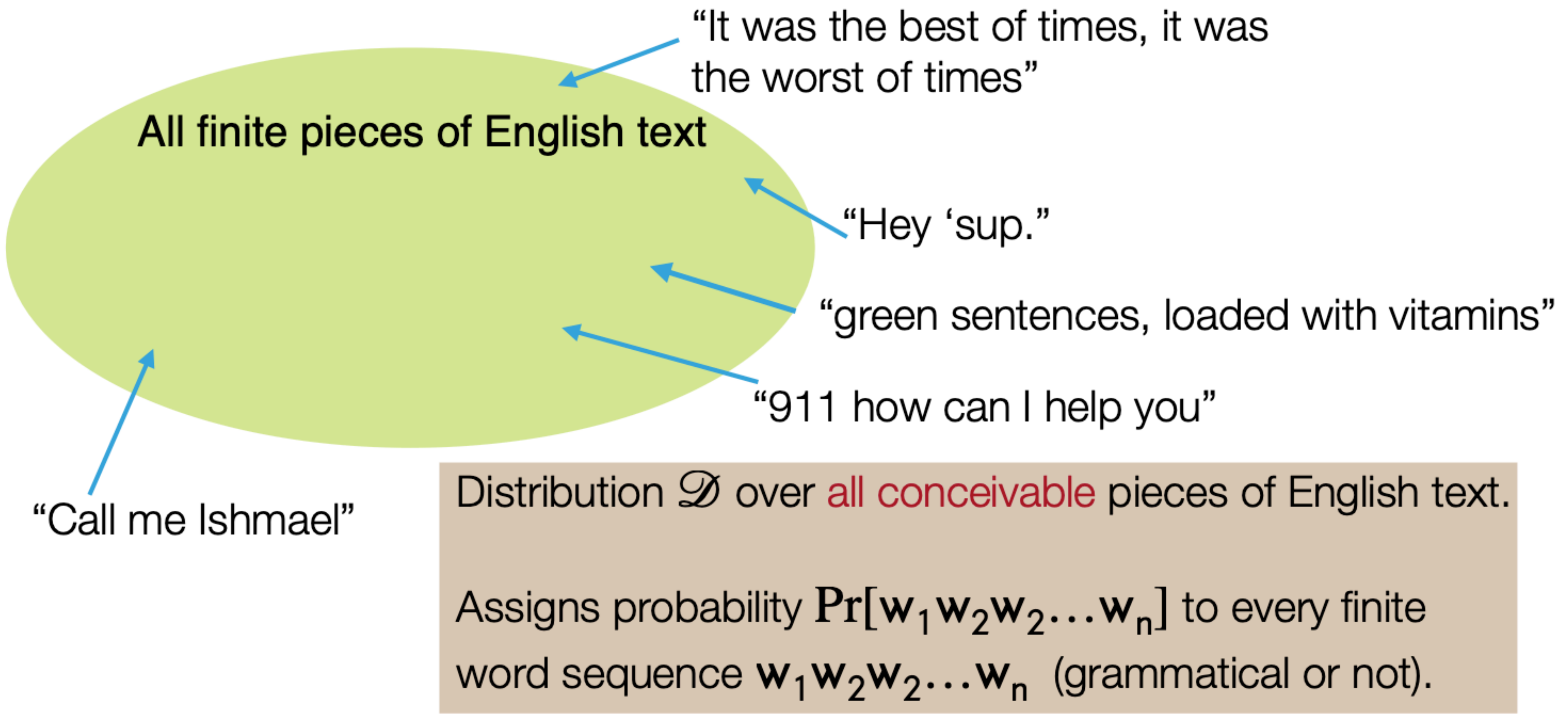
What are large language models (LLMs)?

Language models: narrow sense

- A probabilistic model that assigns a probability $P[w_1, w_2, \ldots, w_n]$ to every finite sequence $w_1, \ldots, w_n$ (grammatical or not)


All finite pieces of English text
"It was the best of times, it was the worst of times"
"Hey 'sup."
"green sentences, loaded with vitamins"
"911 how can I help you"
"Call me Ishmael"

Distribution $\mathscr{D}$ over all conceivable pieces of English text.

Assigns probability $\mathrm{Pr}[w_1 w_2 w_2 \ldots w_n]$ to every finite word sequence $w_1 w_2 w_2 \ldots w_n$ (grammatical or not).

Source: COS 324

Language models: narrow sense

Conditional probability

$$p(w_1, w_2, w_3, \ldots, w_N) =$$
$$p(w_1)\ p(w_2|w_1)\ p(w_3|w_1, w_2) \times \ldots \times p(w_N|w_1, w_2, \ldots w_{N-1})$$

Sentence: “the cat sat on the mat”

$$\begin{aligned} P(\text{the cat sat on the mat}) = {} & P(\text{the}) * P(\text{cat}|\text{the}) * P(\text{sat}|\text{the cat}) \\ & * P(\text{on}|\text{the cat sat}) * P(\text{the}|\text{the cat sat on}) \\ & * P(\text{mat}|\text{the cat sat on the}) \end{aligned}$$

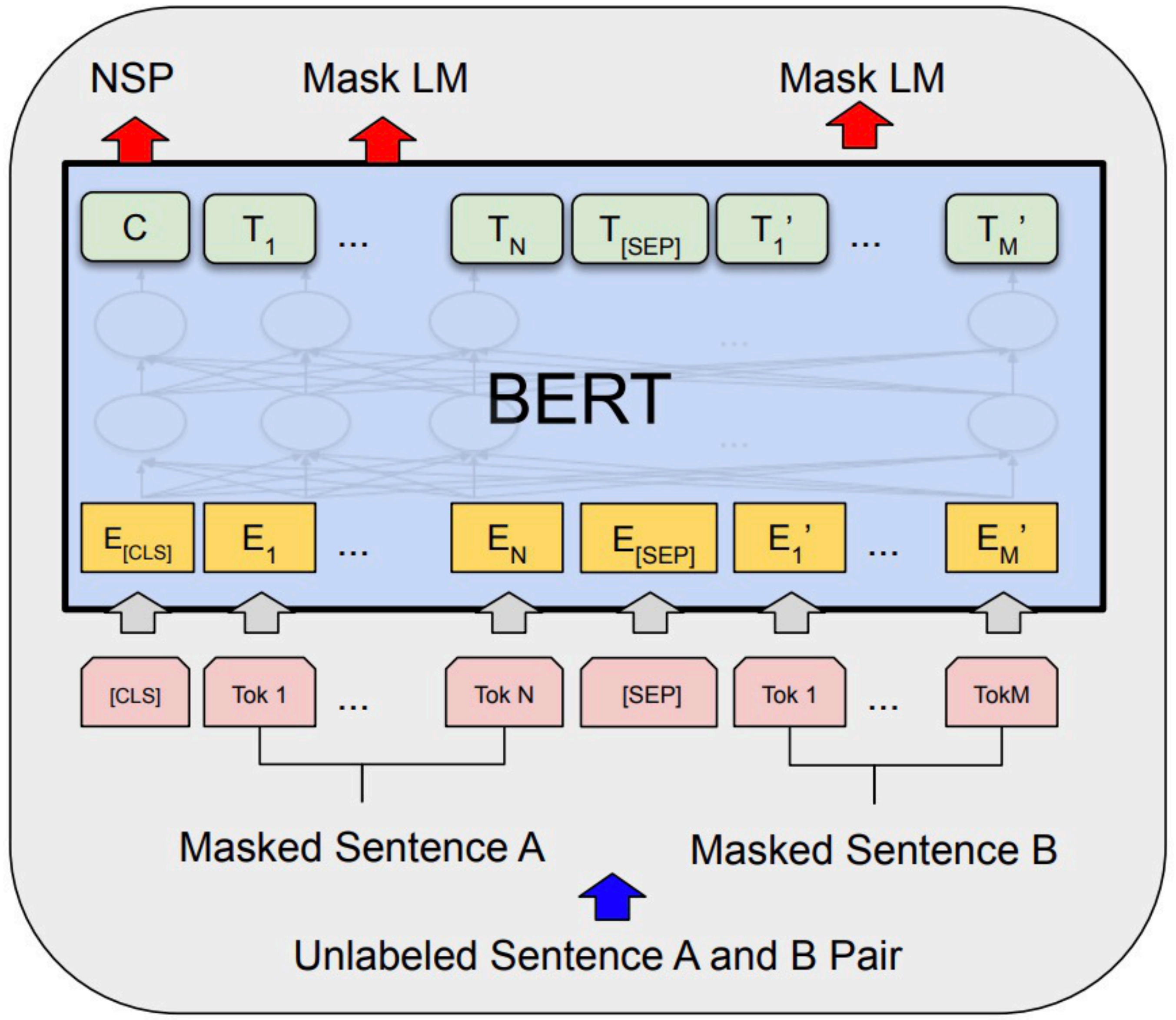
Implicit order

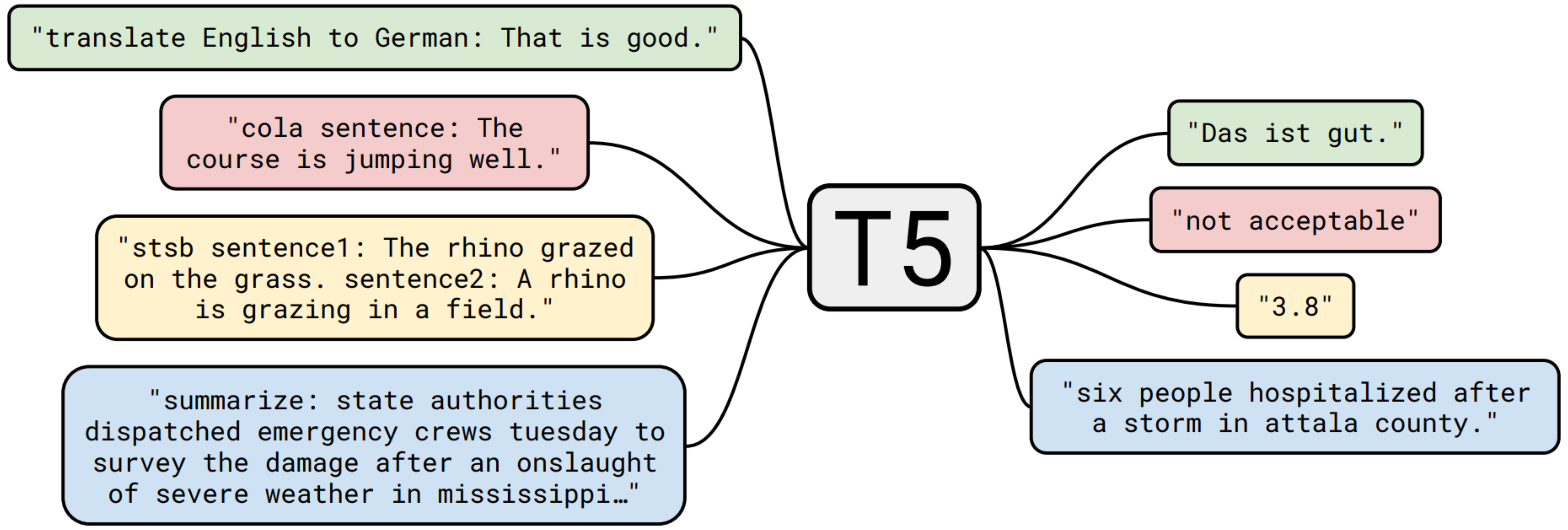
Source: COS 484

GPT-3 still acts in this way but the model is implemented as a very large neural network of 175-billion parameters!

Language models: broad sense

- Decoder-only models (GPT-x models)
- Encoder-only models (BERT, RoBERTa, ELECTRA)
- Encoder-decoder models (T5, BART)




"translate English to German: That is good."

"cola sentence: The course is jumping well."

"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."

"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"

T5

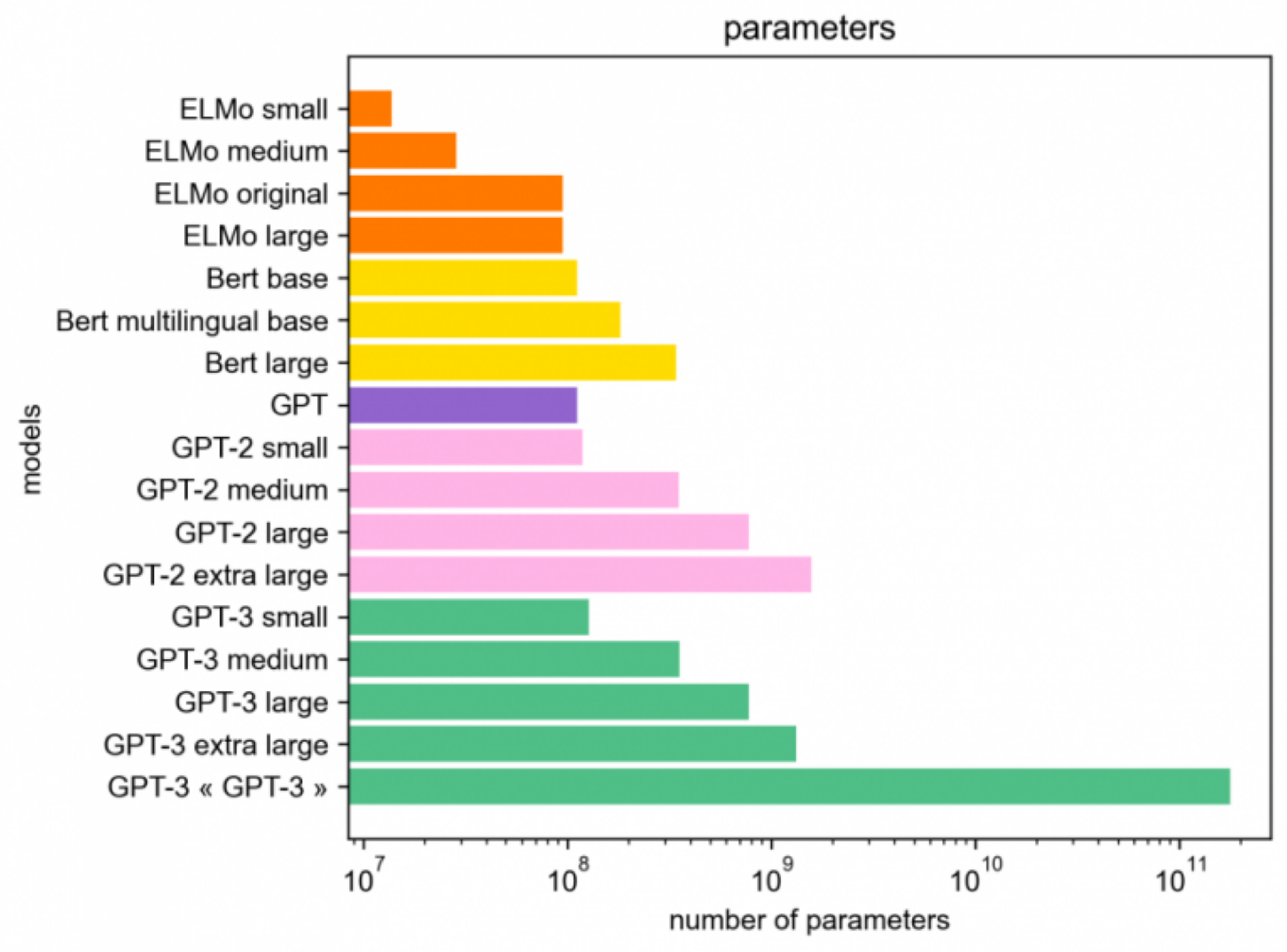
"Das ist gut."

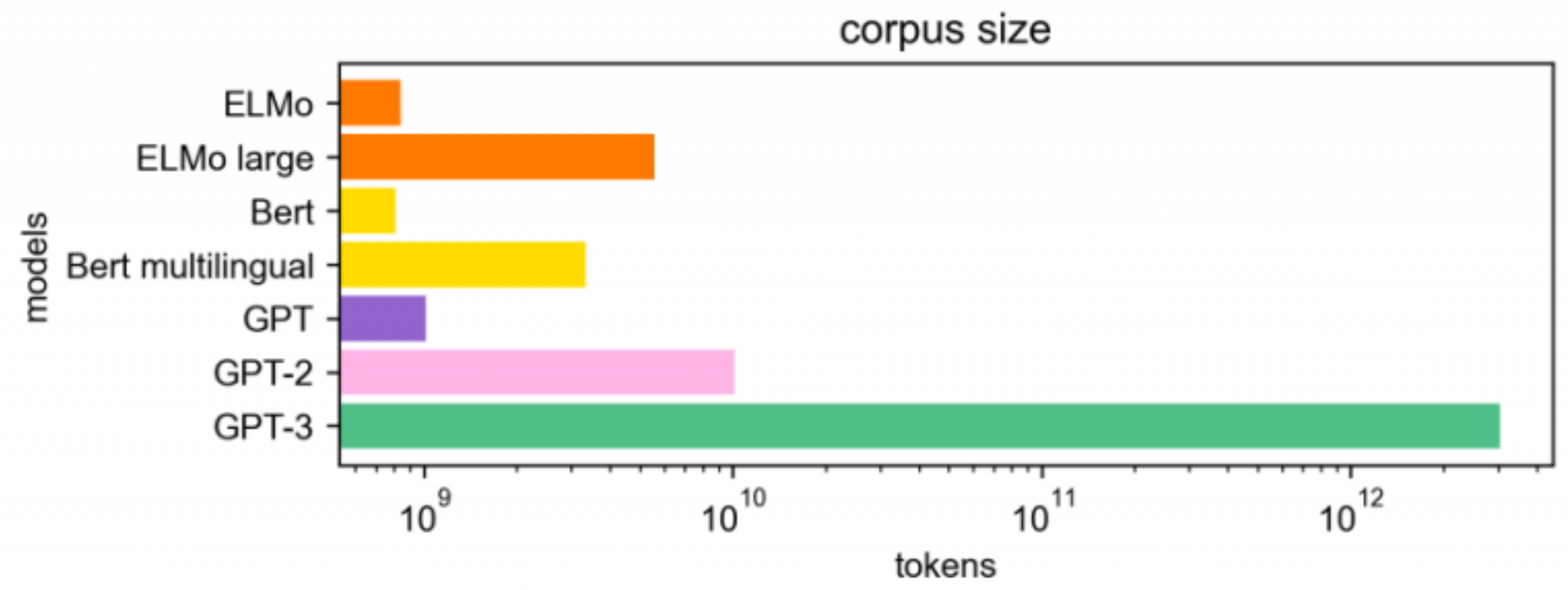
"not acceptable"

"3.8"

"six people hospitalized after a storm in attala county."

How large are “large” LMs?





More recent models: PaLM (540B), OPT (175B), BLOOM (176B)…

Image source: https://hellofuture.orange.com/en/the-gpt-3-language-model-revolution-or-evolution/

How large are “large” LMs?

- Today, we mostly talk about two camps of models:
    - Medium-sized models: BERT/RoBERTa models (100M or 300M), T5 models (220M, 770M, 3B)
    - “Very” large LMs: models of 100+ billion parameters

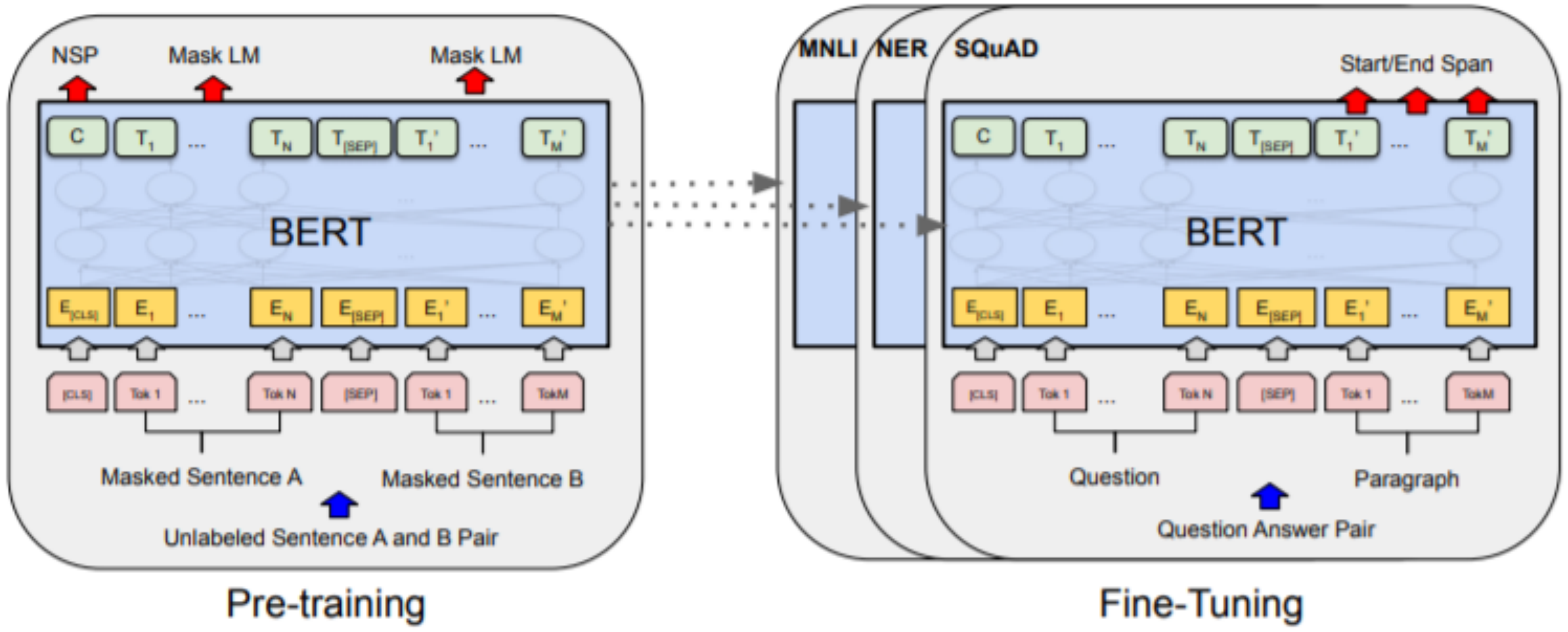
- Larger model sizes ⇒ larger compute, more expensive during inference
- Different sizes of LMs have different ways to adapt and use them
    - Fine-tuning, zero-shot/few-shot prompting, in-context learning…
- Emergent properties arise from model scale
- Trade-off between model size and corpus size

Q: Do largest models always give the best performance today?

Pre-training and adaptation

- **Pre-training**: trained on huge amounts of unlabeled text using “self-supervised” training objectives

- **Adaptation**: how to use a pre-trained model for your downstream task?
    - What types of NLP tasks (input and output formats)?
    - How many annotated examples do you have?



Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // ________

LM

Circulation revenue has increased by 5% in Finland. // Finance

They defeated … in the NFC Championship Game. // Sports

Apple … development of in-house chips. // Tech

The company anticipated its operating profit to improve. // ________

LM

http://ai.stanford.edu/blog/understanding-incontext/

Why LLMs?

- The promise: one single model to solve many NLP tasks

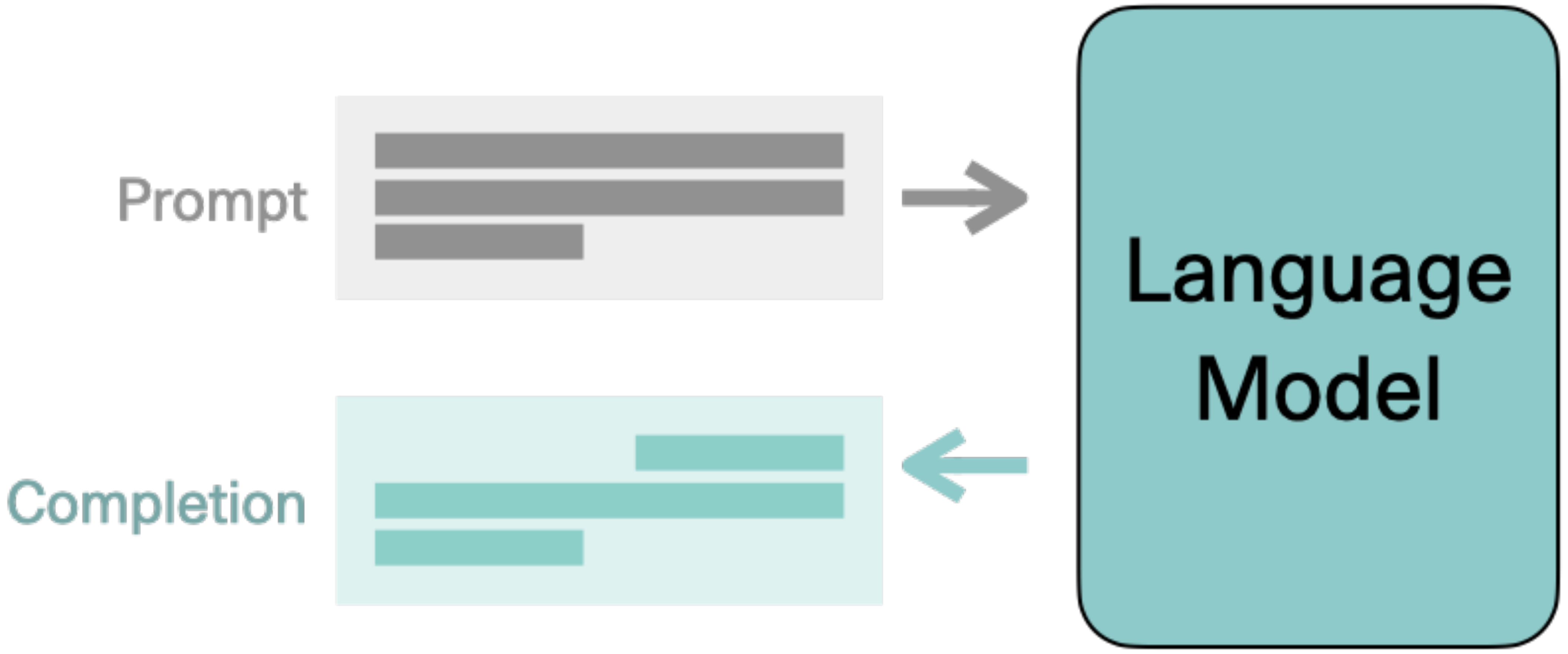
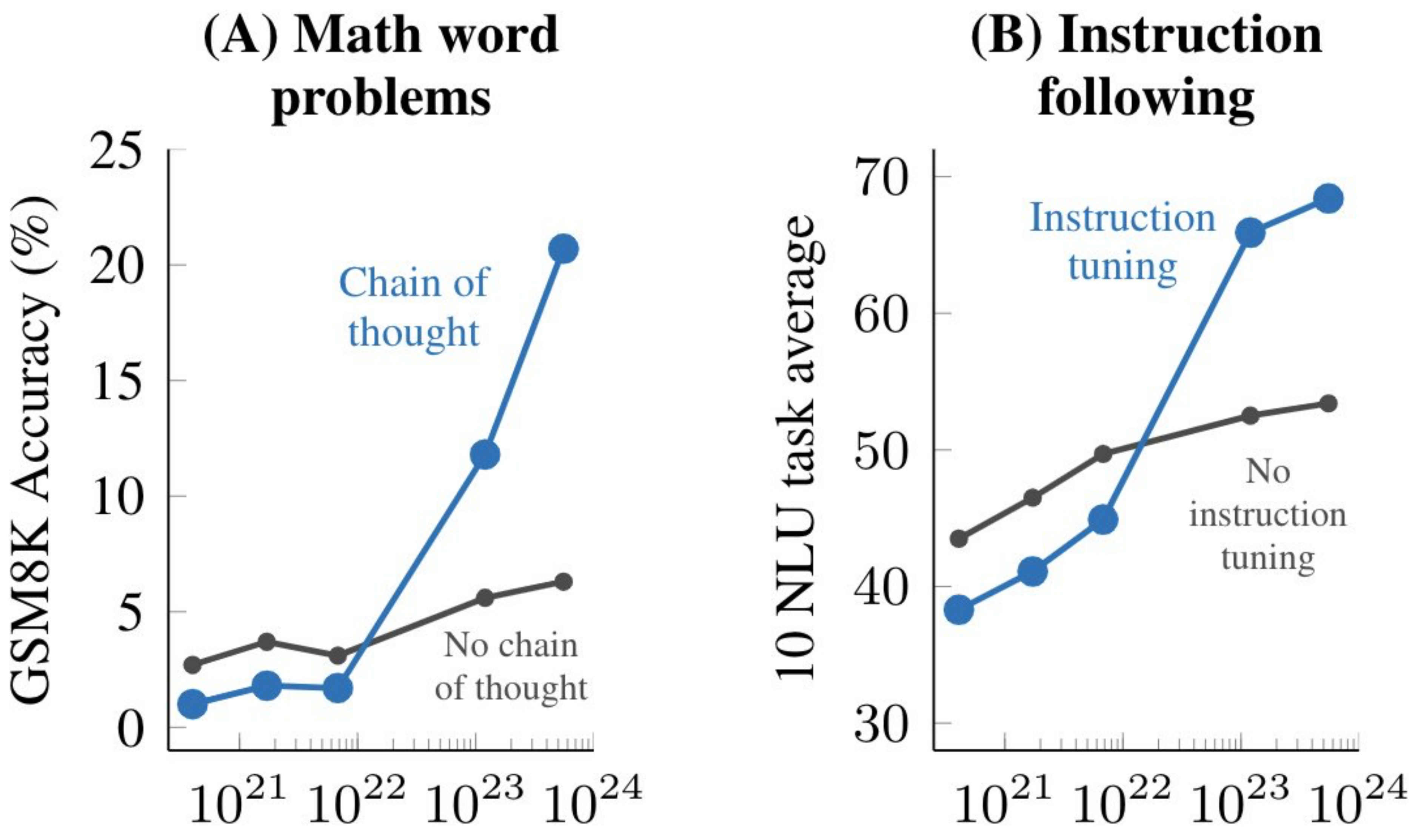


Image credit: Jay Alammar

- Emergent properties in LLMs



(Wei et al., 2022)