How do language models learn facts?
Dynamics, curricula and hallucinations

Nicolas Zucchet

Jörg Bornschein

Stephanie Chan

Andrew Lampinen

Razvan Pascanu

Soham De

Google DeepMind

ETH zürich

Large language models are becoming our main gateway to human knowledge

This knowledge is acquired through learning

Understanding this process and how data impacts it becomes critical

Studying how LLMs learn is challenging

Real-world data is a complicated mix of **different sources** which require learning **different abilities** in parallel

Our approach: train LLMs in a **controlled setting** in which we exactly know which abilities the data requires

A synthetic biography dataset to study knowledge

N individuals with some **attributes** (e.g. birthdate, birthplace)

Bank of **template sentences** for each type of attributes

James Frida Zhu's life began on March 16, 2042.

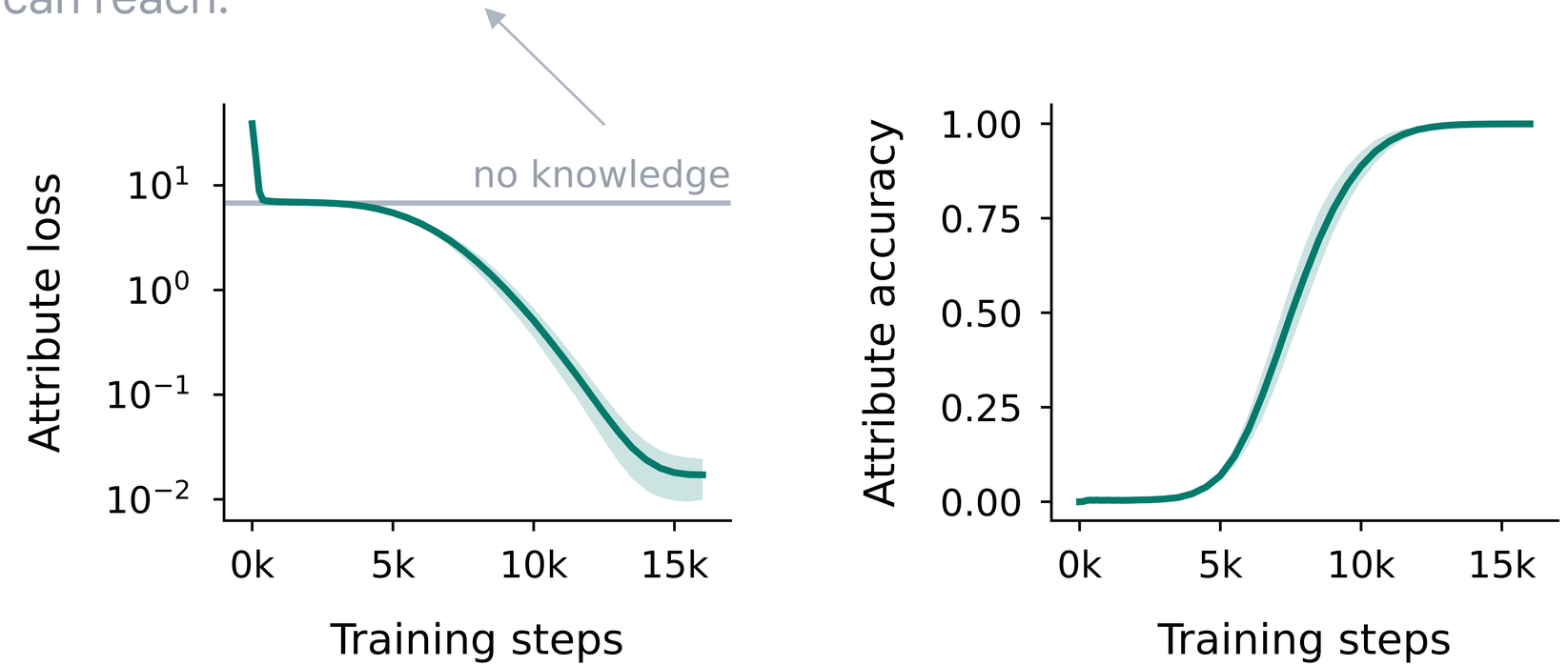
James Frida Zhu is a native of Shanghai.

Predicting attribute tokens is a factual recall task which measures the model's knowledge.

Adaptation of the setup from Allen-Zhu and Li 2023

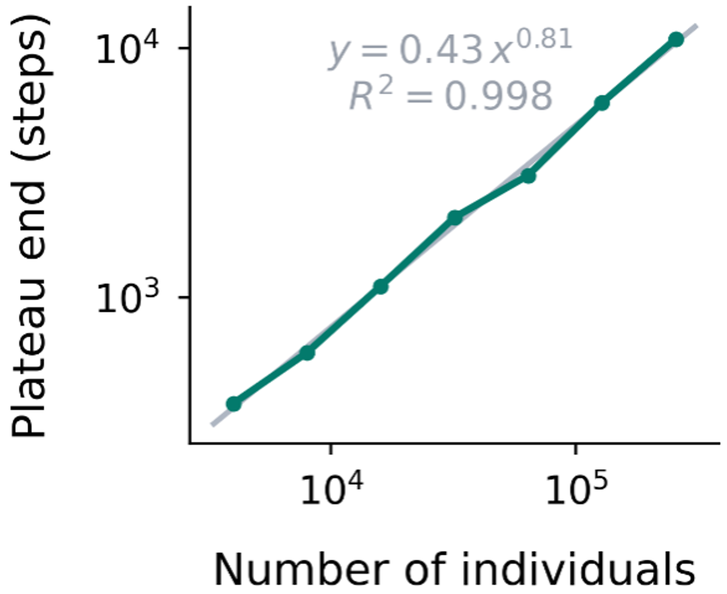
Knowledge acquisition happens in three phases

Best performance a model **without** individual-specific knowledge can reach.

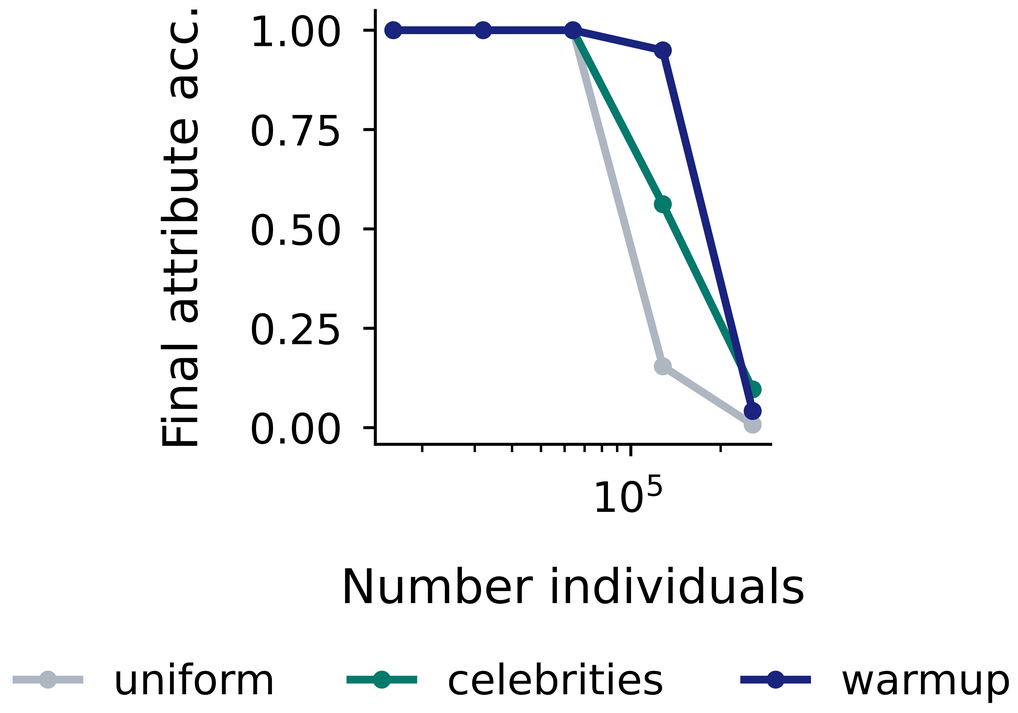


Under the hood: **attention-based circuits** are created during the plateau
In the paper: novel attention patching technique to show this

Lower data diversity can speed up learning



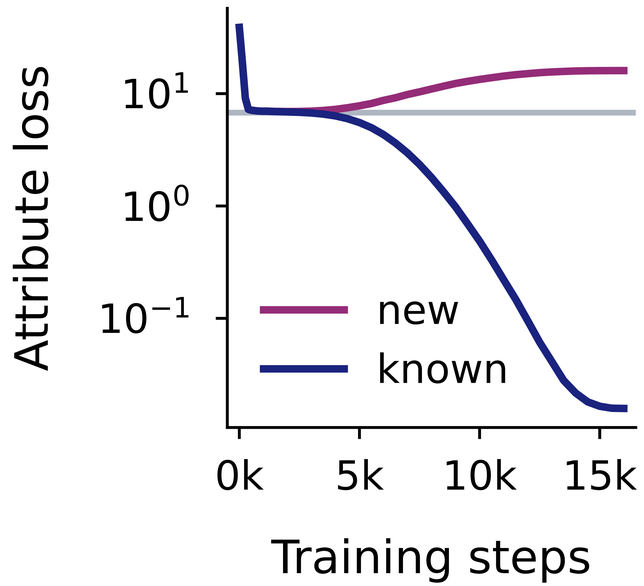
Faster to escape the plateau when there are **fewer individuals** in the training data



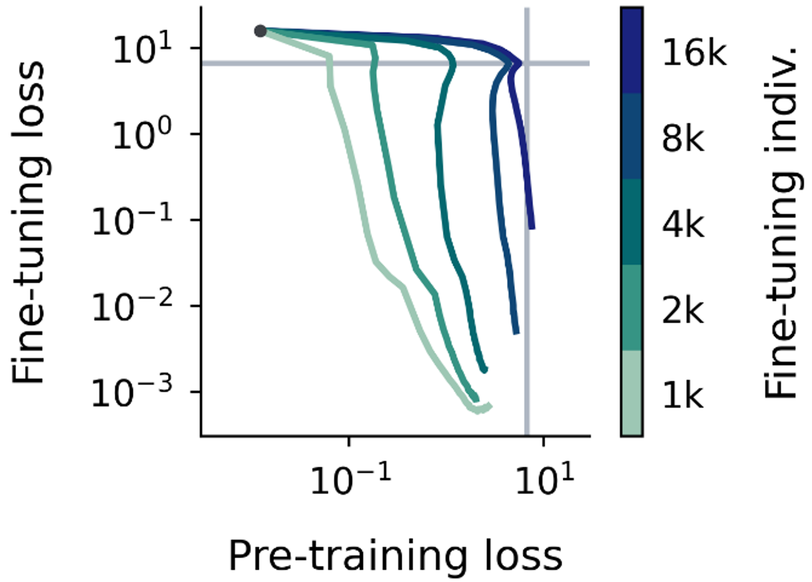
Imbalances can increase final knowledge

Tailored curriculum even more

Fine-tuning on new knowledge is challenging



Hallucinations (overconfident wrong predictions) appear **concurrently** to knowledge



Fine-tuning quickly **destroys** existing knowledge

Implications for LLMs

Language models acquire knowledge in three phases

Implication: data used **early** on during the plateau is **forgotten**

Low data diversity can speed up learning

Implication: LLMs might learn **factual recall faster** because internet data is **skewed**

Implication: **diversity-based curriculum** might be a powerful tool

Incorporating new knowledge is hard

Implication: (naïve) **fine-tuning is not suited** to add new knowledge into LLMs

The case for controlled studies

LLM training is influenced by **many factors**

Controlled studies on synthetic data enable to **focus on specific abilities**

General methodology: we can gain a lot more scientific understanding with similar approaches

How do language models learn facts?
Dynamics, curricula and hallucinations

Nicolas Zucchet

Jörg Bornschein

Stephanie Chan

Andrew Lampinen

Razvan Pascanu

Soham De

Paper