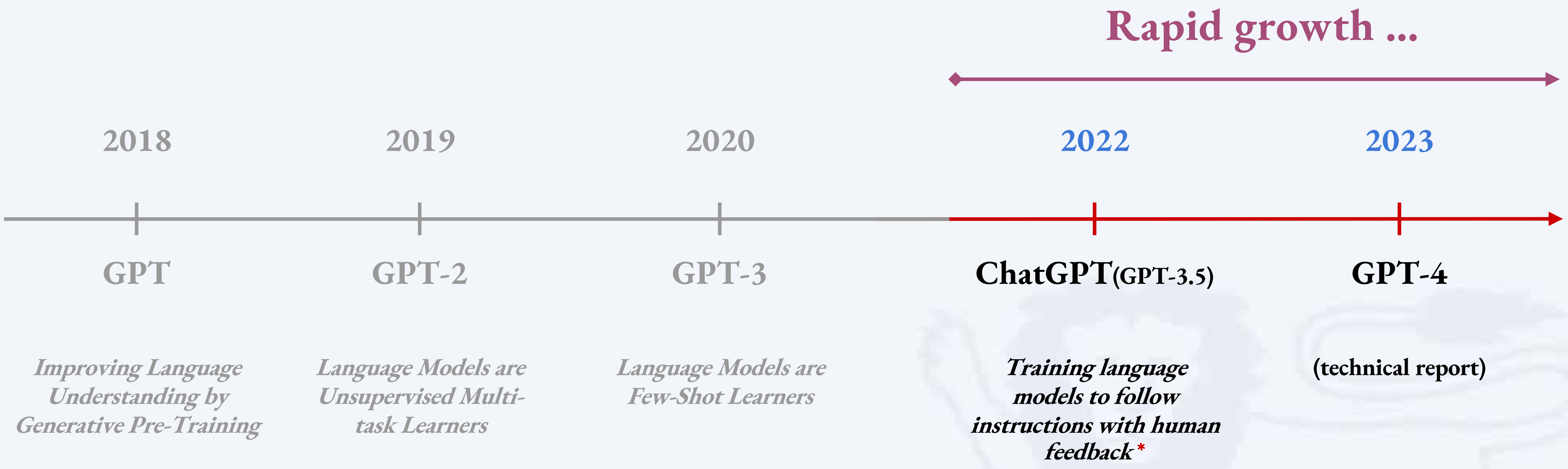
Human in the Loop

Rishabh Anand

@rishabh16_

The GPT Series

GPT → Generative Pretrained Transformer



Rapid growth …

2018 — **GPT** — *Improving Language Understanding by Generative Pre-Training*

2019 — **GPT-2** — *Language Models are Unsupervised Multi-task Learners*

2020 — **GPT-3** — *Language Models are Few-Shot Learners*

2022 — **ChatGPT**(GPT-3.5) — ***Training language models to follow instructions with human feedback****

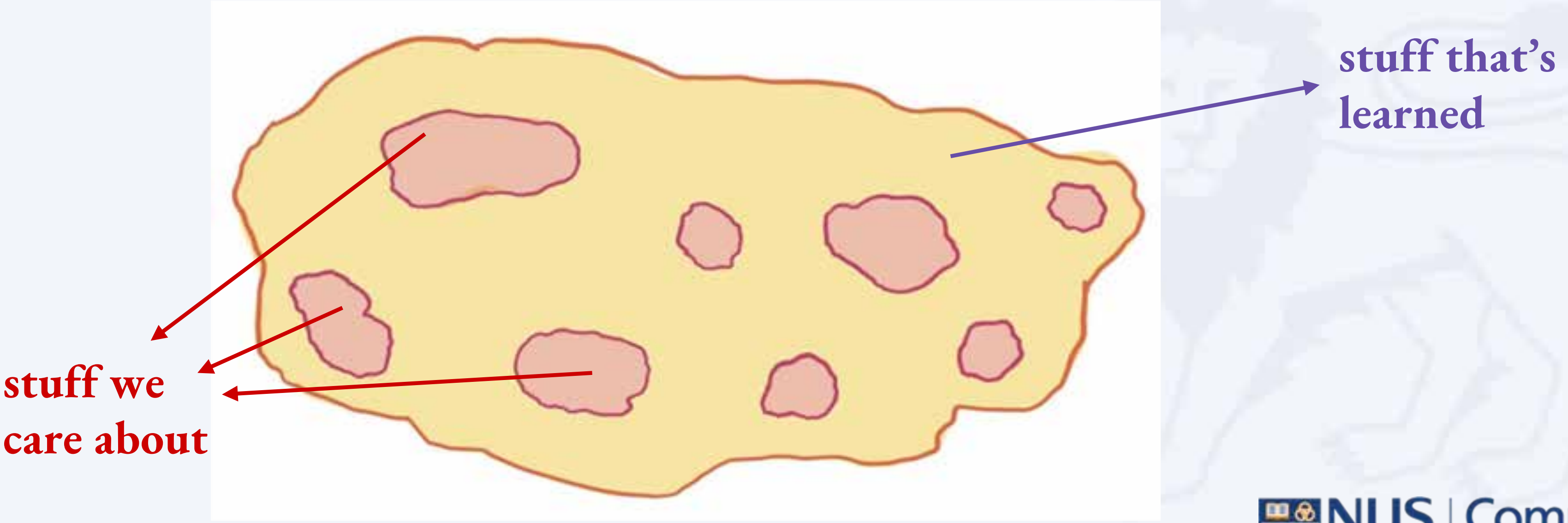
2023 — **GPT-4** — (technical report)

* GPT-3.5 is built on top of *InstructGPT* with a different data collection setup

Reinforcement Learning from Human Feedback

(Large) Language Models

- Language Models (like GPT-X),
    - are chaotic
    - model a "giant mass of people" ~ Minqi Jiang, MetaAI
- For different prompts, you can get wildly different outputs
- We must "snip out" the ugly, less-preferred parts



RL from Human Feedback

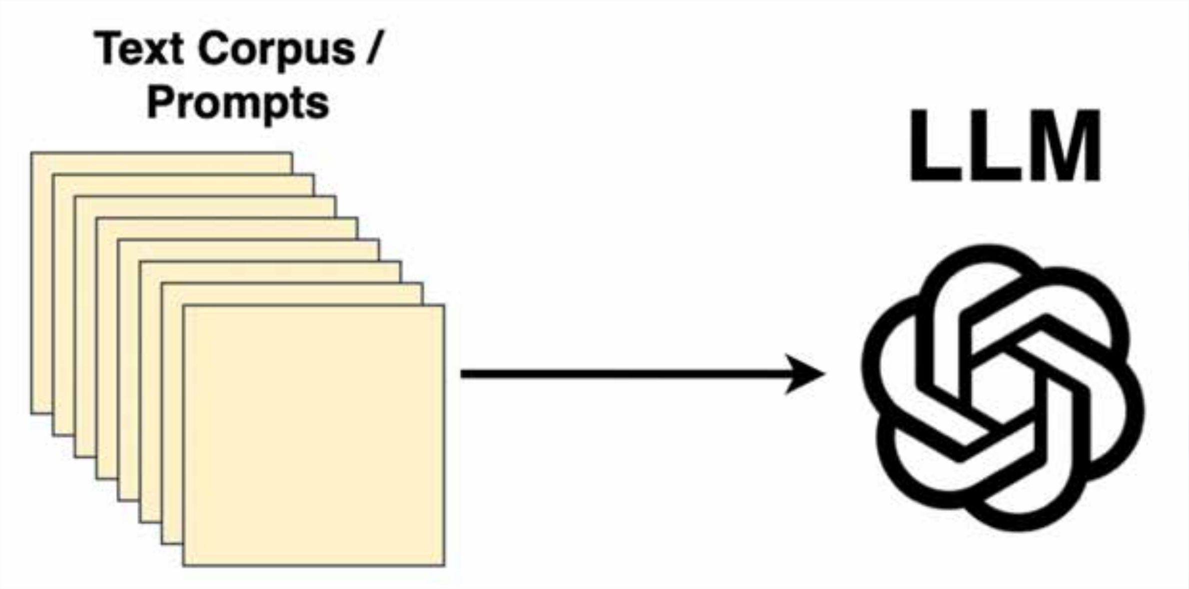
- Provides a friendlier interface to interact with LMs
- Biases the underlying model to generate human-aligned content
- Improves reliability, honesty, and safety of LLMs

"How do we get LLMs to sound more *human*?"

RL from Human Feedback

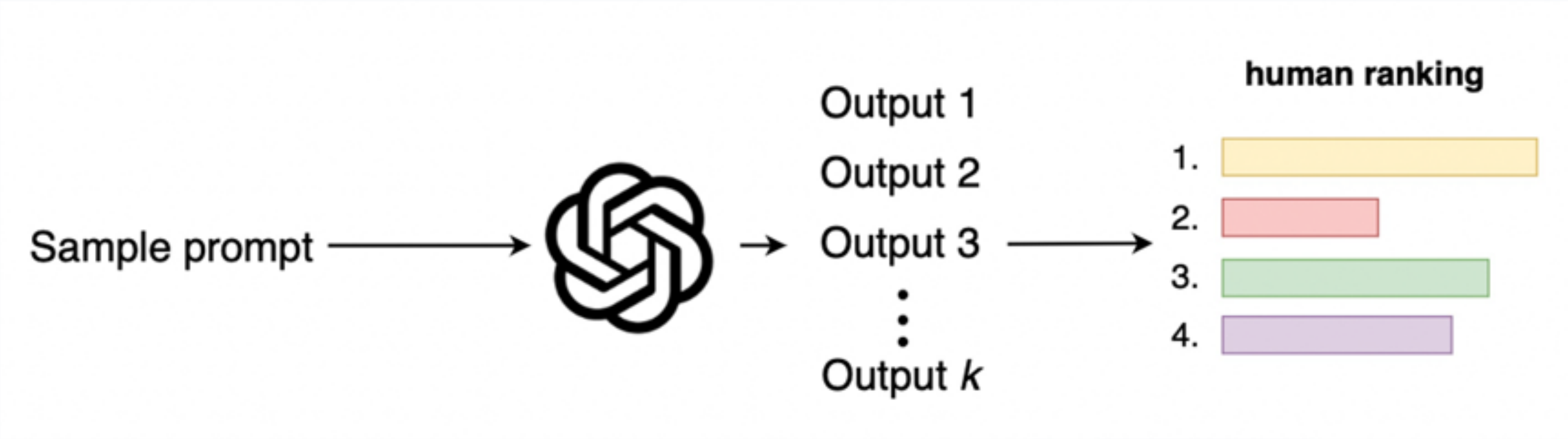
RL from Human Feedback

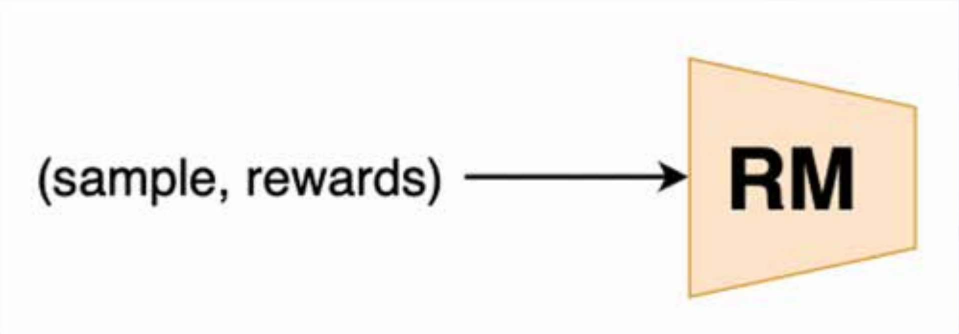
1. **Pretrain a LLM on a body of text [GPT-X, for instance]**



RL from Human Feedback

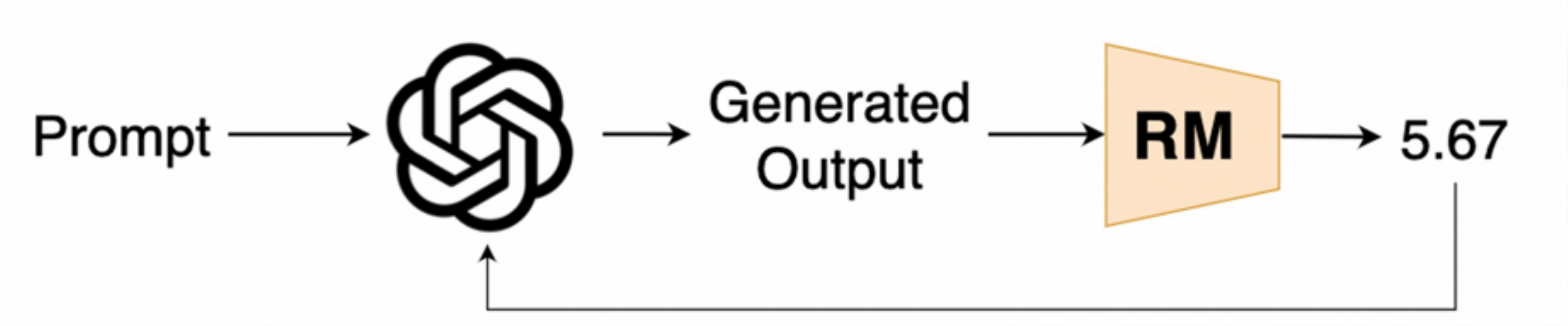
1. **Pretrain a LLM on a body of text [GPT-X, for instance]**
2. **Train a Reward Model (RM) → "how would a human feel?"**



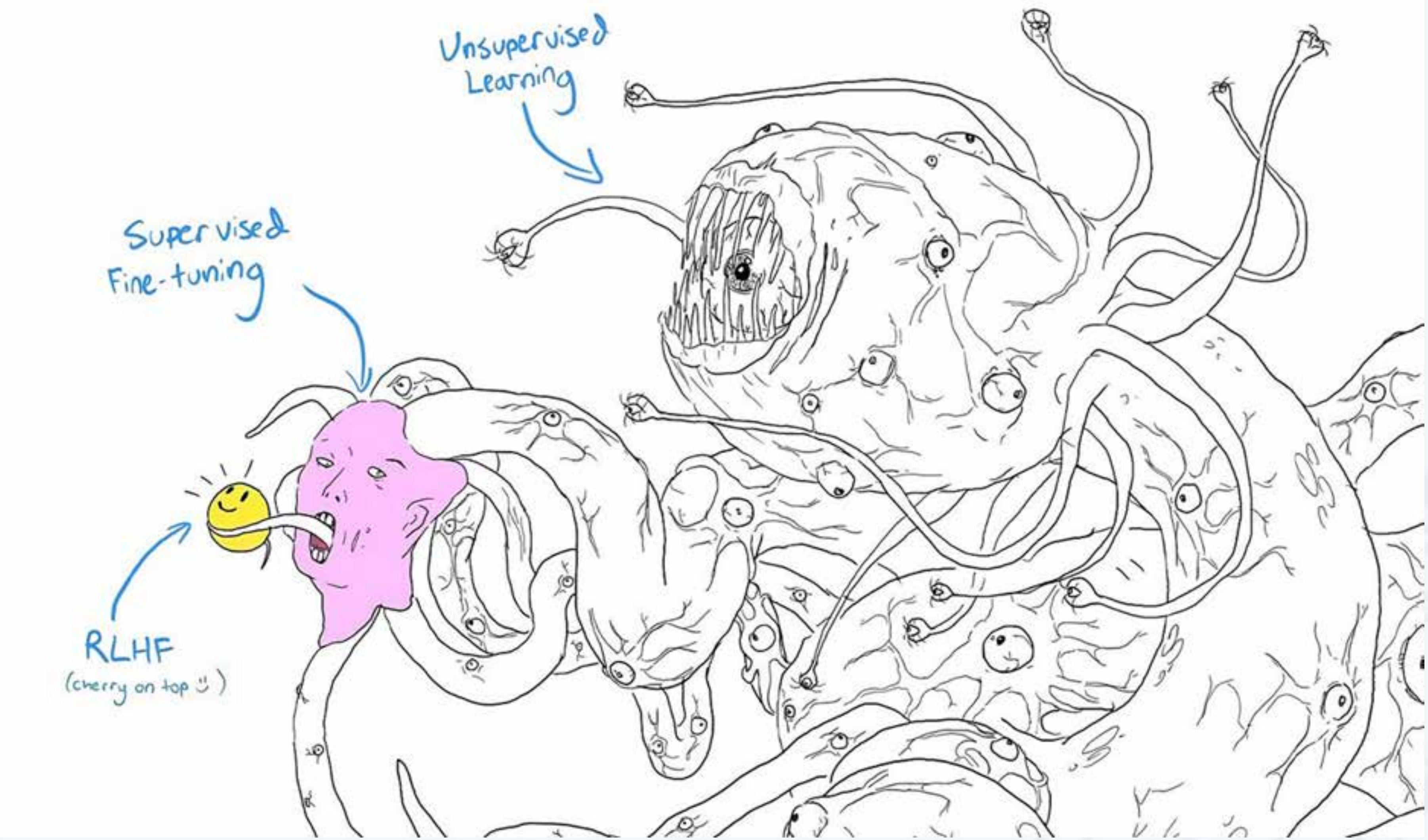


RL from Human Feedback

1. Pretrain a LLM on a body of text [GPT-X, for instance]
2. Train a Reward Model (RM) → "how would a human feel?"
3. Finetune using RL [LLM agent predicts words and is scored]



LLMs + RLHF



[source]

ChatGPT for Students

Ask Away!

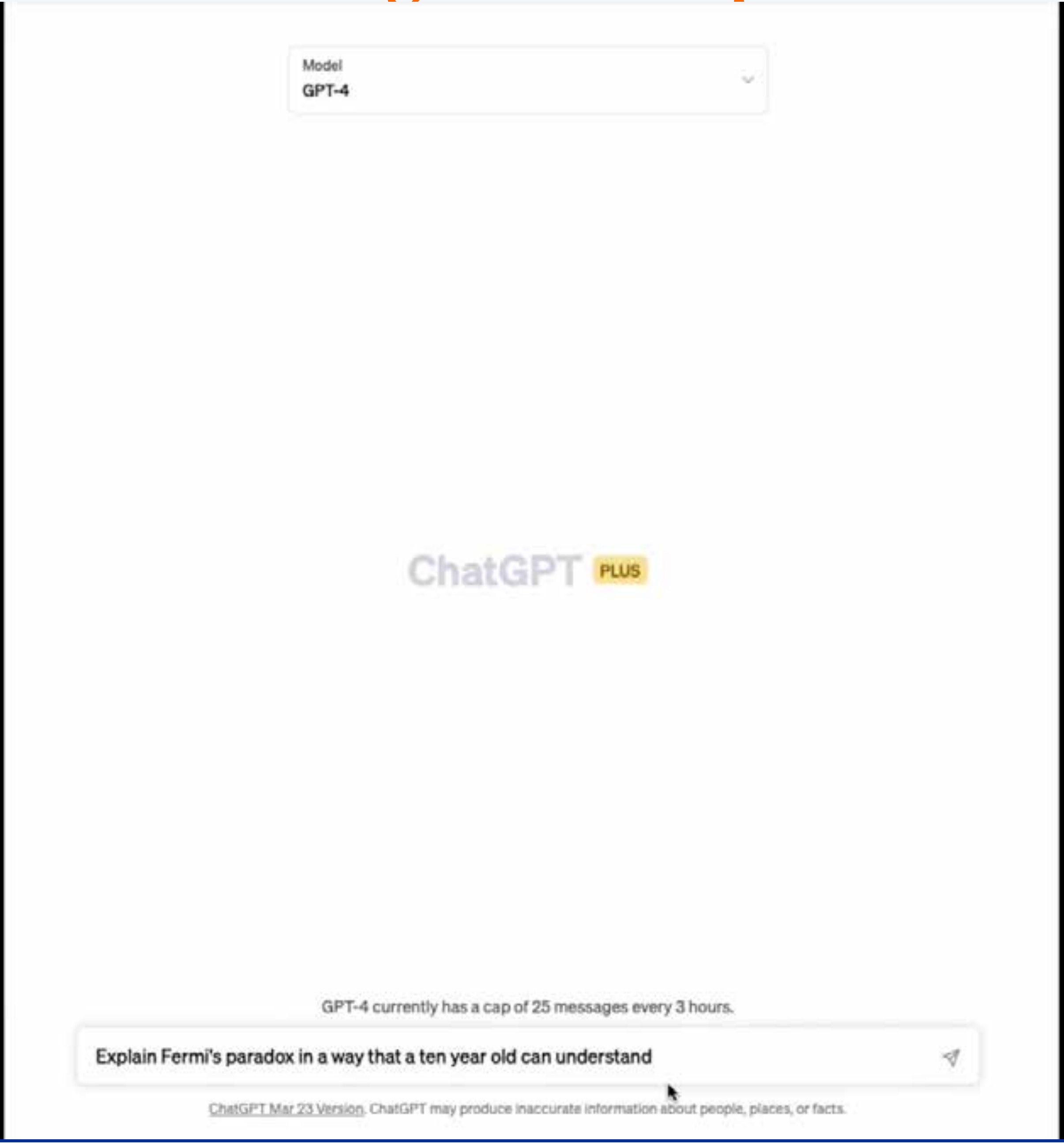
- Treat ChatGPT as you would a friend
- Want something? Just ask for it!
- The art of "Prompt Engineering" with ChatGPT

Use ChatGPT as a personal tutor!

The Possibilities

- Digestible explanations
- Summarising Long-form content
- Peer Review + feedback

Generate Digestible Explanations



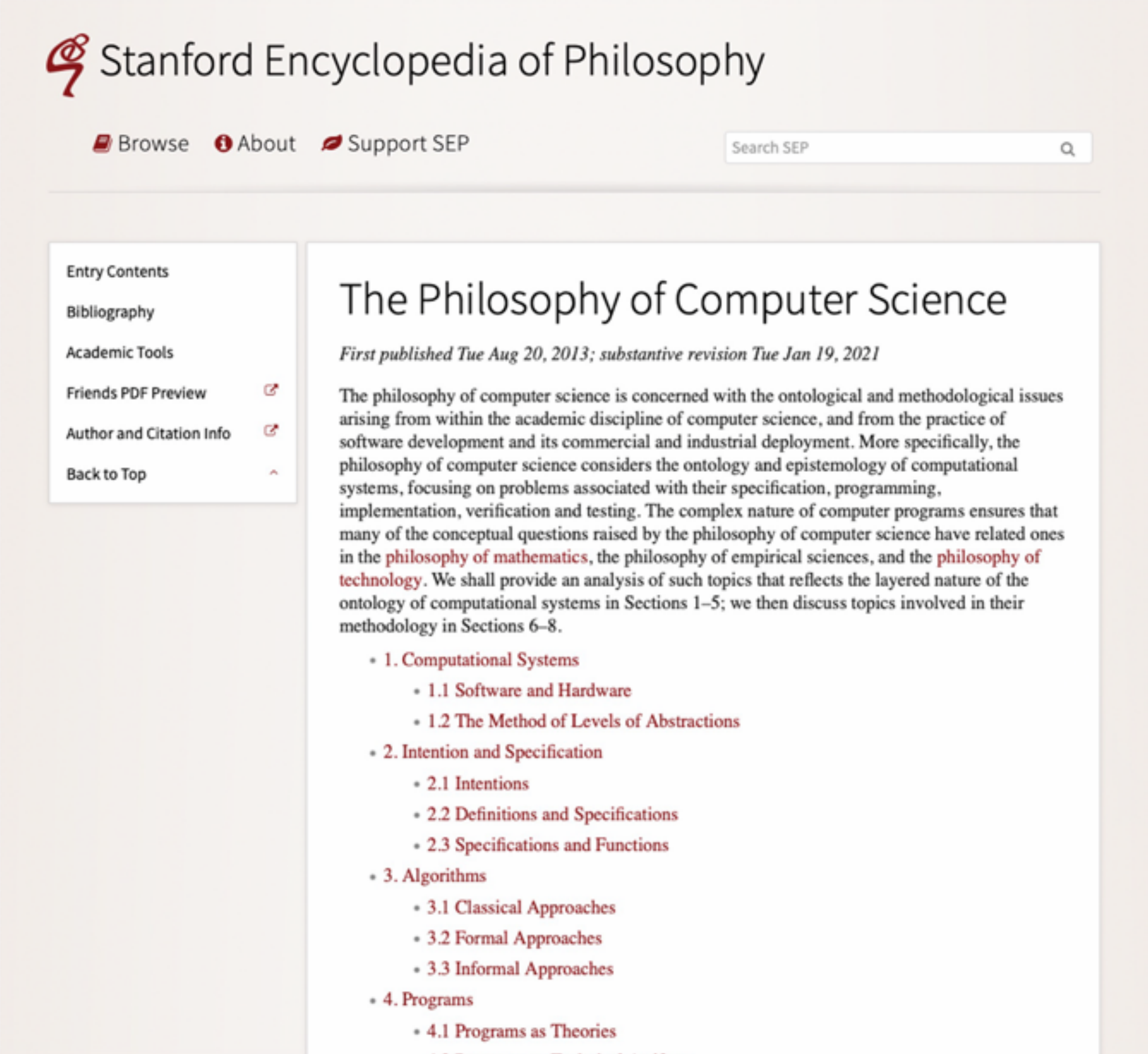
Model
GPT-4

GPT-4 currently has a cap of 25 messages every 3 hours.

Explain Fermi's paradox in a way that a ten year old can understand

ChatGPT Mar 23 Version. ChatGPT may produce inaccurate information about people, places, or facts.

Summarising Content



The Philosophy of Computer Science

First published Tue Aug 20, 2013; substantive revision Tue Jan 19, 2021

The philosophy of computer science is concerned with the ontological and methodological issues arising from within the academic discipline of computer science, and from the practice of software development and its commercial and industrial deployment. More specifically, the philosophy of computer science considers the ontology and epistemology of computational systems, focusing on problems associated with their specification, programming, implementation, verification and testing. The complex nature of computer programs ensures that many of the conceptual questions raised by the philosophy of computer science have related ones in the philosophy of mathematics, the philosophy of empirical sciences, and the philosophy of technology. We shall provide an analysis of such topics that reflects the layered nature of the ontology of computational systems in Sections 1–5; we then discuss topics involved in their methodology in Sections 6–8.

- 1. Computational Systems
  - 1.1 Software and Hardware
  - 1.2 The Method of Levels of Abstractions
- 2. Intention and Specification
  - 2.1 Intentions
  - 2.2 Definitions and Specifications
  - 2.3 Specifications and Functions
- 3. Algorithms
  - 3.1 Classical Approaches
  - 3.2 Formal Approaches
  - 3.3 Informal Approaches
- 4. Programs
  - 4.1 Programs as Theories
  - 4.2 Programs as Technical Artifacts

Given some long-form content that contains a lot to go through …

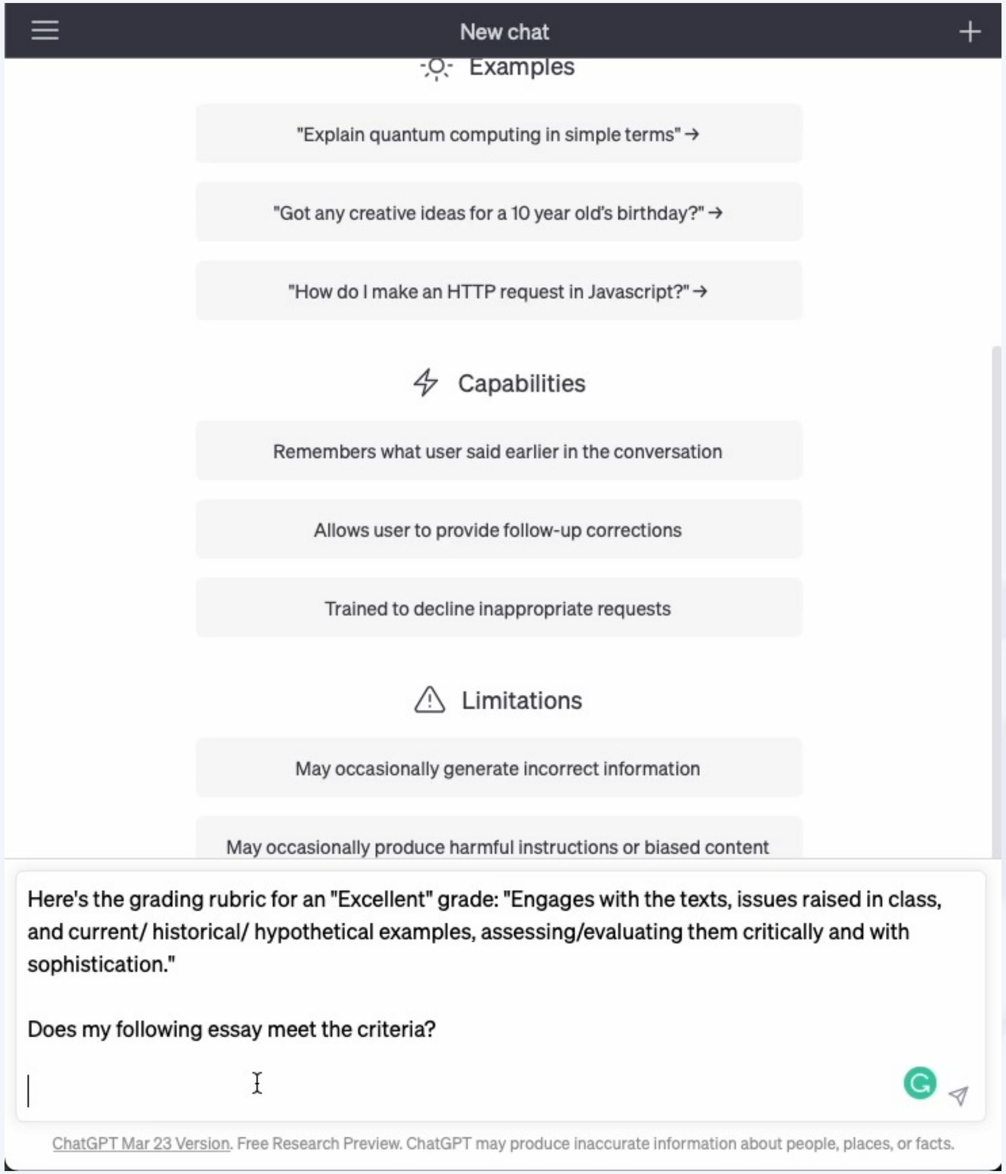
Summarising Content

Model
GPT-4

GPT-4 currently has a cap of 25 messages every 3 hours.

ChatGPT Mar 23 Version. ChatGPT may produce inaccurate information about people, places, or facts.

Peer Review + Feedback



Examples

"Explain quantum computing in simple terms" →

"Got any creative ideas for a 10 year old's birthday?" →

"How do I make an HTTP request in Javascript?" →

Capabilities

Remembers what user said earlier in the conversation

Allows user to provide follow-up corrections

Trained to decline inappropriate requests

Limitations

May occasionally generate incorrect information

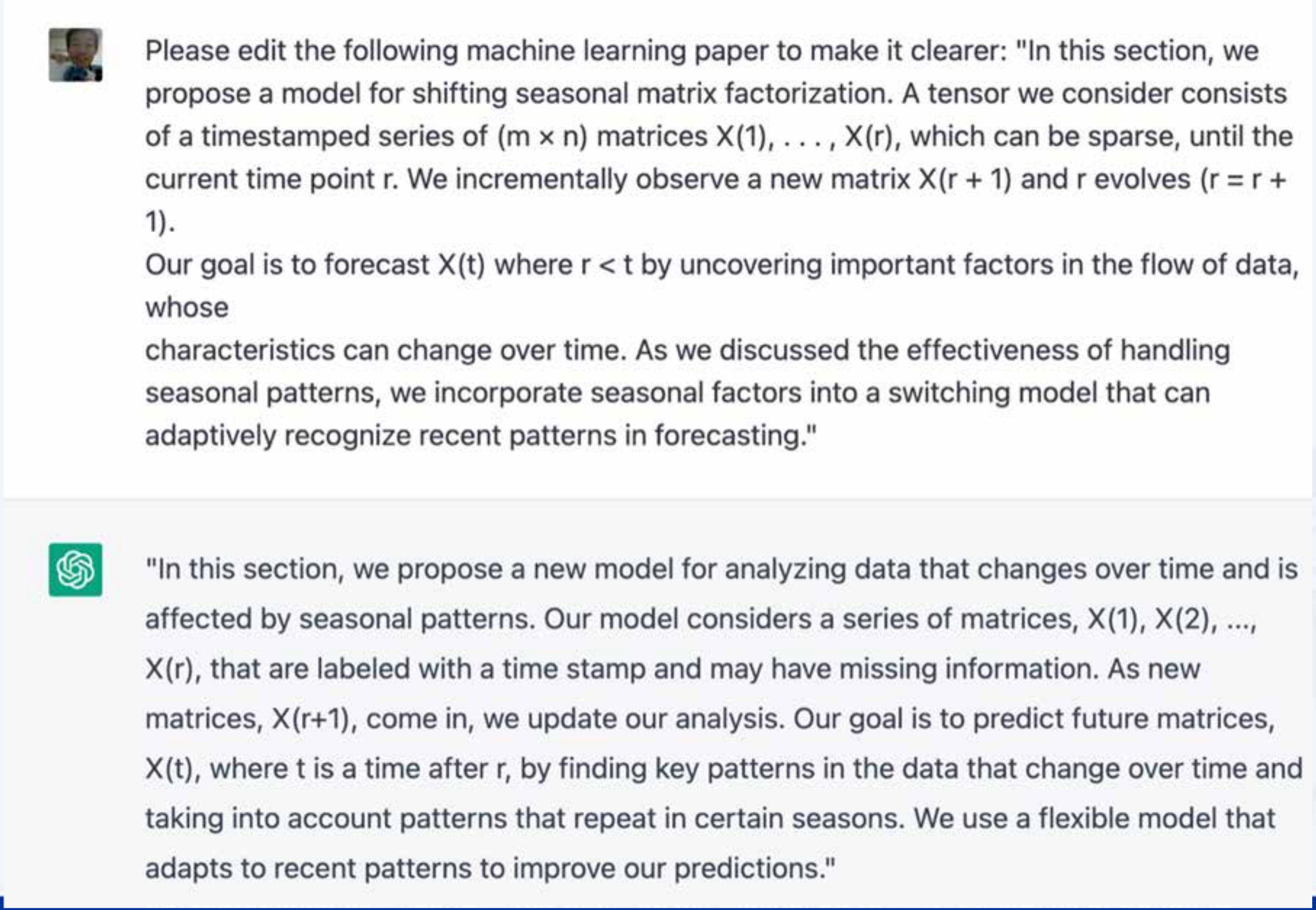
May occasionally produce harmful instructions or biased content

Here's the grading rubric for an "Excellent" grade: "Engages with the texts, issues raised in class, and current/ historical/ hypothetical examples, assessing/evaluating them critically and with sophistication."

Does my following essay meet the criteria?

ChatGPT Mar 23 Version. Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts.

Peer Review + Feedback



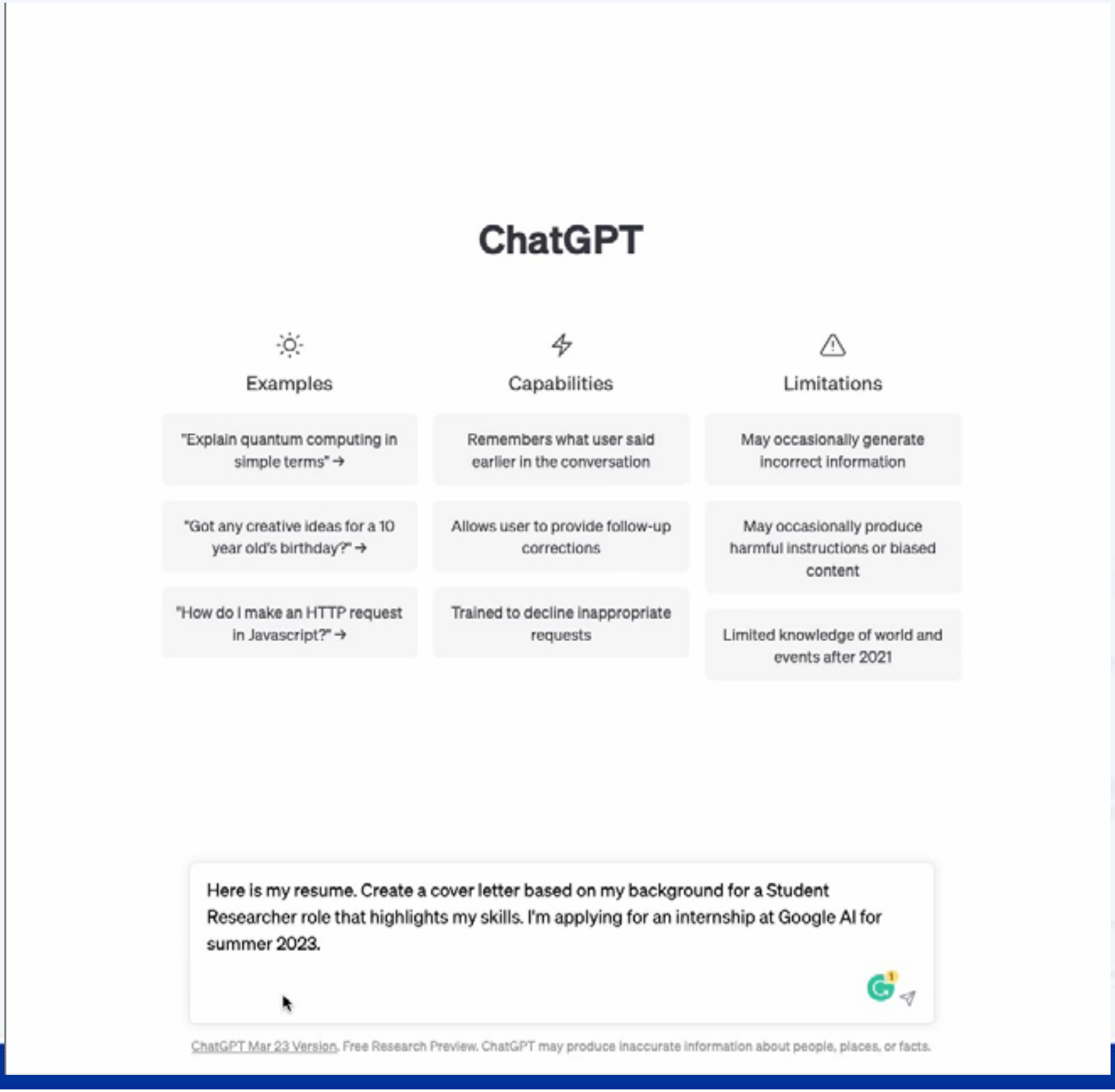
Please edit the following machine learning paper to make it clearer: "In this section, we propose a model for shifting seasonal matrix factorization. A tensor we consider consists of a timestamped series of (m × n) matrices X(1), . . . , X(r), which can be sparse, until the current time point r. We incrementally observe a new matrix X(r + 1) and r evolves (r = r + 1).

Our goal is to forecast X(t) where r < t by uncovering important factors in the flow of data, whose

characteristics can change over time. As we discussed the effectiveness of handling seasonal patterns, we incorporate seasonal factors into a switching model that can adaptively recognize recent patterns in forecasting."

"In this section, we propose a new model for analyzing data that changes over time and is affected by seasonal patterns. Our model considers a series of matrices, X(1), X(2), ..., X(r), that are labeled with a time stamp and may have missing information. As new matrices, X(r+1), come in, we update our analysis. Our goal is to predict future matrices, X(t), where t is a time after r, by finding key patterns in the data that change over time and taking into account patterns that repeat in certain seasons. We use a flexible model that adapts to recent patterns to improve our predictions."

Peer Review + Feedback



ChatGPT

Examples	Capabilities	Limitations
"Explain quantum computing in simple terms" →	Remembers what user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?" →	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?" →	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021

Here is my resume. Create a cover letter based on my background for a Student Researcher role that highlights my skills. I'm applying for an internship at Google AI for summer 2023.

ChatGPT Mar 23 Version. Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts.

ChatGPT for Students

- LLM technology will only get better from here on
- Students ~~should~~ *can* learn how to operate these tools
- While LLMs can improve productivity, it’s not the be-all-end-all

AI tools lower the activation energy to get started!!!

But … shortcomings?

Stay for our panels!