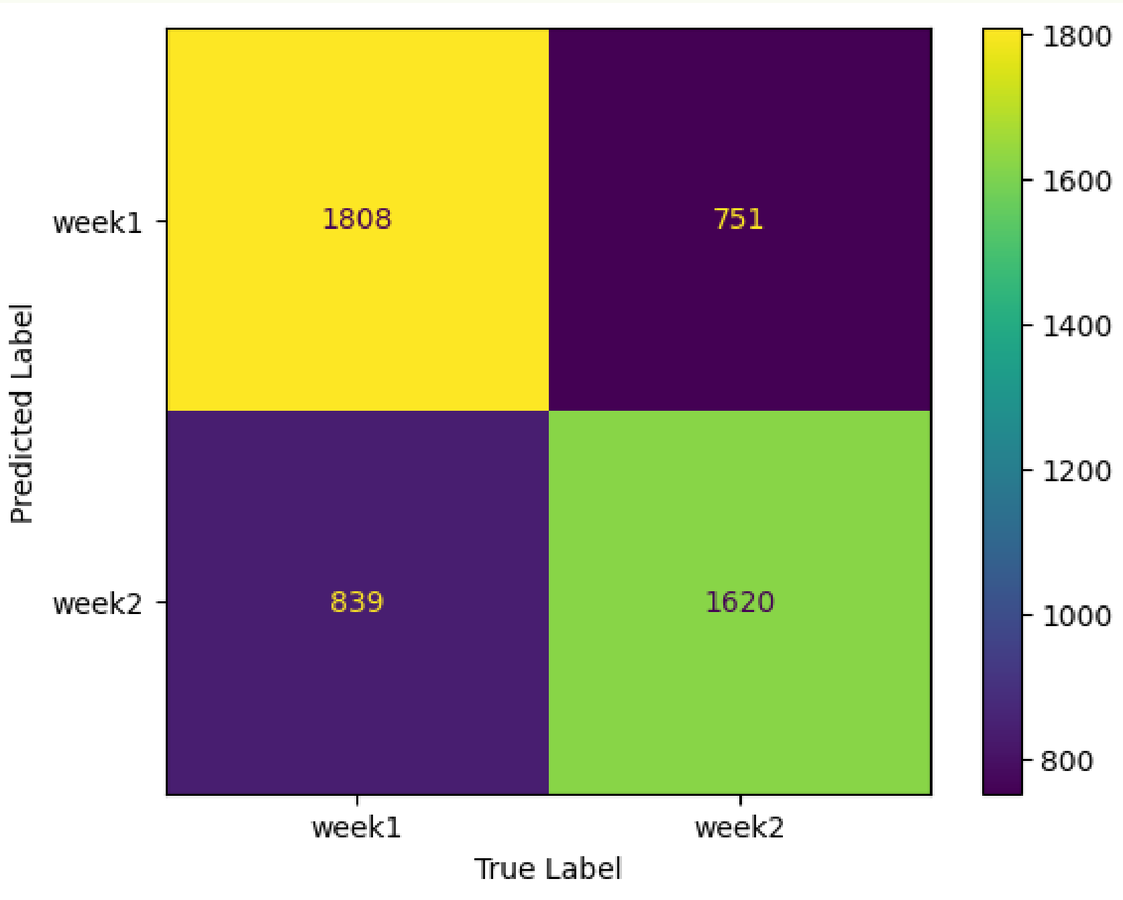
Human Week Recall (Dependent)



Predict W2 when W1: 17.1%

Predict W1 when W2: 15.4%

GPT Week Recall (Dependent)



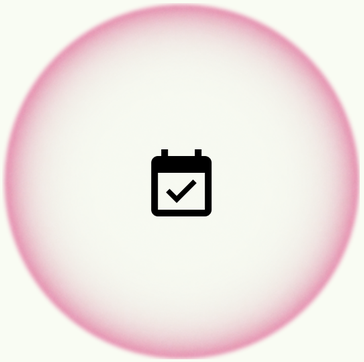
Predict W2 when W1: 31.7%

Predict W1 when W2: 31.6%

Conclusion

CGPT’s accuracy is higher than chance (baseline), but lower than human accuracy

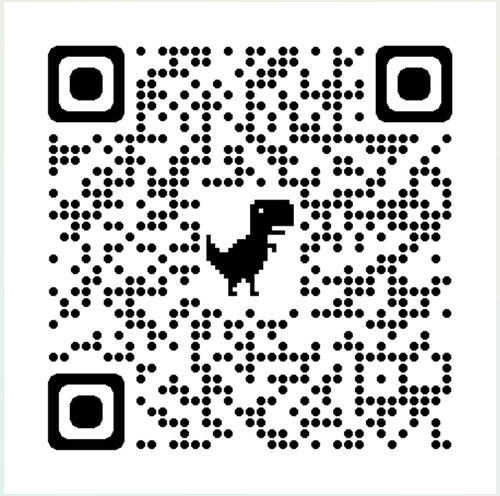
CGPT with Inter-trial dependencies had a higher match when conditioned on the incorrect responses (both humans and CGPT)



Humans show slight forward telescoping whereas the model doesn’t

Thank you!

Do you have any questions?



LinkedIn

Slides & Code

Content

Results

- Higher than chance (baseline), but lower than human accuracy
- Performace is sensitive to small changes
- ~~Accuracy increases as the changes are made (with the exception of adding additional information)~~
    - ~~could be due to the wording of the prompt or additional interference~~

Results

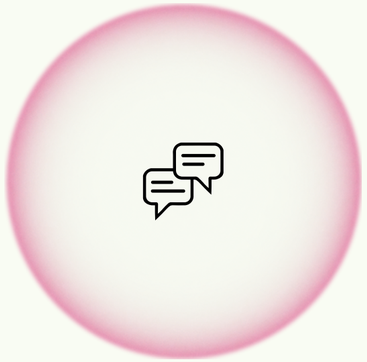
- The baseline is 33.3% since GPT has a 1-in-3 chance of matching the human
  - given that both the human and GPT is both wrong
- Dependency reflects the real world more accurately, it may have a higher match rate if we used the other formats

Results

- GPT is more prone to mistakes (as supported by previous bar charts)
- GPT’s prediction across the week is even whereas Humans have a slightly higher tendency to report an event to be more distant
    - Telescoping (Huttenlocher, 1988): Unbiased encoded information is subjected to error that increases with age of memory
    - Backward Telescope: Recent events are reported as more remote
    - **Forward Telescope: Distant event are reported as more recent**

Limitations, Recommendations & Conclusion

More information doesn’t necessarily give better accuracy



Slight tweaks in the prompt can impact the experiment. Explore the prompts more.

Reducing GPT interference helps improve accuracy but may strays away from mimicking human’s mistakes

Explore Open Source Models to compare and contrast against GPT.