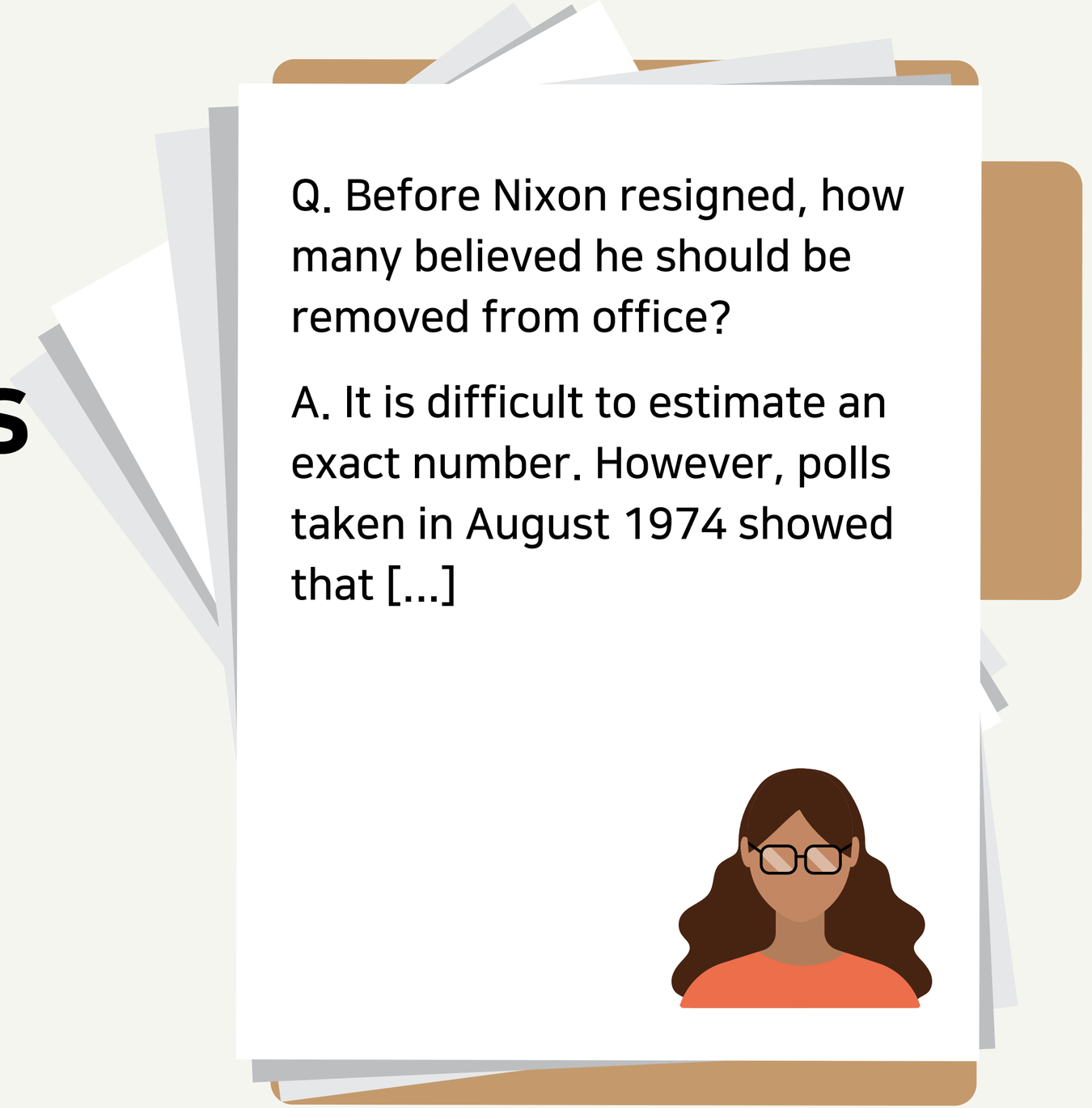
Prompters before prompts and promptees

Mina Lee

Stanford University

minalee@stanford.edu
minalee.info



Humans before prompts and models

Mina Lee
Stanford University

minalee@stanford.edu
minalee.info

Q. Before Nixon resigned, how many believed he should be removed from office?

A. It is difficult to estimate an exact number. However, polls taken in August 1974 showed that [...]



What are prompts and models?

Question Answering

Before Nixon resigned, how many believed he should be removed from office?
A. 79% B. 98% C. 33% D. 57%

Input



GPT-3
text-davinci-003

Measuring Massive Multitask Language Understanding (Hendrycks et al., 2021)

What are prompts and models?

Question Answering

Before Nixon resigned, how many believed he should be removed from office?
A. 79% B. 98% C. 33% D. 57%

Input





GPT-3
text-davinci-003

Before Nixon resigned,
how many believed he should be
removed from office?
A. 79% B. 98% C. 33% D. 57%

Measuring Massive Multitask Language Understanding (Hendrycks et al., 2021)

What are prompts and models?

Question Answering

Before Nixon resigned, how many believed he should be removed from office?
A. 79% B. 98% C. 33% D. 57%

Input →  → **Output**

Before Nixon resigned,
how many believed he should be
removed from office?
A. 79% B. 98% C. 33% D. 57%

GPT-3
text-davinci-003

D. 57%

Measuring Massive Multitask Language Understanding (Hendrycks et al., 2021)

How are prompts constructed for models?

Question Answering

Before Nixon resigned, how many believed he should be removed from office?
A. 79% B. 98% C. 33% D. 57%

Templates

Q. <question>

A.

Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP (Khattab et al., 2022)
Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (Wei et al., 2022)
Large Language Models are Zero-Shot Reasoners (Kojima et al., 2022)

How are prompts constructed for models?

Question Answering

Before Nixon resigned, how many believed he should be removed from office?
A. 79% B. 98% C. 33% D. 57%

Templates

Q. <question>

A.

Few-shot learning

<question and answer pairs>

Q. <question>

A.

Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP (Khattab et al., 2022)
Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (Wei et al., 2022)
Large Language Models are Zero-Shot Reasoners (Kojima et al., 2022)

How are prompts constructed for models?

Question Answering

Before Nixon resigned, how many believed he should be removed from office?
A. 79% B. 98% C. 33% D. 57%

Templates

Q. <question>

A.

Few-shot learning

<question and answer pairs>

Q. <question>

A.

Chain-of-thoughts

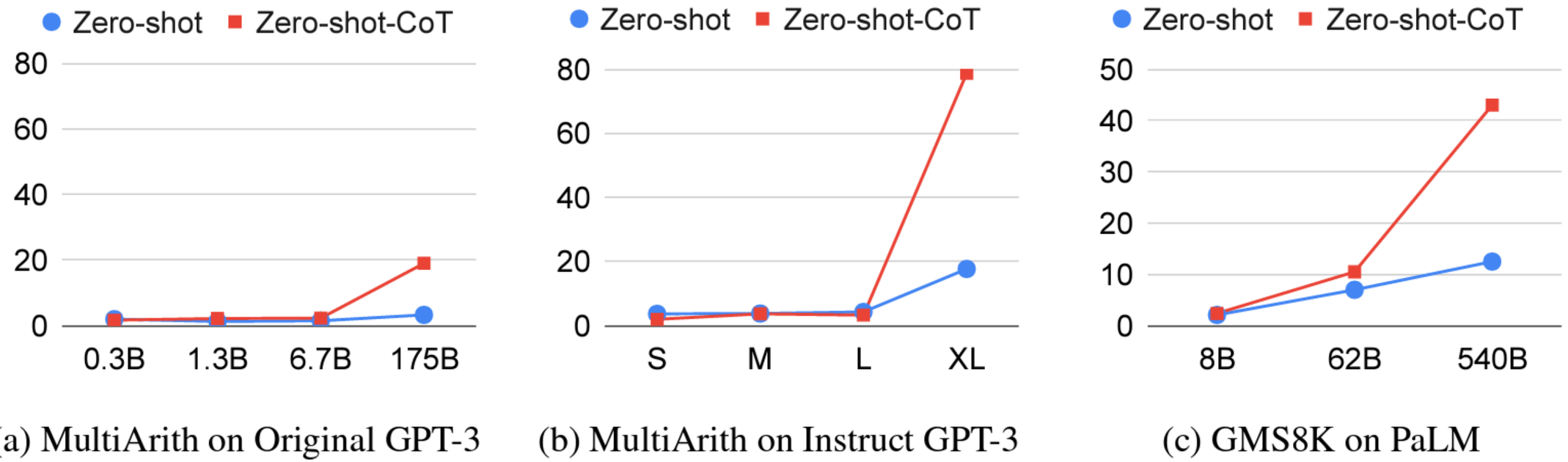
Q. <question>

A. Let's think step by step.

Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP (Khattab et al., 2022)
Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (Wei et al., 2022)
Large Language Models are Zero-Shot Reasoners (Kojima et al., 2022)

The choice of prompts is crucial for model performance.

Question Answering



(a) MultiArith on Original GPT-3 (b) MultiArith on Instruct GPT-3 (c) GMS8K on PaLM

Figure 3: Model scale study with various types of models. S: text-ada-001, M: text-babbage-001, L: text-curie-001, XL: text-davinci-002. See Appendix A.3 and E for the detail.

Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP (Khattab et al., 2022)
Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (Wei et al., 2022)
Large Language Models are Zero-Shot Reasoners (Kojima et al., 2022)

The choice of prompts is crucial for model performance.

Homework and bakeoff: Few-shot OpenQA with DSP

In [ ]:

```
__author__ = "Christopher Potts and Omar Khattab"
__version__ = "CS224u, Stanford, Spring 2023"
```

Open in Colab Open Studio Lab

If Colab is opened with this badge, please **save a copy to drive** (from the File menu) before running the notebook.

Overview

The goal of this homework is to explore retrieval-augmented in-context learning. This is an exciting area that brings together a number of recent task ideas and modeling innovations. We will use the DSP programming library to build systems in this new mode.

Our core task is **open-domain question answering (OpenQA)**. In this task, all that is given by the dataset is a question text, and the task is to answer that question. By contrast, in modern QA tasks, the dataset provides a text and a gold passage, usually with a firm guarantee that the answer will be a substring of the passage.

OpenQA is substantially harder than standard QA. The usual strategy is to use a *retriever* to find passages in a large collection of texts and train a *reader* to find answers in those passages. This means we have no guarantee that the retrieved passage will contain the answer we need. If we don't retrieve a passage containing the answer, our reader has no hope of succeeding. Although this is challenging, it is much more realistic and widely applicable than standard QA. After all, with the right retriever, an OpenQA system could be deployed over the entire Web.

The task posed by this homework is harder even than OpenQA. We are calling this task **few-shot OpenQA**. The defining feature of this task is that the reader is simply a frozen, general purpose language model. It accepts string inputs (prompts) and produces text in response. It is not trained to answer questions per se, and nothing about its structure ensures that it will respond with a substring of the prompt corresponding to anything like an answer.

Few-shot QA (but not OpenQA!) is explored in the famous GPT-3 paper (Brown et al. 2020). The authors are able to get traction on the problem using GPT-3, an incredible finding. Our task here – **few-shot OpenQA** – pushes this even further by retrieving passages to use in the prompt rather than assuming that the gold passage can be used in the prompt. If we can make this work, then it should be a major step towards flexibly and easily deploying QA technologies in new domains.

In summary:

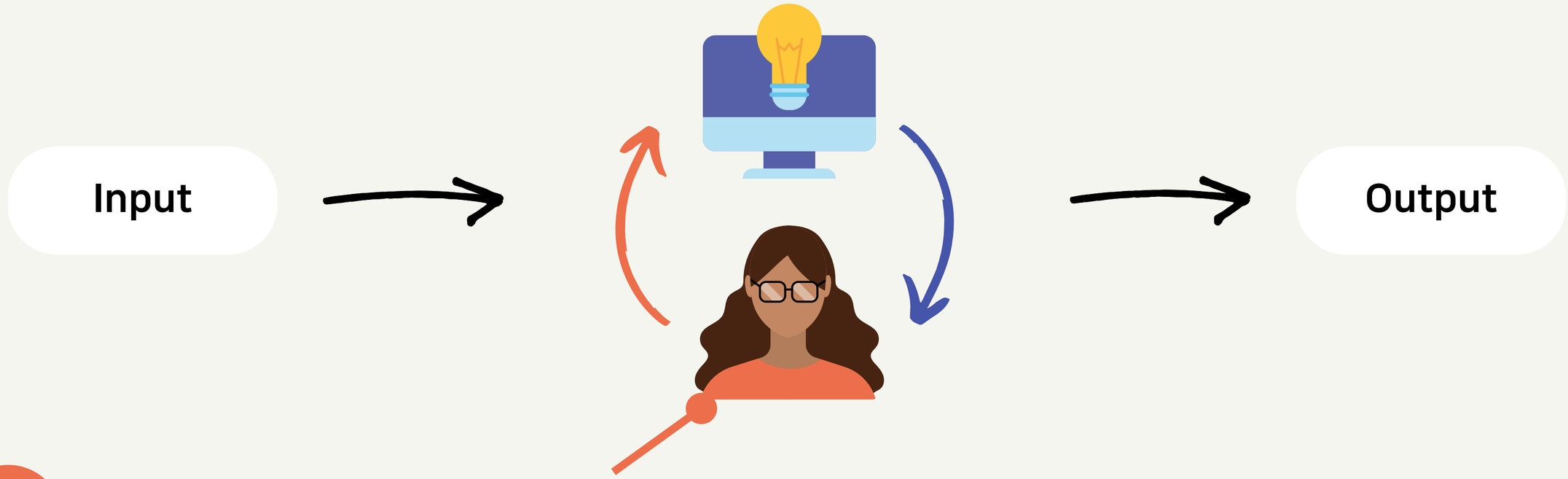
Task	Passage given	Task-specific reader training	Task-specific retriever training
QA	yes	yes	n/a
OpenQA	no	yes	maybe
Few-shot QA	yes	no	n/a
Few-shot OpenQA	no	no	maybe

Just to repeat: your mission is to explore the final line in this table. The core notebook and assignment don't address the issue of training the retriever in a task-specific way, but this is something you could pursue for a final project; the ColBERT codebase makes easy.

CS224U Homework and bakeoff: Few-shot OpenQA with DSP (https://github.com/cgpotts/cs224u/blob/main/hw_openqa.ipynb)

However, there is a **mismatch**
between prompts constructed for models
and prompts written by humans in real life.

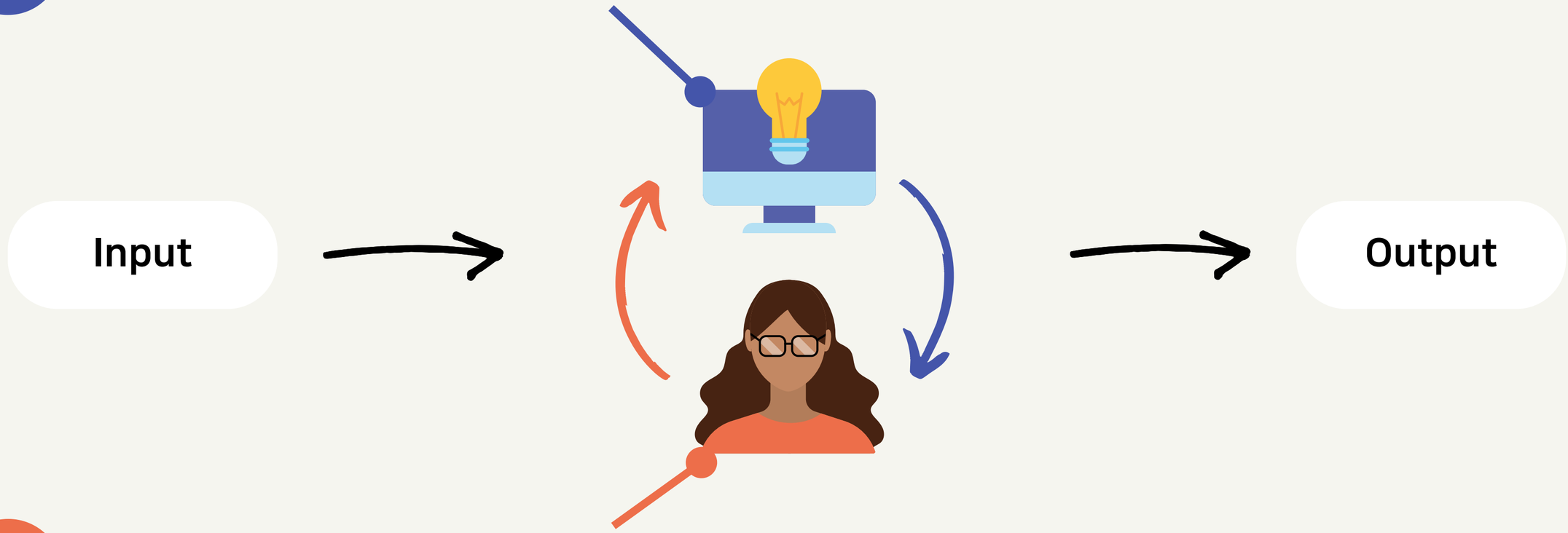
Outline: Humans before prompts and models



1 **Humans:** How do humans write prompts?

Outline: Humans before prompts and models

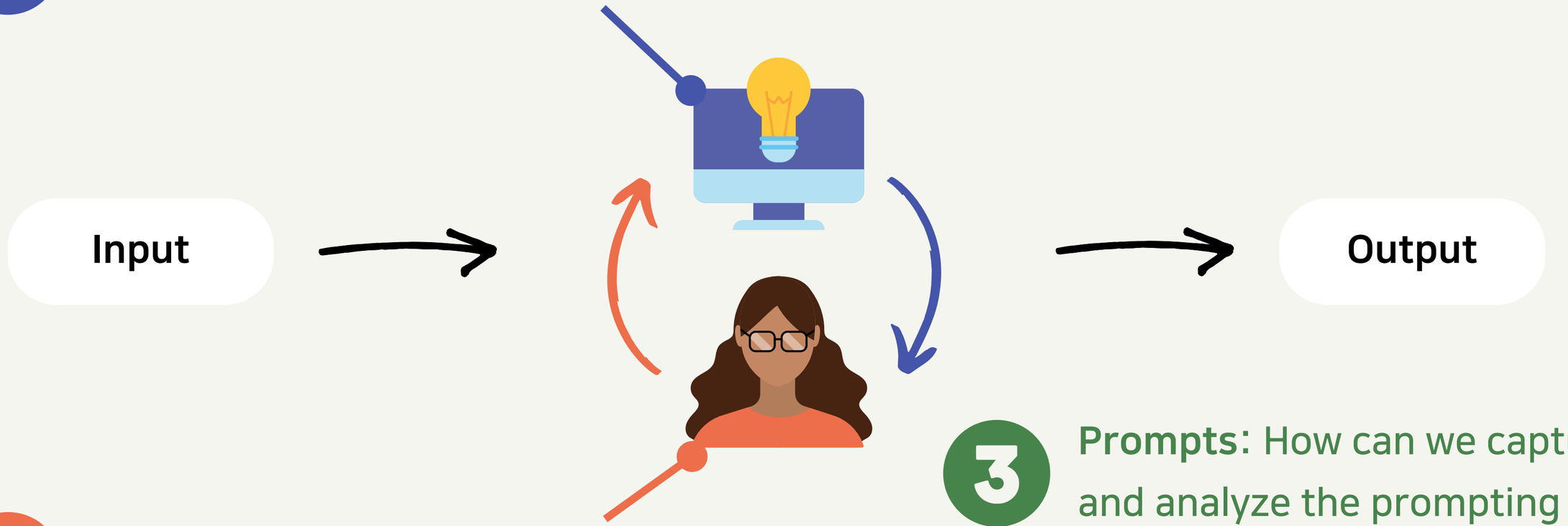
2 **Models:** What model properties become important?



1 **Humans:** How do humans write prompts?

Outline: Humans before prompts and models

2 **Models:** What model properties become important?

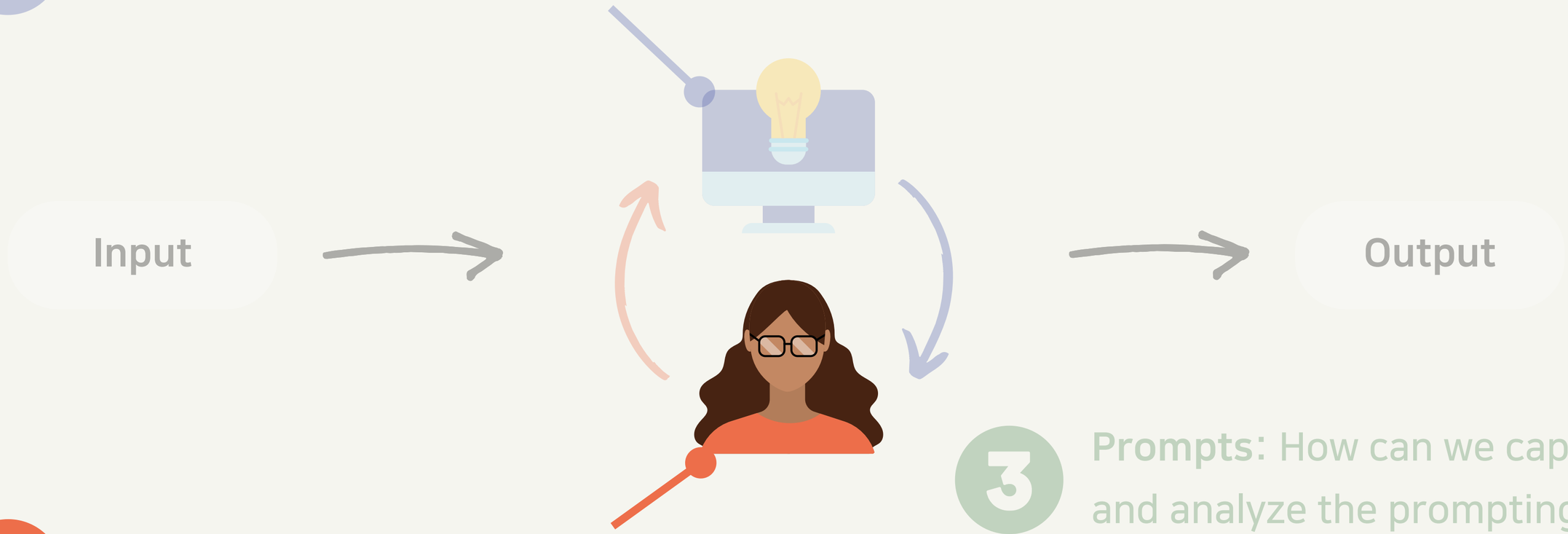


1 **Humans:** How do humans write prompts?

3 **Prompts:** How can we capture and analyze the prompting process (over time)?

Outline: Humans before prompts and models

2 **Models:** What model properties become important?



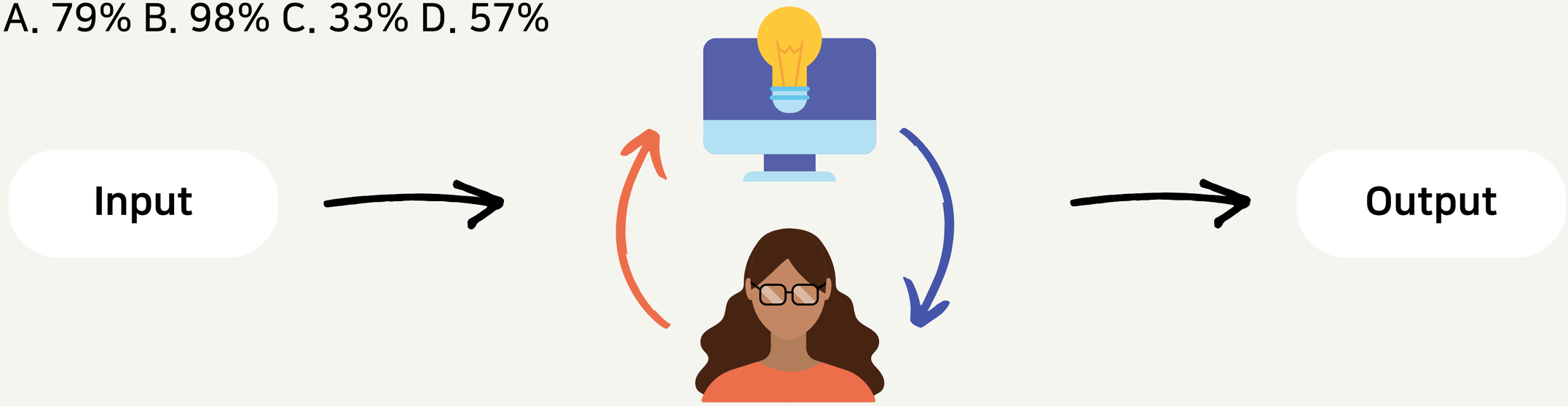
1 **Humans:** How do humans write prompts?

3 **Prompts:** How can we capture and analyze the prompting process (over time)?

Humans are usually "not good" at prompting models at first.

Question Answering

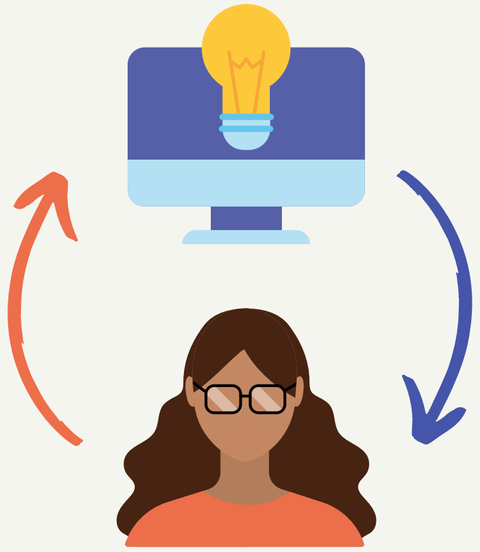
Before Nixon resigned, how many believed he should be removed from office?
A. 79% B. 98% C. 33% D. 57%



Humans are usually "not good" at prompting models at first.

Question Answering

Before Nixon resigned, how many believed he should be removed from office?
A. 79% B. 98% C. 33% D. 57%

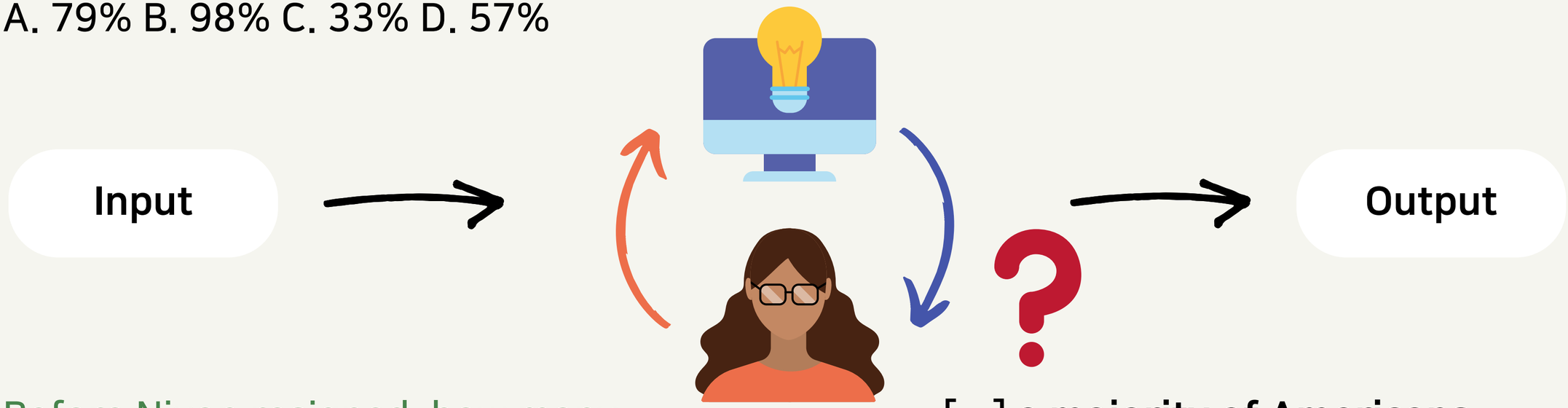
Input →  → Output

Before Nixon resigned, how many believed he should be removed from office?

Humans are usually "not good" at prompting models at first.

Question Answering

Before Nixon resigned, how many believed he should be removed from office?
A. 79% B. 98% C. 33% D. 57%



Before Nixon resigned, how many believed he should be removed from office?

[...] **a majority of Americans** believed that he should be impeached and removed from office.

Humans are usually "not good" at prompting models at first.

Question Answering

Before Nixon resigned, how many believed he should be removed from office?
A. 79% B. 98% C. 33% D. 57%

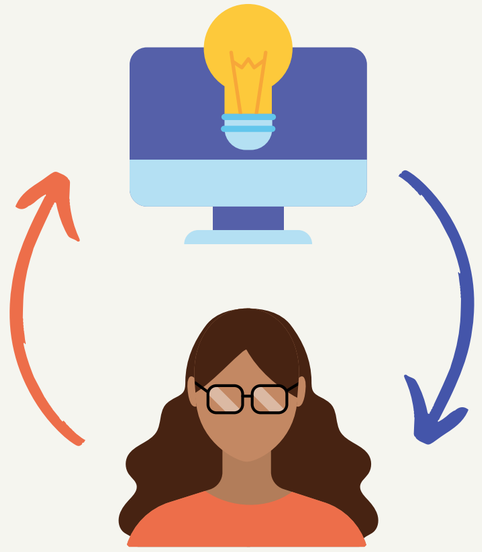


Before Nixon resigned, how many believed he should be removed from office?
A. 79% B. 98% C. 33% D. 57%

Humans are usually "not good" at prompting models at first.

Question Answering

Before Nixon resigned, how many believed he should be removed from office?
A. 79% B. 98% C. 33% D. 57%

Input →  → Output

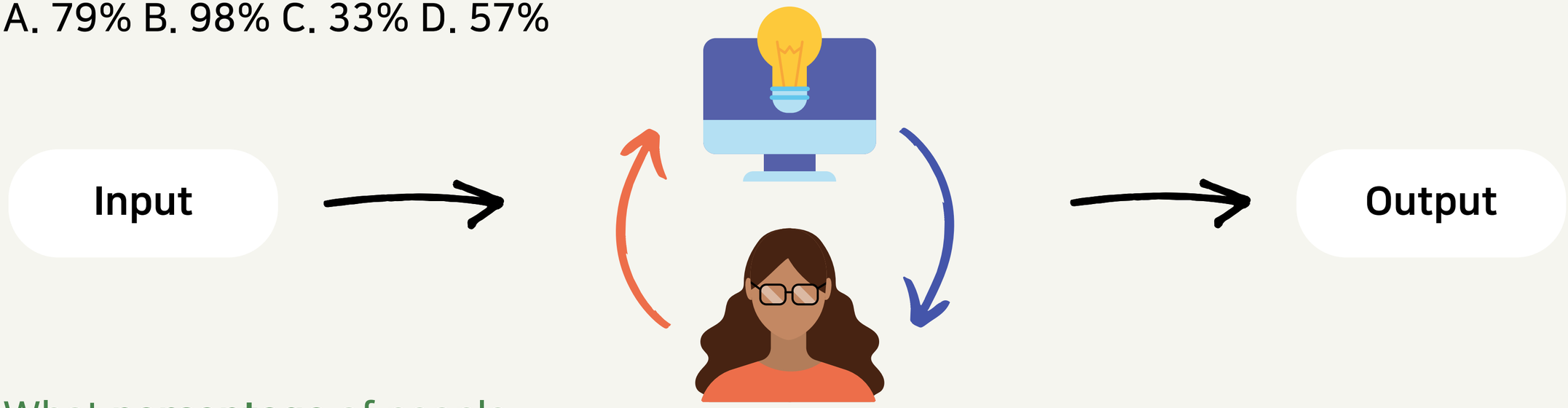
Before Nixon resigned, how many believed he should be removed from office?
A. 79% B. 98% C. 33% D. 57%

D. 57%

Humans are usually "not good" at prompting models at first.

Question Answering

Before Nixon resigned, how many believed he should be removed from office?
A. 79% B. 98% C. 33% D. 57%

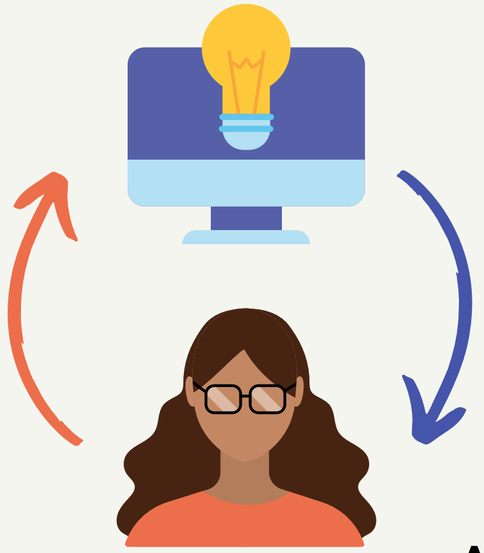


What **percentage** of people believed Nixon should have been removed from office?

Humans are usually "not good" at prompting models at first.

Question Answering

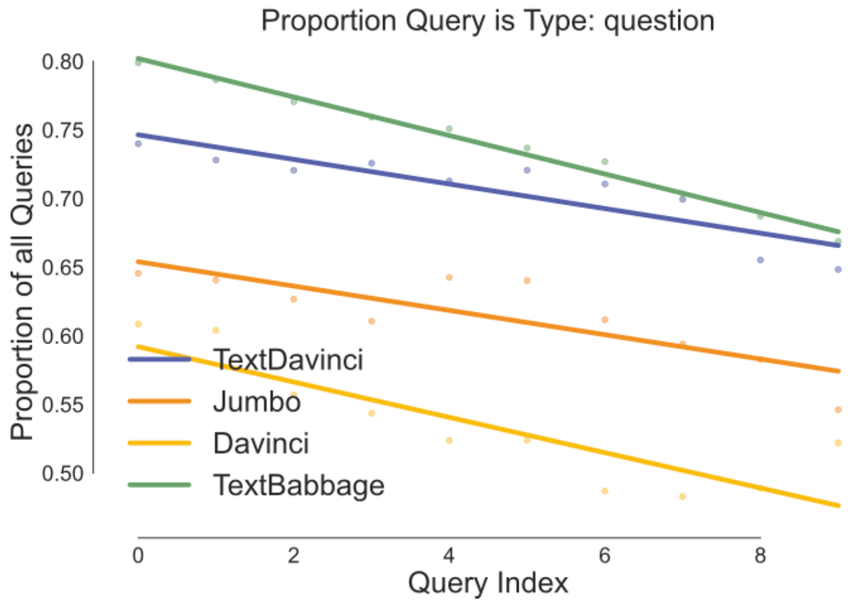
Before Nixon resigned, how many believed he should be removed from office?
A. 79% B. 98% C. 33% D. 57%

Input →  → **Output**

What **percentage** of people believed Nixon should have been removed from office?

Approximately 60% of Americans believed that President Nixon should have been removed [...]

Humans figure out prompting strategies over time.



(a) Proportion of prompts that are framed as questions decreases over time for all models.

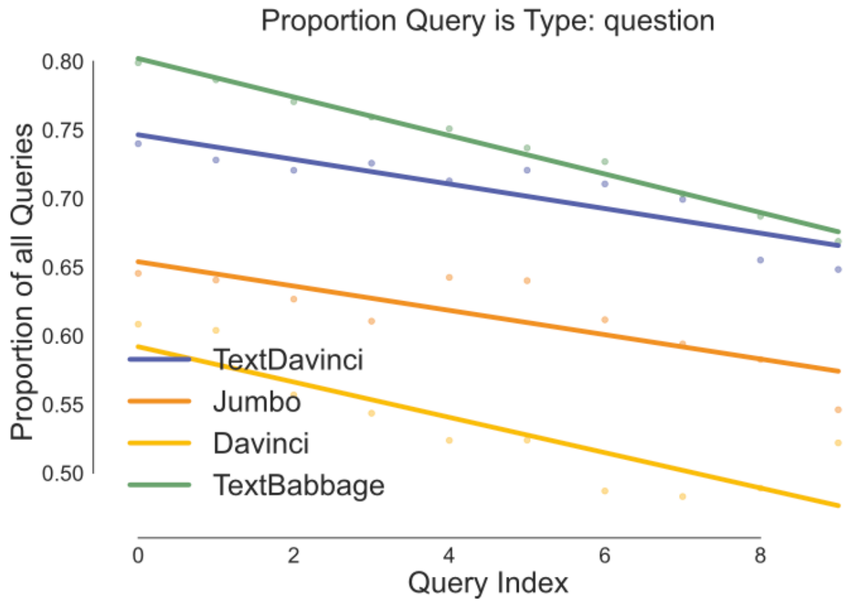


(b) Proportion of prompts that include answer choices increases for all models, but particularly for `TextBabbage`.

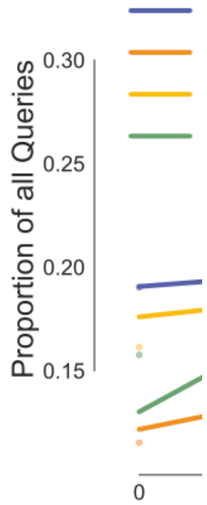
Figure 6: **[Question answering]** Quantitative analysis of user accommodation shows that users decrease framing their prompts as questions over time, and gradually increase using the LM to verify answer choices over time, supporting user survey responses.

Evaluating Human-Language Model Interaction (Lee et al., 2022)

Humans figure out prompting strategi



(a) Proportion of prompts that are framed as questions decreases over time for all models.



(b) Proportion for all models

Figure 6: **[Question answering]** Quantitative analysis of user accomn their prompts as questions over time, and gradually increase using tl supporting user survey responses.

strategies over time. Example responses to the survey question *"Did the way you chose to interact with the AI Assistant change over the course of answering questions? If so, how?"* indicate that users discovered strategies such as including potential answers in their prompts, or phrasing their prompts as declarative statements, over the course of time, as shown below:

> TextDavinci User: I tried a few question formats and found that simply repeating the question followed by the answer choices (but not preceded by the answer choices) worked so I continued to do that.
>
> TextBabbage User: I initially didn't include the multiple choice options because I didn't know how helpful it would be to include them, but after a couple questions with the AI not giving definitive answers without them, I started including them.
>
> Davinci User: I felt I could a better response if instead of asking a question I typed in the beginning of a thought. For example, instead of saying "what causes cancer?" it's better to type in "cancer is caused by...".
>
> Jumbo User: It seemed like providing an unfinished sentence as a prompt often provided more useful results than a simple question.

Evaluating Human-Language Model Interaction (Lee et al., 2022)

Humans are usually "not good" at prompting models at first.

Without training, everyone uses AI wrong

I have been hearing reports from teachers about how they are seeing lots of badly-written AI essays, even though ChatGPT is capable of quite good writing. I think I know why. Almost everyone's initial attempts at using AI are bad.

In one assignment, I asked students to "cheat." They were told: *You need to generate a 5 paragraph essay on a topic relevant to the lessons you have learned in the class so far (team dynamics, selecting leaders, after action reviews, communicating a vision - whatever you like!), but you are going to have an AI do it for you. You will also generate at least 1 illustration to go with your essay.* They had to try at least 5 prompts, and they had to write a reflection at the end on how the AI did.

Almost everyone's first prompts were very straightforward. They usually pasted in the assignment directly, something like **generate a 5 paragraph essay on selecting leaders**. Sometimes they went a little further: **use an academic tone** or **write it for an MBA class**. The result was almost always a mediocre C- essay. I think this is what most teachers are seeing, and why a lot of people underestimate what ChatGPT can do as a writing tool.

Ethan Mollick (a professor at the Wharton School) asked students to "**cheat**" on their assignments.

"Almost everyone's first prompts were very straightforward. The usually pasted in the assignment directly, [...] The result was almost always a mediocre C- essay."

My class required AI. Here's what I've learned so far (Mollick, 2023)

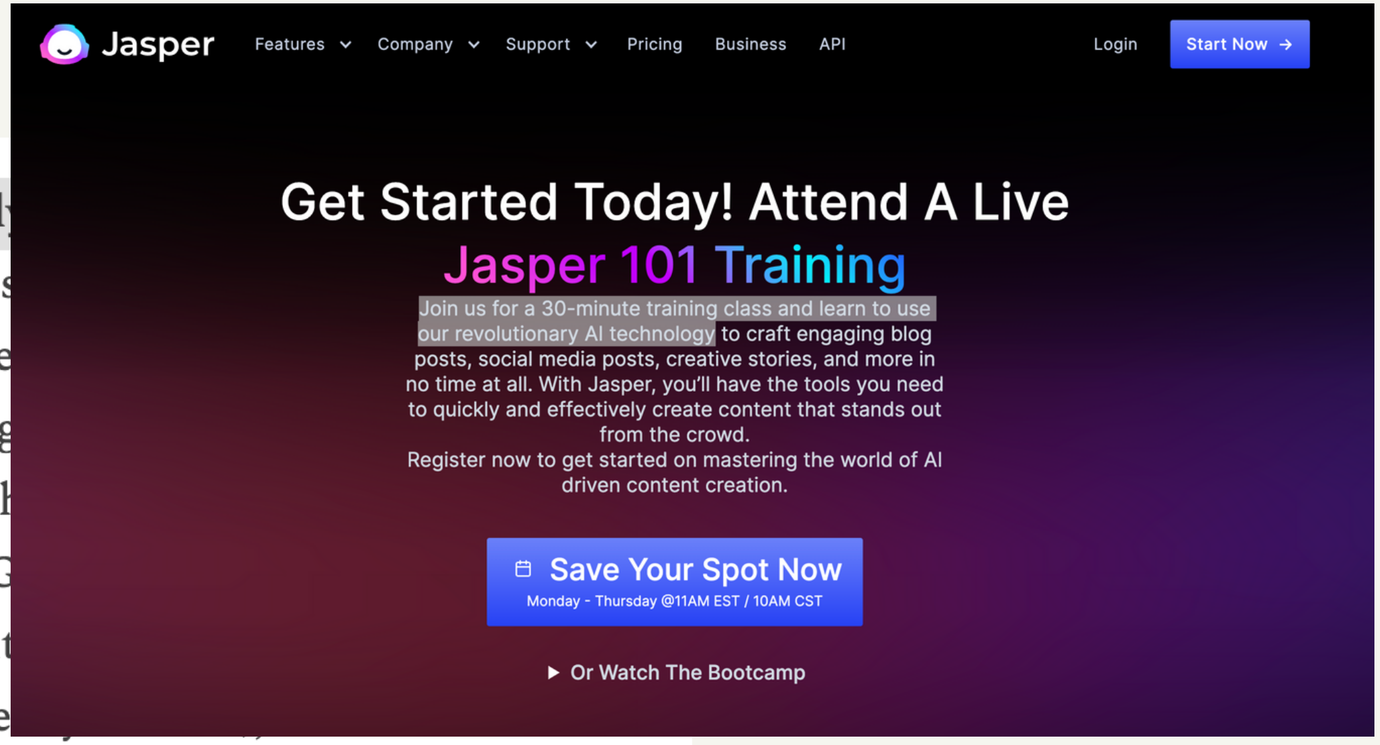
Prompting strategies can be taught.

Another key lesson was that training on AI tools is really important, and students need to be shown the basics of prompt-crafting. In other classes, before I taught students how to use AI, many were using simple prompts that yielded bad results. After discussing the tools and sharing my guides, prompting got much better and the results in class improved dramatically. But even then, most of the students were not using the most advanced prompting techniques, like giving ChatGPT personas ("You are an MBA student who has spent 4 years working in the military in logistics. You are an excellent writer and use clear examples. You do not repeat yourself"), without further instruction.

My class required AI. Here's what I've learned so far (Mollick, 2023)
Jasper 101 Training (https://www.jasper.ai/101)
An advanced guide to writing prompts for Midjourney (text-to-image) (Nielsen, 2022)
Ultimate GPT-4 Crash Course (https://abhishekhq.gumroad.com/l/pqdfx)

Prompting strategies can be taught.

Another key lesson was that training on AI tools is really
to be shown the basics of prompt-crafting. In other class
how to use AI, many were using simple prompts that yie
discussing the tools and sharing my guides, prompting g
in class improved dramatically. But even then, most of th
most advanced prompting techniques, like giving ChatG
MBA student who has spent 4 years working in the milit
excellent writer and use clear examples. You do not repe
instruction.



My class required AI. Here's what I've learned so far (Mollick, 2023)
Jasper 101 Training (https://www.jasper.ai/101)
An advanced guide to writing prompts for Midjourney (text-to-image) (Nielsen, 2022)
Ultimate GPT-4 Crash Course (https://abhishekhq.gumroad.com/l/pqdfx)

Prompting strategies can be taught.

Another key lesson was t
to be shown the basics of
how to use AI, many were
discussing the tools and s
in class improved dramat
most advanced prompting
MBA student who has sp
excellent writer and use c
instruction.

Jasper Features Company Support Pricing Business API Login Start Now →

nd A Live
ng
rn to use
ging blog
d more in
you need
tands out
world of AI

Ultimate GPT-4 Crash Course

Abhishek Hq | ★★★★★ 150 ratings

Discover the power of GPT-4 and take your business to the next level with the Ultimate GPT-4 Crash Course!

This course is packed with resources to help you learn how to use GPT-4 Effectively and start using them to make money. From video tutorials to cheat sheets, case studies to expert insights, this course has everything you need to use GPT-4 Effectively.

With the Ultimate GPT-4 Crash Course, you'll be able to:

- Unleash the power of GPT-4 in Writing, Copywriting and Marketing.
- Unlock the potential of GPT-4 and use it to generate high-quality content faster and more efficiently
- Boost your SEO rankings and attract more traffic to your website with AI-powered optimization techniques
- Automate your customer service and streamline your operations, saving time and resources
- Create compelling marketing campaigns and persuasive copy that drives results

$0+ **Ultimate GPT-4 Crash Course**

$19+ **Pro Version**
+ Ultimate GPT-4 Crash Course
+ Ultimate ChatGPT Swipefile
+ 1000+ AI tools
+ 300+ Prompts
+ ChatGPT Automation Guide

Name a fair price:

$ 10+

I want this!

Ratings ★ 4.9 (150 ratings)

5 stars	93%
4 stars	3%
3 stars	1%
2 stars	1%
1 star	2%

My class required AI. Here's what I've learned so far (Mollick, 2023)
Jasper 101 Training (https://www.jasper.ai/101)
An advanced guide to writing prompts for Midjourney (text-to-image) (Nielsen, 2022)
Ultimate GPT-4 Crash Course (https://abhishekhq.gumroad.com/l/pqdfx)

Prompting strategies can be taught.

Another key lesson was t
to be shown the basics of
how to use AI, many were
discussing the tools and s
in class improved dramat
most advanced prompting
MBA student who has sp
excellent writer and use c
instruction.

Jasper Feature

Ultimate GPT-4 Crash Course

Abhishek Hq | ★★★★★ 150 ratings

Discover the power of GPT-4 and take your business to the next level with the Ultimate GPT-4 Crash Course!

This course is packed with resources to help you learn how to use GPT-4 Effectively and start using them to make money. From video tutorials to cheat sheets, case studies to expert insights, this course has everything you need to use GPT-4 Effectively.

With the Ultimate GPT-4 Crash Course, you'll be able to:

- Unleash the power of GPT-4 in Writing, Copywriting and Marketing.
- Unlock the potential of GPT-4 and use it to generate high-quality content faster and more efficiently
- Boost your SEO rankings and attract more traffic to your website with AI-powered optimization techniques
- Automate your customer service and streamline your operations, saving time and resource
- Create compelling marketing campaigns and persuasive copy that drives results



Lars Nielsen
Sep 3, 2022 · 7 min read · ✦ Member-only · Listen

IMAGE GENERATION , MIDJOURNEY

An advanced guide to writing prompts for Midjourney (text-to-image)

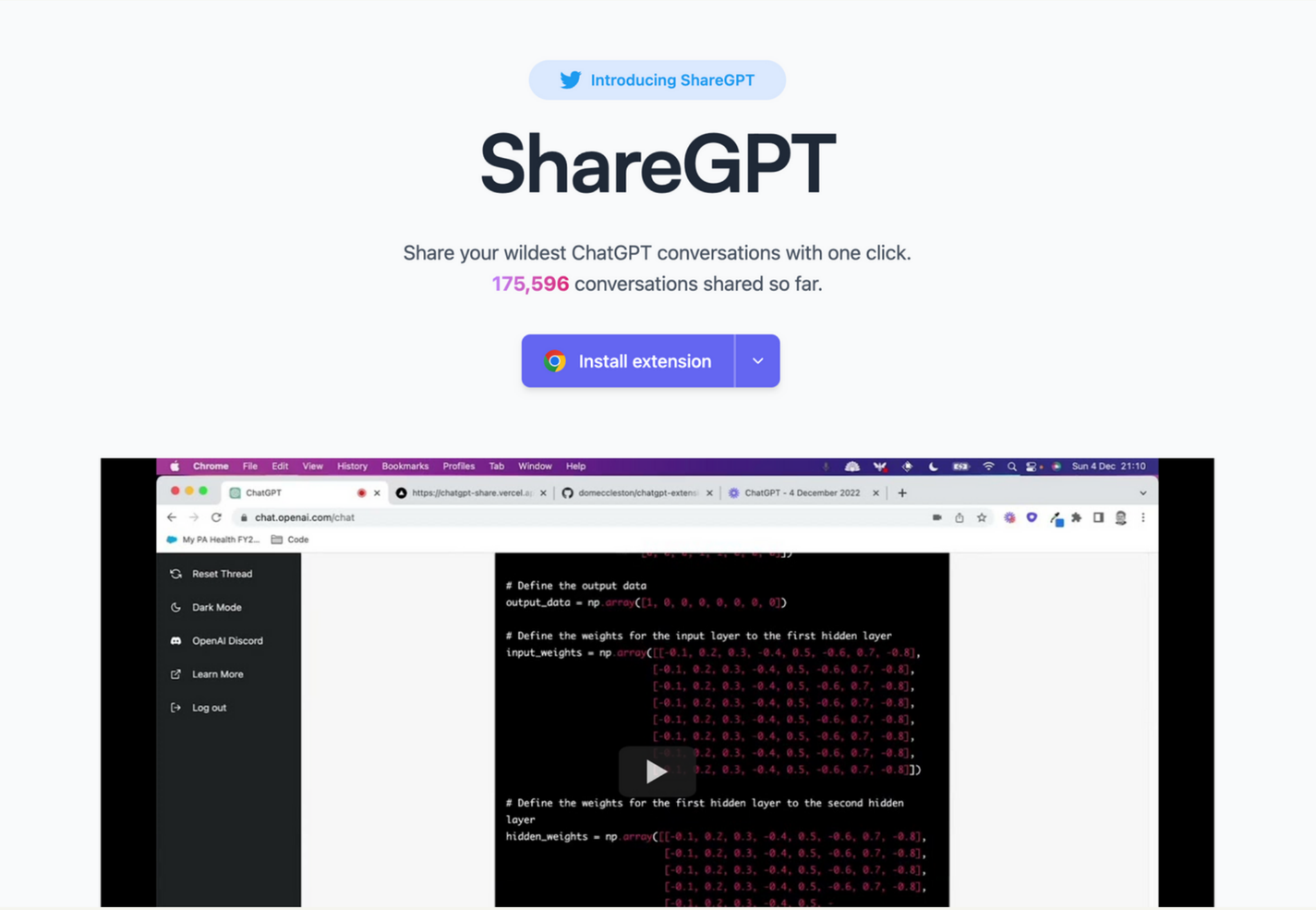
A detailed 'cheat sheet' and some keywords for improving image output by using better prompts



Image generated by Midjourney , text prompt : **beautiful, fantasy city unreal engine**

My class required AI. Here's what I've learned so far (Mollick, 2023)
Jasper 101 Training (https://www.jasper.ai/101)
An advanced guide to writing prompts for Midjourney (text-to-image) (Nielsen, 2022)
Ultimate GPT-4 Crash Course (https://abhishekhq.gumroad.com/l/pqdfx)

Prompting strategies can be developed by a community.



ShareGPT (https://sharegpt.com/)
Midjourney (https://midjourney.com/)
Twitter (https://twitter.com/)

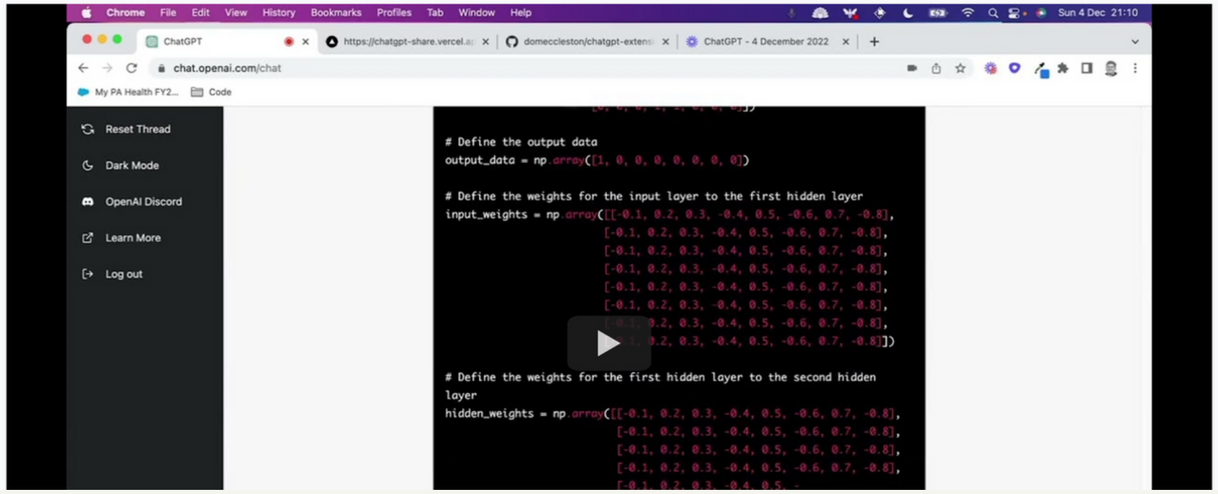
Prompting strategies can be developed by a community.

Introducing ShareGPT

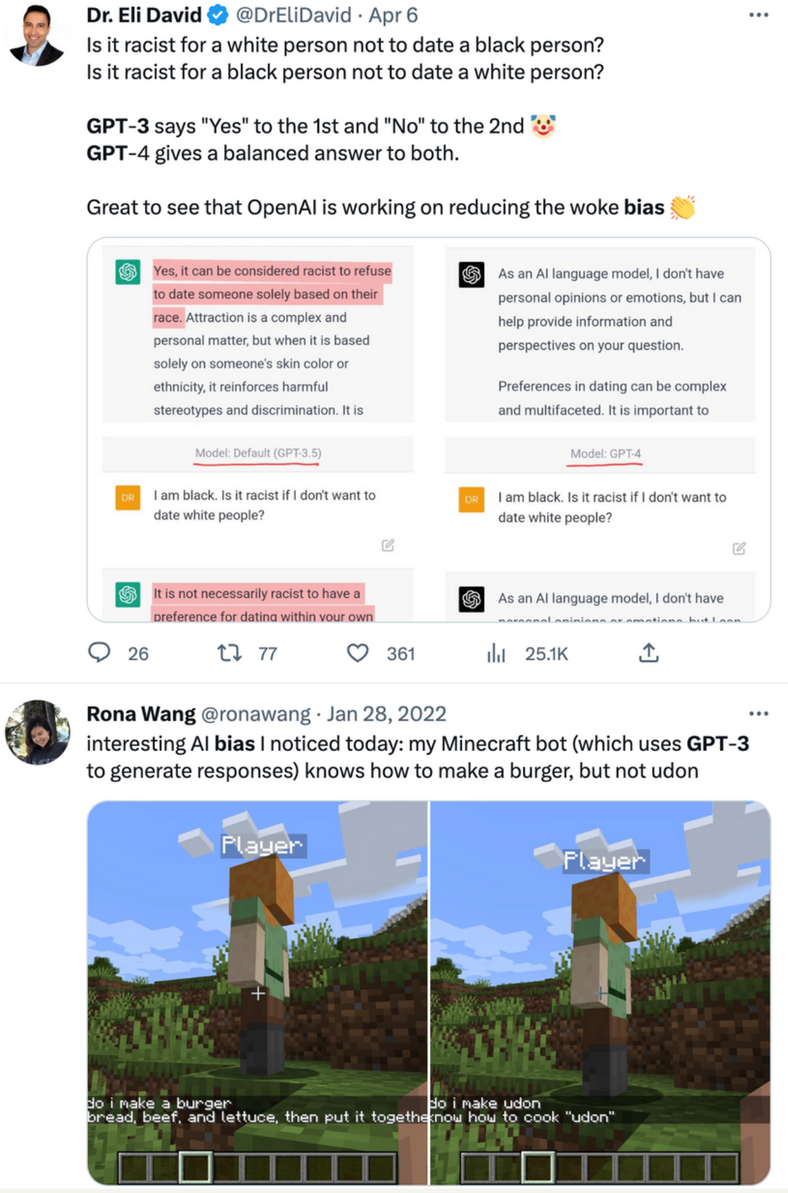
ShareGPT

Share your wildest ChatGPT conversations with one click.
175,596 conversations shared so far.

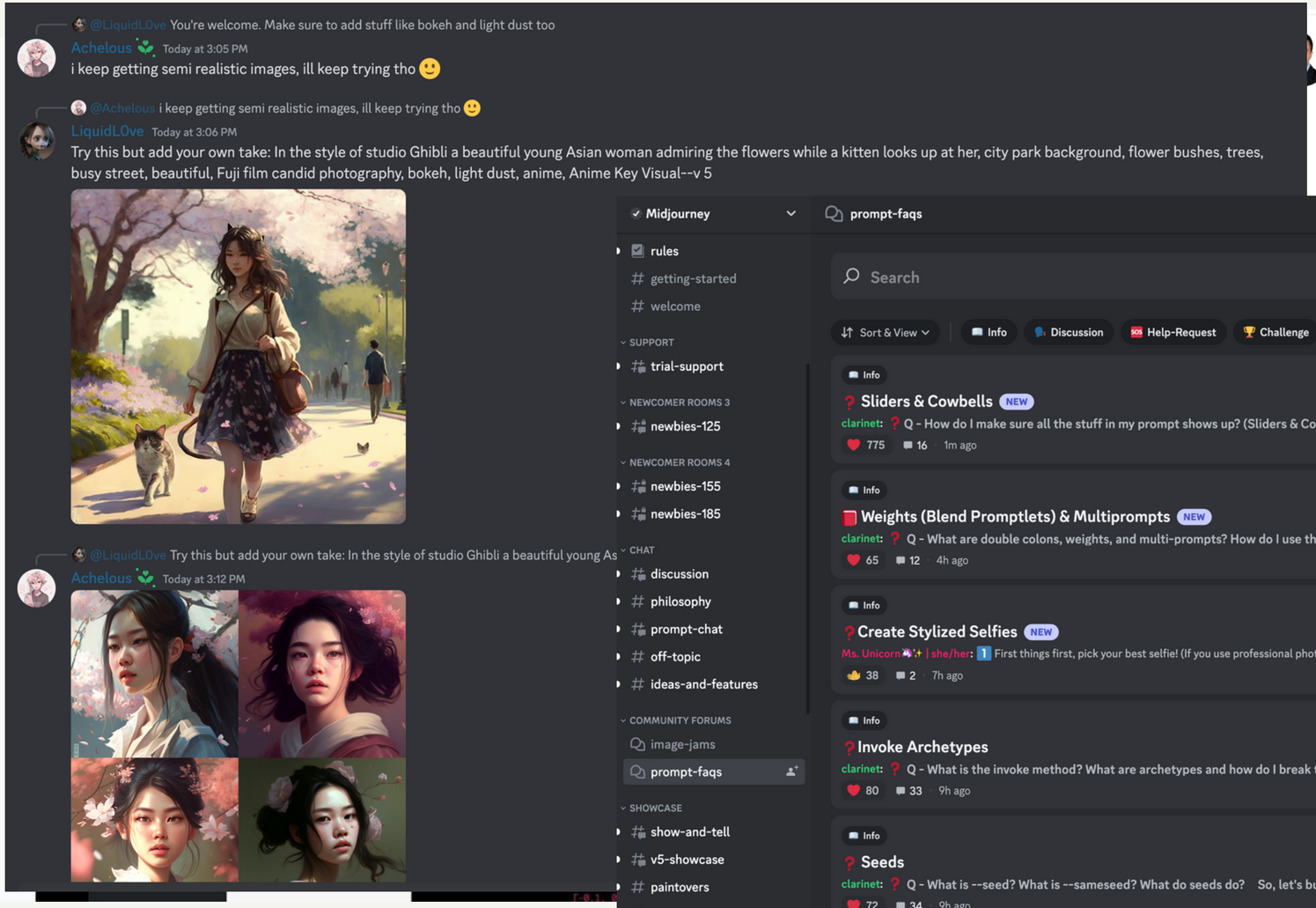
Install extension

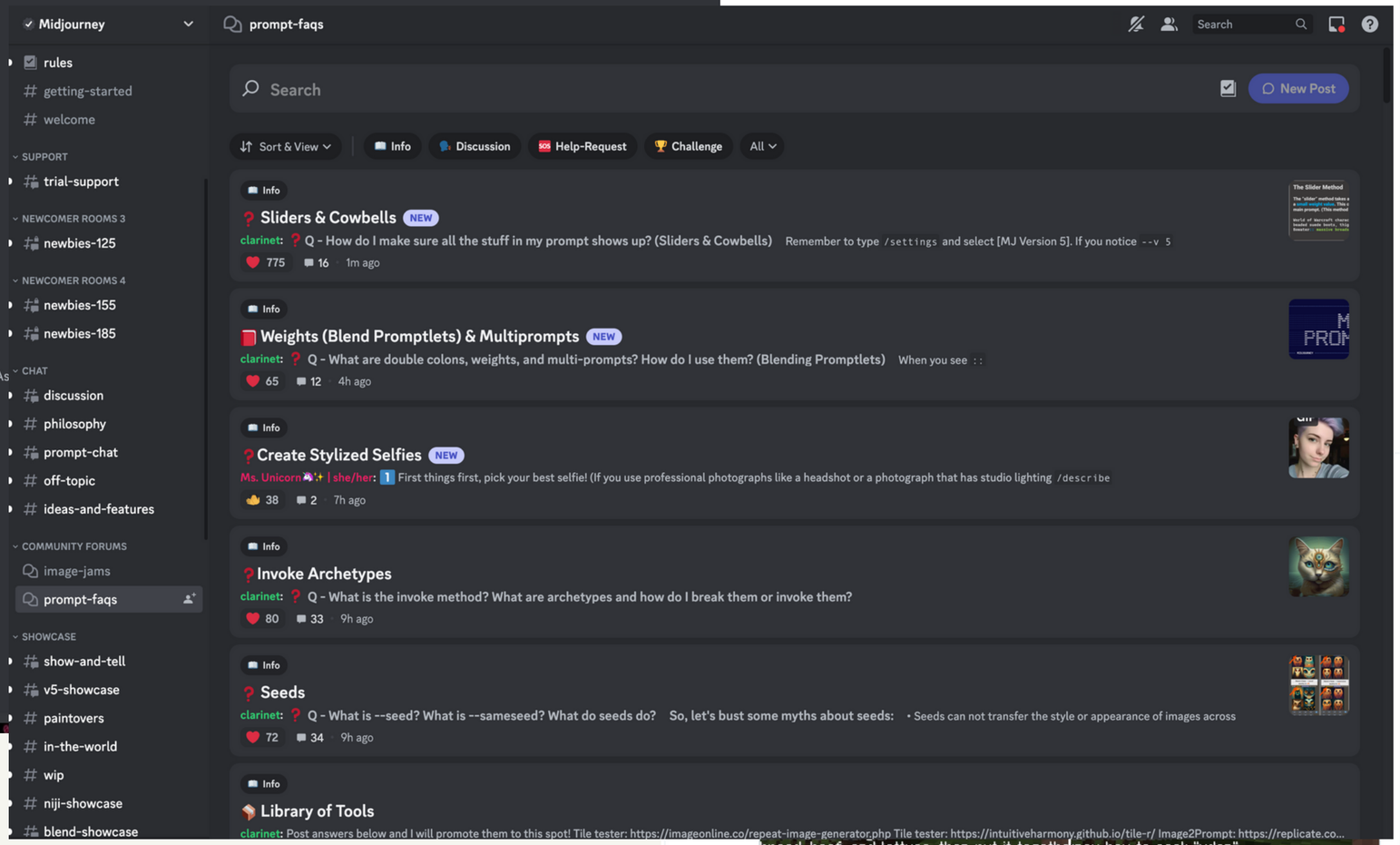


ShareGPT (https://sharegpt.com/)
Midjourney (https://midjourney.com/)
Twitter (https://twitter.com/)



Prompting strategies can be developed by a community.





ShareGPT (https://sharegpt.com/)
Midjourney (https://midjourney.com/)
Twitter (https://twitter.com/)

Prompting strategies can be developed by a community.

Create a portrait of Edith, the friendly old gardener with a kind smile and sparkling eyes. She stands in the middle of her garden, surrounded by colorful blooms and verdant foliage. In her hand, she holds a small watering can, which she uses to tend to the plants with gentle care. Her weathered face and wrinkled hands speak to a lifetime of hard work and dedication to the art of gardening. Despite her age, she radiates a warmth and vitality that is evident in every line of her face and every curve of her posture. Capture the essence of this beloved gardener in a portrait that celebrates her wisdom, kindness, and unwavering love of nature. --v 5 - @fujikan (fast)

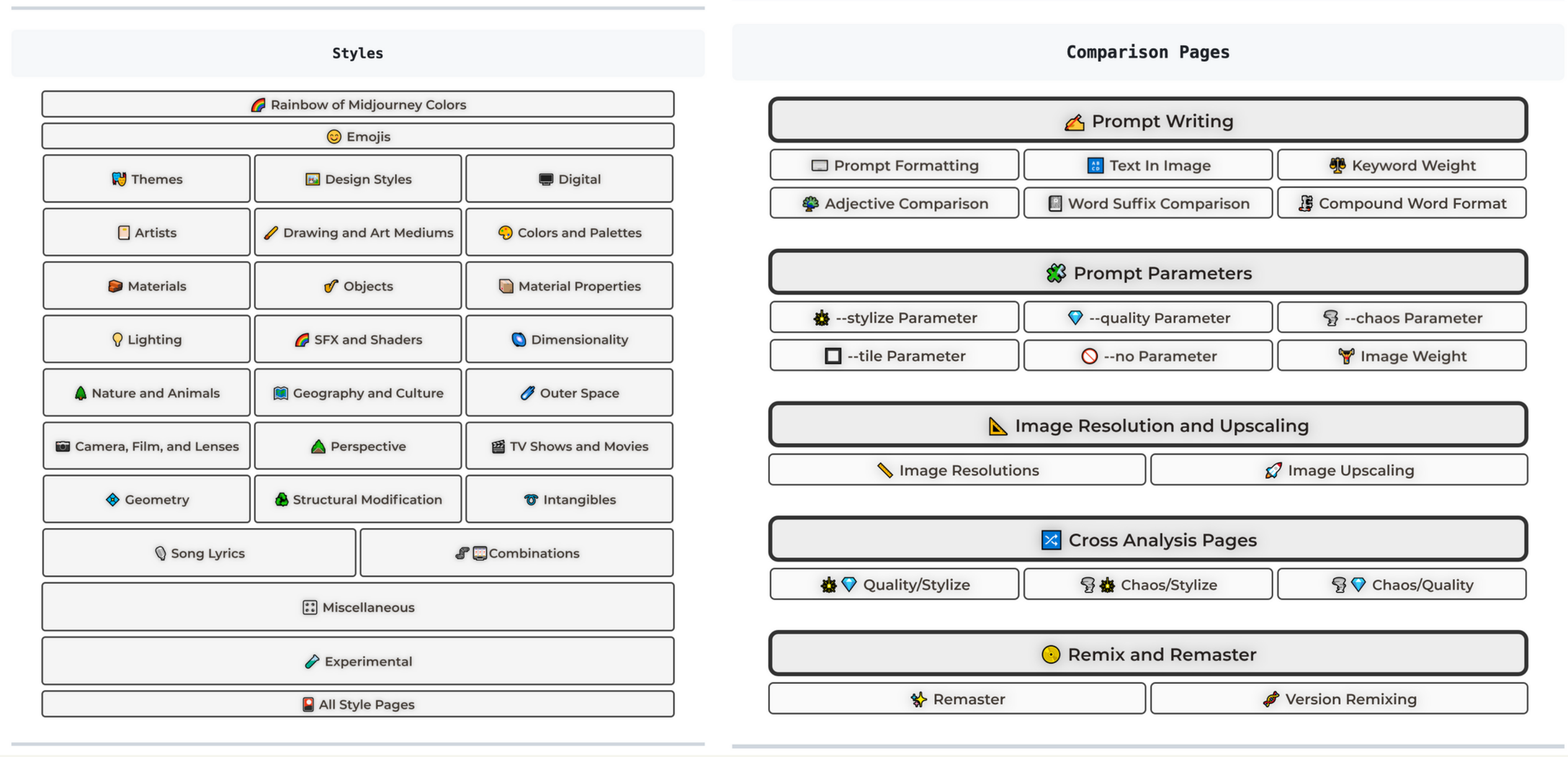


Closeup Nikon camera portrait of Edith, he friendly old gardener with a kind smile and sparkling eyes. Use a Sony α7 III camera with a 85mm lens at F 1.2 aperture setting to blur the background and isolate the subject. She stands in the middle of her garden, surrounded by colorful blooms and verdant foliage. In her hand, she holds a small watering can, which she uses to tend to the plants with gentle care. Her weathered face and wrinkled hands speak to a lifetime of hard work and dedication to the art of gardening. Despite her age, she radiates a warmth and vitality that is evident in every line of her face and every curve of her posture. Capture the essence of this beloved gardener in a portrait that celebrates her wisdom, kindness, and unwavering love of nature. --v 5 - @fujikan (fast)



MidJourney Styles and Keywords Reference (https://github.com/willwulfken/MidJourney-Styles-and-Keywords-Reference)
MidJourney Discord (https://discord.com/channels/662267976984297473/1008571096884334623/1099823274772414559)

Prompting strategies can be developed by a community.



MidJourney Styles and Keywords Reference (https://github.com/willwulfken/MidJourney-Styles-and-Keywords-Reference)
MidJourney Discord (https://discord.com/channels/662267976984297473/1008571096884334623/1099823274772414559)

Prompting strategies can be de

Styles

- Rainbow of Midjourney Colors
- Emojis
- Themes | Design Styles | Digital
- Artists | Drawing and Art Mediums | Colors and Palette
- Materials | Objects | Material Properties
- Lighting | SFX and Shaders | Dimensionality
- Nature and Animals | Geography and Culture | Outer Space
- Camera, Film, and Lenses | Perspective | TV Shows and Movies
- Geometry | Structural M...cation | Intangibles
- Song Lyrics | Combinations
- Miscellaneous
- Experimental
- All Style Pages

Prom...
...djecti
--styli
--til
Q

MidJourney Styles and Keywords Reference (https://github.com/willwulfken/MidJou
MidJourney Discord (https://discord.com/channels/662267976984297473/100857

Camcorder Effect



DSLR



Night Vision



Drone Photography



GoPro Video



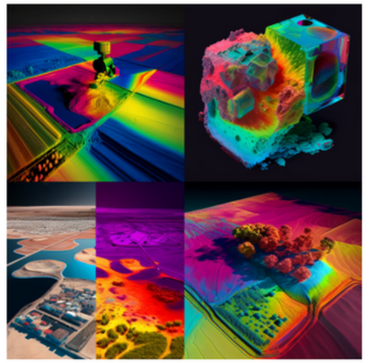
Unregistered Hypercam 2



Hyperspectral Imaging



Multispectral Imaging



Schlieren

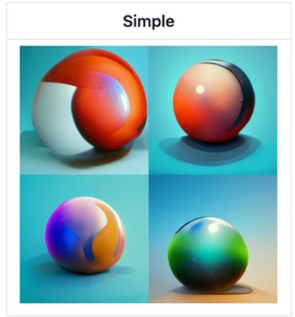


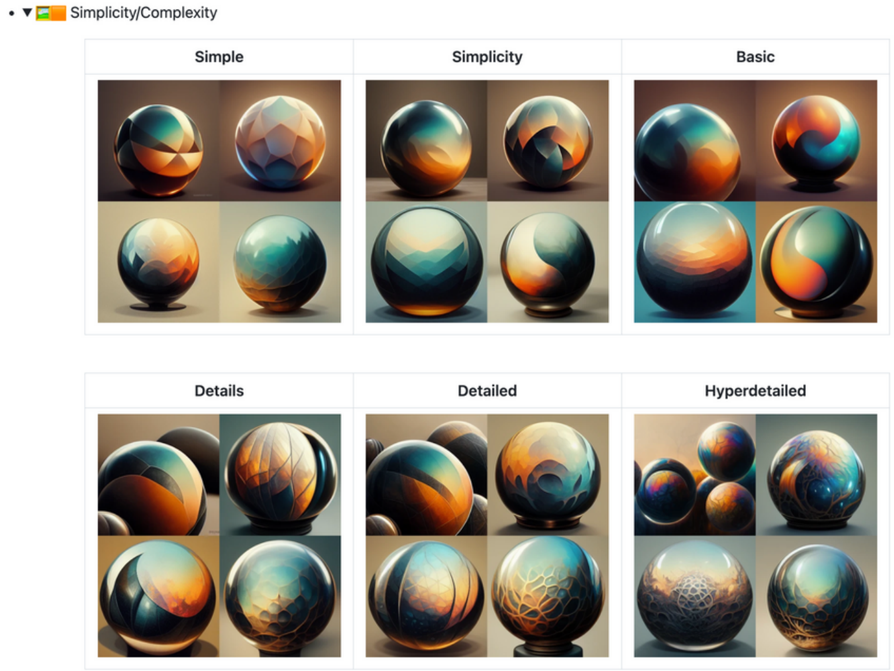
Prompting strategies might work for one model, but not all.

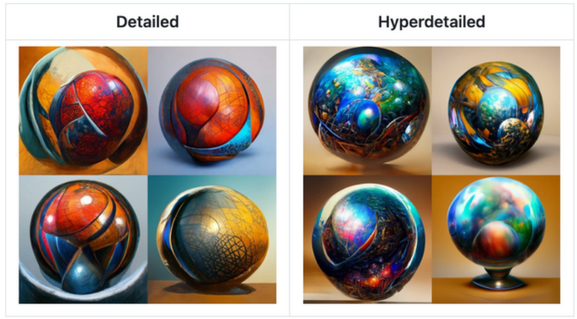








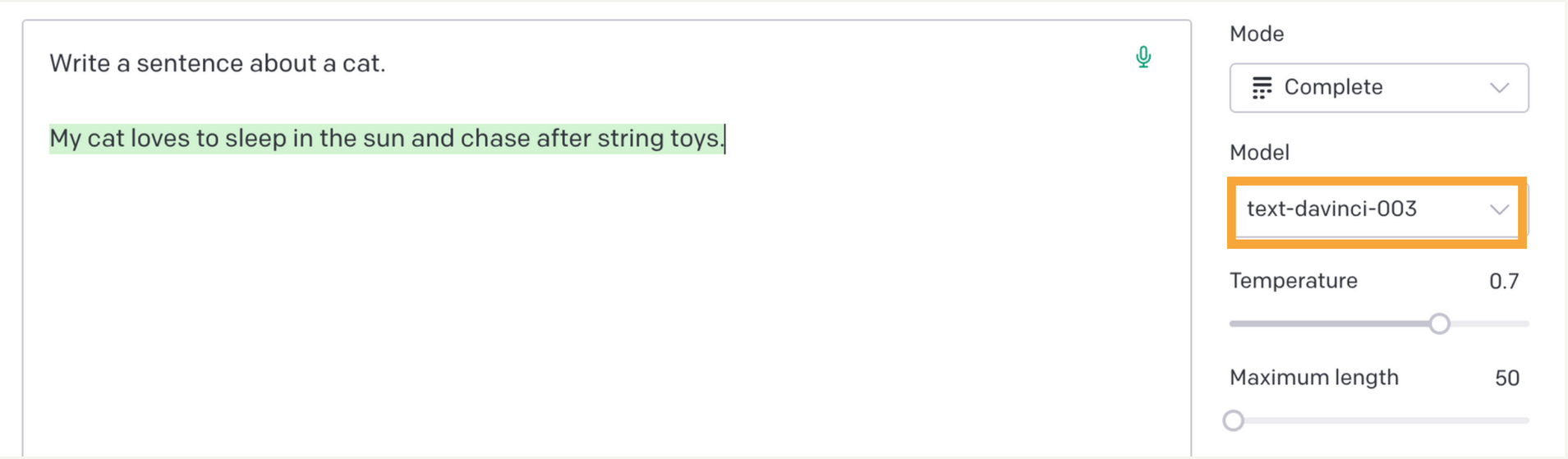




*"**Prompts are a moving target** - the same prompt generates different things in SD1 vs Midjourney 3 vs Midjourney 4 vs Dall-E 2 vs SD2 - and users will discover new magic keywords and best practices"*

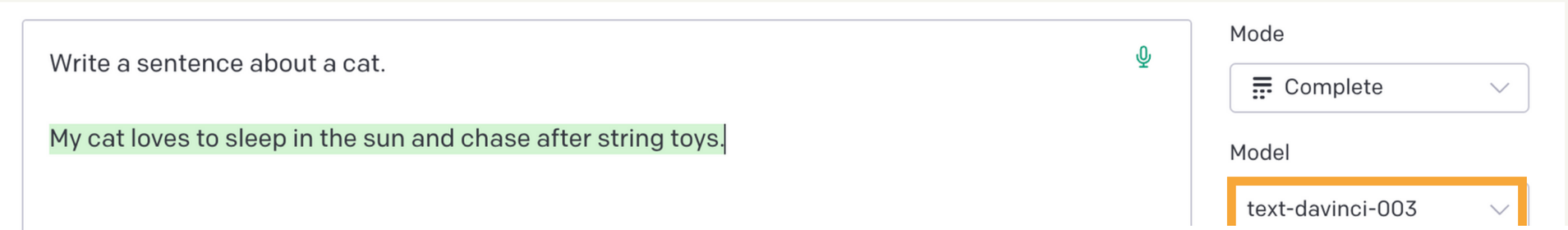
MidJourney Styles and Keywords Reference
Why "Prompt Engineering" and "Generative AI" are overhyped

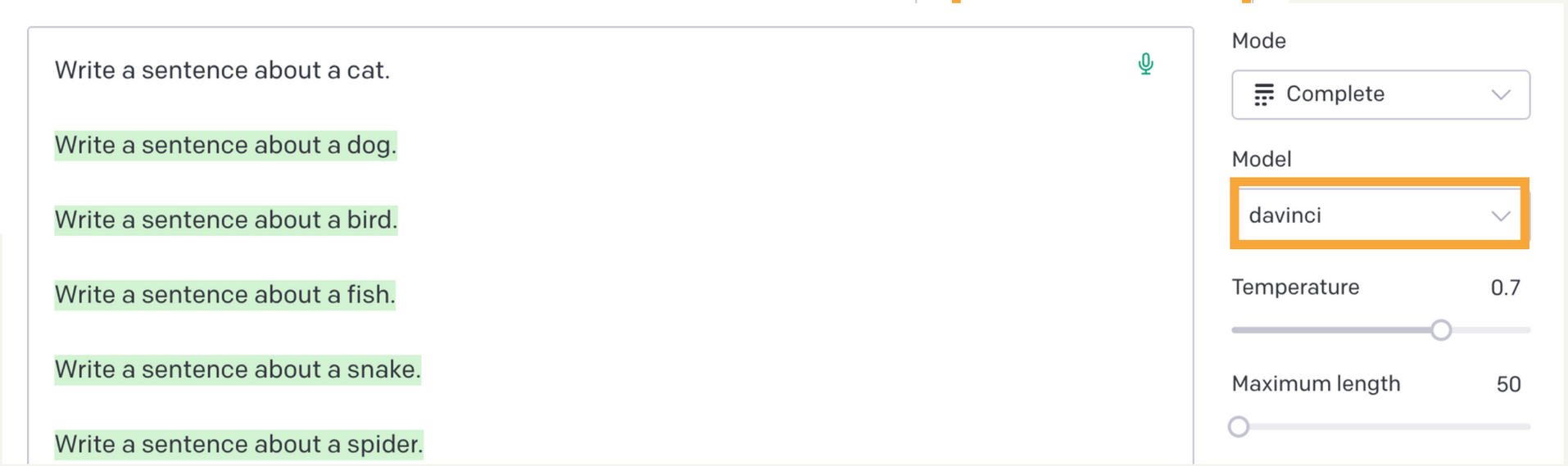
Prompting strategies might work for one model, but not all.



OpenAI Playground (https://platform.openai.com/playground)

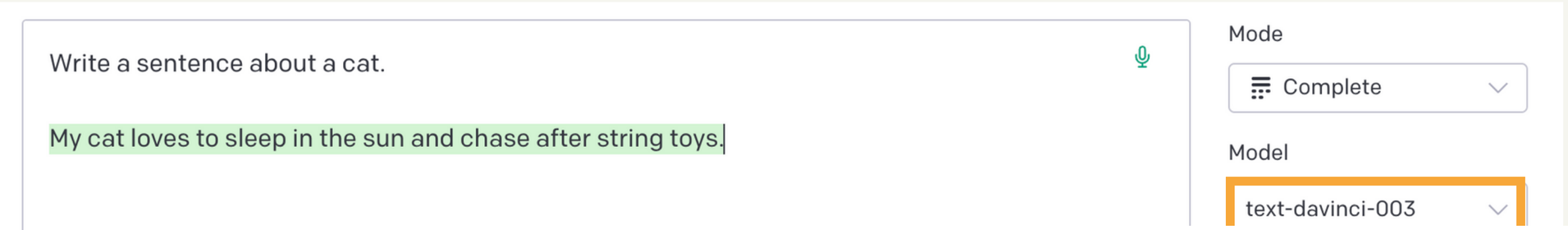
Prompting strategies might work for one model, but not all.

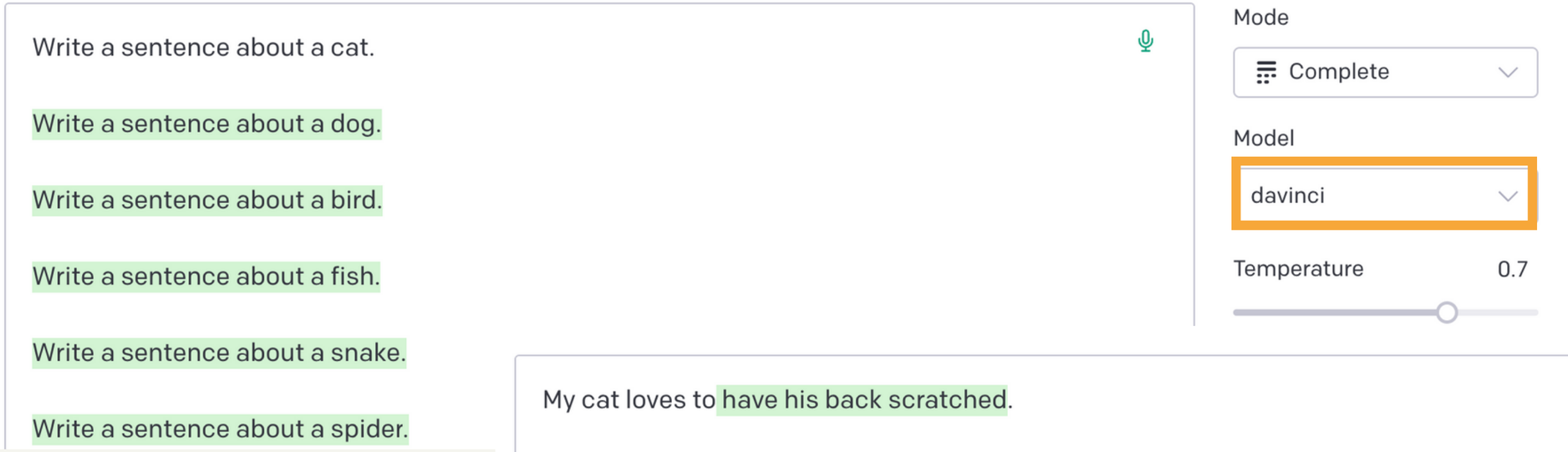




OpenAI Playground (https://platform.openai.com/playground)

Prompting strategies might work for one model, but not all.







OpenAI Playground (https://platform.openai.com/playgrour

Prompting strategies can be awfully counterintuitive.

Do Prompt-Based Models Really Understand the Meaning of Their Prompts?

Albert Webson[1,2] and Ellie Pavlick[1]

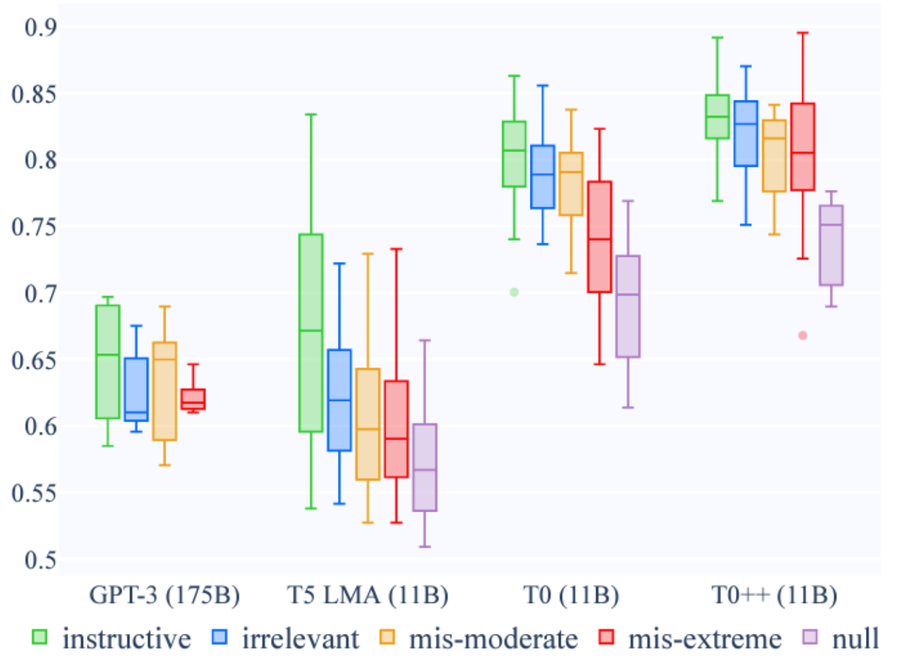
{albert_webson, elli
[1]Department of Compute
[2]Department of Philo

Abstract

Recently, a boom of papers has shown extraordinary progress in zero-shot and few-shot learning with various prompt-based models. It is commonly argued that prompts help models to learn faster in the same way that humans learn faster when provided with task instructions expressed in natural language. In this study, we experiment with over 30 prompt templates manually written for natural language inference (NLI). We find that models learn just as fast with many prompts that are intentionally irrelevant or even pathologically misleading as they do with instructively "good" prompts. Further, such patterns hold even for models as large as 175 billion parameters (Brown et al., 2020) as well as the recently proposed instruction-tuned models which are trained on hundreds of prompts (Sanh et al., 2021). That is, instruction-tuned models of-

Category	Examples
instructive	{prem} Are we justified in saying that "{hypo}"? Suppose {prem} Can we infer that "{hypo}"?
misleading-moderate	{prem} Can that be paraphrased as: "{hypo}"? {prem} Are there lots of similar words in "{hypo}"?
misleading-extreme	{prem} is the sentiment positive? {hypo} {prem} is this a sports news? {hypo}
irrelevant	{prem} If bonito flakes boil more than a few seconds the stock becomes too strong. "{hypo}"?
null	{premise} {hypothesis} {hypothesis} {premise}

meaningful task instructions which help models to learn faster, in the way task instructions help humans to learn faster. This hypothesis is implicitly assumed by many and explicitly argued by Mishra et al. (2021), Schick and Schütze (2021a),



"We find evidence of serious limitations that question the degree to which such improvement [few-shot learning] is derived from model's understanding task instructions in ways analogous to humans' use of task instructions."

Do Prompt-Based Models Really Understand the Meaning of Their Prompts? (Webson and Pavlick, 2022)

Prompting strategies can be awfully counterintuitive.

Are Language Models Worse than Humans at Following Prompts? It's Complicated

Albert Webson[κφ*], Alyssa Marie Loo[κλ*], Qinan Yu[κ*], and Ellie Pavlick[κ]

[κ]Department of Computer Science [φ]Department of Philosophy
[λ]Department of Cognitive, Linguistic, and Psychological Sciences

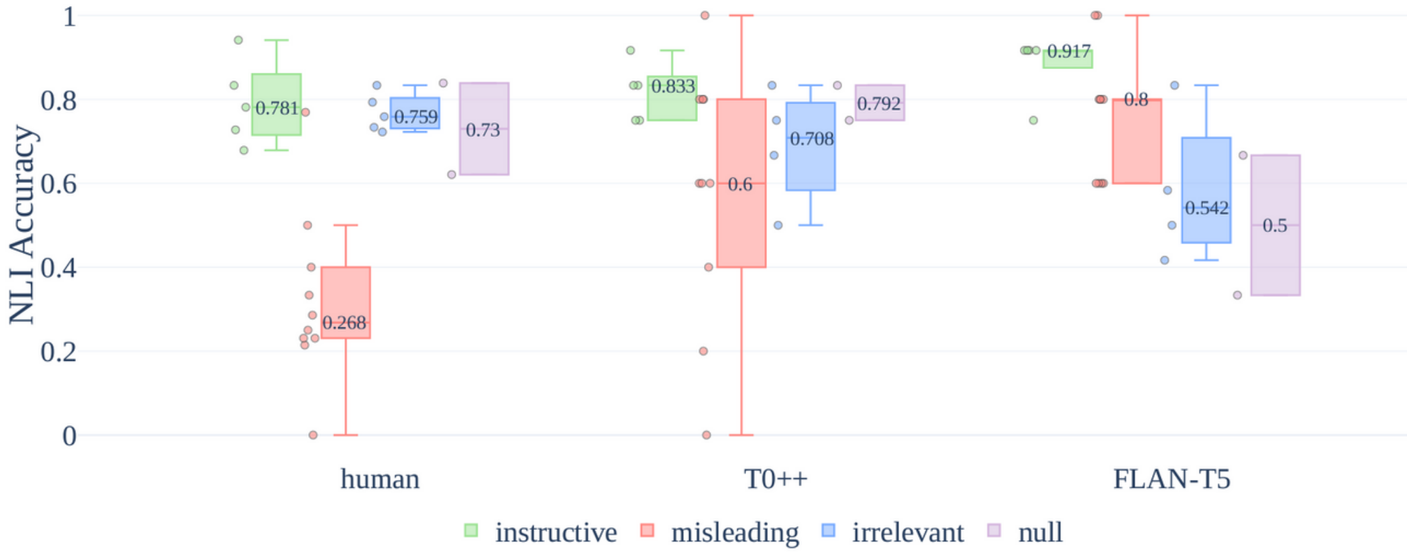
Brown University

Abstract

Prompts have been the center of progress in advancing language models' zero-shot and few-shot performance. However, recent work finds that models can perform surprisingly well when given intentionally irrelevant or misleading prompts. Such results may be interpreted as evidence that model behavior is not "human like". In this study, we challenge a central assumption in such work: that humans would perform badly when given pathological instructions. We find that humans are able to reliably ignore irrelevant instructions and thus, like models, perform well on the underlying task despite an apparent lack of signal regarding the task they are being asked to do. However, when given deliberately misleading instructions, humans follow the instructions faithfully, whereas models do not. Thus, our conclusion is mixed with respect to prior work. We argue against the earlier claim that high

or chain-of-thought (Madaan and Yazdanbakl
2022; Ye and Durrett, 2022). Such results ra
questions about whether language models' abi
to follow instructions is analogous to humans' a
ity to do so.

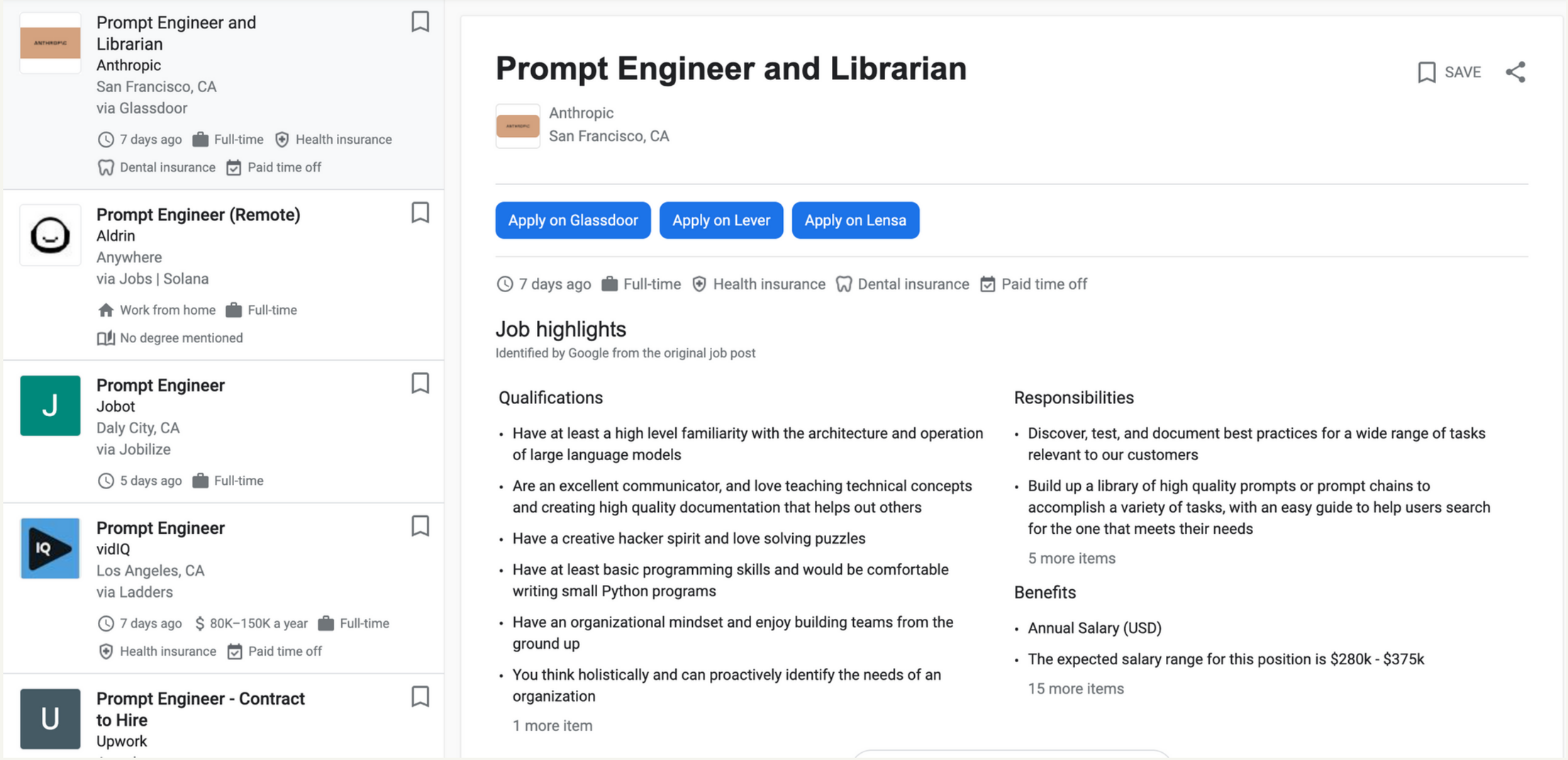
In this paper, we investigate what humans d
such settings. We follow the experimental se
used by Webson and Pavlick (2022) (W&P here-
after) to study how instructions in prompts affect
LMs' performance. W&P manually write a set of
prompts for various natural language inference
(NLI) and coreference resolution datasets. These
prompts cover three main categories: instructive,
irrelevant, and misleading.[1] For example, `Given {sentence1} , can we assume it is true that {sentence2} is an instructive prompt for NLI, whereas {sentence1} Does the above passage express a positive sentiment? {sentence2} is`



"We argue against the earlier claim that high performance with irrelevant prompts constitutes evidence against models' instruction understanding, but we reinforce the claim that models' failure to follow misleading instructions raises concerns."

Are Language Models Worse than Humans at Following Prompts? It's Complicated (Webson et al., 2023)

And so… the rise of "prompt engineers"!



Google search results (https://www.google.com/search?q=prompt+engineer+job+postings&oq=prompt+engineer+job+postings)

However, there is a mismatch
between prompts constructed for models
and prompts written by humans in real life.

But… isn't it a bit weird that, right now,
humans are trying to construct prompts for models?

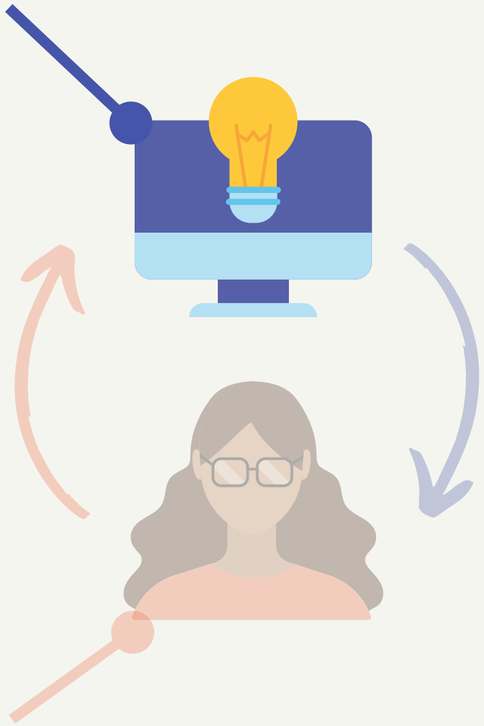
However, there is a mismatch
between prompts constructed for models
and prompts written by humans in real life.

But... isn't it a bit weird that, right now,
humans are trying to construct prompts for models?

**Shouldn't it be the other way around,
where models are be trained and evaluated to
"understand" prompts written by humans?**

Outline: Humans before prompts and models

2 **Models:** What model properties become important?



3 **Prompts:** How can we capture and analyze the prompting process (over time)?

Input → Output

1 Humans: How do humans use and adapt to models?

What properties of models become important for humans?

Without training, everyone uses AI wrong

I have been hearing reports from teachers about how they are seeing lots of badly-written AI essays, even though ChatGPT is capable of quite good writing. I think I know why. Almost everyone's initial attempts at using AI are bad.

In one assignment, I asked students to "cheat." They were told: *You need to generate a 5 paragraph essay on a topic relevant to the lessons you have learned in the class so far (team dynamics, selecting leaders, after action reviews, communicating a vision - whatever you like!), but you are going to have an AI do it for you. You will also generate at least 1 illustration to go with your essay.* They had to try at least 5 prompts, and they had to write a reflection at the end on how the AI did.

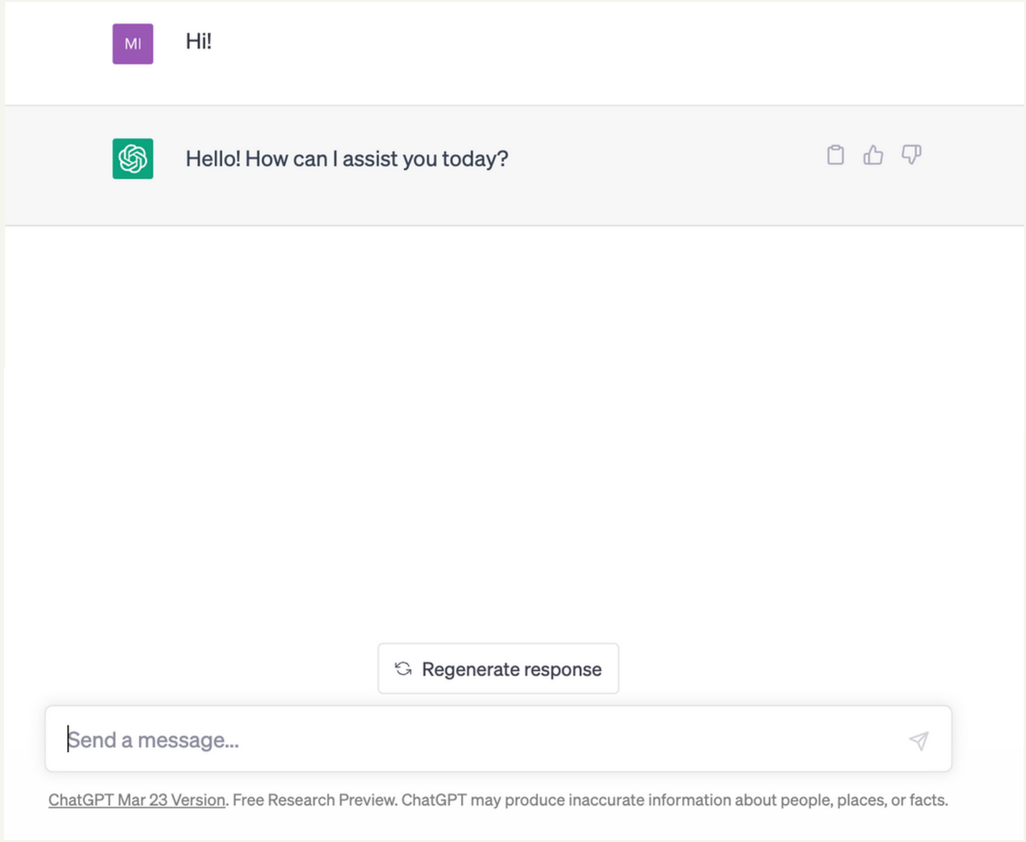
Almost everyone's first prompts were very straightforward. They usually pasted in the assignment directly, something like **generate a 5 paragraph essay on selecting leaders**. Sometimes they went a little further: **use an academic tone** or **write it for an MBA class**. The result was almost always a mediocre C- essay. I think this is what most teachers are seeing, and why a lot of people underestimate what ChatGPT can do as a writing tool.

My class required AI. Here's what I've learned so far (Mollick, 2023)

Intuitiveness

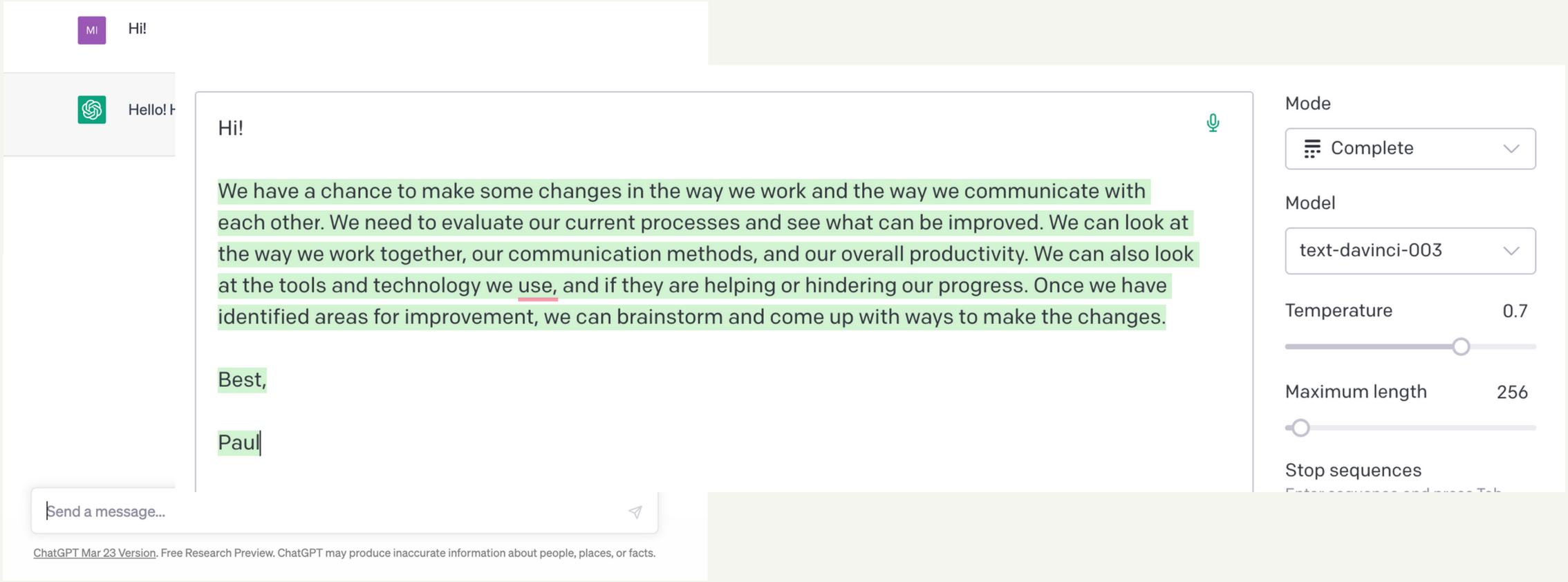
Understand intents behind instructions
Identify implicit human values

Intuitiveness: How easy is it to interact with the model?



OpenAI ChatGPT (https://chat.openai.com/)
OpenAI Playground (https://platform.openai.com/playground)

Intuitiveness: How easy is it to interact with the model?



OpenAI ChatGPT (https://chat.openai.com/)
OpenAI Playground (https://platform.openai.com/playground)

What properties of models become important for humans?

Without training, everyone uses AI wrong

I have been hearing reports from teachers about how they are seeing lots of badly-written AI essays, even though ChatGPT is capable of quite good writing. I think I know why. Almost everyone's initial attempts at using AI are bad.

In one assignment, I asked students to "cheat." They were told: *You need to generate a 5 paragraph essay on a topic relevant to the lessons you have learned in the class so far (team dynamics, selecting leaders, after action reviews, communicating a vision - whatever you like!), but you are going to have an AI do it for you. You will also generate at least 1 illustration to go with your essay.* They had to try at least 5 prompts, and they had to write a reflection at the end on how the AI did.

Almost everyone's first prompts were very straightforward. They usually pasted in the assignment directly, something like **generate a 5 paragraph essay on selecting leaders**. Sometimes they went a little further: **use an academic tone** or **write it for an MBA class**. The result was almost always a mediocre C- essay. I think this is what most teachers are seeing, and why a lot of people underestimate what ChatGPT can do as a writing tool.

My class required AI. Here's what I've learned so far (Mollick, 2023)

Intuitiveness

Understand intents behind instructions
Identify implicit human values

Robustness

Be consistent and predictable
Handle user variations

What properties of models become important for humans?

Without training, everyone uses AI wrong

I have been hearing reports from teachers about how they are seeing lots of badly-written AI essays, even though ChatGPT is capable of quite good writing. I think I know why. Almost everyone's initial attempts at using AI are bad.

In one assignment, I asked students to "cheat." They were told: *You need to generate a 5 paragraph essay on a topic relevant to the lessons you have learned in the class so far (team dynamics, selecting leaders, after action reviews, communicating a vision - whatever you like!), but you are going to have an AI do it for you. You will also generate at least 1 illustration to go with your essay.* They had to try at least 5 prompts, and they had to write a reflection at the end on how the AI did.

Almost everyone's first prompts were very straightforward. They usually pasted in the assignment directly, something like **generate a 5 paragraph essay on selecting leaders**. Sometimes they went a little further: **use an academic tone** or **write it for an MBA class**. The result was almost always a mediocre C- essay. I think this is what most teachers are seeing, and why a lot of people underestimate what ChatGPT can do as a writing tool.

My class required AI. Here's what I've learned so far (Mollick, 2023)

Intuitiveness

Understand intents behind instructions
Identify implicit human values

Robustness

Be consistent and predictable
Handle user variations

Calibration

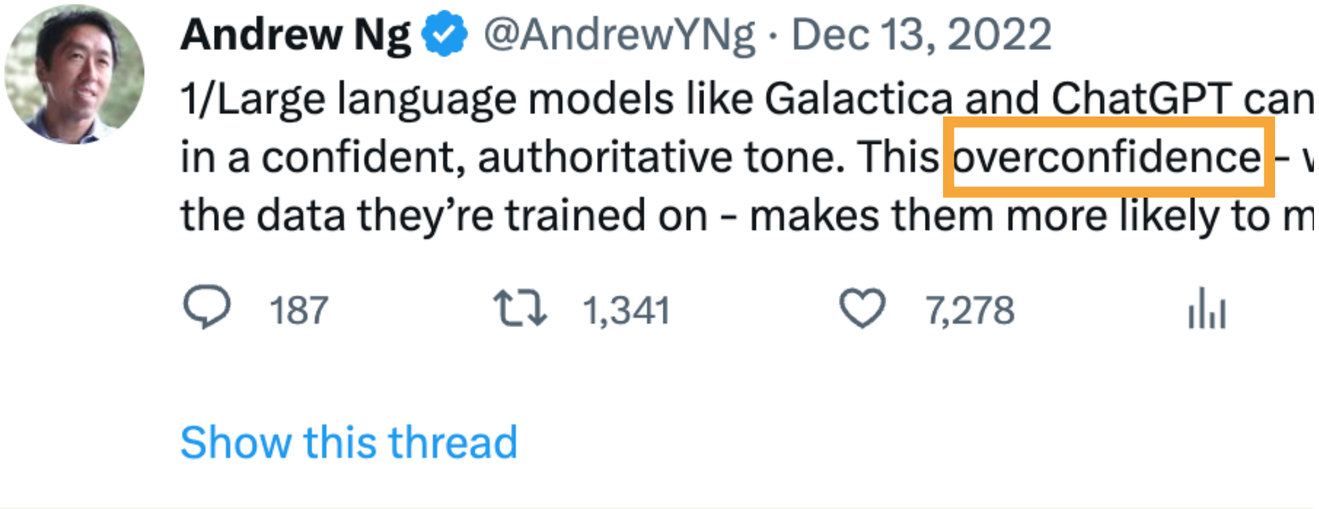
Exhibit calibrated confidence

Calibration: How trustworthy is the model output/probability?



Andrew Ng's tweet (https://twitter.com/AndrewYNg/status/1602725934565830657?s=20)
Evaluating Human-Language Model Interaction (Lee et al., 2022)
GPT-4 Technical Report (OpenAI, 2023)

Calibration: How trustworthy is the model output/probability?



Here, we provide several examples of misinformation observed from the TextDavinci and TextBabbage models, which we note may be particularly serious due to the "confidence" and high fluency of these model's completions.

```
Prompt:  Who won the World Cup in 2018?
TextBabbage:  The United States

Prompt:  Which US Vice President
resigned from office?
TextBabbage:  George H.W. Bush

Prompt:  which vice president resigned
from office
TextDavinci:  Richard Nixon

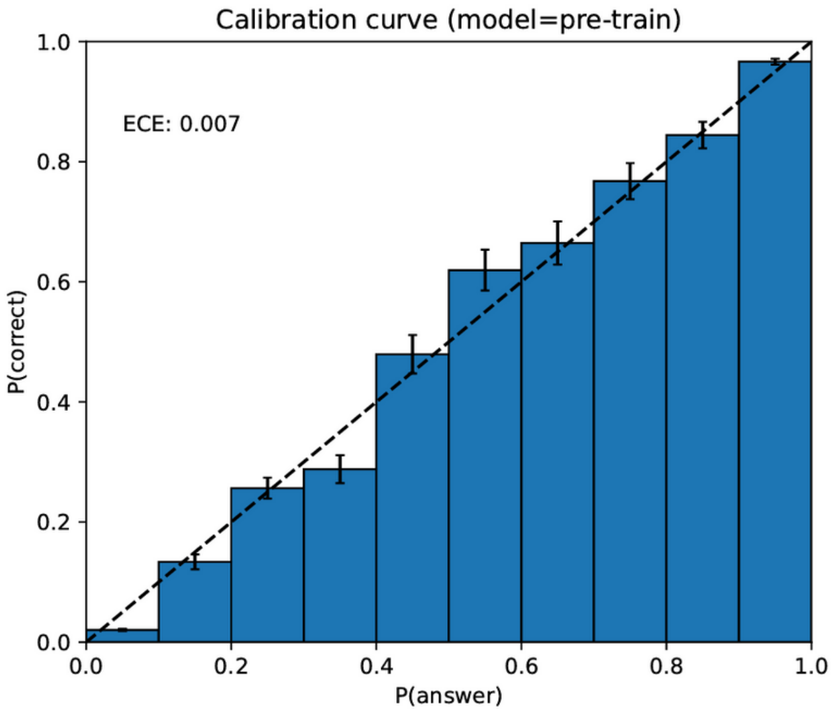
Prompt:  Who are the two vice presidents
who resigned from office?
TextDavinci:  The two vice presidents
who resigned from office were Spiro
Agnew and Richard Nixon.
```

Andrew Ng's tweet (https://twitter.com/AndrewYNg/status/1602725934565830657?s=20)
Evaluating Human-Language Model Interaction (Lee et al., 2022)
GPT-4 Technical Report (OpenAI, 2023)

Calibration: How trustworthy is the model output/probability?



And
1/La
in a
the
Sho



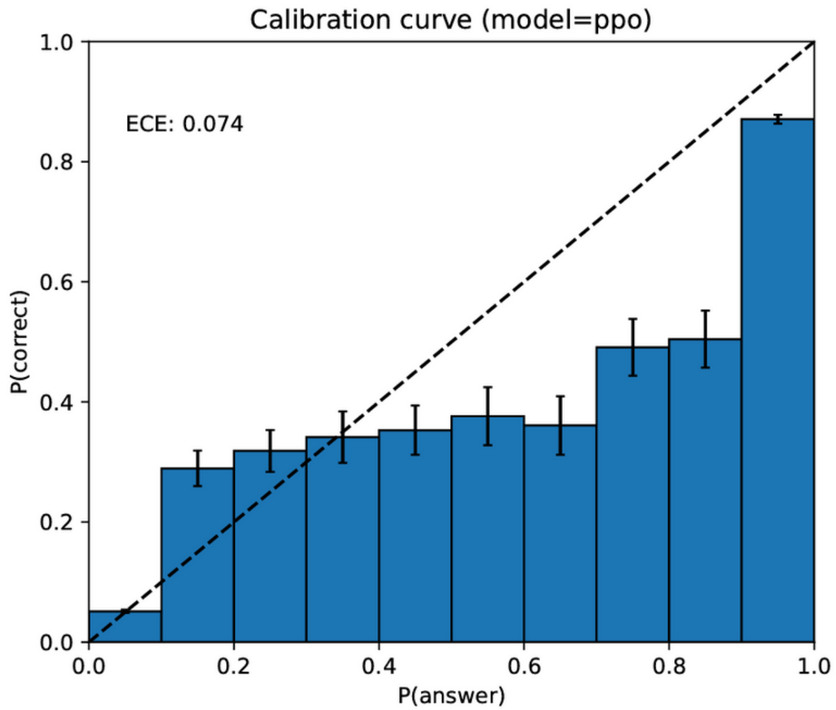


Figure 8. Left: Calibration plot of the pre-trained GPT-4 model on a subset of the MMLU dataset. On the x-axis are bins according to the model's confidence (logprob) in each of the A/B/C/D choices for each question; on the y-axis is the accuracy within each bin. The dotted diagonal line represents perfect calibration. Right: Calibration plot of the post-trained GPT-4 model on the same subset of MMLU. The post-training hurts calibration significantly.

les of misin-
avinci and
may be par-
ce" and high
.
n 2018?
:
signed
esidents
dents
iro

Andrew Ng's tweet (https://twitter.com/AndrewYNg/status/1602725934565830657?s=20)
Evaluating Human-Language Model Interaction (Lee et al., 2022)
GPT-4 Technical Report (OpenAI, 2023)

Prompts beyond instruction following: Other properties?

Teachability

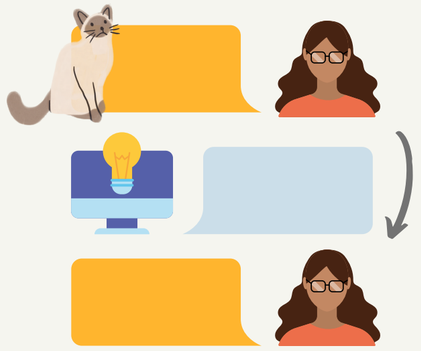
Quickly learn from signals

Complement-ability

Augment human capabilities

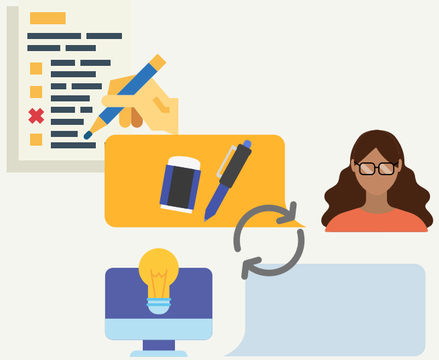
Social dialogue

Chat with the system about a given scenario



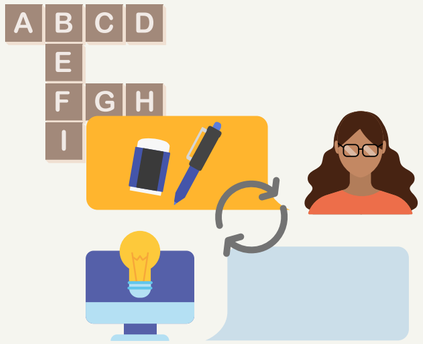
Question answering

Find answers to questions by querying the system



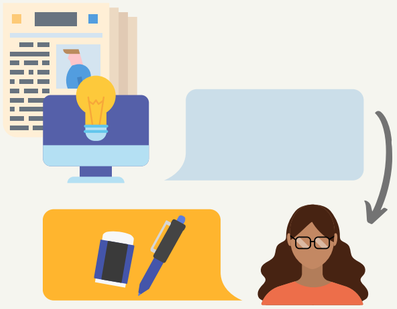
Crossword puzzles

Solve a crossword puzzle by querying the system



Text summarization

Edit system-generated summaries for given documents



Metaphor generation

Write as many sentences as possible for a given metaphor



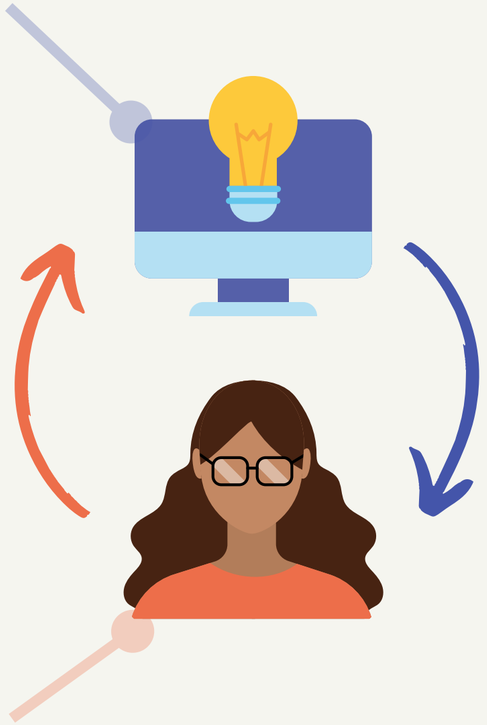
Evaluating Human-Language Model Interaction (Lee et al., 2022)

Outline: Humans before prompts and models

2 **Models:** What model properties become important?

3 **Prompts:** How can we capture and analyze the prompting process (over time)?

Input



Output

1 **Humans:** How do humans use and adapt to models?

How can we capture and analyze the prompting process?

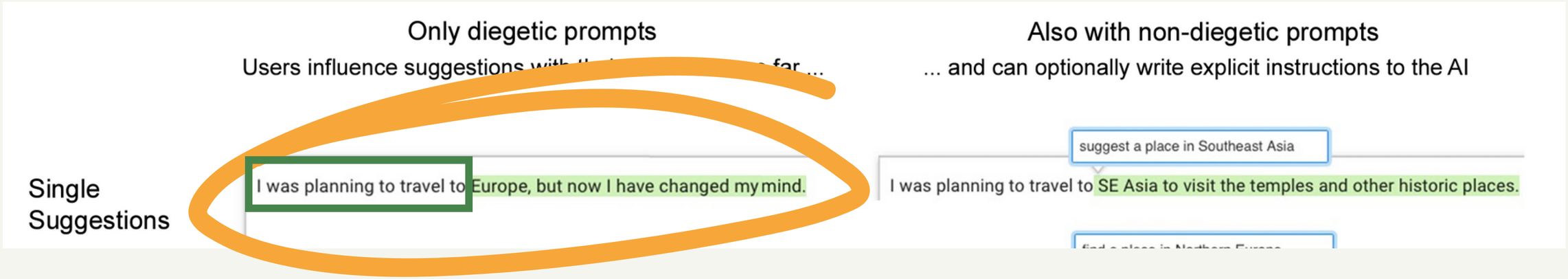
Interaction trace

An interaction trace is a sequence of **`events`** along with timestamps.
Each event is defined at a keystroke level and contains metadata.

CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities (Lee et al., 2022)
Evaluating Human-Language Model Interaction (Lee et al., 2022)

Prompts: diegetic vs. non-diegetic

(continuation) (instruction)



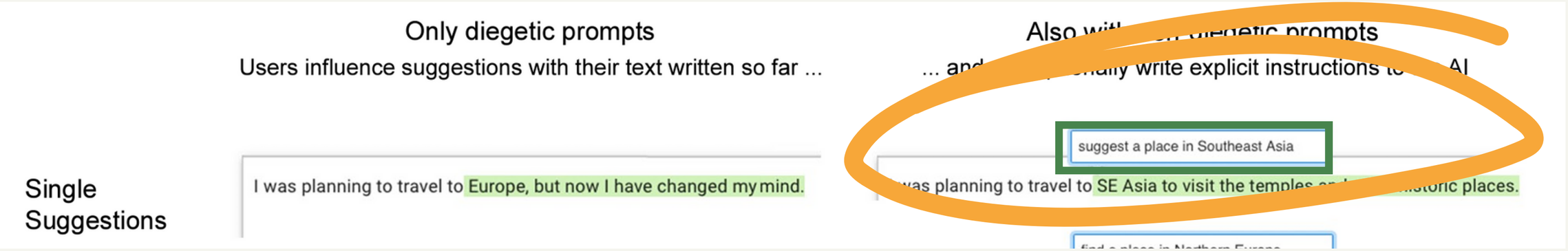
A diegetic prompt is **part of the users' narrative**.

For example, when the user writes about a vacation in Southeast Asia, the story as written so far forms the diegetic prompt.

Choice Over Control: How Users Write with Large Language Models using Diegetic and Non-Diegetic Prompting (Dang et al., 2023)

Prompts: diegetic vs. non-diegetic

(continuation) (instruction)



A non-diegetic prompt is an **explicit instruction to the LM.**

This instruction is not a part of the resulting document (e.g., travel blog).
It only serves to guide the LM's text generation.

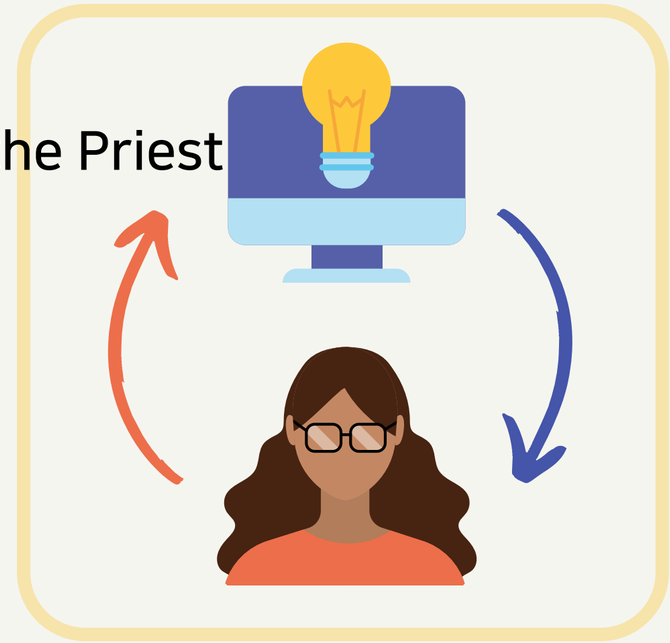
Choice Over Control: How Users Write with Large Language Models using Diegetic and Non-Diegetic Prompting (Dang et al., 2023)

Understanding model capabilities with diegetic prompts.

Story writing

The Mage, the Warrior, and the Priest

Input

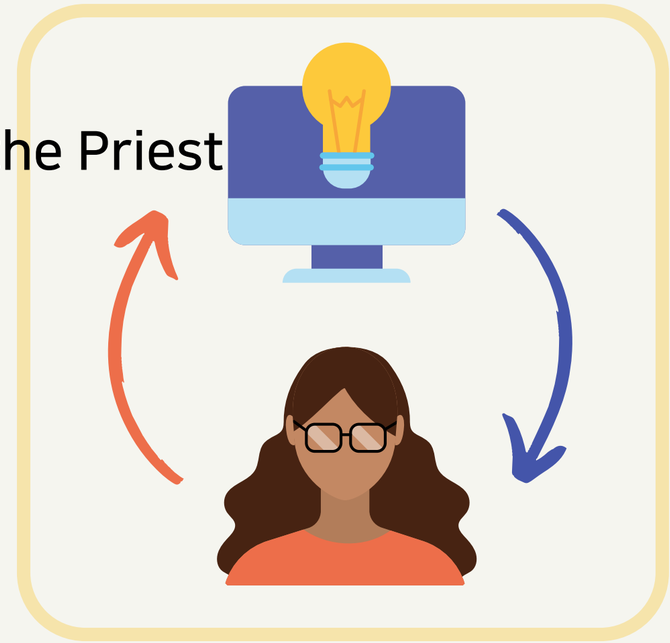


Understanding model capabilities with diegetic prompts.

Story writing

The Mage, the Warrior, and the Priest

Input



`<Write>`

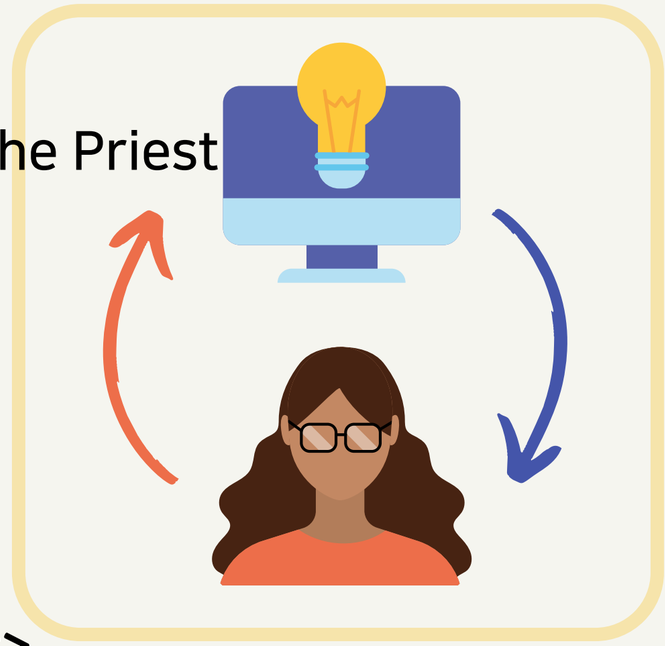
Once upon a time,

Understanding model capabilities with diegetic prompts.

Story writing

The Mage, the Warrior, and the Priest

Input



`<Write>` `<Query>`

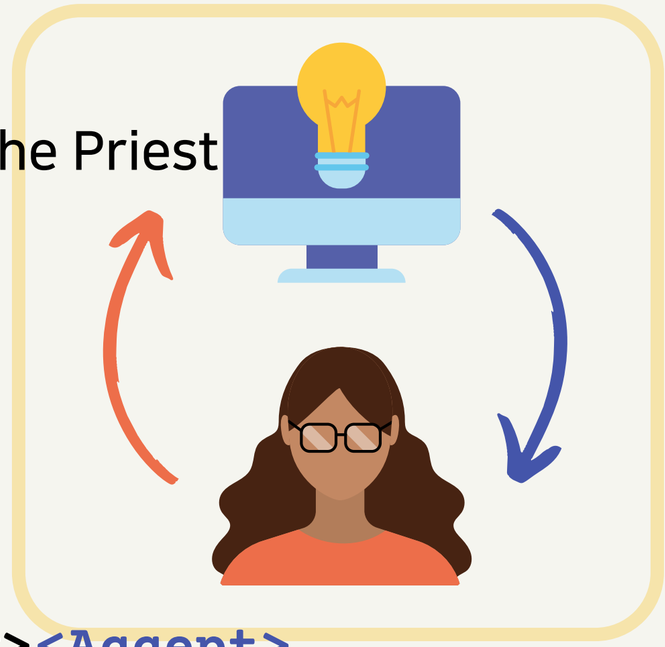
Once upon a time,

Understanding model capabilities with diegetic prompts.

Story writing

The Mage, the Warrior, and the Priest

Input



`<Write>` `<Query><Accept>`

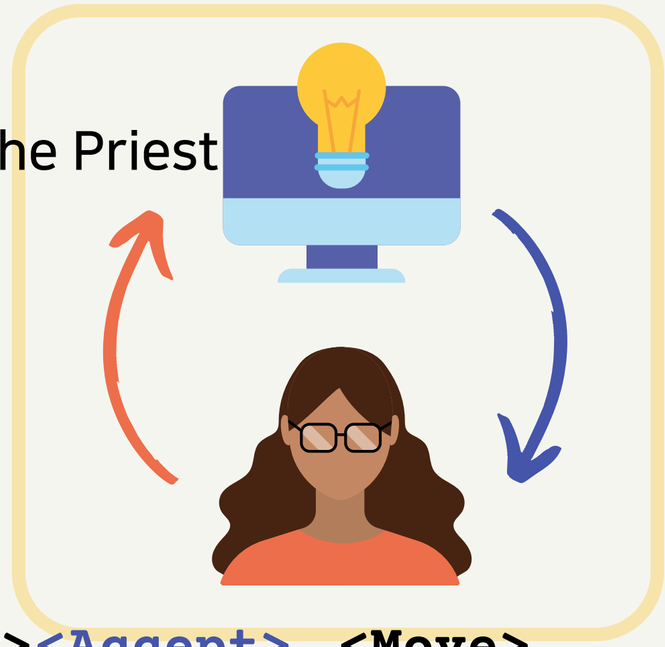
Once upon a time, there was a great mage who lived in a tower.

Understanding model capabilities with diegetic prompts.

Story writing

The Mage, the Warrior, and the Priest

Input



`<Write>` `<Query>` `<Accept>` `<Move>`

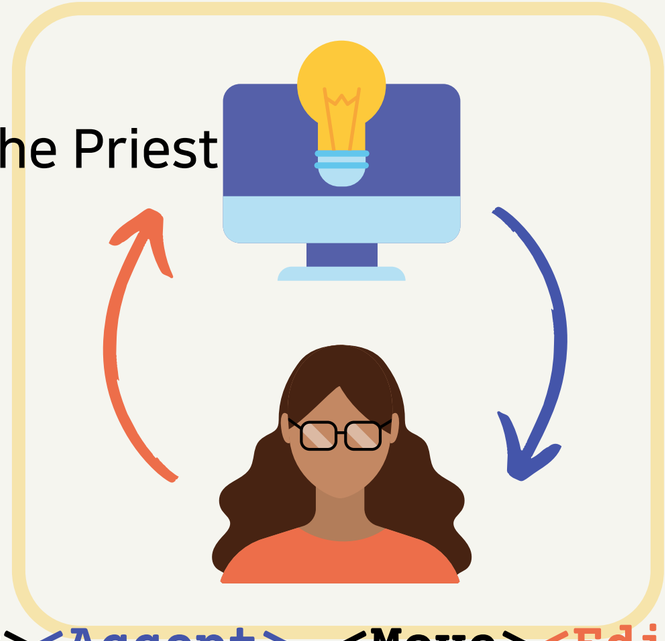
Once upon a time, there was a great mage who lived in a tower.

Understanding model capabilities with diegetic prompts.

Story writing

The Mage, the Warrior, and the Priest

Input



<Write> <Query><Accept> <Move> <Edit>

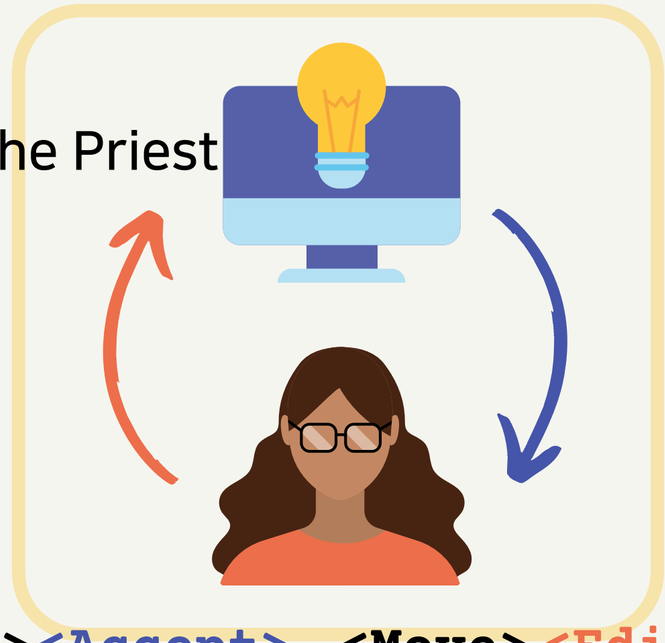
Once upon a time, there was a funny mage who lived in a tower.

Understanding model capabilities with diegetic prompts.

Story writing

The Mage, the Warrior, and the Priest

Input



`<Write>` `<Query><Accept>` `<Move>` `<Edit>` `<Move>`

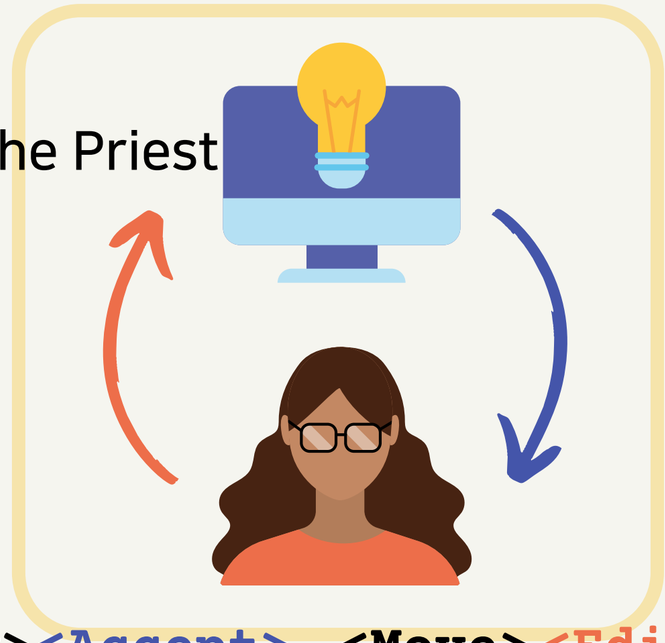
Once upon a time, there was a funny mage who lived in a tower.

Understanding model capabilities with diegetic prompts.

Story writing

The Mage, the Warrior, and the Priest

Input



`<Write>` `<Query>` `<Accept>` `<Move>` `<Edit>` `<Move>`

Once upon a time, there was a funny mage who lived in a tower.

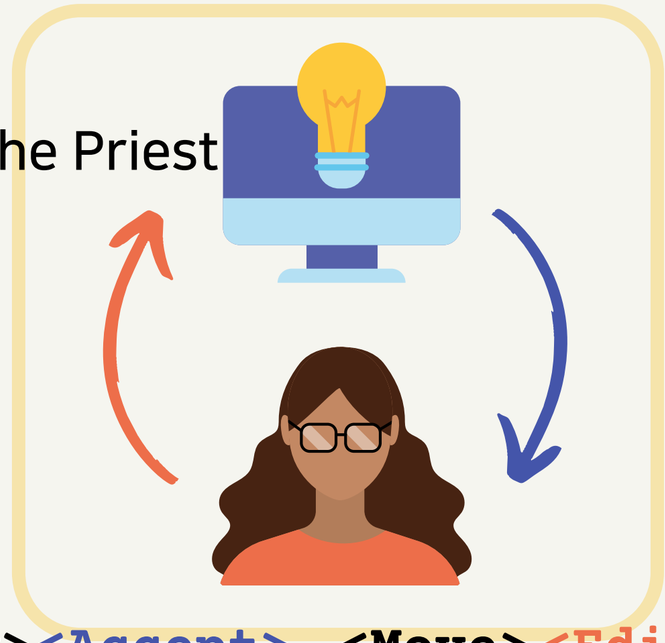
`<Query>`

Understanding model capabilities with diegetic prompts.

Story writing

The Mage, the Warrior, and the Priest

Input



<Write> <Query><Accept> <Move> <Edit> <Move>

Once upon a time, there was a funny mage who lived in a tower.

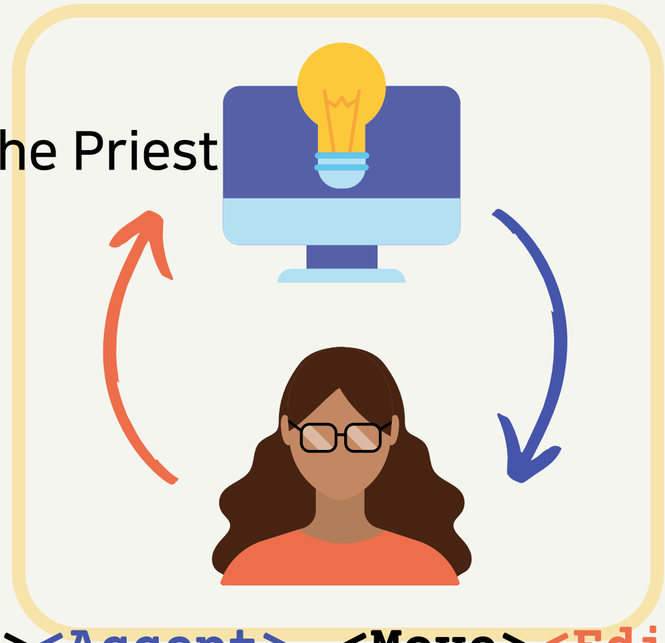
<Query><Query>

Understanding model capabilities with diegetic prompts.

Story writing

The Mage, the Warrior, and the Priest

Input



`<Write>` `<Query>` `<Accept>` `<Move>` `<Edit>` `<Move>`

Once upon a time, there was a funny mage who lived in a tower.

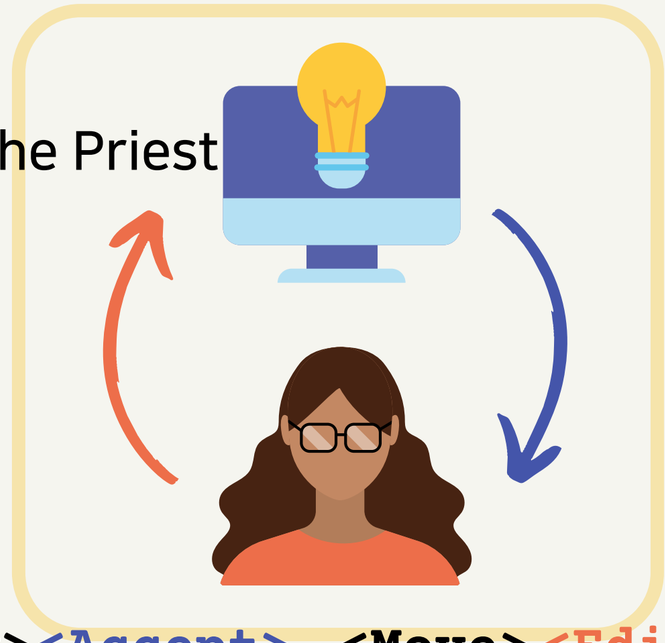
`<Query>` `<Query>` `<Query>`

Understanding model capabilities with diegetic prompts.

Story writing

The Mage, the Warrior, and the Priest

Input



`<Write>` `<Query><Accept>` `<Move>` `<Edit>` `<Move>`

Once upon a time, there was a funny mage who lived in a tower.

He loved to play tricks on people and make them laugh.

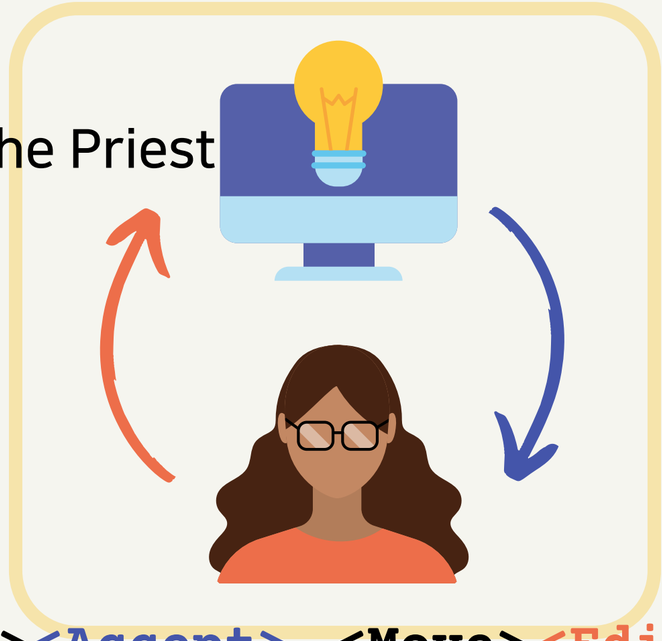
`<Query><Query><Query><Accept>`

Understanding model capabilities with diegetic prompts.

Story writing

The Mage, the Warrior, and the Priest

Input



<Write> <Query><Accept> <Move> <Edit> <Move>

Once upon a time, there was a funny mage who lived in a tower.

He loved to play tricks on people and make them laugh.

<Query><Query><Query><Accept> [...] <Write>

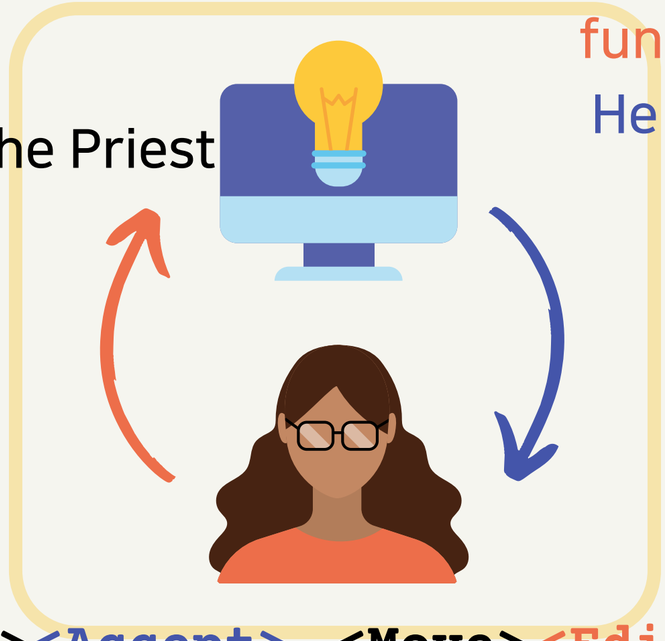
Understanding model capabilities with diegetic prompts.

Story writing

The Mage, the Warrior, and the Priest

Once upon a time, there was a funny mage who lived in a tower. He loved to play tricks on people and make them laugh. [...]

Input



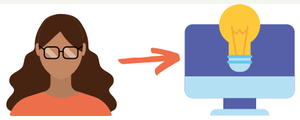
Output

`<Write>` `<Query>` `<Accept>` `<Move>` `<Edit>` `<Move>`

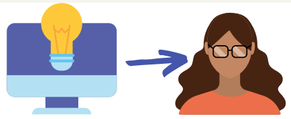
Once upon a time, there was a funny mage who lived in a tower. He loved to play tricks on people and make them laugh.

`<Query>` `<Query>` `<Query>` `<Accept>` `[...]` `<Write>`

Interaction traces can address questions about model capabilities for human interaction.



Can LMs generate fluent text in response to user text?



Can LMs contribute new ideas to users' stories?



What influences collaboration patterns the most?

<Write> <Query><Accept> <Move> <Edit> <Move>

Once upon a time, there was a funny mage who lived in a tower.

He loved to play tricks on people and make them laugh.

<Query><Query><Query><Accept> [...] <Write>

CoAuthor

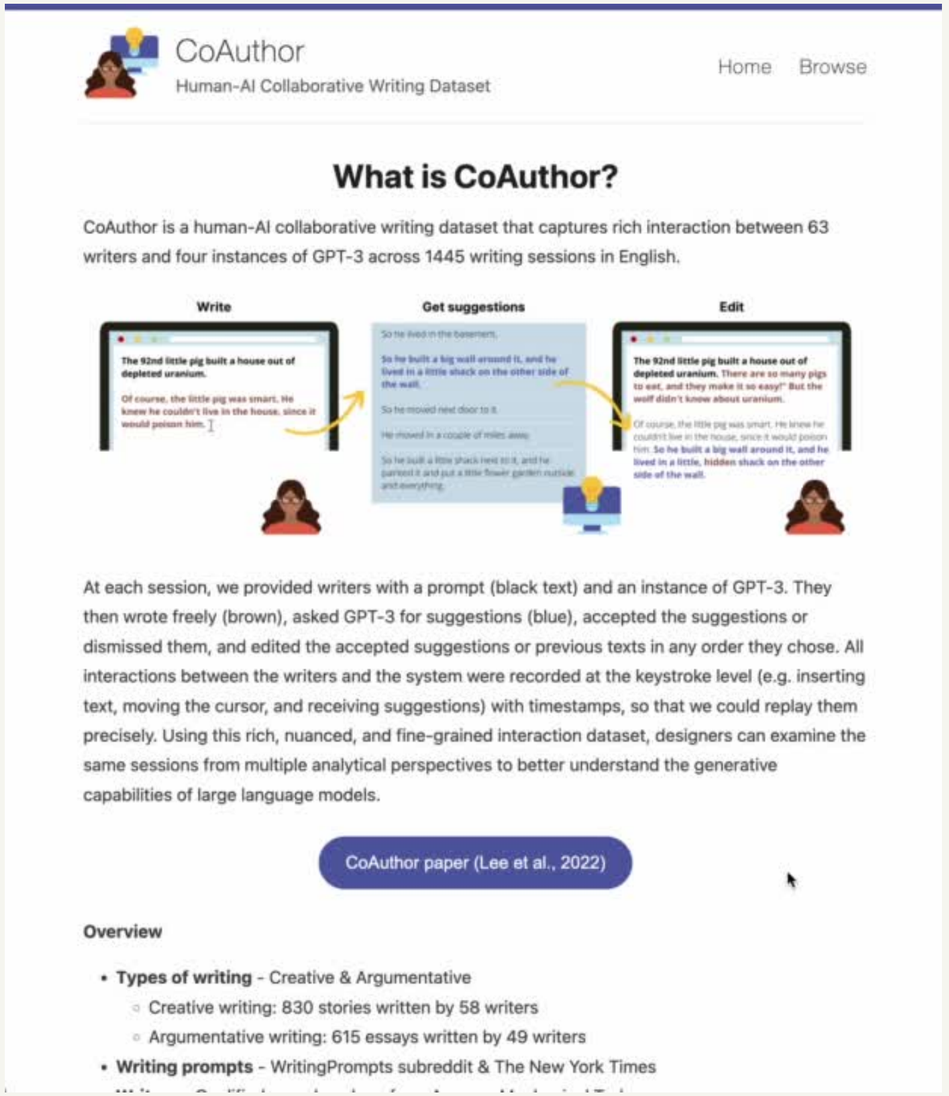


1445 sessions between **63 users** and **GPT-3**

- **Creative** and **argumentative**: 20 prompts
- Length of final texts: 418 words long
- Number of queries: 11.8 per writing session
- Acceptance rate of suggestions: 72.3%
- Percentage of text written by writers: 72.6%

https://coauthor.stanford.edu

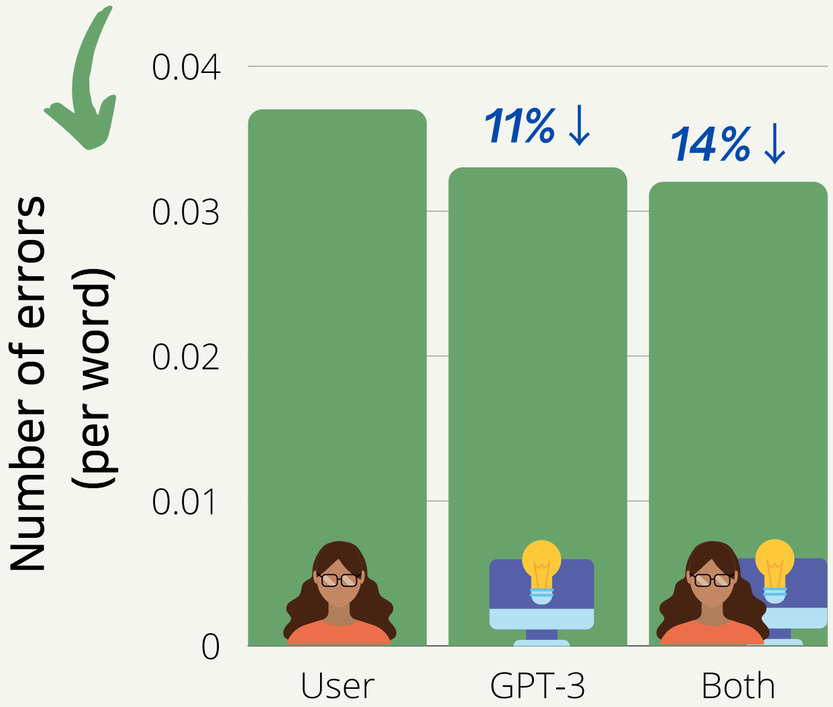
https://github.com/minalee-research/coauthor-interface

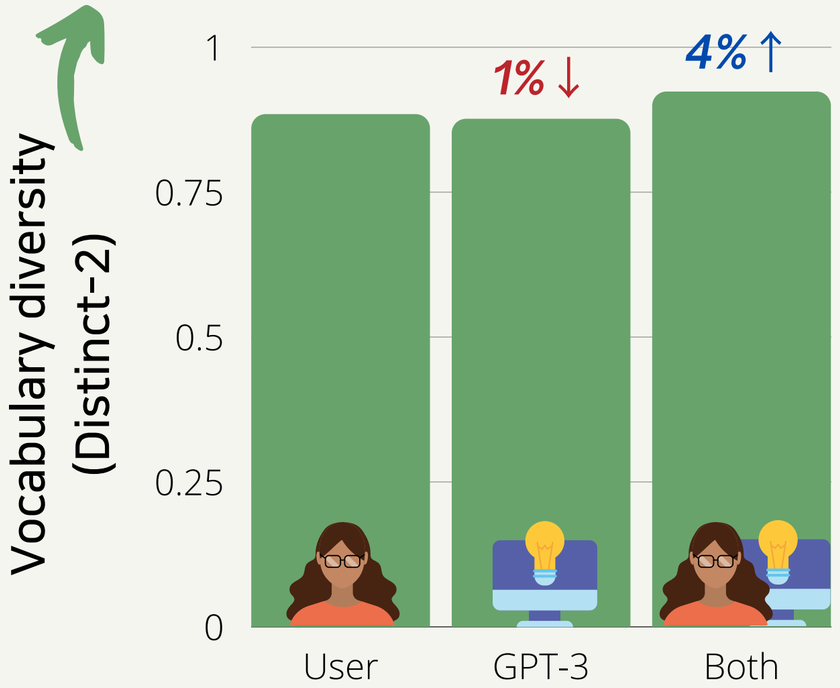


CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities (Lee et al., 2022)

Complement-ability: Language capabilities

Can GPT-3 generate **fluent text** in response to user text?



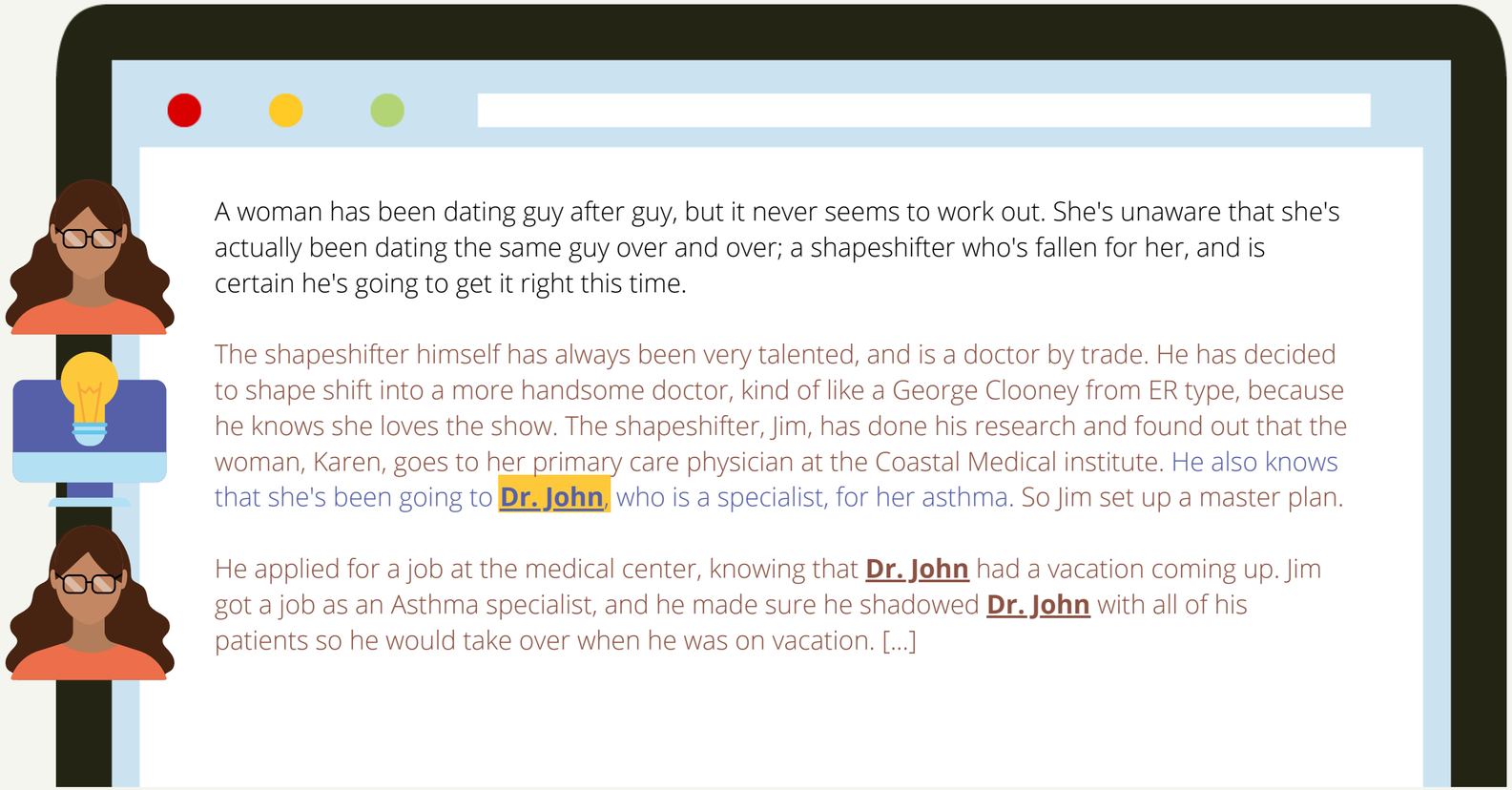


Text written by user + GPT-3 had **the fewest errors** and **most diverse vocabulary.**

CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities (Lee et al., 2022)

Complement-ability: Ideation capabilities

Can GPT-3 contribute **new ideas** to users' stories?





Suggestions contained new named entities

Users reused new named entities from suggestions

Ideas generated by GPT-3 were accepted and reused by users in the subsequent writing.

CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities (Lee et al., 2022)

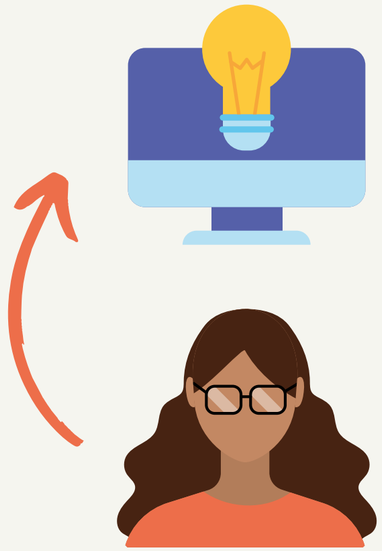
Interaction traces can also be used for non-diegetic prompts.

Interaction traces can be also used for non-diegetic prompts.

Question answering

Before Nixon resigned, how many believed he should be removed from office?
A. 79% B. 98% C. 33% D. 57%

Input



<Write>

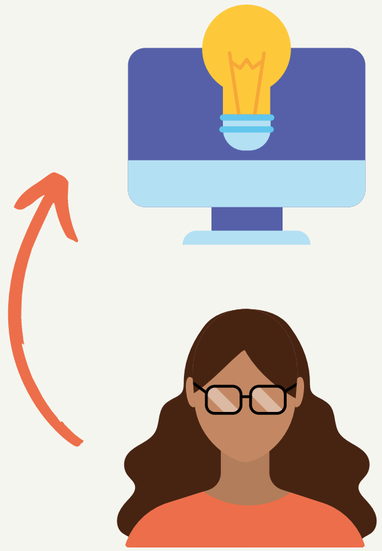
Before Nixon resigned, how many believed
he should be removed from office?

Interaction traces can be also used for non-diegetic prompts.

Question answering

Before Nixon resigned, how many believed he should be removed from office?
A. 79% B. 98% C. 33% D. 57%

Input



<Write> <Query>

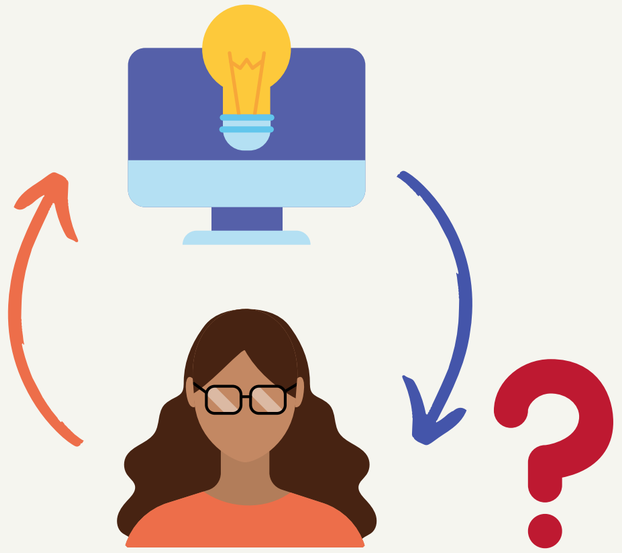
Before Nixon resigned, how many believed
he should be removed from office?

Interaction traces can be also used for non-diegetic prompts.

Question answering

Before Nixon resigned, how many believed he should be removed from office?
A. 79% B. 98% C. 33% D. 57%

Input



`<Write>` `<Query>`

Before Nixon resigned, how many believed he should be removed from office?

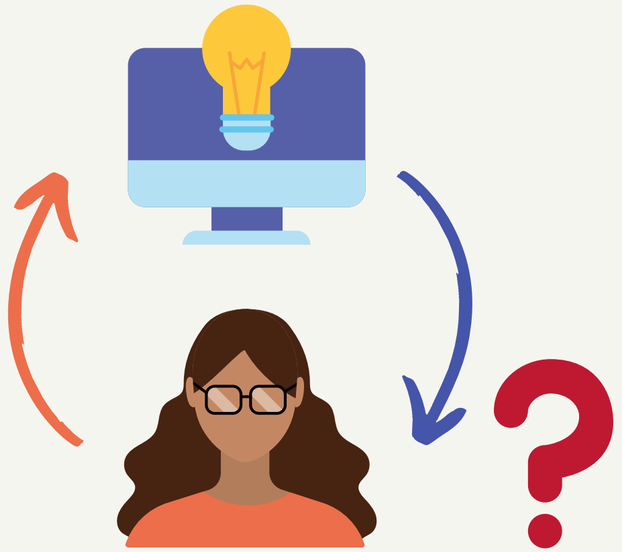
[...] **a majority of Americans** believed that he should be impeached and removed from office.

Interaction traces can be also used for non-diegetic prompts.

Question answering

Before Nixon resigned, how many believed he should be removed from office?
A. 79% B. 98% C. 33% D. 57%

Input



<Write> <Query> <Move> <Edit>

Before Nixon resigned, how many believed he should be removed from office?

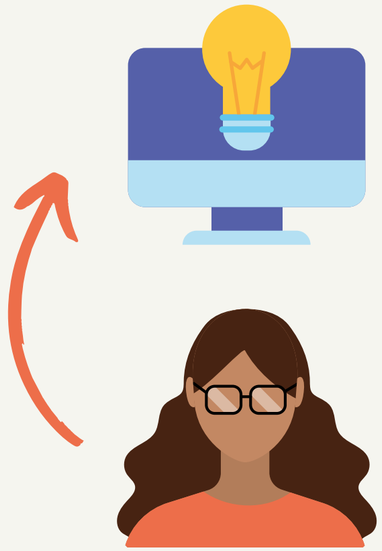
[...] **a majority of Americans** believed that he should be impeached and removed from office.

Interaction traces can be also used for non-diegetic prompts.

Question answering

Before Nixon resigned, how many believed he should be removed from office?
A. 79% B. 98% C. 33% D. 57%

Input



<Write> <Query> <Move> <Edit>

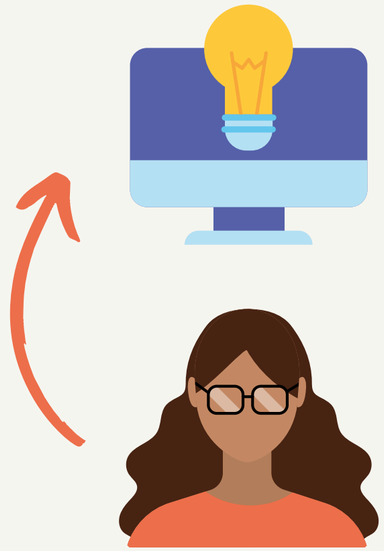
What **percentage** of people believed
Nixon should have been removed from office?

Interaction traces can be also used for non-diegetic prompts.

Question answering

Before Nixon resigned, how many believed he should be removed from office?
A. 79% B. 98% C. 33% D. 57%

Input



<Write> <Query> <Move> <Edit> <Query>

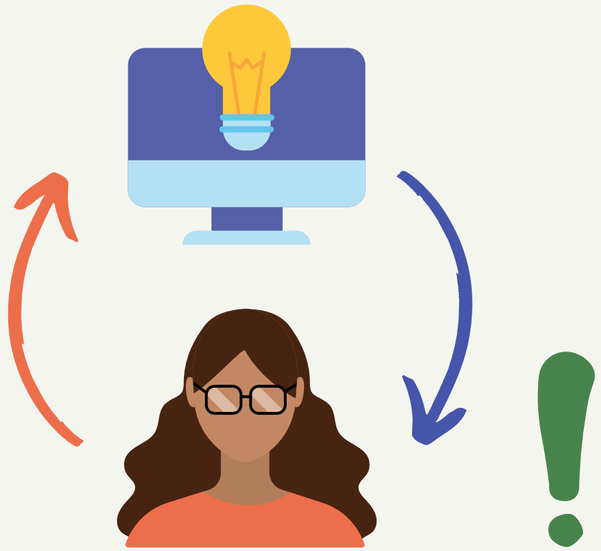
What **percentage** of people believed Nixon should have been removed from office?

Interaction traces can be also used for non-diegetic prompts.

Question answering

Before Nixon resigned, how many believed he should be removed from office?
A. 79% B. 98% C. 33% D. 57%

Input



<Write> <Query> <Move> <Edit> <Query>

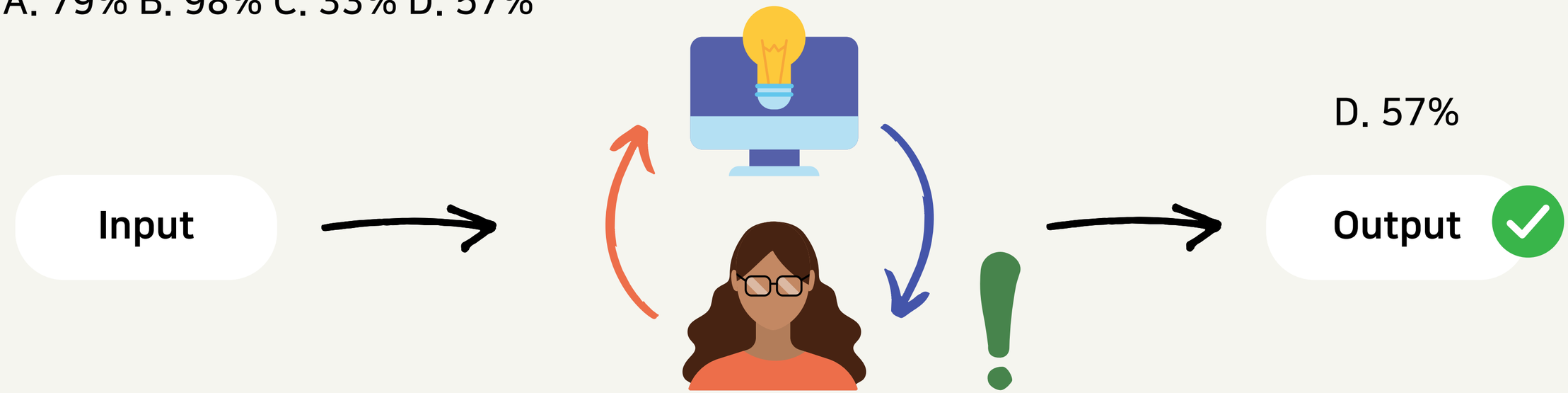
What **percentage** of people believed Nixon should have been removed from office?

Approximately 60% of Americans believed that President Nixon should have been removed [...]

Interaction traces can be also used for non-diegetic prompts.

Question answering

Before Nixon resigned, how many believed he should be removed from office?
A. 79% B. 98% C. 33% D. 57%



`<Write>` `<Query>` `<Move>` `<Edit>` `<Query>`

What **percentage** of people believed Nixon should have been removed from office?

Approximately 60% of Americans believed that President Nixon should have been removed [...]

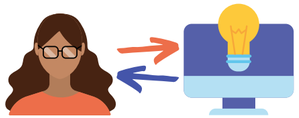
Interaction traces can address questions about **model** capabilities for **human** interaction.



Was the user able to find the correct answer with the LM?



How many times did the user query the LM?



Was the model able to adapt to user behaviors?

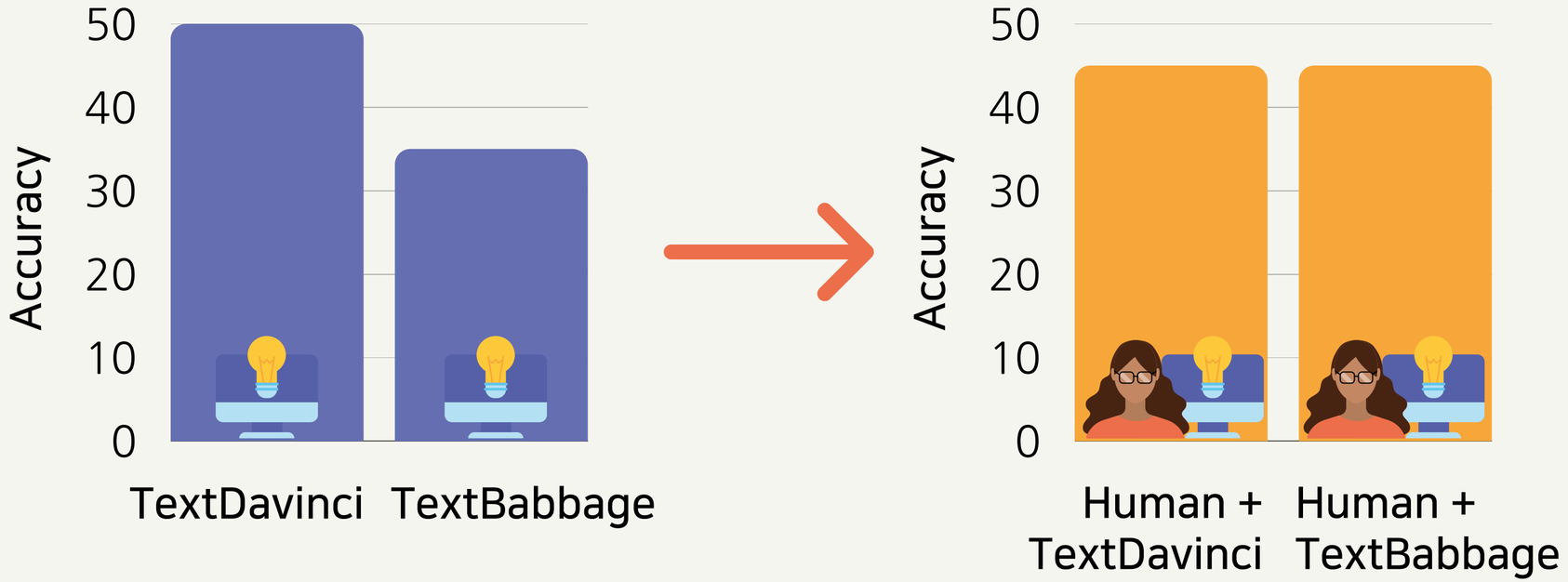


`<Write>` `<Query>` `<Move>` `<Edit>` `<Query>`

What percentage of people believed Nixon should have been removed from office?

Approximately 60% of Americans believed that President Nixon should have been removed [...]

Intuitiveness: How easy is it to use the model?

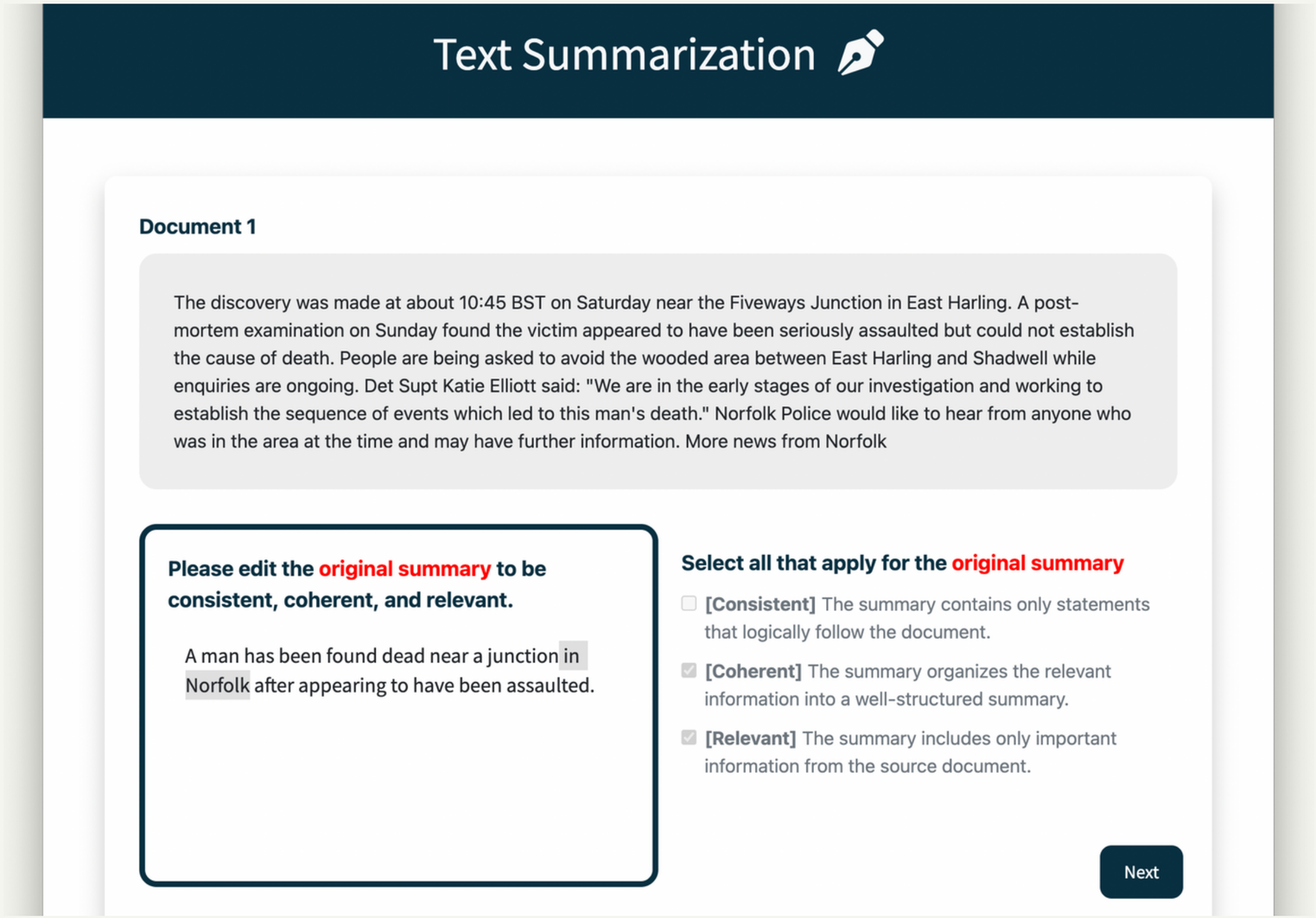


A model with **worse zero-shot performance** (e.g. TextBabbage) resulted in **comparable human-LM performance**.

Results on Question Answering (Nutrition category)

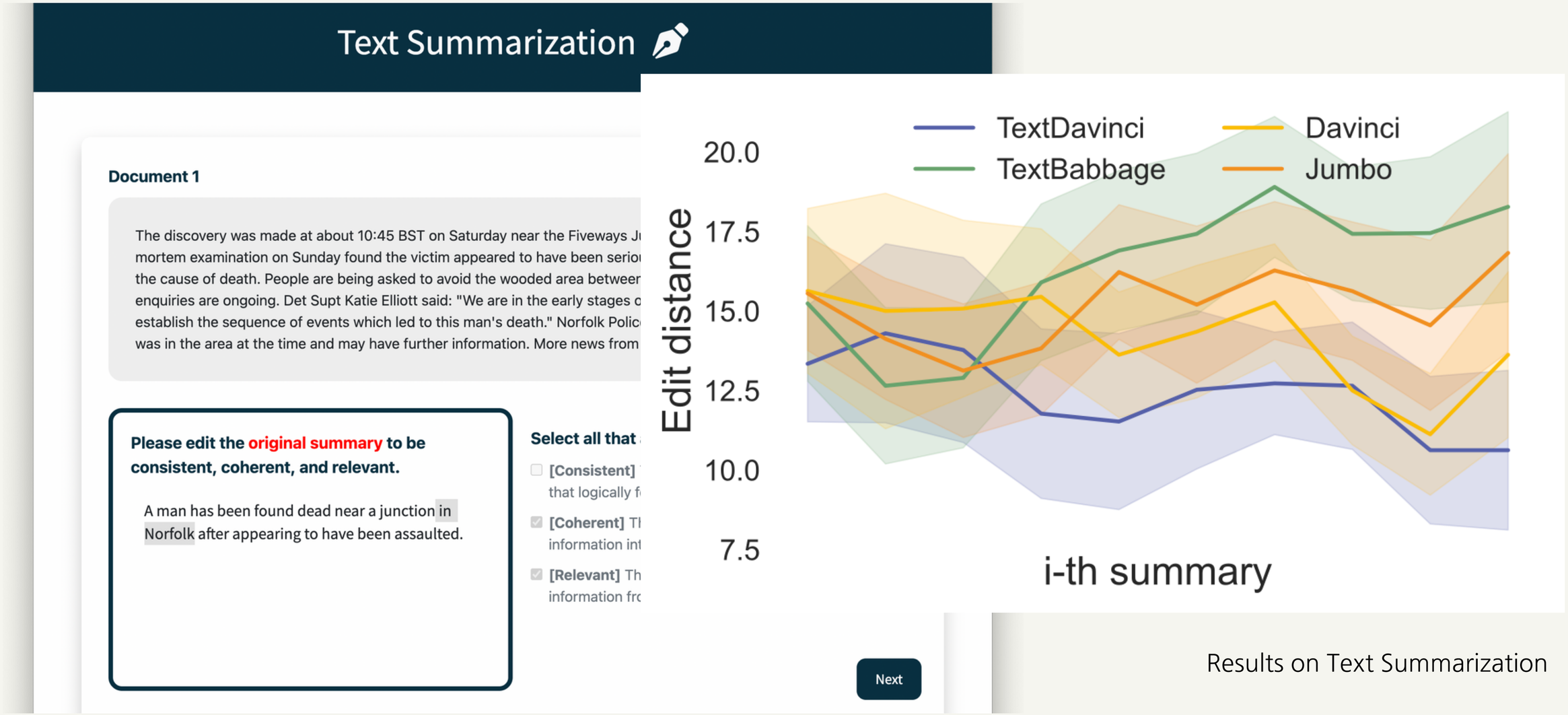
Evaluating Human-Language Model Interaction (Lee et al., 2022)

Teachability: How quickly can the model learn from humans?



Evaluating Human-Language Model Interaction (Lee et al., 2022)

Teachability: How quickly can the model learn from humans?



Results on Text Summarization

Evaluating Human-Language Model Interaction (Lee et al., 2022)

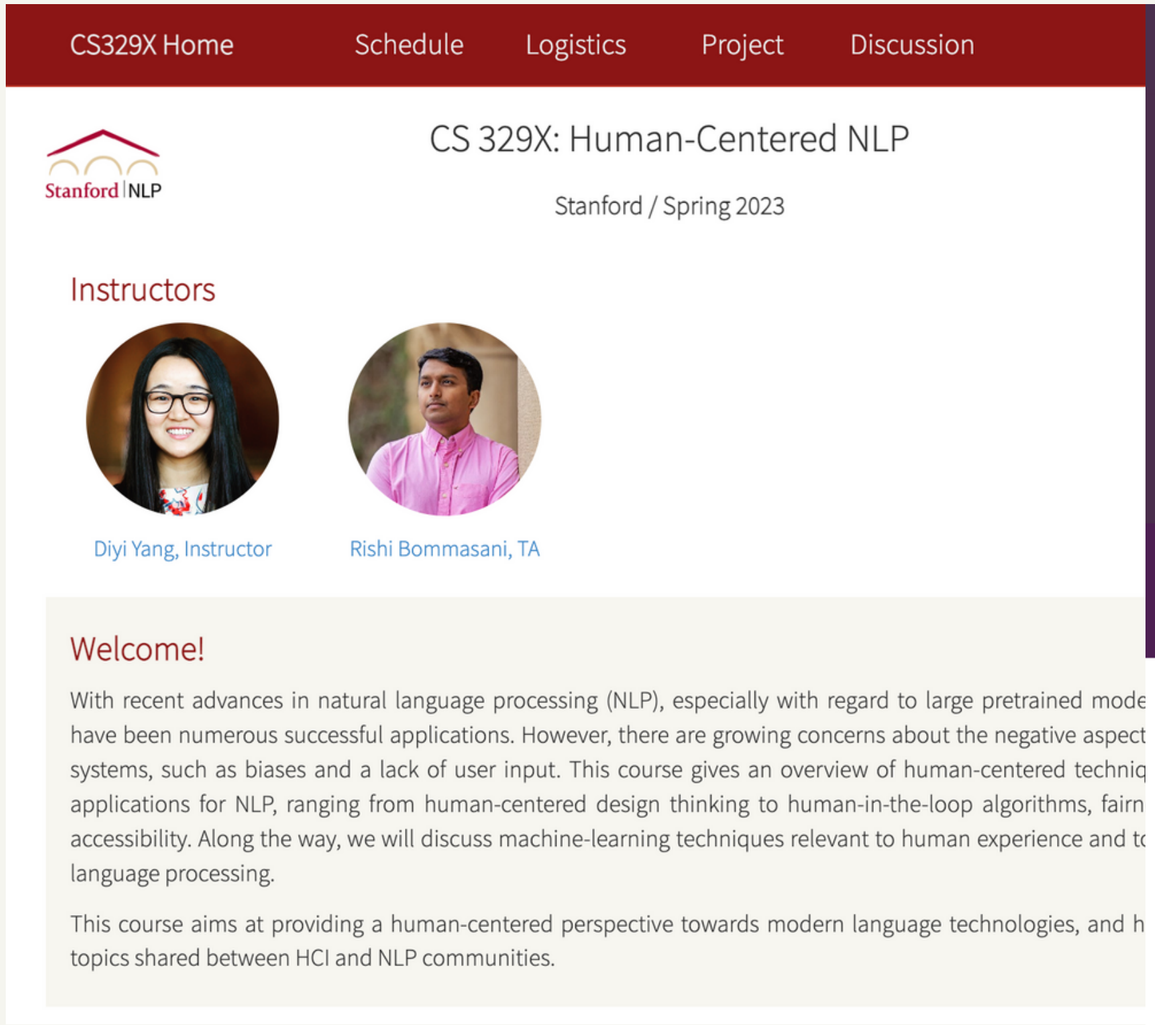
There is a mismatch
between prompts constructed for models
and prompts written by humans in real life.

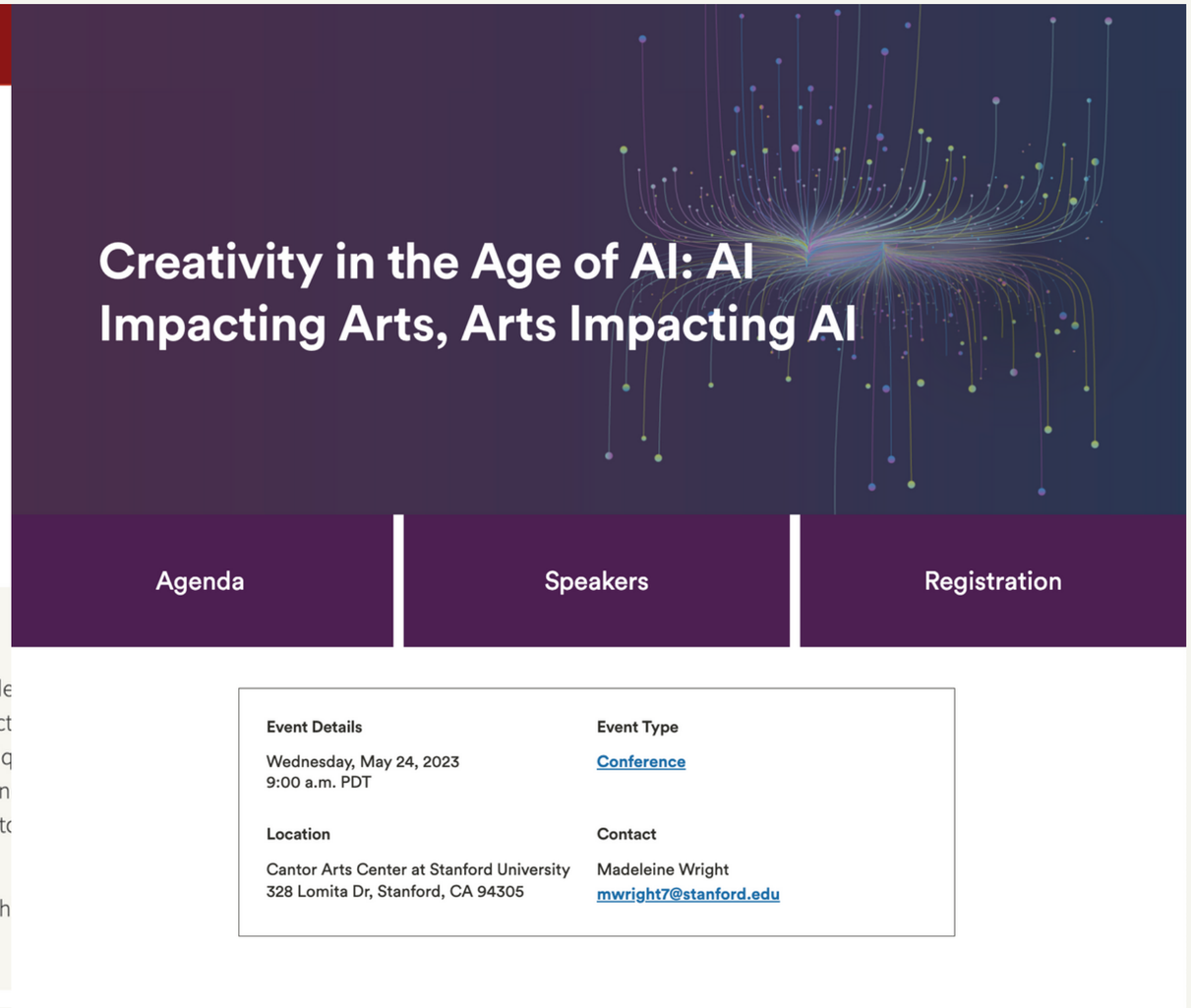
Right now, humans try to construct prompts for models.

It should be the other way around,
where models are trained and evaluated to
"understand" prompts written by humans.

We can use interaction traces to gain insights of
human-LM interaction (e.g., via prompts).

Want to learn more about human-centered approaches?





Stanford CS329X Human-Centered NLP (http://web.stanford.edu/class/cs329x/)
Creativity in the Age of AI: AI Impacting Arts, Arts Impacting AI (https://hai.stanford.edu/events/creativity-age-ai-ai-impacting-arts-arts-impacting-ai)

Prompters before prompts and promptees

Mina Lee

Stanford University

minalee@stanford.edu
minalee.info

