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

Nicolas Zucchet, Jörg Bornschein, Stephanie Chan, Andrew Lampinen, Razvan Pascanu, Soham De

ETH zürich

Google DeepMind

Oral
Thursday
3:45pm

Motivation

Large language models are becoming our main gateway on human knowledge

This knowledge is acquired through learning

Understanding this process and how data impacts it become increasingly important

It will improve our trust in LLMs and may help us design better training strategies

Studying how LLMs learn is challenging

Real-world data is a complicated mix of different sources which require different abilities

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

Using synthetic data: precise control on

1. abilities needed to solve the task
2. data the model has seen + its properties

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

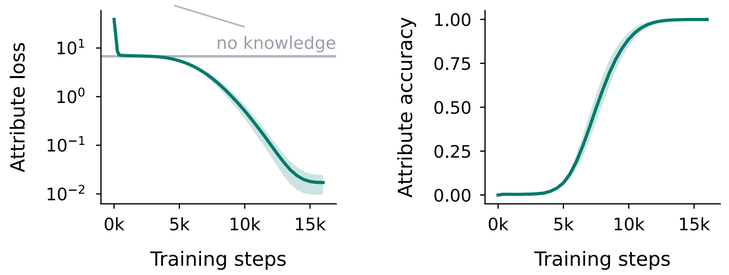
Some important technical details:

- name and attribute type always before attribute value
- test model on unseen templates (no pure memorization)

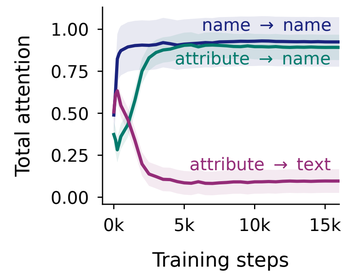
Allen-Zhu & Li, 2023

Knowledge acquisition happens in 3 phases

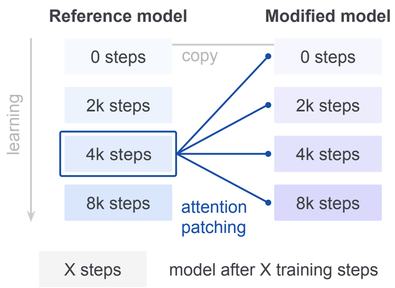
Best performance a model **without** individual-specific knowledge can reach

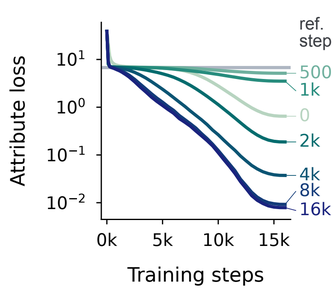


Attention-based circuits are created during the plateau

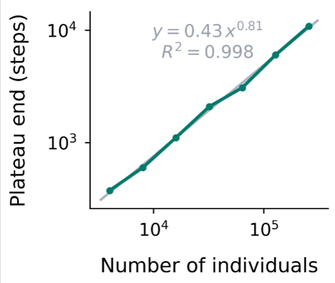


name	text
James Frida Zhu is a native of Shanghai.	
attribute	





Lower data diversity can speed up learning

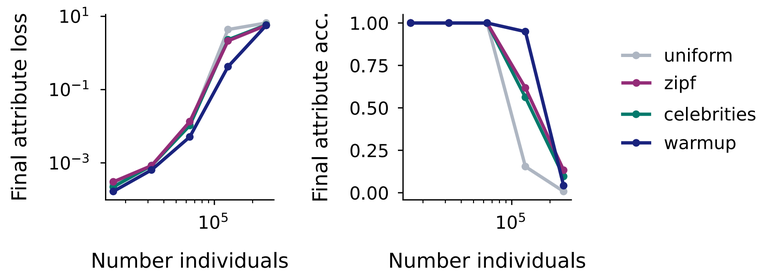


Faster to escape the plateau when there are fewer individuals in the training data

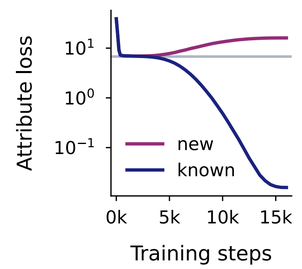
What if the individual distribution becomes more skewed? Includes more individuals over time?

Imbalances (zipf / celebrities) can increase final knowledge

Tailored curriculum (warmup) even more



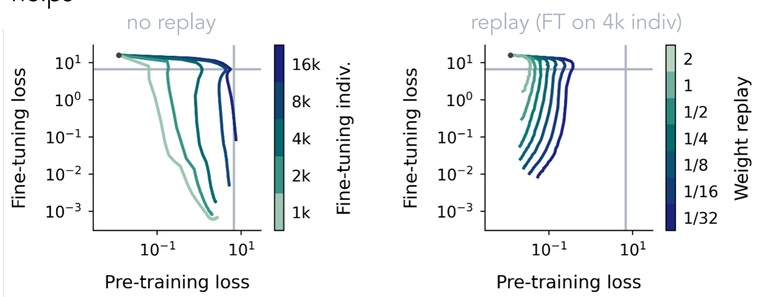
Fine-tuning on new knowledge is challenging



Hallucinations (overconfident wrong predictions) appear concurrently with knowledge

Creates a tough start for fine-tuning…

Fine-tuning quickly destroys existing knowledge, replay partially helps



Takeaways for LLM training

Controlled studies on synthetic data brings detailed understanding of the learning process

Data used early on (during the plateau) is forgotten

Diversity-based curriculum might be a powerful tool

(Naïve) fine-tuning is not suited to adding new knowledge