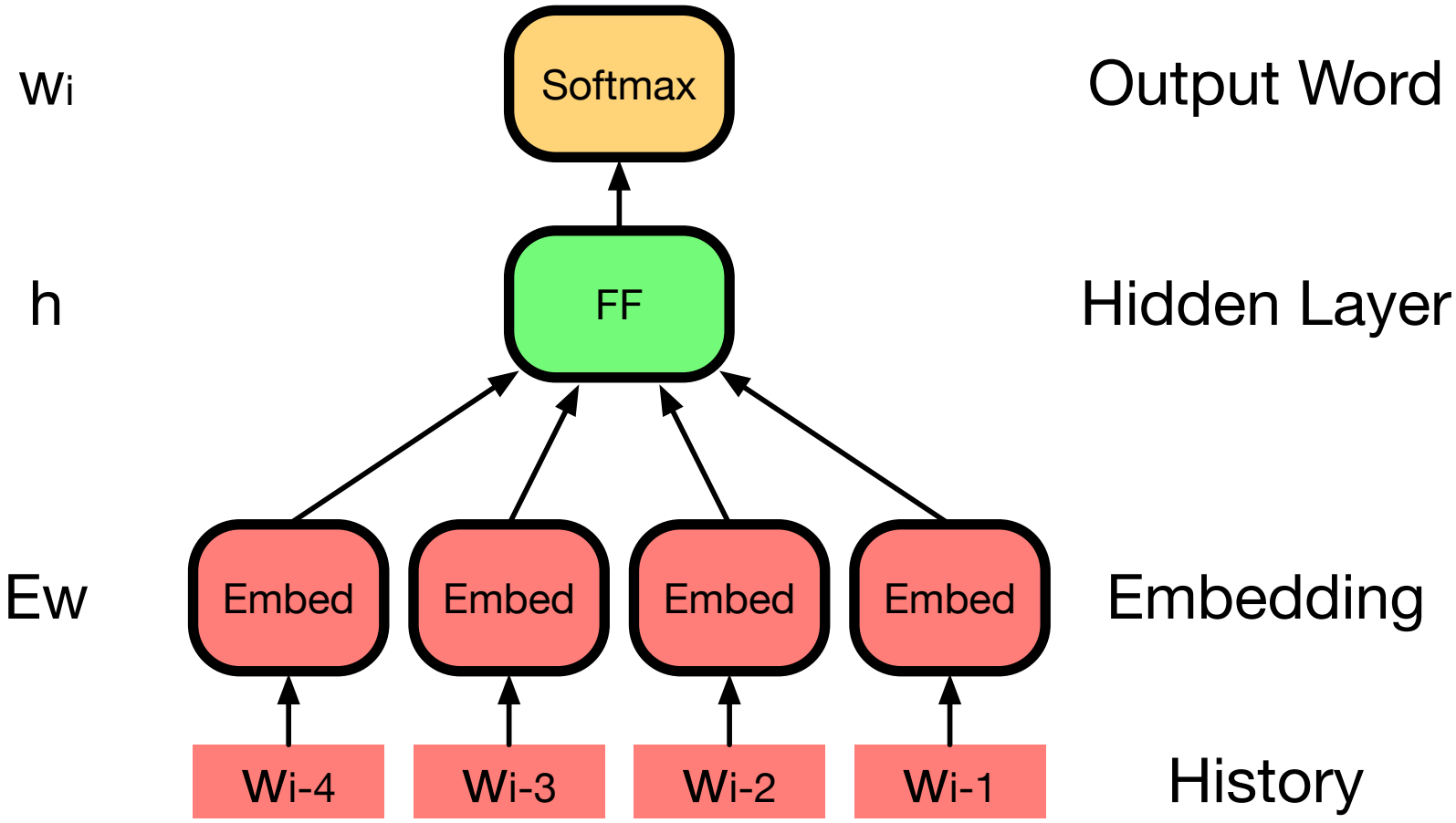
Large Language Models

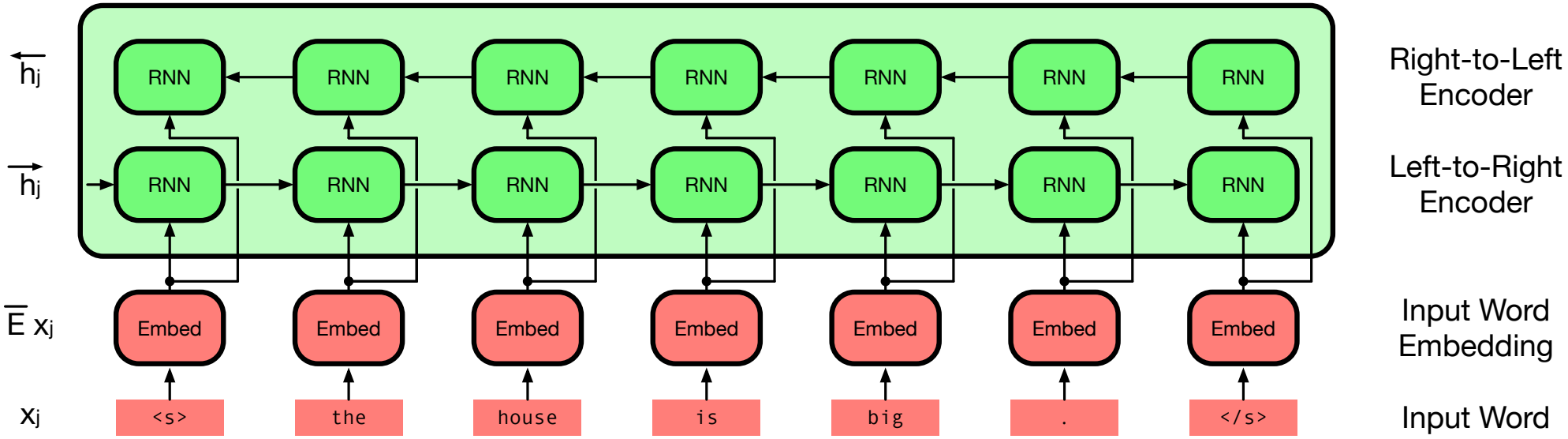
Philipp Koehn

23 September 2025

Word Embeddings

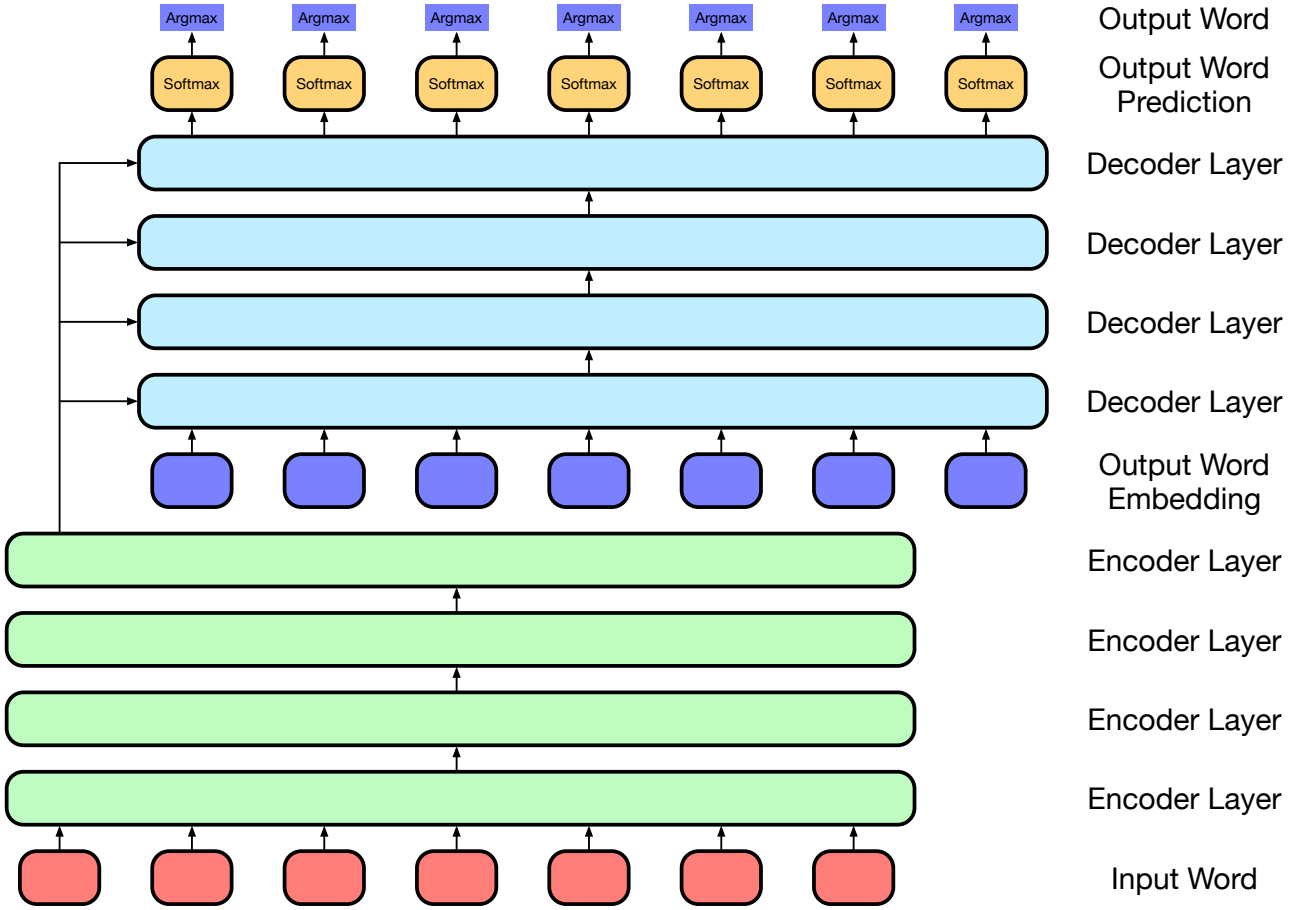


Contextualized Word Embeddings



- ELMo: Embeddings from Language Models (2018)
- Bidirectional LSTM

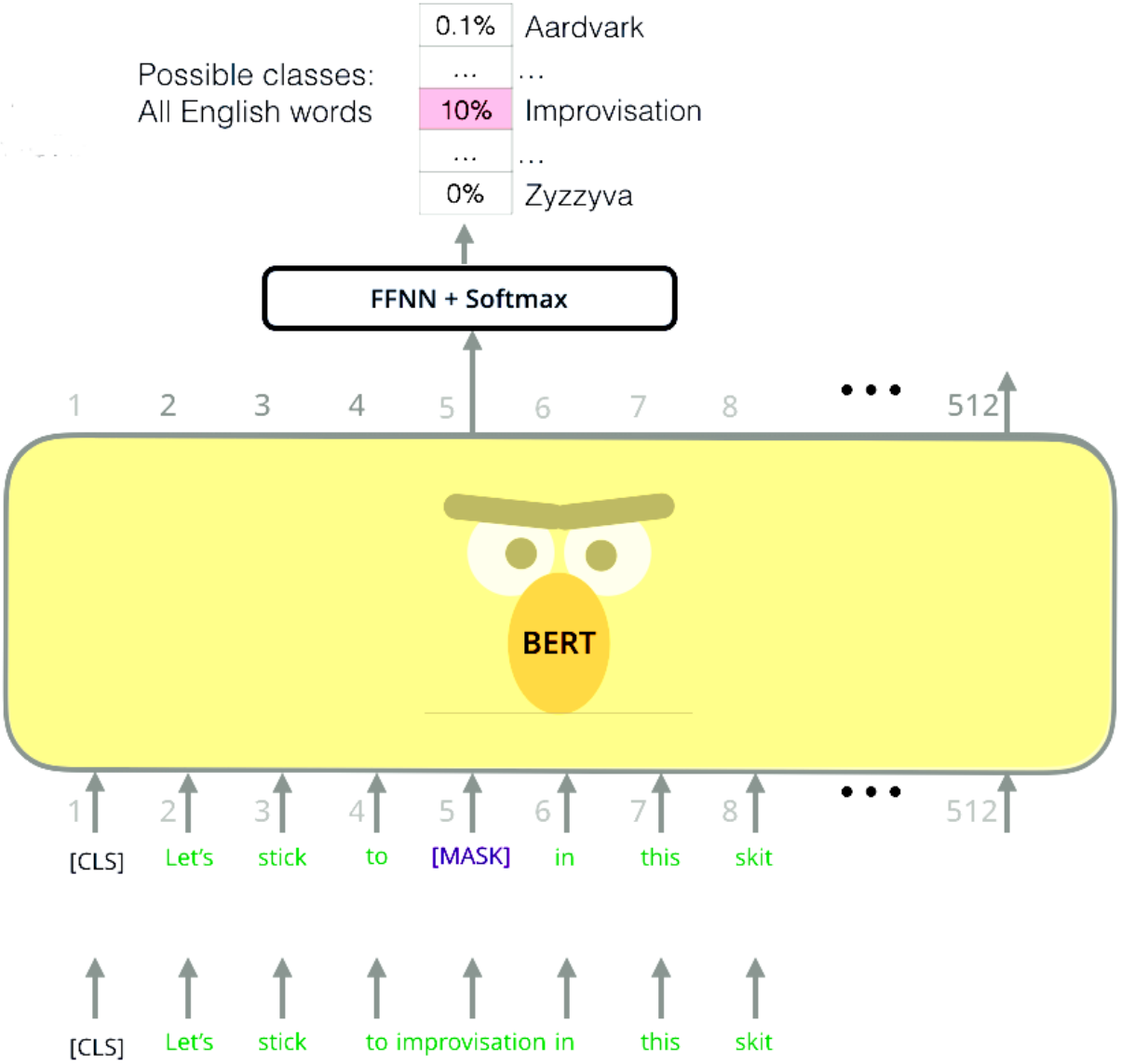
Contextualized Word Embeddings



- BERT: Bidirectional Encoder Representations from Transformers (2019)

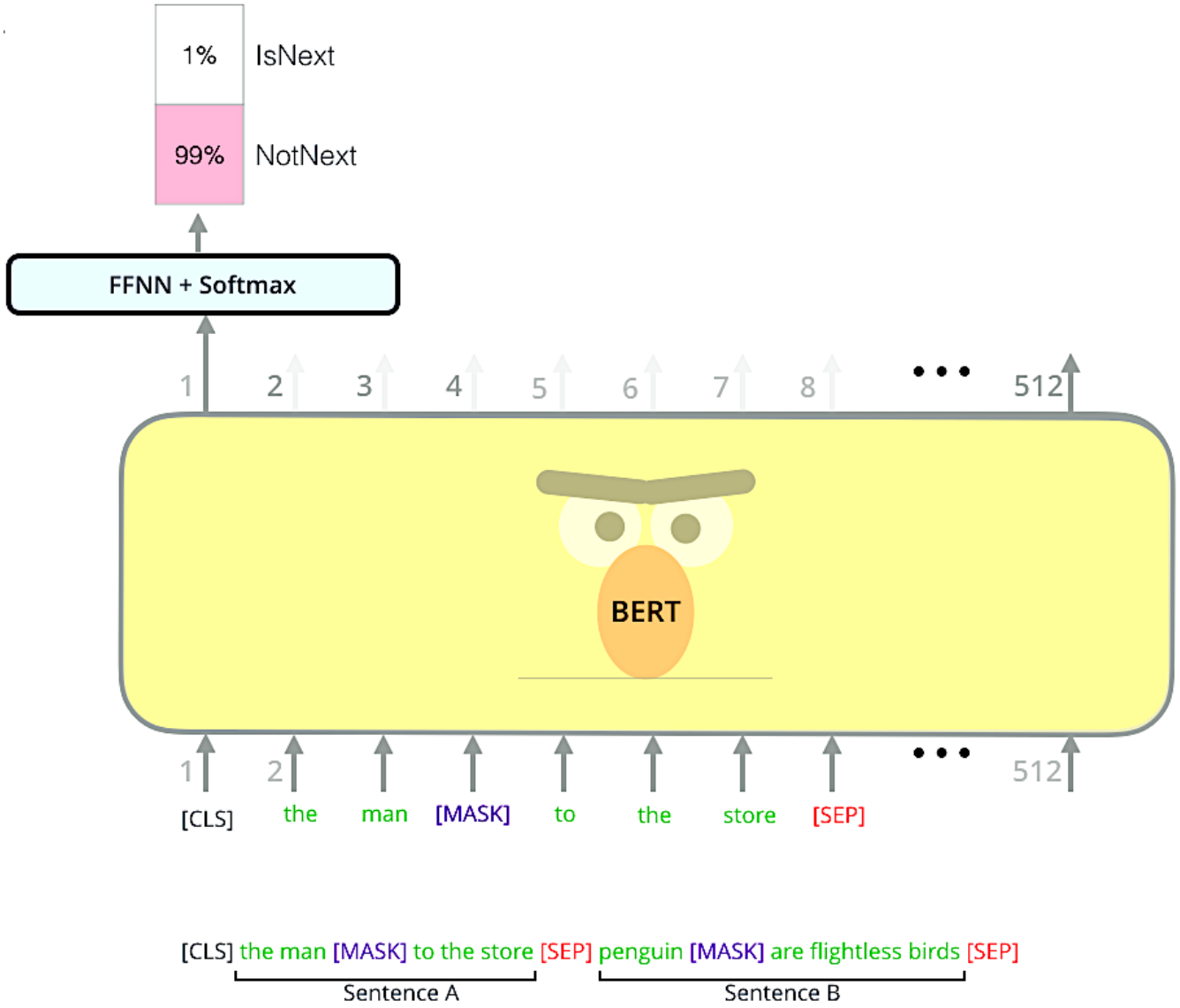
Masked Language Model Training

- Transformer expects an input and an output sequence
- Masked training
  - output sequence: one sentence of text
  - input sequence: same sentence, with some words masked out

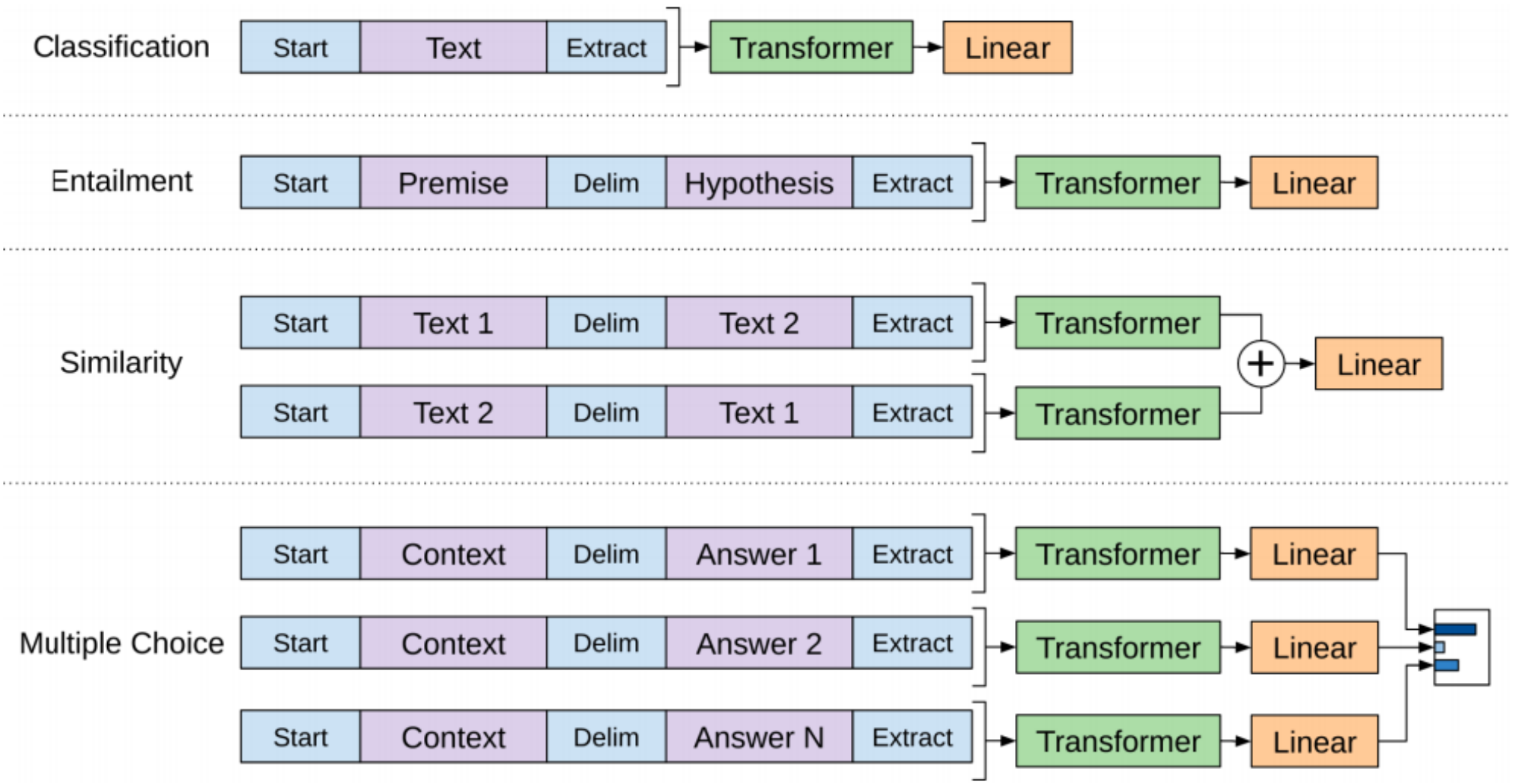


Next Sentence Prediction

- Next sentence prediction
- Input: two sentences
- Output: prediction that they were in sequence



Using Sentence Representations



LMs as Unsupervised Learners (2019)

- Train language models on relatively clean text data (GPT-2)
- Such text contains **naturally occurring demonstrations** of many tasks
- Convert any NLP problem into a text continuation problem

”I’m not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile** [**I’m not a fool**].

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: ”**Mentez mentez, il en restera toujours quelque chose**,” which translates as, ”**Lie lie and something will always remain.**”

”I hate the word ’**perfume**,”’ Burr says. ’It’s somewhat better in French: ’**parfum**.’

If listened carefully at 29:55, a conversation can be heard between two guys in French: ”**Comment on fait pour aller de l’autre cote? Quel autre coté?’** ”, which means ”**How do you get to the other side? What side?**”.

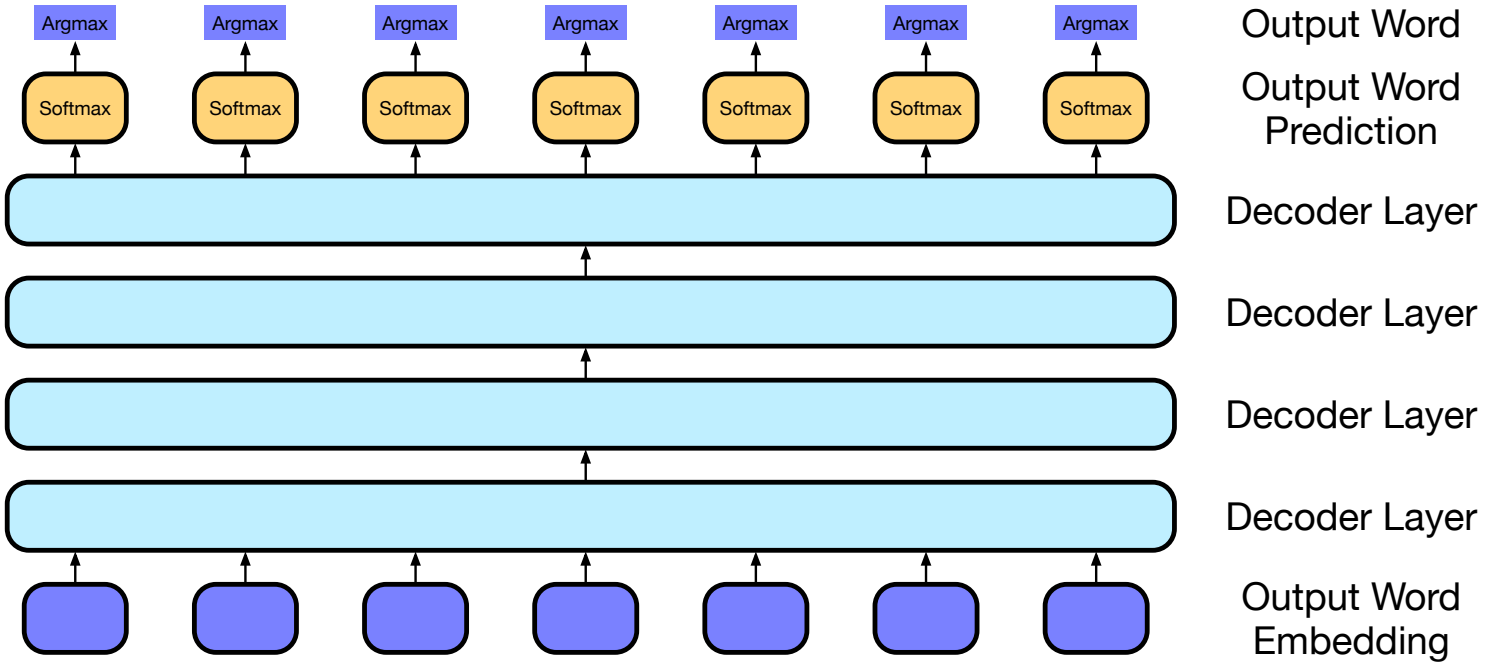
If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinema?** , or **Did you go to the movies?**, which literally translates as **Have-you to go to movies/theater?**

Decoder-Only Models

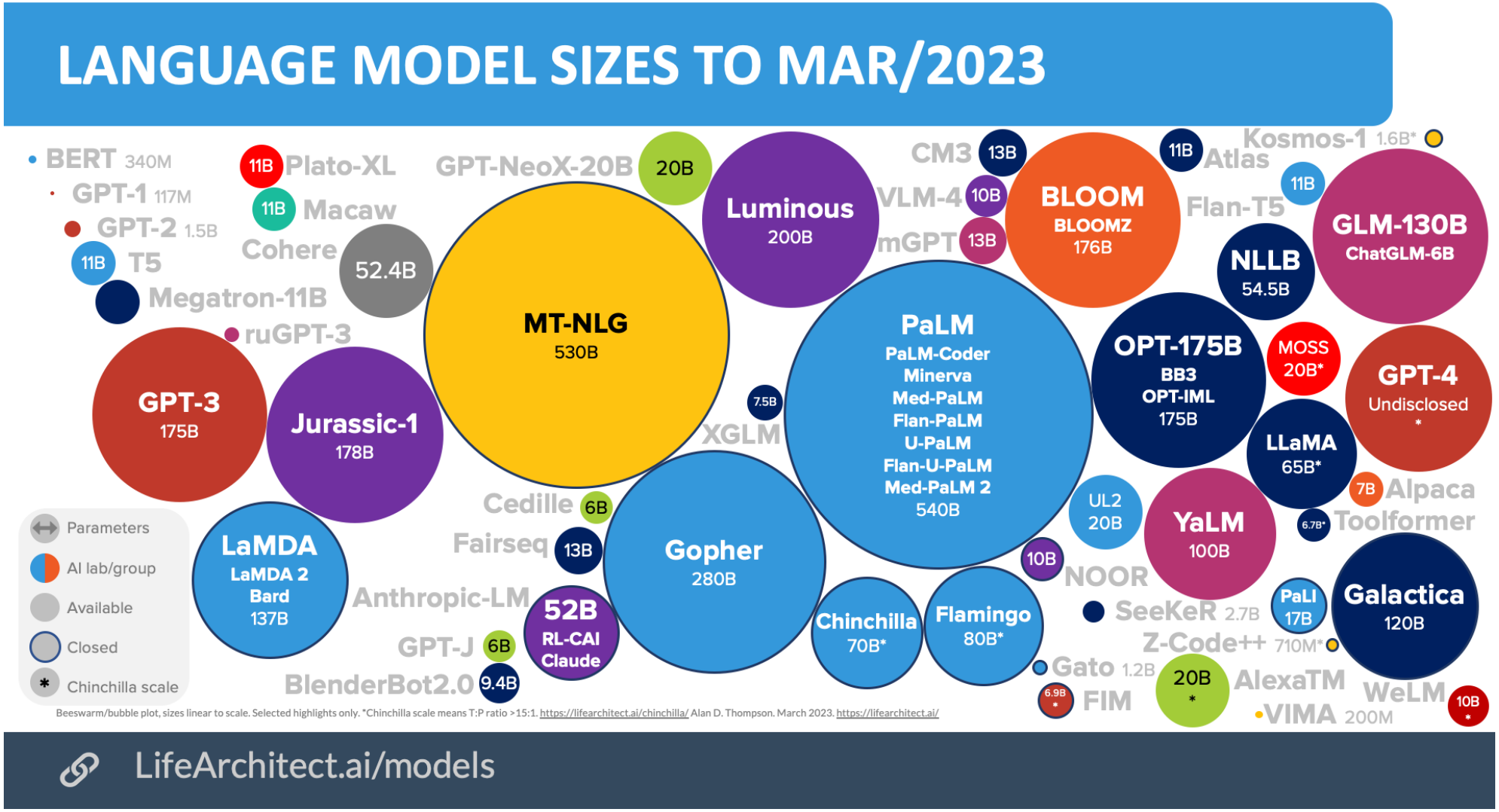
- Alternative architecture: Just decoder of Transformer model

$\Rightarrow$ no input, only self-attention

- Trained with next-word prediction



Size of Large Language Models



evaluation

SQuAD

- Reading comprehension
- Given: a short text, questions
- Expected answer:
  span of words in text
- SQuAD V2: Also added
  unanswerable questions

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Computer Code Generation


```
def solution(lst):
    """Given a non-empty list of integers, return the sum of all of the odd elements
    that are in even positions.

    Examples
    solution([5, 8, 7, 1]) ==>12
    solution([3, 3, 3, 3, 3]) ==>9
    solution([30, 13, 24, 321]) ==>0
    """
    return sum(lst[i] for i in range(0,len(lst)) if i % 2 == 0 and lst[i] % 2 == 1)
```

- Generation of computer code from a textual description of the task
- Example HumanEval [Chen et al., 2021]: Hand-written evaluation set
- Evaluation: run the code and see if the answers are correct (unit tests)
- Very similar test set: Mostly Basic Programming Problems (MBPP) [Austin et el., 2021]

Common Sense

Relation	Formulated question example
AtLocation	*Where would I not want a fox?* **A.** hen house, **B.** england, **C.** mountains, **D.** ...
Causes	*What is the hopeful result of going to see a play?* **A.** being entertained, **B.** meet, **C.** sit, **D.** ...
CapableOf	*Why would a person put flowers in a room with dirty gym socks?* **A.** smell good, **B.** many colors, **C.** continue to grow , **D.** ...
Antonym	*Someone who had a very bad flight might be given a trip in this to make up for it?* **A.** first class, **B.** reputable, **C.** propitious , **D.** ...
HasSubevent	*How does a person begin to attract another person for reproducing?* **A.** kiss, **B.** genetic mutation, **C.** have sex , **D.** ...
HasPrerequisite	*If I am tilting a drink toward my face, what should I do before the liquid spills over?* **A.** open mouth, **B.** eat first, **C.** use glass , **D.** ..
CausesDesire	*What do parents encourage kids to do when they experience boredom?* **A.** read book, **B.** sleep, **C.** travel , **D.** ...
Desires	*What do all humans want to experience in their own home?* **A.** feel comfortable, **B.** work hard, **C.** fall in love , **D.** ...
PartOf	*What would someone wear to protect themselves from a cannon?* **A.** body armor, **B.** tank, **C.** hat , **D.** ...
HasProperty	*What is a reason to pay your television bill?* **A.** legal, **B.** obsolete, **C.** entertaining , **D.** ...

- Questions about commonsense knowledge
- Example COMMONSENSEQA [Talmor et al., 2019]: Questions derived from CONCEPTNET
- Evaluation: multiple choice, highest probability assigned to A, B, C or D

- Similar test set obtained with crowdsourcing: WinoGrande [Sakaguchi et al., 2021]

The trophy doesn't fit into the brown suitcase because **it**'s too *large*.	**trophy** / suitcase
The trophy doesn't fit into the brown suitcase because **it**'s too *small*.	trophy / **suitcase**

Mathematical Reasoning

- Ability to work through mathematical problems
- Example: MATH [Hendrycks et al., 2021]
- Taken from math competition problem sets
- Evaluation of the final solution (box in figure)

Problem: Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles. How many different groups of two marbles can Tom choose?

Solution: There are two cases here: either Tom chooses two yellow marbles (1 result), or he chooses two marbles of different colors ($\binom{4}{2} = 6$ results). The total number of distinct pairs of marbles Tom can choose is $1 + 6 = \boxed{7}$.

Problem: The equation $x^2 + 2x = i$ has two complex solutions. Determine the product of their real parts.

Solution: Complete the square by adding 1 to each side. Then $(x+1)^2 = 1 + i = e^{\frac{i\pi}{4}}\sqrt{2}$, so $x + 1 = \pm e^{\frac{i\pi}{8}}\sqrt[4]{2}$. The desired product is then $\left(-1 + \cos\left(\frac{\pi}{8}\right)\sqrt[4]{2}\right)\left(-1 - \cos\left(\frac{\pi}{8}\right)\sqrt[4]{2}\right) = 1 - \cos^2\left(\frac{\pi}{8}\right)\sqrt{2} = 1 - \frac{\left(1+\cos\left(\frac{\pi}{4}\right)\right)}{2}\sqrt{2} = \boxed{\frac{1-\sqrt{2}}{2}}$.

Academic Knowledge and Reasoning

One of the reasons that the government discourages and regulates monopolies is that
(A) producer surplus is lost and consumer surplus is gained. ✗
(B) monopoly prices ensure productive efficiency but cost society allocative efficiency. ✗
(C) monopoly firms do not engage in significant research and development. ✗
(D) consumer surplus is lost with higher prices and lower levels of output. ✓

- Questions about humanities, social science, STEM
- Example: MMLU [https://arxiv.org/pdf/2009.03300]
- Multiple choice questions
- Collected from practice questions for college or certification exams
- Multilingual version MMMLU: human translated into 14 languages

MMLU Reasoning Example

As Seller, an encyclopedia salesman, approached the grounds on which Hermit's house was situated, he saw a sign that said, "No salesmen. Trespassers will be prosecuted. Proceed at your own risk." Although Seller had not been invited to enter, he ignored the sign and drove up the driveway toward the house. As he rounded a curve, a powerful explosive charge buried in the driveway exploded, and Seller was injured. Can Seller recover damages from Hermit for his injuries?

(A) Yes, unless Hermit, when he planted the charge, intended only to deter, not harm, intruders. ✗

(B) Yes, if Hermit was responsible for the explosive charge under the driveway. ✓

(C) No, because Seller ignored the sign, which warned him against proceeding further. ✗

(D) No, if Hermit reasonably feared that intruders would come and harm him or his family. ✗

Test on Train?

- Grave concerns about training data contamination
- If test sets are built on web data $\rightarrow$ very likely in the training data
- Larger models $\rightarrow$ higher capacity to memorize

Question Answering Without the Question

- Can LLMs answer multiple-choice questions without the question? [Balepur et al., 2024]

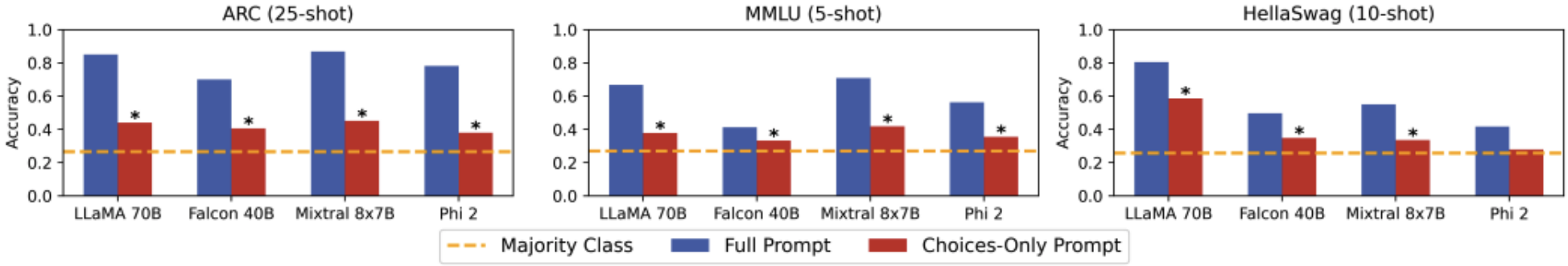
No Choices

Question: Which of these contains only a solution?
Answer: (B)

Empty Choices

Question: Which of these contains only a solution?
Choices: (A) \n (B) \n (C) \n (D) \n
Answer: (B)

- Results

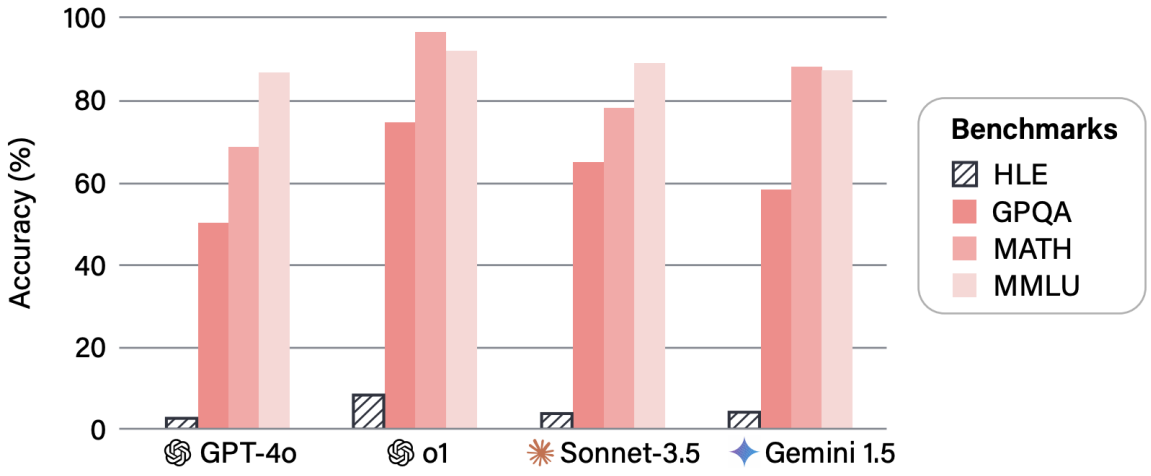


Text Generation?

- Note that none of these require generation of long fluent text
- Why? Evaluation of responses is difficult
  Write a story about a cow who wants to be a pig.
- Response similarity metrics (akin to BLEU) exist
  - e.g., ROUGE for summarization
  - ... but not very reliable
- A currently popular solution:
  ask a language model to score against reference response

Humanity's Last Exam

- 2,700 questions across dozens of subjects, including mathematics, humanities, and the natural sciences
- Written by subject-matter experts
- Multiple-choice and short-answer questions → suitable for automated grading



√x Mathematics

Question:
The set of natural transformations between two functors $F, G : C \to D$ can be expressed as the end

$$Nat(F, G) \cong \int_A Hom_D(F(A), G(A)).$$

Define set of natural cotransformations from F to G to be the coend

$$CoNat(F, G) \cong \int^A Hom_D(F(A), G(A)).$$

Let:
- $F = B_\bullet(\Sigma_4)_{*/}$ be the under ∞-category of the nerve of the delooping of the symmetric group Σ_4 on 4 letters under the unique 0-simplex $*$ of $B_\bullet\Sigma_4$.
- $G = B_\bullet(\Sigma_7)_{*/}$ be the under ∞-category nerve of the delooping of the symmetric group Σ_7 on 7 letters under the unique 0-simplex $*$ of $B_\bullet\Sigma_7$.

How many natural cotransformations are there between F and G?

Emily S
University of São Paulo

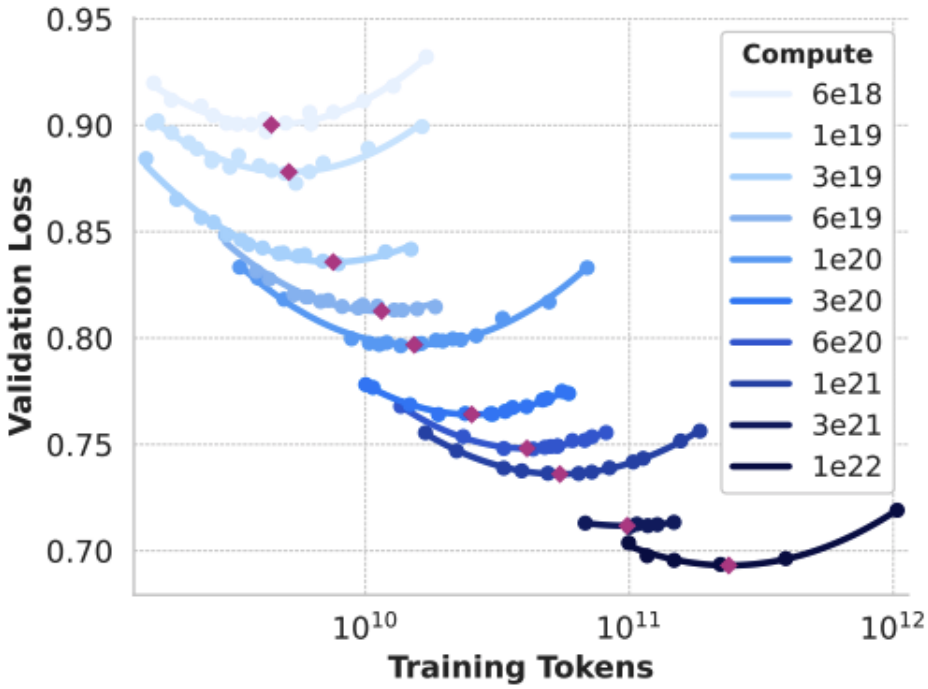
training

Three Stages of Training

- Stage 1: Train on massive amounts of text (up to a trillion words)
- Stage 2: Instruction training
- Stage 3: Reinforcement learning from human feedback

pretraining

Training



- Scaling laws: more data → bigger models → better performance
- Today: trillions of words → 10s to 100s of billions of parameters
- Llama3 405B: trained on 16,384 GPUs — available open source

Massive Amounts of Text

- Web crawls
  - publicly available raw data: CommonCrawl
  - filtered and cleaned data: Fineweb
- eBooks
- Compute code (from github)
- Trillions of words

Massive Engineering Effort

- Example: Llama3
  - 16K H100 GPUs
  - 54 days
- Example: Deepseek V3
  - 2048 H800 GPUs
  - 2.8 million H800 GPU hours
- Rarely done — even big company do this maybe once a year

instruction tuning

Instruction Training

- Examples of requests and responses constructed by human annotators
- May be collected from actual user requests and edited by experts
- May be generated from existing data sets

Question Answering	**Summarization**	**Translation**
What is the highest mountain in the world? *Answer:* The highest mountain in the world is Mount Everest.	*Summarize the following paragraph into one sentence.* The Federal Reserve paused its campaign of interest rate increases for the first time in more than a year. But officials suggested that rates would rise more in 2023, as inflation remains "well above" the central bank's target. *Summary:* No interest rate rise for now but maybe later in the year.	*Translate from English to German.* *English:* My name is Ozymandias, King of Kings; Look on my Works, ye Mighty, and despair! *German:* Mein Name ist Ozymandias, König der Könige; Schau auf meine Werke, du Mächtiger, und verzweifle!

Datasets from Tasks

- Long-running research efforts into specific NLP tasks
  - machine translation
  - questions answering (about a provided document / open-ended)
  - summarization (given one or more documents)
  - named entity detection (finding company, person, etc. names)
  - sentiment detection (positive review?)
  - grammar correction
  - ... and many more

- For all these training data and test sets were created

$\Rightarrow$ These can be converted into LLM training data

Chat Interactions

- We expect the model to answer user questions

$\Rightarrow$ Text needs to be adapted

Instruction

User: "Write a summary of the movie Interstellar."

Model: "Interstellar is a sci-fi movie directed by Christopher Nolan. It follows a group of astronauts who travel through a wormhole in search of a new home for humanity."

Chat

User: "Okay, let's say I'm writing a story. Can you help me brainstorm some ideas for a fantasy setting?"

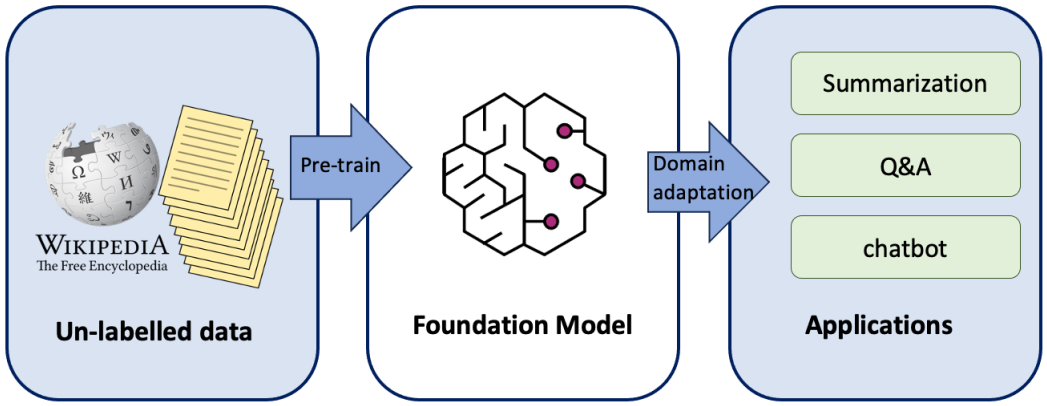
Model: "Sure, I can help with that. What kind of fantasy setting are you interested in? Do you have any specific elements in mind, like a particular type of magic system or a specific type of monster?"

Human Answers to Prompts

- Developers of LLMs collect a lot user questions
- These are evaluated for quality control
- Good responses ⇒ use as instruction training data
- Bad responses ⇒ (expert) human create acceptable answers
  use as instruction training data

Supervised Fine-Tuning

- Take pre-training model
- Continue training with instruction data

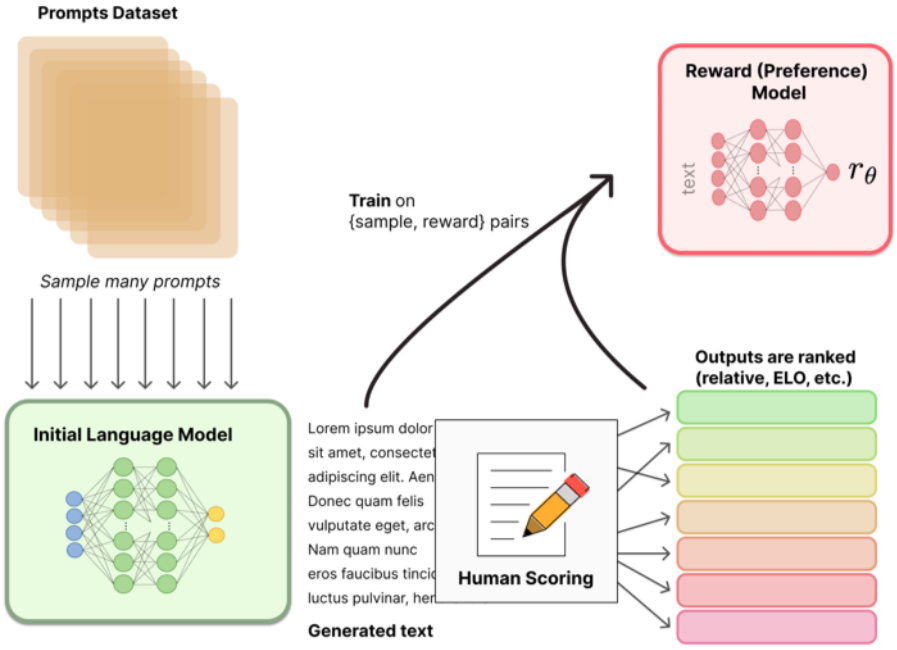


- This can be done many times, in different ways
  ... even by application builders with modest resources

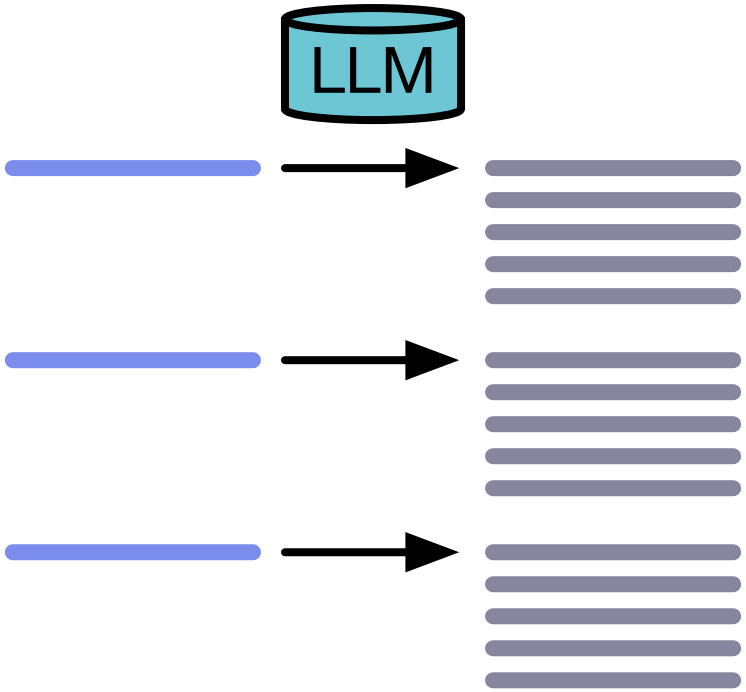
preference training

Reinforcement Learning from Human Feedback

- Machine generates multiple responses to a prompt
- Human annotators rank them
- Train a reward model
  - predict human annotation
  - can be applied to any text generated by model
  - normalized as reward function
- Fine-tune model with reward model
  - model generates response to prompt
  - reward function assesses response
  - if low reward, model needs to change

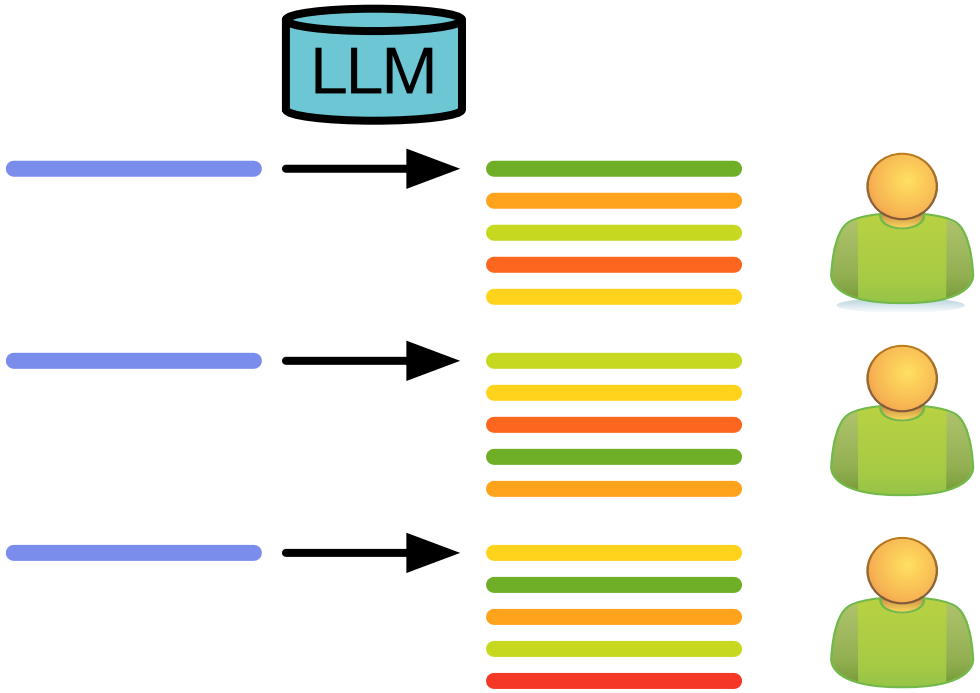


Learning from Human Preferences



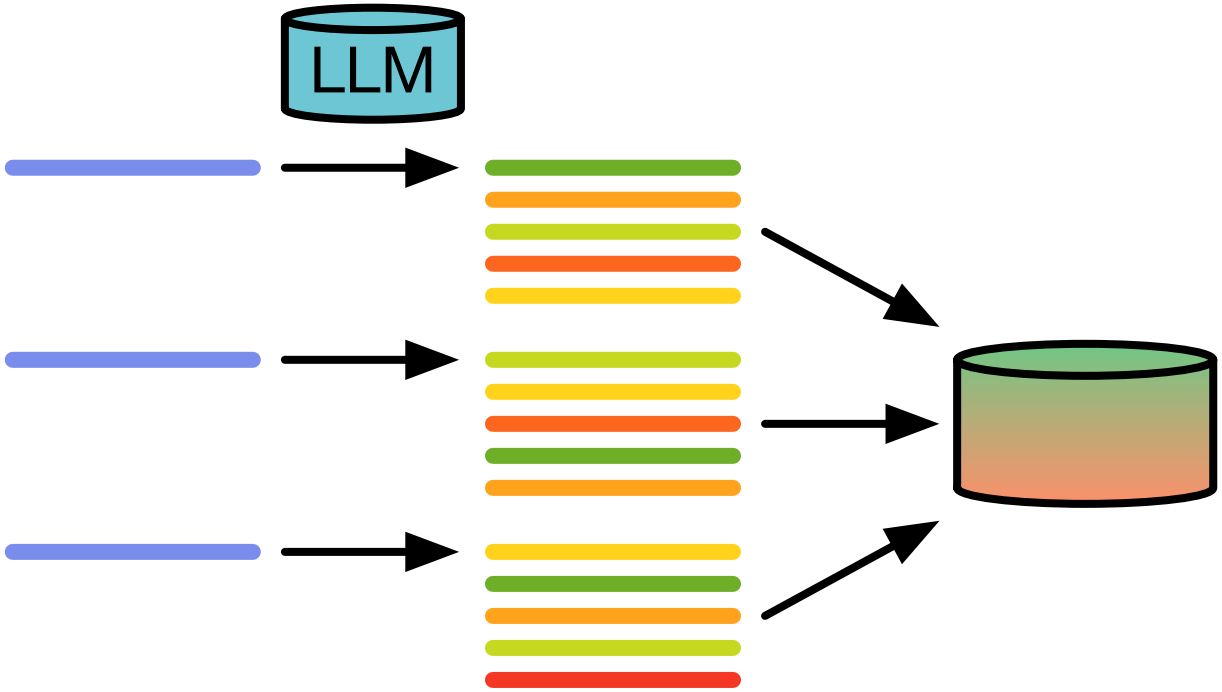
- Generate responses from a prompt by sampling
  - greedy decoding: always choose word prediction with 80% probability
  - Monte Carlo decoding: choose it 80% of the time

Learning from Human Preferences



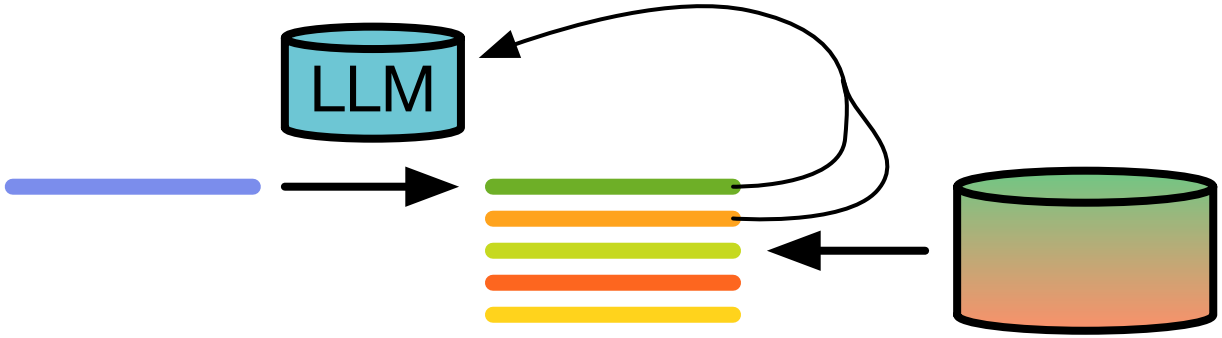
- Human annotators rank the responses
- This is easier to do that authoring responses but still expensive

Learning from Human Preferences



- Train a preference model
- Typically based on sequence representations from language models

Learning from Human Preferences



- Use the preference model during training original model
  - for a prompt, generate responses
  - score the responses with the preference model
  - update model to
    - promote higher-scoring responses (winner)
    - demote lower-scoring responses (loser)

Machine Translation: Build on Existing Work

- Where do we get human preference judgments from?

- Long-running WMT evaluation campaign (since 2006)
  - participants submit translations of their system
  - human evaluators score or rank these translations

  → this is the human preference data we need

- Quality estimation models
  - long-standing interest in models that can score translations without a reference
  - example: confidence estimation on how good machine translation is, to, e.g.,
    - decide between output from different systems
    - decide if it should be post-edited by human translator
    - inform end user

  → publicly released quality estimation models, e.g., CometKiwi [Rei et al., 2022]

Reinforcement Learning from Human Feedback

- This idea was originally introduced as a form of reinforcement learning
- The idea of a reward model stems from reinforcement learning
- Method: Proximal Policy Optimization (PPO)
- Recently, simpler methods are more common

Direct Preference Optimization (DPO)

- First train a reward model r^*
- Sample two possible responses for an input x
- Score them with the reward model
  - higher scoring translation is the winner y_w
  - higher scoring translation is the loser y_l
- Train a new model π_θ from an original model π_{ref} (using a hyper parameter β)

$$\text{Loss}(x, y_w, y_l) = \log \text{sigmoid}\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)$$

[from Rafailov et al., 2023]

Contrastive Preference Optimization (CPO)

- Computing probabilities with both
  - the reference model $\pi_{\text{ref}}(y|x)$ and
  - the new model $\pi_\theta(y|x)$

  is expensive

  $\rightarrow$ twice the memory requirements
  $\rightarrow$ twice the number of computations

- Simplification: only score with new model

$$\text{Loss}(x, y_w, y_l) = \log \text{sigmoid}\Big(\beta \log \pi_\theta(y_w|x) - \beta \log \pi_\theta(y_l|x)\Big)$$

[from Xu et al., 2024]

compact models

Large Language Models Very Costly

- 10-1000 times as many parameters as dedicated MT models
- More powerful machines needed (with multiple $20,000 GPUs)
- Slower, each translation request more expensive
- Very costly to adapt to particular user cases

Large Language Models for Mere Mortals

- Considering the size of language models
  - parameters are typically stored as 16-bit floats
  - during training also gradients and optimizer states need to be stored

  ⇒ 6 bytes per parameter
  - Also need to store the state of training examples (depends on sequence length and batch size)

- Size of GPUs
  - A100: 40-80GB RAM ($15,000)
  - RTX2080ti: 11GB RAM ($800)

- Only a few billion parameters models fit on single GPU

Compressed Models

QLoRA: Efficient Finetuning of Quantized LLMs

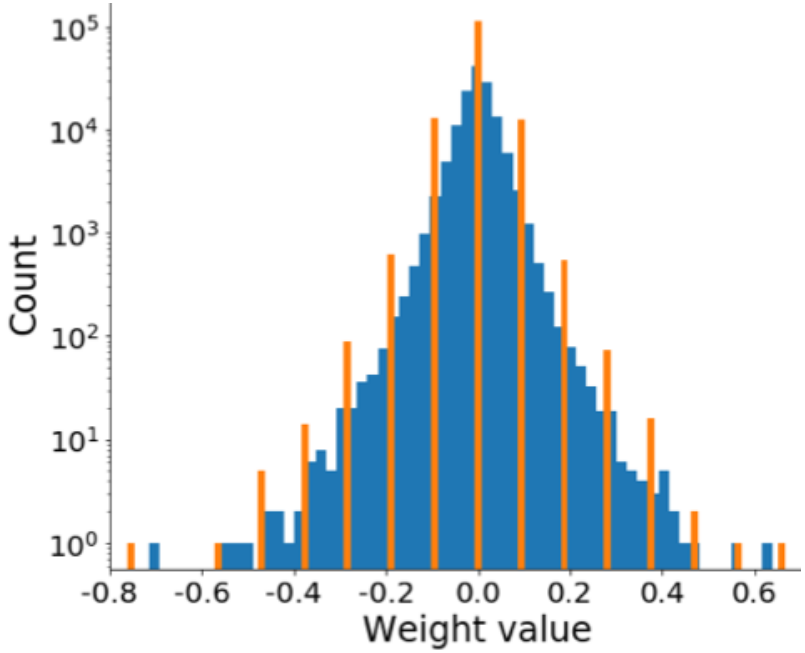
Tim Dettmers* **Artidoro Pagnoni*** **Ari Holtzman**

Luke Zettlemoyer

University of Washington
{dettmers,artidoro,ahai,lsz}@cs.washington.edu

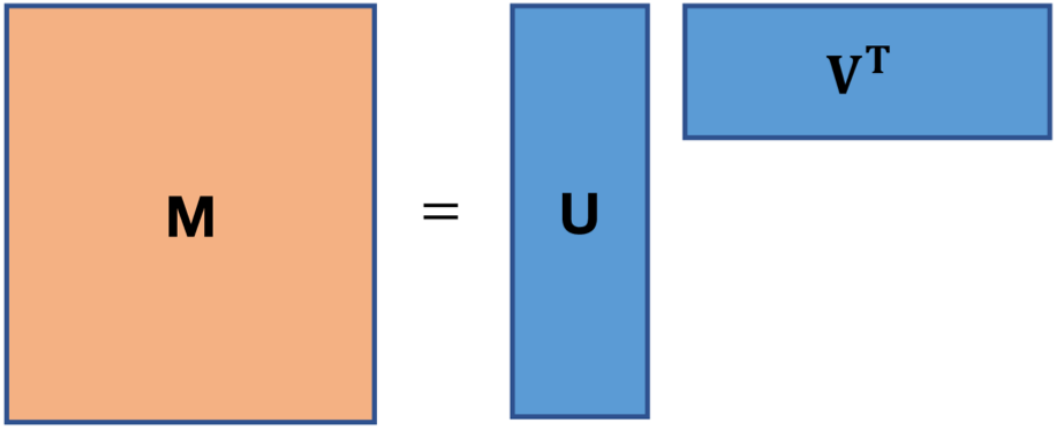
- Low-rank adaptation parameter matrices
- Quantize all values into 4-bit floats
- Integrated into Huggingface, may adapt any model

Solution 1: Quantization



- Store values in 4 bit floats (or less)
- Computation still in 16 bits
- Additional tricks: double quantization, paged optimizers

Solution 2: Low-Rank Adaptation (LoRA)



- Keep original model intact during adaptation
- Add adaptation parameters in form of low-rank matrices
  - original: n^2 parameter matrix M
  - adaptation:
    - nr and rn matrices U, V
    - with $r << n$
    - e.g., n=2048, r=16

Solution 3: Knowledge Distillation

- Large language model as Teacher
- Small language model as Student
- Data distillation
  - process task-relevant data with Teacher model → good responses
  - use this synthetic data to train Student model
- Model distillation
  - Train Student model directly on predictions of Teacher model

questions?