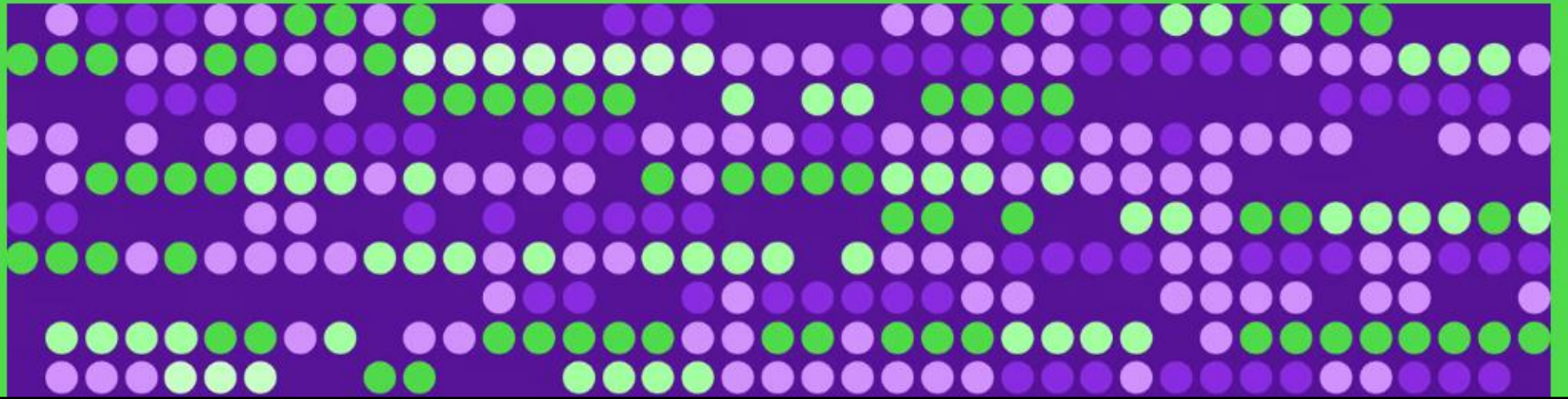




用 AI 來解釋 AI

Language models can explain neurons in language models



Blog: <https://openai.com/research/language-models-can-explain-neurons-in-language-models>

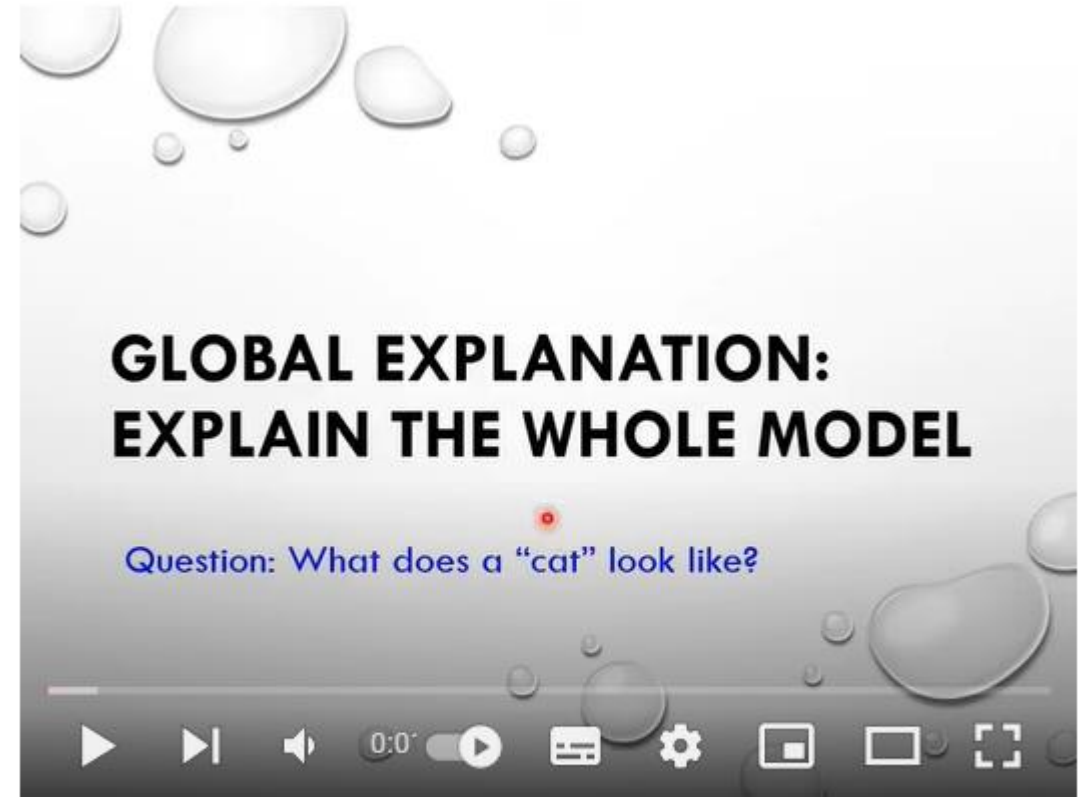
Paper: <https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html>

可解釋的 AI



【機器學習2021】機器學習模型的可解釋性
(Explainable ML) (上) - 為什麼類神經網路可以正確分...

<https://youtu.be/WQY85vaQfTI>

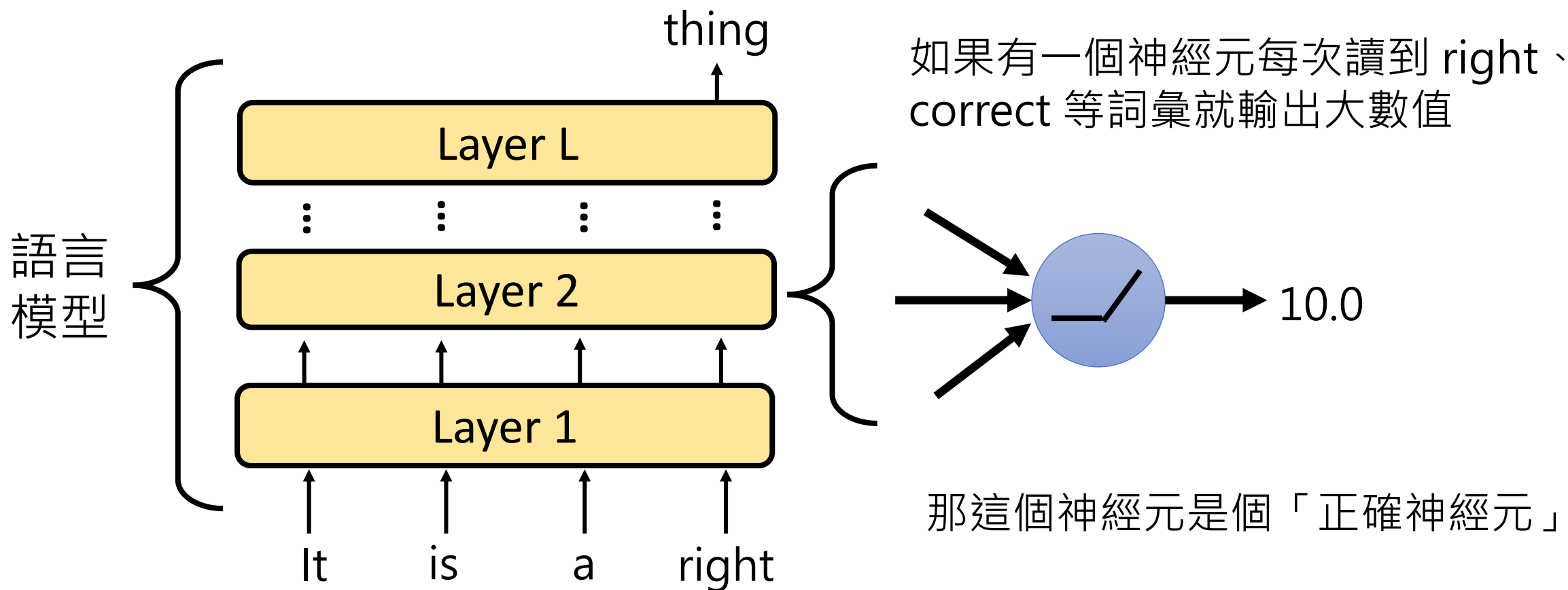


【機器學習2021】機器學習模型的可解釋性
(Explainable ML) (下) - 機器心中的貓長什麼樣子？

<https://youtu.be/0aylPqbdHYQ>

知道一個神經元 (Neuron) 的作用

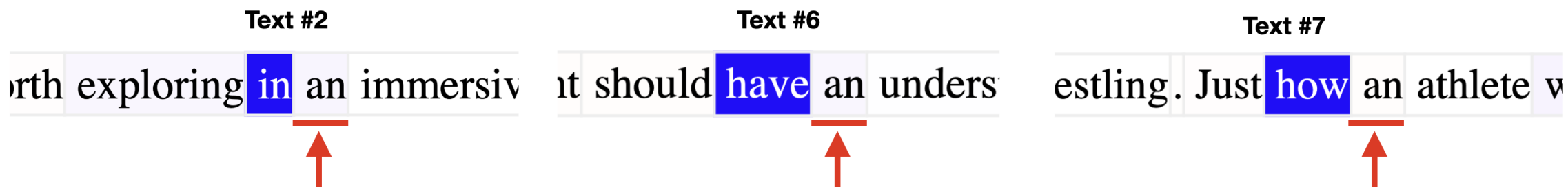
根據神經元何時起反應 (Activate) ，可以猜測它的作用



We Found An Neuron in GPT-2

Source of image: <https://clementneo.com/posts/2023/02/11/we-found-an-neuron>

Layer 31 Neuron 892



有些神經元的行為比較複雜，交給 GPT4 解釋

GPT4 如何解釋神經元的行為呢？

to meditate on emptiness, the ultimate nature of reality, so that we experience the union of bliss and emptiness or Mahamudra. This mind of the clear light of bliss and its main object emptiness mix together like water mixed with water, they go together very well, they belong together. In fact, once you

cy Jones Taught Snoop How To Make Hit Records" It's an instinct. It's like a kid that rides a bike and then he ain't rode a bike in 20 years and you give it back to him. He's gonna master that shit. It's a technique. We have an instinct about

請問你覺得這個神經元的功用是甚麼？

這個神經元的功用是尋找





Neuron 1

Activations:

```

<start>
the 0
sense 0
of 0
together 3
ness 7
in 0
our 0
town 1
is 0
strong 0
. 0
<end>

```

neurons in a neural network. Each neuron looks for some particular thing in a document. Look at the parts of the document the neuron activates for and summarize what the neuron is looking for. Don't list examples of words.


Same activations, but with all zeros filtered out:

```

<start>
together 3
ness 7
town 1
<end>

```

Explanation of neuron 1 behavior: the main thing this neuron does is find

 instances of community or unity related concepts in the text.

成功的例子

but with a statistical significance of about 65 percent rather than the usual 95 percent standard. That's because, when the mathematics is done properly, it shows that 15 years is almost always too short a time interval to make meaningful conclusions about climate. Temperature is sticky, correlated to itself; a warm year is more likely

do, if not at least one of the better options ... if it's done right. (Take whatever stance you like on the hit by Nicklas Backstrom - I think it's probably a penalty, but given the situation can see it going uncalled - but we're not here to debate that. Moving

請問你覺得這個神經元的功用是甚麼？

這個神經元的功用是尋找 *words and phrases related to performing actions correctly or properly.*



成功的例子

to meditate on emptiness, the ultimate nature of reality, so that we experience the union of bliss and emptiness or Mahamudra. This mind of the clear light of bliss and its main object emptiness mix together like water mixed with water, they go together very well, they belong together. In fact, once you

cy Jones Taught Snoop How To Make Hit Records" It's an instinct. It's like a kid that rides a bike and then he ain't rode a bike in 20 years and you give it back to him. He's gonna master that shit. It's a technique. We have an instinct about

請問你覺得這個神經元的功用是甚麼？

這個神經元的功用是尋找 *descriptive comparisons, especially similes* (明喻).



成功的例子

From the side	Advertisement
From the front	Advertisement
From behind	Advertisement
From the side	Advertisement
From the front	Advertisement

Map02, Map03, Map04, Map05, Map06, Map07, Map08, Map09 Single Player : Designed for Cooperative 2-4 Player : No Other game styles : Difficulty Settings : Yes * Construction * Base : New from

From the front	Advertisement
From behind	Advertisement
From the side	Advertisement
Waste of a shell	Advertisement

這個神經元的功用是尋找 *repetitions of a similar word or an evolving sequence of words.*





失敗的例子

人類答案：前面有錯字

the next starbound update. View On reddit.com submitted 1 year ago by nikolai_8 posted in /r/starbound

This poll has some of my ideas and there is other things. if you think I missed something big please comment. to see the live vote tally look at the

need to focus on incorporating people as people; not having to look for what we didn't acknowledge about their culture. Articles like the just add to the general belief that people play the culture card or race card to get attention instead of seeing that their culture was incorporated into something cool and fun. So over this bull!

the Klintholm Klint are suggested to constitute   ne GeoSite together.

26. Supplementary description:

Denudation processes at the "Store Taler" section of the cliff ("Før" means before, "Nu" means now): Photo: Ole Bang (Camp

這個神經元的功用是尋找 *words related to general concepts, titles, and partial terms.*



失敗的例子

人類答案：規律被破壞

1 2 3 4 5 red

Algeria Malaysia Venezuela Bhutan Luxury

Red 21, Blue 33, Orange 18, Green truck

Table Table Table Table Italics

like it, love it, hate it, try it, when does it end?

alpha 1, alpha 2, alpha 3, alpha 5

1 Washington, 2 Adams, 3 Jefferson, 4 Madison, 5
ebullient

bravo 2, bravo 3, bravo 5, bravo 7, bravo 11, bravo 12

這個神經元的功用是尋找 *numbers, ordinal terms, and possessive constructions.*



怎麼知道 GPT4 如何解釋的好不好呢？

叫 GPT4 根據自己的解釋扮演神經元

神經元 #1938 的功用是尋找 *words and phrases related to performing actions correctly or properly.*

給一個句子 if their applications are executed properly

請問神經元 #1938 讀到最後一個 token 會輸出多少數值？

9



怎麼知道 GPT4 如何解釋的好不好呢？

神經元 #1938 的功用是尋找 words and phrases related to performing actions correctly or properly.

給一個句子 if their applications are executed properly

請問神經元 #1938 讀到最後一個 token 會輸出多少數值？

解釋精確

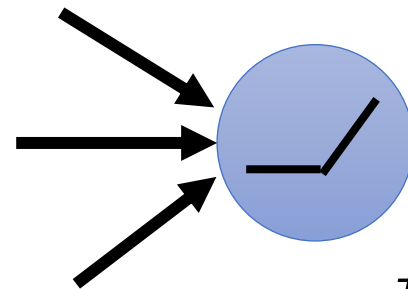
9



相近

if their applications are executed properly

GPT2



#1938

?

Explanation Score (0 ~ 1)



We're studying neurons in a neural network. Each neuron looks for some particular thing in a short document. Look at an explanation of what the neuron does, and try to predict its activations on a particular token.

The activation format is token<tab>activation, and activations range from 0 to 10. Most activations will be 0.

Neuron 4

Explanation of neuron 4 behavior: the main thing this neuron does is find present tense verbs ending in 'ing'

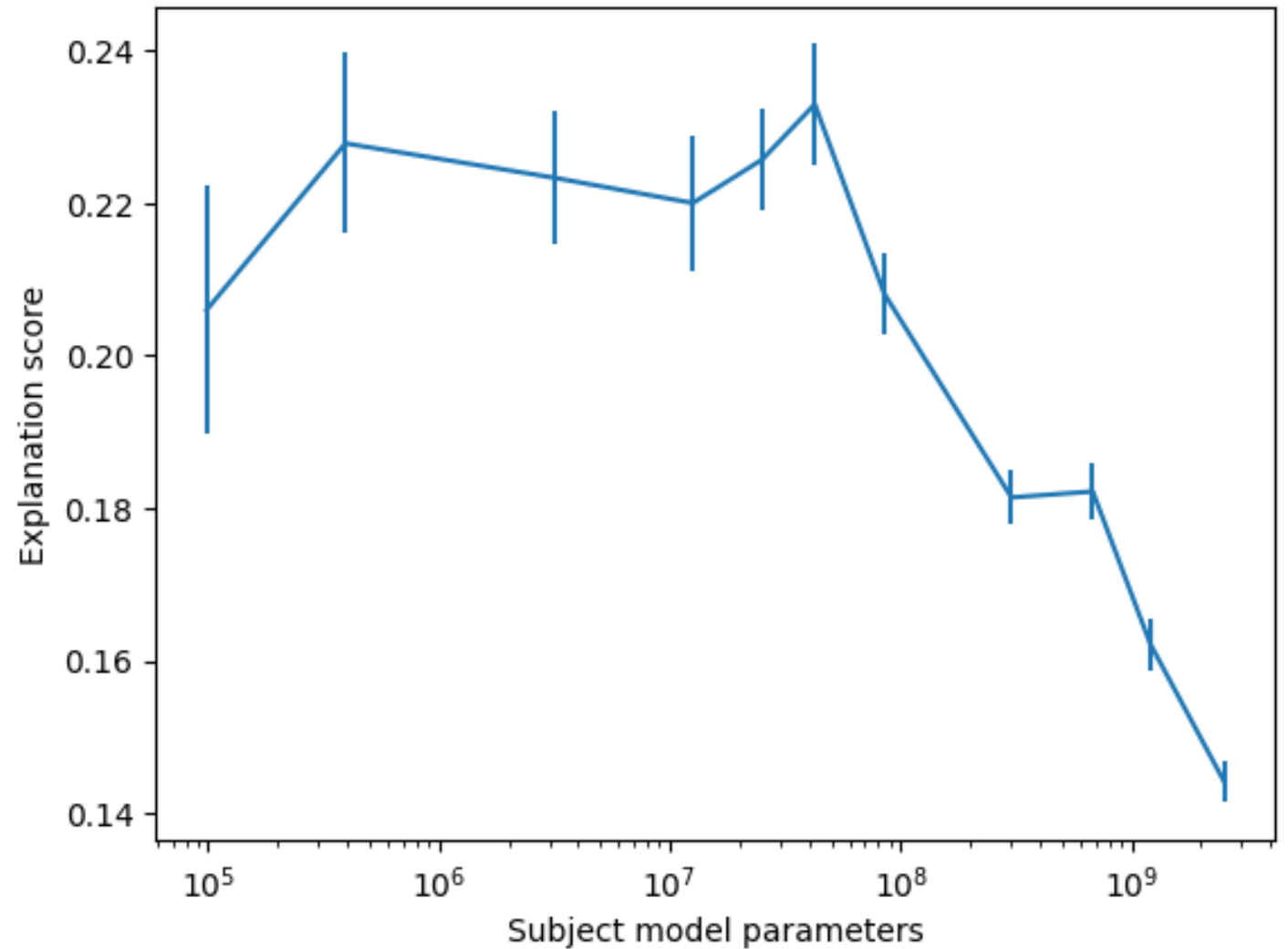
Text: I am **running** to

Last token activation, considering the token in the context in which it appeared in the text: :

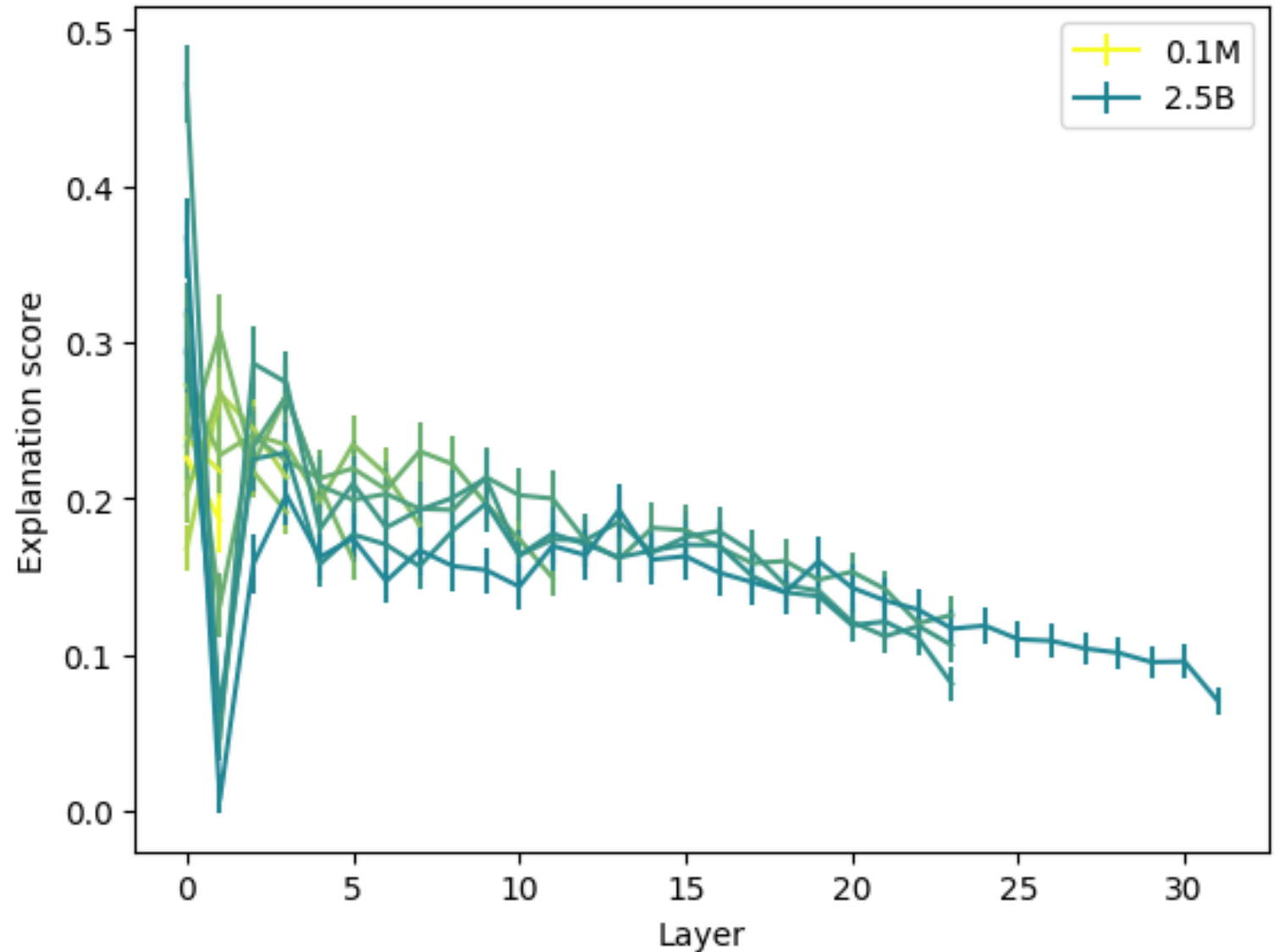


am 0 **running 10** to 0

GPT4 能夠成功解釋神經元嗎？



GPT4 能夠成功解釋神經元嗎？



GPT4 能夠成功解釋神經元嗎？

這個神經元的功用是尋找 *words and phrases related to performing actions correctly or properly.* **0.42**

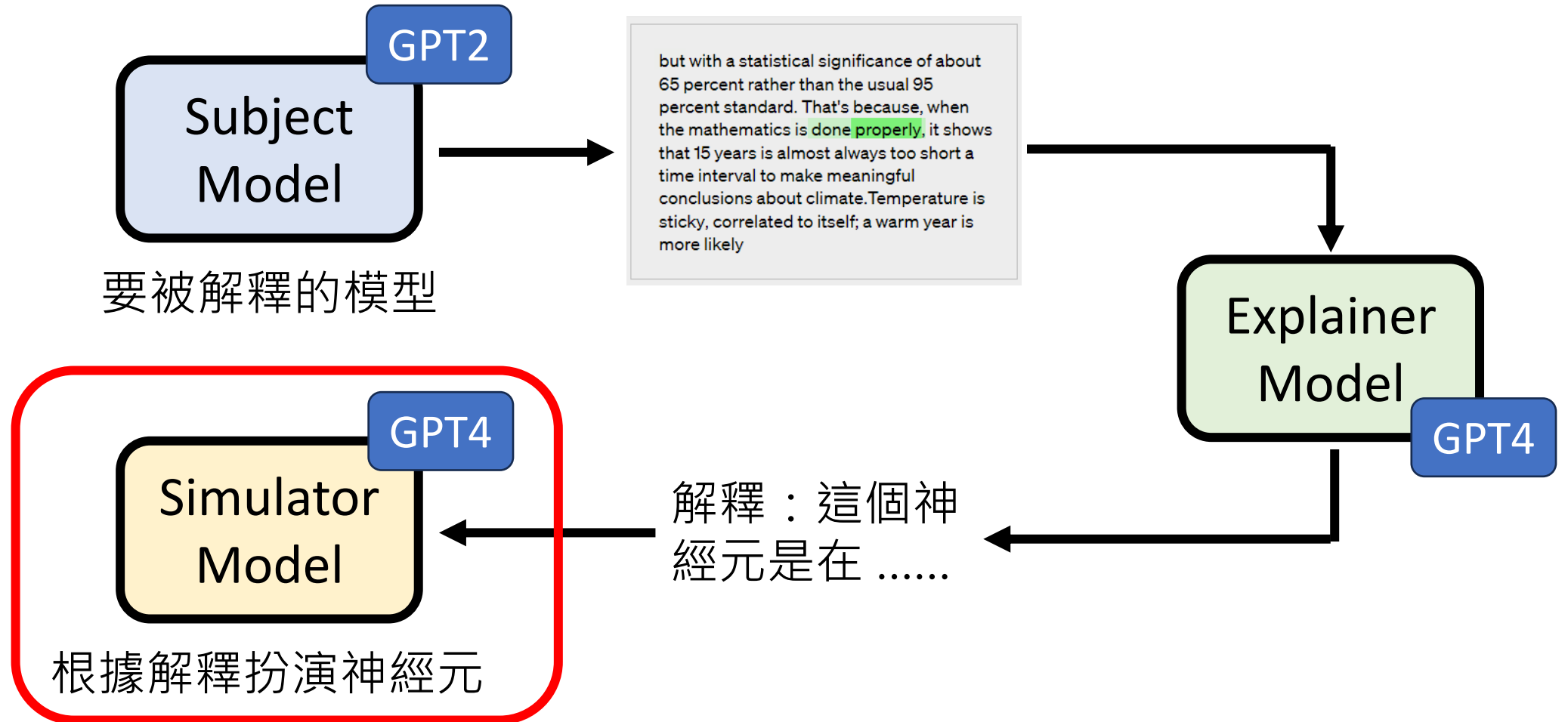
這個神經元的功用是尋找 *words related to general concepts, titles, and partial terms.* **0.14**



GPT4 提供解釋，GPT2 中神經元 explanation score 平均為 **0.15**
(也就是說其實多數的神經元都沒有好的解釋)

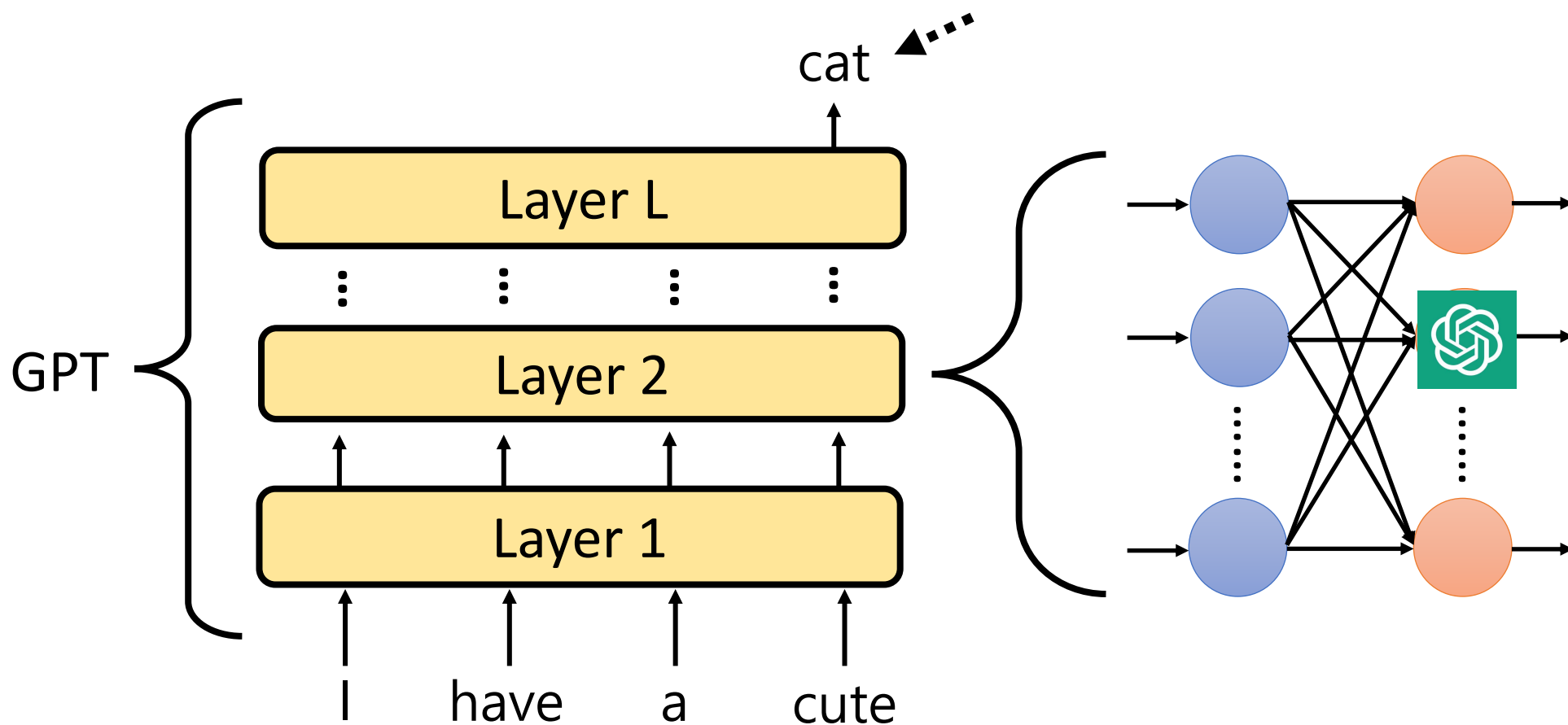
以人類提供解釋，GPT2 中神經元 explanation score 平均為 **0.18**

用 AI 解釋 AI 方法概覽

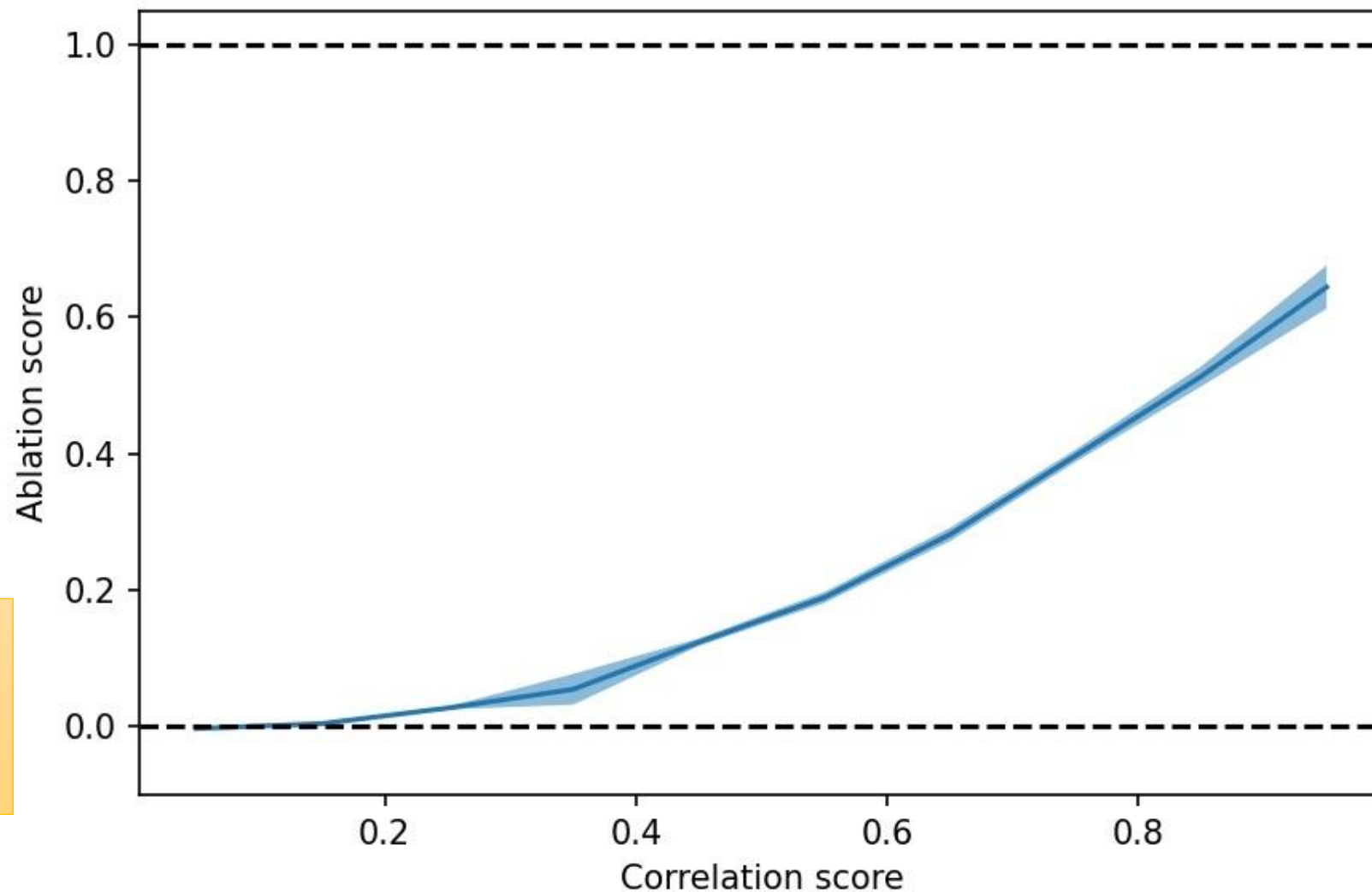


用 GPT4 扮演神經元取代 GPT2 的神經元

是否改變？

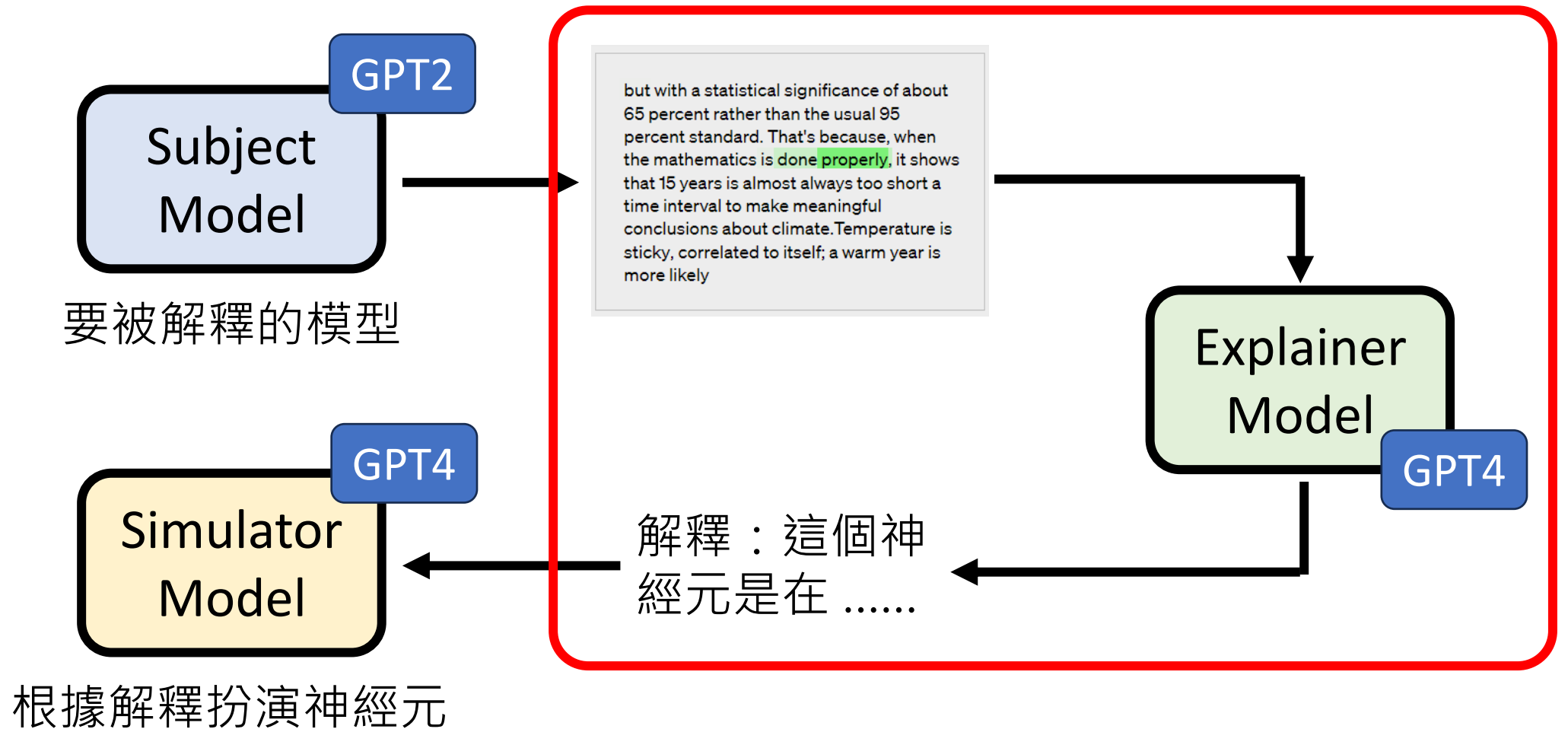


用 GPT4 扮演神經元取代 GPT2 的神經元



Humans prefer higher-scoring explanations over lower-scoring ones.

用 AI 解釋 AI 方法概覽



Neuron 4

Explanation of neuron 4 behavior: the main thing this neuron does is find present tense verbs ending in 'ing'

Activations:

<start>

Star unknown

ting unknown

from unknown

a unknown

position unknown

of unknown

strength unknown

<end>



<start>

Star 0

ting 10

from 0

a 0

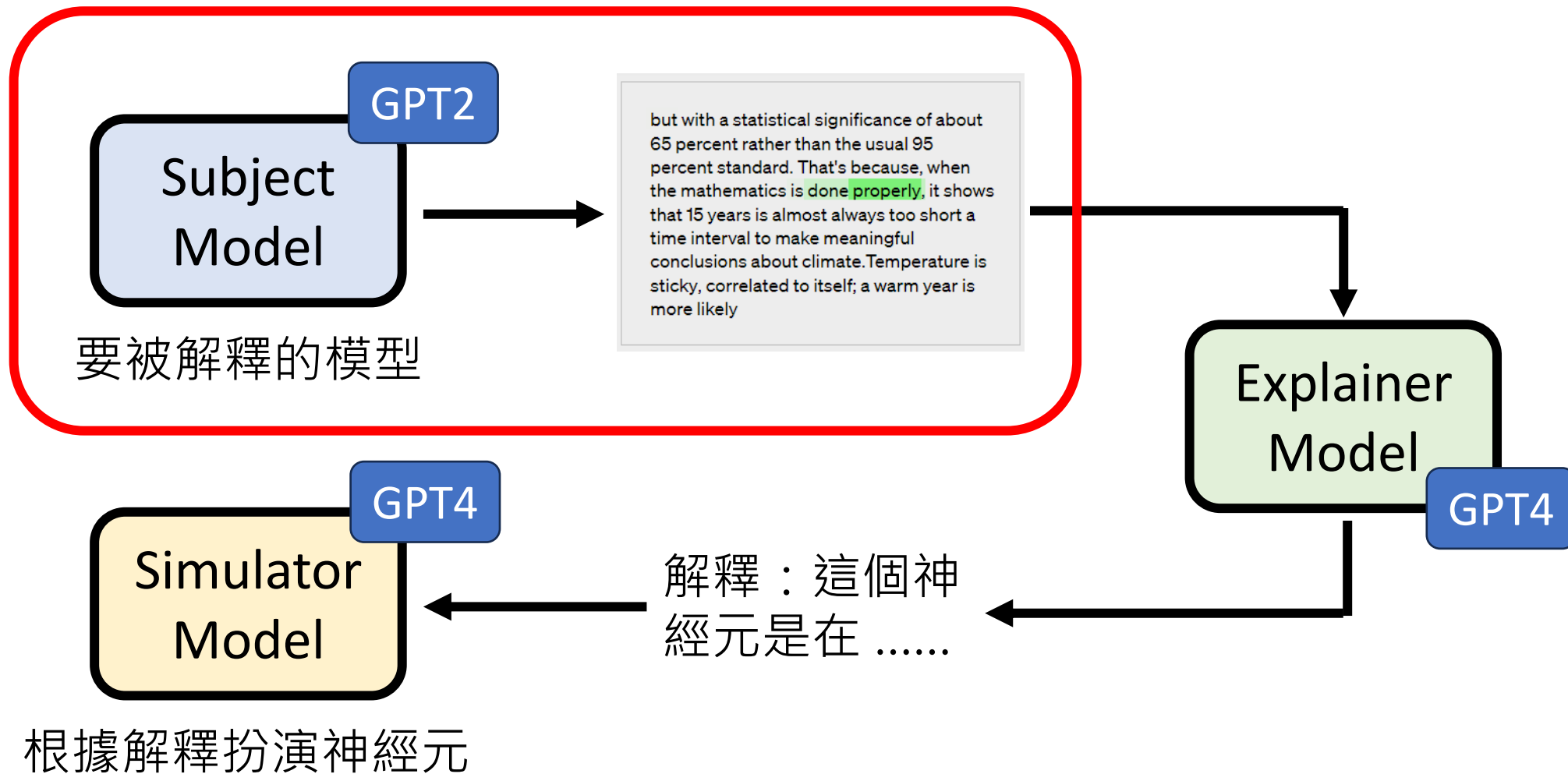
position 0

of 0

strength 0

<end>

用 AI 解釋 AI 方法概覽



提供給 Explainer 的資料：直覺選擇待解釋神經元有輸出較大的句子

is updated every hour.

Please note: not all stray animal pictures are posted on this site. Please visit the shelter regularly to look for your lost pet.

Sorry, none of our We will do our best to help with your search, but as the owner, you are ultimately responsible to look for and identify your pet.

Please check to make sure you checked off the correct filters at the top of the page.

Try a more general search.

Not all films from our collection are a

It's possible that the film you're search

your destination. In the unfortunate event of shipping damage, complimentary wedges cannot be replaced for free if they are damaged in transit.

NOTE ABOUT COMPLIMENTARY WEDGES

All AA grade handles are shipped with complimentary wedges - if wedges are required.

Not all handles

這個神經元的功用是尋找 "all"



The task format is as follows. description :: <answer>example sentence that fits that description</answer>

This task has exactly 10 answer(s) each enclosed in <answer></answer> tags.

Remember, the answer is always at least one full sentence, not just a word or a phrase.

the term "all" along with related contextual phrases. ::

沒有輸出大數值

這個神經元的功用是尋找 "not all"



<answer>Everyone is excited about the event.</answer>

<answer>All the students attended the lecture.</answer>

<answer>She manage your destination. In the unfortunate event of shipping damage, complimentary wedges cannot be replaced

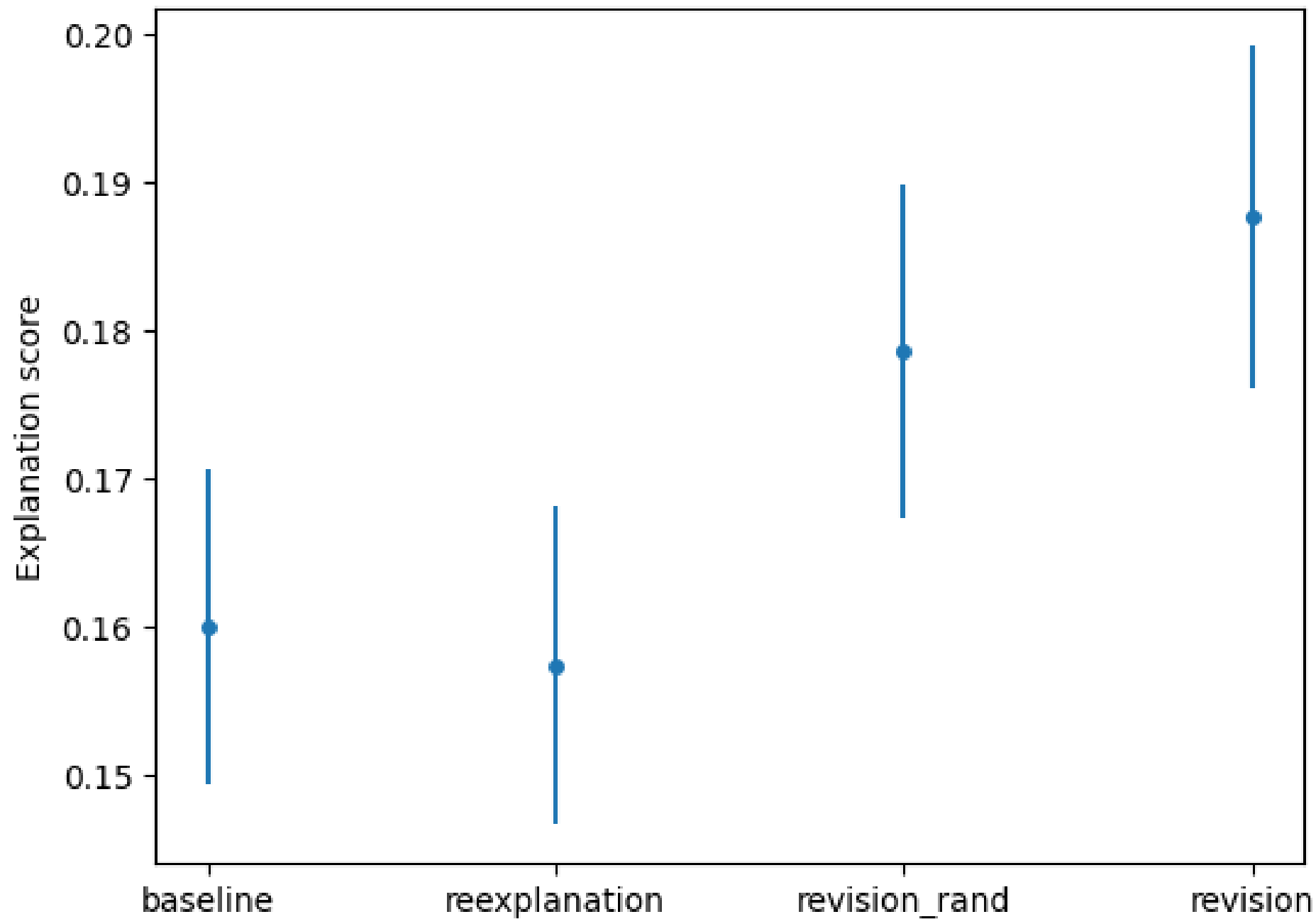
<answer>They invited for free if they are damaged in transit.

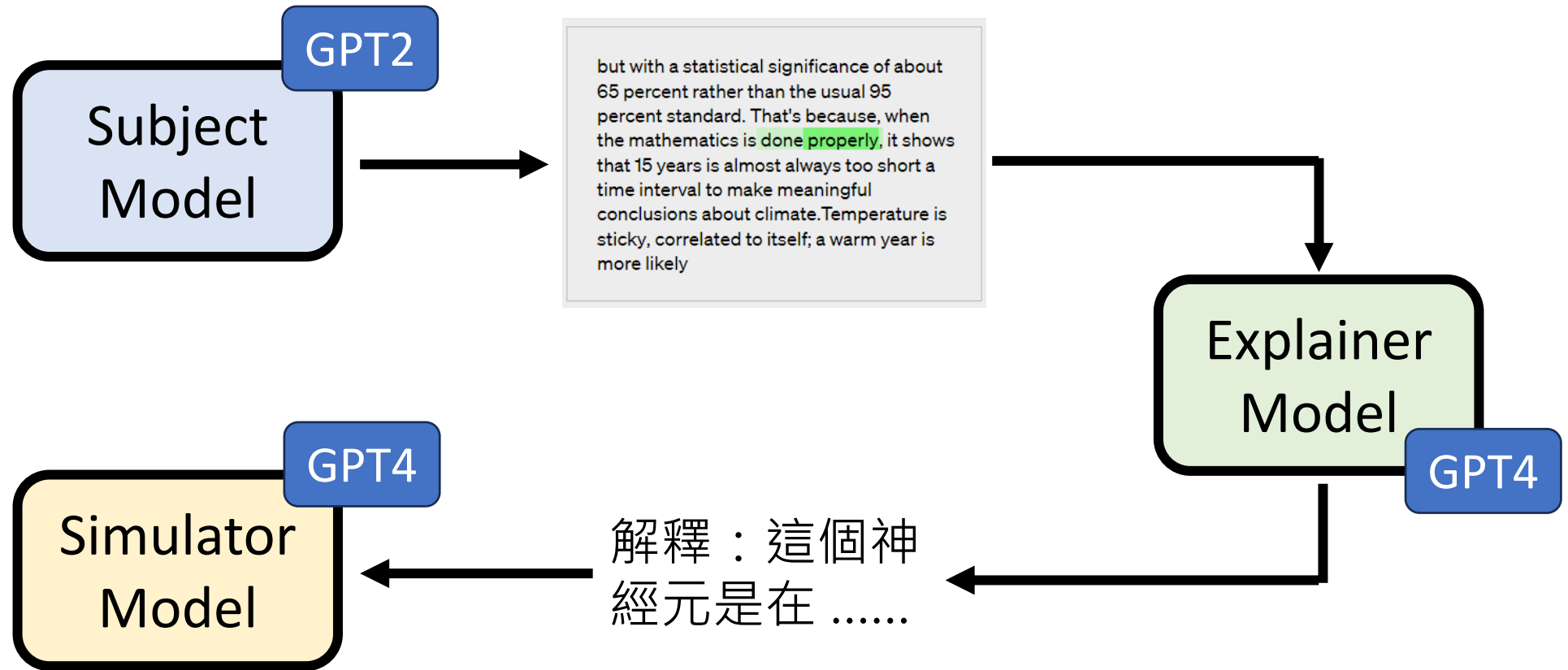
<answer>All the emplo NOTE ABOUT COMPLIMENTARY WEDGES

<answer>He finished a All AA grade handles are shipped with complimentary wedges - if wedges are required.

<answer>All the ingred Not all handles

<answer>All the lights in the house were turned off.</answer>





- 目標上的質疑：單一神經元有可解釋的功能嗎？能用語言來解釋嗎？
- 框架上的質疑：
 - 如果 Simulator 很弱，那麼就算 Explainer 有精確的解釋也拿不到 explanation score
 - Explainer 和 Simulator 會不會串通好？用他們之間的祕密用語來解釋