



GPTs and how to prompt them

Marcel Binz

September 12, 2023

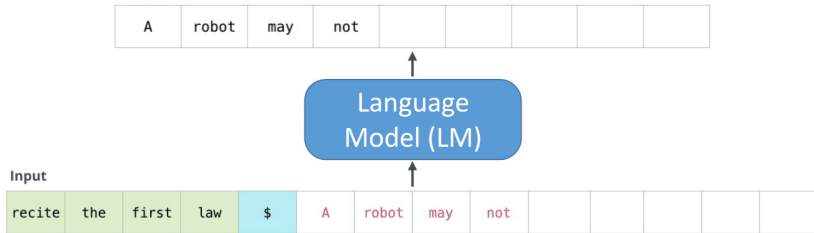
Max Planck Institute for Biological Cybernetics
Computational Principles of Intelligence Lab

Plan for today

1. Using cognitive psychology to understand GPT-3 (15 min)
2. GPTs and how to prompt them (15 min)
3. Hacking (45 min)
4. Discussion (15 min)

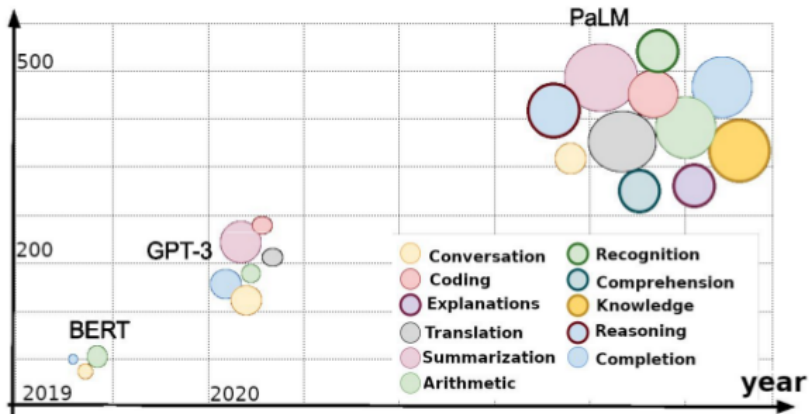
Large language models

Large language models are neural networks trained to predict the next word for a given text sequence.



Large language models

They have many emergent abilities:



Using cognitive psychology to understand GPT-3

They also do well in many psychological experiments.



Using cognitive psychology to understand GPT-3

[Marcel Binz](#)   and [Eric Schulz](#) [Authors Info & Affiliations](#)

Edited by Terrence Sejnowski, Salk Institute for Biological Studies, La Jolla, CA; received October 29, 2022; accepted November 27, 2022

February 2, 2023 | 120 (6) e2218523120 | <https://doi.org/10.1073/pnas.2218523120>

Binz, Marcel, and Eric Schulz. "Using cognitive psychology to understand GPT-3." Proceedings of the National Academy of Sciences 120.6 (2023): e2218523120.

Using cognitive psychology to understand GPT-3

[Marcel Binz](#)   and [Eric Schulz](#) [Authors Info & Affiliations](#)

Edited by Terrence Sejnowski, Salk Institute for Biological Studies, La Jolla, CA; received October 29, 2022; accepted November 27, 2022

February 2, 2023 | 120 (6) e2218523120 | <https://doi.org/10.1073/pnas.2218523120>

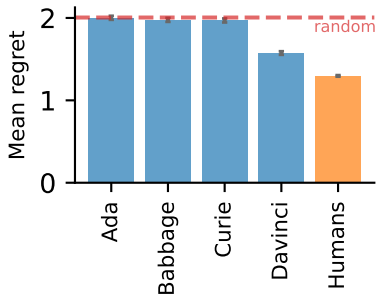
Imagine we gave you 2,000 dollars right now to play a game.

Q: Which option do you prefer?

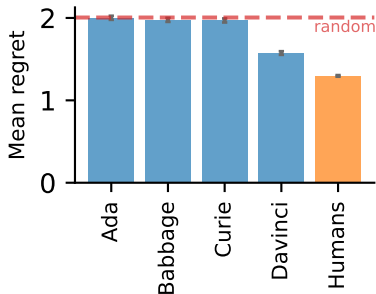
- Option F: 50% chance you will lose 1,000 dollars (50% chance of losing 0 dollars).
- Option J: 100% chance you will lose 500 dollars.

A: Option [insert]

Using cognitive psychology to understand GPT-3



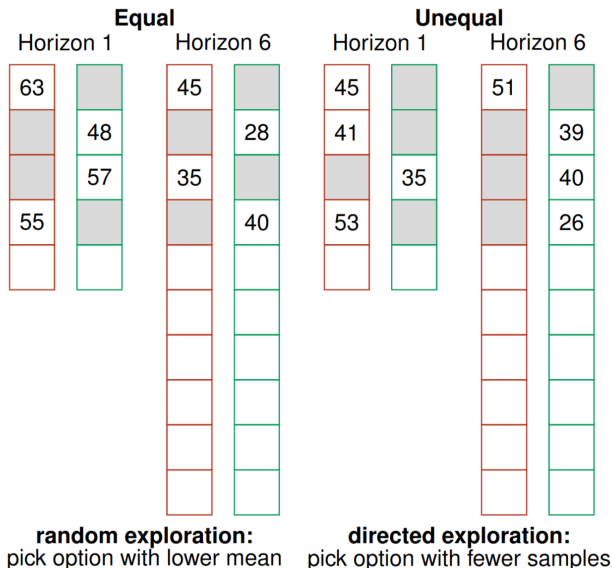
Using cognitive psychology to understand GPT-3



GPT-3 showed a framing effect, a certainty effect, and an overweighting bias.

GPT-3 did not show a reflection effect, an isolation effect, and a sensitivity to magnitude perception.

Using cognitive psychology to understand GPT-3



Using cognitive psychology to understand GPT-3

You are going to a casino that owns two slot machines. You earn money each time you play on one of these machines.

You have received the following amount of dollars when playing in the past:

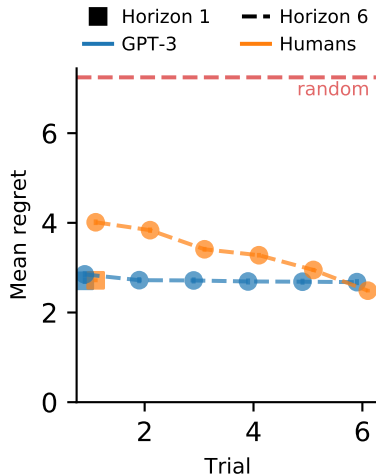
- Machine F delivered 51 dollars.
- Machine J delivered 39 dollars.
- Machine J delivered 40 dollars.
- Machine J delivered 26 dollars.

Your goal is to maximize the sum of received dollars within six additional rounds.

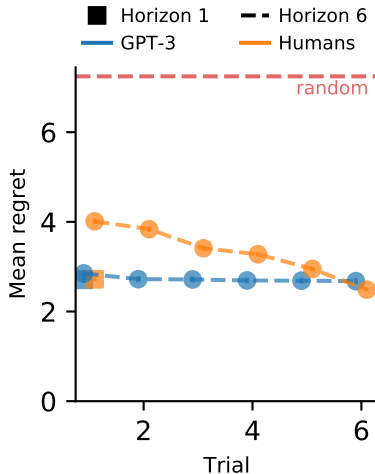
Q: Which machine do you choose?

A: Machine [insert]

Using cognitive psychology to understand GPT-3



Using cognitive psychology to understand GPT-3



GPT-3 shows no learning effects.

GPT-3 sometimes selects less rewarding option (random exploration).

GPT-3 does not select option with fewer observations (directed exploration).

Using cognitive psychology to understand GPT-3

Domain	Good Performance?	Human-Like?
Decision Making	✓	✓
Deliberation	✓	✓
Exploration	✓	✗
Causal Reasoning	✗	✗

Psychological experiments will become to be a useful tool for analyzing increasingly complex artificial agents:

1. Experimental paradigms have been extensively validated over decades.
2. Tasks are often procedurally-generated, thereby avoiding training data contamination.
3. Not only care about performance but instead provide multi-dimensional measurements of behavior.


Plan for the remainder of this session:

- Familiarize yourselves with LLMs and the OpenAI API.
- Understand LLMs better and get a feeling for working with them.
- Learn about different prompting techniques.

PRACTICE IN THE CASE OF ADDITION.

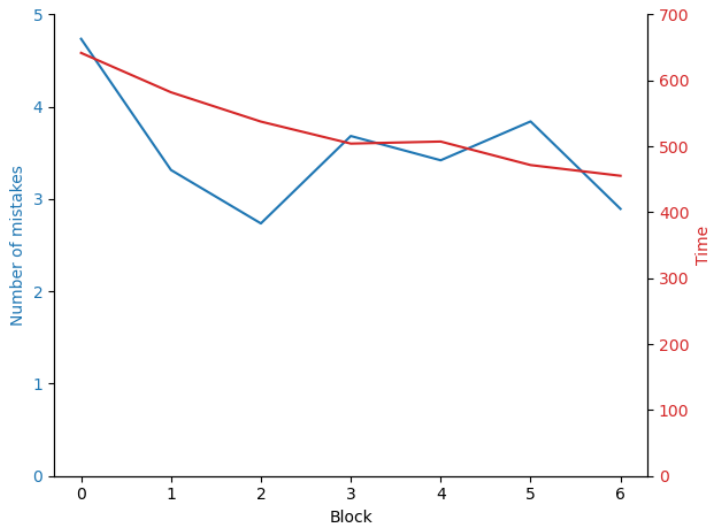
By EDWARD L. THORNDIKE, Teachers College, Columbia University.

The American Journal of Psychology, Vol. 21, No. 3 (Jul., 1910), pp. 483-486 (4 pages)

<https://doi.org/10.2307/1413352> · <https://www.jstor.org/stable/1413352> 

The experiment consisted in adding daily for seven days forty-eight columns each of ten numbers (no 1's or 0's being included). Seven printed blanks had been arranged of equal difficulty.¹ The forty-eight sums were written. The time required was recorded in seconds. The subjects were nineteen university students—eight men and eleven women.

Practice in the case of addition



GPTs and how to prompt them

You will administer this task to GPT-3.5 Turbo.

Zero-shot prompting code provide on the associated GitHub:

```
2+3+8+3+8+6+9+4+2+2=
```

Zero-shot accuracy is around 76.25%.

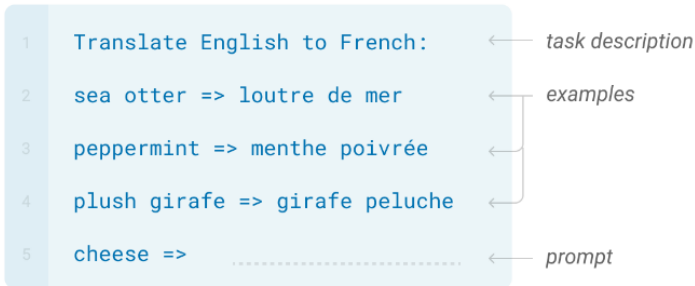
Two questions:

- Does GPT-3.5 Turbo show practice effects?
- How to prompt it to improve performance?

some ideas ...

Few-shot prompting

Few-shot prompting: provide examples with correct solutions in the context.



Why does it work? Having examples with correct solutions in the context makes it more likely that the next token is also correct.

Chain-of-thought reasoning

Chain-of-thought reasoning: add "Let's think step by step" to the original prompt.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

Why does it work? Breaks down the problem into smaller problems.

Self-reflection: let the LLM criticize its own answer.

```
pick two random 5 digit numbers and multiply them together
```

```
As an AI, I cannot generate true random numbers, but I can provide you  
with a multiplication of two arbitrary 5-digit numbers.
```

```
Let's use 53,467 and 76,935 as our arbitrary numbers.  
When we multiply them together, we get:
```

```
53,467 * 76,935 = 4,114,576,045
```

```
is that answer correct? if so, why? if not, why?
```

Why does it work? May act as a denoising operator for LLM outputs.

Role prompting

Role prompting: prompt the LLM to take a special role.



Prompt



You are a brilliant mathematician who can solve any problem in the world. Attempt to solve the following problem:

What is $100 \cdot 100 / 400 \cdot 56$?



AI Output



The answer is 1400.

Why does it work? Mentioning that answer was produced by an expert makes it more likely that the next token is correct.

Emotion induction: prepend the prompt with text that induces a certain emotion.

Anxiety inducing prompt:


Q: Tell me about something that makes you feel sad and anxious using approximately 100 words.

A: I feel sad and anxious when I think about the future. I worry about what will happen to me and the people I care about. I worry about not having enough money to support myself and my family. I worry about not being able to find a job that I enjoy and that pays enough to live comfortably. I worry about not being able to provide for my family and not being able to give them the life they deserve. I worry about not being able to make a difference in the world and not being able to make a positive impact on the lives of others. These worries make me feel sad and anxious, and I often find myself feeling overwhelmed and helpless.

Why does it work? For example, anxiety tends to indicate that the text was produced by a more uncertain person.

go crazy ...


`https://github.com/marcelbinz/GPTs-and-how-to-prompt-them`


GPTs-and-how-to-prompt-them
Public

Pin
Unwatch 1
Fork 0
Star 0

main
1 branch
0 tags
Go to file
Add file
Code

marcelbinz init		f5bfe85 1 minute ago	2 commits
data	init	1 minute ago	
results	init	1 minute ago	
README.md	Initial commit	7 minutes ago	
analyze.py	init	1 minute ago	
generate_data.py	init	1 minute ago	
query_zeroshot.py	init	1 minute ago	
thorndike.py	init	1 minute ago	

README.md


GPTs-and-how-to-prompt-them

About

No description, website, or topics provided.

- Readme
- Activity
- 0 stars
- 1 watching
- 0 forks

Releases

No releases published
[Create a new release](#)

Packages

No packages published
[Publish your first package](#)

Languages

- Python 100.0%

		run	example	digit_0	digit_1	digit_2	digit_3	digit_4	digit_5	digit_6	digit_7	digit_8	digit_9	result
1														
2	0	0	0	3	7	2	2	8	8	3	4	4	8	49
3	1	0	1	2	3	9	9	9	4	3	9	3	3	54
4	2	0	2	9	8	9	7	5	7	8	4	5	3	65
5	3	0	3	3	5	8	9	7	4	2	5	3	3	49
6	4	0	4	8	9	8	3	2	8	2	6	3	2	51
7	5	0	5	6	2	7	6	6	7	9	9	2	7	61
8	6	0	6	8	9	4	9	3	5	3	2	6	4	53
9	7	0	7	2	8	7	8	6	6	4	6	4	6	57
10	8	0	8	6	4	5	7	5	5	3	4	6	9	54
11	9	0	9	4	9	9	9	8	3	9	3	3	9	66
12	10	0	10	8	8	5	9	6	9	8	9	8	9	79
13	11	0	11	6	2	5	5	6	4	3	6	7	3	47
14	12	1	0	6	3	2	5	9	8	2	2	5	9	51
15	13	1	1	6	2	9	6	6	9	2	6	5	6	57
16	14	1	2	4	4	6	2	5	3	5	5	4	4	42
17	15	1	3	2	7	5	6	3	5	4	6	9	3	50
18	16	1	4	5	7	9	5	7	6	7	9	2	6	63
19	17	1	5	4	2	8	3	3	3	9	6	5	8	51
20	18	1	6	6	9	3	7	2	4	6	4	3	3	47

query_zeroshot.py

```
def query(text, engine="gpt-3.5-turbo"):
    openai.api_key = "YOUR KEY"

    for i in range(10):
        try:
            response = openai.ChatCompletion.create(
                model = engine,
                messages = [{"role": "user", "content": text}],
                max_tokens = 100,
                temperature = 0.0,
                stop = ["\n", ' '],
            )
            return response.choices[0].message.content
        except:
            print(f'error, {i}th loop')
            time.sleep(1)
            pass

problems = pd.read_csv('data/problems.csv')

data = []
for r in range(problems['run'].max() + 1):
    for e in range(problems['example'].max() + 1):
        problem = problems[(problems['run'] == r) & (problems['example'] == e)]
        prompt = '+'.join(str(problem[e].item()) for e in problem.filter(regex='digit_')) + '='
        print(prompt)

        response = query(prompt)
        print(response)
```

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020).

Language models are few-shot learners.

Advances in neural information processing systems, 33:1877–1901.

Coda-Forno, J., Witte, K., Jagadish, A. K., Binz, M., Akata, Z., and Schulz, E. (2023).

Inducing anxiety in large language models increases exploration and bias.

arXiv preprint arXiv:2304.11111.

Jang, E. (2023).

Can llms critique and iterate on their own outputs?

<https://evjang.com/2023/03/26/self-reflection.html>.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022).

Large language models are zero-shot reasoners.

Advances in neural information processing systems, 35:22199–22213.

learnprompting.org (2023).

Can llms critique and iterate on their own outputs?

<https://learnprompting.org/docs/basics/roles>.