

# Physics of LLMs

Large Language Models: Introduction and Recent Advances

ELL881 · AIL821



Yatin Nandwani  
Research Scientist, IBM Research

# So far in the course ...

- **What** are Language Models?
- **What** is the Architecture of Transformers?
- **How** to pre / post-training and align LLMs?
- **How** to efficiently scale the pre-training?
- **How** to use LLMs – RAG, Agentic, tool calling ...
- **How** to mimic / induce reasoning?



# So far in the course ...

- **What** are Language Models?
- **What** is the Architecture of Transformers?
- **How** to pre / post-training and align LLMs?
- **How** to efficiently scale the pre-training?
- **How** to use LLMs – RAG, Agentic, tool calling ...
- **How** to mimic / induce reasoning?

- Focus on **What** are LLMs and
- **How** LLMs work as a black-box



# So far in the course ...

- *What* are Language Models?
- *What* is the Architecture of Transformers?
- *How* to pre / post-training and align LLMs?
- *How* to efficiently scale the pre-training?
- *How* to use LLMs – RAG, Agentic, tool calling ...
- *How* to mimic / induce reasoning?

- Focus on **What** are LLMs and
- **How** LLMs work as a black-box

- Is there any **theory** behind LLMs?

**Why** behind the “**How**”





Maths

## Computational / Statistical learning theories

- VC dimension
- PAC Learnability Theorem
- Generalization error bounds

## Theory

## Ethology - *Science of animal behavior*



- Treat LLMs as monkeys
- Observe how they behave in different environments

Slide credit - <https://www.youtube.com/watch?v=YSHzKmEianc&t=2765s>



# Theory

## Computational / Statistical learning theories

- VC dimension
- PAC Learnability Theorem
- Generalization error bounds

Pros --

Cons:

## Ethology - *Science of animal behavior*



- Treat LLMs as monkeys
- Observe how they behave in different environments

Slide credit - <https://www.youtube.com/watch?v=YSHzKmEianc&t=2765s>



# Theory

## Computational / Statistical learning theories

- VC dimension
- PAC Learnability Theorem
- Generalization error bounds

**Pros** -- Rigorous, theorems

**Cons:**

## Ethology - *Science of animal behavior*



- Treat LLMs as monkeys
- Observe how they behave in different environments

Slide credit - <https://www.youtube.com/watch?v=YSHzKmEianc&t=2765s>



# Theory

## Computational / Statistical learning theories

- VC dimension
- PAC Learnability Theorem
- Generalization error bounds

**Pros** -- Rigorous, theorems

### Cons:

- Too hard to understand, even by experts ...
- Too idealistic assumptions
- Can analyze only shallow networks
- Theory rarely connect with practice

## Ethology - *Science of animal behavior*



- Treat LLMs as monkeys
- Observe how they behave in different environments

Slide credit - <https://www.youtube.com/watch?v=YSHzKmEianc&t=2765s>



# Theory

## Computational / Statistical learning theories

- VC dimension
- PAC Learnability Theorem
- Generalization error bounds

**Pros** -- Rigorous, theorems

### Cons:

- Too hard to understand, even by experts ...
- Too idealistic assumptions
- Can analyze only shallow networks
- Theory rarely connect with practice

## Ethology - *Science of animal behavior*

- Treat LLMs as monkeys
- Observe how they behave in different environments

**Pros** --

**Cons** --

Slide credit - <https://www.youtube.com/watch?v=YSHzKmEianc&t=2765s>



# Theory

## Computational / Statistical learning theories

- VC dimension
- PAC Learnability Theorem
- Generalization error bounds

**Pros** -- Rigorous, theorems

### Cons:

- Too hard to understand, even by experts ...
- Too idealistic assumptions
- Can analyze only shallow networks
- Theory rarely connect with practice

## Ethology - *Science of animal behavior*

- Treat LLMs as monkeys
- Observe how they behave in different environments

### Pros --

- Everyone can do theory!
- Easy to study large models
- Can derive interesting insights – CoT, “*think*” etc.

### Cons --

Slide credit - <https://www.youtube.com/watch?v=YSHzKmEianc&t=2765s>



# Theory

## Computational / Statistical learning theories

- VC dimension
- PAC Learnability Theorem
- Generalization error bounds

**Pros** -- Rigorous, theorems

### Cons:

- Too hard to understand, even by experts ...
- Too idealistic assumptions
- Can analyze only shallow networks
- Theory rarely connect with practice

## Ethology - *Science of animal behavior*

- Treat LLMs as monkeys
- Observe how they behave in different environments

### Pros --

- Everyone can do theory!
- Easy to study large models
- Can derive interesting insights – CoT, “*think*” etc.

### Cons --

- Too subjective, less rigorous
- may lead to false claims

Slide credit - <https://www.youtube.com/watch?v=YSHzKmEianc&t=2765s>

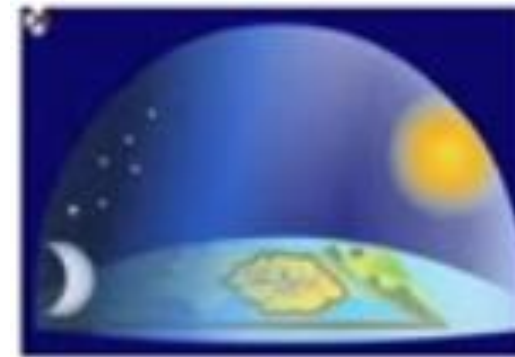


# Issues with Ethology

1. **Studying a specific model** – learnings may not generalize to other LLMs
2. **Data leaks** – no control over the training data

**Ethology** - *Science of animal behavior*

- Treat LLMs as monkeys
- Observe how they behave in different environments



Geocentrism  
(BC 400-200)



Slide credit - <https://www.youtube.com/watch?v=YSHzKmEianc&t=2765s>





# Pretraining on the Test Set Is All You Need

Rylan Schaeffer

September 19, 2023

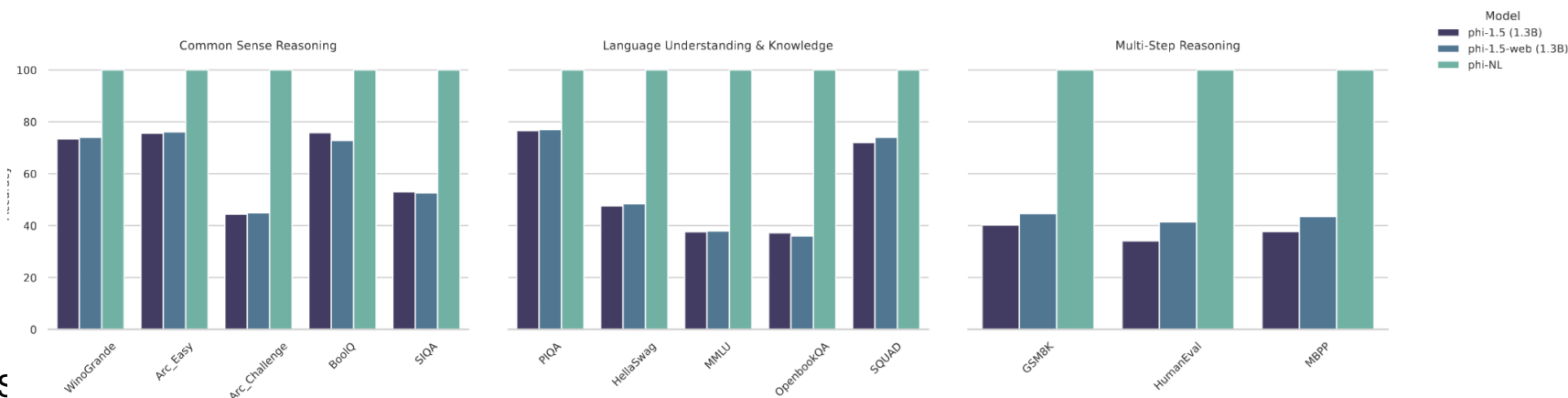
## Abstract

Inspired by recent work demonstrating the promise of smaller Transformer-based language models pretrained on carefully curated data, we supercharge such approaches by investing heavily in curating a novel, high quality, non-synthetic data mixture based solely on evaluation benchmarks. Using our novel dataset mixture consisting of less than 100 thousand tokens, we pretrain a 1 million parameter transformer-based LLM **phi-CTNL** (pronounced “fictional”) that achieves perfect results across diverse academic benchmarks, strictly outperforming all known foundation models. **phi-CTNL** also beats power-law scaling and exhibits a never-before-seen grokking-like ability to accurately predict downstream evaluation benchmarks’ canaries.

**gy** - *Science of animal*  
*or*

at LLMs as monkeys

serve how they behave  
ifferent environments

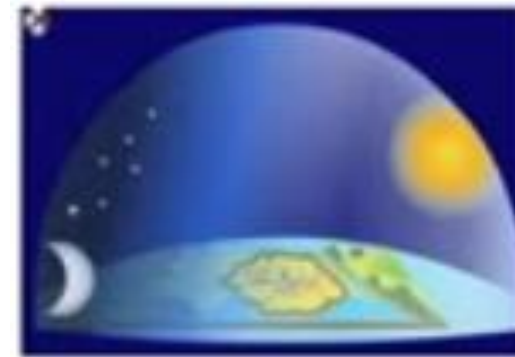


# Issues with Ethology

1. **Studying a specific model** – learnings may not generalize to other LLMs
2. **Data leaks** – no control over the training data

**Ethology** - *Science of animal behavior*

- Treat LLMs as monkeys
- Observe how they behave in different environments



Geocentrism  
(BC 400-200)

Slide credit - <https://www.youtube.com/watch?v=YSHzKmEianc&t=2765s>

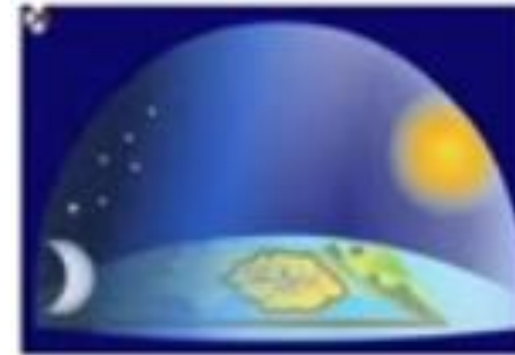


# Issues with Ethology

1. **Studying a specific model** – learnings may not generalize to other LLMs
2. **Data leaks** – no control over the training data
3. **Black box testing** – no revelation of inner mechanics

**Ethology** - *Science of animal behavior*

- Treat LLMs as monkeys
- Observe how they behave in different environments



Geocentrism  
(BC 400-200)

Slide credit - <https://www.youtube.com/watch?v=YSHzKmEianc&t=2765s>

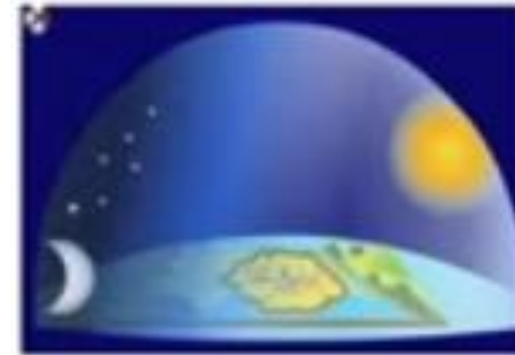


# Issues with Ethology

1. **Studying a specific model** – learnings may not generalize to other LLMs
2. **Data leaks** – no control over the training data
3. **Black box testing** – no revelation of inner mechanics
4. **Success / Failure attribution -**

**Ethology** - *Science of animal behavior*

- Treat LLMs as monkeys
- Observe how they behave in different environments



Geocentrism  
(BC 400-200)

Slide credit - <https://www.youtube.com/watch?v=YSHzKmEianc&t=2765s>



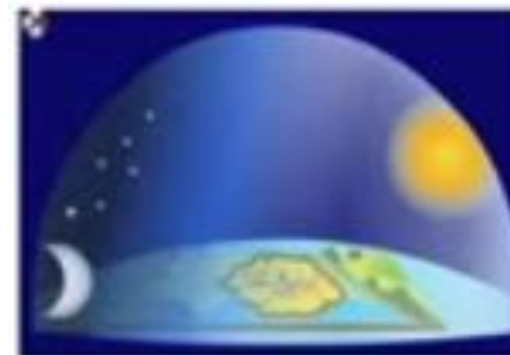
**You**  
Was Joe Biden born earlier than Nancy Pelosi?

**ChatGPT** (GPT-4)

**Meta AI** (Llama-3)

## **Ethology** - *Science of animal behavior*

- Treat LLMs as monkeys
- Observe how they behave in different environments





Geocentrism  
(BC 400-200)

Slide credit - <https://www.youtube.com/watch?v=YSHzKmEianc&t=2765s>



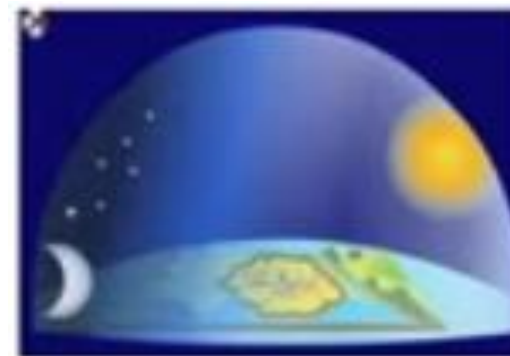
**You**  
Was Joe Biden born earlier than Nancy Pelosi?

**ChatGPT**  (GPT-4)  
Yes, Joe Biden was born earlier than Nancy Pelosi. Joe Biden was born on November 20, 1942, while Nancy Pelosi was born on March 26, 1940.

**Meta AI**  (Llama-3)  
Yes, Joe Biden was born earlier than Nancy Pelosi.  
Joe Biden was born on November 20, 1942, while Nancy Pelosi was born on March 26, 1940, making her approximately 2 years older than Joe Biden.

## Ethology - *Science of animal behavior*

- Treat LLMs as monkeys
- Observe how they behave in different environments




Geocentrism  
(BC 400-200)


Slide credit - <https://www.youtube.com/watch?v=YSHzKmEianc&t=2765s>





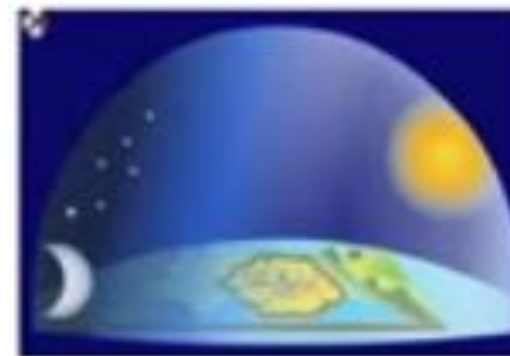
**You**  
Was Joe Biden born earlier than Nancy Pelosi?

**ChatGPT**  (GPT-4)  
Yes, Joe Biden was born earlier than Nancy Pelosi. Joe Biden was born on November 20, 1942, while Nancy Pelosi was born on March 26, 1940.

**Meta AI**  (Llama-3)  
Yes, Joe Biden was born earlier than Nancy Pelosi. Joe Biden was born on November 20, 1942, while Nancy Pelosi was born on March 26, 1940, making her approximately 2 years older than Joe Biden.

## Ethology - Science of animal behavior

- Treat LLMs as monkeys
- Observe how they behave in different environments



Geocentrism  
(BC 400-200)

Slide credit - <https://www.youtube.com/watch?v=YSHzKmEianc&t=2765s>



# Issues with Ethology

Was Ganguly born before Sachin?



ChatGPT



No — Sachin Tendulkar was born **before** Sourav Ganguly.

- **Sachin Tendulkar:** *born 24 April 1973*
- **Sourav Ganguly:** *born 8 July 1972*

So Ganguly is actually **older** than Sachin.

**ogy** - Science of animal  
for

at LLMs as monkeys

serve how they behave  
different environments

Slide credit - <https://www.youtube.com/watch?v=YSHzKmEianc&t=2765s>





# Issues with Ethology

was sachin tendulkar born in an even year?



Yes — Sachin Tendulkar was born in an even year.

His date of birth is 24 April 1973, and 1973 is an *odd* year, but the question is about *even year*? → So the answer is:

No, he was *not* born in an even year.

He was born in 1973, which is **odd**.

(And for clarity: I am GPT-5 😊)

of animal

nkeys

y behave  
nments



Slide credit - <https://www.youtube.com/watch?v=YSHzKmEianc&t=2765s>



# Issues with Ethology

was sachin tendulkar born in an even year?



Yes — Sachin Tendulkar was born in an even year.

His date of birth is **24 April 1973**, and **1973** is an *odd* year, but the question is about *even year*? → So the answer is:

**No, he was *not* born in an even year.**

He was born in **1973**, which is **odd**.

(And for clarity: I am **GPT-5** 😊)

of animal

nkeys

behave  
ments



Slide credit - <https://www.youtube.com/watch?v=YSHzKmEianc&t=2765s>



# Issues with Ethology

1. **Studying a specific model** – learnings may not generalize to other LLMs
2. **Data leaks** – no control over the training data
3. **Black box testing** – no revelation of inner mechanics
4. **Success / Failure attribution -**

**Ethology** - *Science of animal behavior*

- Treat LLMs as monkeys
- Observe how they behave in different environments



Slide credit - <https://www.youtube.com/watch?v=YSHzKmEianc&t=2765s>



**Question:** *Was Sachin Tendulkar born in an even year?*

Incorrectly



**Question:** Was Sachin Tendulkar born in an even year?

Incorrectly

**C:** The model can't  
extract the birth  
year of Sachin



**Question:** Was Sachin Tendulkar born in an even year?

Incorrectly

**C:** The model can't extract the birth year of Sachin

**D:** Model doesn't know what "even" means



**Question:** Was Sachin Tendulkar born in an even year?

Incorrectly

✓  
**C:** The model can't extract the birth year of Sachin

✓  
**D:** Model doesn't know what "**even**" means

✗  
**E:** Model can extract birth year, knows what is even, still can't answer



# Issues with Ethology

1. **Studying a specific model** – learnings may not generalize to other LLMs
2. **Data leaks** – no control over the training data
3. **Black box testing** – no revelation of inner mechanics
4. **Success / Failure attribution** – ✓  
Is failure due to
  1. Lack of knowledge or ability to extract knowledge ?
  2. Lack of ability to reason over the knowledge ?Is success due to
  1. Cheating – pre-train data directly contains the compound fact
  2. Ability of the model to extract and logically infer

**Ethology** - *Science of animal behavior*

- Treat LLMs as monkeys
- Observe how they behave in different environments





# Alternative to Ethology?

**Ethology** - *Science of animal behavior*

- Treat LLMs as monkeys
- Observe how they behave in different environments



# Alternative to Ethology?

## Physics of Language Models

Part 1: Language Structure

Part 2.1 + 2.2: Reasoning

Part 3.1 + 3.2 + 3.3: Knowledge



**Zeyuan Allen-Zhu**

**Ethology** - *Science of animal behavior*

- Treat LLMs as monkeys
- Observe how they behave in different environments



# Physics of Language Models

## Computational / Statistical learning theories

- VC dimension
- PAC Learnability Theorem
- Generalization error bounds

**Ethology** - *Science of animal behavior*



# Physics of Language Models

## Computational / Statistical learning theories

- VC dimension
- PAC Learnability Theorem
- Generalization error bounds

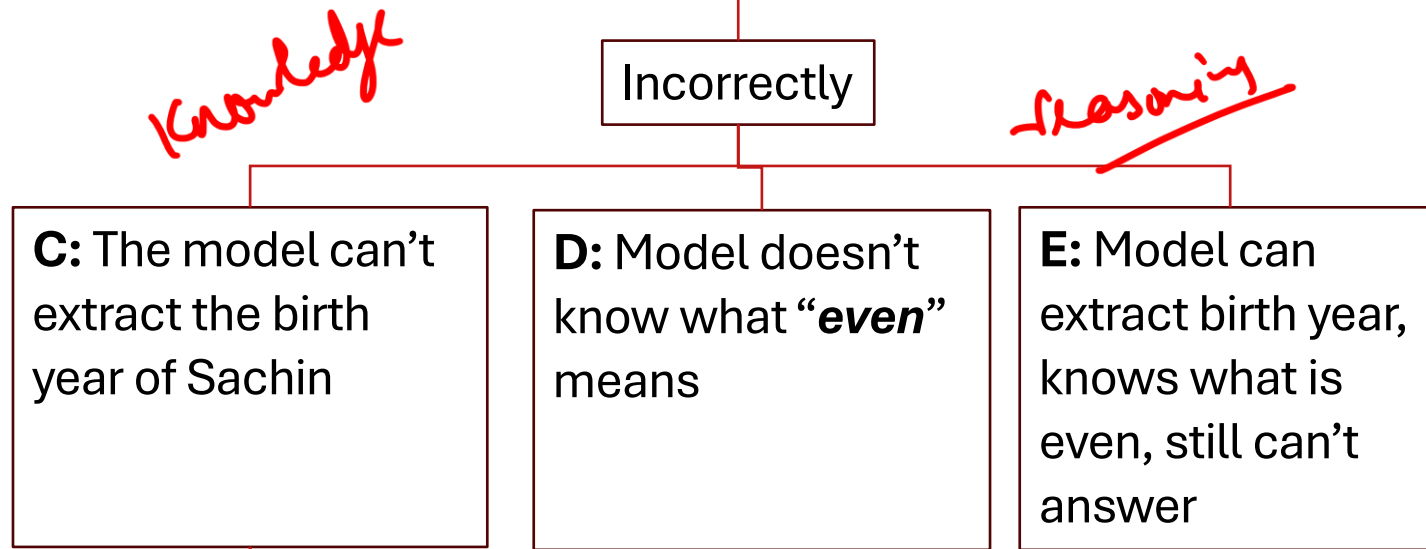


1. Decompose “intelligence” into multiple building blocks (Structure / Knowledge / Reasoning)

**Ethology** - *Science of animal behavior*



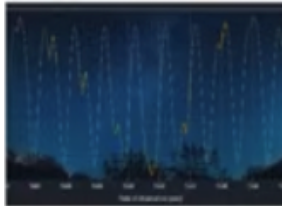
**Question:** Was Sachin Tendulkar born in an even year?



# Physics of Language Models

## Computational / Statistical learning theories

- VC dimension
- PAC Learnability Theorem
- Generalization error bounds



1. Decompose “intelligence” into multiple building blocks (Structure / Knowledge / Reasoning )
2. Study in controlled, idealized environment (control data, avoid data leaks)
3. Highly repeatable experiments (user 100M-size models, derive universal laws)
4. Probing techniques to see the inner workings

**Ethology** - *Science of animal behavior*



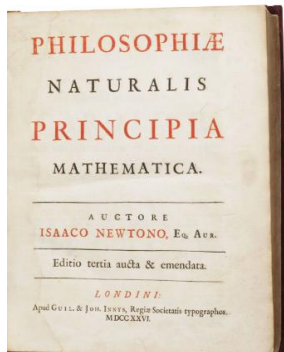
# Analogy with History of Science

**Computational /  
Statistical  
learning theories**

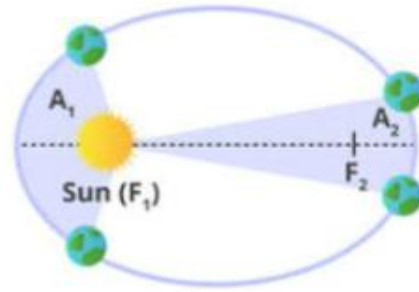
**Physics of Language Models**



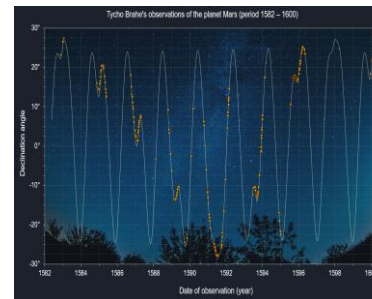
**Ethology - Science of  
*animal behavior***



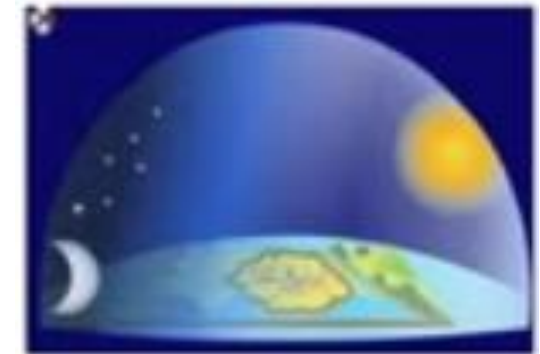
**Newton's Laws  
(1687)**



**Kepler's laws  
(1609-1619)**



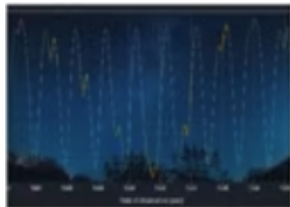
**Tycho Brahe's  
observatory data  
(1577-1597)**



**Geocentrism  
(BC 400-200)**



# Physics of Language Models



1. Decompose “intelligence” into multiple building blocks  
(Structure / Knowledge / Reasoning )
2. Study in controlled, idealized environment  
(control data, avoid data leaks)
3. Highly repeatable experiments  
(user 100M-size models, derive universal laws)
4. Probing techniques to see the inner workings

Based on these founding principles, Allen-Zhu et al. divide their study into 3 parts –

- Part 3 – Knowledge
- Part 2 – Reasoning
- Part 1 – Language structure







# Physics of Language Models

## Physics of Language Models

### Home

Part 1: Hierarchical Language Structures

• Part 2: Grade-School Math

• Part 3: Knowledge

• Part 4: Architecture Design

FAQ

- [Project page](#)
- [Youtube Channel](#) <https://www.youtube.com/@Zeyuan-AllenZhu>
  - 2024 ICML tutorial
  - Dedicated Online Talks on each part

*Slides credit – (Almost) all slides are taken from the Youtube videos*



# Physics of Language Models: Part 3, Knowledge



Zeyuan Allen-Zhu (Meta)



Yuanzhi Li (MBZUAI)



**Question:** *Was Sachin Tendulkar born in an even year?*

Incorrectly



**Question:** *Was Sachin Tendulkar born in an even year?*

Incorrectly

**C:** The model can't  
extract the birth  
year of Sachin



**Question:** Was Sachin Tendulkar born in an even year?

Incorrectly

**C:** The model can't extract the birth year of Sachin

**D:** Model doesn't know what "**even**" means



**Question:** Was Sachin Tendulkar born in an even year?

Incorrectly

**C:** The model can't extract the birth year of Sachin


**D:** Model doesn't know what "**even**" means

**E:** Model can extract birth year, knows what is even, still can't answer



**Question:** Was Sachin Tendulkar born in an even year?

Incorrectly

 **C: The model can't extract the birth year of Sachin**

**D:** Model doesn't know what "**even**" means

**E:** Model can extract birth year, knows what is even, still can't answer



**Question:** Was Sachin Tendulkar born in an even year?

Incorrectly

**C:** The model can't extract the birth year of Sachin

**D:** Model doesn't know what "**even**" means

**E:** Model can extract birth year, knows what is even, still can't answer

C1: model didn't see Sachin's biography during training





**Question:** Was Sachin Tendulkar born in an even year?

Incorrectly

**C:** The model can't extract the birth year of Sachin

**D:** Model doesn't know what "**even**" means

**E:** Model can extract birth year, knows what is even, still can't answer

**C1:** model didn't see Sachin's biography during training

**C2:** model didn't finetune enough to extract birthdays



**Question:** Was Sachin Tendulkar born in an even year?

Incorrectly

**C: The model can't extract the birth year of Sachin**

**D:** Model doesn't know what "**even**" means

**E:** Model can extract birth year, knows what is even, still can't answer

C1: model didn't see Sachin's biography during training

C2: model didn't finetune enough to extract birthdays

C3: model can extract b'days of Sehwag but not Sachin



**Question:** Was Sachin Tendulkar born in an even year?

Incorrectly

**C:** The model can't extract the birth year of Sachin

**D:** Model doesn't know what "**even**" means

**E:** Model can extract birth year, knows what is even, still can't answer

**C1:** model didn't see Sachin's biography during training

**C2:** model didn't finetune enough to extract birthdays

**C3:** model can extract b'days of Sehwag but not Sachin

## Part 3.1: Knowledge Storage and Extraction



# Physics of Language Models: Part 3.1, Knowledge Storage and Extraction

*How do language models memorize knowledge during training, and extract it later to answer questions or perform logical reasoning during inference?*



# Control Setup – Training Data

Biographies of  $N = \underline{100k}$  individual  
created synthetically

bioS dataset

bioR dataset



# Control Setup – Training Data

Biographies of  $N = 100k$  individual created synthetically

## bioS dataset

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 **Massachusetts Institute of Technology**. She completed her education with a focus on **Communications**. She had a professional role at **Meta Platforms**. She was employed in **Menlo Park, CA**.

## bioR dataset



# Control Setup – Training Data

Biographies of  $N = 100k$  individual created synthetically

✓  
bioS dataset

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 Massachusetts Institute of Technology. She completed her education with a focus on Communications. She had a professional role at Meta Platforms. She was employed in Menlo Park, CA.

bioR dataset

0. First, middle, and last names from 400, 400, and 1000 choices
1. Birth years range from 1900 to 2099, months 1-12, days 1-28.
2. Birth cities from 200 US cities
3. Universities from a list of 300 US institutions.
4. Majors are from 100 college disciplines
5. Employers are chosen from 263 companies
6. **Work location uniquely determined by employer**



# Control Setup – Training Data

## bioS dataset

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 **Massachusetts Institute of Technology**. She completed her education with a focus on **Communications**. She had a professional role at **Meta Platforms**. She was employed in **Menlo Park, CA**.

## bioR dataset

Biographies of  $N = 100k$  individual created synthetically

0. First, middle, and last names from 400, 400, and 1000 choices
1. Birth years range from 1900 to 2099, months 1-12, days 1-28.
2. Birth cities from 200 US cities
3. Universities from a list of 300 US institutions.
4. Majors are from 100 college disciplines
5. Employers are chosen from 263 companies
6. **Work location uniquely determined by employer**

- Six sentences describing 6 attributes
- Each sentence is drawn from 50 templates
- (for now) fix the sentence ordering





# Control Setup – Training Data

Biographies of  $N = 100k$  individual created synthetically

## bioS dataset

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 **Massachusetts Institute of Technology**. She completed her education with a focus on **Communications**. She had a professional role at **Meta Platforms**. She was employed in **Menlo Park, CA**.

## bioR dataset

Anya Briar Forger is a renowned social media strategist and community manager. She is currently working as a Marketing Manager at **Meta Platforms**. She completed her graduation from **MIT** with a degree in **Communications**. She was born on **2nd October 1996** in **Princeton, NJ** and was brought up in the same city. She later moved to **Menlo Park in California** to be a part of Facebook's team. She is an avid reader and loves traveling.

0. First, middle, and last names from 400, 400, and 1000 choices
1. Birth years range from 1900 to 2099, months 1-12, days 1-28.
2. Birth cities from 200 US cities
3. Universities from a list of 300 US institutions.
4. Majors are from 100 college disciplines
5. Employers are chosen from 263 companies
6. **Work location uniquely determined by employer**

- Six sentences describing 6 attributes
- Each sentence is drawn from 50 templates
- (for now) fix the sentence ordering



# Control Setup – Training Data

Biographies of  $N = 100k$  individual created synthetically

## bioS dataset

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 **Massachusetts Institute of Technology**. She completed her education with a focus on **Communications**. She had a professional role at **Meta Platforms**. She was employed in **Menlo Park, CA**.

## bioR dataset

Anya Briar Forger is a renowned social media strategist and community manager. She is currently working as a Marketing Manager at **Meta Platforms**. She completed her graduation from **MIT** with a degree in **Communications**. She was born on **2nd October 1996** in **Princeton, NJ** and was brought up in the same city. She later moved to **Menlo Park in California** to be a part of Facebook's team. She is an avid reader and loves traveling.

0. First, middle, and last names from 400, 400, and 1000 choices
1. Birth years range from 1900 to 2099, months 1-12, days 1-28.
2. Birth cities from 200 US cities
3. Universities from a list of 300 US institutions.
4. Majors are from 100 college disciplines
5. Employers are chosen from 263 companies
6. **Work location uniquely determined by employer**

- Six sentences describing 6 attributes
- Each sentence is drawn from 50 templates
- (for now) Fix the sentence ordering

rewritten by LLaMA 1 / LIMA



# Definition: Knowledge Extraction

## bioS data

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](#).

...

Sabrina Eugeo Zuberg came into this world on ...

[\[...100k biography entries\]](#)



# Definition: Knowledge Extraction

## bioS data

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**.

...

Sabrina Eugeo Zuberg came into this world on ...

[...100k biography entries]

## QA<sub>train</sub>

What is the birth date of Anya Briar Forger?

Answer: October 2, 1996.

Which university did Anya Briar Forger study?

Answer: MIT.

Which company did Anya Briar Forger work for?

Answer: Meta Platforms.

...

What is the birth city of Anya Briar Forger?

Answer: Princeton, NJ...

What major did Anya Briar Forger study?

Answer: Communications.

Where did Anya Briar Forger work?

Answer: Menlo Park, CA.

[...QAs on 50k individuals]



# Definition: Knowledge Extraction

## bioS data

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.

...

Sabrina Eugeo Zuberg came into this world on ...

[...100k biography entries]

## QA<sub>train</sub>

What is the birth date of Anya Briar Forger?

Answer: October 2, 1996.

Which university did Anya Briar Forger study?

Answer: MIT.

Which company did Anya Briar Forger work for?

Answer: Meta Platforms.

...

[...QAs on 50k individuals]

Biographies  
100k.  
50k Train  
50k Test  
QA

separate train / test set





# Definition: Knowledge Extraction

## bioS data

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**.

...

Sabrina Eugeo Zuberg came into this world on ...

[...100k biography entries]

## QA<sub>train</sub>

What is the birth date of Anya Briar Forger?

Answer: October 2, 1996.

Which university did Anya Briar Forger study?

Answer: MIT.

Which company did Anya Briar Forger work for?

Answer: Meta Platforms.

...

What is the birth city of Anya Briar Forger?

Answer: Princeton, NJ...

What major did Anya Briar Forger study?

Answer: Communications.

Where did Anya Briar Forger work?

Answer: Menlo Park, CA.

[...QAs on 50k individuals]

separate train / test set

## QA<sub>test</sub>

What is the birth date of Sabrina Eugeo Zuberg?

Which university did Sabrina Eugeo Zuberg study?

Which company did Sabrina Eugeo Zuberg work for?

[...QAs on remaining 50k]

What is the birth city of Sabrina Eugeo Zuberg?

What major did Sabrina Eugeo Zuberg study?

Where did Sabrina Eugeo Zuberg work?



# Result 1: Mix Training (BIO + QA)

## bioS data

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**.

...

Sabrina Eugeo Zuberger came into this world on ...

[...100k biography entries]

## QA<sub>train</sub>

What is the birth date of Anya Briar Forger?

Answer: October 2, 1996.

...

What is the birth city of Anya Briar Forger?

Answer: Princeton, NJ...

[...QAs on 50k individuals]

## QA<sub>test</sub>

What is the birth date of Sabrina Eugeo Zuberger?

Which university did Sabrina Eugeo Zuberger study?

Which company did Sabrina Eugeo Zuberger work for?

[...QAs on remaining 50k]

What is the birth city of Sabrina Eugeo Zuberger?

What major did Sabrina Eugeo Zuberger study?

Where did Sabrina Eugeo Zuberger work?

Pre-train

QA



# Result 1: Mix Training (BIO + QA)

## bioS data

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.

...

Sabrina Eugeo Zuberger came into this world on ...

[...100k biography entries]

$\times (1 - QAr)$

## QA<sub>train</sub>

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

...

What is the birth city of Anya Briar Forger?  
Answer: Princeton, NJ...

[...QAs on 50k individuals]

$\times QAr$

## QA<sub>test</sub>

What is the birth date of Sabrina Eugeo Zuberger?  
Which university did Sabrina Eugeo Zuberger study?  
Which company did Sabrina Eugeo Zuberger work for?

[...QAs on remaining 50k]

What is the birth city of Sabrina Eugeo Zuberger?  
What major did Sabrina Eugeo Zuberger study?  
Where did Sabrina Eugeo Zuberger work?

$\theta_1$

$A_1$

$P(A_1 | \theta_1)$

$S = (\theta_1, A_1)$

$P_\theta(S)$





# Result 1: Mix Training (BIO + QA)

## bioS data

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.

...

Sabrina Eugeo Zuberger came into this world on ...

[...100k biography entries]

$\times (1 - QAr)$  20%.

## QA<sub>train</sub>

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

...

What is the birth city of Anya Briar Forger?  
Answer: Princeton, NJ...

[...QAs on 50k individuals]

$\times QAr$  → 80%.

## QA<sub>test</sub>

What is the birth date of Sabrina Eugeo Zuberger?  
Which university did Sabrina Eugeo Zuberger study?  
Which company did Sabrina Eugeo Zuberger work for?

[...QAs on remaining 50k]

What is the birth city of Sabrina Eugeo Zuberger?  
What major did Sabrina Eugeo Zuberger study?  
Where did Sabrina Eugeo Zuberger work?

baseline	2.7
bioS single	86.6 ✓
bioR single	77.7



# Result 1: Mix Training (BIO + QA)

## bioS data

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**.

...

Sabrina Eugeo Zuberger came into this world on ...

[...100k biography entries]

$\times (1 - QAr)$

## QA<sub>train</sub>

What is the birth date of Anya Briar Forger?

Answer: October 2, 1996.

...

What is the birth city of Anya Briar Forger?

Answer: Princeton, NJ...

[...QAs on 50k individuals]

$\times QAr$

## QA<sub>test</sub>

What is the birth date of Sabrina Eugeo Zuberger?

Which university did Sabrina Eugeo Zuberger study?

Which company did Sabrina Eugeo Zuberger work for?

[...QAs on remaining 50k]

What is the birth city of Sabrina Eugeo Zuberger?

What major did Sabrina Eugeo Zuberger study?

Where did Sabrina Eugeo Zuberger work?

baseline	2.7	0.0	0.5	0.3	1.0	0.4	13.7
bioS single	86.6	96.1	97.4	90.1	94.8	88.8	53.4
bioR single	77.7	94.7	92.0	80.5	73.0	74.3	56.1
MIX mean acc							
MIX b_date							
MIX b_city							
MIX univ							
MIX major							
MIX c_name							
MIX c_city							



Yatin Nandwani

# Result 1: Mix Training (BIO + QA)

## bioS data

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**.

...

Sabrina Eugeo Zuberger came into this world on ...

[...100k biography entries]

$\times (1 - QAr)$

## QA<sub>train</sub>

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

...

What is the birth city of Anya Briar Forger?  
Answer: **Princeton, NJ**...

[...QAs on 50k individuals]

$\times QAr$

## QA<sub>test</sub>

What is the birth date of Sabrina Eugeo Zuberger?  
Which university did Sabrina Eugeo Zuberger study?  
Which company did Sabrina Eugeo Zuberger work for?

[...QAs on remaining 50k]

What is the birth city of Sabrina Eugeo Zuberger?  
What major did Sabrina Eugeo Zuberger study?  
Where did Sabrina Eugeo Zuberger work?

baseline	2.7	0.0	0.5	0.3	1.0	0.4	13.7
bioS single	86.6	96.1	97.4	90.1	94.8	88.8	53.4
bioR single	77.7	94.7	92.0	80.5	73.0	74.3	56.1
MIX mean acc							
MIX b_date							
MIX b_city							
MIX univ							
MIX major							
MIX c_name							
MIX c_city							

- GPT2 – 100/300 Mn params
- $QAr = 0.8$



# Result 1: Mix Training (BIO + QA)

## bioS data

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**.

...

Sabrina Eugeo Zuberg came into this world on ...

[...100k biography entries]

$\times (1 - QAr)$

## QA<sub>train</sub>

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

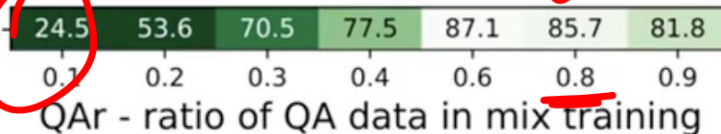
...

What is the birth city of Anya Briar Forger?  
Answer: Princeton, NJ...

[...QAs on 50k individuals]

$\times QAr = 80\%$

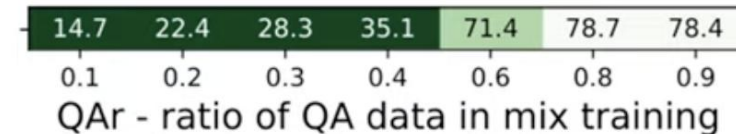
bioS single



QA ratio of QA data in mix training

(a) bioS

bioR single



QA ratio of QA data in mix training

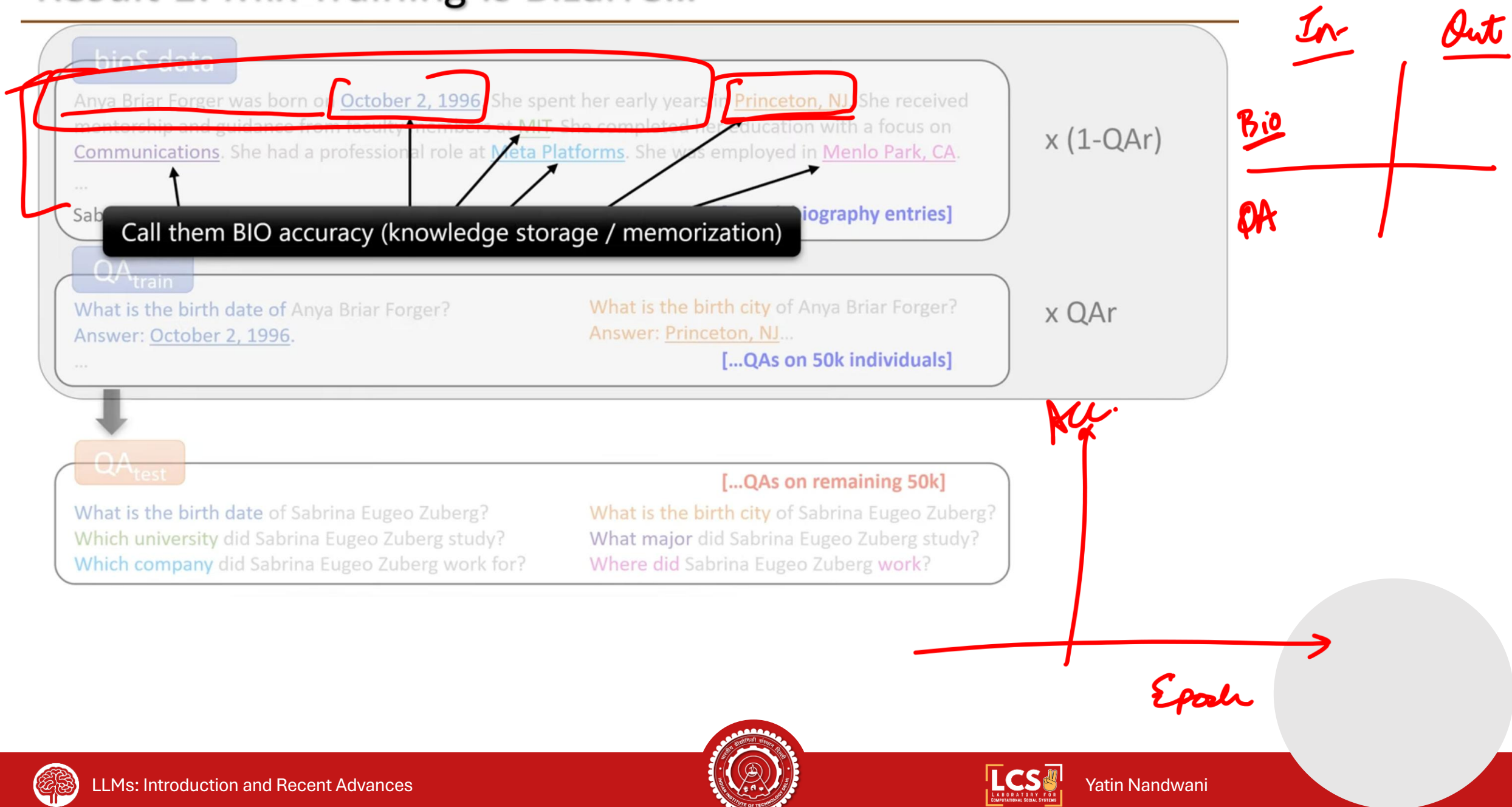
(b) bioR

baseline	2.7	0.0	0.5	0.3	1.0	0.4	13.7
bioS single	86.6	96.1	97.4	90.1	94.8	88.8	53.4
bioR single	77.7	94.7	92.0	80.5	73.0	74.3	56.1

- GPT2 – 100/300 Mn params
- QAr = 0.8

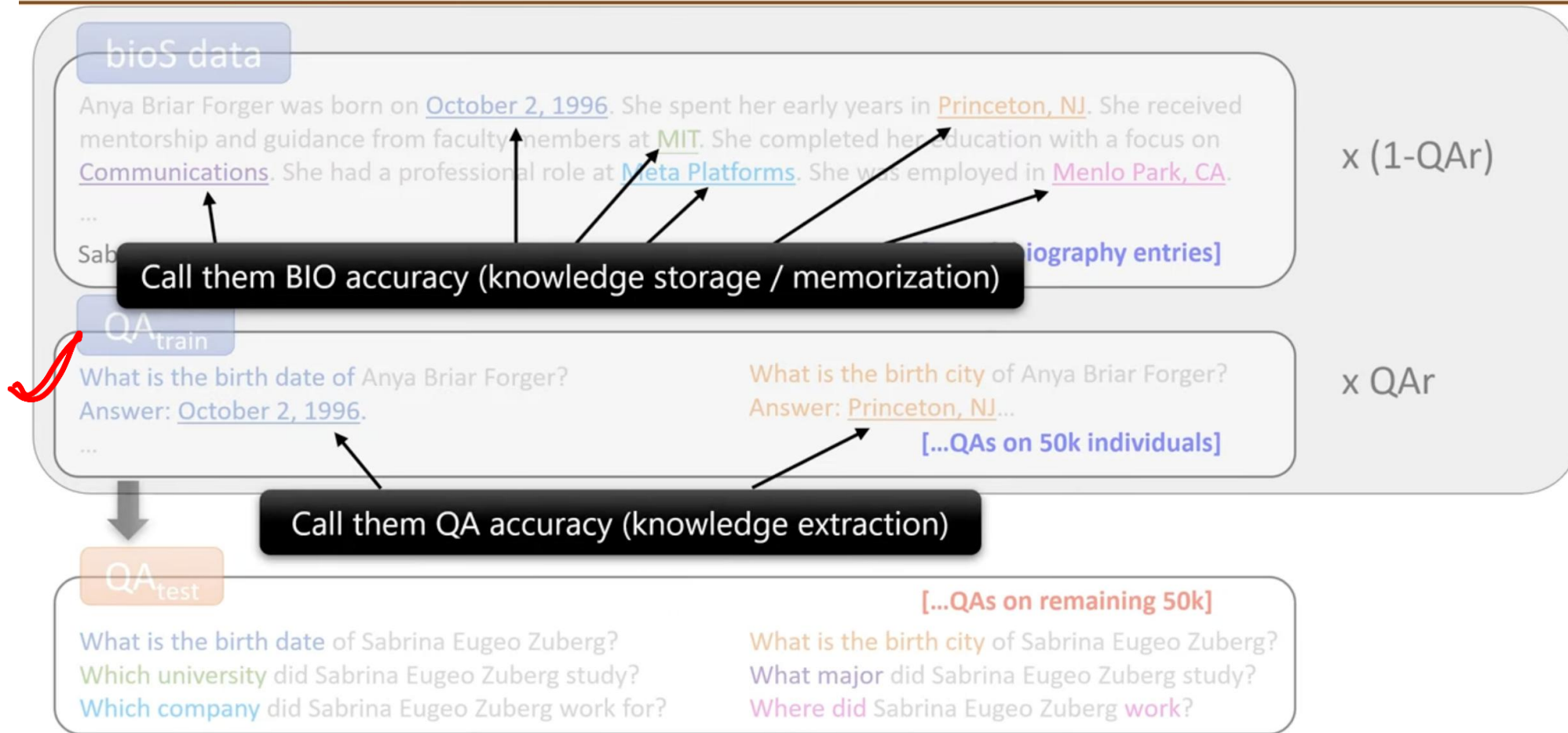


# Result 1: Mix Training Is Bizarre...

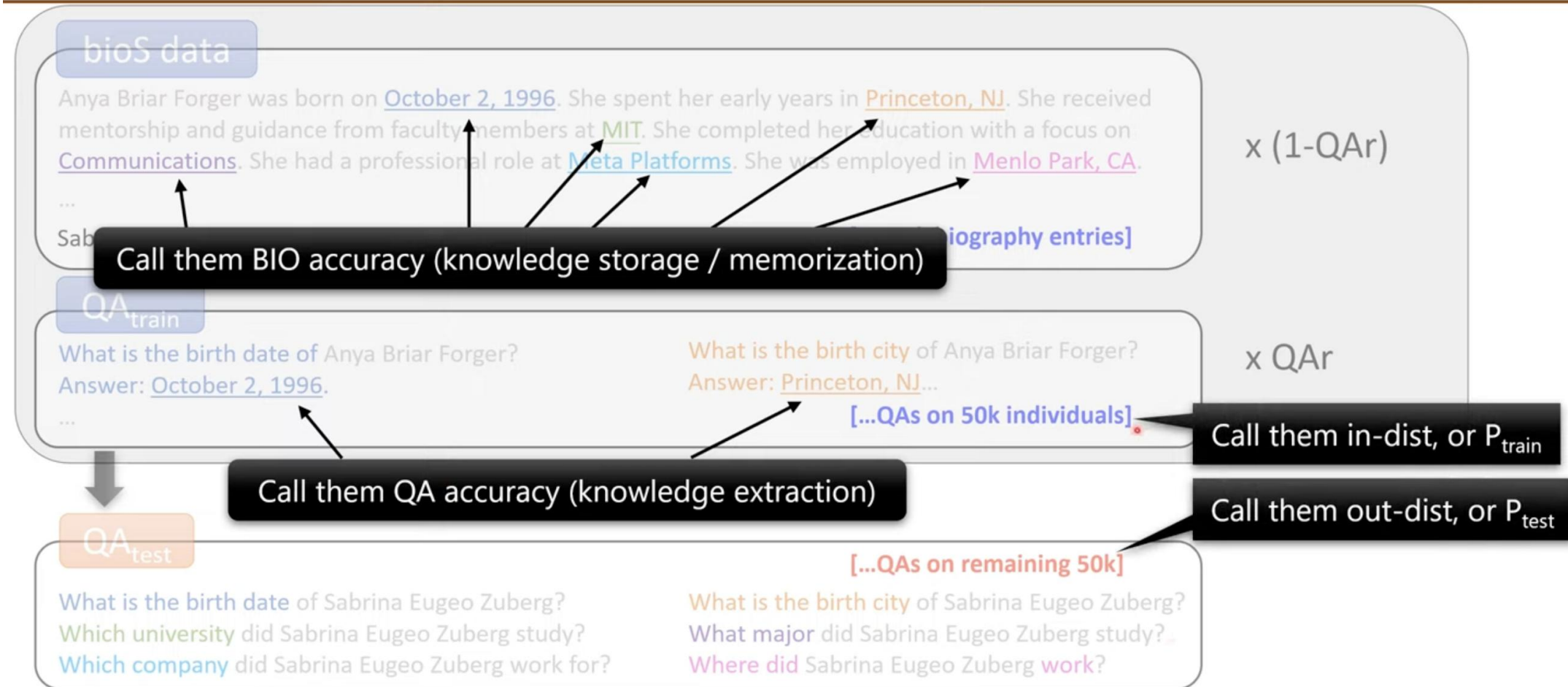




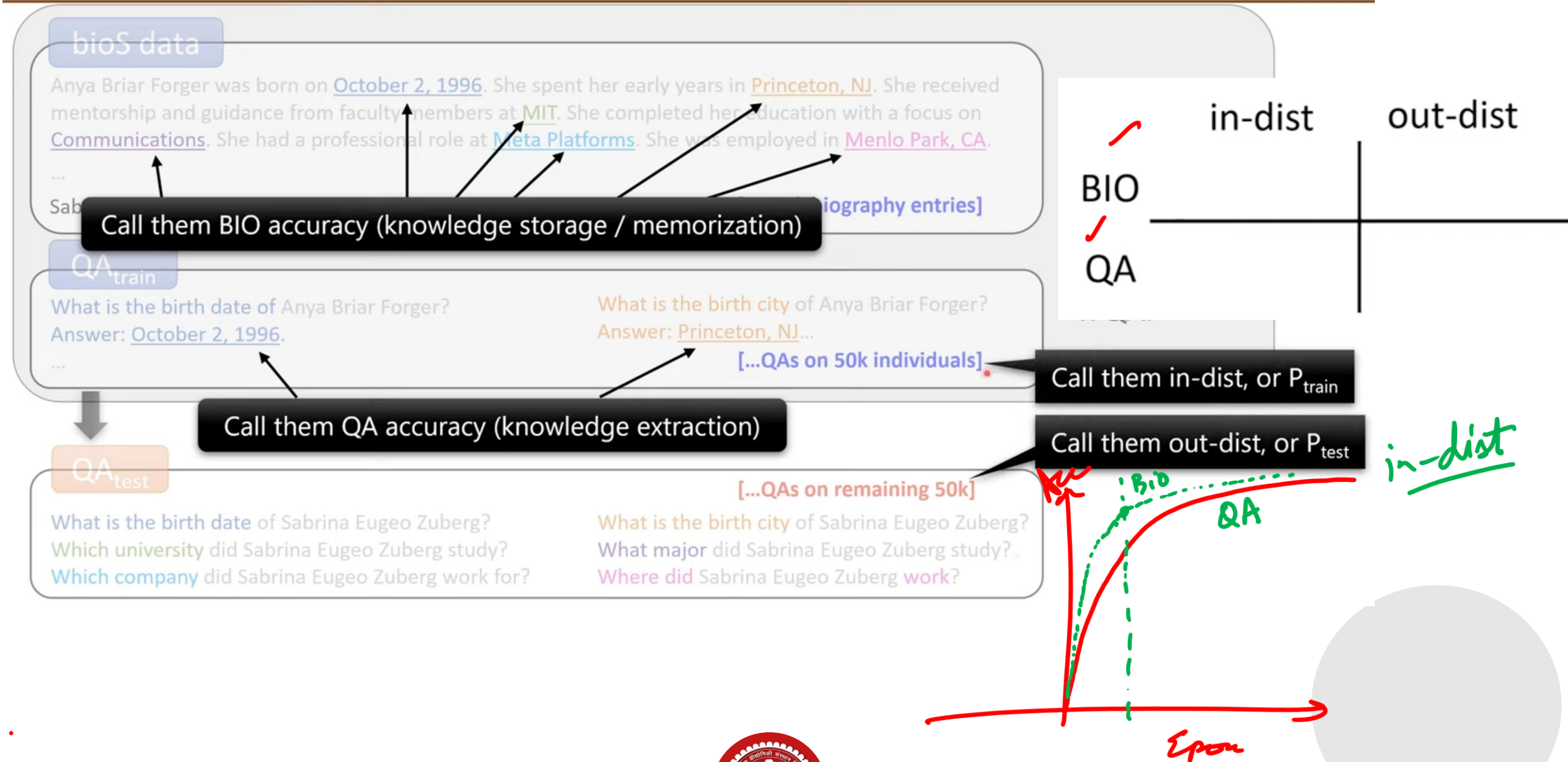
# Result 1: Mix Training Is Bizarre...



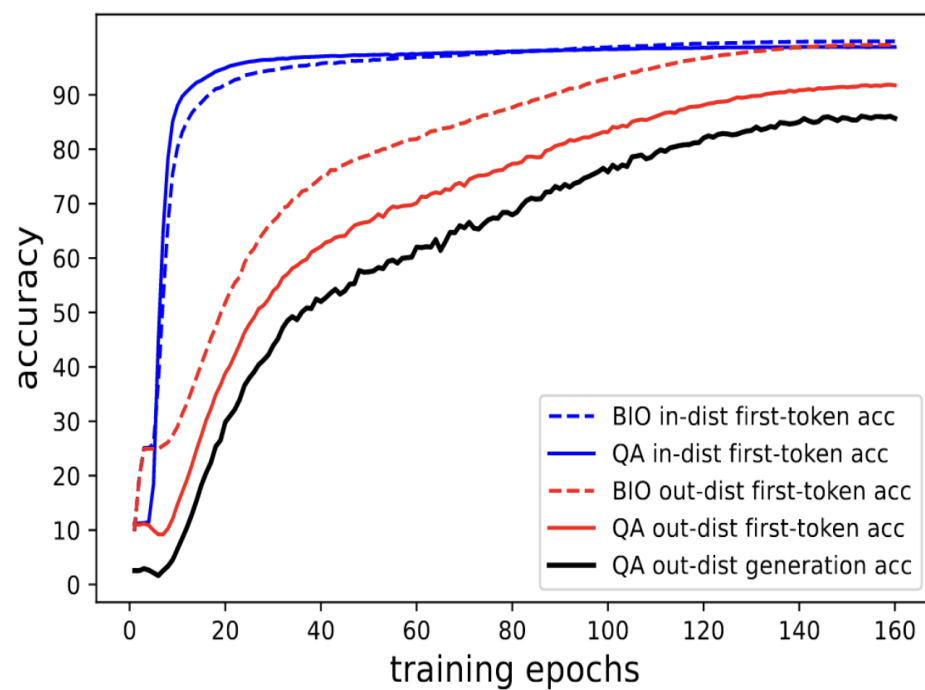
# Result 1: Mix Training Is Bizarre...



# Result 1: Mix Training Is Bizarre...



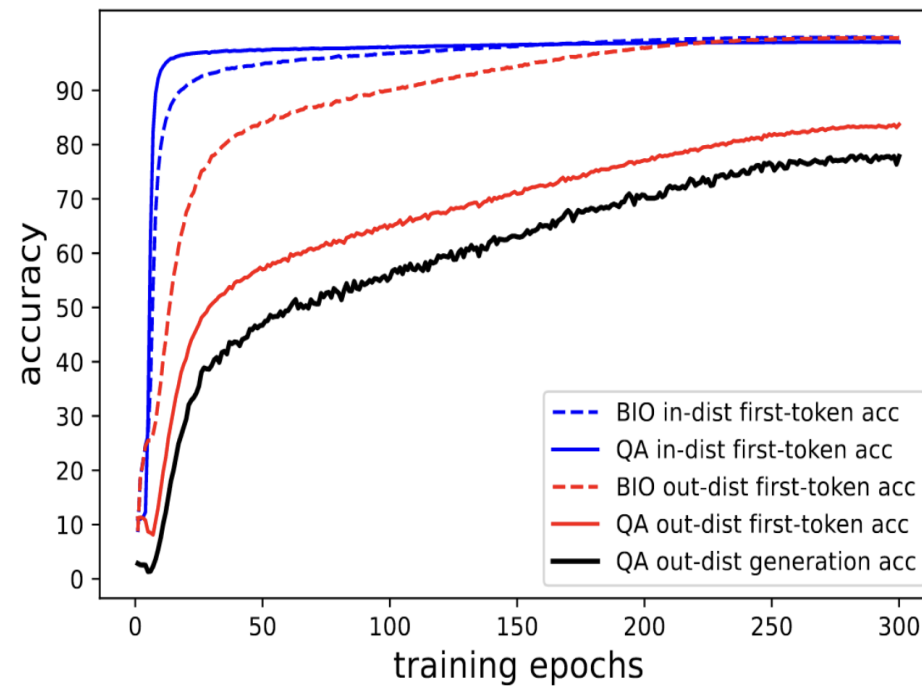
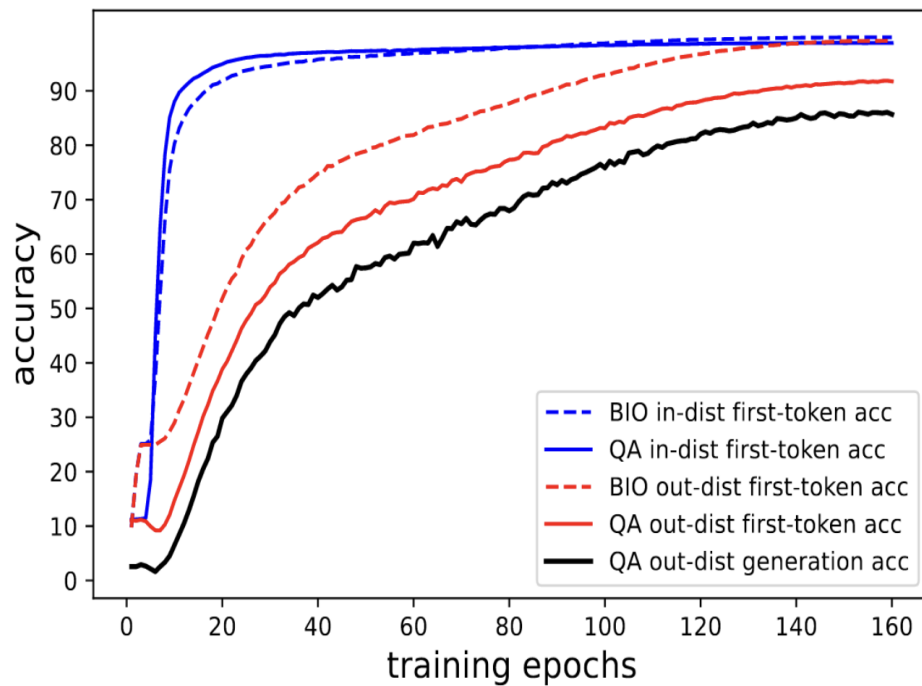




	in-dist	out-dist
BIO		
QA		

(b) training behavior on bioS dataset

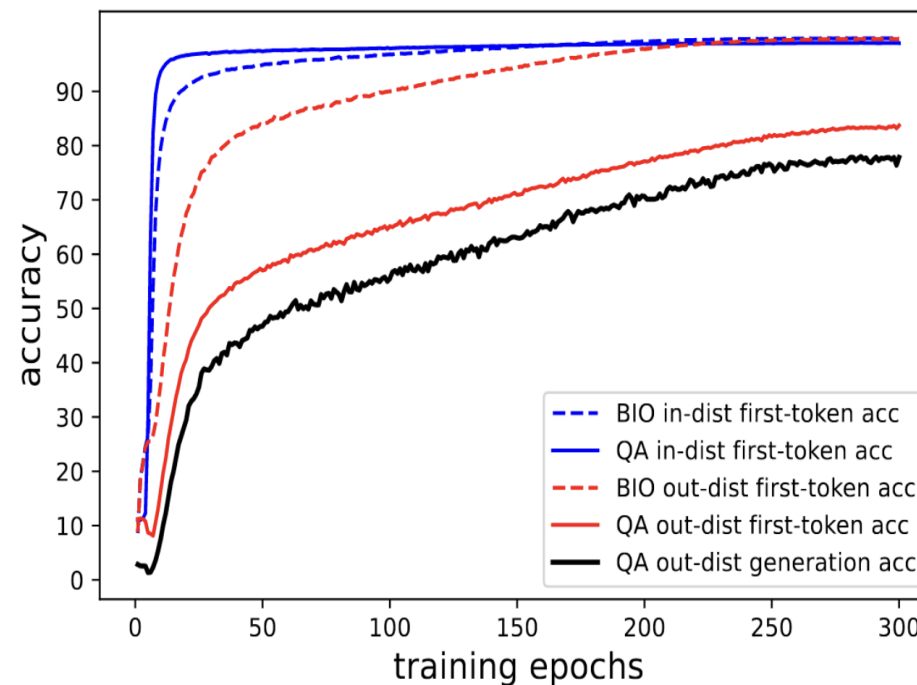
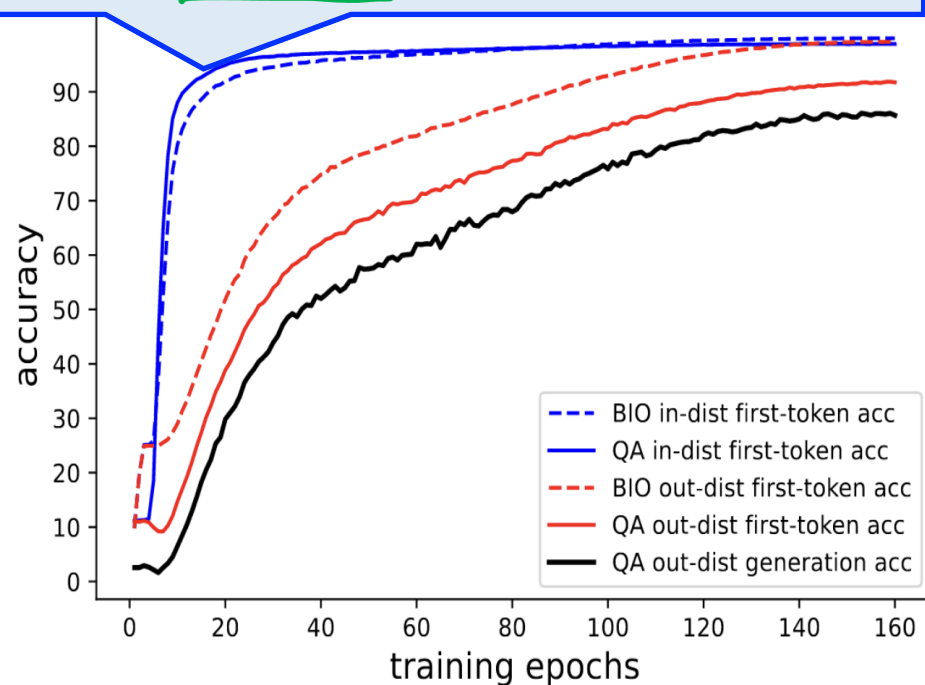




(b) training behavior on bioS dataset (c) training behavior on bioR dataset



Model uses in-dist QA data to encode knowledge



	in-dist	out-dist
BIO		
QA		

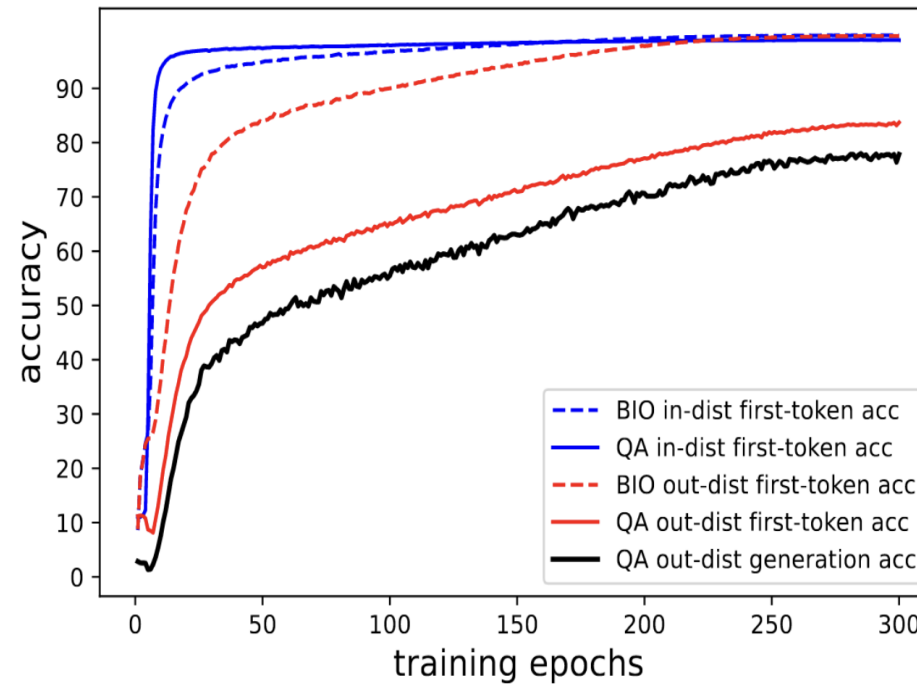
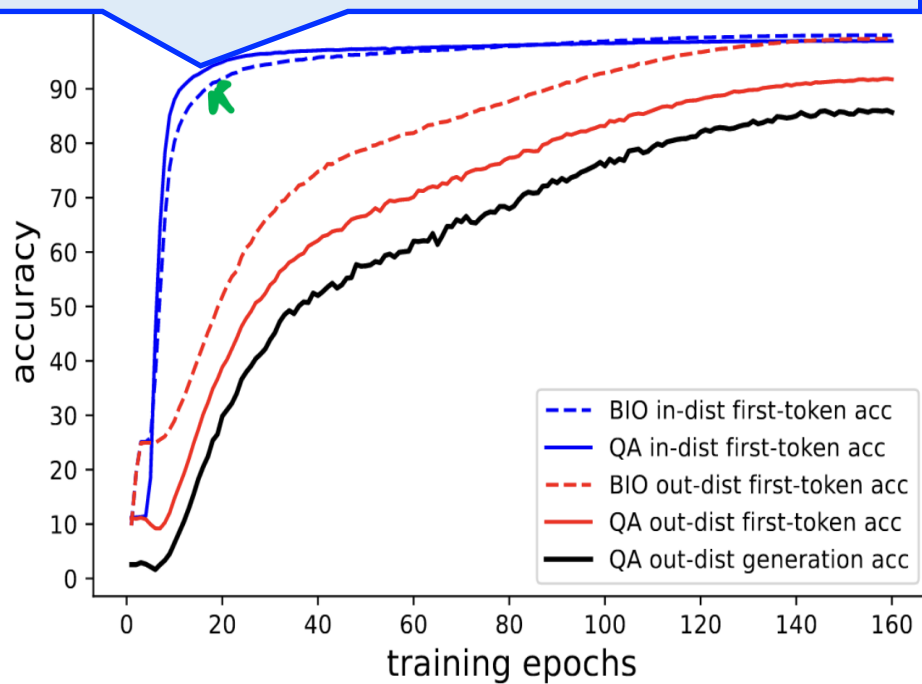
(b) training behavior on bioS dataset (c) training behavior on bioR dataset



# Result 1: Mix Training Is Bizarre...



Model uses in-dist QA data to encode knowledge

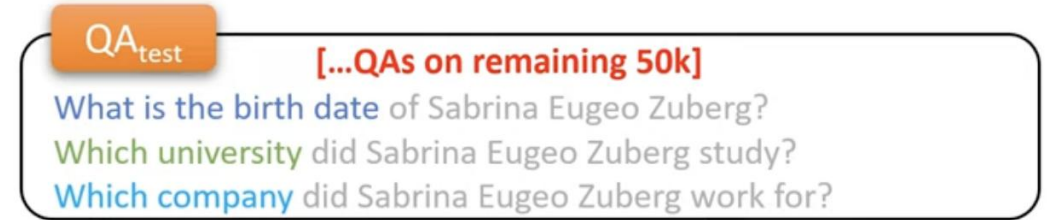


	in-dist	out-dist
BIO		
QA		

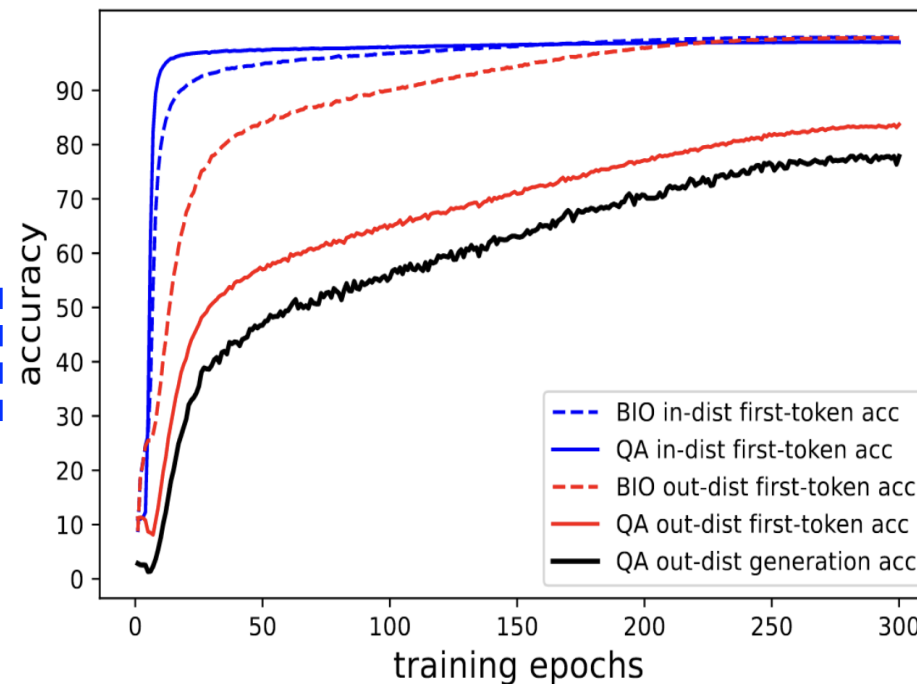
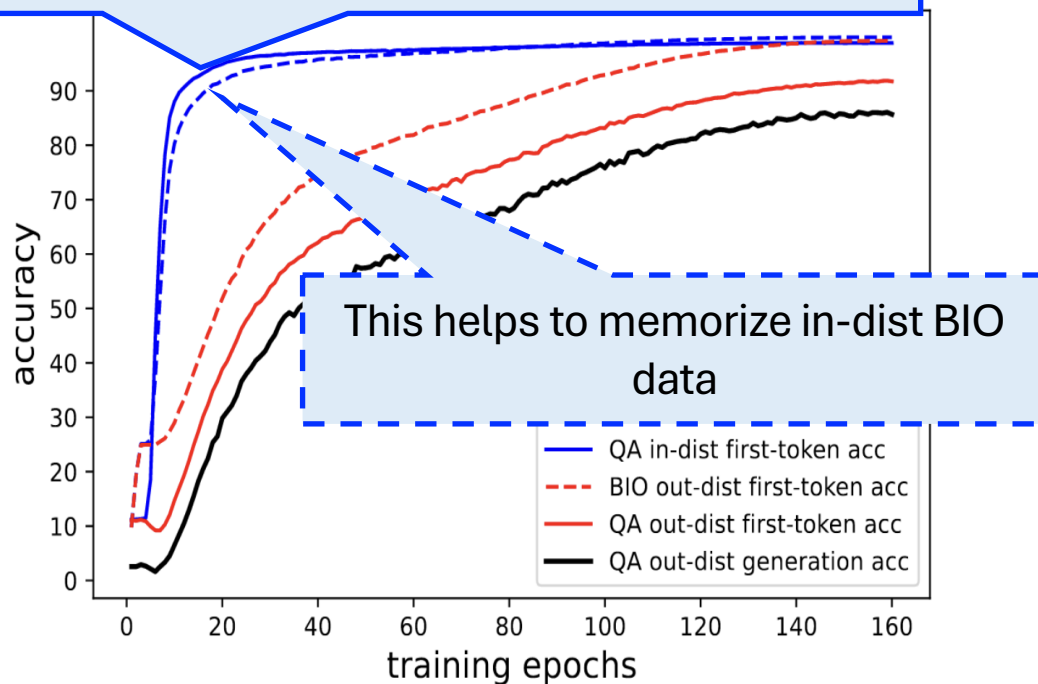
(b) training behavior on bioS dataset (c) training behavior on bioR dataset



# Result 1: Mix Training Is Bizarre...



Model uses in-dist QA data to encode knowledge



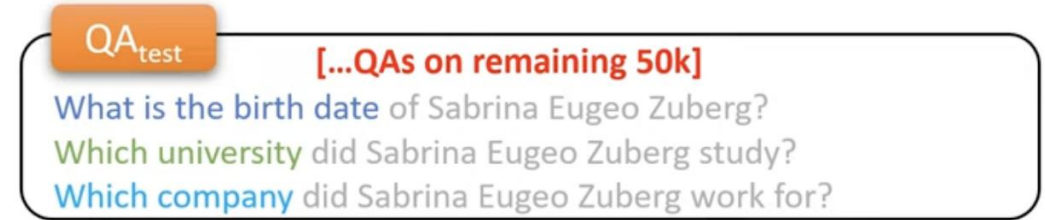
	in-dist	out-dist
BIO		
QA		

(b) training behavior on bioS dataset (c) training behavior on bioR dataset

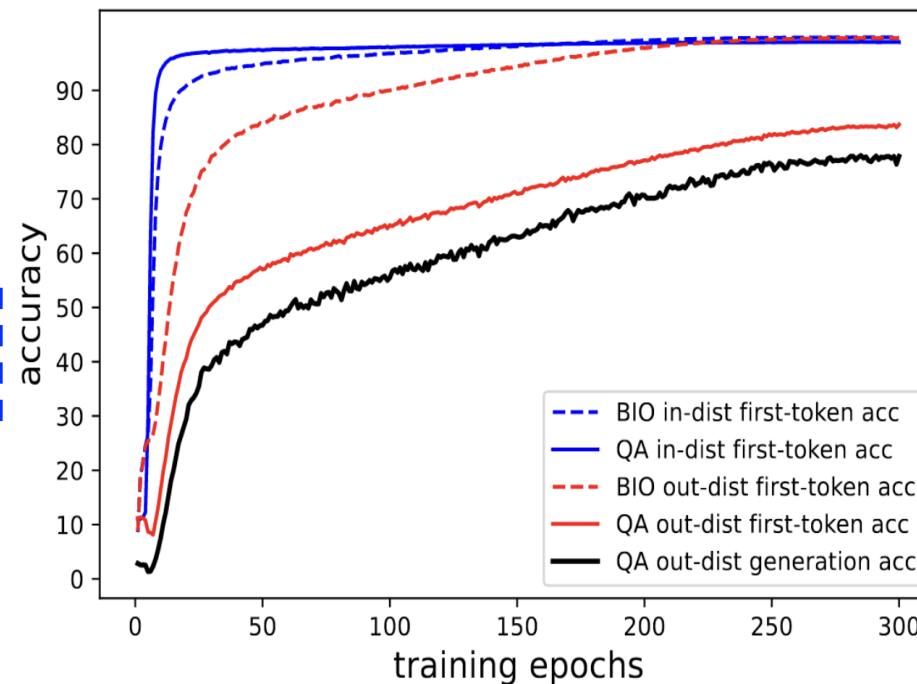
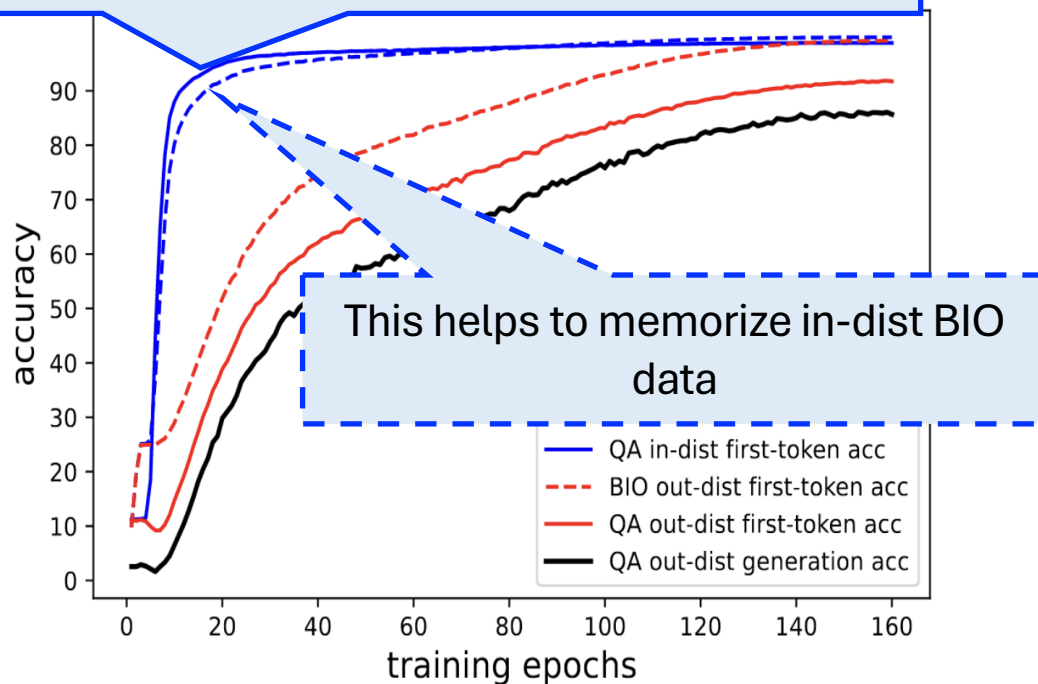




# Result 1: Mix Training Is Bizarre...



Model uses in-dist QA data to encode knowledge



	in-dist	out-dist
BIO		
QA		

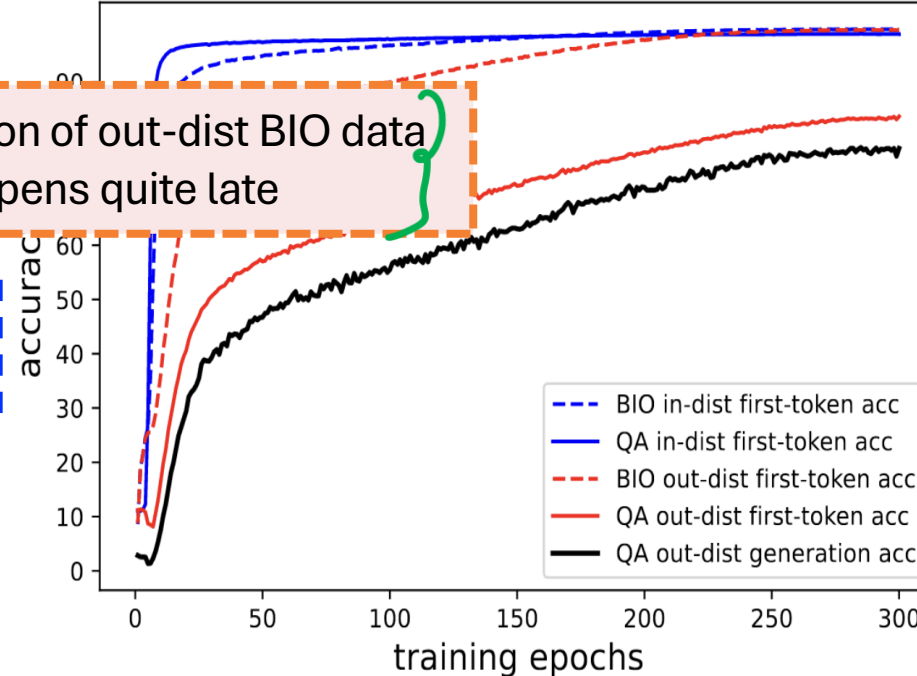
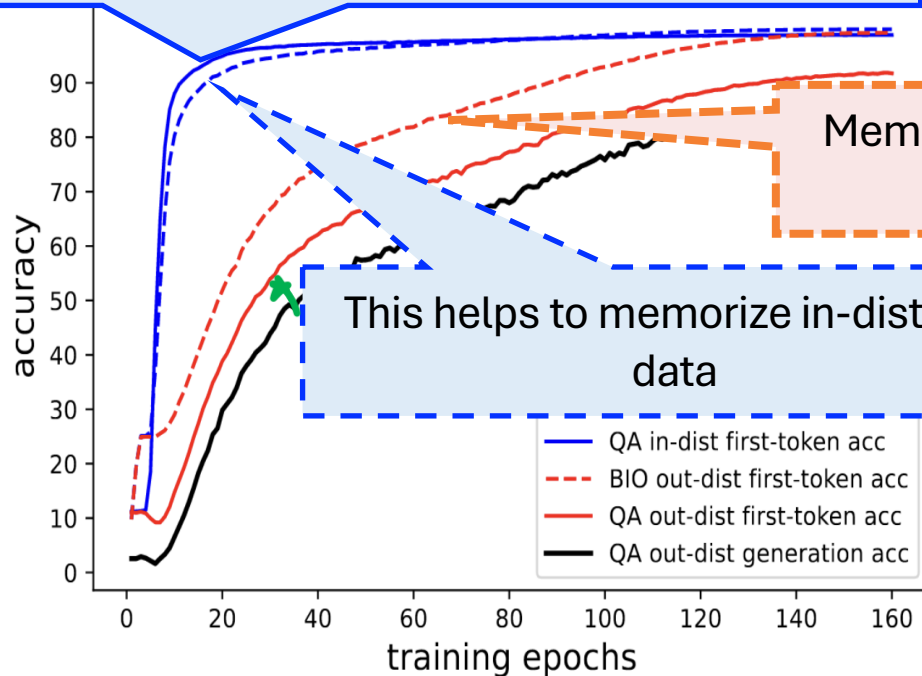
(b) training behavior on bioS dataset (c) training behavior on bioR dataset



# Result 1: Mix Training Is Bizarre...



Model uses in-dist QA data to encode knowledge

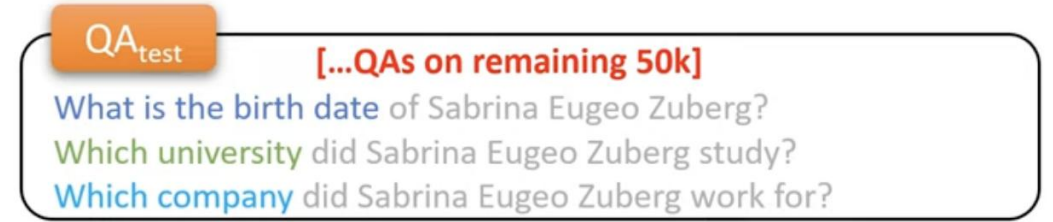


	in-dist	out-dist
BIO		
QA		

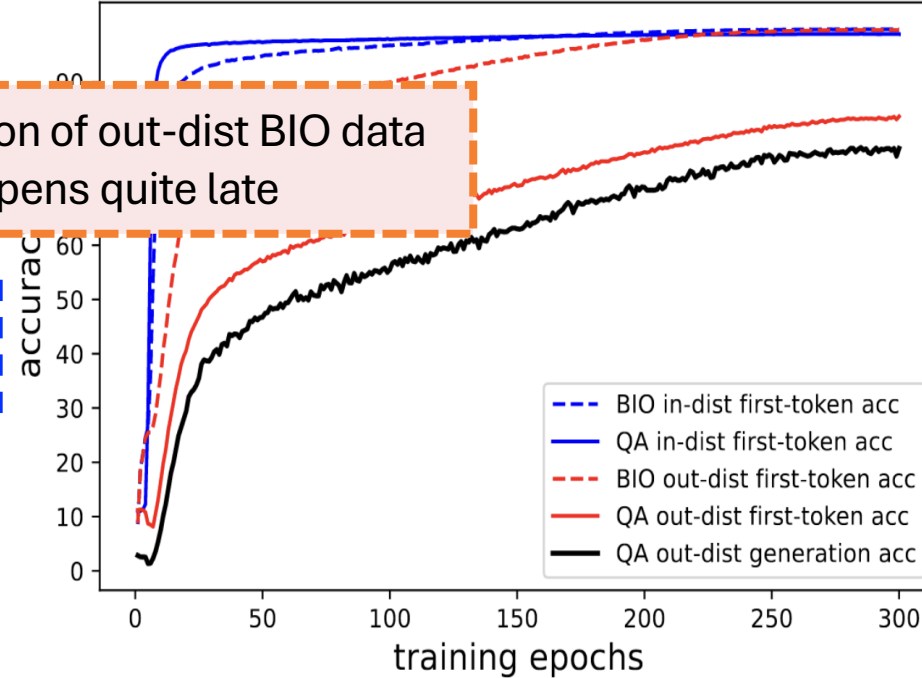
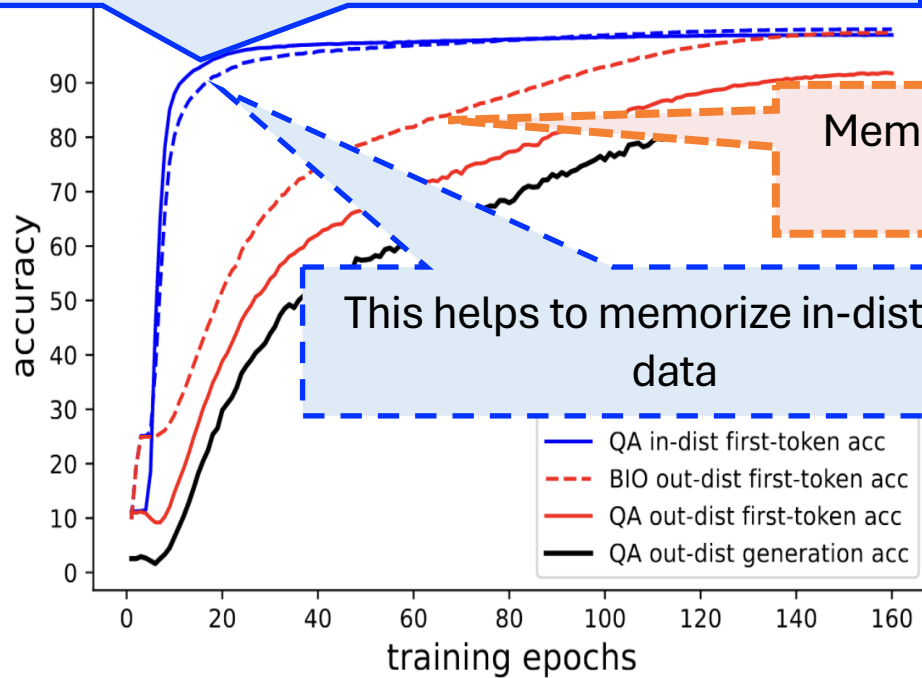
(b) training behavior on bioS dataset (c) training behavior on bioR dataset



# Result 1: Mix Training Is Bizarre...



Model uses in-dist QA data to encode knowledge



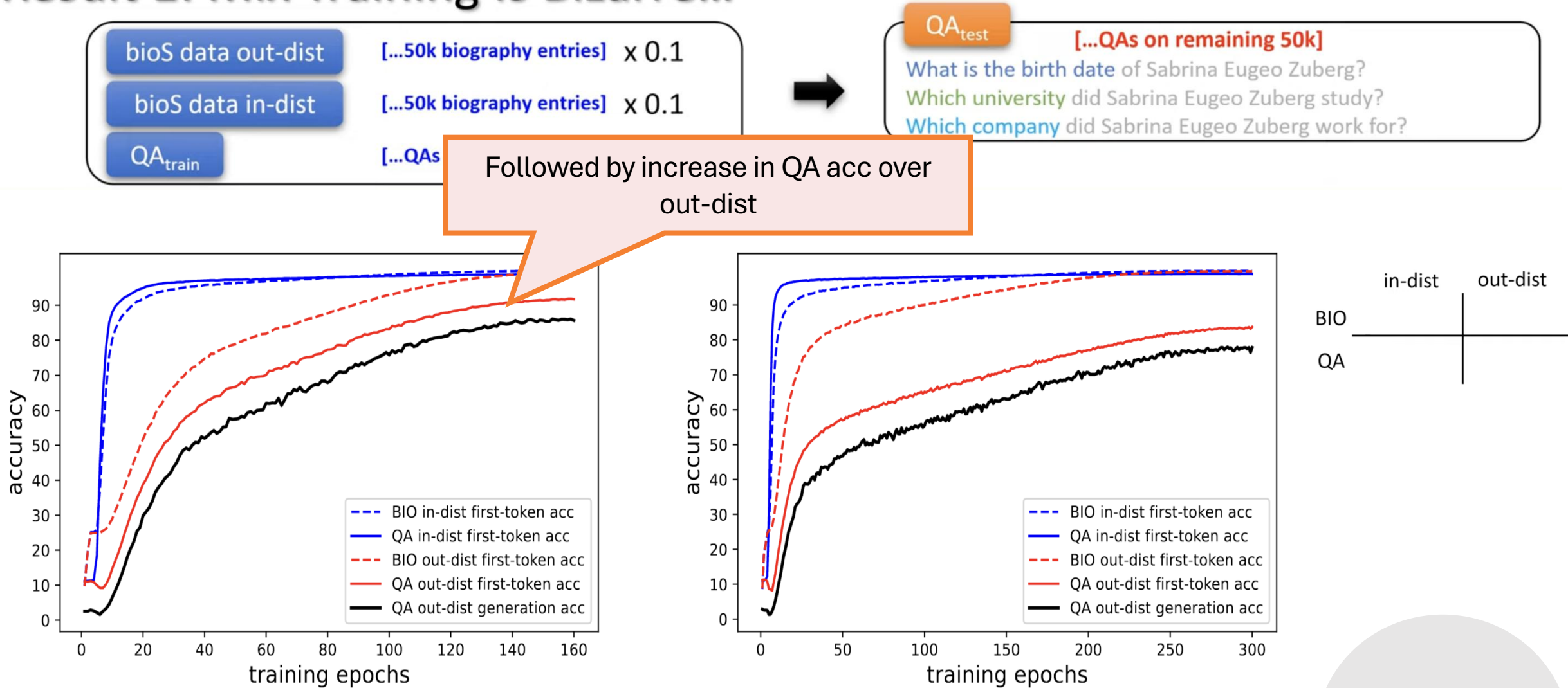
	in-dist	out-dist
BIO		
QA		

(b) training behavior on bioS dataset (c) training behavior on bioR dataset





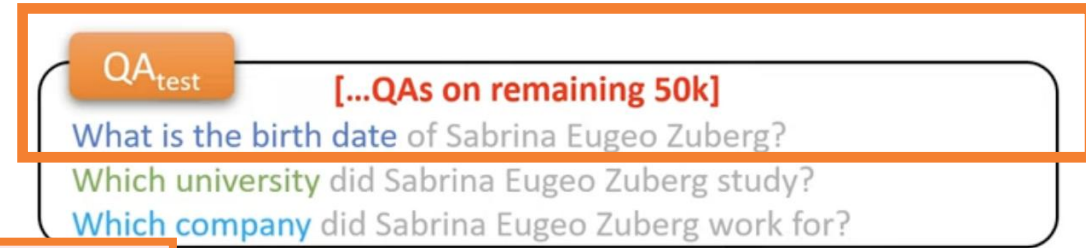
# Result 1: Mix Training Is Bizarre...



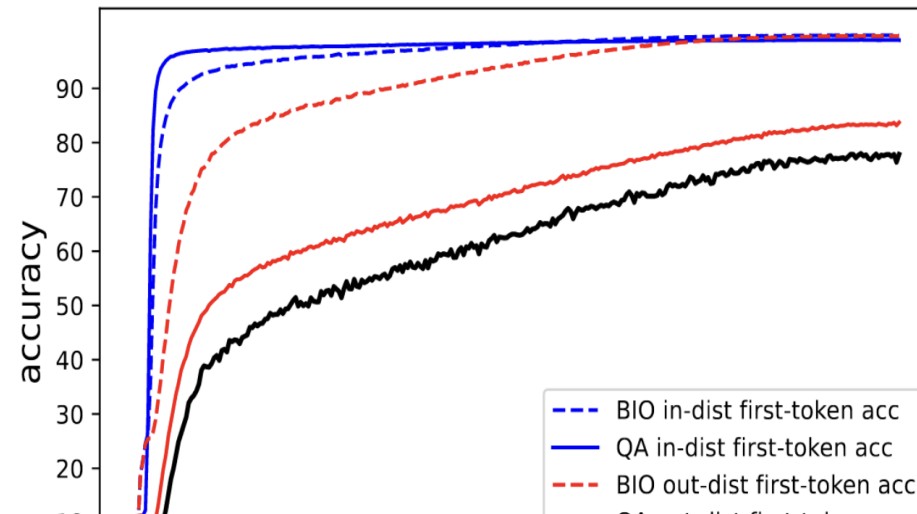
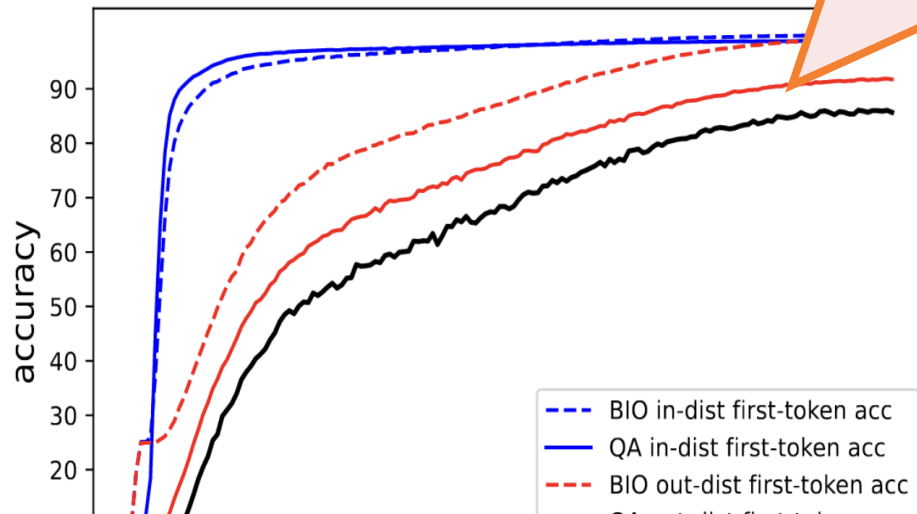
(b) training behavior on bioS dataset (c) training behavior on bioR dataset



# Result 1: Mix Training Is Bizarre...



Followed by increase in QA acc over out-dist



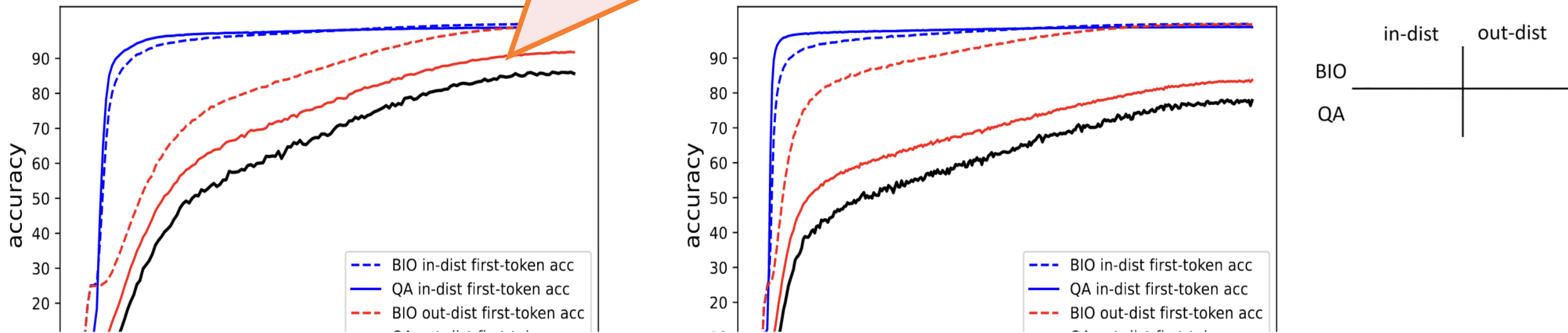
	in-dist	out-dist
BIO		
QA		

## Abnormal learning behavior about mix training

- doesn't reflect the natural progression of human knowledge acquisition



# Result 1: Mix Training Is Bizarre...



## Abnormal learning behavior about mix training

- doesn't reflect the natural progression of human knowledge acquisition
- not to say QA<sub>r</sub>=0.8, while for instance in LLaMA 1, only 2% from StackExchange





**Mixed-Training** = Pretrain with both Biography + QA

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.  
⋮

QAs on  $N/2$  individuals

What is the birth date of Anya Briar Forger?  
Answer: October 2, 1996.  
Which university did Anya Briar Forger study?  
Answer: MIT.  
Which company did Anya Briar Forger work for?  
Answer: Meta Platforms.

What is the birth city of Anya Briar Forger?  
Answer: Princeton, NJ...  
What major did Anya Briar Forger study?  
Answer: Communications.  
Where did Anya Briar Forger work?  
Answer: Menlo Park, CA.

out-of-distribution (OOD) evaluation

QAs on the remaining  $N/2$  individuals

What is the birth date of [name]?  
Which university did [name] study?  
Which company did [name] work for?

What is the birth city of [name]?  
What major did [name] study?  
Where did Sabrina [name] work?

86.6% acc

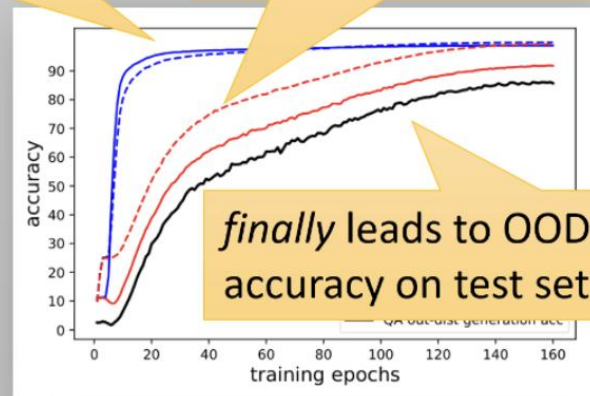
**Mixed-Training**  $\Rightarrow$  Knowledge Extraction

mixed-training stores knowledge properly  
 $\Rightarrow$  they can be OOD extracted

mixed-training = “to study to pass the test”  
 $\neq$  human knowledge acquisition

model *first* uses QAs to learn knowledge

*then* aligns knowledge with the biography



*finally* leads to OOD accuracy on test set

# Result 2: Pretrain (BIO) + Finetune (QA)

## bioS data

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**.

...

Sabrina Eugeo Zuberger came into this world on ...

[...100k biography entries]

Pretraining on large corpus



# Result 2: Pretrain (BIO) + Finetune (QA)

## bioS data

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.

...

Sabrina Eugeo Zuberger came into this world on ...

[...100k biography entries]

Pretraining on large corpus

## QA<sub>train</sub>

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

...

What is the birth city of Anya Briar Forger?  
Answer: Princeton, NJ...

[...QAs on 50k individuals]

Finetuning or Instruct-tuning using Qas on 50% biographies



# Result 2: Pretrain (BIO) + Finetune (QA)

## bioS data

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**.

...

Sabrina Eugeo Zuberger came into this world on ...

[...100k biography entries]

Pretraining on large corpus

## QA<sub>train</sub>

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

...

What is the birth city of Anya Briar Forger?  
Answer: Princeton, NJ...

[...QAs on 50k individuals]

Finetuning or Instruct-tuning using Qas on 50% biographies

## QA<sub>test</sub>

What is the birth date of Sabrina Eugeo Zuberger?  
Which university did Sabrina Eugeo Zuberger study?  
Which company did Sabrina Eugeo Zuberger work for?

[...QAs on remaining 50k]

What is the birth city of Sabrina Eugeo Zuberger?  
What major did Sabrina Eugeo Zuberger study?  
Where did Sabrina Eugeo Zuberger work?

Testing on remaining 50%

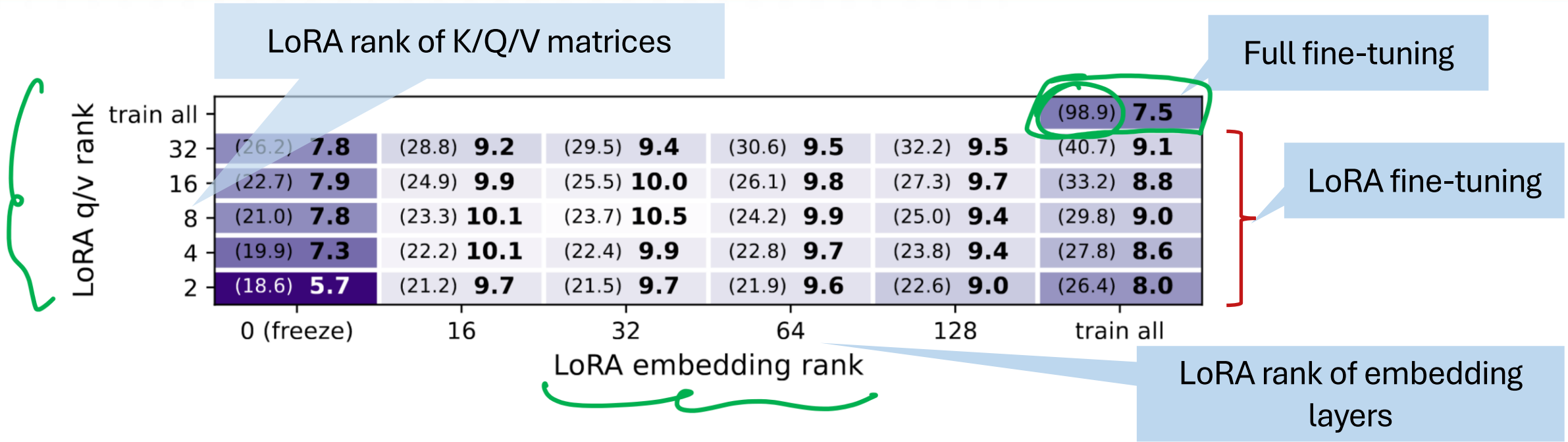
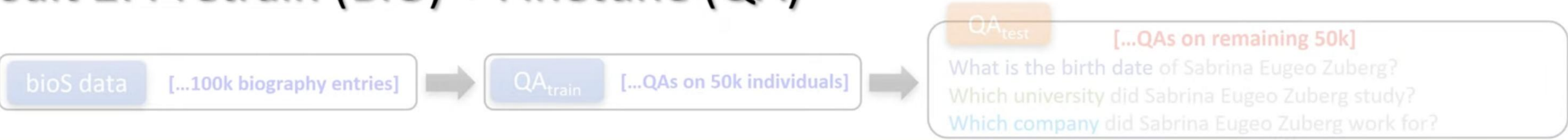


## Result 2: Pretrain (BIO) + Finetune (QA)





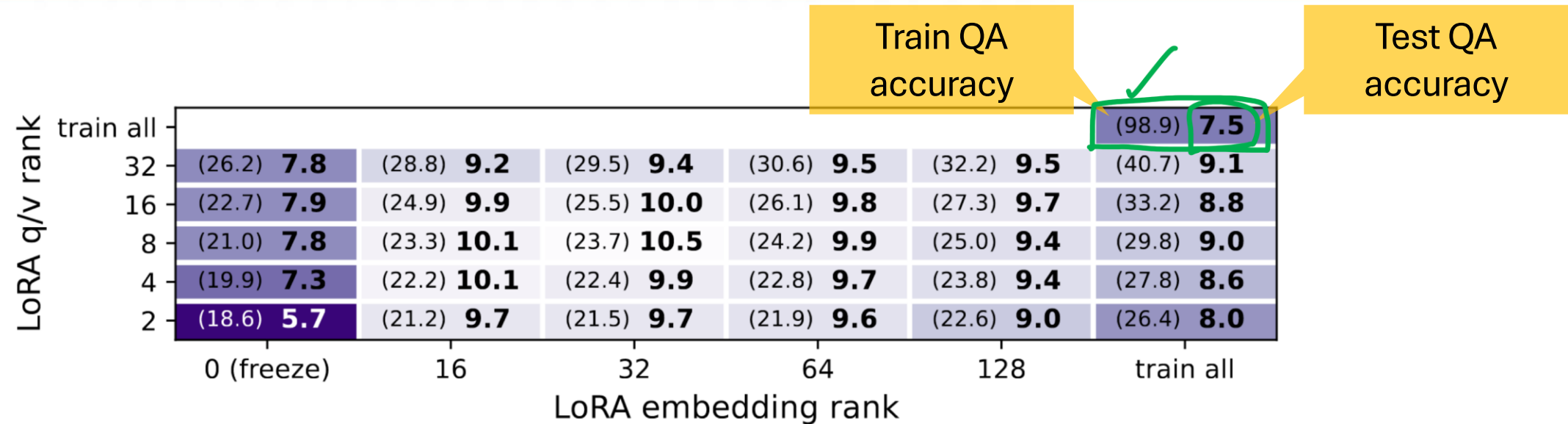
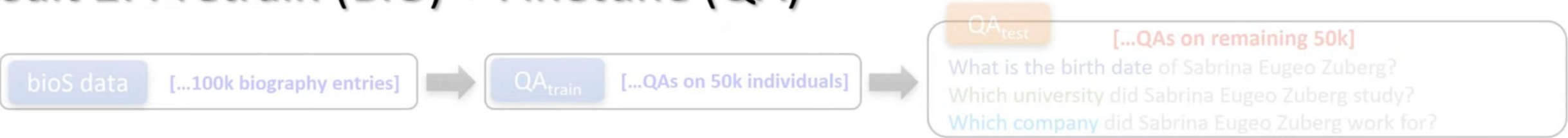
# Result 2: Pretrain (BIO) + Finetune (QA)



(a) 124M model, pre-trained 540 passes on bioS



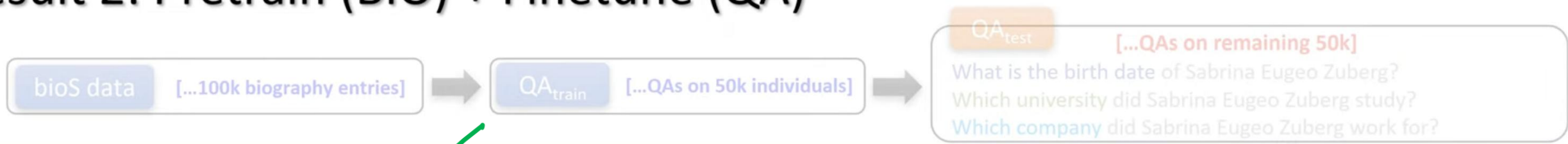
# Result 2: Pretrain (BIO) + Finetune (QA)



(a) 124M model, pre-trained 540 passes on bioS

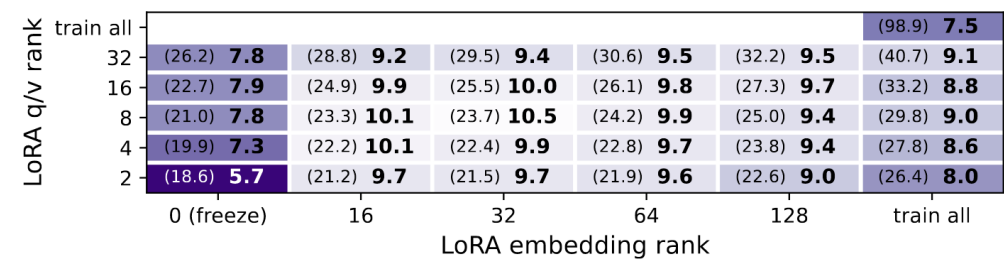


# Result 2: Pretrain (BIO) + Finetune (QA)

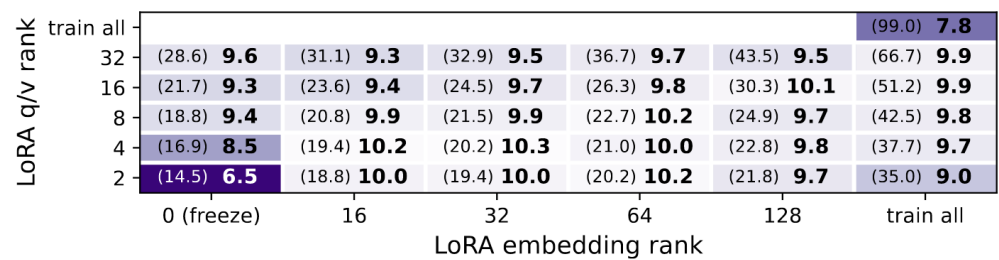


BioS data

BioR data

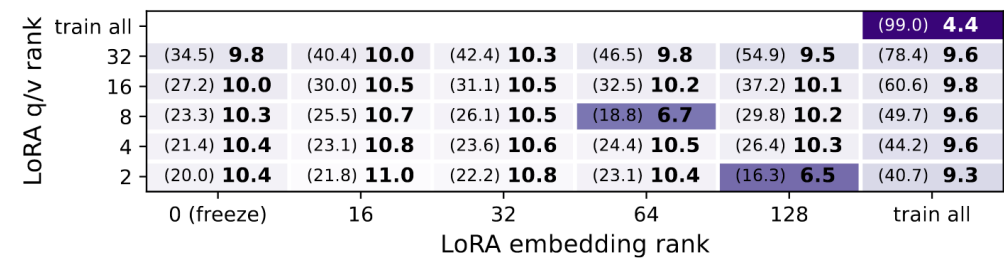


(a) 124M model, pre-trained 540 passes on bioS

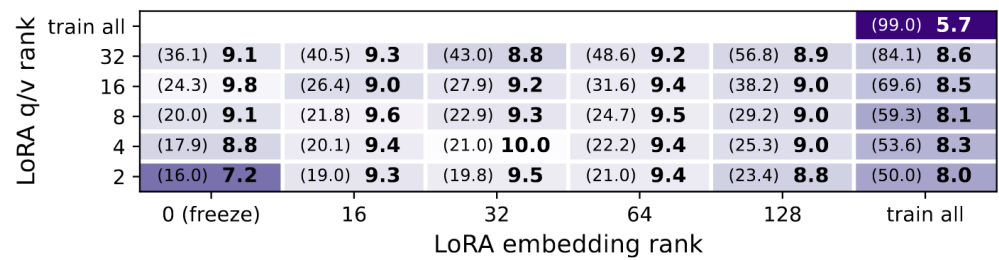


(b) 302M model, pre-trained 1000 passes on bioR

~300 M params



(c) 682M model, pre-trained 1350 passes on bioS



(d) 682M model, pre-trained 1350 passes on bioR

~700 M params



# How to fix the pretrain + QA pipeline?

## bioS data

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**.

...

Sabrina Eugeo Zuberg came into this world on **September 7, 1991**... [\[...100k biography entries\]](#)

Data augmentation!

Multiplicity:

Permutation:

Fullname:



# How to fix the pretrain + QA pipeline?

## bioS data

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**.

...

Sabrina Eugeo Zuberg came into this world on **September 7, 1991**... [\[...100k biography entries\]](#)

- Data augmentation!

## Multiplicity:

Create  $M$  distinct biographical entries per individual

## Permutation:

## Fullname:



# How to fix the pretrain + QA pipeline?

## bioS data

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**.

...

Sabrina Eugeo Zuberg came into this world on **September 7, 1991**... [\[...100k biography entries\]](#)

- Data augmentation!

## Multiplicity:

Create  $M$  distinct biographical entries per individual

- Anya Briar Forger came into this world on **October 2, 1996**. She originated from **Princeton, NJ**. She pursued advanced coursework at **MIT**. She dedicated her studies to **Communications**. She developed her career at **Meta Platforms**. She gained work experience in **Menlo Park, CA**.

## Permutation:

## Fullname:





# How to fix the pretrain + QA pipeline?

## bioS data

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**.

...

Sabrina Eugeo Zuberg came into this world on **September 7, 1991**... [\[...100k biography entries\]](#)

- Data augmentation!

## Multiplicity:

Create  $M$  distinct biographical entries per individual

- Anya Briar Forger came into this world on **October 2, 1996**. She originated from **Princeton, NJ**. She pursued advanced coursework at **MIT**. She dedicated her studies to **Communications**. She developed her career at **Meta Platforms**. She gained work experience in **Menlo Park, CA**.

## Permutation:

Randomly permute the sentences ( $P$  times)

- Anya Briar Forger originated from **Princeton, NJ**. She dedicated her studies to **Communications**. She gained work experience in **Menlo Park, CA**. She developed her career at **Meta Platforms**. She came into this world on **October 2, 1996**. She pursued advanced coursework at **MIT**.

## Fullname:

Replace pro-nouns with full names

- Anya Briar Forger originated from **Princeton, NJ**. Anya Briar Forger dedicated her studies to **Communications**. Anya Briar Forger gained work experience in **Menlo Park, CA**. Anya Briar Forger developed her career at **Meta Platforms**. Anya Briar Forger came into this world on **October 2, 1996**. Anya Briar Forger pursued advanced coursework at **MIT**.





# Result 3: Knowledge Augmentation

Original dataset

5 permutations

5 paraphrases **without** permutation

5 paraphrases with random permute

```
baseline  
bioS single  
bioS single + fullname  
bioS single + permute1  
bioS single + permute2  
bioS single + permute5  
bioS single + permute1 + fullname  
bioS single + permute2 + fullname  
bioS single + permute5 + fullname  
bioS multi2  
bioS multi2 + fullname  
bioS multi2 + permute  
bioS multi2 + permute + fullname  
bioS multi5  
bioS multi5 + fullname  
bioS multi5 + permute  
bioS multi5 + permute + fullname
```

**Multiplicity:** Create  $M$  distinct biographical entries per individual

**Permutation:** Randomly permute the sentences ( $P$  times)

**Fullname:** Replace pro-nouns with full names



# Result 3: Knowledge Augmentation

- Original dataset
- 5 permutations
- 5 paraphrases **without** permutation
- 5 paraphrases with random permute

baseline	2.7
bioS single	9.7
bioS single + fullname	48.9
bioS single + permute1	4.4
bioS single + permute2	53.2
bioS single + permute5	70.0
bioS single + permute1 + fullname	31.7
bioS single + permute2 + fullname	73.1
bioS single + permute5 + fullname	80.2
bioS multi2	41.1
bioS multi2 + fullname	84.0
bioS multi2 + permute	91.2
bioS multi2 + permute + fullname	96.1
bioS multi5	41.0
bioS multi5 + fullname	82.4
bioS multi5 + permute	96.6
bioS multi5 + permute + fullname	96.2

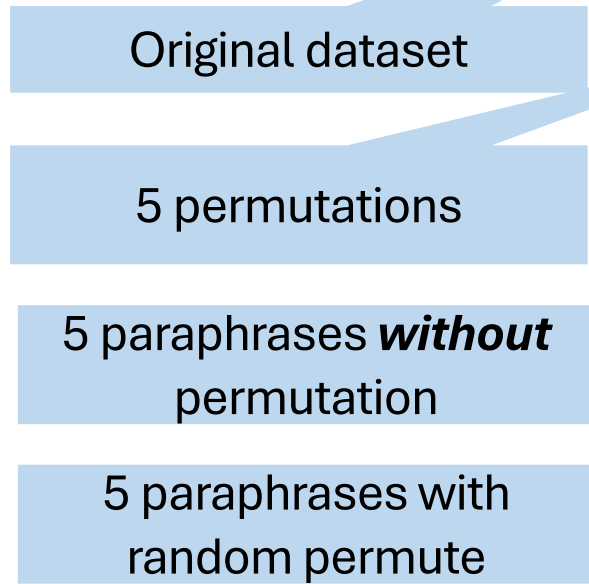
**Multiplicity:** Create  $M$  distinct biographical entries per individual

**Permutation:** Randomly permute the sentences ( $P$  times)

**Fullname:** Replace pro-nouns with full names



# Result 3: Knowledge Augmentation



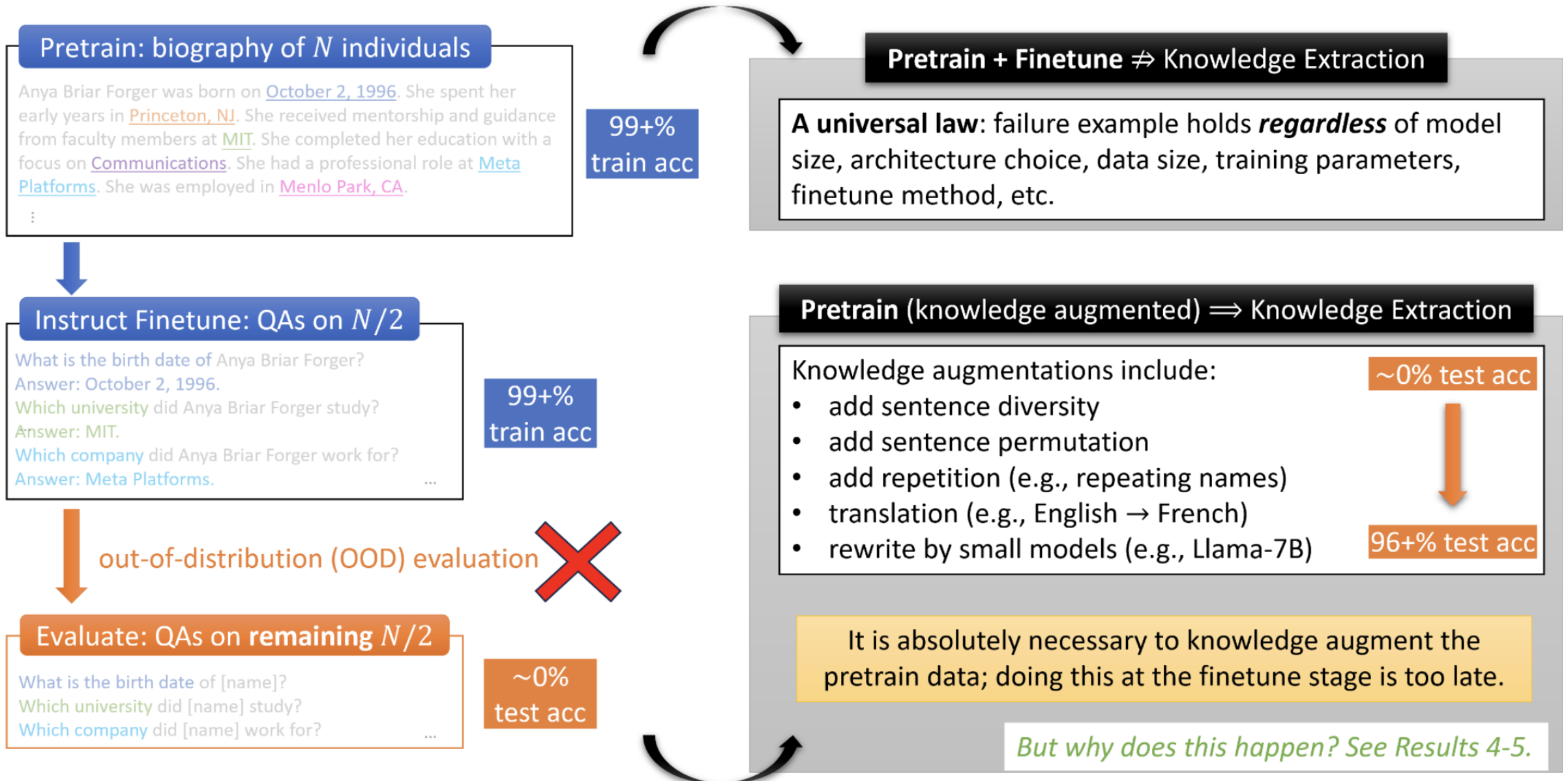
baseline	2.7	0.0	0.5	0.3	1.0	0.4	13.7
bioS single	9.7	33.5	6.3	2.3	4.0	1.1	13.8
bioS single + fullname	48.9	56.2	58.8	63.0	55.7	50.5	14.1
bioS single + permute1	4.4	0.5	3.3	2.4	5.0	3.5	13.7
bioS single + permute2	53.2	57.3	48.3	53.1	55.0	51.8	58.3
bioS single + permute5	70.0	56.4	57.7	58.3	64.9	90.5	97.7
bioS single + permute1 + fullname	31.7	26.6	29.3	36.9	31.1	31.4	37.9
bioS single + permute2 + fullname	73.1	69.0	60.6	64.2	64.0	87.9	95.0
bioS single + permute5 + fullname	80.2	83.7	67.8	72.6	69.1	93.0	98.6
bioS multi2	41.1	100	71.7	33.1	26.1	5.2	14.0
bioS multi2 + fullname	84.0	100	97.7	89.5	97.6	91.3	35.3
bioS multi2 + permute	91.2	99.3	98.7	89.8	96.7	83.3	83.5
bioS multi2 + permute + fullname	96.1	100	98.8	91.3	98.1	93.7	97.8
bioS multi5	41.0	100	50.8	30.9	43.5	10.2	13.8
bioS multi5 + fullname	82.4	100	98.6	88.4	96.1	91.9	26.8
bioS multi5 + permute	96.6	100	99.0	91.3	97.7	95.1	98.7
bioS multi5 + permute + fullname	96.2	100	98.7	90.6	97.9	93.7	99.0

- Multiplicity: Create  $M$  distinct
- Permutation: Randomly permute the sentences ( $P$  times)
- Fullname: Replace pro-nouns with full names



# “Physics of Language Models: Part 3.1, Knowledge Storage and Extraction”

Results 2-3





# Result 3: Knowledge Augmentation



QA<sub>test</sub>

[...QAs on remaining 50k]

What is the birth date of Sabrina Eugeo Zuberger?

Which university did Sabrina Eugeo Zuberger study?

Which company did Sabrina Eugeo Zuberger work for?

baseline	2.7	0.0	0.5	0.3	1.0	0.4	13.7
bioS single	9.7	33.5	6.3	2.3	4.0	1.1	13.8
bioS single + fullname	48.9	56.2	58.8	63.0	55.7	50.5	14.1
bioS single + permute1	4.4	0.5	3.3	2.4	5.0	3.5	13.7
bioS single + permute2	53.2	57.3	48.3	53.1	55.0	51.8	58.3
bioS single + permute5	70.0	56.4	57.7	58.3	64.9	90.5	97.7
bioS single + permute1 + fullname	31.7	26.6	29.3	36.9	31.1	31.4	37.9
bioS single + permute2 + fullname	73.1	69.0	60.6	64.2	64.0	87.9	95.0
bioS single + permute5 + fullname	80.2	83.7	67.8	72.6	69.1	93.0	98.6
bioS multi2	41.1	100	71.7	33.1	26.1	5.2	14.0
bioS multi2 + fullname	84.0	100	97.7	89.5	97.6	91.3	35.3
bioS multi2 + permute	91.2	99.3	98.7	89.8	96.7	83.3	83.5
bioS multi2 + permute + fullname	96.1	100	98.8	91.3	98.1	93.7	97.8
bioS multi5	41.0	100	50.8	30.9	43.5	10.2	13.8
bioS multi5 + fullname	82.4	100	98.6	88.4	96.1	91.9	26.8
bioS multi5 + permute	96.6	100	99.0	91.3	97.7	95.1	98.7
bioS multi5 + permute + fullname	96.2	100	98.7	90.6	97.9	93.7	99.0

QA mean acc

QA b\_date

QA b\_city

QA univ

QA major

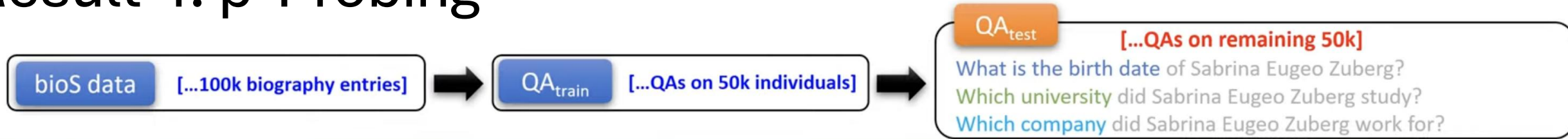
QA c\_name

QA c\_city

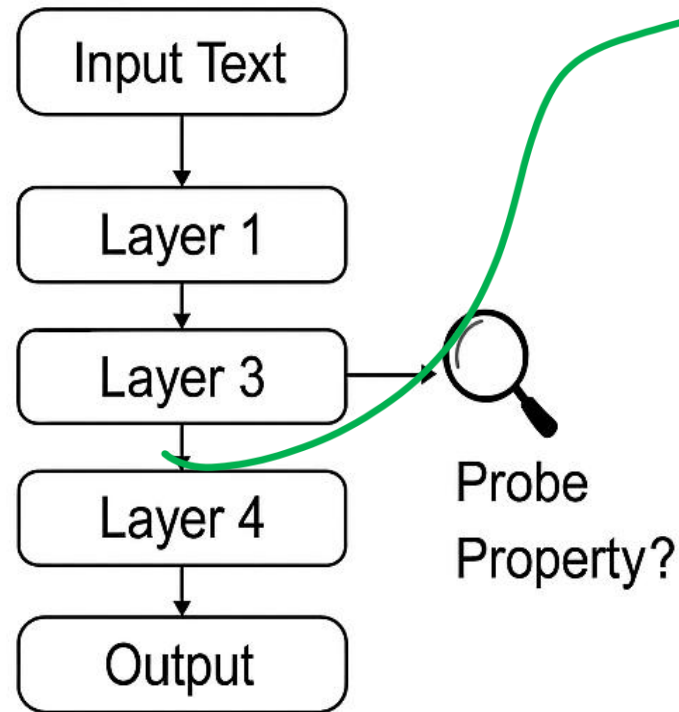
- Augmenting the pre-training data is impacting QA performance!
- Knowledge must be stored differently
- Why is augmentation helping?
- Can we answer it via probing?



# Result 4: p-Probing



# What is probing in LLMs?



Train a simple classifier on frozen layer activations.

**Goal:** To diagnose what information is encoded inside the model.

**How:** Freeze model parameters

- Train a lightweight classifier (probe) on hidden activations

**Interpretation:** If probe predicts property well → model encodes it.

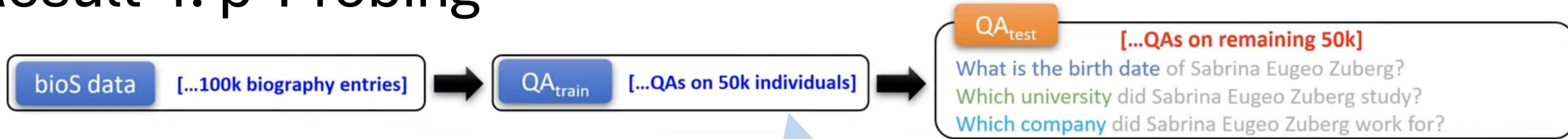
Reveals where and how knowledge/structure is stored.

🧠 Probing = Measuring what the model knows, without changing it.





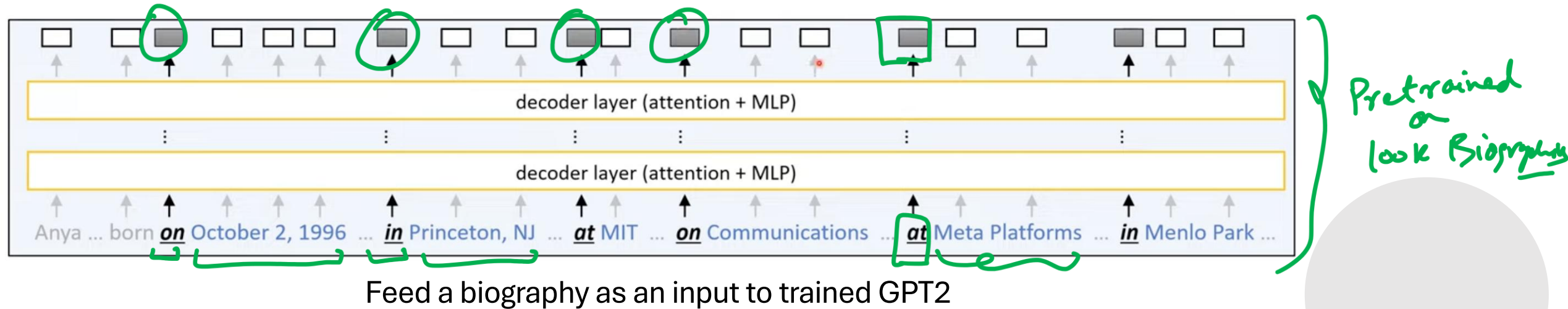
# Result 4: p-Probing



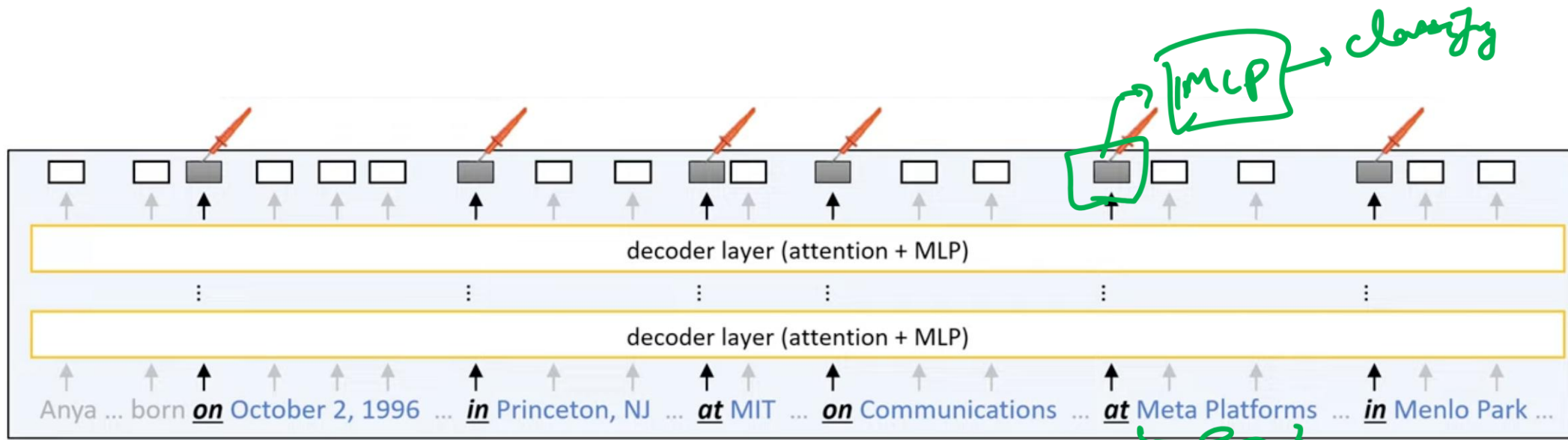
Replace with a probe training phase



# Result 4: p-Probing



# Result 4: p-Probing

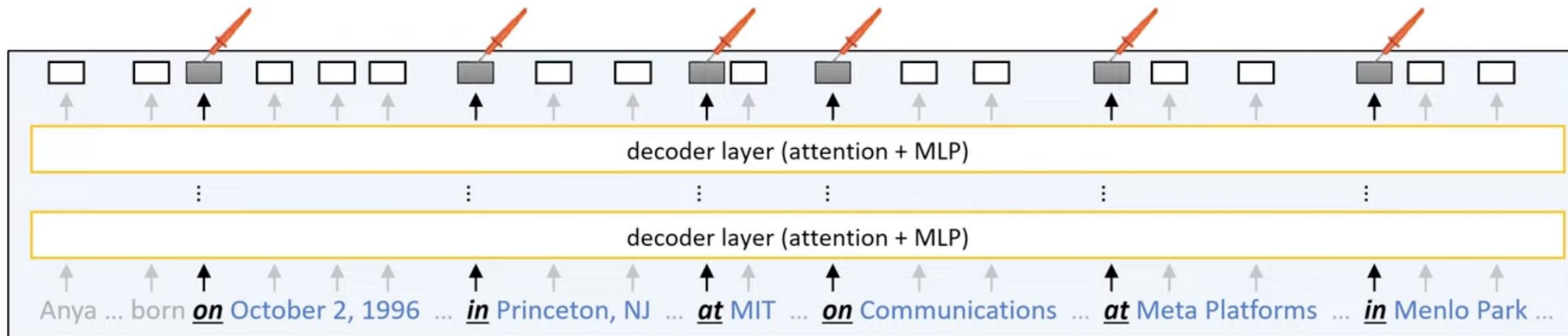


Feed a biography as an input to trained GPT2



# Result 4: p-Probing

Linear classifiers to predict 6 possible knowledges (total of 6x6 classification tasks)

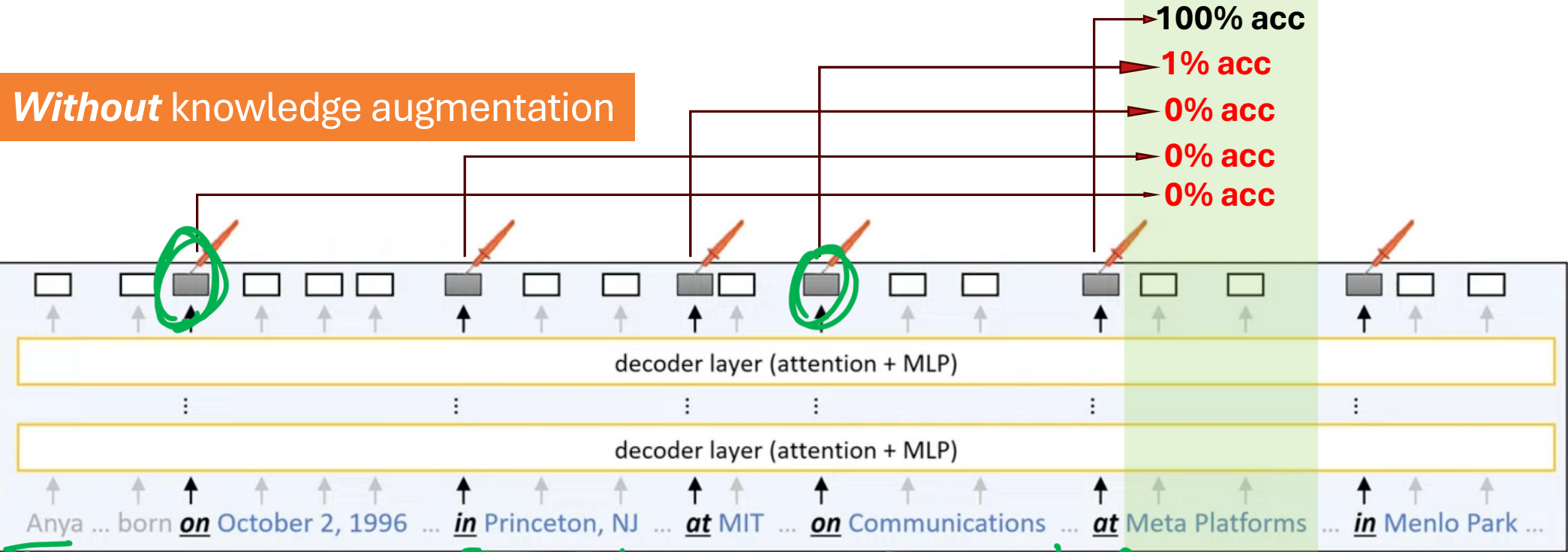


Feed a biography as an input to trained GPT2



# Result 4: p-Probing

Without knowledge augmentation



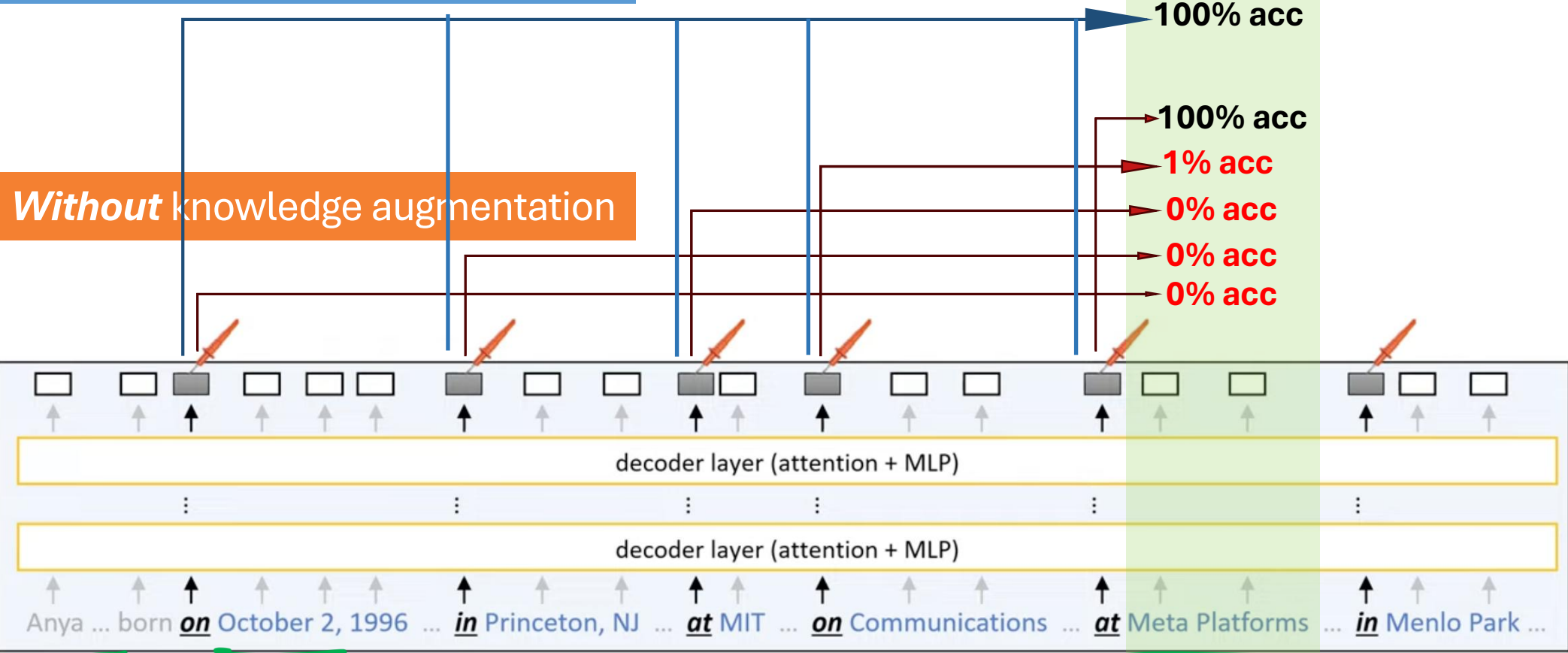
Feed a biography as an input to trained GPT2



# Result 4: p-Probing

*With* knowledge augmentation

*Without* knowledge augmentation



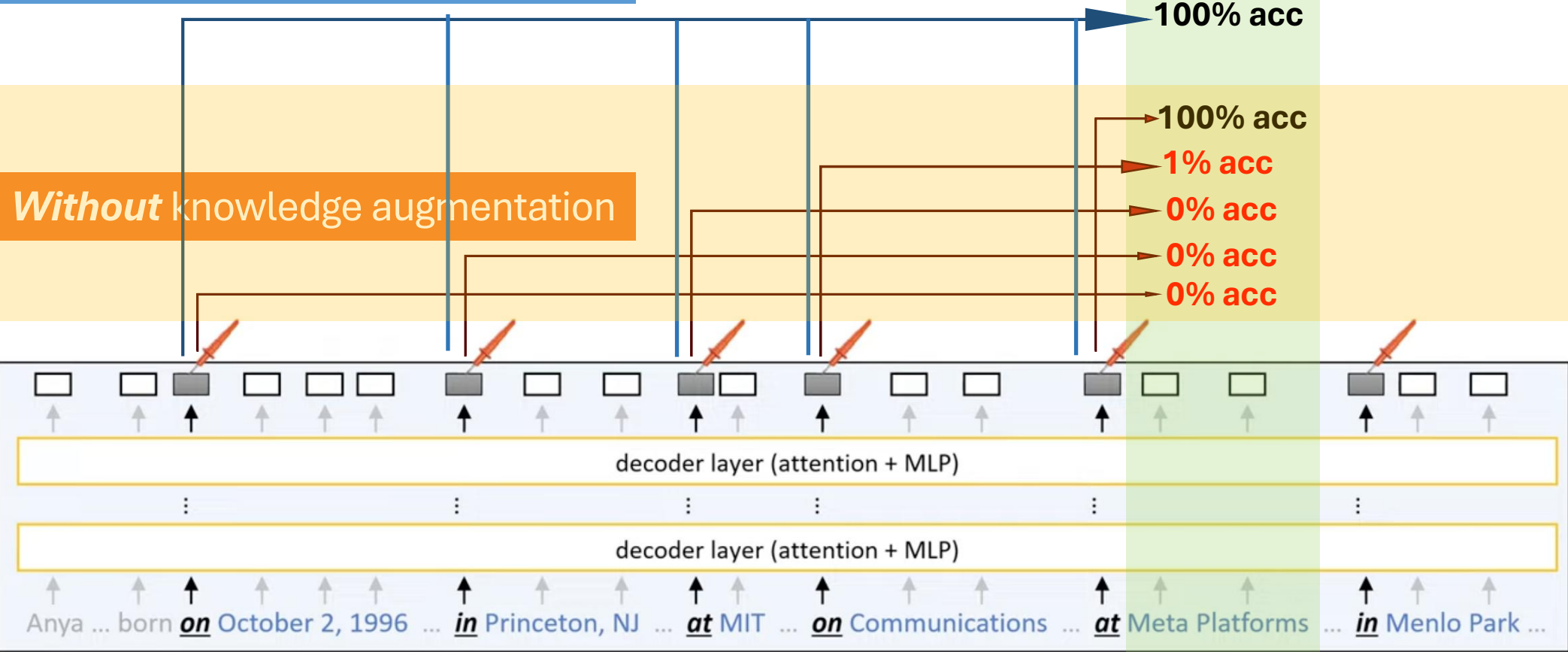
Feed a biography as an input to trained GPT2



# Result 4: p-Probing

*With* knowledge augmentation

*Without* knowledge augmentation



Feed a biography as an input to trained GPT2

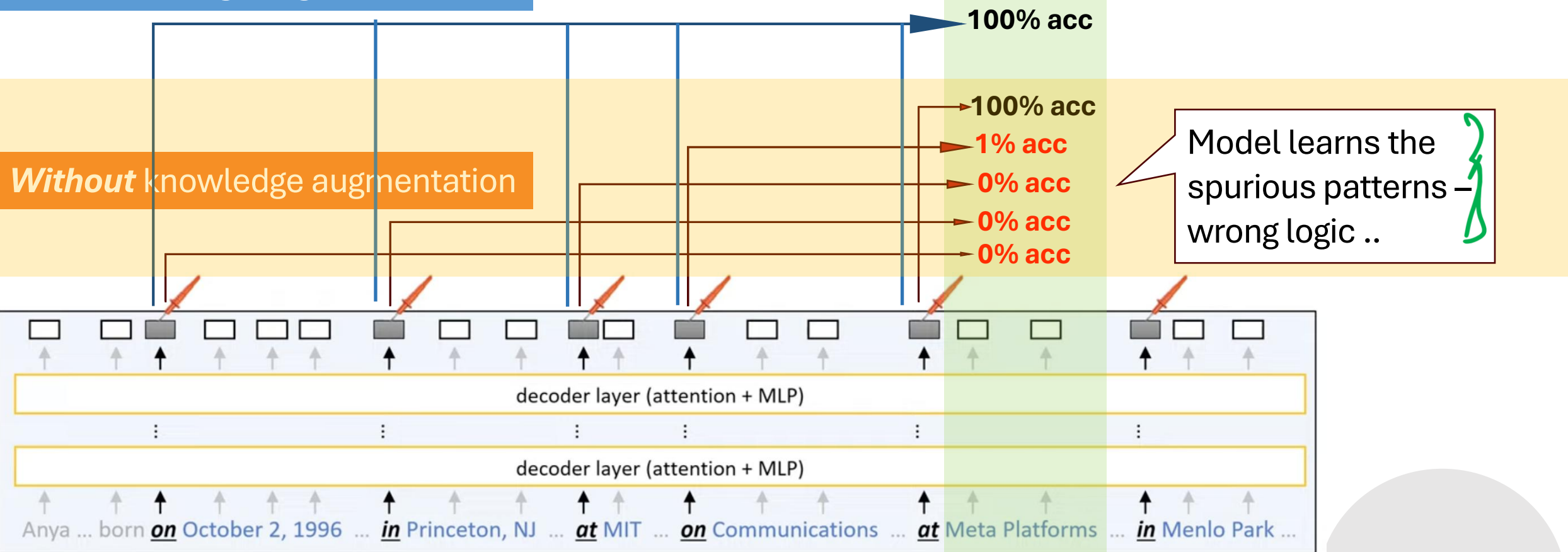




# Result 4: p-Probing

*With* knowledge augmentation

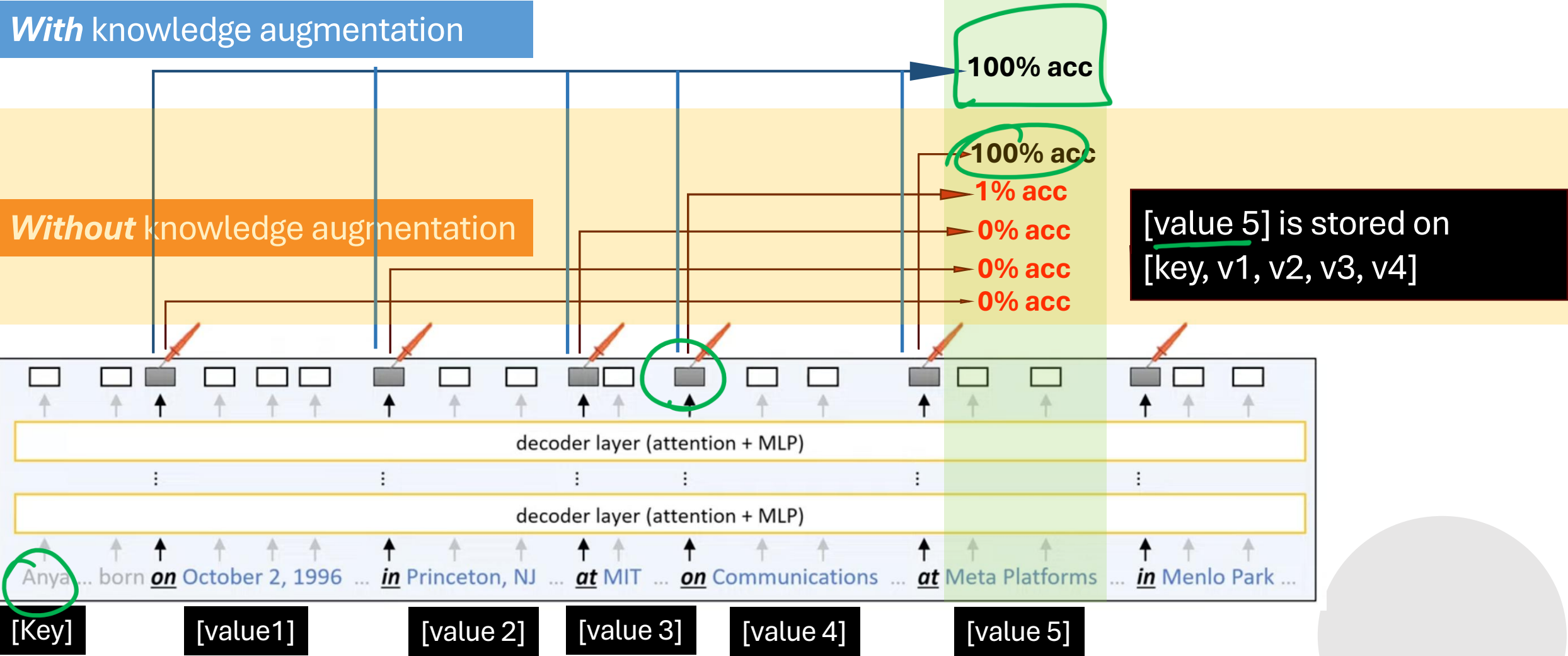
*Without* knowledge augmentation



# Result 4: p-Probing

With knowledge augmentation

Without knowledge augmentation



# Result 4: p-Probing

*With* knowledge augmentation

*Without* knowledge augmentation

Model learns the right logic & stores info. on the key

100% acc

100% acc

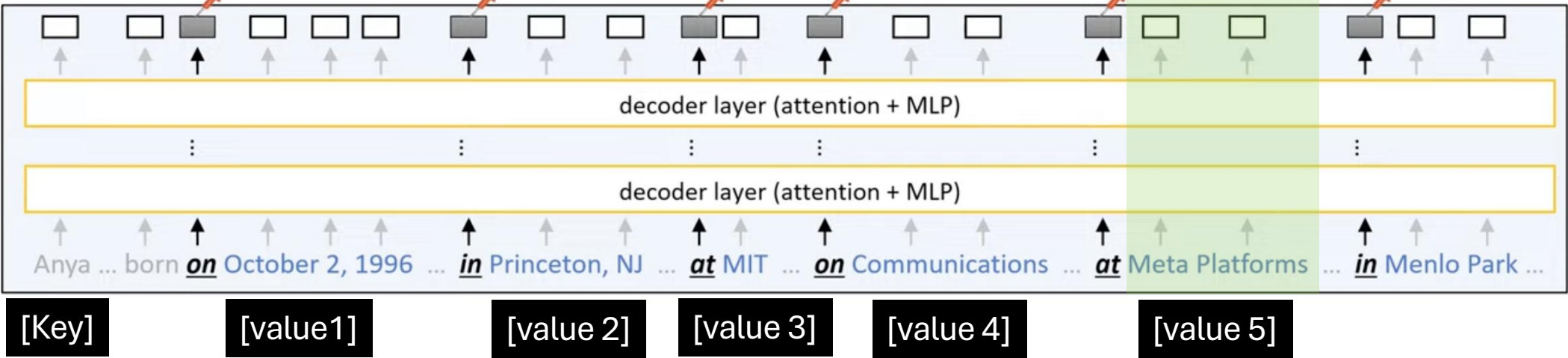
1% acc

0% acc

0% acc

0% acc

[value 5] is stored on [key, v1, v2, v3, v4]



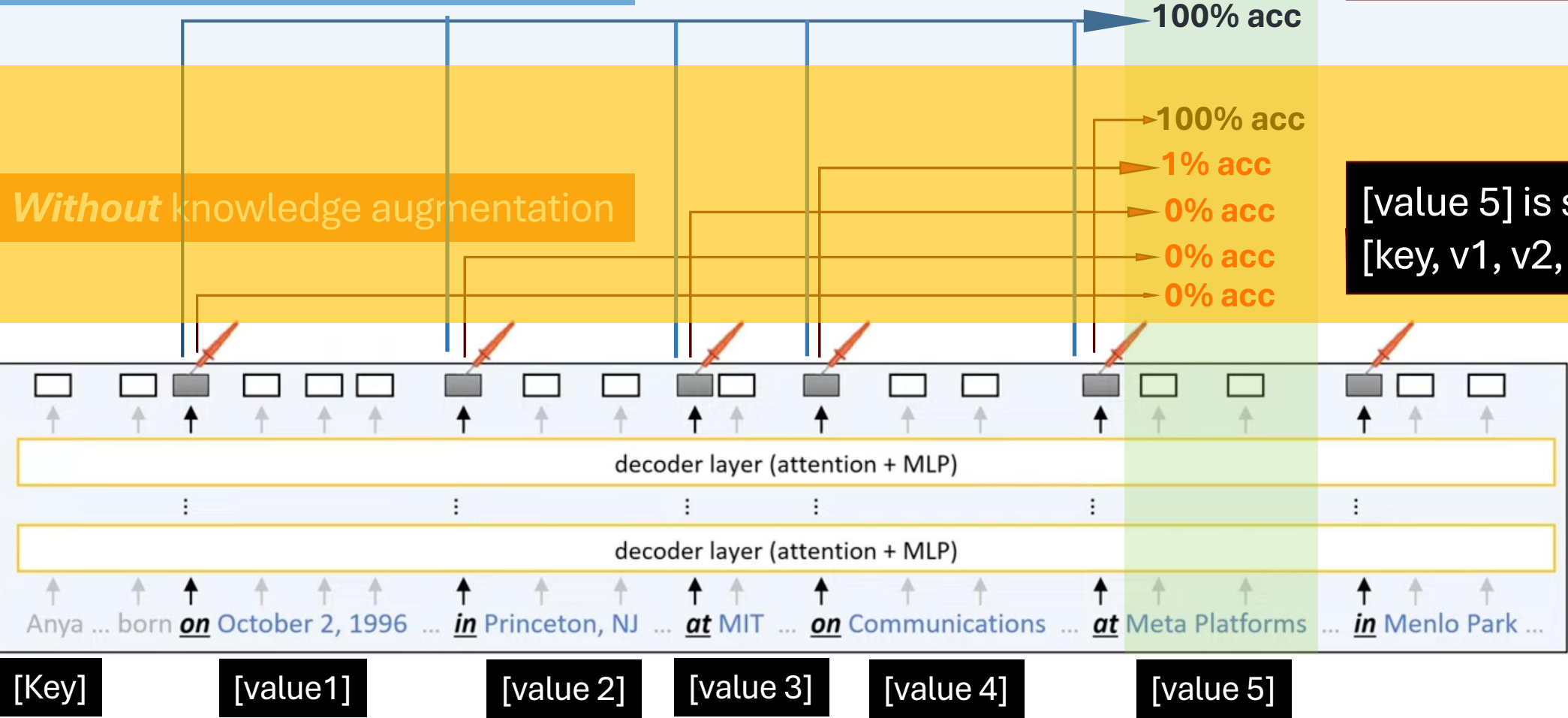
# Result 4: p-Probing

With knowledge augmentation

Without knowledge augmentation

[value 5] is stored only on [key]

[value 5] is stored on [key, v1, v2, v3, v4]



# Result 4: p-Probing

*With* knowledge augmentation

*Without* knowledge augmentation

[value 5] is stored only on [key]

[value 5] is stored on [key, v1, v2, v3, v4]

100% acc

100% acc

1% acc

0% acc

0% acc

0% acc

Pretrained with knowledge augmentation

→ Changes how knowledge is stored

→ Affects whether knowledge can be extracted via QA finetuning or not

[Key]

[value 1]

[value 2]

[value 3]

[value 4]

[value 5]



# Result 4: p-Probing

**With** knowledge augmentation

**Without** knowledge augmentation

[value 5] is stored only on [key]

[value 5] is stored on [key, v1, v2, v3, v4]

100% acc

100% acc

1% acc

0% acc

0% acc

0% acc

Pretrained with knowledge augmentation

→ Changes how knowledge is stored

→ Affects whether knowledge can be extracted via QA finetuning or not



[value 1]

[value 2]

[value 3]

[value 4]

[value 5]





# Result 4: p-Probing

- We claim that for extraction of knowledge, the information must be stored directly on the name
- Can we test this directly?
- Let's design a probe on the name.

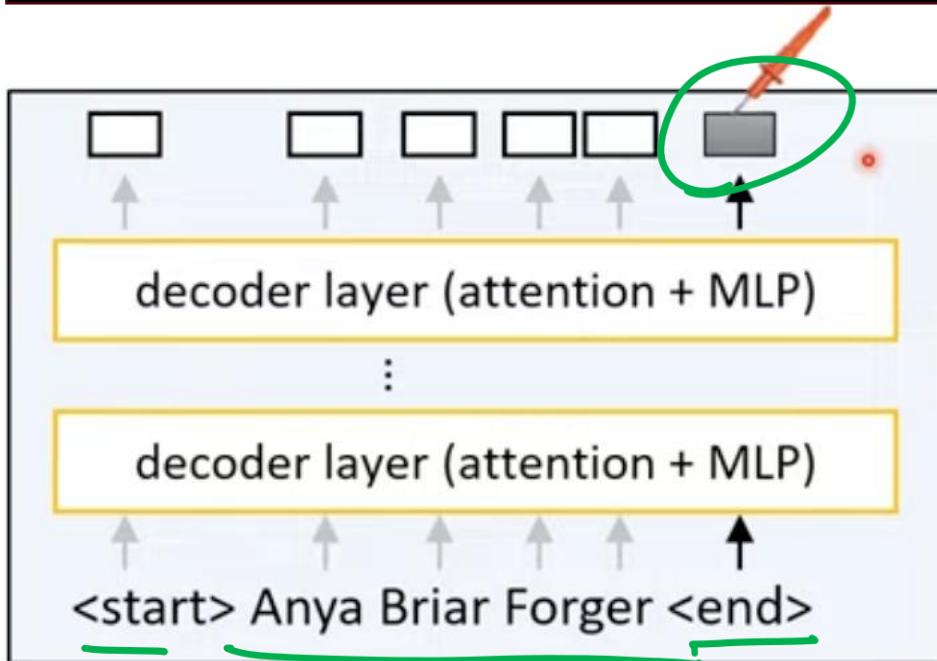


# Result 5: q-Probing

probe training phase



Linear classifier to predict 6 possible knowledges



# Result 5: q-Probing

bioS data [...100k biography entries]

QA<sub>train</sub> [...QAs on 50k individuals]

QA<sub>test</sub>

[...QAs on remaining 50k]

What is the birth date of Sabrina Eugeo Zuberger?

Which university did Sabrina Eugeo Zuberger study?

Which company did Sabrina Eugeo Zuberger work for?

baseline	2.7	0.0	0.5	0.3	1.0	0.4	13.7	8.3	2.5	37.0	4.0	1.5	14.8
bioS single	9.7	33.5	6.3	2.3	4.0	1.1	13.8	63.4	1.9	37.5	3.1	0.2	13.1
bioS single + fullname	48.9	56.2	58.8	63.0	55.7	50.5	14.1	78.8	47.5	65.4	51.0	47.4	28.9
bioS single + permute1	4.4	0.5	3.3	2.4	5.0	3.5	13.7	10.2	1.2	37.5	2.9	0.7	12.8
bioS single													9.1
bioS single													9.1
bioS single													4.7
bioS single													6.7
bioS single													9.9
bioS multi2													3.0
bioS multi2 + fullname	84.0	100	97.7	89.5	97.6	91.5	55.5	100	100	99.7	99.7	99.8	89.2
bioS multi2 + permute	91.2	99.3	98.7	89.8	96.7	83.3	83.5	99.9	99.9	99.6	99.4	93.9	90.2
bioS multi2 + permute + fullname	96.1	100	98.8	91.3	98.1	93.7	97.8	99.7	100	100	99.9	99.9	99.4
bioS multi5	41.0	100	50.8	30.9	43.5	10.2	13.8	100	39.8	40.3	40.1	3.2	13.2
bioS multi5 + fullname	82.4	100	98.6	88.4	96.1	91.9	26.8	100	100	98.3	99.4	98.5	63.1
bioS multi5 + permute	96.6	100	99.0	91.3	97.7	95.1	98.7	100	100	99.9	100	99.9	99.8
bioS multi5 + permute + fullname	96.2	100	98.7	90.6	97.9	93.7	99.0	100	100	99.8	100	100	99.8

“attribute directly saved to the person’s name” is a crucial factor for effective knowledge extration

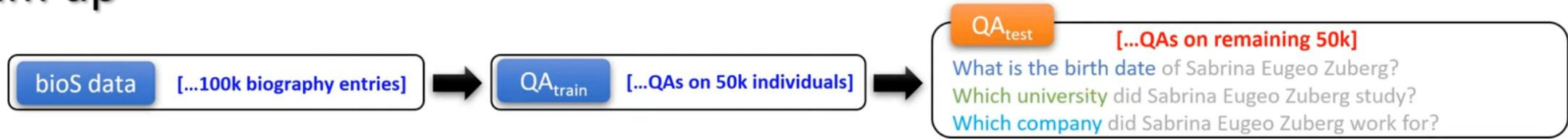
QA accuracy

QA mean acc	QA b_date	QA b_city	QA univ	QA major	QA c_name	QA c_city
Q-first b_month	Q-first b_city	Q-first univ	Q-first major	Q-first c_name	Q-first c_city	

Q-prob accuracy



# Sum up



knowledge **augmented** in pretrain data ↔ attribute is **nearly entirely** stored on person's names (after pretrain)

↔ knowledge **can be extracted** (after instruct finetune)

**no** knowledge augmentation in pretrain data ↔ attribute is **not** entirely stored on person's names (after pretrain)

↔ knowledge **cannot be extracted** (after instruct finetune)

Do we have to augment all biographies?



# Result 6: Celebrity Can Help Minority

pretrain

bioS data

[100k minority biography entries,  
*non-augmented*]

e.g bioS single+permute

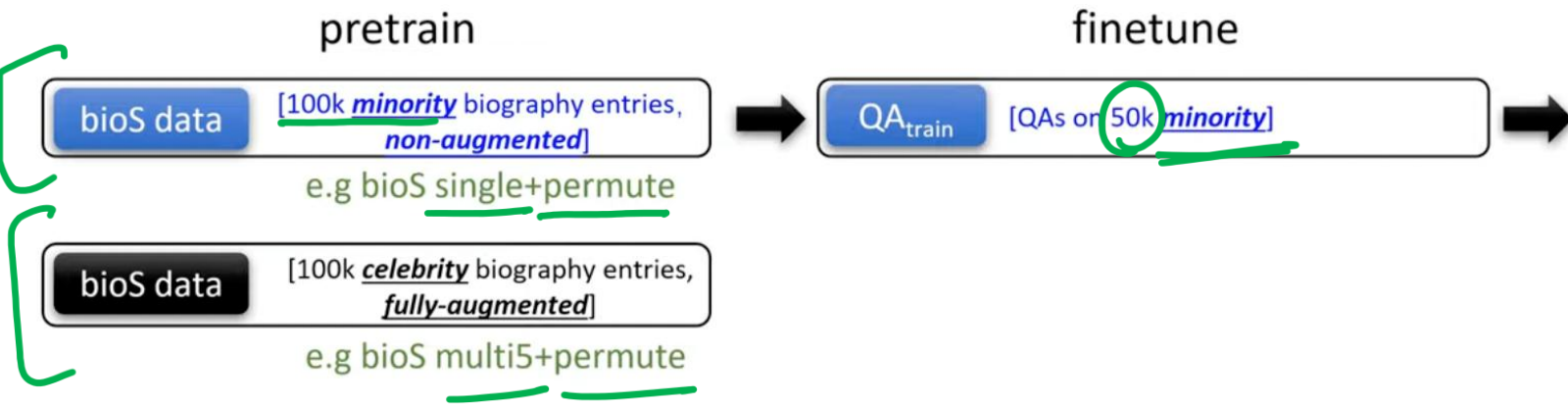
bioS data

[100k celebrity biography entries,  
*fully-augmented*]

e.g bioS multi5+permute

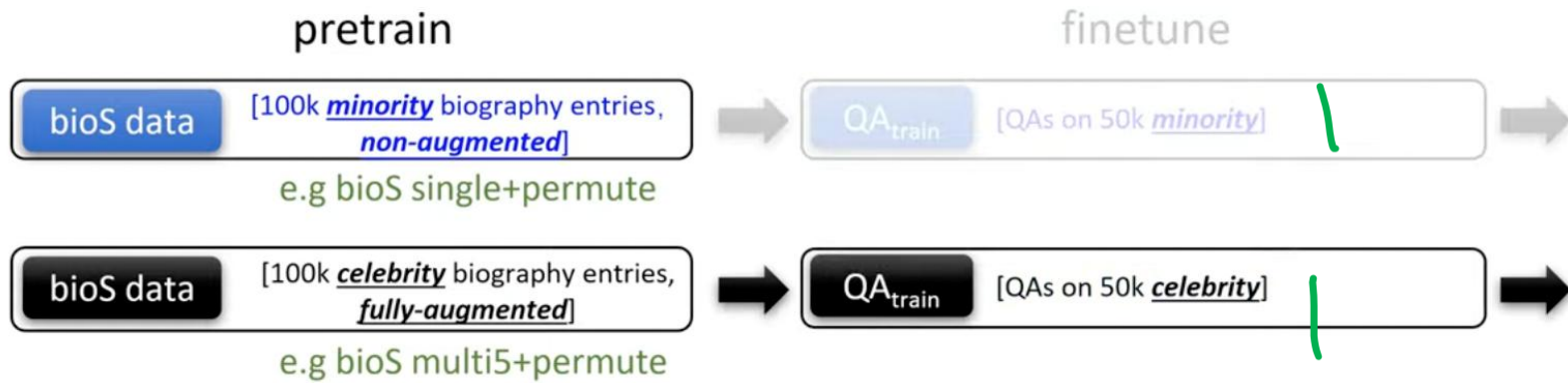


# Result 6: Celebrity Can Help Minority





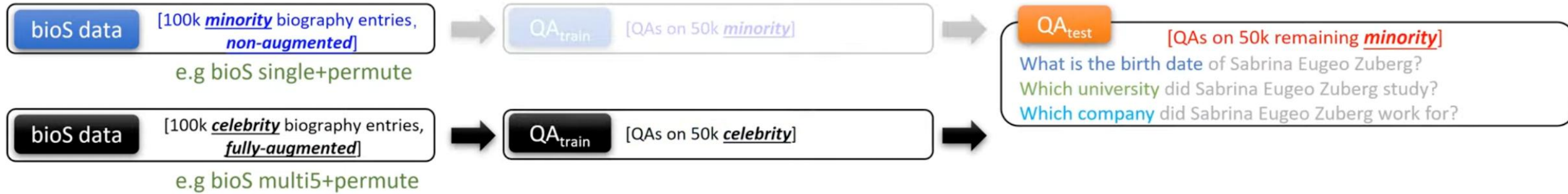
# Result 6: Celebrity Can Help Minority



# Result 6: Celebrity Can Help Minority

pretrain

finetune



# Result 6: Celebrity Can Help Minority

pretrain

bioS data [100k minority biography entries, non-augmented]

e.g bioS single+permute

bioS data [100k celebrity biography entries, fully-augmented]

e.g bioS multi5+permute

QA<sub>train</sub>

[QAs on 50k celebrity]

QA<sub>test</sub>

[QAs on 50k remaining minority]

What is the birth date of Sabrina Eugeo Zuberger?  
Which university did Sabrina Eugeo Zuberger study?  
Which company did Sabrina Eugeo Zuberger work for?

baseline	2.7	0.0	0.5	0.3	1.0	0.4	13.7
bioS single + permute1	4.4	0.5	3.3	2.4	5.0	3.5	13.7
bioS single + permute1 + CEL	86.8	98.3	96.8	90.7	90.2	71.7	80.1
bioR single	10.0	25.1	13.9	2.4	5.5	2.0	14.1
bioR single + wiki	7.3	18.4	5.2	2.6	4.3	1.8	14.1
bioR single + CEL	76.3	94.3	85.3	82.9	79.4	67.0	56.6
QA mean acc							
QA b_date							
QA b_city							
QA univ							
QA major							
QA c_name							
QA c_city							



# Result 6: Celebrity Can Help Minority

pretrain

bioS data [100k minority biography entries, non-augmented]

e.g bioS single+permute

bioS data [100k celebrity biography entries, fully-augmented]

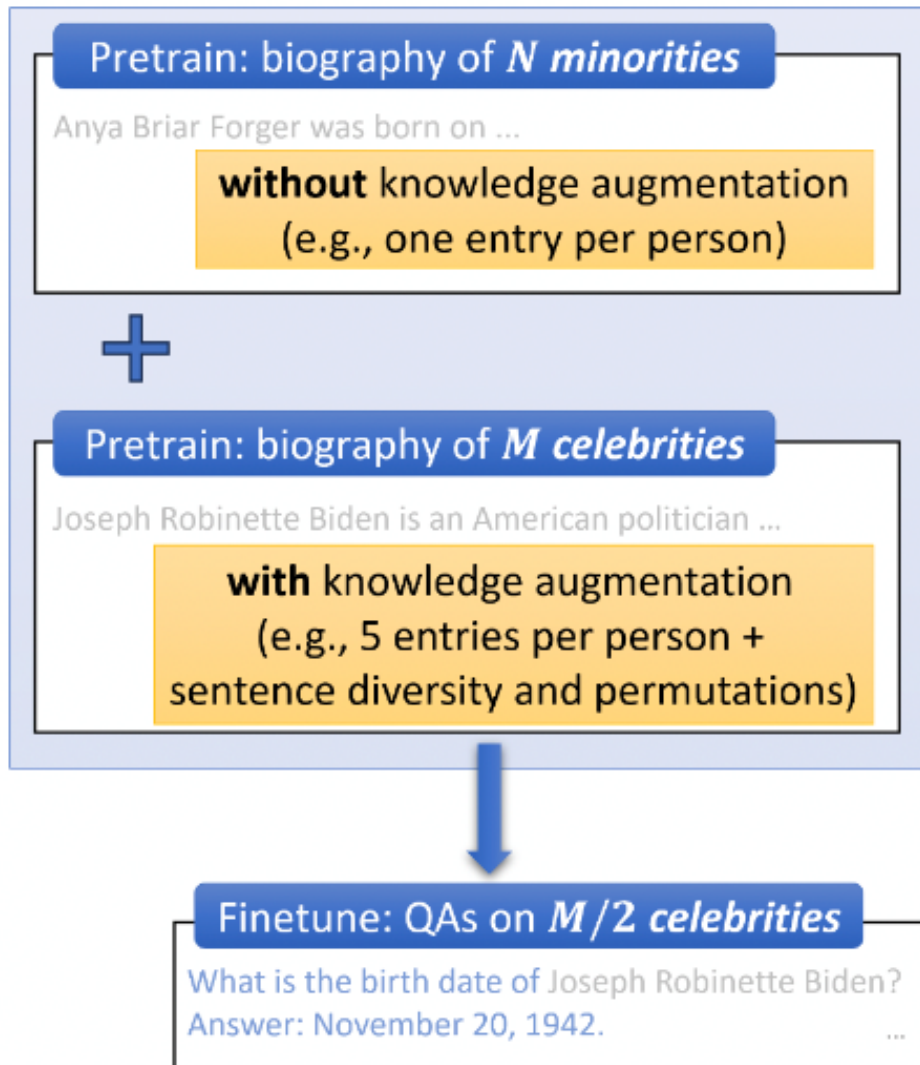
e.g bioS multi5+permute

QA<sub>train</sub> [QAs on 50k celebrity]

QA<sub>test</sub> [QAs on 50k remaining minority]  
What is the birth date of Sabrina Eugeo Zuberger?  
Which university did Sabrina Eugeo Zuberger study?  
Which company did Sabrina Eugeo Zuberger work for?

baseline	2.7	0.0	0.5	0.3	1.0	0.4	13.7
bioS single + permute1	4.4	0.5	3.3	2.4	5.0	3.5	13.7
bioS single + permute1 + CEL	86.8	98.3	96.8	90.7	90.2	71.7	80.1
bioR single	10.0	25.1	13.9	2.4	5.5	2.0	14.1
bioR single + wiki	7.3	18.4	5.2	2.6	4.3	1.8	14.1
bioR single + CEL	76.3	94.3	85.3	82.9	79.4	67.0	56.6
QA mean acc		QA b_date	QA b_city	QA univ	QA major	QA c_name	QA c_city





Knowledge augmentation on *celebrity*  $\Rightarrow$   
Knowledge extraction for *minority*

Even if knowledge augmentation is applied to a subset of individuals, what we call celebrities, test accuracy for others also increases.

The mere inclusion of celebrity data (e.g., people with plentiful online biographical data of diverse writing styles) in pre-training enhances the model's knowledge extraction for minorities.

**Evaluate: QAs on  $N$  minorities**

What is the birth date of Anya Briar Forger? ...  
**~80% test acc**



# Summary of Part 3.1

knowledge storage

100% word-by-word memorization (biography)



knowledge extraction

QA: What's the birthdate of Anya?





# Summary of Part 3.1

## knowledge storage

100% word-by-word memorization (biography)



## knowledge extraction

QA: What's the birthdate of Anya?

## mixed training (BIO+QA)

"add instruct-finetune data to pretrain"



80%.



# Summary of Part 3.1

knowledge storage

100% word-by-word memorization (biography)



knowledge extraction

QA: What's the birthdate of Anya?

**mixed training (BIO+QA)**

"add instruct-finetune data to pretrain"



**pretrain (BIO) + finetune (QA)**



without knowledge augment



knowledge augmentation  
(sufficient on the celebrities)



# Summary of Part 3.1

## knowledge storage

100% word-by-word memorization (biography)



## knowledge extraction

QA: What's the birthdate of Anya?

## mixed training (BIO+QA)

"add instruct-finetune data to pretrain"



## pretrain (BIO) + finetune (QA)



without knowledge augment



knowledge augmentation  
(sufficient on the celebrities)

## bi-directional model (BERT, DeBERTa)



even with mixed training  
+ knowledge augment





# Physics of Language Models: Part 3.1

## Knowledge **Storage and Extraction**

### Result 1

mixed training  $\Rightarrow$  knowledge extraction  
mixed training means “add instruct-finetune data to pretrain stage”

### Result 2-3

instruct finetune  $\nRightarrow$  knowledge extraction  
unless pretrain data is augmented (e.g. diversified writing styles)

### Result 4-5

two probing techniques to explain why this happens  
can detect *how* and *where* knowledge is stored in an LLM

### Result 6

knowledge on “celebrity” helps “minority”  
augmenting pretrain data for some knowledge helps others



# Physics of Language Models: Part 3.1

## Knowledge **Storage and Extraction**

### Result 1

mixed training  $\Rightarrow$  knowledge extraction

mixed training means “add instruct-finetune data to pretrain stage”

### Result 2-3

instruct finetune  $\nRightarrow$  knowledge extraction

unless pretrain data is augmented (e.g. diversified writing styles)

### Result 4-5

two probing techniques to explain why this happens

can detect *how* and *where* knowledge is stored in an LLM

### Result 6

knowledge on “celebrity” helps “minority”

augmenting pretrain data for some knowledge helps others

### Result 7

encoder models like BERT  $\nRightarrow$  knowledge extraction

discover and explain why this happens



**Question:** Was Sachin Tendulkar born in an even year?

Incorrectly

**C:** The model can't extract the birth year of Sachin

**D:** Model doesn't know what "**even**" means

**E:** Model can extract birth year, knows what is even, still can't answer

**C1:** model didn't see Sachin's biography during training

**C2:** model didn't finetune enough to extract birthdays

**C3:** model can extract b'days of Sehwag but not Sachin

## Part 3.1: Knowledge Storage and Extraction





**Question:** Was Sachin Tendulkar born in an even year?

Incorrectly

**C:** The model can't extract the birth year of Sachin

**D:** Model doesn't know what "**even**" means

**E:** Model can extract birth year, knows what is even, still can't answer

Let us assume that model can extract knowledge

C1: model didn't see Sachin's biography during training

C2: model didn't finetune enough to extract birthdays

C3: model can extract b'days of Sehwag but not Sachin

Part 3.1: Knowledge Storage and Extraction



**Question:** Was Sachin Tendulkar born in an even year?

Incorrectly

~~D: Model doesn't know what "even" means~~

E: Model can extract birth year, knows what is even, still can't answer

Let us assume that model can extract knowledge

How good the model is in *manipulating this knowledge*?



**Question:** Was Sachin Tendulkar born in an even year?

Incorrectly

**D:** Model doesn't know what **"even"** means

**E:** Model can extract birth year, knows what is even, still can't answer

Let us assume that model can extract knowledge

How good the model is in *manipulating this knowledge*?

e.g. to answer this question, model has to 1<sup>st</sup> extract birth date and then do *classification* into even and odd



**Question:** Was Sachin Tendulkar born in an even year?

Incorrectly

Correctly

**D:** Model doesn't know what "even" means

**E:** Model can extract birth year, knows what is even, still can't answer

**A:** Model infers from "1973 is odd" and "Sachin was born in 1973"

**B:** Pretraining data directly contains the information that "Sachin was born in even year"

### Part 3.2: Knowledge Manipulation

Classification & comparison

Inverse search

Partial / Dual Retrieval



# Physics of Language Models: Part 3.2

## Knowledge Manipulation

### Results 1-2

knowledge partial / dual retrievals may be difficult

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

### Results 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”

### Results 6/8/9

Turing tests to distinguish Humans from contemporary AIs

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



