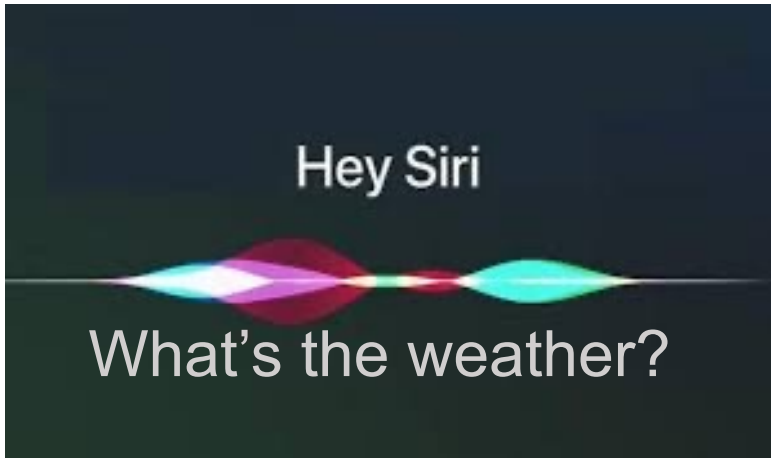
Situated Natural Language Explanations

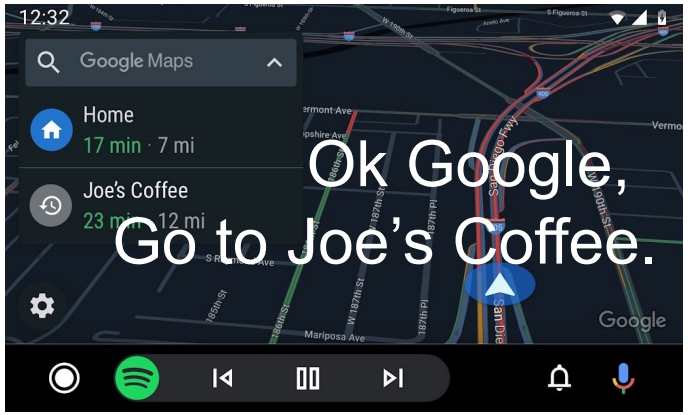
Zining Zhu, Haoming Jiang, Jingfeng Yang, Sreyashi Nag, Chao Zhang, Jie Huang, Yifan Gao, Frank Rudzicz, Bing Yin

AI technologies are prevalent

Speech recognition

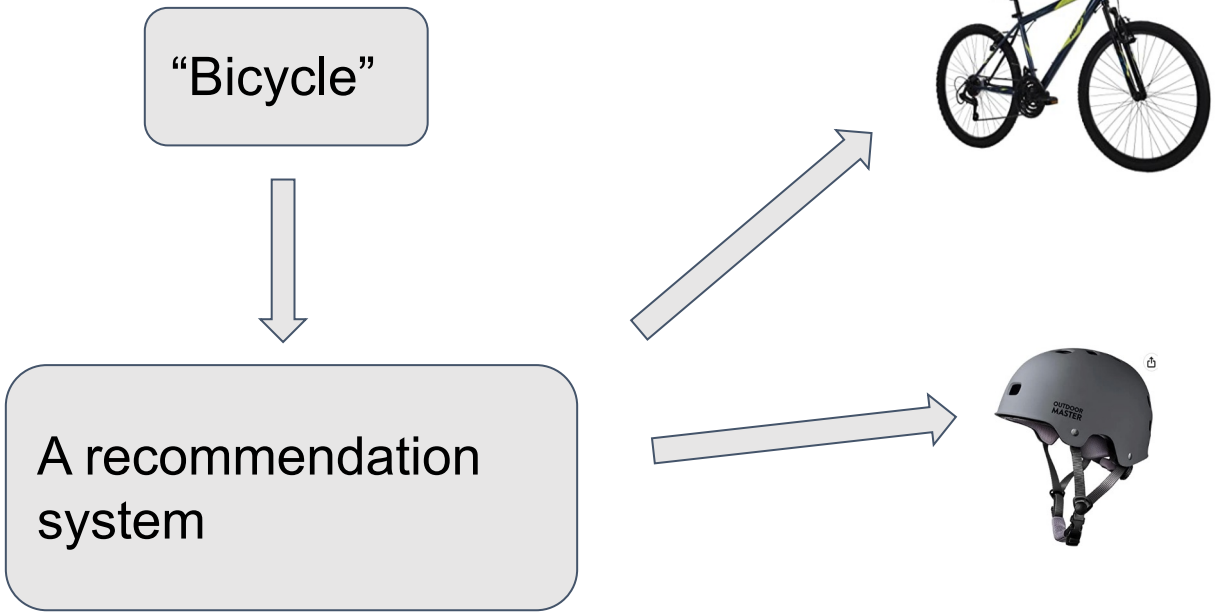


Smart assistants



Search engine

Some behaviors can be mysterious



Huffy Hardtail Mountain Bike, Stone Mountain 26 inch, 21-Speed, Lightweight, Dark Blue

Why?

OutdoorMaster Skateboard Cycling Helmet - Two Removable Liners Ventilation Multi-Sport Scooter Roller Skate Inline Skating Rollerblading for Kids, Youth & Adults

Why?

Let’s explain the results

Search term: “Bicycle”

Because this helmet has MIPS technology, which makes the ride safe and comfortable.

Because a helmet makes riding bicycles safer.

Because California requires all riders below 18 to legally wear a helmet, and riders above 18 to legally wear a helmet in regions including Bidwell Park, Chico, California.

Which one is more preferred?

Imagine two customers

I searched for “bike”. Why does the app recommend a helmet?

Because wearing a helmet makes bike riding safe.

🤦

Because the app searched in a database for the product with the highest vector similarity to the query, and found this helmet.

🤦

Good explanations are situated

I searched for “bike”. Why does the app recommend a helmet?

Because the app searched in a database for the product with the highest vector similarity to the query, and found this helmet.

😀

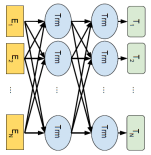
Because wearing a helmet makes bike riding safe.

😀

Prompt LLMs to generate explanations

Language Model

Prompt



Explanation

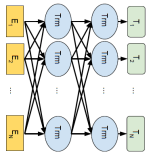
The sky is dark *because*

the sun has gone down.

Adapt explanations towards the situations

Language Model

Situated prompt



Situated explanation

I care about the details.
The sky is dark *because*

it is nighttime. The dark sky is caused by the rotation of the Earth, which causes the sun, which is the source of light, to set below the horizon and thus make the sky dark.

A pipeline for situated explanation

Step 1
Identify a desired property of explanation in the situation.

Explain to 

Shorter is better.

Step 2
Prompt a PLM using a collection of *situated* prompts.

[…] because, in short
[…] because, basically
[…] because, essentially
I'm busy. […] because

Step 3
Select the generated explanations using the identified property.

[…] because, in short, wearing a helmet makes bike riding safe.

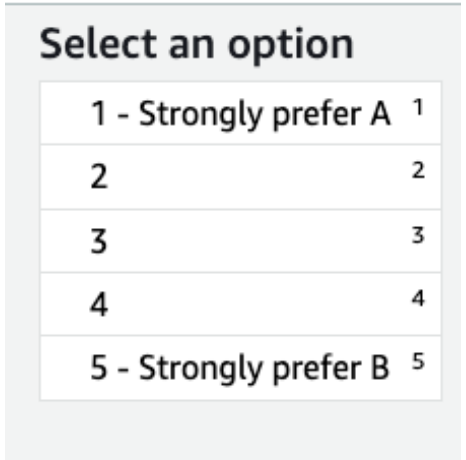
Humans prefer the situated explanations

I searched for “bike” and was recommended a helmet.

Which explanation better explains the recommendation?

Explanation A is generated with the hint “in short” in the prompt.

Explanation B is generated with the hint “in detail” in the prompt.



Try to finish this question in 20 seconds.

Preference: $2.98\ (\mathrm{sd} = 0.74)$

Take your time and read the texts carefully.

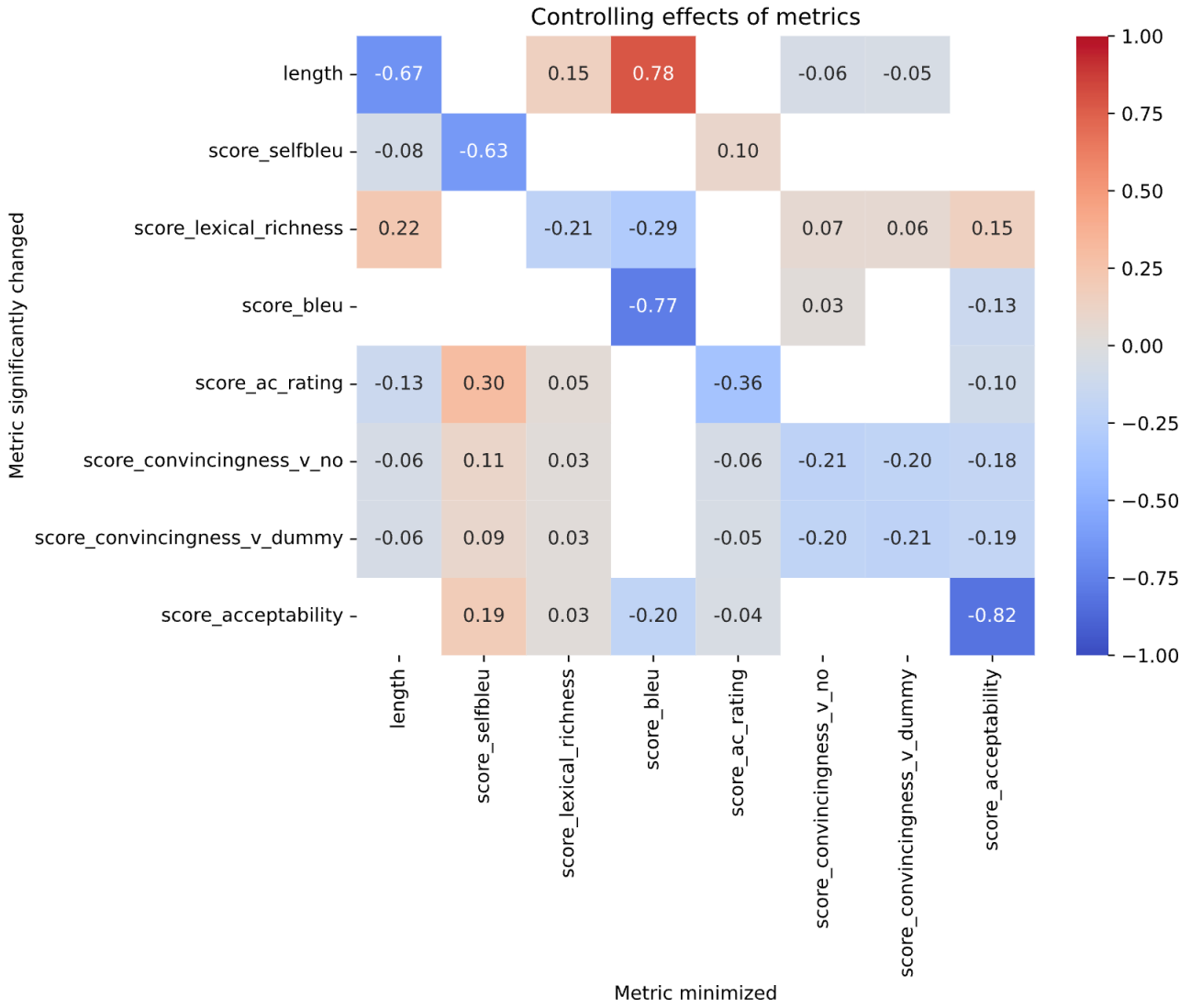
Preference: $3.10\ (\mathrm{sd} = 0.73)$

$p = 0.038^{*}$

What are the situated controlling effects?

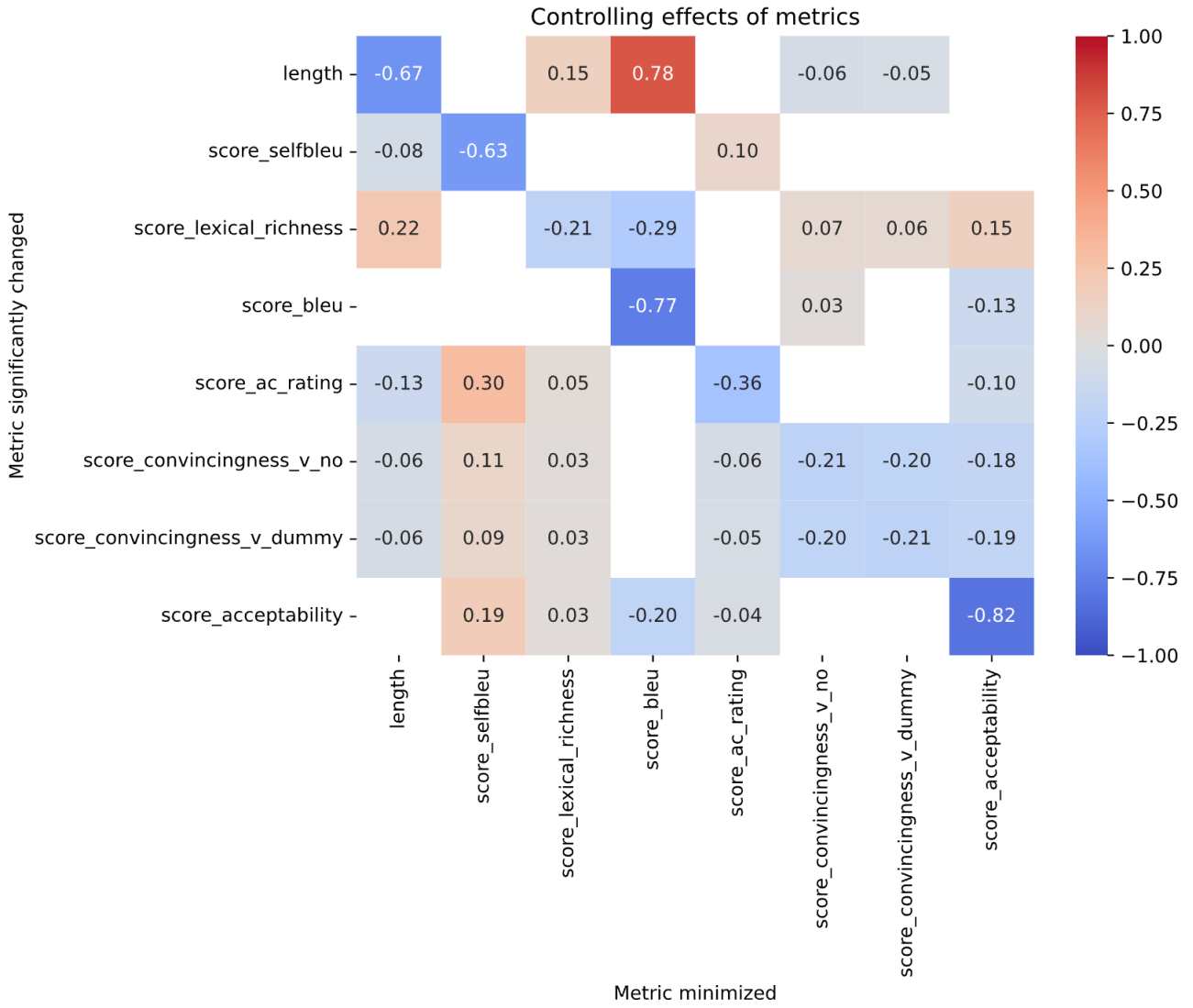
We compute scores in the following categories:

- Lexical
  - Length, self-BLEU, lexical richness
- Semantic
  - BLEU, abstract/concreteness rating
- Pragmatic
  - Convincingness (vs no explanation)
  - Convincingness (vs “it is what it is”)
  - Acceptability



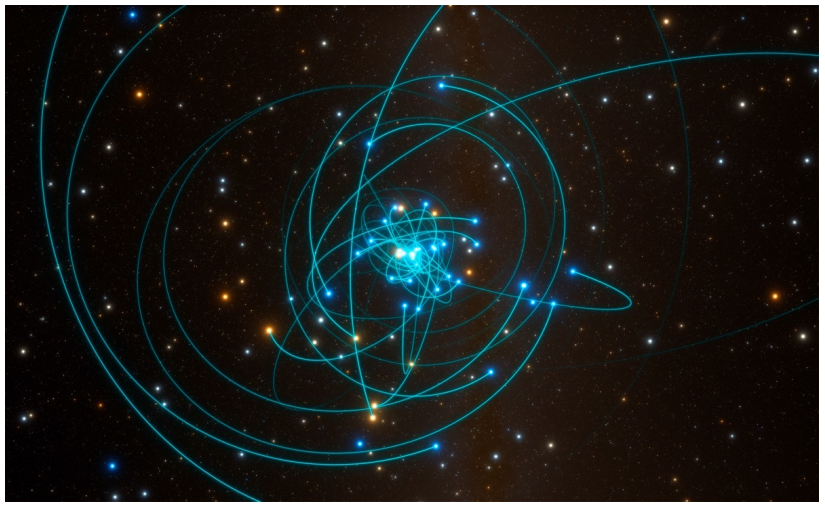
What are the situated controlling effects?

- Example: busy customer.
- Want short explanations.
  - Minimize the length.
  - Look at the first column.
- Query the effects from the figure.
  - The lexical richness score increased by 22%.
  - compared to the baseline.
- “Shorter explanations contain more unique words”.



Explanations can make systems trustworthy

Complex systems

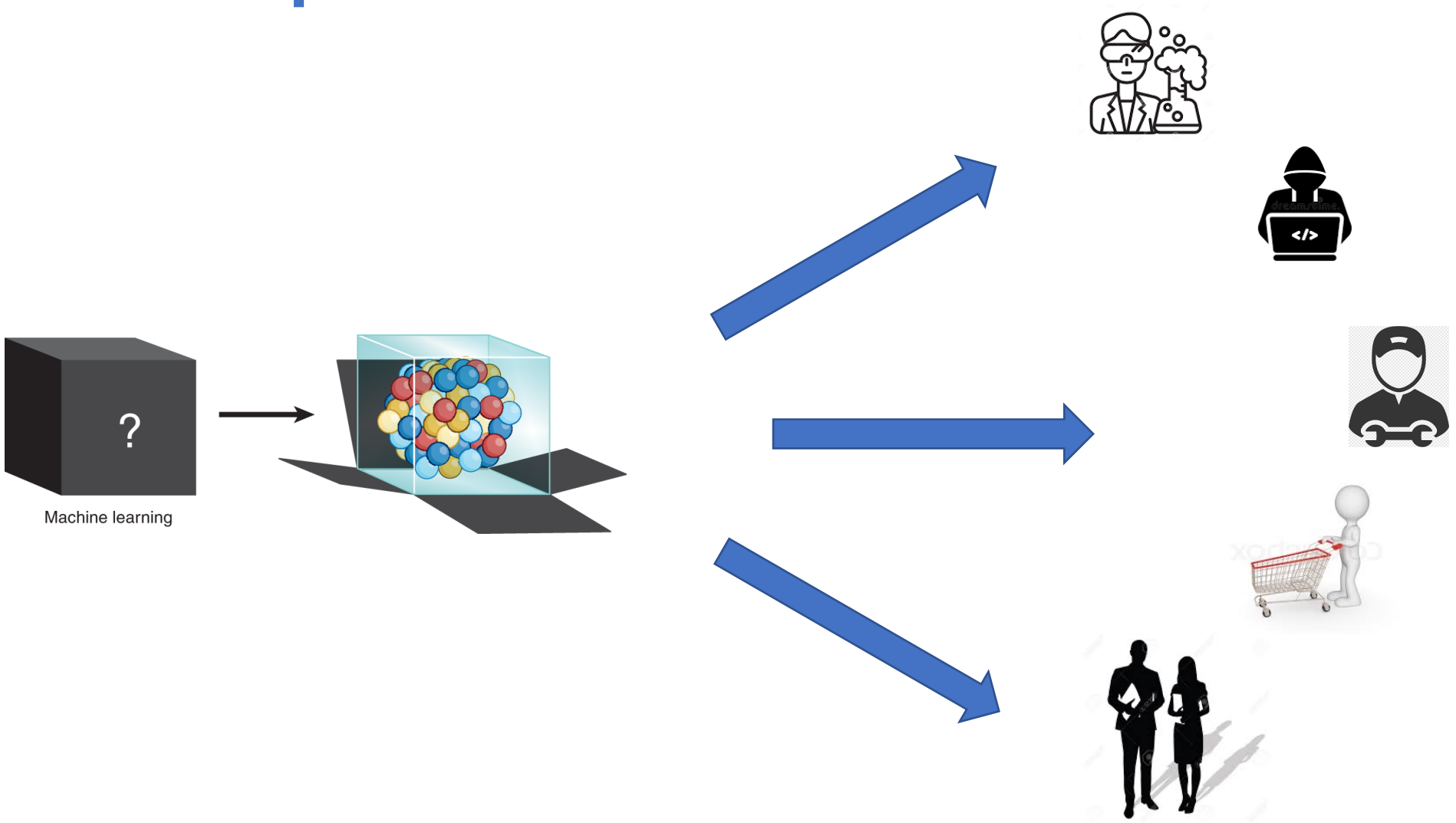


because

Regularities

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = 8\pi G T_{\mu\nu}$$

Future of explanations



Summary

- Prompting techniques can produce high-quality explanations.
- Situated prompts can create suitable explanations.