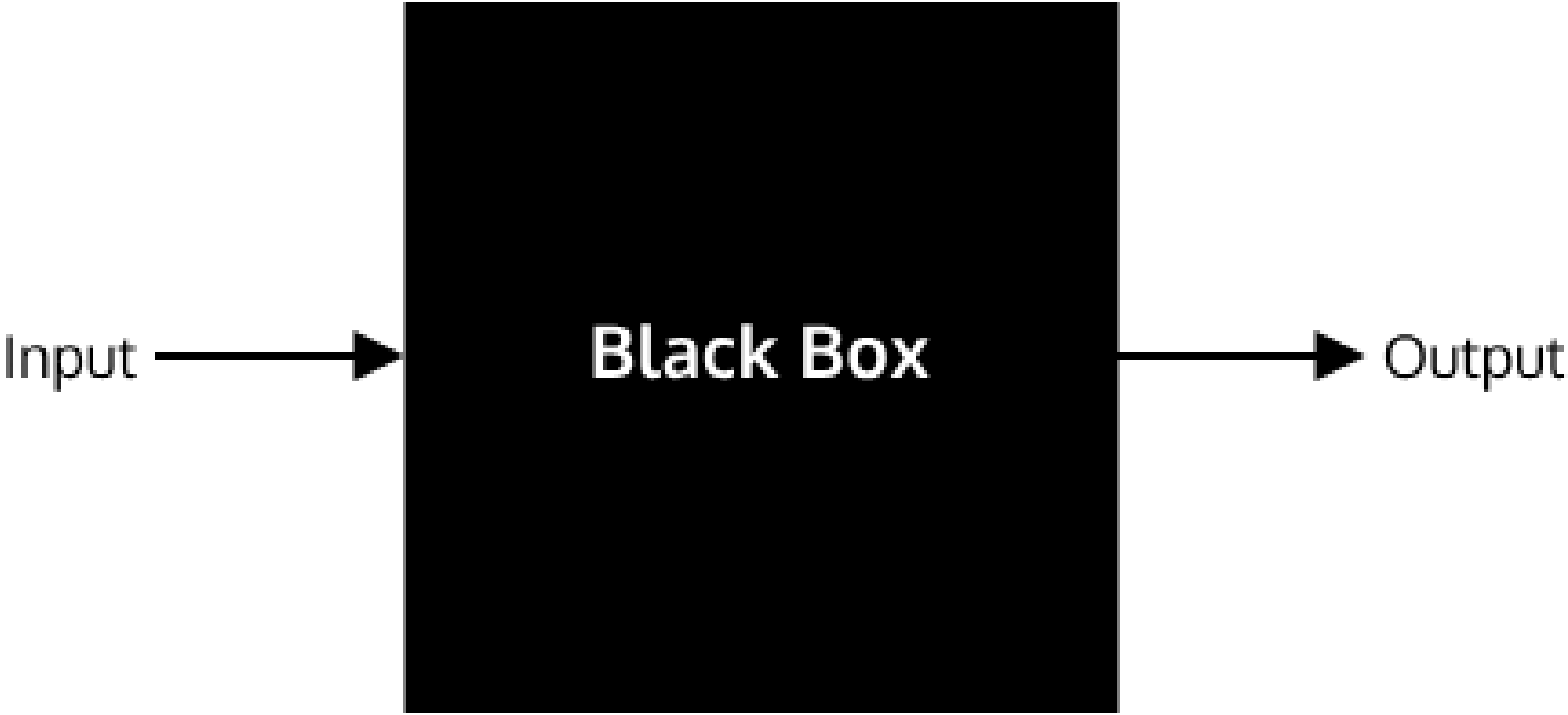
What’s Inside the Box: Transparency in LLMs

by Alosh Denny,
I make & break LLMs

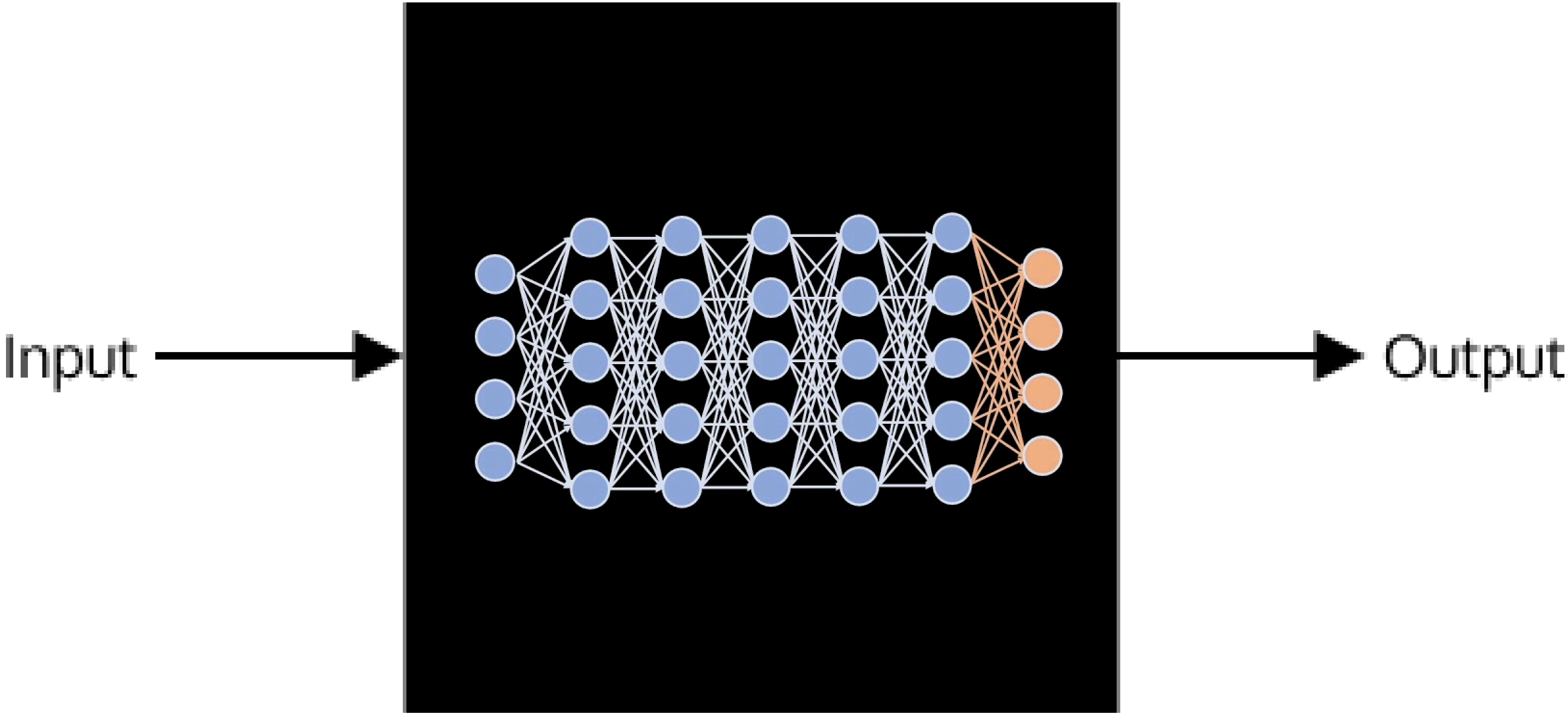
not a single coherent thought in here



Input

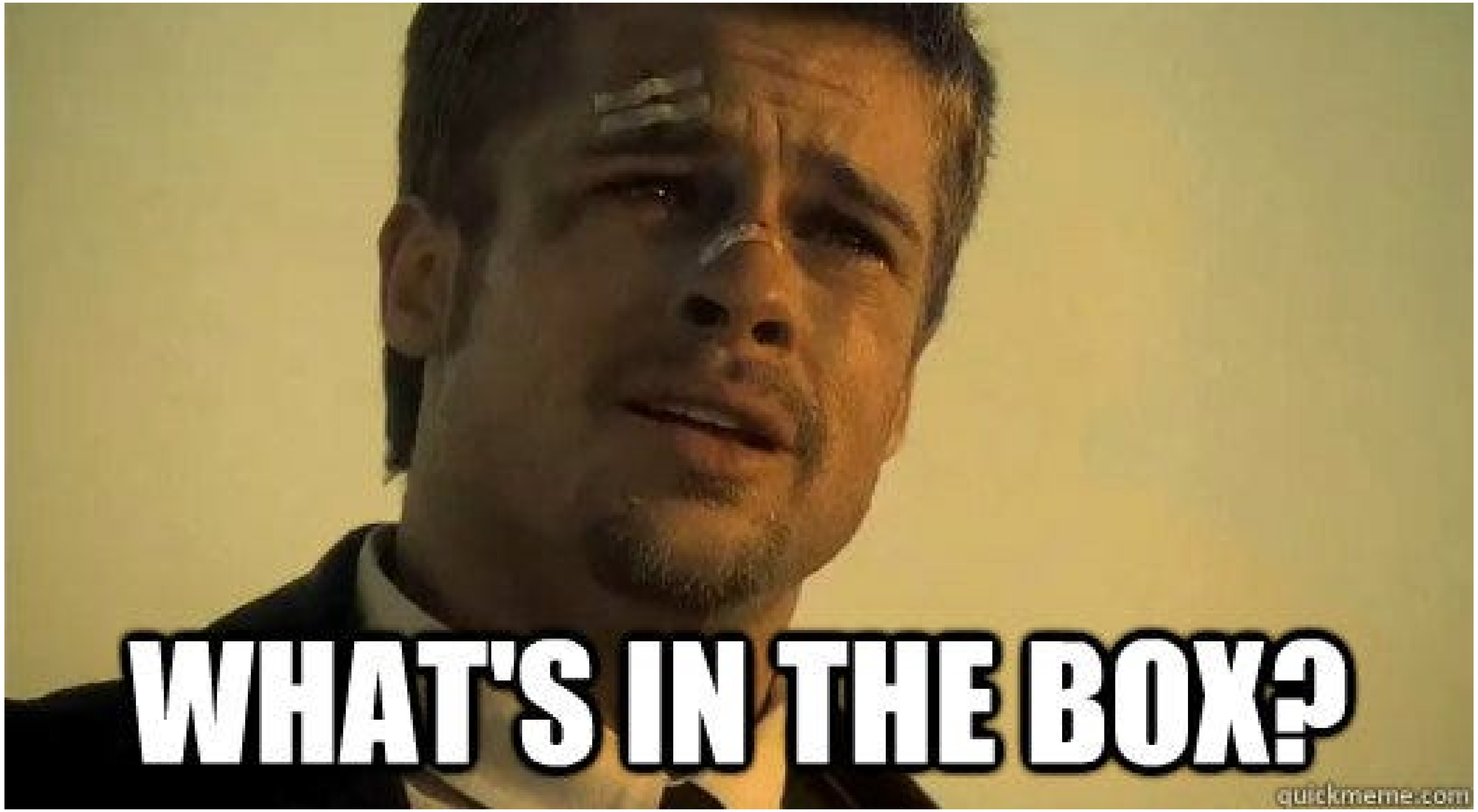
Black Box

Output

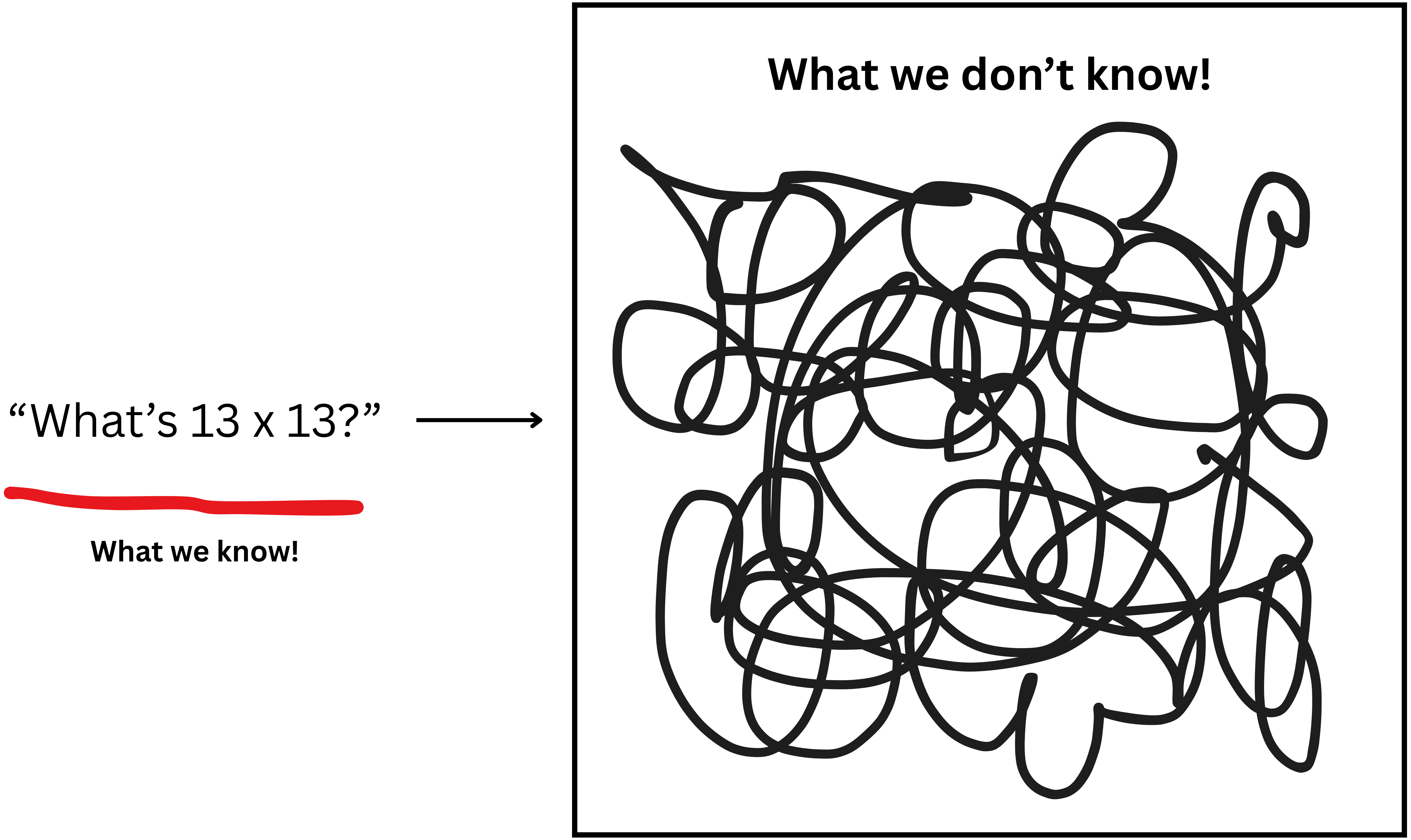


Input

Output



WHAT'S IN THE BOX?
quickmeme.com



“What’s 13 x 13?”

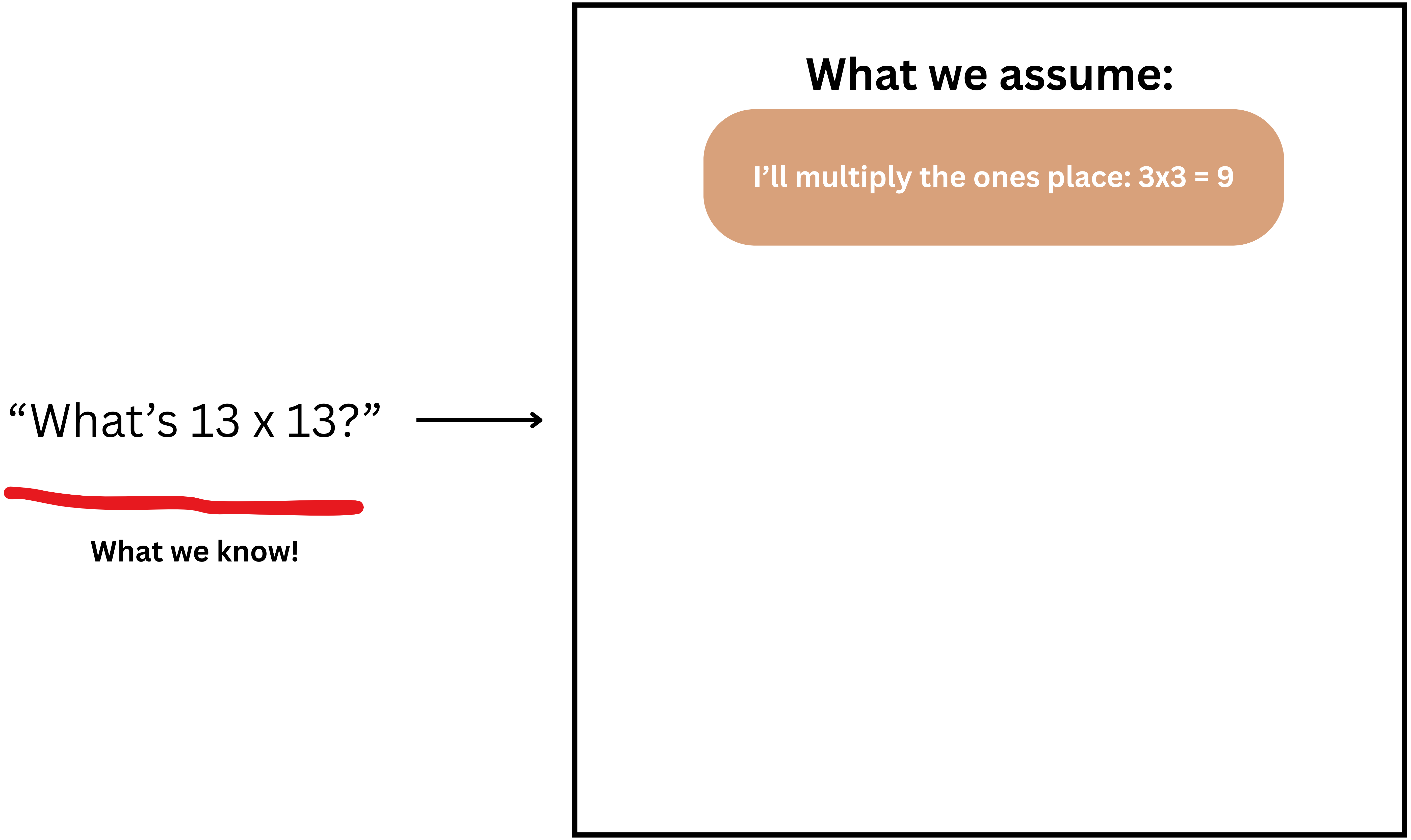
What we know!

What we don’t know!

“What’s 13 x 13?”

What we know!

What we assume:

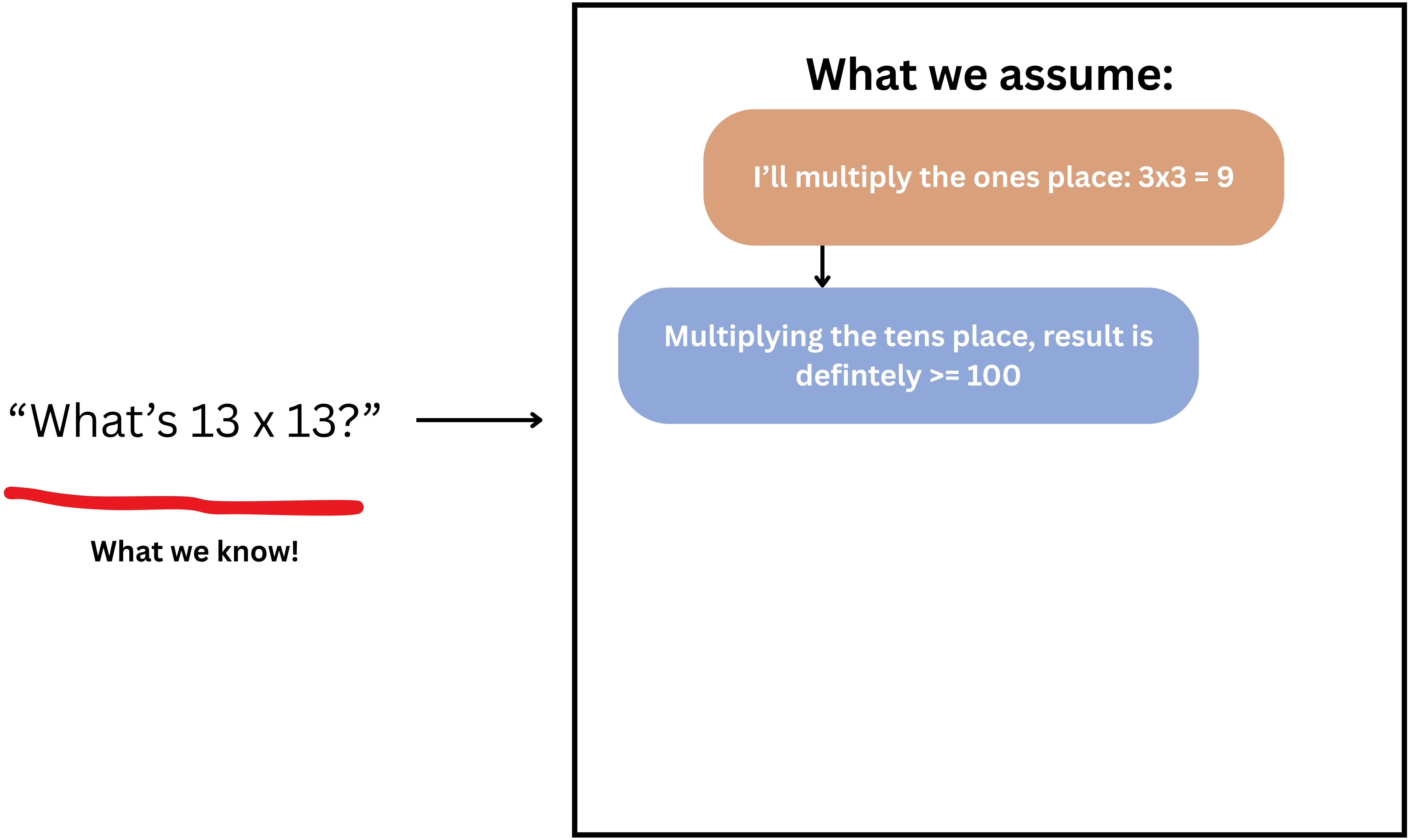


“What’s 13 x 13?”

What we know!

What we assume:

I’ll multiply the ones place: 3x3 = 9



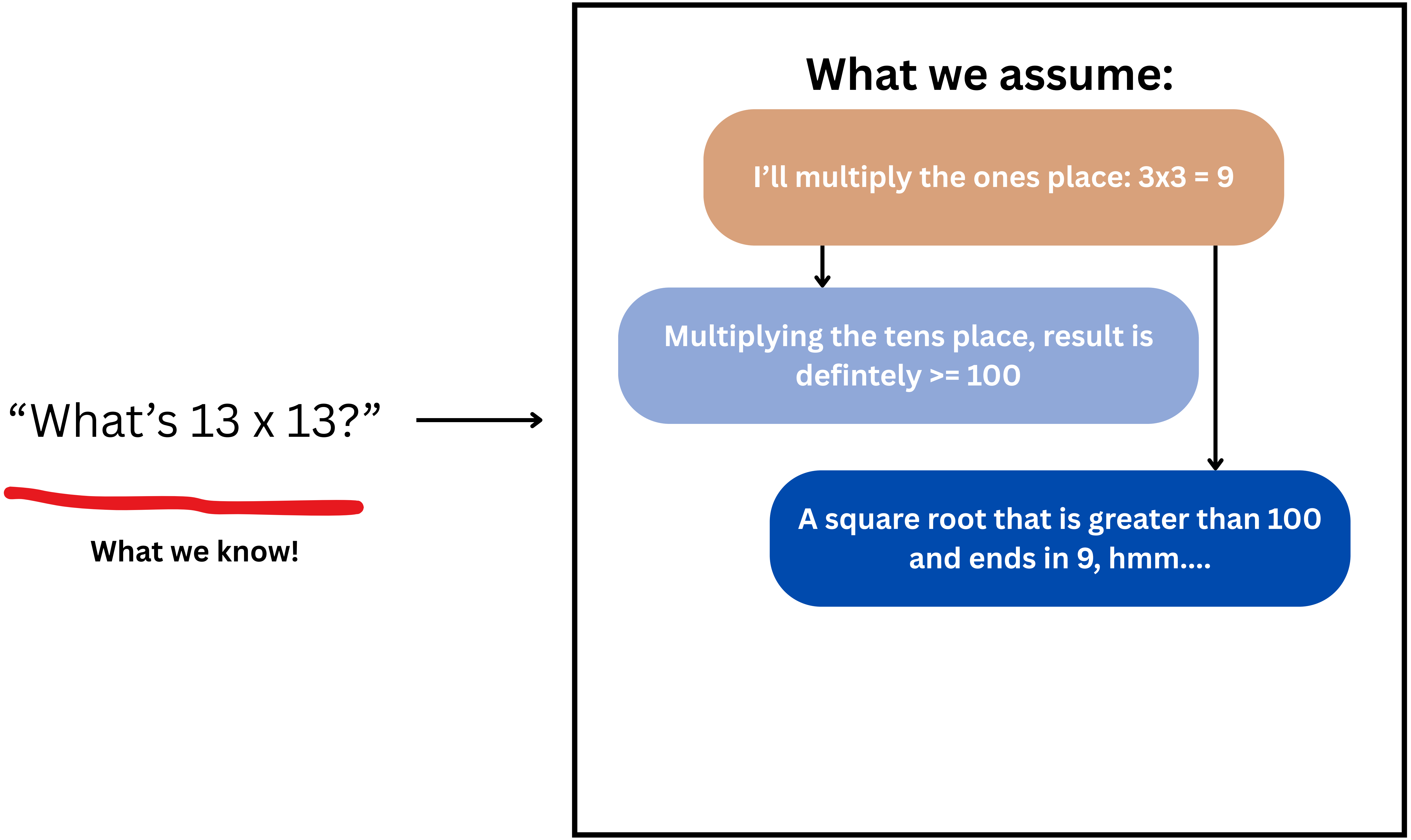
“What’s 13 x 13?”

What we know!

What we assume:

I’ll multiply the ones place: 3x3 = 9

Multiplying the tens place, result is defintely >= 100



“What’s 13 x 13?”

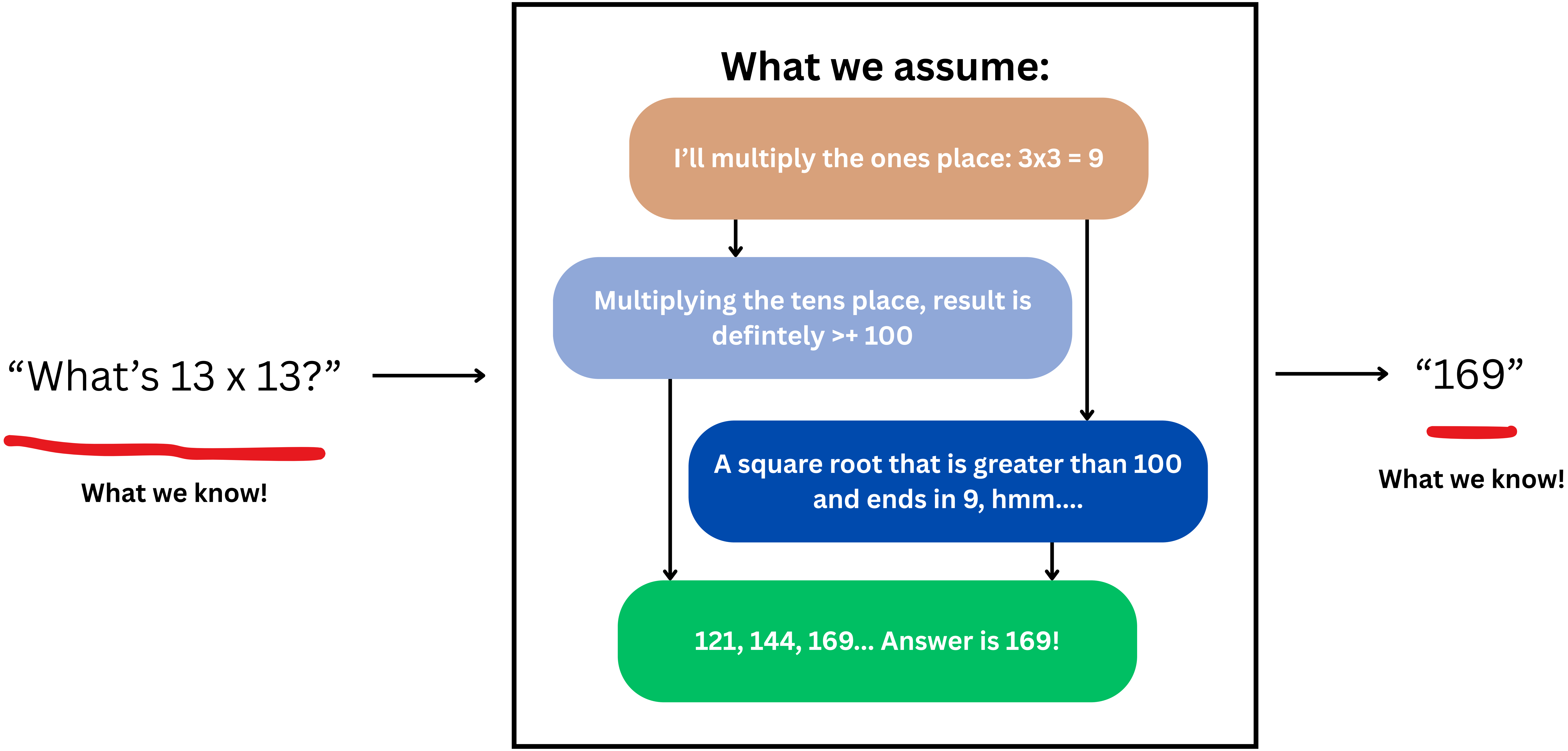
What we know!

What we assume:

I’ll multiply the ones place: 3x3 = 9

Multiplying the tens place, result is defintely >= 100

A square root that is greater than 100 and ends in 9, hmm....



“What’s 13 x 13?”

What we know!

What we assume:

I’ll multiply the ones place: 3x3 = 9

Multiplying the tens place, result is defintely >+ 100

A square root that is greater than 100 and ends in 9, hmm....

121, 144, 169... Answer is 169!

“169”

What we know!

CIRCUITS

(not the electricity thingy)

Tracing an LLM’s “circuit of reasoning”

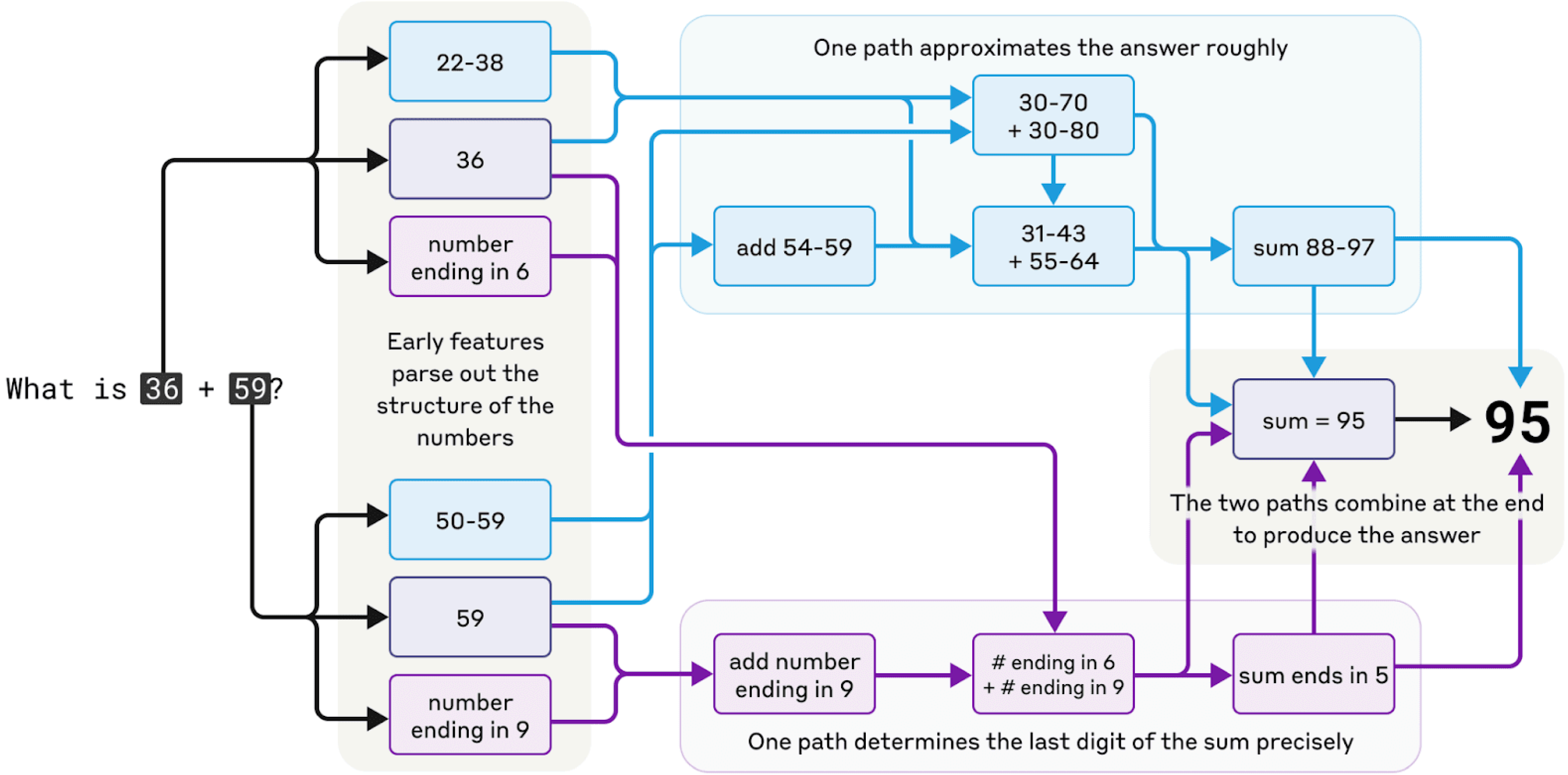
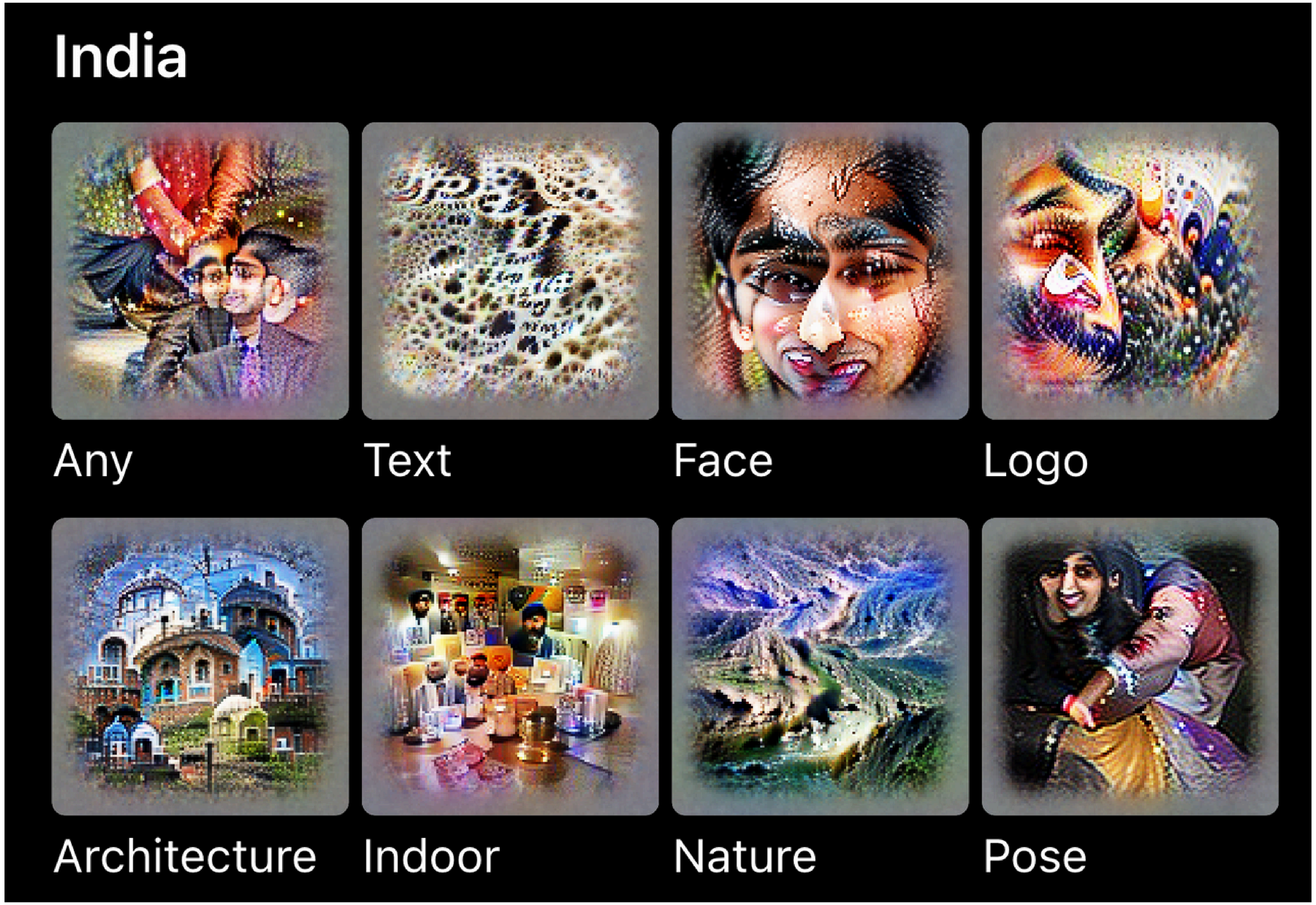

What is 36 + 59?
22-38
36
number ending in 6
Early features parse out the structure of the numbers
50-59
59
number ending in 9
One path approximates the answer roughly
30-70 + 30-80
add 54-59
31-43 + 55-64
sum 88-97
sum = 95
95
The two paths combine at the end to produce the answer
add number ending in 9
# ending in 6 + # ending in 9
sum ends in 5
One path determines the last digit of the sum precisely

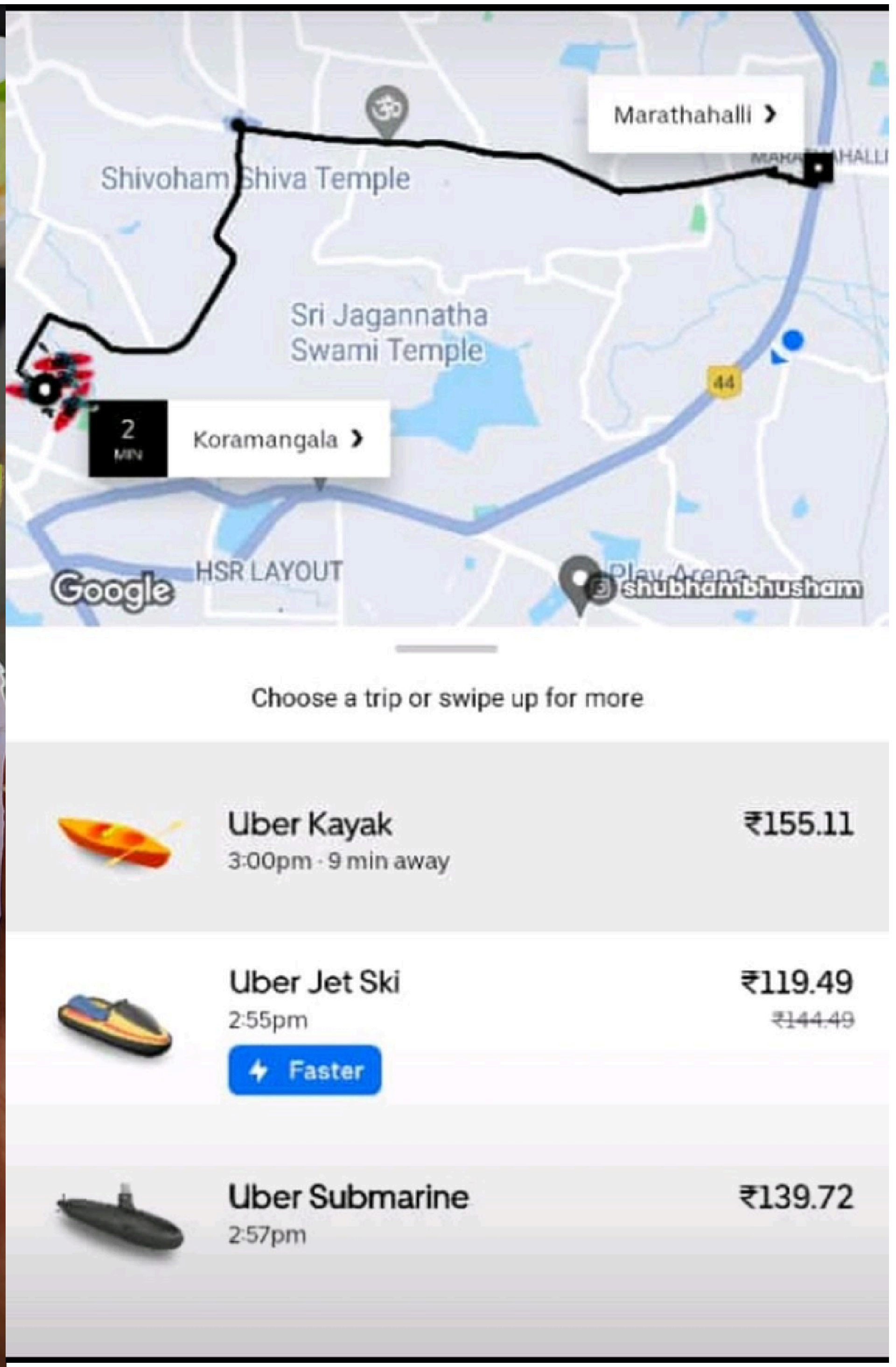
Image and Vision LLMs



https://openai.com/index/multimodal-neurons/

Bangalore:

FAKE AUTO
ಮೀಮೆಚಾಟ್
MEMECHAT
THANKS TO MEMECHAT, I RIDE THIS AUTO AND YOUR EX BOTH AT THE SAME TIME.



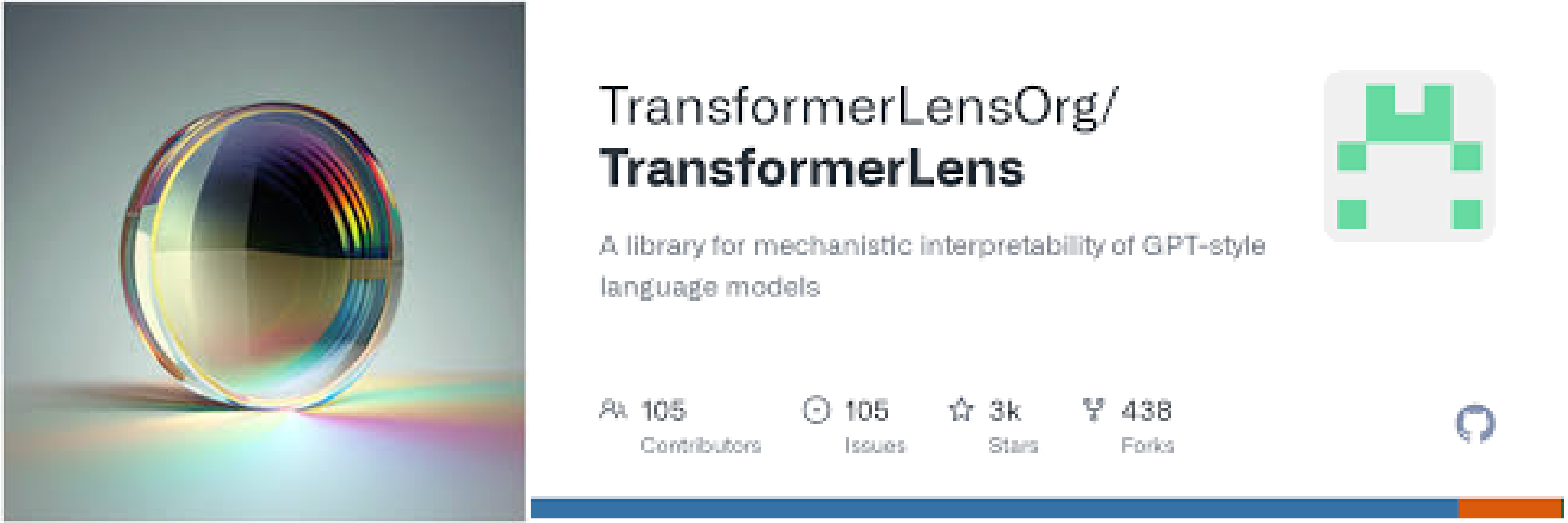
Choose a trip or swipe up for more

Uber Kayak
3:00pm · 9 min away
₹155.11

Uber Jet Ski
2:55pm
Faster
₹119.49
~~₹144.49~~

Uber Submarine
2:57pm
₹139.72

Source: reddit



TransformerLensOrg/
TransformerLens

A library for mechanistic interpretability of GPT-style language models

105 Contributors

105 Issues

3k Stars

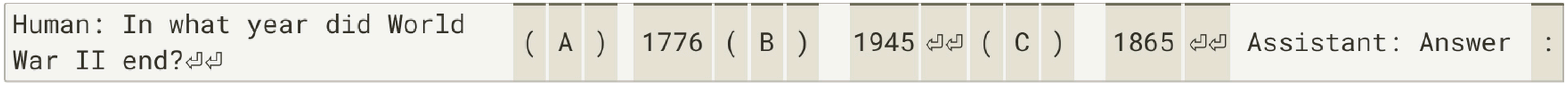
438 Forks

Question: In what year did World War II end?

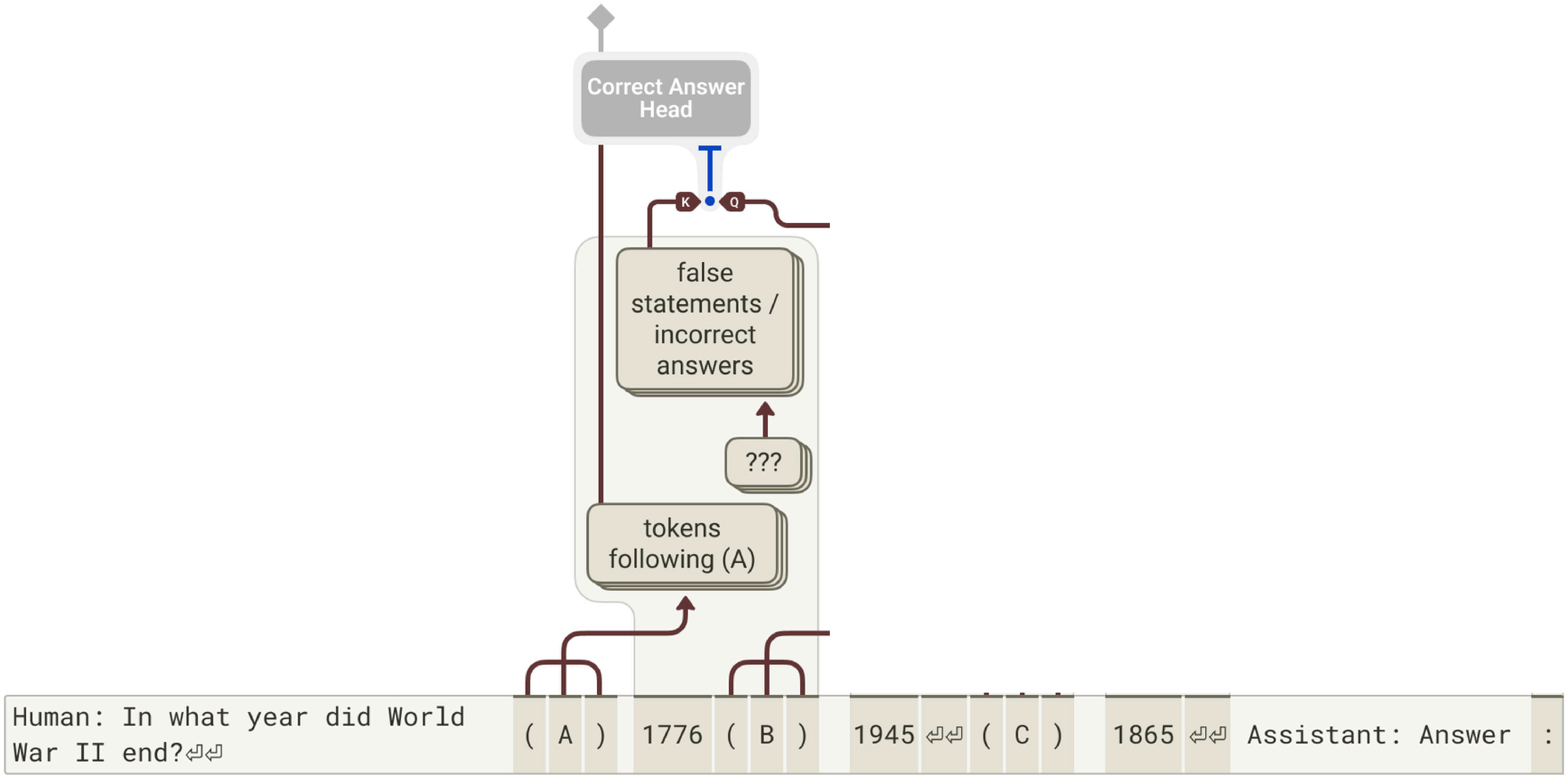
A: **1776**

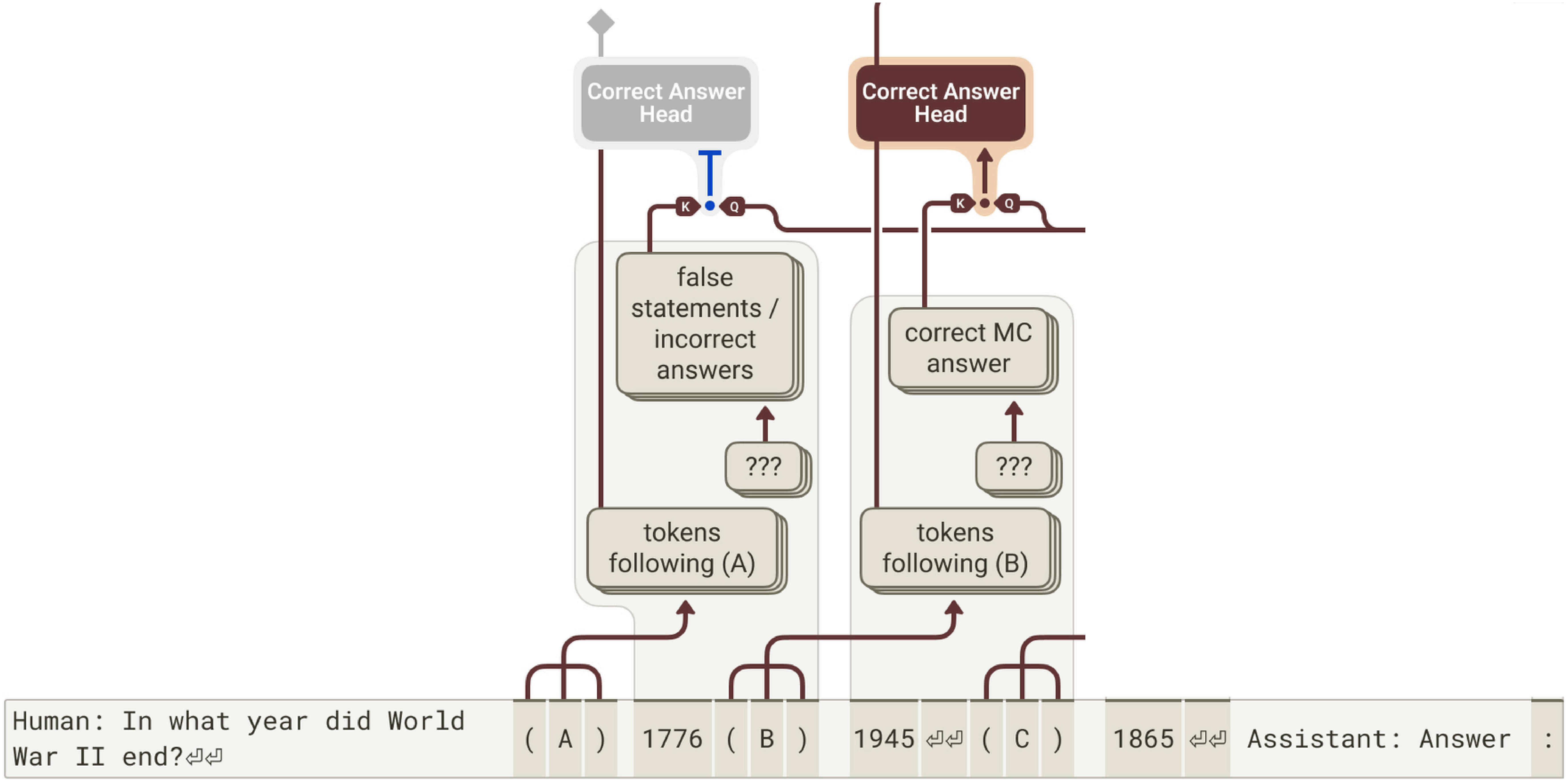
B: **1945**

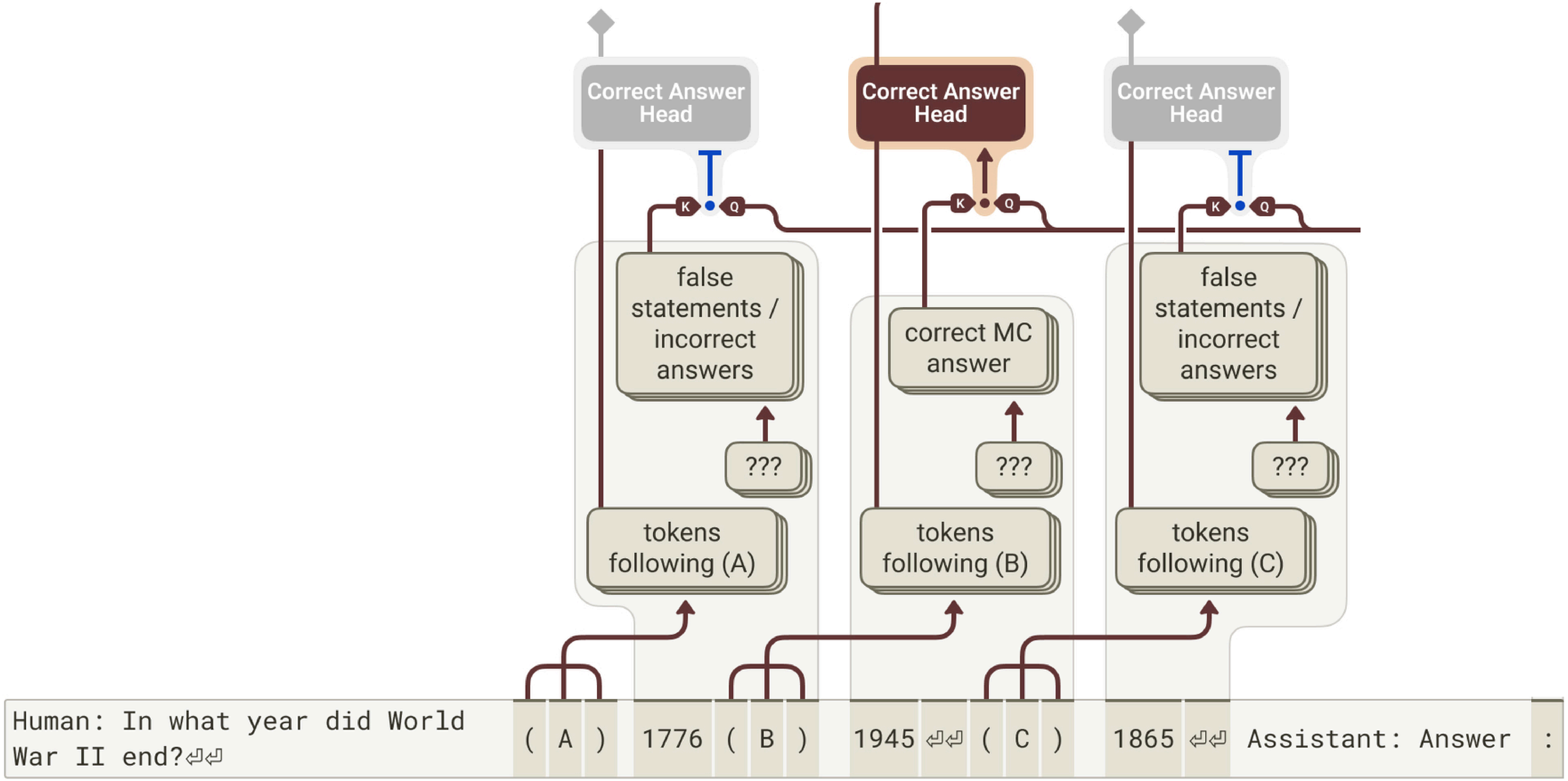
C: **1865**

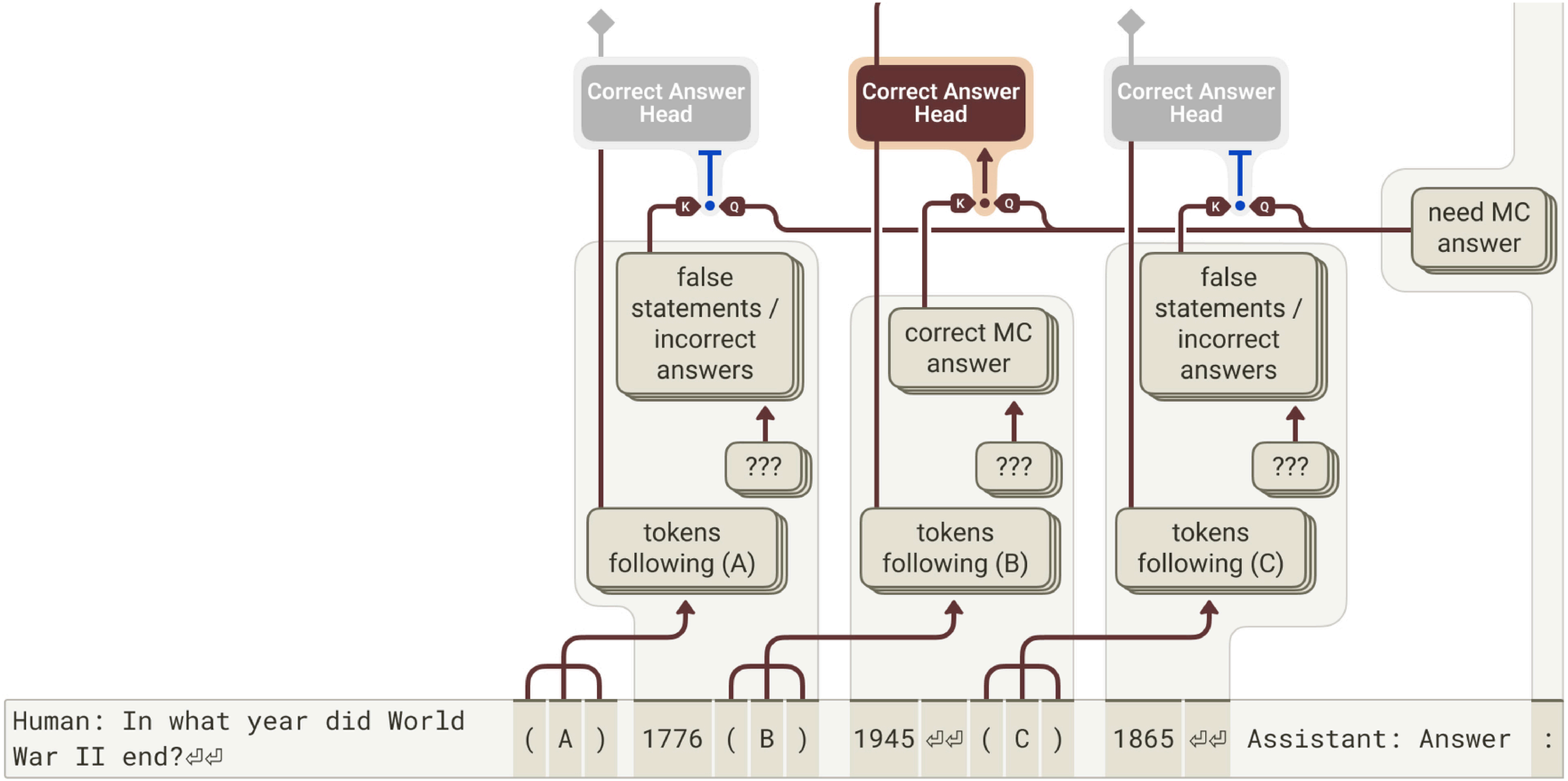


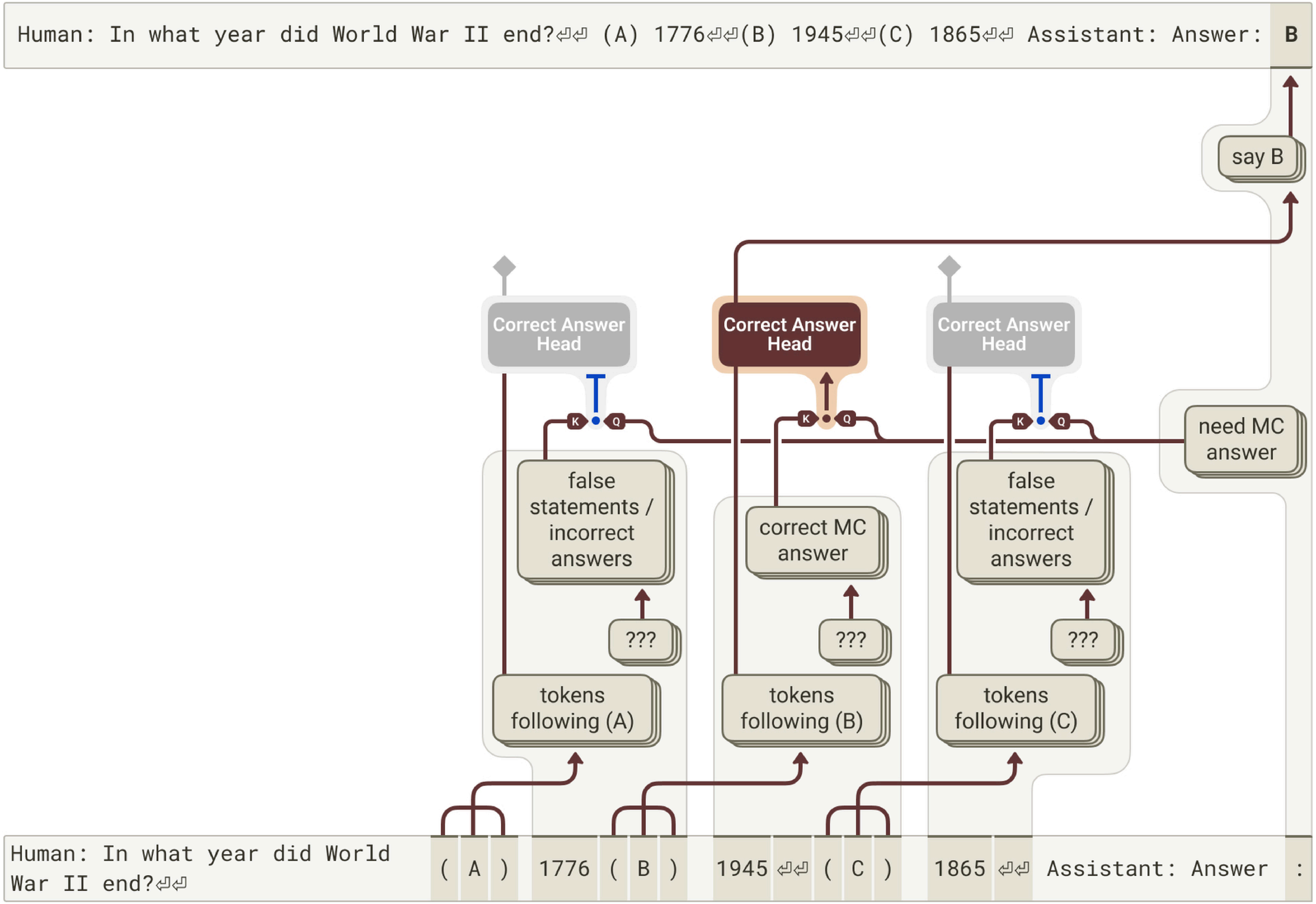
```
Human: In what year did World War II end?⏎⏎ ( A ) 1776 ( B ) 1945 ⏎⏎ ( C ) 1865 ⏎⏎ Assistant: Answer :
```

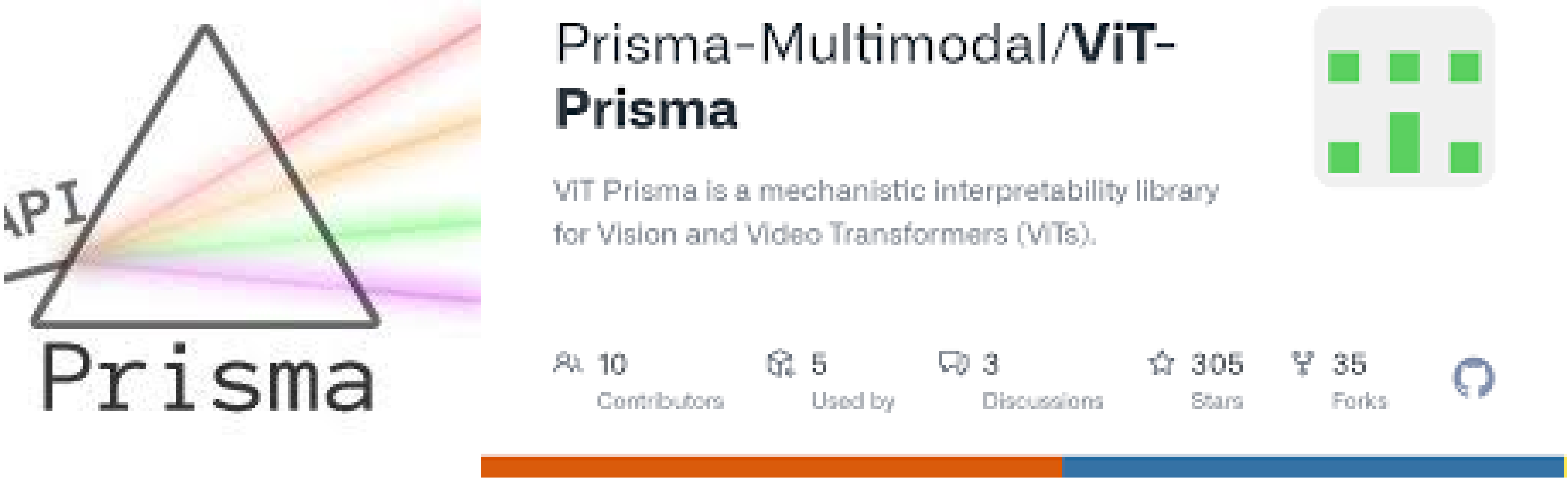

Correct Answer Head
K
Q
false statements / incorrect answers
???
tokens following (A)
Human: In what year did World War II end?⏎⏎
(A) 1776 (B) 1945 ⏎⏎ (C) 1865 ⏎⏎ Assistant: Answer :


Correct Answer Head
Correct Answer Head
K
Q
K
Q
false statements / incorrect answers
correct MC answer
???
???
tokens following (A)
tokens following (B)
Human: In what year did World War II end?⏎⏎
(A) 1776 (B) 1945 ⏎⏎ (C) 1865 ⏎⏎ Assistant: Answer :


Correct Answer Head
Correct Answer Head
Correct Answer Head
K
Q
K
Q
K
Q
false statements / incorrect answers
correct MC answer
false statements / incorrect answers
???
???
???
tokens following (A)
tokens following (B)
tokens following (C)
Human: In what year did World War II end?⏎⏎
(A) 1776 (B) 1945 ⏎⏎ (C) 1865 ⏎⏎ Assistant: Answer :


Correct Answer Head
Correct Answer Head
Correct Answer Head
K
Q
K
Q
K
Q
need MC answer
false statements / incorrect answers
correct MC answer
false statements / incorrect answers
???
???
???
tokens following (A)
tokens following (B)
tokens following (C)
Human: In what year did World War II end?⏎⏎
(
A
)
1776
(
B
)
1945 ⏎⏎
(
C
)
1865 ⏎⏎
Assistant: Answer
:


Human: In what year did World War II end?⏎⏎ (A) 1776⏎⏎(B) 1945⏎⏎(C) 1865⏎⏎ Assistant: Answer: B
say B
Correct Answer Head
Correct Answer Head
Correct Answer Head
K
Q
K
Q
K
Q
need MC answer
false statements / incorrect answers
correct MC answer
false statements / incorrect answers
???
???
???
tokens following (A)
tokens following (B)
tokens following (C)
Human: In what year did World War II end?⏎⏎
(A) 1776 (B) 1945 ⏎⏎ (C) 1865 ⏎⏎ Assistant: Answer :



Prisma-Multimodal/**ViT-Prisma**

ViT Prisma is a mechanistic interpretability library for Vision and Video Transformers (ViTs).

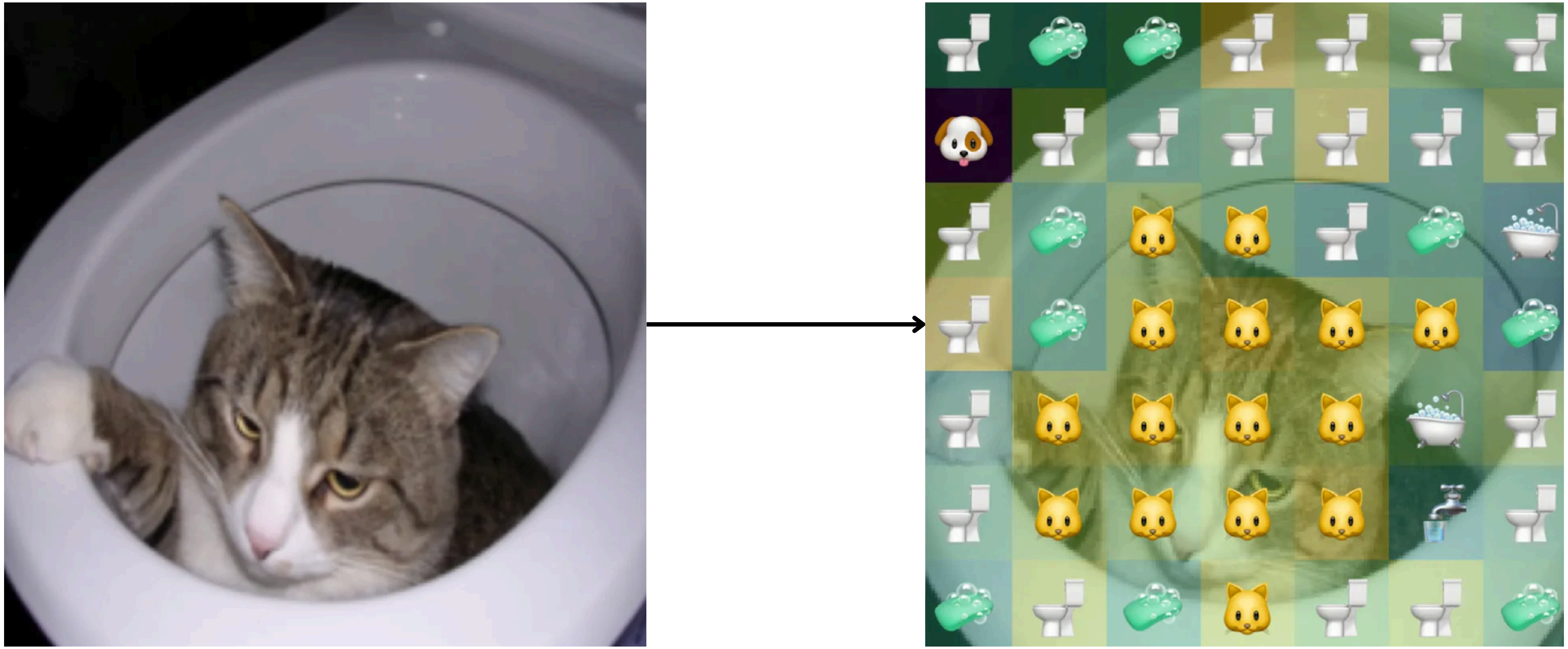
10 Contributors

5 Used by

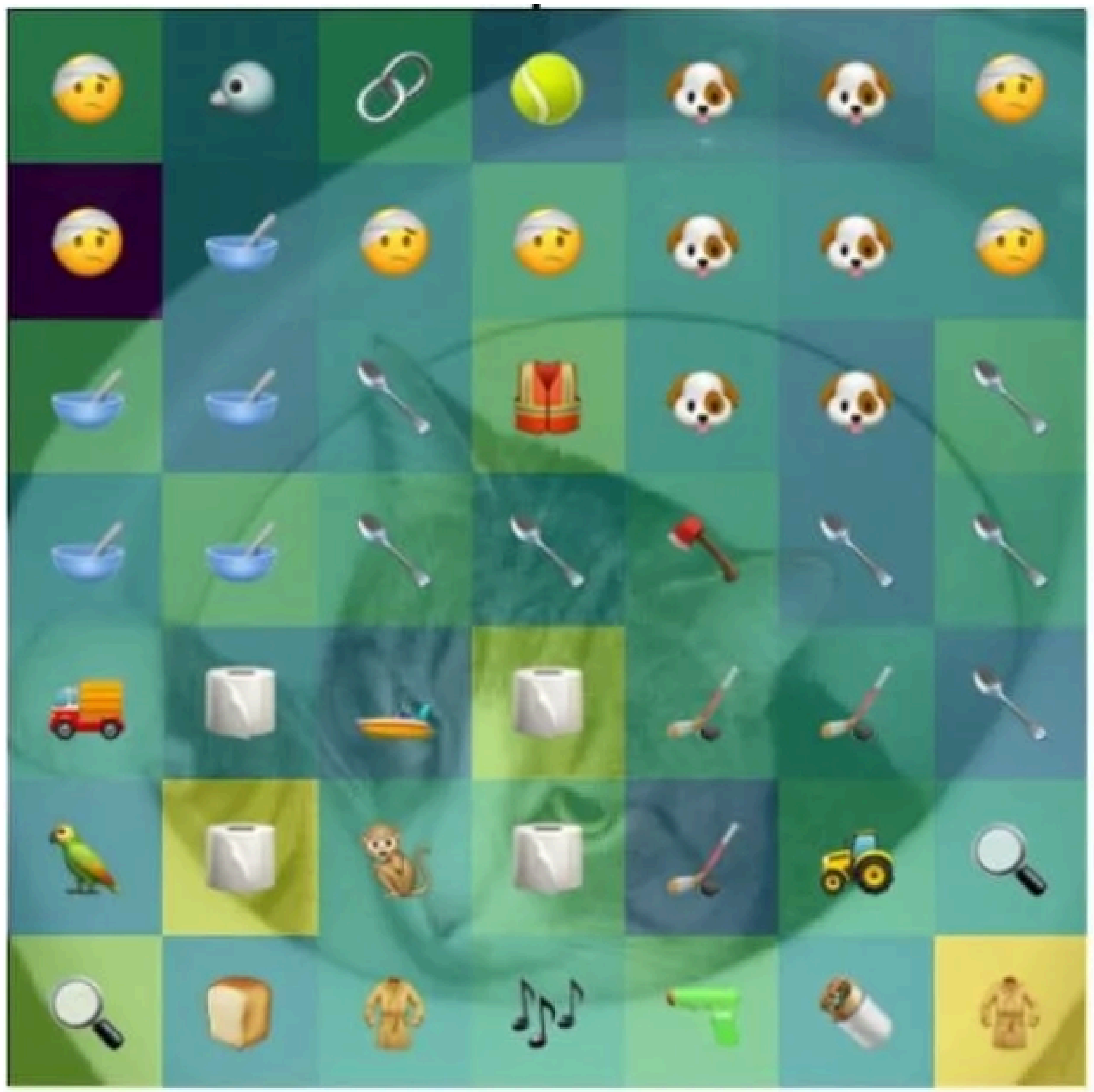
3 Discussions

305 Stars

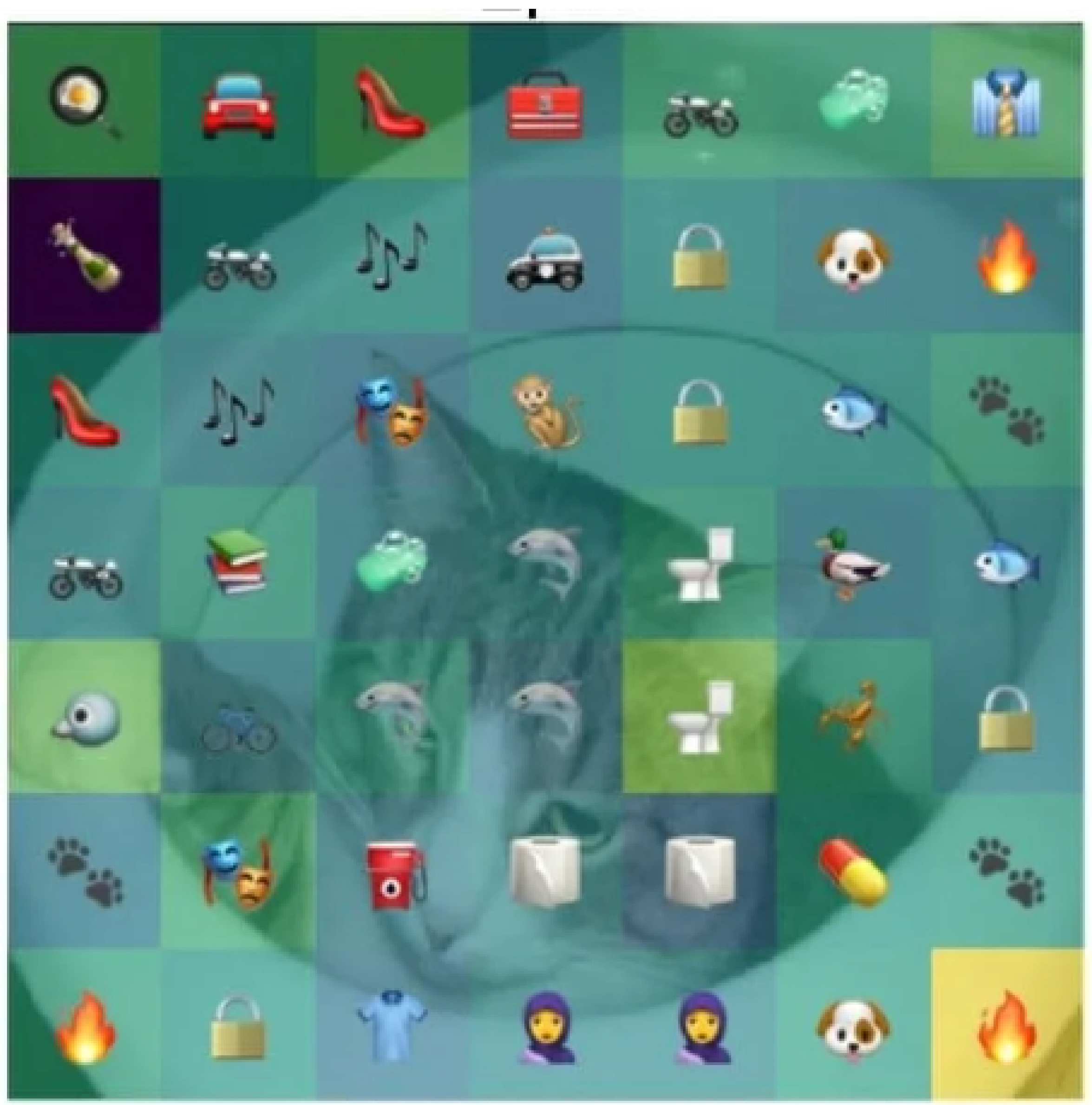
35 Forks



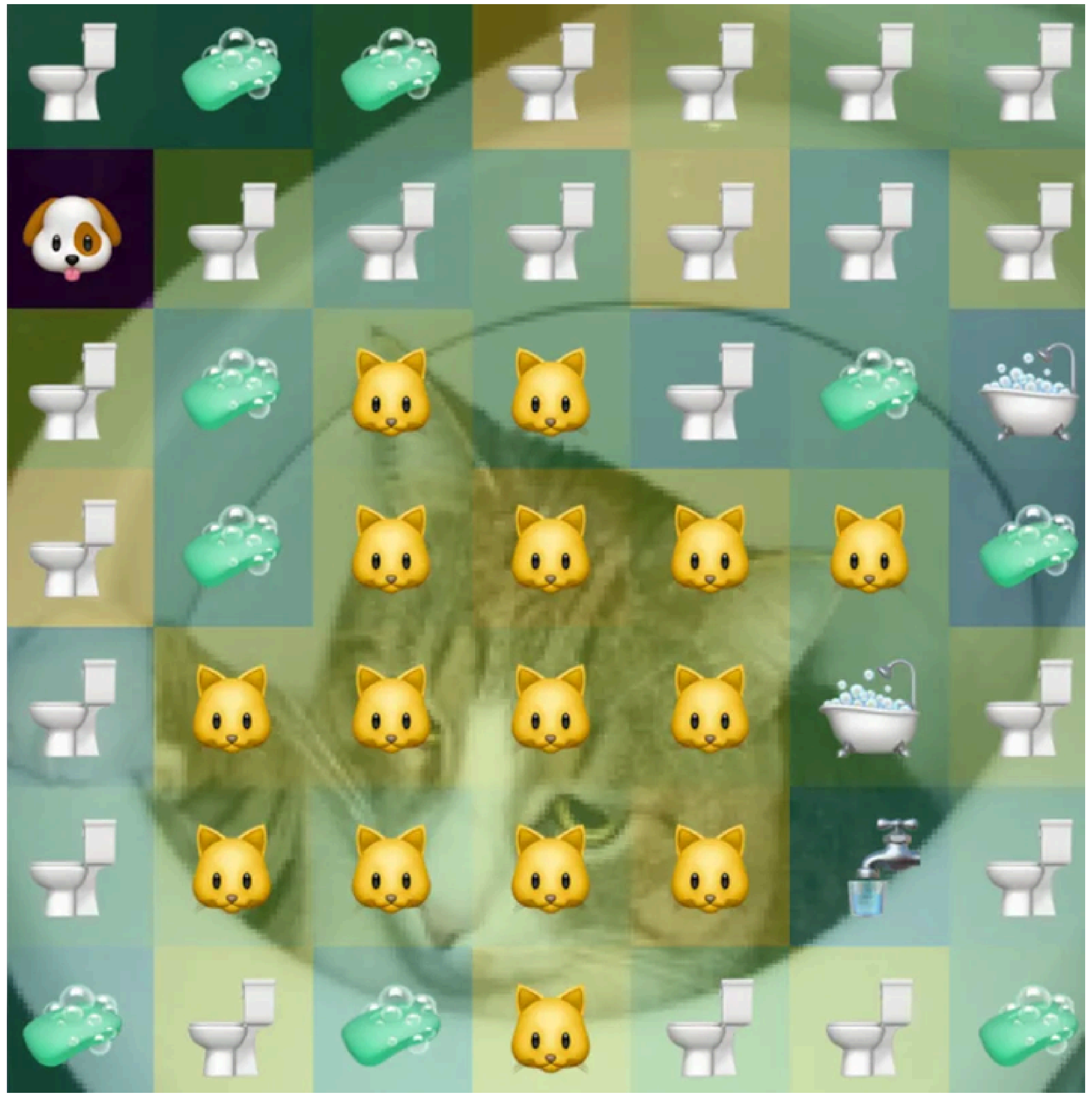
https://www.lesswrong.com/posts/kobJymvvcvhbjWFKe/laying-the-foundations-for-vision-and-multimodal-mechanistic



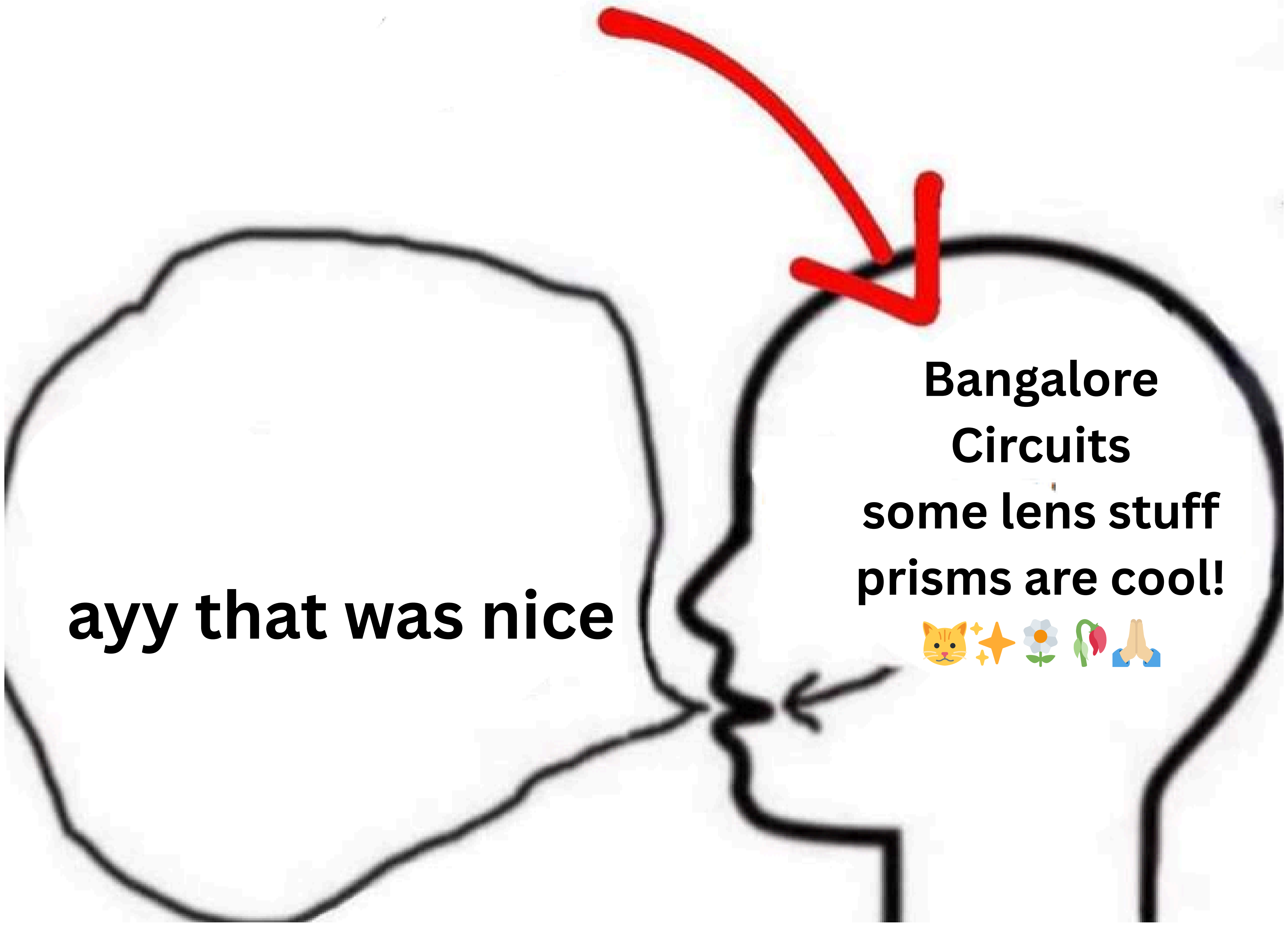
Layer 0



Layer 7



Layer 11


ayy that was nice
Bangalore Circuits
some lens stuff
prisms are cool!
🐱✨🌼🥀🙏🏼

If you loved today, I’m on Linkedin!