How do LMs learn facts?
Dynamics, curricula and hallucinations

Nicolas Zucchet[1,+], Jörg Bornschein[2], Stephanie Chan[2], Andrew Lampinen[2], Razvan Pascanu[2] and Soham De[2]
[1]ETH Zürich, [2]Google DeepMind, [+]Work done at Google DeepMind

Google DeepMind

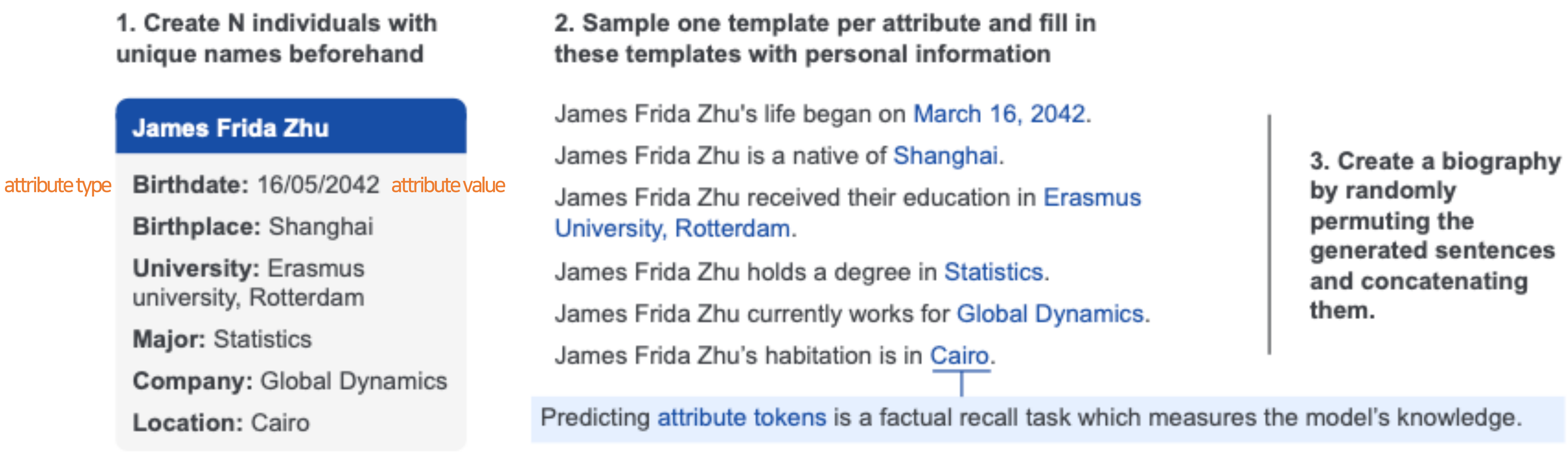
HUMANE Lab 박현빈

25.04.11

Key Findings

- LMs learn in 3 phases
  - distribution statistics -> performance plateaus -> acquiring knowledge
- The training data distribution impacts learning dynamics, as imbalanced distribution lead to shorter plateaus
- Hallucinations emerge simultaneously with knowledge

Synthetic Biographies



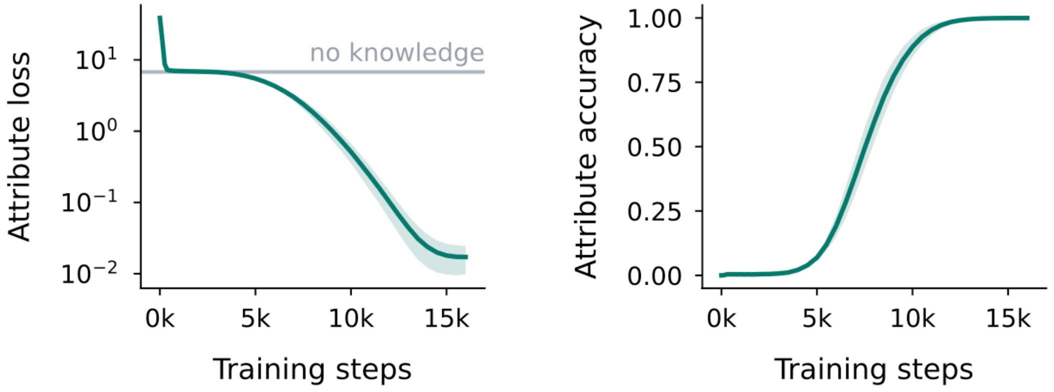
- 25 templates per attribute type (20 for training, 5 for evaluation)
- Keep 5 for evaluation to measuring knowledge rather than memorization
- Evaluate the model using *attribute loss* and *attribute accuracy*

Training Setup

Model (decoder-only Transformer)		
parameters	44M	Number of parameters.
n_layers	8	Number of layers.
n_heads	8	Number of attention heads per attention layer.
d_model	512	Model dimension (residual stream).
d_hidden	2048	Hidden dimension of the multi-layer perception.
key_size	64	Dimension of the key and values.
sequence_mixer	Attention	Which sequence mixing block we are using.
Training		
training_steps	16k	Number of training steps.
batch_size	128	Batch size.
lr_scheduler	cosine	Learning rate scheduler, default is cosine scheduler (no warm-up, final learning rate: 10^{-7}).
lr	$4 \cdot 10^{-4}$	Maximum learning rate.
weight_decay	0.1	Weight decay.
optimizer	AdamW	Optimizer (momentum parameters for the AdamW optimizer are $\beta_1 = 0.9$, $\beta_2 = 0.95$).

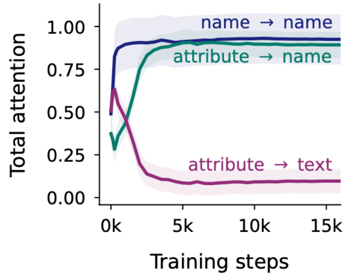
How language models acquire knowledge

- Initial language understanding
  - the network learns the overall attribute value distribution
  - at the end of this phase, it performs like an optimal model without individual-specific knowledge
- Performance plateaus
- Knowledge emergence



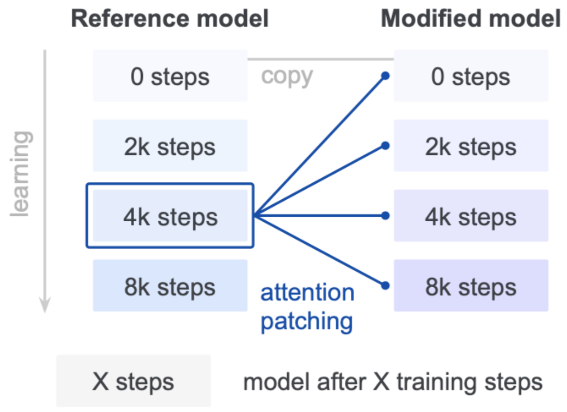
How language models acquire knowledge

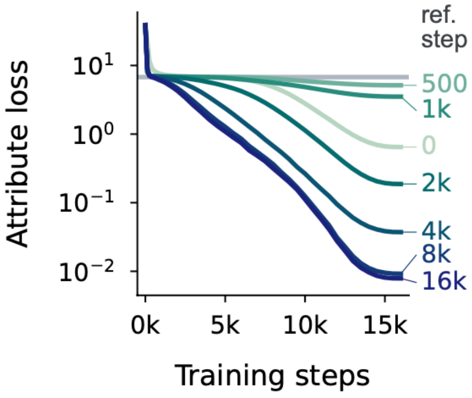
- Why occur performance plateaus
  - losses are not properly backpropagated from attribute value tokens to name tokens
- Performing attention circuit during the plateau
  - name tokens grouping circuit
    - the first attention layer aggregates name tokens together to form a representation of the individual's name at position of the last name token
  - extraction circuit
    - the final attention layer give high attention to name tokens when predicting the first attribute value token



Attention Patching Experiment

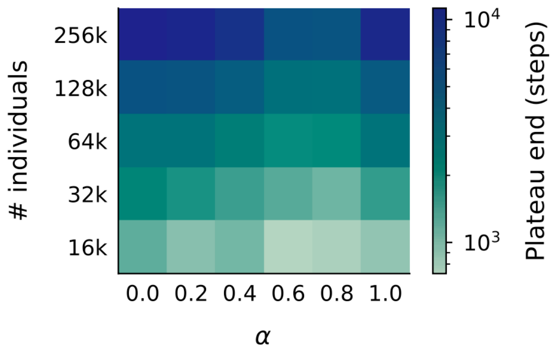
- Hypothesis: a model with learned attention patterns should learn faster than one pre-learning patterns (shorter plateaus)
- First train a reference model, and save reference checkpoints
- Then restart training from scratch, but replace the model's attention patterns with those produced from one of the reference checkpoints





Data Curricula

- While each individual appears with equal frequency in the preceding analysis, real-world data can be significantly less balanced
- Plateau length minimized in highly imbalanced distributions, whereas knowledge acquisition speed maximized in uniform distributions

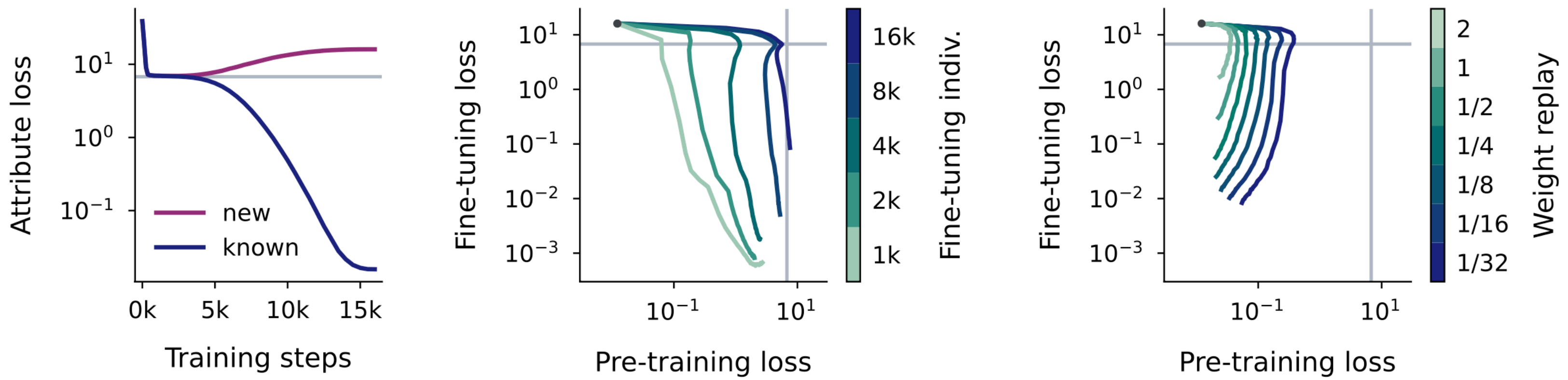


$\alpha = 0$ means uniform distribution, $\alpha = 1$ means highly imbalanced distribution

- warm-up strategy: the model initially trains on a subset of individuals for a fixed number of steps, and on all individuals afterwards

Hallucination

- Hallucinations emerge simultaneously with knowledge acquisition
- Performance on rapidly collapse during the initial stages of fine-tuning, with very little corresponding acquisition of new knowledge
- Incorporating replay data about pre-training individuals in the fine-tuning set partially mitigates the collapse and facilitates the restoration of corrupted knowledge



Conclusion

- Language models first learn statistical patterns, then form attention circuits during a plateau phase before acquiring knowledge
- The more imbalanced the probability of individuals appearing in training data, the sooner the plateau ends. Conversely, the more uniform the probabilities, the more extensive the knowledge acquisition
- Dynamically adjusting the probability distribution during training can yield optimal outcomes
- Hallucinations occur simultaneously with knowledge acquisition and hinder learning new data
- When fine-tuning a model, a substantial amount of previously learned knowledge is lost; however, incorporating pre-training data into the fine-tuning dataset can mitigate this knowledge loss

My Review

- This study offers valuable insights into the internal mechanisms of LMs, which have traditionally been treated as black boxes
- It conducts the analysis using a simplified dataset and small-scale models
- However, it would have been more informative if broader datasets had also been considered.