FORGETTER with forgetful hyperparameters and recurring sleeps can continue to learn beyond normal overtfitting limits

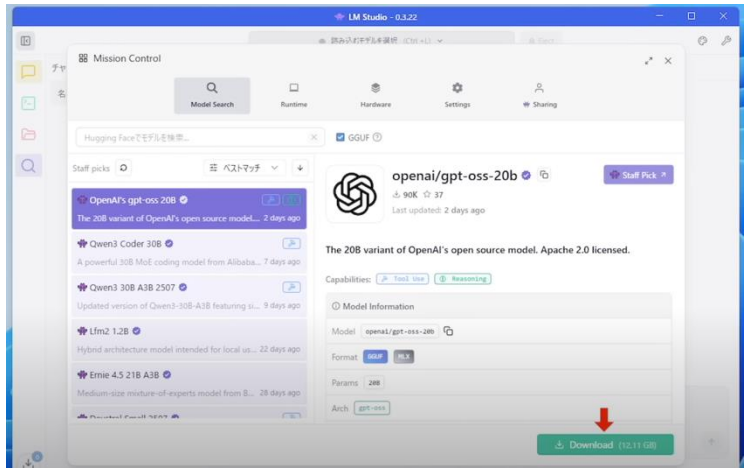
Rui Yamamoto and Keiji Miura (Kwansei Gakuin Univ)

Nov 8, 2025

EMNLP 2025 BabyLM Workshop, Suzhou, China

ChatGPT forced us to reconsider what tasks only humans can perform

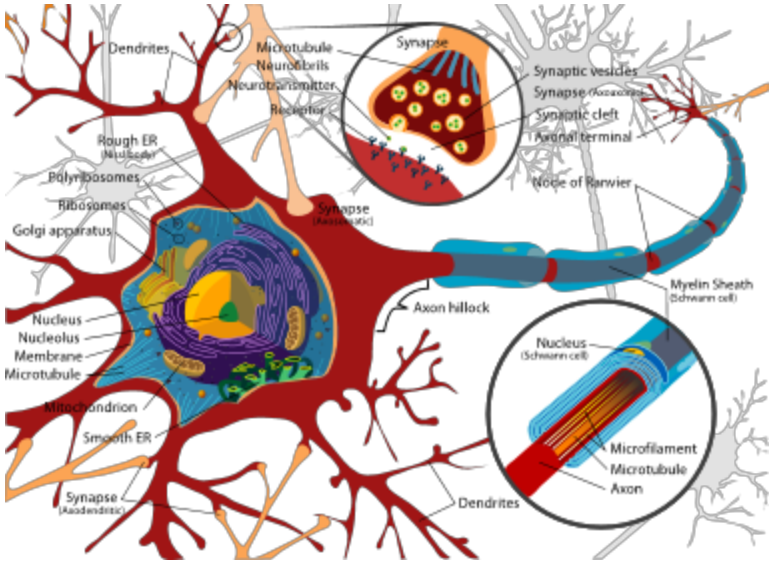
(Priyanka Deo, *TIMESOFINDIA.COM*, Mar 7, 2023)



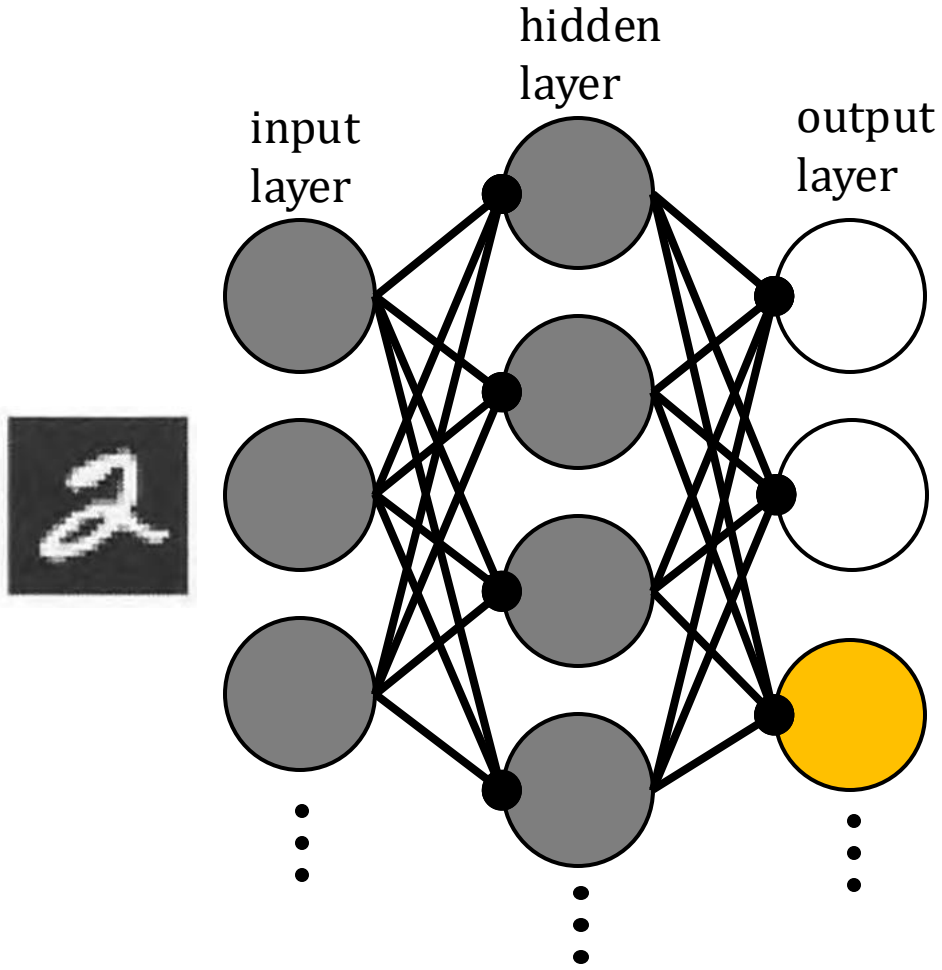
Free local app version was released on August 5, 2025!

A neural network mimics the brain

Neuron

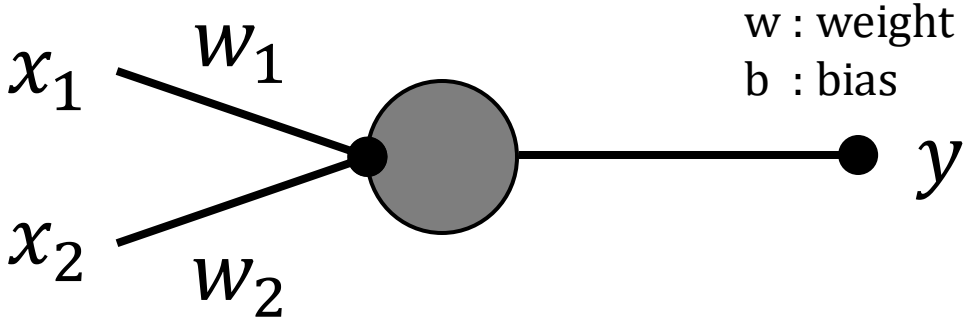


Neural network



Artificial neuron

$$y = f(\Sigma w_i x_i + b)$$



w : weight
b : bias

Understanding ChatGPT may help us understand how the brain generates language

Room for more efficient learning?

- But, LLMs still suffer from considerable computational costs in training
- A biologically plausible curriculum learning that reduces the learning cost is desired (especially for training GPTs)

Goal: exploring efficiency in pretraining

- Optimize the network structure of minGemma (a variant of GPT) for pretraining with BabyLM dataset
- FORGETTER, a new curriculum learning, in which a model forgets a part of state variables for optimization after every sleep and all the hyperparameters are set toward forgetting memory

BabyLM dataset

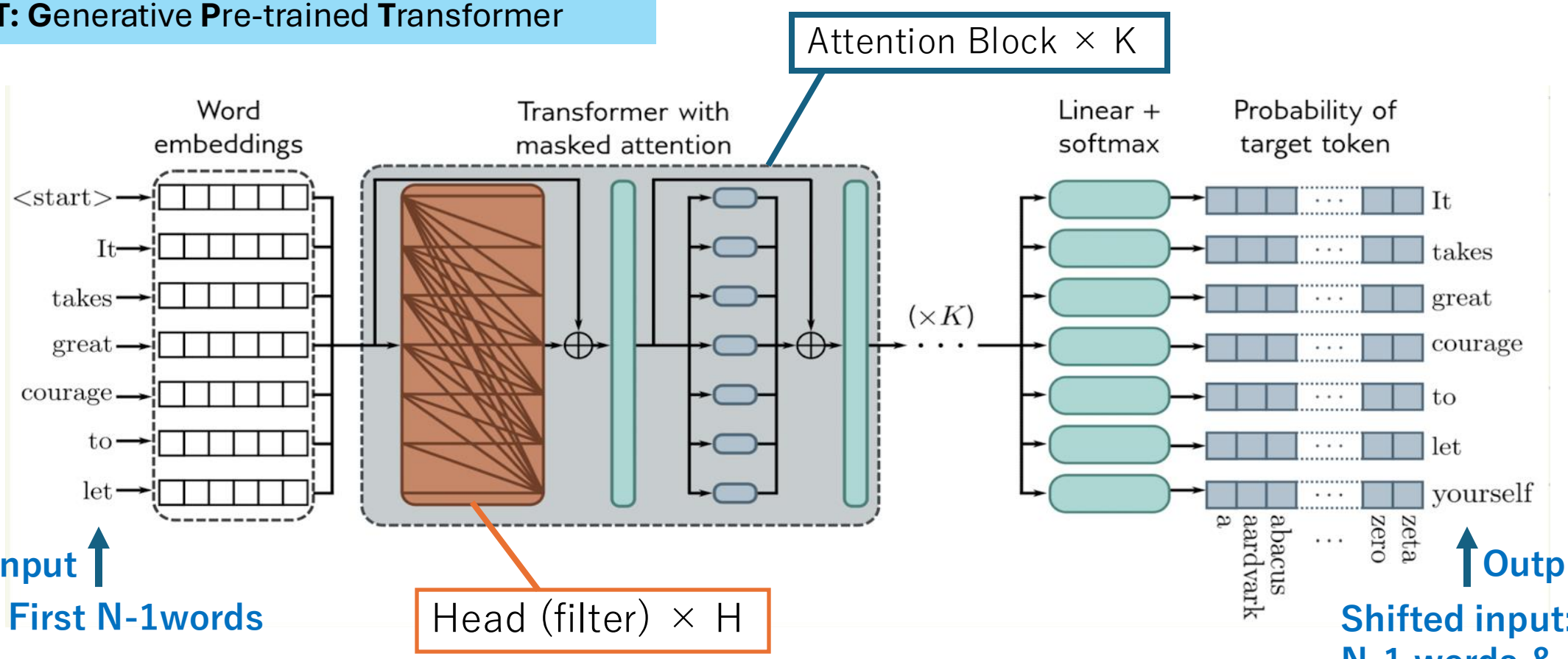
BabyLM contest:

Simulating the language acquisition process by training a neural network to learn baby language

	# tokens	contents
BabyLM 10M	**10Million**	**Texts and everyday conversations that children usually come across (Intended for infants and toddlers aged 1 to 4 years old)**
BabyLM 100M	**100Million**	
WikiText-2	**2Million**	**Articles in Wikipedia**

GPT consists of attention layers

GPT: Generative **P**re-trained **T**ransformer

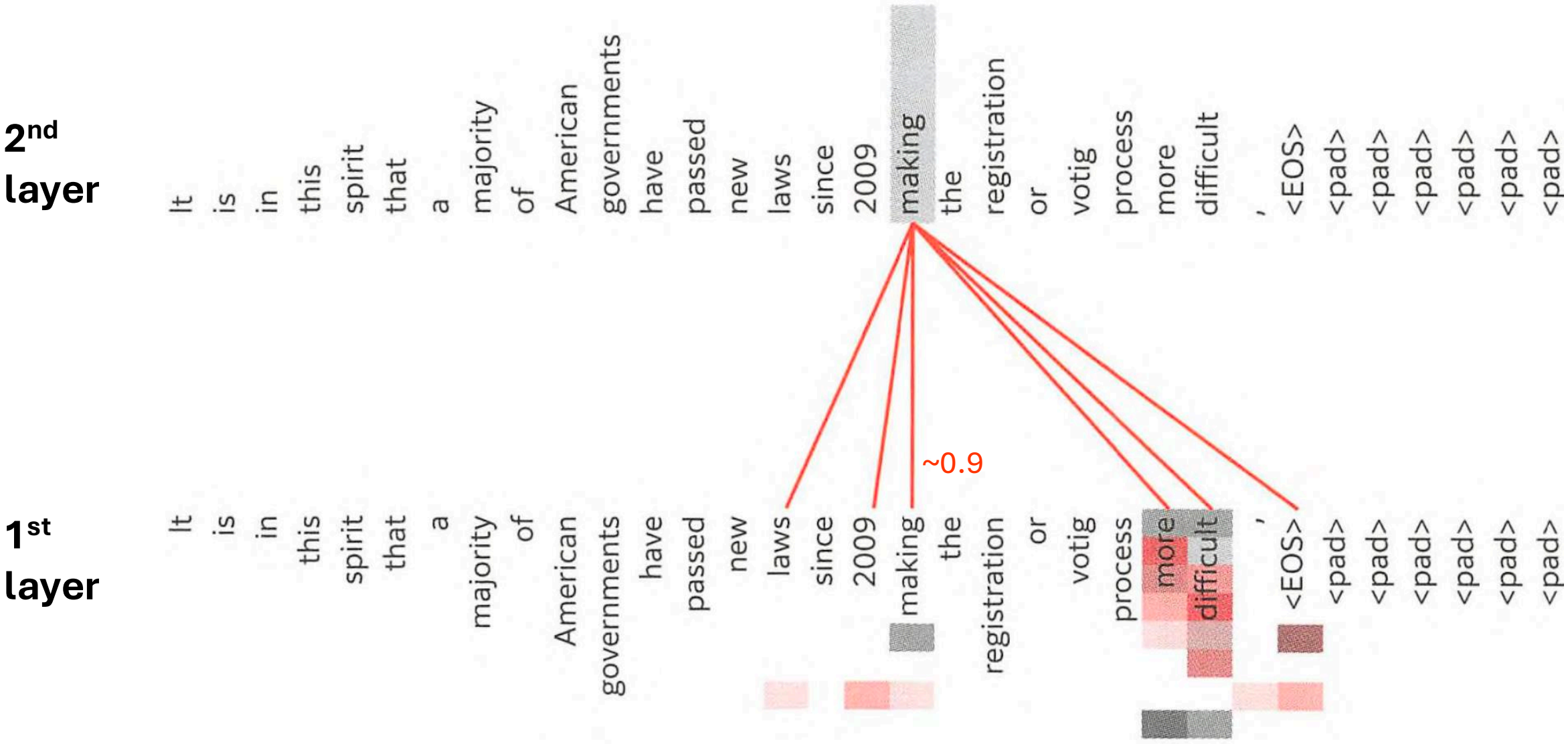


Number of words (tokens) in the input and output are equal !

If you can devise training data, you can chat and solve written problems.
→ It's all just a matter of outstanding accuracy in next-word prediction

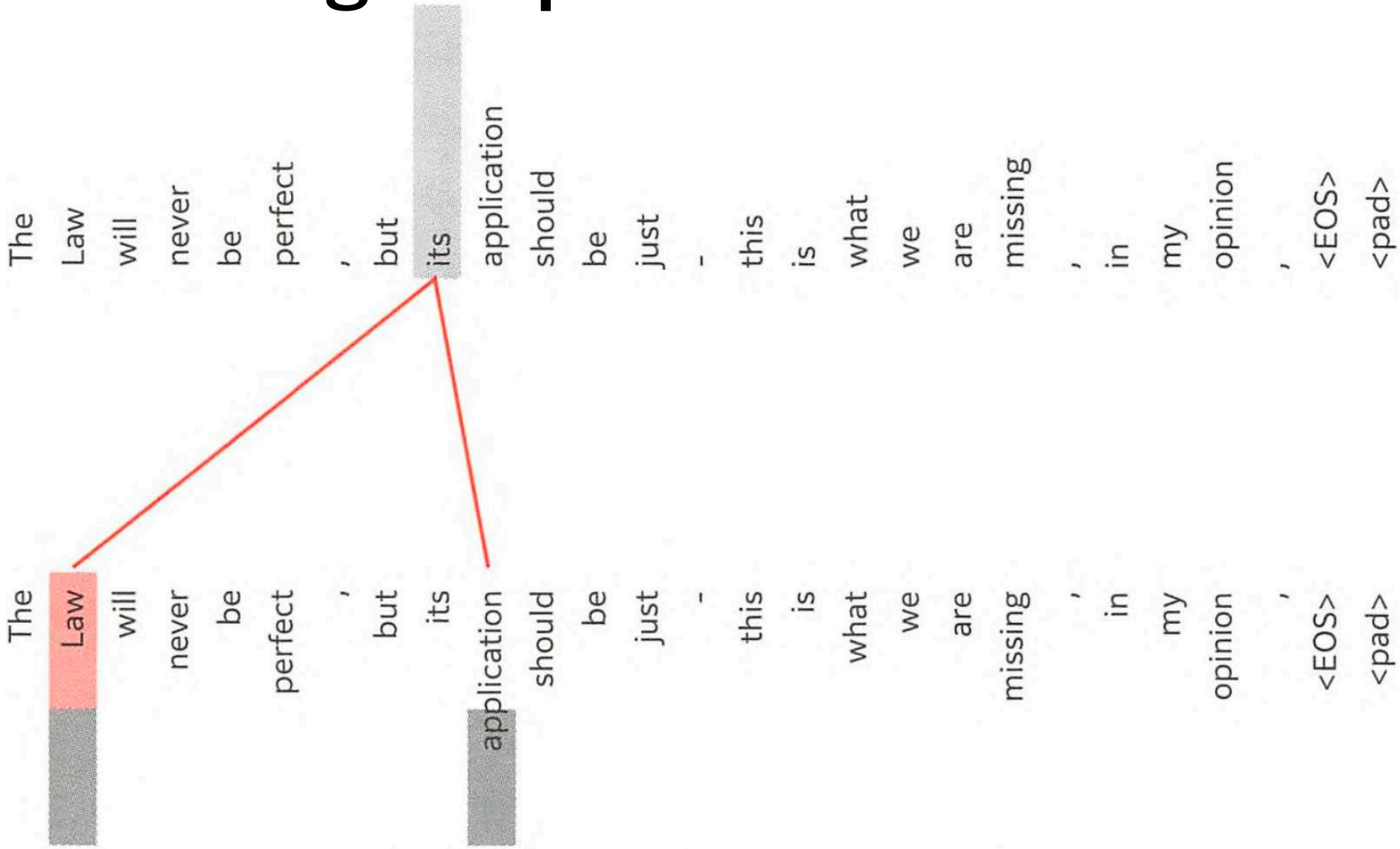
(Understanding Deep Learning, Prince, 2023)

Understanding phrases with attention



word = 600-dim vector → Better representation in next layer

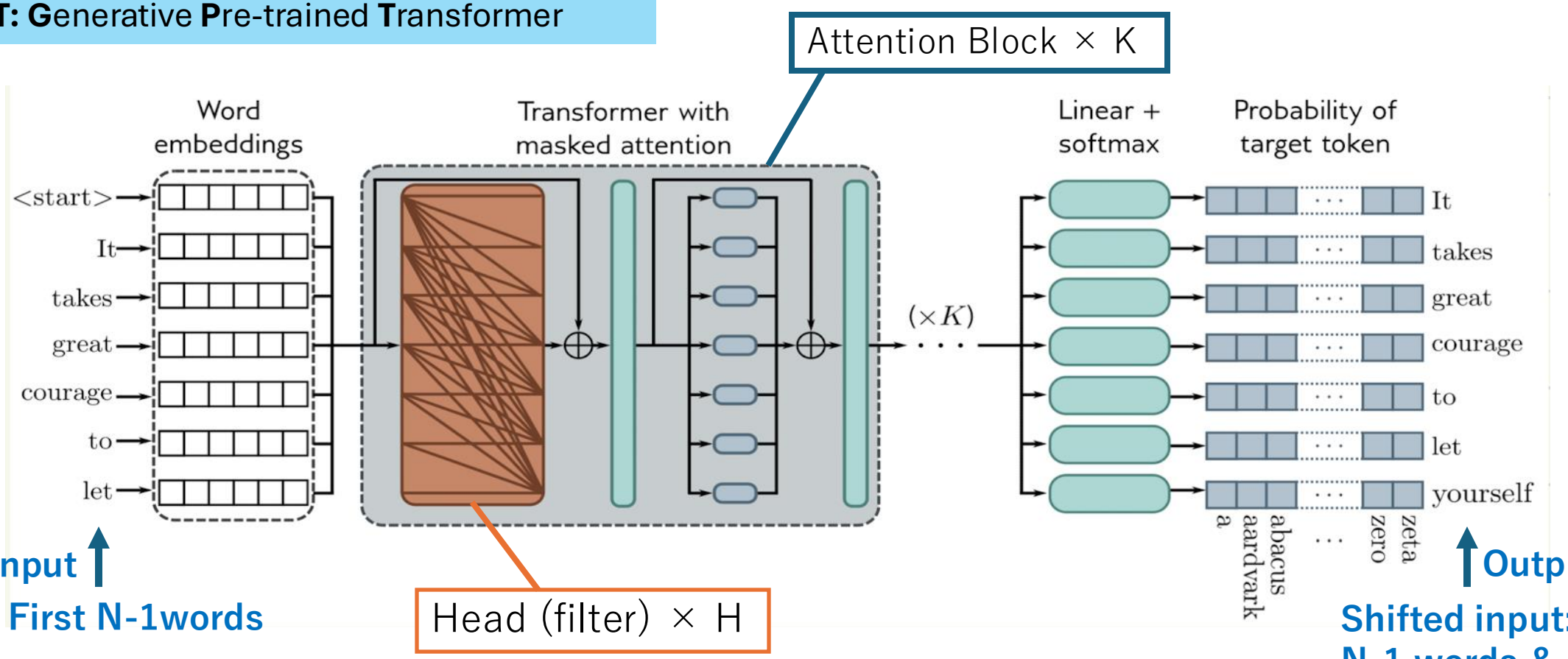
Understanding anaphora with attention



word = 600-dim vector → Better representation in next layer

Exploring optimal structure (minGemma)

GPT: Generative **P**re-trained **T**ransformer

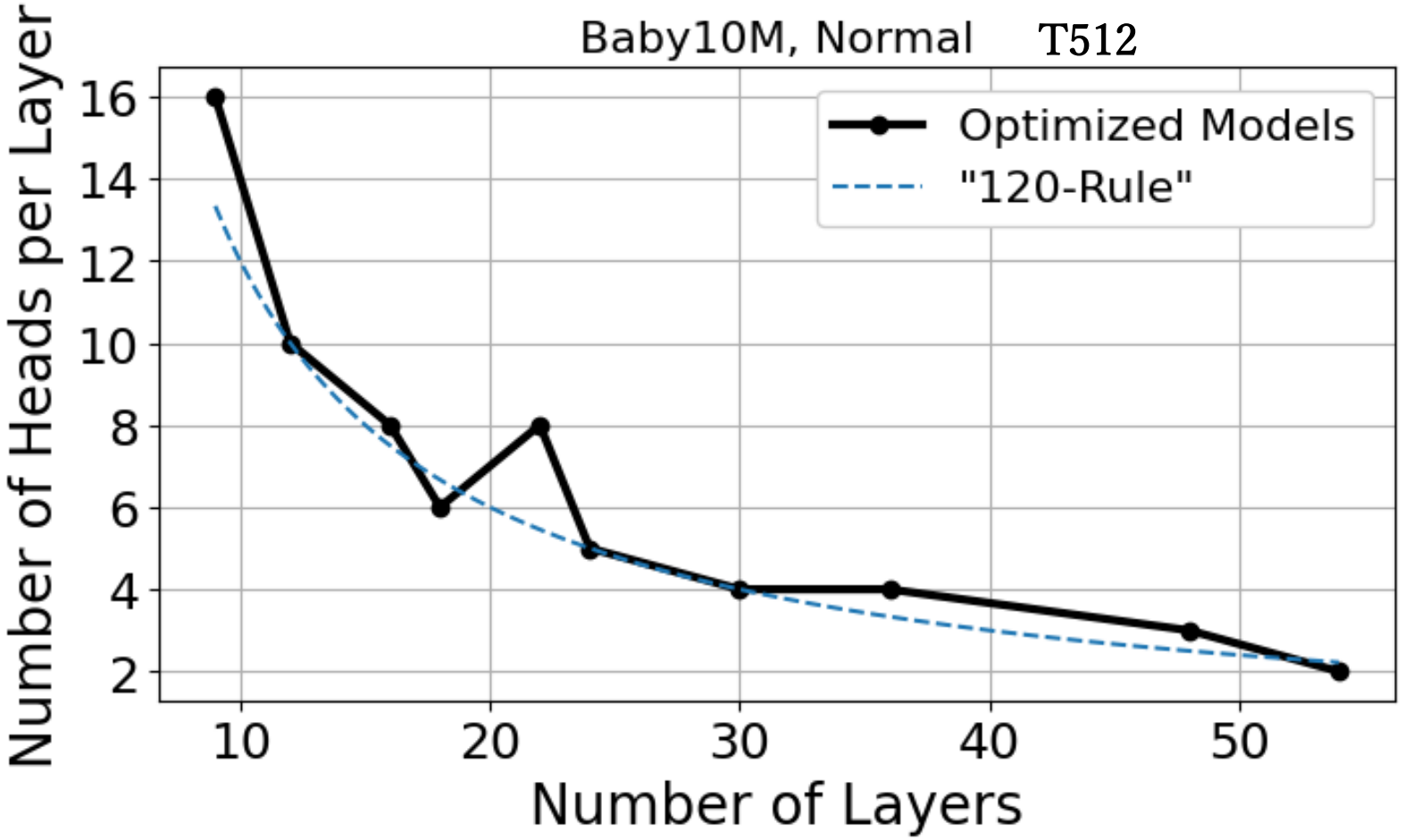


Number of words (tokens) in the input and output are equal !

If you can devise training data, you can chat and solve written problems.
→ It's all just a matter of outstanding accuracy in next-word prediction

(Understanding Deep Learning, Prince, 2023)

120-rule for #heads



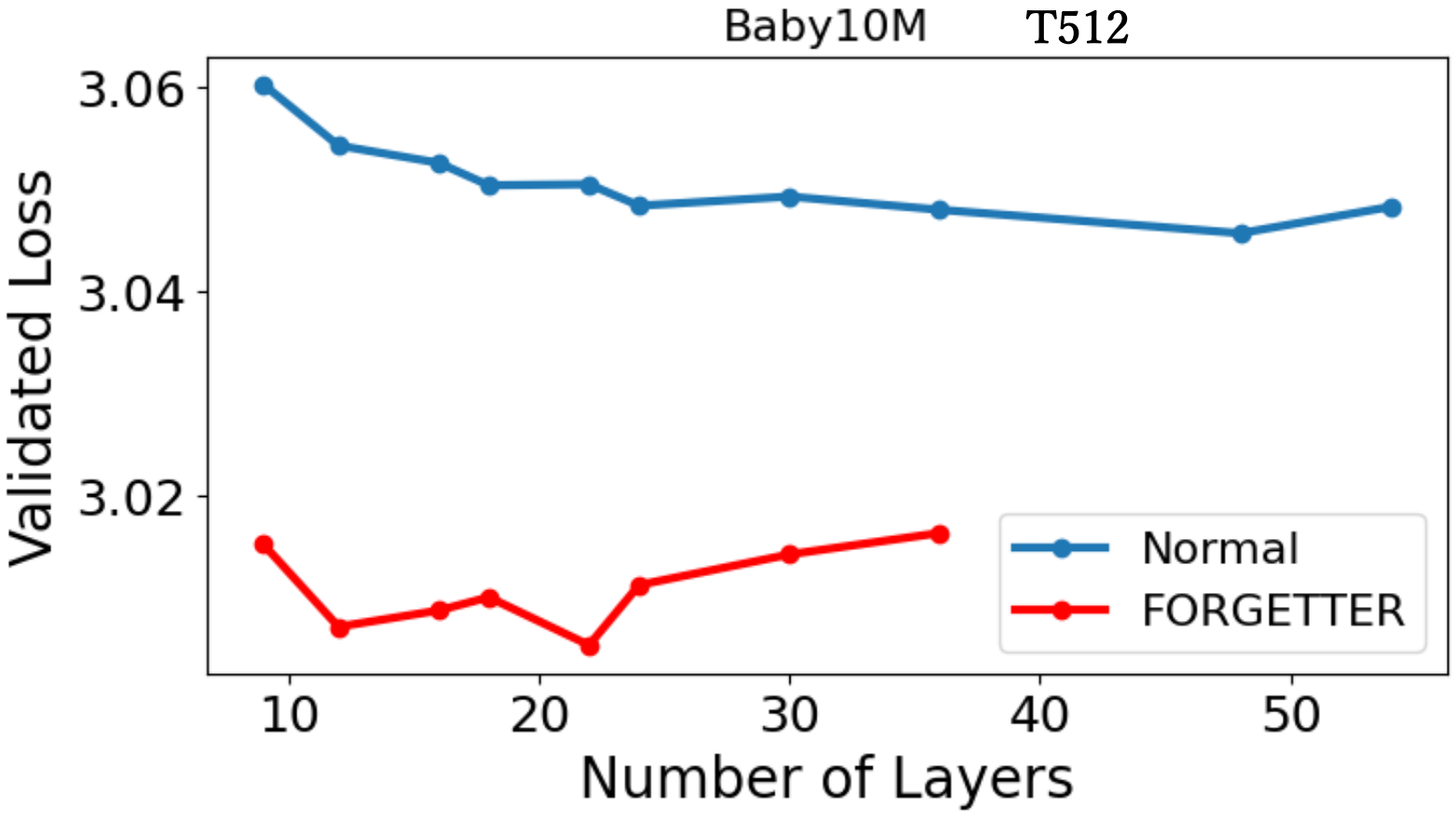
Models with about 120 heads outperformed! Ex) 5 heads x 24 layers

Methods: FORGETTER model

- Learning rates and state variables for optimization are initialized after every sleep
  - (Usually, the learning rate decreases linearly to zero over the training and the state variables for AdamW are never reset.)
- All the hyperparameters are set toward forgetting:

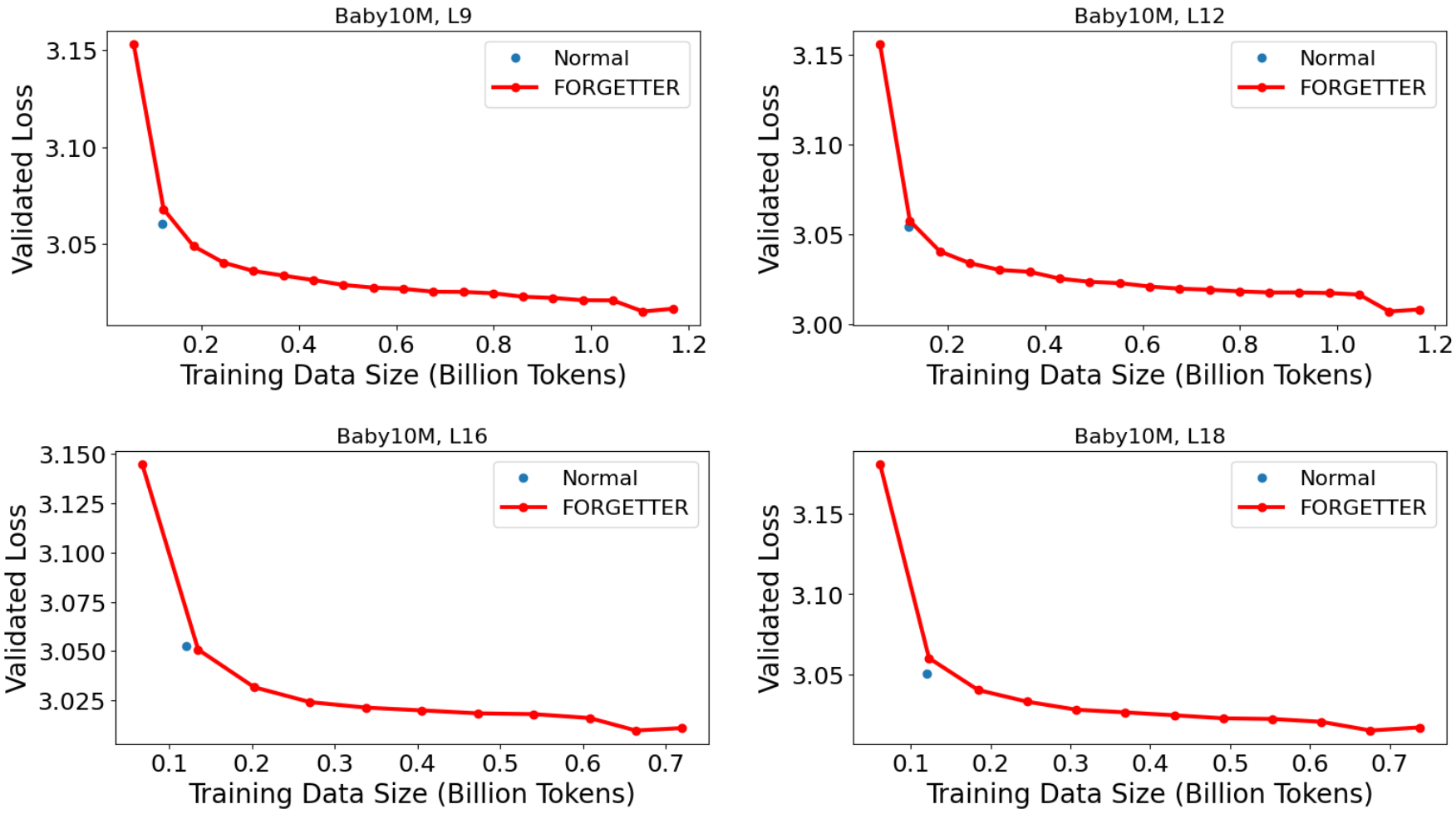
(BabyLM-10M)	Forgetter model	Normal model
Reset after sleep	**Yes**	No
Weight Decay (How fast memories fade)	high (1)	high (1)
Batch Size (How many past sentences to store)	Small (12)	Small (12)
Learning Rate (How fast to learn and forget)	high (10^{-3})	high (1.35×10^{-3})

Normal vs FORGETTER learning



FORGETTER is much better than the normal model

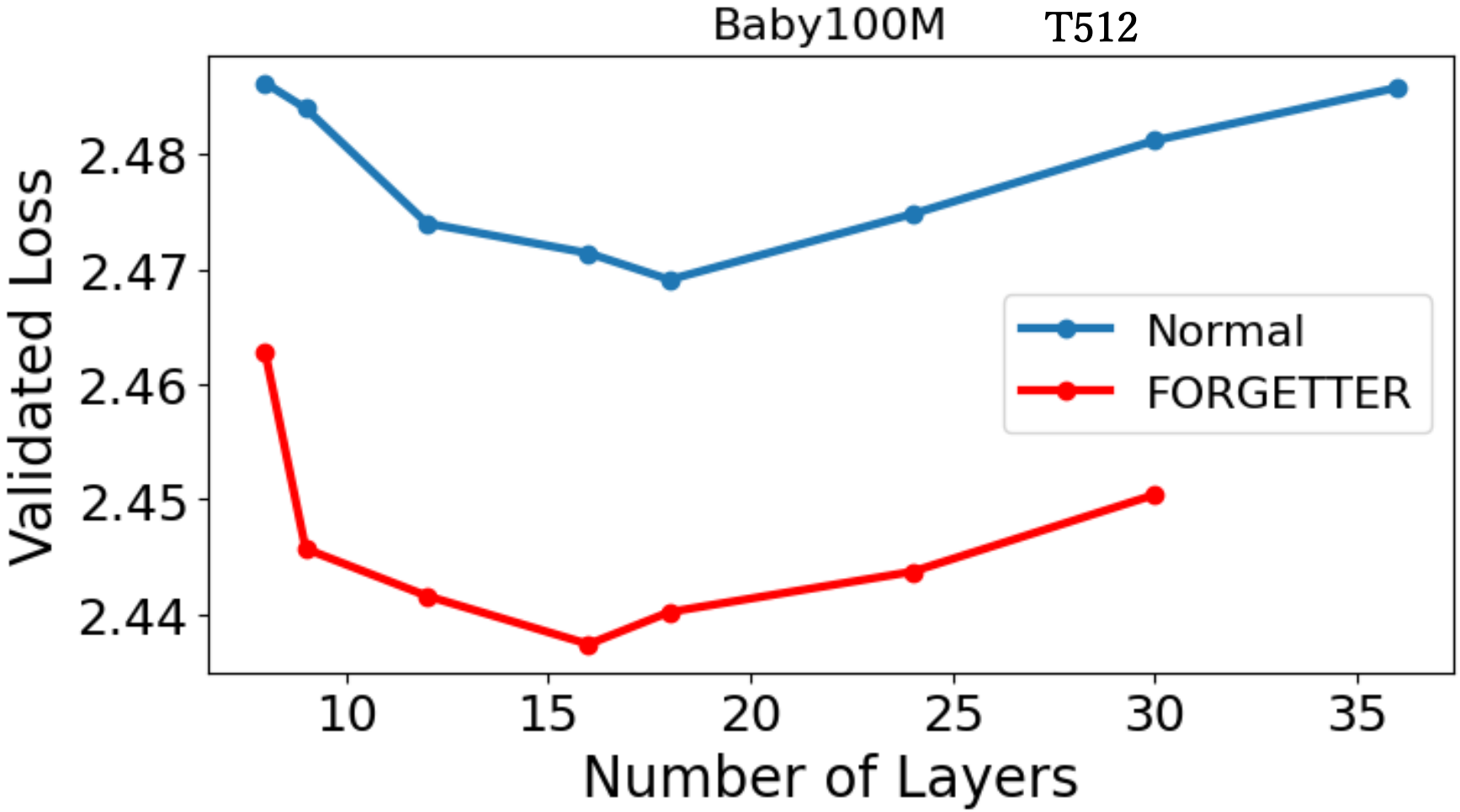
Examples of FORGETTER learning



T512

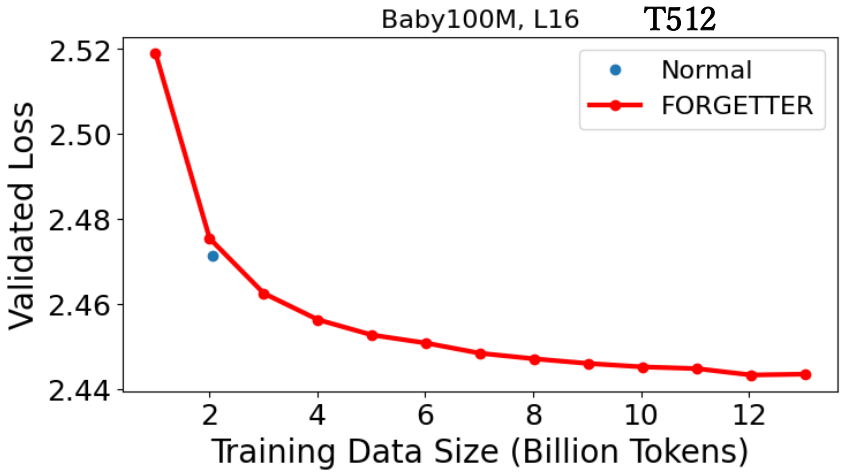
FORGETTER can learn beyond the normal over fitting limit

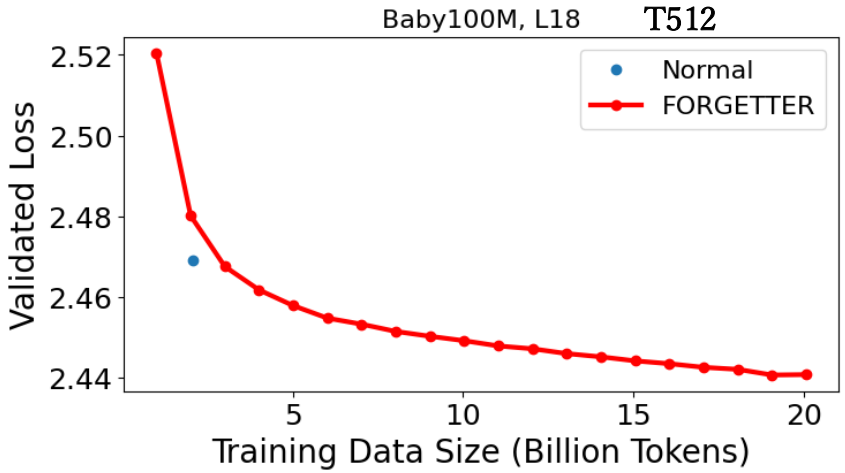
Normal vs FORGETTER learning: 100M dataset

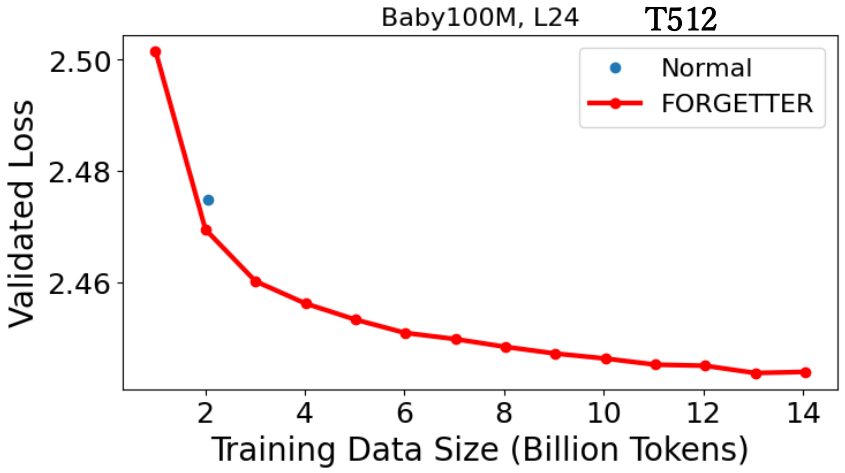


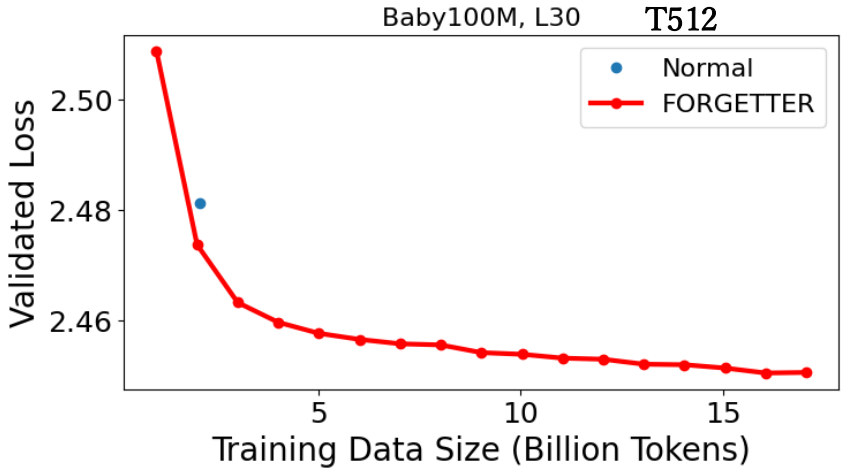
FORGETTER is much better than the normal model

Examples of FORGETTER: 100M dataset



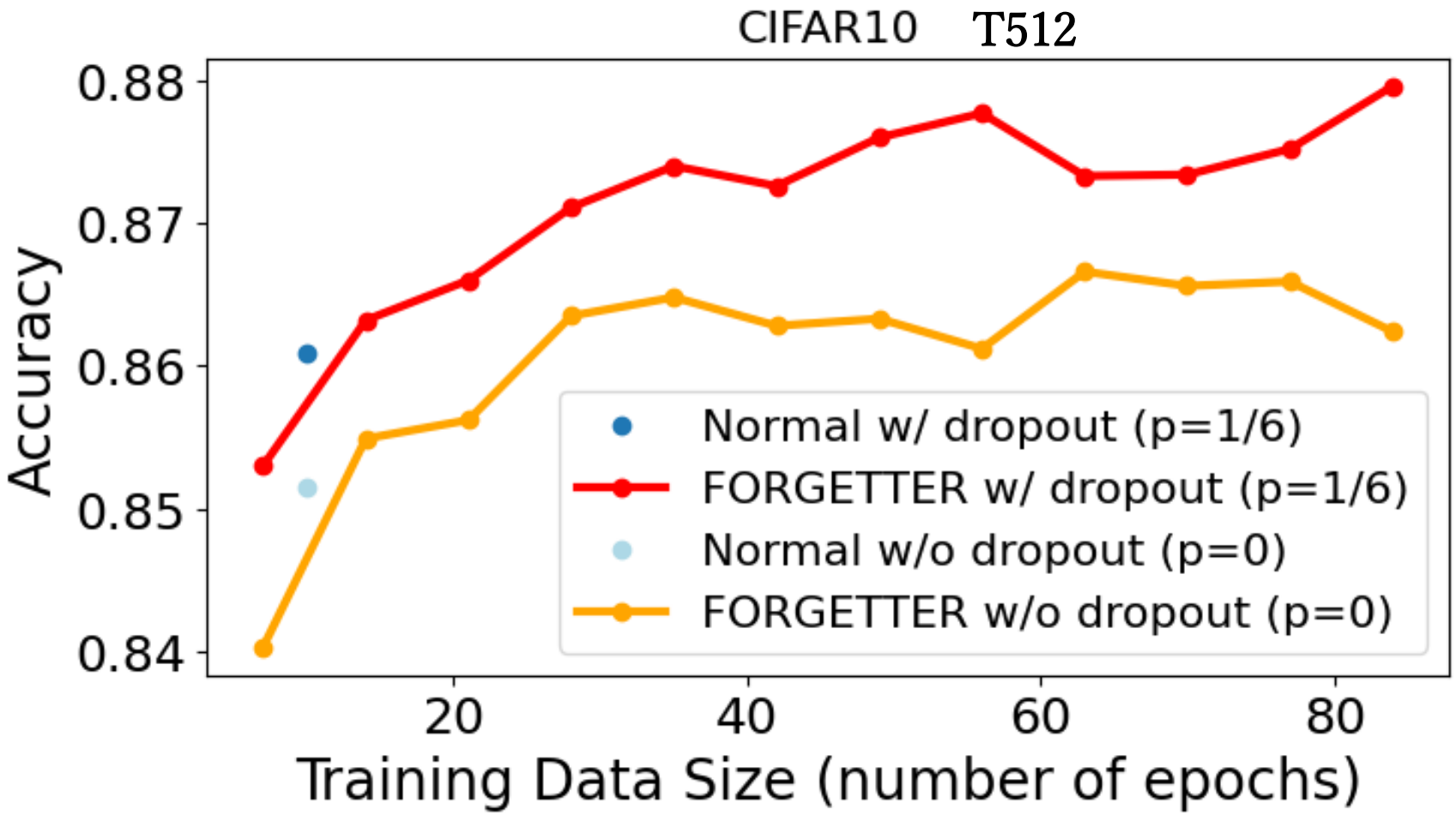






FORGETTER can learn beyond the normal over fitting limit

FORGETTER outperforms even for image classification



FORGETTER is much better than the normal model

Summary: Forgetting is beneficial

- 120 rule: models with about 120 heads outperformed
- FORGETTER can continue to learn beyond the normal overfitting limit.
- The benefit of FORGETTER seems general. Applicable, at least, to BabyLM-10M, BabyLM-100M, WikiText103, WikiText02, CIFAR10.