Physics of Language Models: Part 3.2
Knowledge Manipulation

Result 1-2

knowledge partial / dual retrievals may be difficult

model can say birthday “October 2, 1996” but not “1996”

Result 3-5

knowledge classification / comparison is hard without CoT

model cannot say “even year” without saying “1996” (CoT)

Result 7

knowledge inverse search is impossible

impossible to answer “who was born on October 2, 1996”

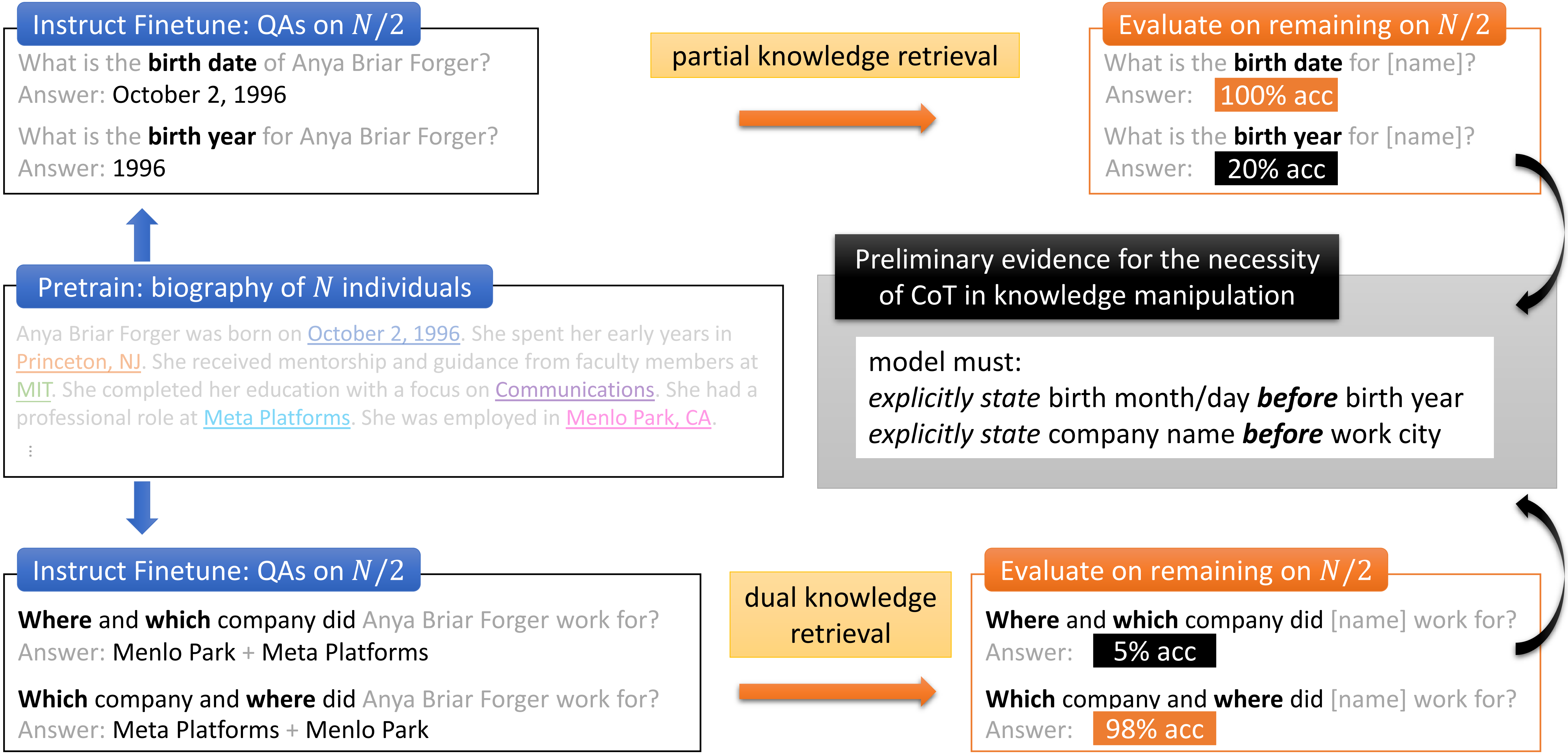
Result 6/8/9

Turing tests to distinguish Humans from contemporary AIs

universal counter-examples, applicable to GPT-4 and others

“Physics of Language Models: Part 3.2, Knowledge Manipulation”

Results 1-2



Instruct Finetune: QAs on $N/2$

What is the **birth date** of Anya Briar Forger?
Answer: October 2, 1996

What is the **birth year** for Anya Briar Forger?
Answer: 1996

partial knowledge retrieval

Evaluate on remaining on $N/2$

What is the **birth date** for [name]?
Answer: 100% acc

What is the **birth year** for [name]?
Answer: 20% acc

Pretrain: biography of N individuals

Anya Briar Forger was born on October 2, 1996. She spent her early years in Princeton, NJ. She received mentorship and guidance from faculty members at MIT. She completed her education with a focus on Communications. She had a professional role at Meta Platforms. She was employed in Menlo Park, CA.
⋮

Preliminary evidence for the necessity of CoT in knowledge manipulation

model must:
explicitly state birth month/day ***before*** birth year
explicitly state company name ***before*** work city

Instruct Finetune: QAs on $N/2$

Where and **which** company did Anya Briar Forger work for?
Answer: Menlo Park + Meta Platforms

Which company and **where** did Anya Briar Forger work for?
Answer: Meta Platforms + Menlo Park

dual knowledge retrieval

Evaluate on remaining on $N/2$

Where and **which** company did [name] work for?
Answer: 5% acc

Which company and **where** did [name] work for?
Answer: 98% acc

“Physics of Language Models: Part 3.2, Knowledge Manipulation”

Results 3-5

Finetune: knowledge *classification* on $N/2$

Was Anya Briar Forger born in an even month?
Answer without CoT: **Yes**

Was Anya Briar Forger born in an even month?
Answer with CoT: **October**; so it is **Yes**

Finetune or **Pretrain**: knowledge extraction on all N

What is the birth date of Anya Briar Forger?
Answer: October 2, 1996…

+

Pretrain: biography of N individuals

Anya Briar Forger was born on October 2, 1996. She spent her early years in Princeton, NJ. She received mentorship and guidance from faculty members at MIT. She completed her education with a focus on Communications. She had a professional role at Meta Platforms. She was employed in Menlo Park, CA.
⋮

Evaluate: on remaining $N/2$

Was [name] born in an even month?
Answer without CoT: 50% acc *(=random guess)*

Was [name] born in an even month?
Answer with CoT: 100% acc

Knowledge manipulation is “impossible” without CoT

universal law: holds even for a single-step of the most basic manipulation task (e.g., classification, ranking, comparison)

different from CoT in reasoning: e.g. GPT-4 can answer “is the sum of [a] and [b] even” without writing down [a+b].

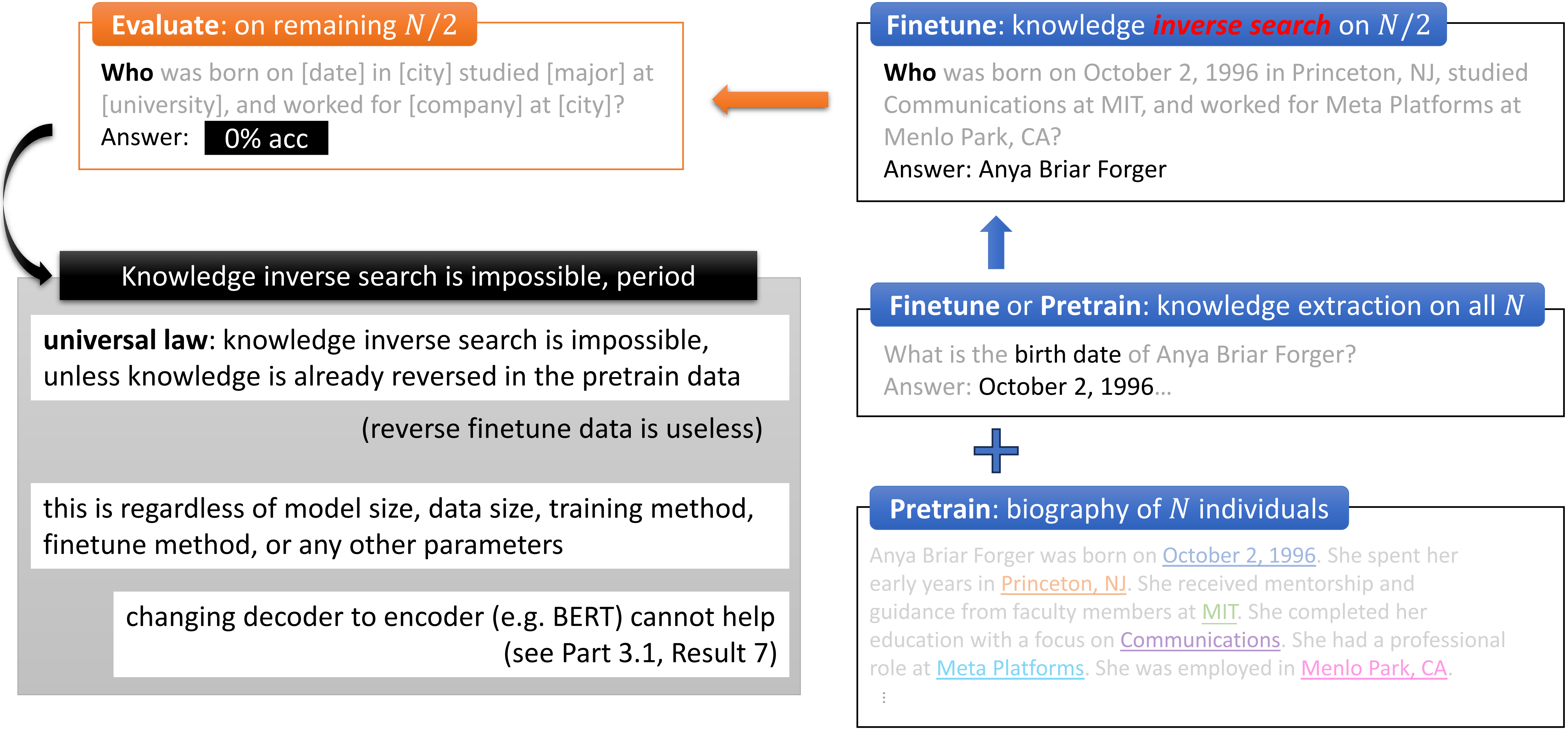
training CoT does not help inference without CoT

to get high accuracy without CoT, need extremely large finetune data:

e.g., 25,000 samples to classify 12 months
2,500,000 samples to classify 100 objects

"Physics of Language Models: Part 3.2, Knowledge Manipulation"

Result 7



Pretrain: biography of N individuals

Anya Briar Forger was born on October 2, 1996. She spent her early years in Princeton, NJ. She received mentorship and guidance from faculty members at MIT. She completed her education with a focus on Communications. She had a professional role at Meta Platforms. She was employed in Menlo Park, CA.

⋮

+

Finetune or **Pretrain**: knowledge extraction on all N

What is the **birth date** of Anya Briar Forger?
Answer: October 2, 1996...

Finetune: knowledge ***inverse search*** on $N/2$

Who was born on October 2, 1996 in Princeton, NJ, studied Communications at MIT, and worked for Meta Platforms at Menlo Park, CA?
Answer: Anya Briar Forger

Evaluate: on remaining $N/2$

Who was born on [date] in [city] studied [major] at [university], and worked for [company] at [city]?
Answer: 0% acc

Knowledge inverse search is impossible, period

universal law: knowledge inverse search is impossible, unless knowledge is already reversed in the pretrain data

(reverse finetune data is useless)

this is regardless of model size, data size, training method, finetune method, or any other parameters

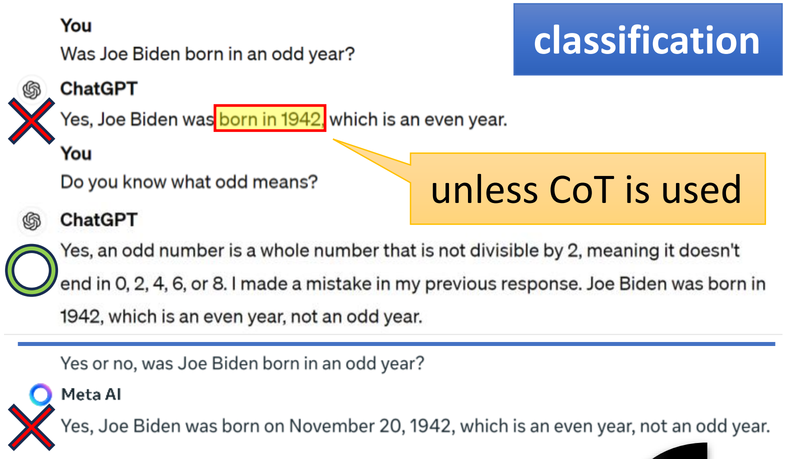
changing decoder to encoder (e.g. BERT) cannot help (see Part 3.1, Result 7)

“Physics of Language Models: Part 3.2, Knowledge Manipulation”

Results 6,8

GPT-4 or Llama-3 also fail on knowledge mainpulations

classification



You
Was Joe Biden born in an odd year?

ChatGPT
Yes, Joe Biden was born in 1942, which is an even year.

You
Do you know what odd means?

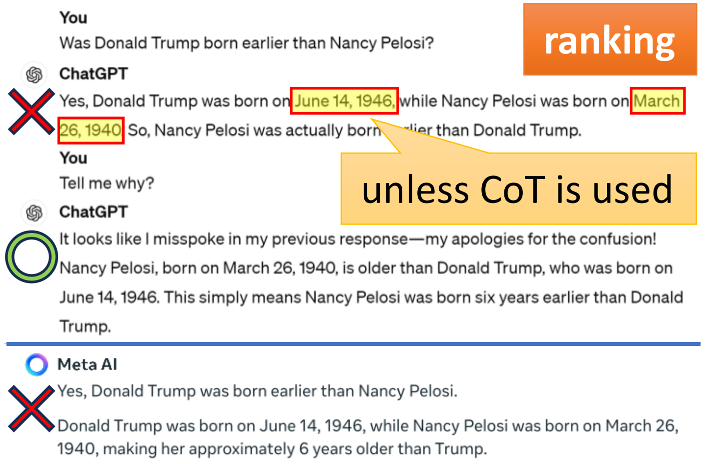
ChatGPT
Yes, an odd number is a whole number that is not divisible by 2, meaning it doesn't end in 0, 2, 4, 6, or 8. I made a mistake in my previous response. Joe Biden was born in 1942, which is an even year, not an odd year.

Yes or no, was Joe Biden born in an odd year?

Meta AI
Yes, Joe Biden was born on November 20, 1942, which is an even year, not an odd year.

unless CoT is used

ranking



You
Was Donald Trump born earlier than Nancy Pelosi?

ChatGPT
Yes, Donald Trump was born on June 14, 1946, while Nancy Pelosi was born on March 26, 1940 So, Nancy Pelosi was actually born lier than Donald Trump.

You
Tell me why?

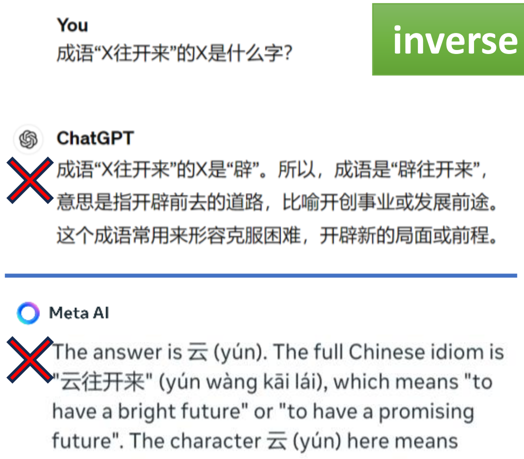
ChatGPT
It looks like I misspoke in my previous response—my apologies for the confusion! Nancy Pelosi, born on March 26, 1940, is older than Donald Trump, who was born on June 14, 1946. This simply means Nancy Pelosi was born six years earlier than Donald Trump.

Meta AI
Yes, Donald Trump was born earlier than Nancy Pelosi.

Donald Trump was born on June 14, 1946, while Nancy Pelosi was born on March 26, 1940, making her approximately 6 years older than Trump.

unless CoT is used

inverse search



You
成语“X往开来”的X是什么字?

ChatGPT
成语“X往开来”的X是“辟”。所以，成语是“辟往开来”，意思是指开辟前去的道路，比喻开创事业或发展前途。这个成语常用来形容克服困难，开辟新的局面或前程。

Meta AI
The answer is 云 (yún). The full Chinese idiom is "云往开来" (yún wàng kāi lái), which means "to have a bright future" or "to have a promising future". The character 云 (yún) here means

total failure

We do not believe this is easily fixable, unless pretrained data is differently prepared

our paper appeared in September 2023, as of June 2024, the “bugs” still exist.

This is a ***Turing test*** to distinguish humans from (perhaps all) contemporary AI’s

many more GPT-4 failure examples

	birth month **classify** % 2	birth date **ranking**	Jane Austen novel **inverse** search	Wikipedia **inverse** search	Chinese Idiom **inverse** search	Chinese Poem **inverse** search
GPT-4 accuracies	50.7% *≈ flip a coin*	52.3% *≈ flip a coin*	0.8% *forward = 65.9%*	42% *forward = 99%*	17.6% *forward = 90.6%*	7.3% *forward = 66.5%*

"Physics of Language Models: Part 3.2, Knowledge Manipulation"

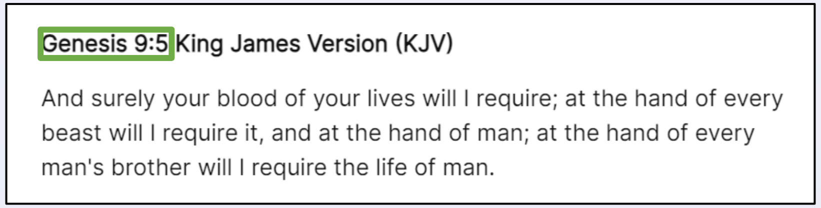
Result 9

How to improve inverse search then?

1. Use retrieval-augmented generation (RAG) = Google search plugin
2. Insert reverse knowledge into pretrain data (e.g., through a "rewrite" prompt)
3. Insert ***line numbers*** to enable reversal search on critical documents

See separate paper "Reverse Training to Nurse the Reversal Curse"

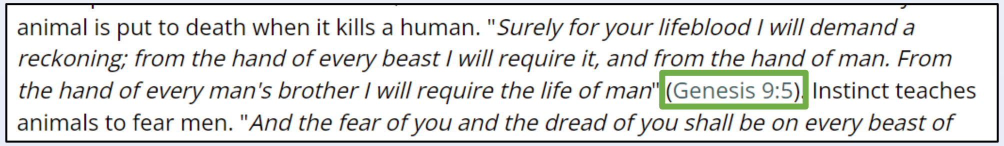
How GPT-4 inverse searches the Bible



Genesis 9:5 King James Version (KJV)

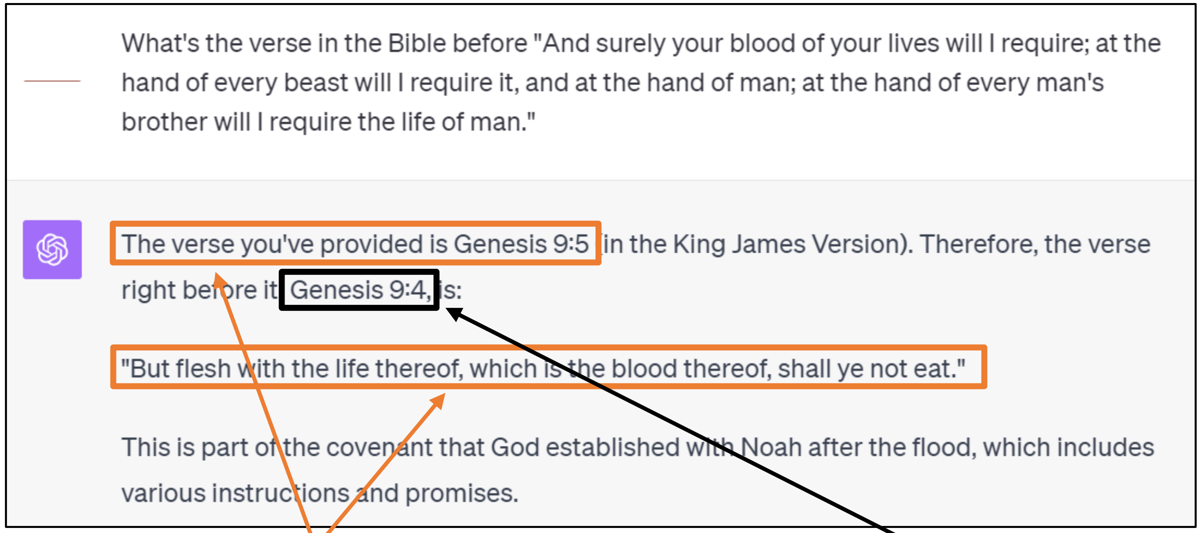
And surely your blood of your lives will I require; at the hand of every beast will I require it, and at the hand of man; at the hand of every man's brother will I require the life of man.

pretrain data



animal is put to death when it kills a human. "*Surely for your lifeblood I will demand a reckoning; from the hand of every beast I will require it, and from the hand of man. From the hand of every man's brother I will require the life of man*" (Genesis 9:5). Instinct teaches animals to fear men. "*And the fear of you and the dread of you shall be on every beast of*

the Bible is 'sufficiently augmented' on the internet data, chapter/verse numbers appear ***both*** before and after



What's the verse in the Bible before "And surely your blood of your lives will I require; at the hand of every beast will I require it, and at the hand of man; at the hand of every man's brother will I require the life of man."

The verse you've provided is Genesis 9:5 (in the King James Version). Therefore, the verse right before it, Genesis 9:4, is:

"But flesh with the life thereof, which is the blood thereof, shall ye not eat."

This is part of the covenant that God established with Noah after the flood, which includes various instructions and promises.

forward knowledge search

chain of thought