VIS 2023

NLP4Vis: Natural Language Processing for Information Visualization

Half-day Tutorial

Enamul Hoque, York University

https://nlp4vis.github.io/

Tutorial Overview

- **Part 1**: Introduction [15 mins]
  - Why NLP + Vis?
  - An overview of NLP + Vis Research
  - An overview of the tutorial
- **Part 2**: Deep Learning for NLP [50 mins]
  - Background
  - Large language models (LLMs)
- **Part 3**: NLP4Vis applications [50 mins]
- **Part 4**: Future challenges and research opportunities [25 mins]

Part 2: Deep Learning for NLP

Agenda

- **Background**
  - Introduction to NLP
  - Language modeling
  - **Model architectures**
    - Transformer architecture
    - Encoder, decoder, encoder-decoder
    - Pre-training and fine-tuning
- **Large language models (LLMs)**
  - Scaling LMs to LLMs
  - Prompt engineering
  - In context learning
  - Instruction tuning

What is NLP?

We study formalisms, models and algorithms to allow computers to perform useful tasks involving knowledge about human languages.

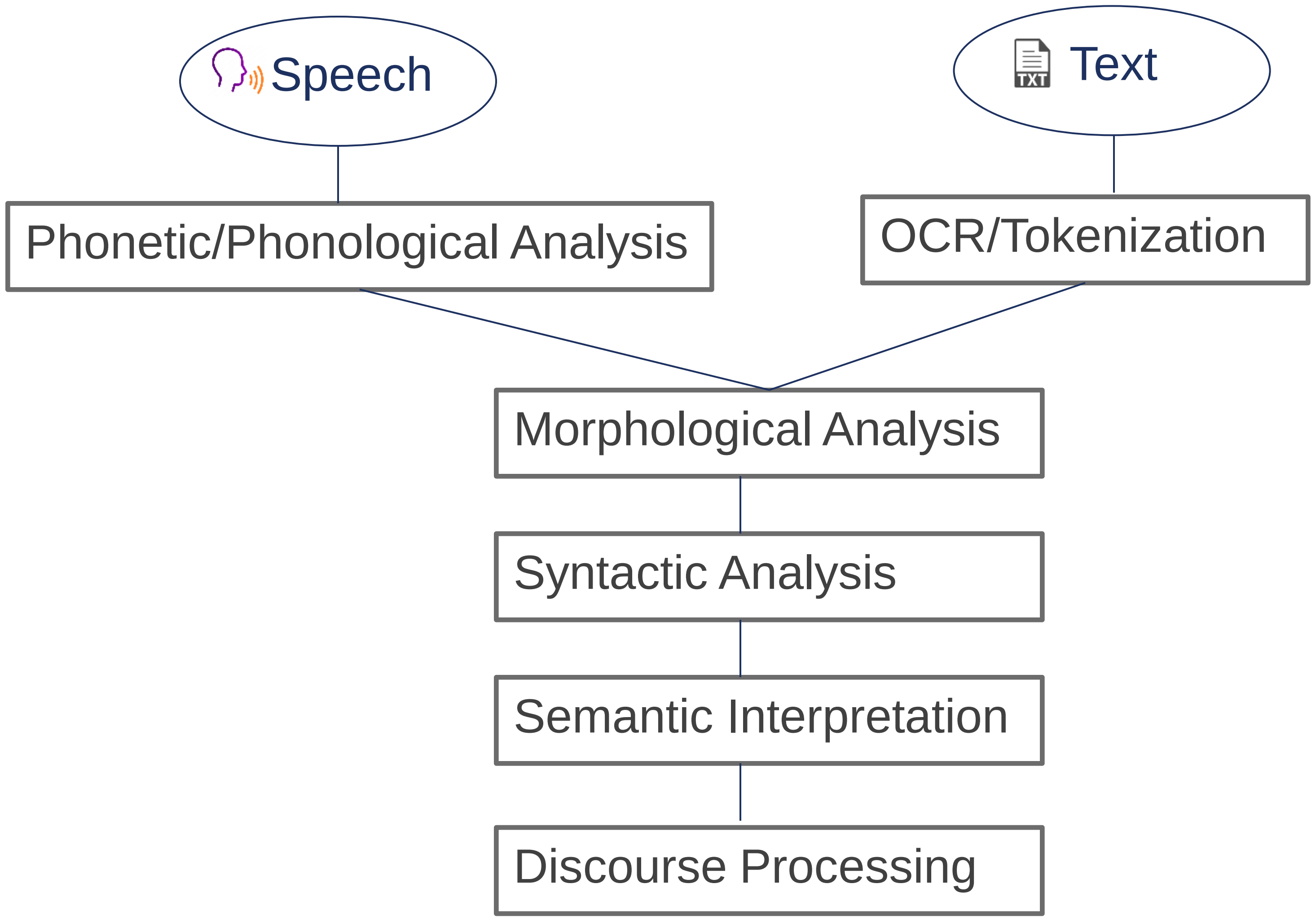
Useful Tasks

- **Conversational agents**:
  - AT&T "How may I help you?" technology
  - Apple SIRI, Amazon's alexa, Microsoft's Cortana.
- **Summarization**:
  - "Please summarize my discussion with Sue about NLP" "What people say about the new Nikon 5000?"
- **Machine Translation**:
  - Google translate (100B USD$ industry)
- **Text Generation**:
  - Data2Text, Table2Text, Chart2Text
- **Question answering**:
  - "Was 1991 an El Nino year? ….Was it the first one after 1982?" "Why was it so intense?"
- **Document Classification**:
  - spam detection, news filtering
- **Speech**:
  - speech recognition, text to speech synthesis.
- **Multimodal**:
  - Video/Image2Text, text2image (captioning, VQA)

What is NLP?

We study formalisms, models and algorithms to allow computers to perform useful tasks involving knowledge about human languages.

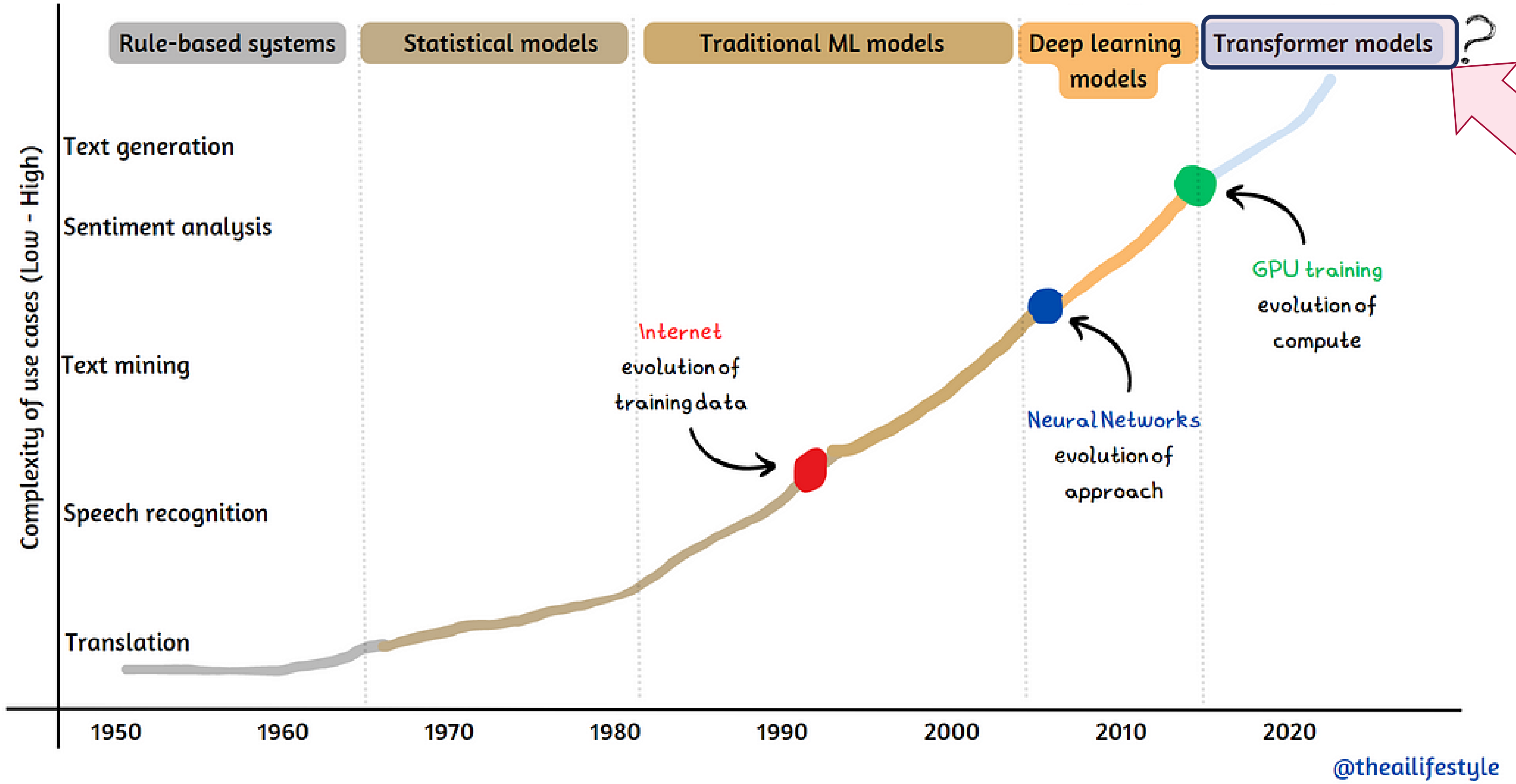
Knowledge about Language



What is NLP?

We study formalisms, models and algorithms to allow computers to perform useful tasks involving knowledge about human languages.

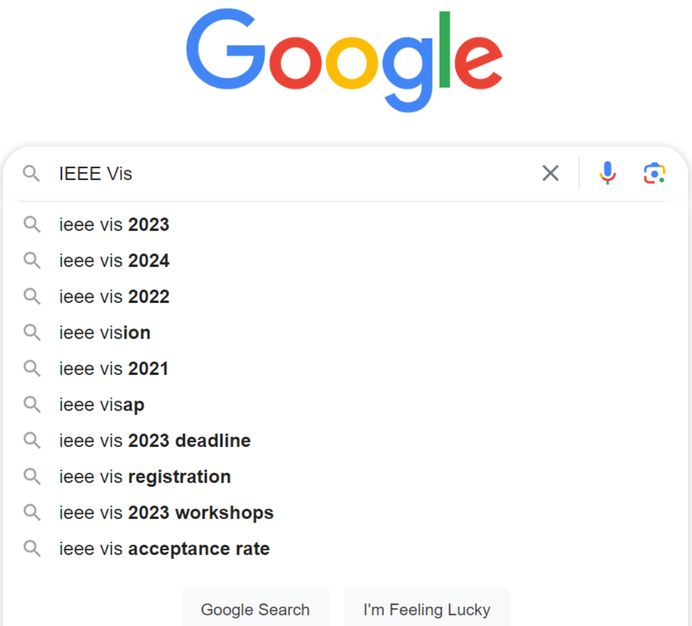
What is NLP?

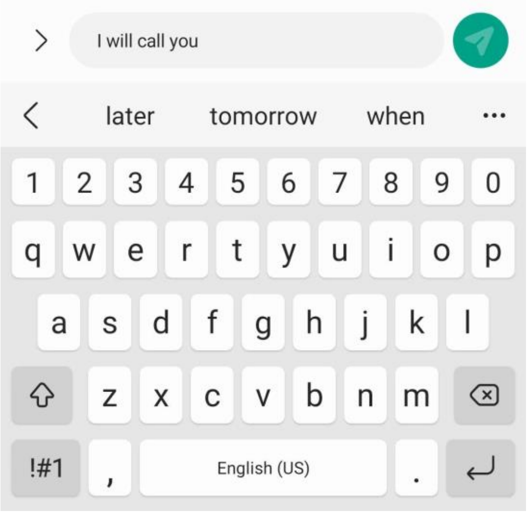


This tutorial

Background: Language Modeling

- A language model takes a list of words (history/context/prompt), and attempts to predict the word that follows them





Why Language Modeling is Important?

- A benchmark task to track our progress on understanding language
- An important component of many NLP tasks, especially those involving generating text or estimating the probability of a text
  - Speech recognition
  - Spelling/grammar correction
  - Machine translation
  - Summarization
  - Dialogue etc.
- Language modeling is by far the most successful self-supervision objective to (pre)train large language models (LLMs)
  - Cheap!

Background: Language Modeling

- A language model takes a list of words (history/context), and attempts to predict the word that follows them

Causal Language Model: predicts the next token

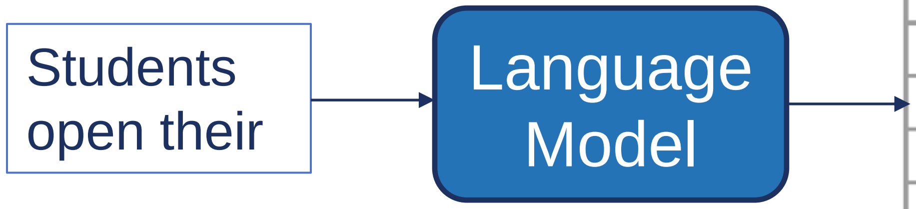
More formally: given a sequence of words $x_{(1)}, x_{(2)}, \dots x_{(t)}$, compute the probability distribution of the next word $x_{(t+1)}$:

$$P(x_{(t+1)} \mid x_{(t)}, \dots, x_{(1)})$$

where $x_{(t+1)}$ can be any word in the vocabulary $V = \{w_1, \dots, w_{|V|}\}$

Background: Language Modeling

- Causal Language Modeling



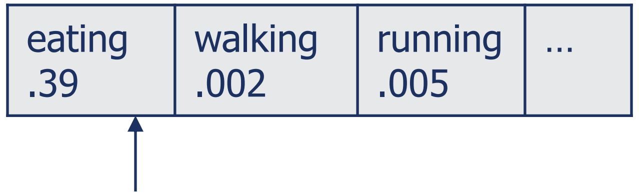
(hypothetical)

Word	Probability
a	*0.00001*
aardvark	*0.000004*
...	
books	*0.5*
...	
laptops	*0.23*
...	
zucchini	*0.000002*

The best language model is the one that best predicts an unseen test case (i.e., best test loss)

Background: Language Modeling

- Masked Language Modeling
  - aka fill in the blanks/cloze



The cat is [MASK] some food.

What Can LMs Learn From Word Prediction?

- **Grammar** In my free time, I like to **{run, banana}**
- **Lexical semantics** I went to the zoo to see giraffes, lions, and **{zebras, spoon}**
- **World knowledge** The capital of Denmark is **{Copenhagen, London}**
- **Sentiment analysis** Movie review: I was engaged and on the edge of my seat the whole time. The movie was **{good, bad}**
- **Translation** The word for “pretty” in Spanish is **{bonita, hola}**
- **Spatial reasoning** [...] Iroh went into the kitchen to make some tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the **{kitchen, store}**
- **Math question** First grade arithmetic exam: 3 + 8 + 4 = **{15, 11}**

Part 2: Deep Learning for NLP

Agenda

- **Background**
  - Introduction to NLP
  - Language modeling
- **Model architectures**
  - Transformer architecture
  - Encoder, decoder, encoder-decoder
  - Pre-training and fine-tuning
- **Large language models (LLMs)**
  - Scaling LMs to LLMs
  - Prompt engineering
  - In context learning
  - Instruction tuning

Part 2: Deep Learning for NLP

Agenda

- **Background**
  - Introduction to NLP
  - Language modeling
- **Model architectures**
  - Transformer architecture
  - Encoder, decoder, encoder-decoder
  - Pre-training and fine-tuning
- **Large language models (LLMs)**
  - Scaling LMs to LLMs
  - Prompt engineering
  - In context learning
  - Instruction tuning

Transformers: Transforming the NLP Field

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

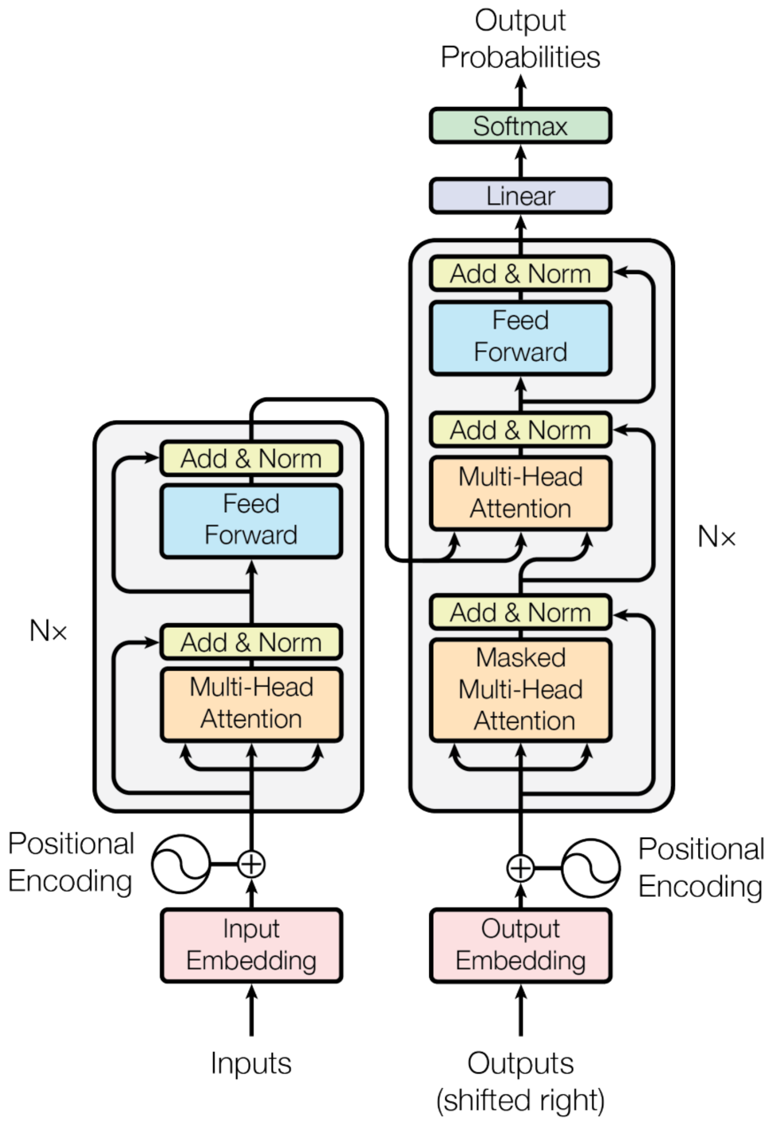
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

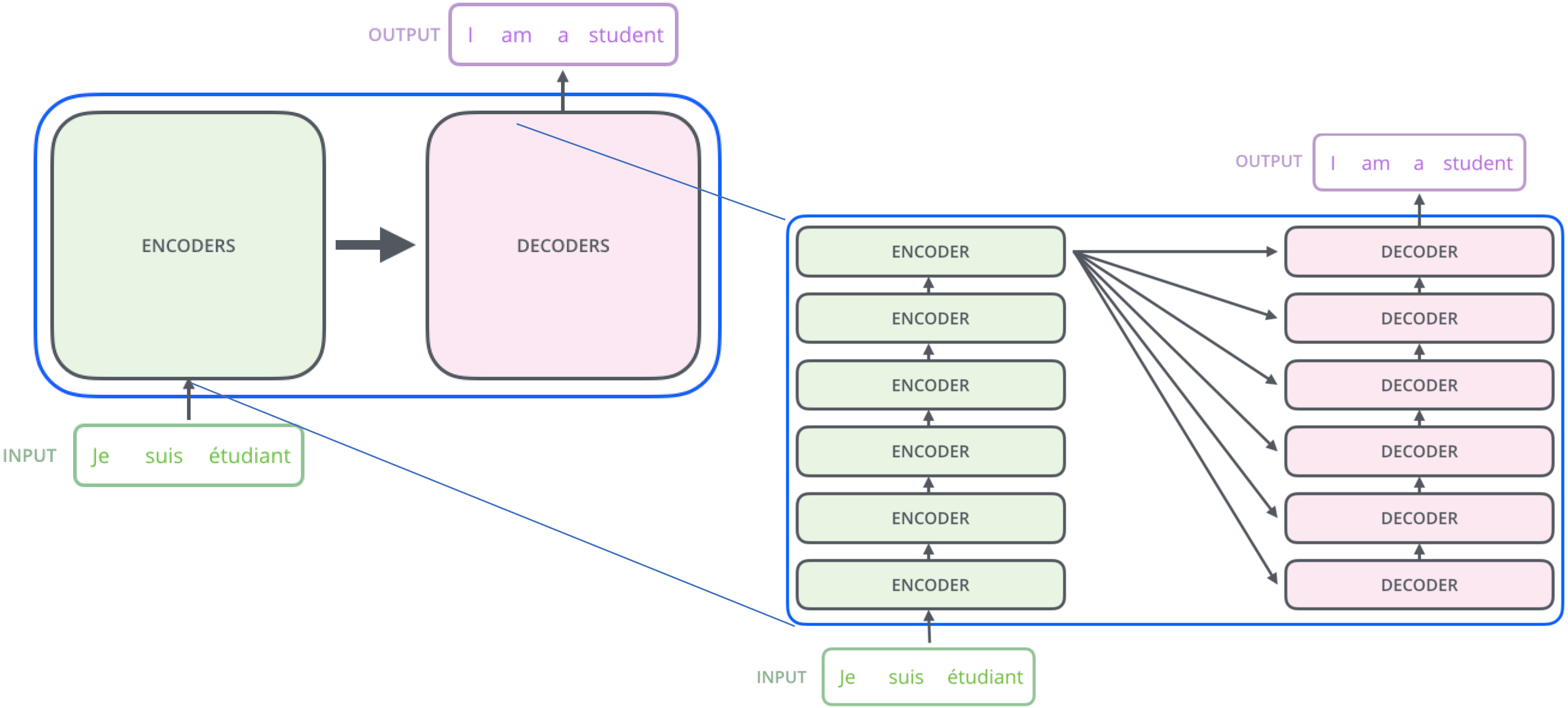
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

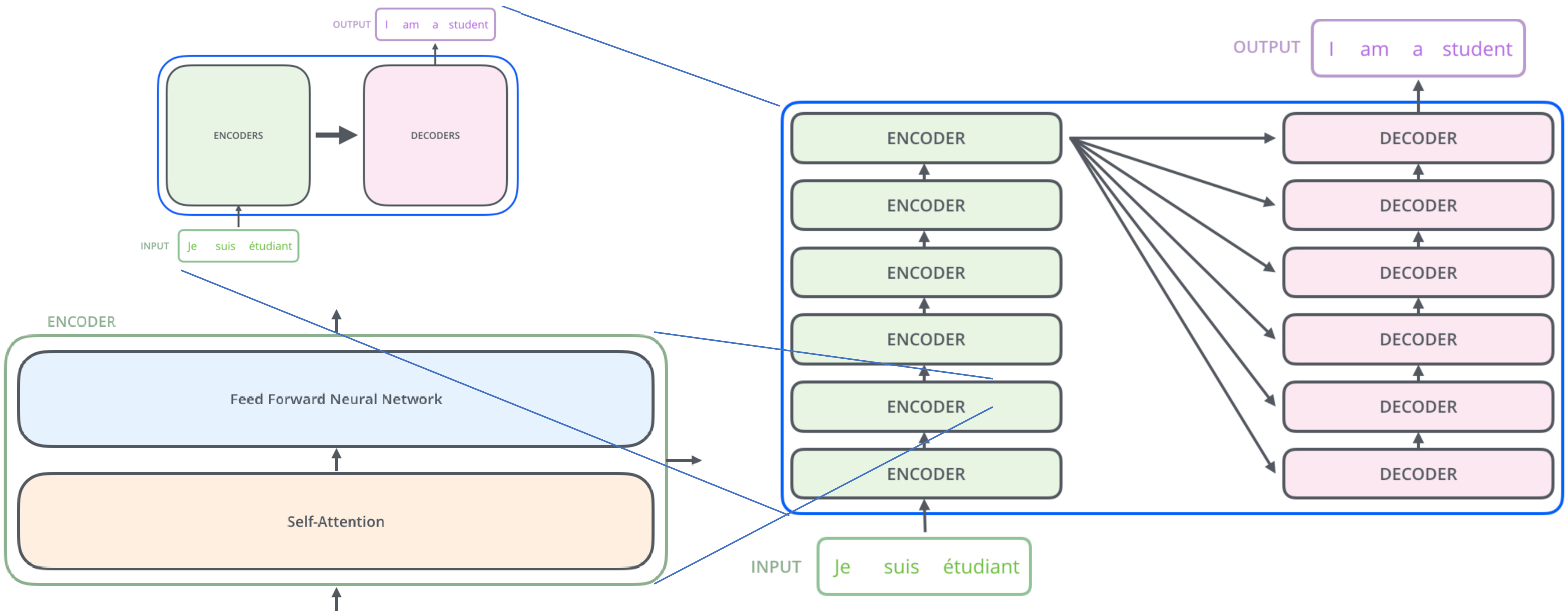


Transformer: High-level Architecture



https://jalammar.github.io/illustrated-transformer/

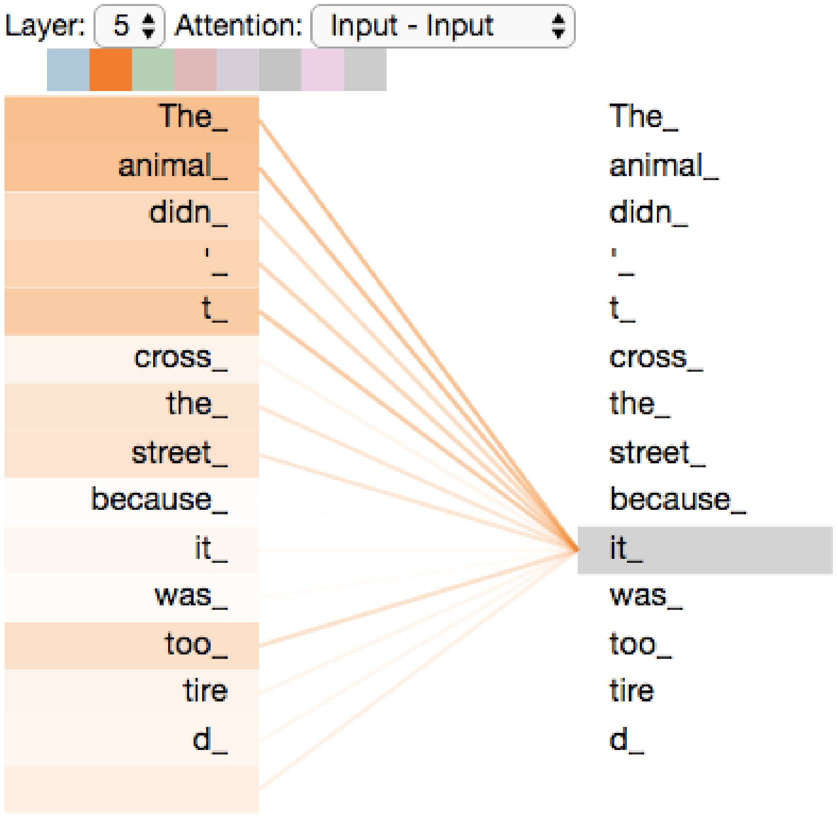
Transformer: High-level Architecture



https://jalammar.github.io/illustrated-transformer/

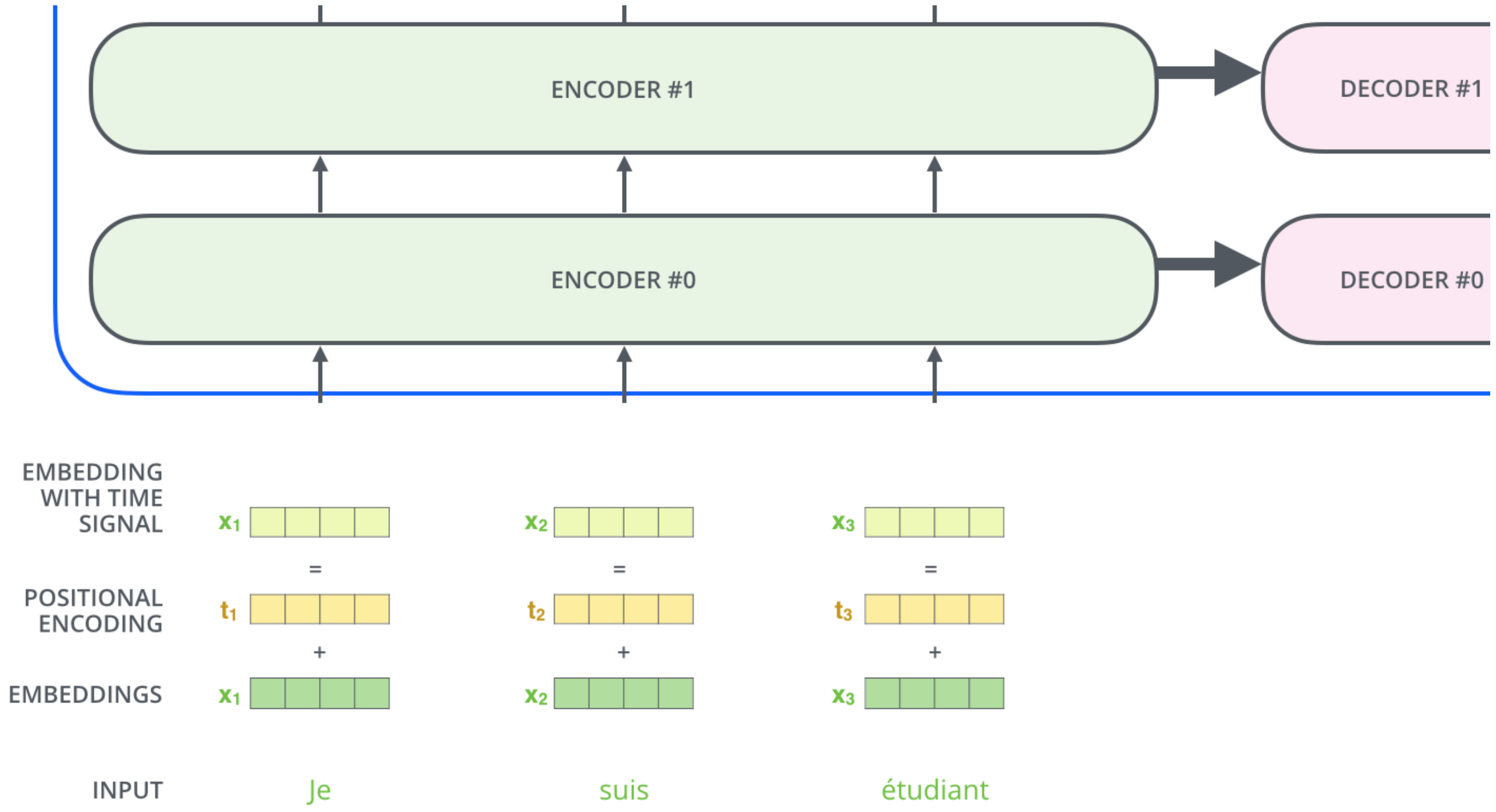
Transformer: Self-Attention at a High Level

"The animal didn't cross the street because **it** was too tired"



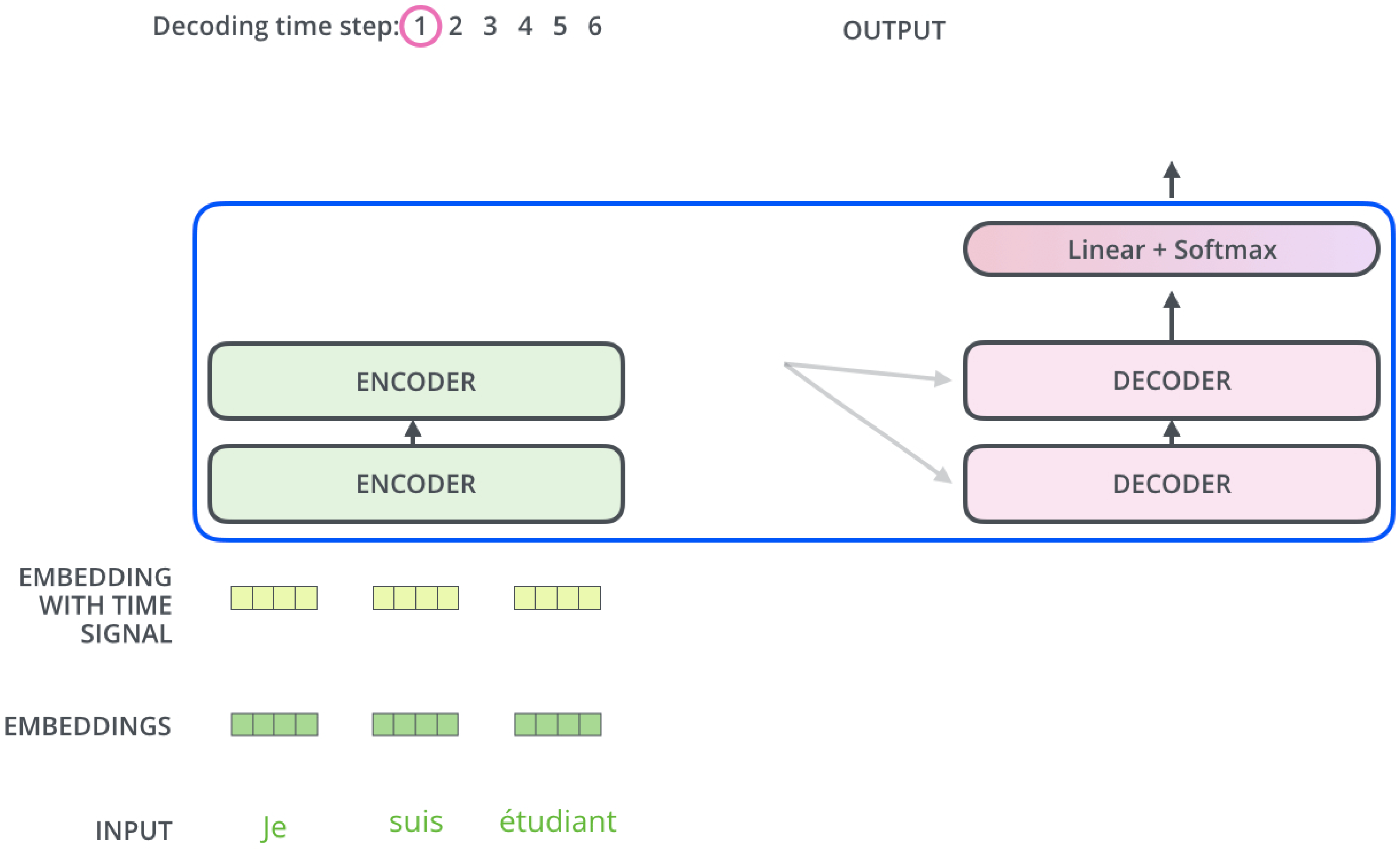
https://jalammar.github.io/illustrated-transformer/

Transformer: Embeddings

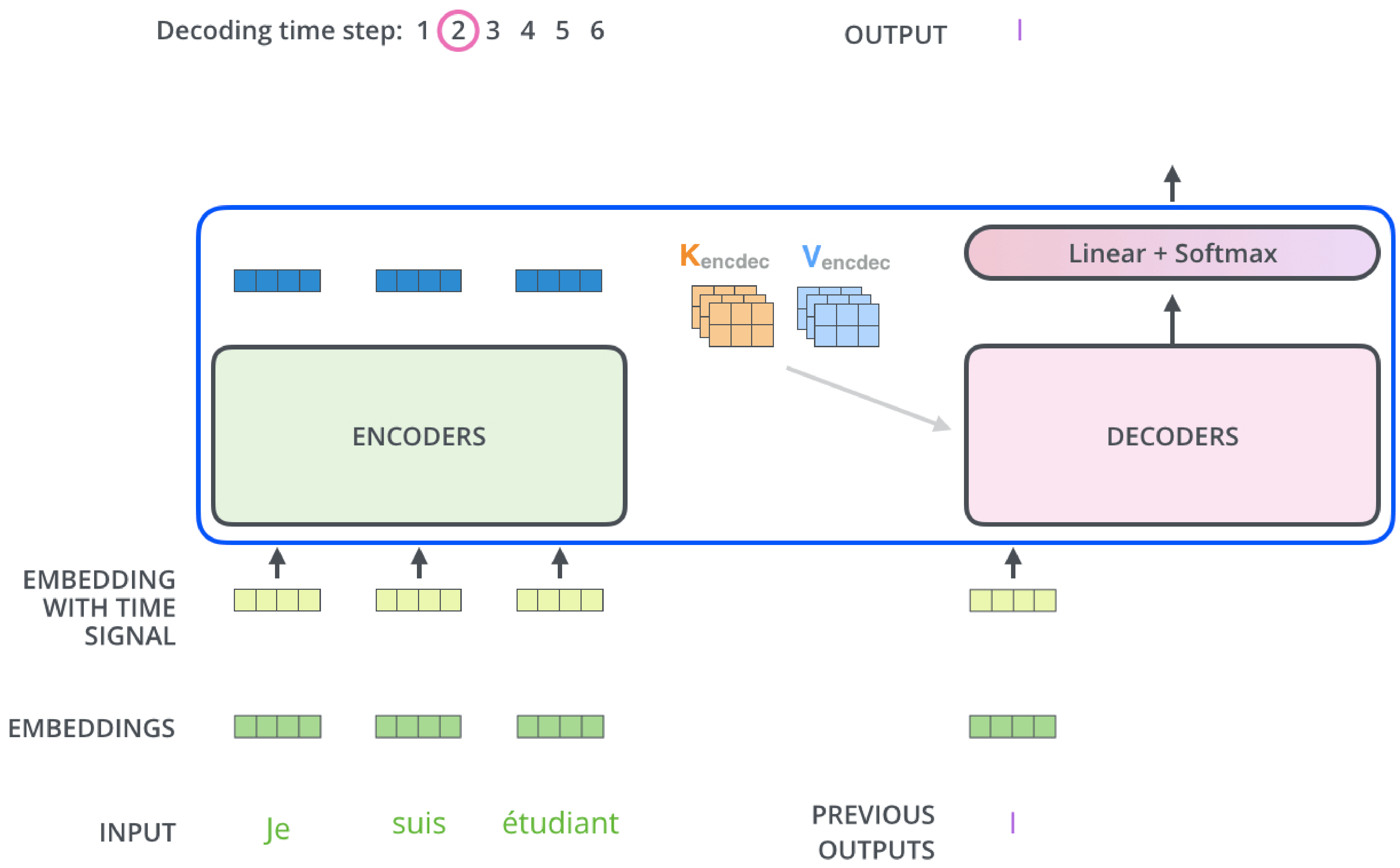


https://jalammar.github.io/illustrated-transformer/

Transformer: Decoders



Transformer: Decoders



https://jalammar.github.io/illustrated-transformer/

Part 2: Deep Learning for NLP

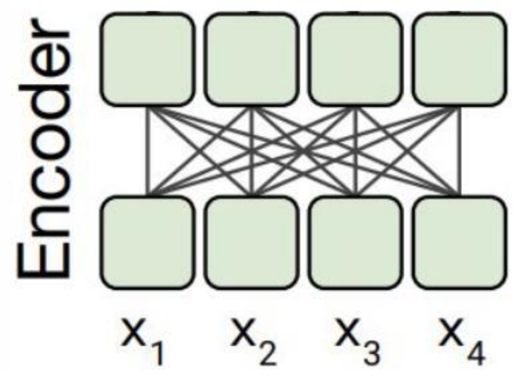
Agenda

- **Background**
  - Introduction to NLP
  - Language modeling
- **Model architectures**
  - Transformer architecture
  - Encoder, decoder, encoder-decoder
  - Pre-training and fine-tuning
- **Large language models (LLMs)**
  - Scaling LMs to LLMs
  - Prompt engineering
  - In context learning
  - Instruction tuning

Model Architectures for LM

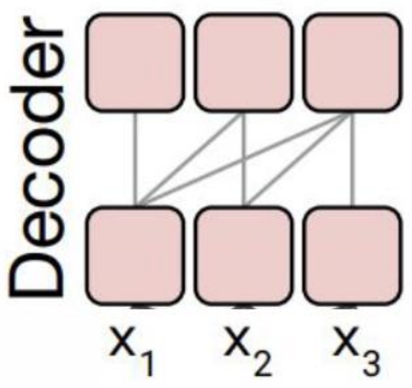
Encoder

- Bi-directional attention
- Entire input
- Prediction
- e.g., BERT, RoBERTa



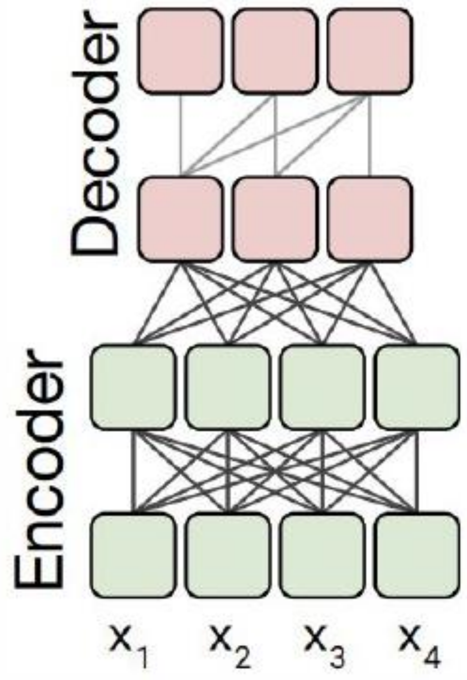
Decoder

- Causal attention
- One at a time
- Generation
- e.g., GPT, LLaMA



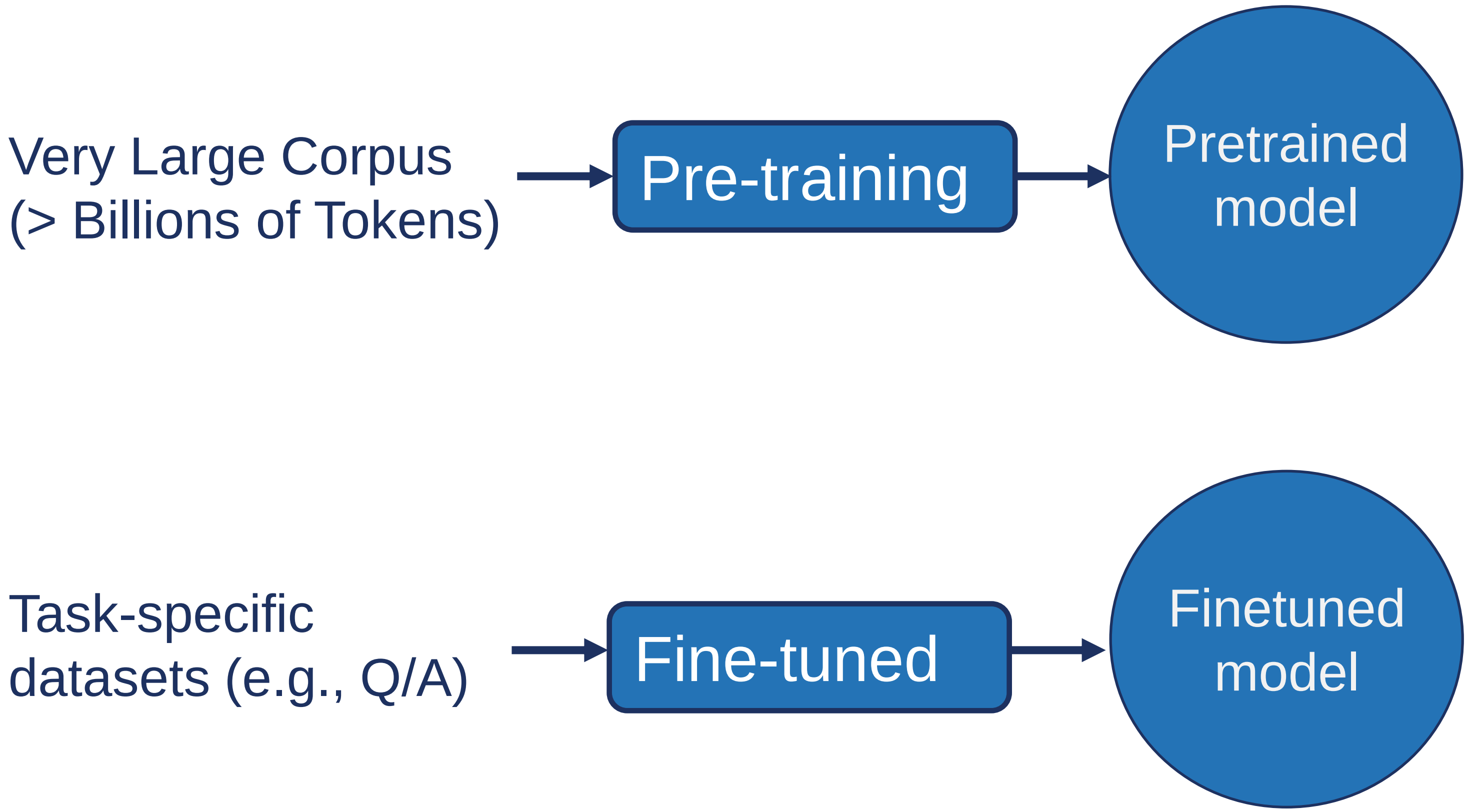
Encoder-Decoder

- Cross attention
- Self attention
- e.g., BART, T5



Training Phases of Language Models

- **Pre-training**: trained on huge amounts of unlabeled text using "self-supervised" training objectives
- **Adaptation**: how to use a pretrained model for your downstream task?

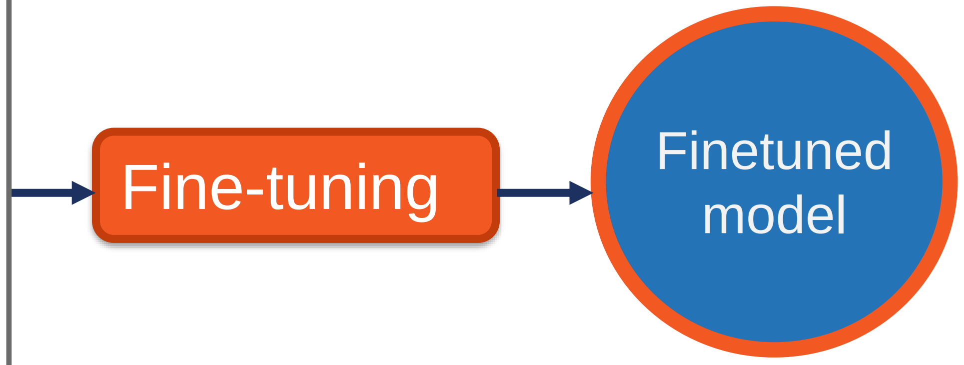


Single task (full-model) fine-tuning

- Bring your own dataset and retrain the model by tuning every weight in the pretrained model.

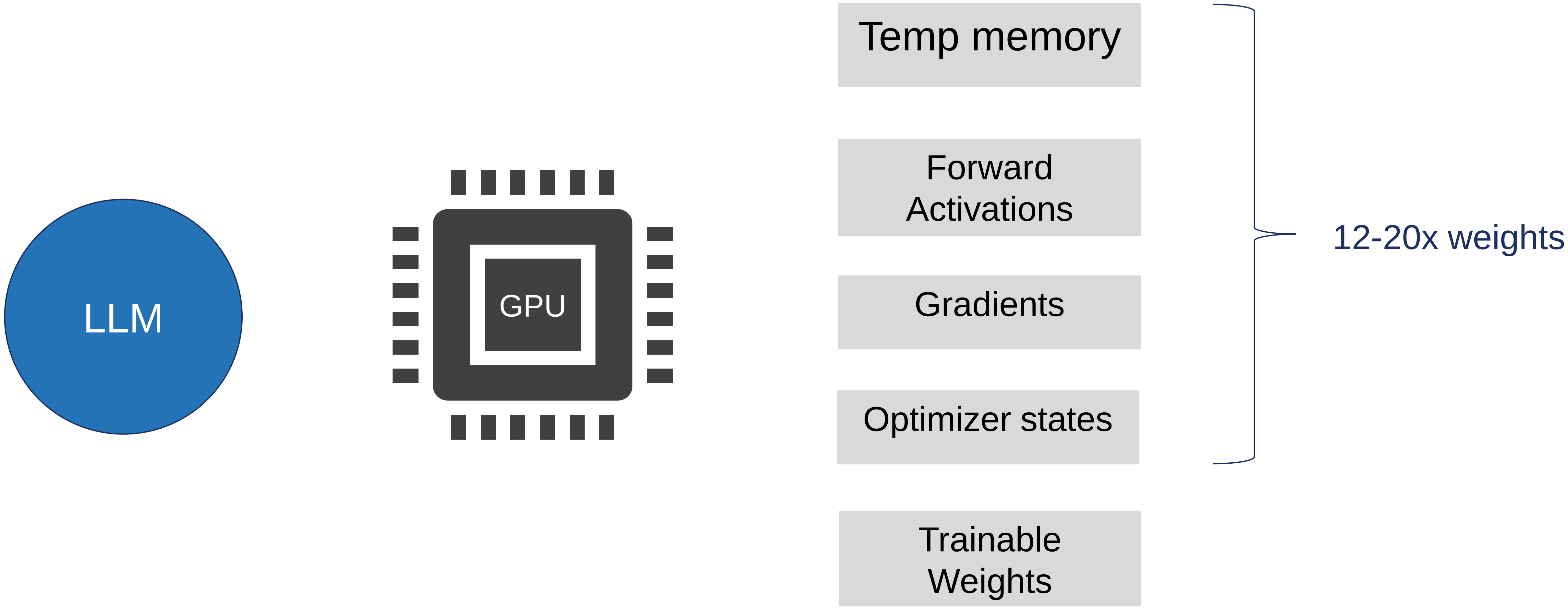
> **Example Input:** Analyze the following bar chart in one paragraph.
> Chart Data: Characteristics| Public| Private & 2021| 149| 209| & 2020 | 146| 202| 2019| 154 | 250 & …
> Chart title: Number of public and private hospitals in Malaysia from 2017 to 2021
> **Example Output:** In 2021, there were around 146 government hospitals and 209 private licensed hospitals in Malaysia. During the COVID-19 pandemic, the Malaysian hospitals were prepared by the government to accommodate infected patients by increasing bed numbers…

…



- **Problem 1**: Requires a lot of computing resources.
  - expensive and not realistic in many cases
  - Are there more efficient methods?
- **Problem 2**: Catastrophic Forgetting

Full fine-tuning of Large LLMs is Challenging



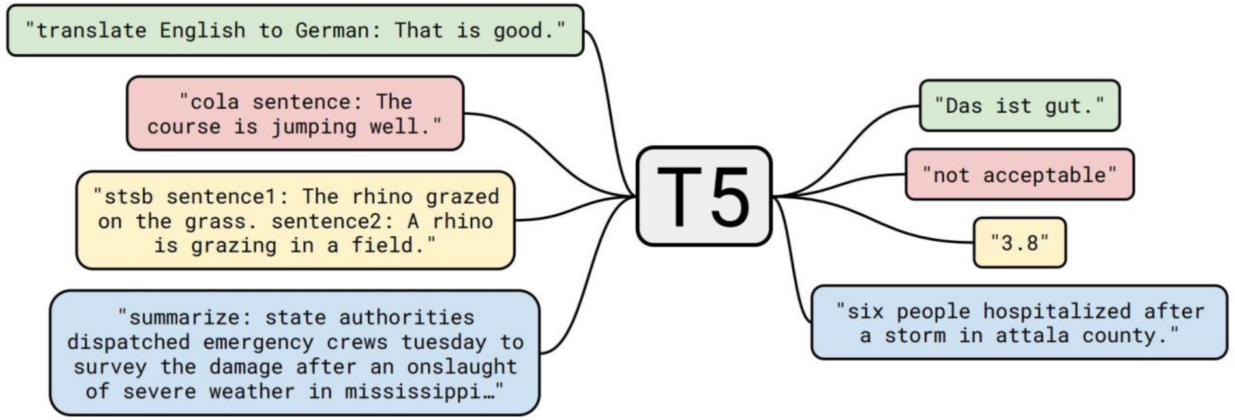
How to Avoid Catastrophic Forgetting?

- How to Avoid Catastrophic Forgetting?
  - Fine-tune on multiple tasks at the same time
  - Consider parameter efficient fine-tuning

Multi-task (full-model) fine-tuning

Question	Context	Answer
What is a major importance of Southern California in relation to California and the US?	...Southern California is a major economic center for the state of California and the US....	major economic center
What is the translation from English to German?	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser
What is the summary?	Harry Potter star Daniel Radcliffe gains access to a reported £320 million fortune...	Harry Potter star Daniel Radcliffe gets £320M fortune...
Hypothesis: Product and geography are what make cream skimming work. Entailment, neutral, or contradiction?	Premise: Conceptually cream skimming has two basic dimensions – product and geography.	Entailment
Is this sentence positive or negative?	A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.	positive

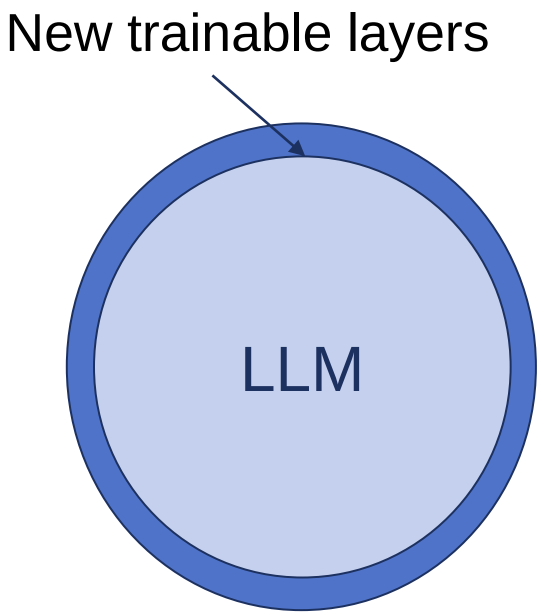
Question	Context	Answer
What has something experienced?	Areas of the Baltic that have experienced eutrophication.	eutrophication
Who is the illustrator of Cycle of the Werewolf?	Cycle of the Werewolf is a short novel by Stephen King, featuring illustrations by comic book artist Bernie Wrightson.	Bernie Wrightson
What is the change in dialogue state?	Are there any Eritrean restaurants in town?	food: Eritrean
What is the translation from English to SQL?	The table has column names... Tell me what the notes are for South Australia	SELECT notes from table WHERE 'Current Slogan' = 'South Australia'
Who had given help? Susan or Joan?	Joan made sure to thank Susan for all the help she had given.	Susan



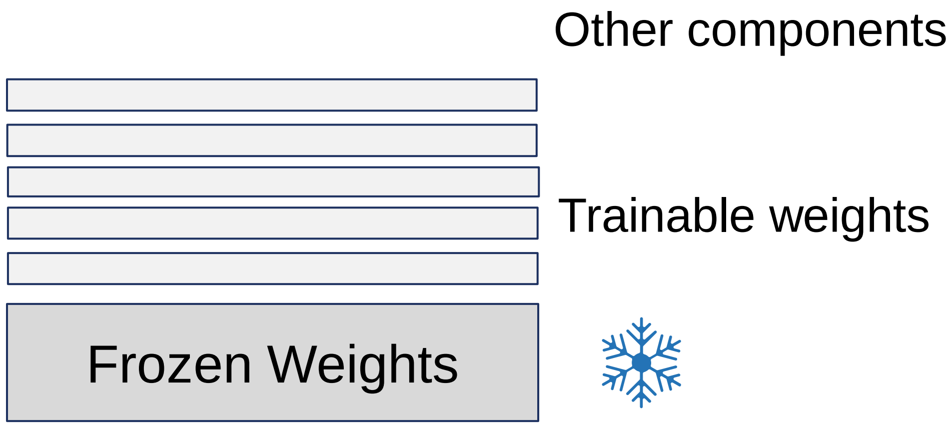
The Natural Language Decathlon: Multitask Learning as Question Answering, 2018
Raffel, et al., Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, 2020

Parameter Efficient Fine-tuning (PEFT)

- Less prone to catastrophic forgetting

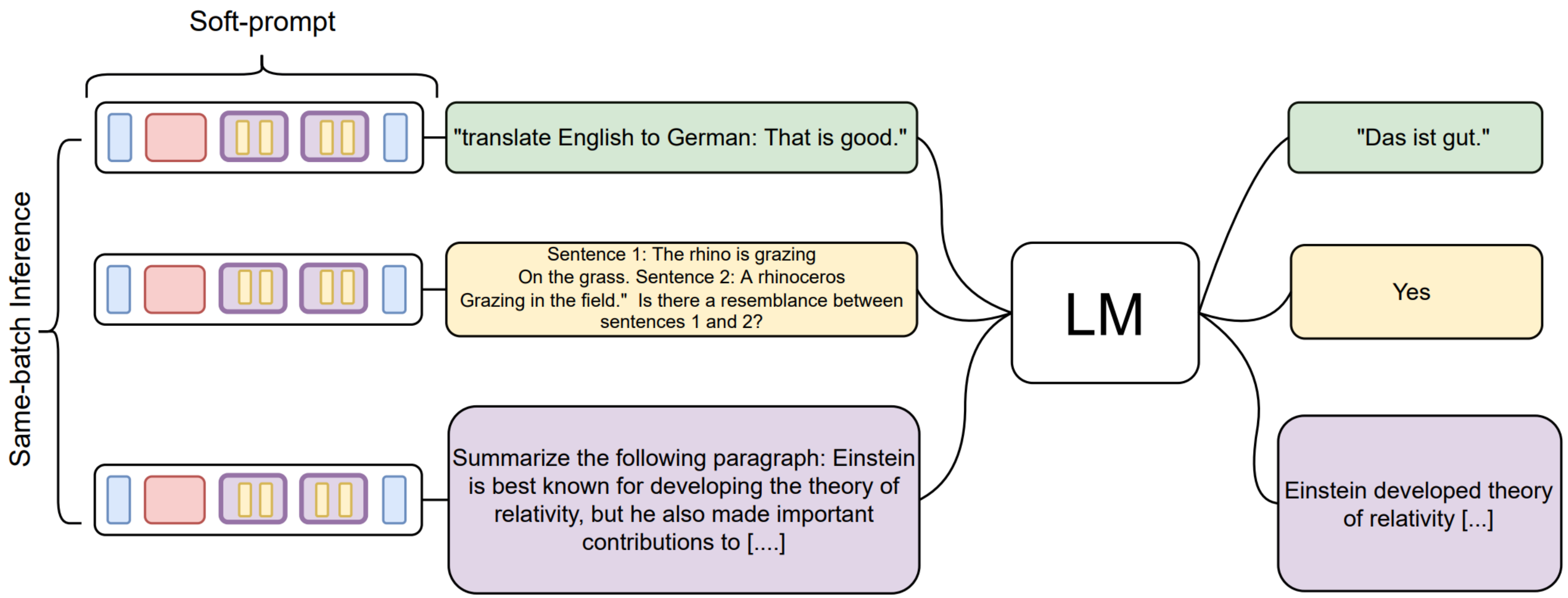


LLM with additional
layers for PEFT



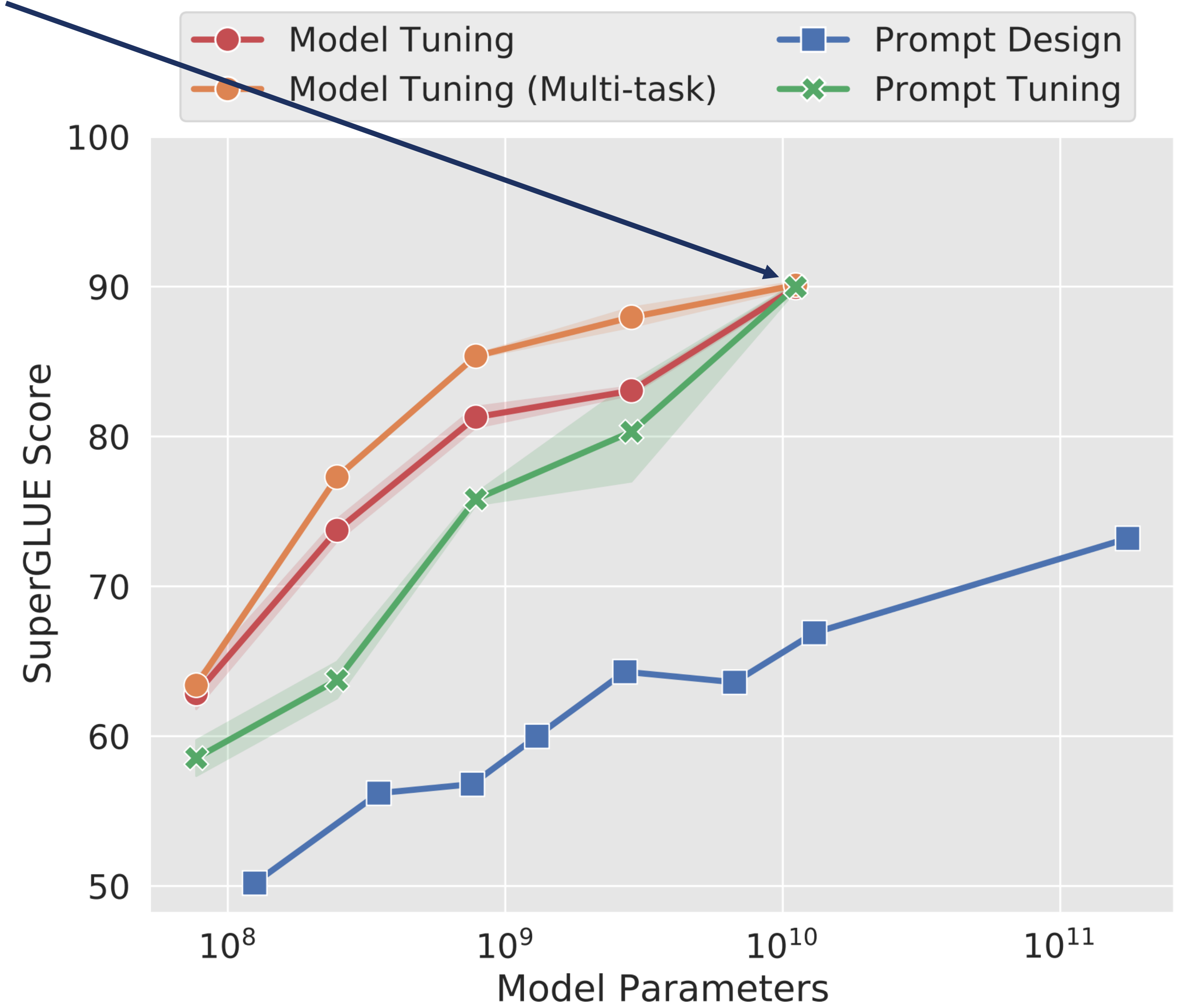
Parameter Efficient Fine-tuning

- Prompt tuning
  - Additional embedding with tunable parameter
  - <0.1% total parameter
  - Avoids forgetting
  - Scales efficiently
  - Performance tradeoff



Performance of Prompt Tuning

- Prompt tuning can be as effective as full fine tuning for large models.



Lester et al., The Power of Scale for Parameter-Efficient Prompt Tuning, 2021

Part 2: Deep Learning for NLP

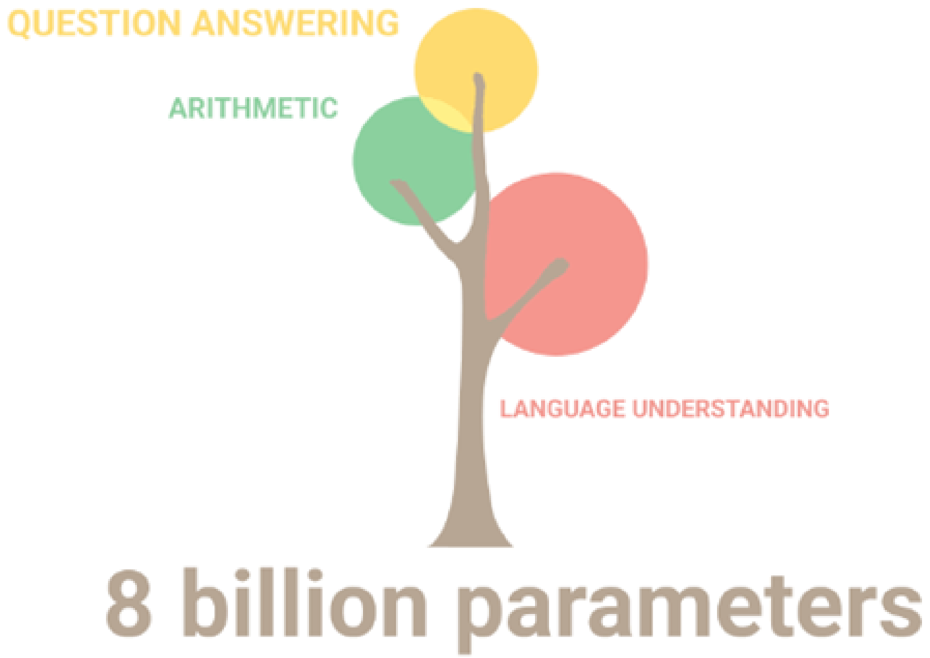
Agenda

- **Background**
  - Introduction to NLP
  - Language modeling
- **Model architectures**
  - Transformer architecture
  - Encoder, decoder, encoder-decoder
  - Pre-training and fine-tuning

Large language models (LLMs)
- Scaling LMs to LLMs
- Prompt engineering
- In context learning
- Instruction tuning

Scaling Up!

- Performance improves across tasks while also unlocking new capabilities.



Credit: Google AI Blog

Why LLMs?

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

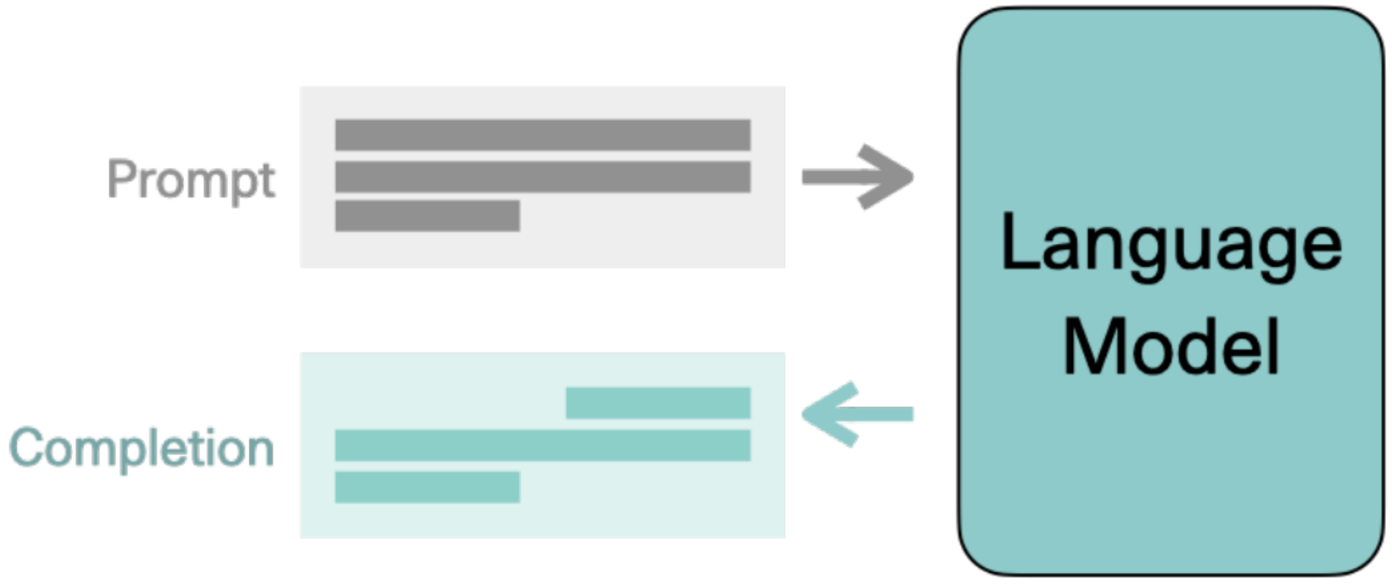
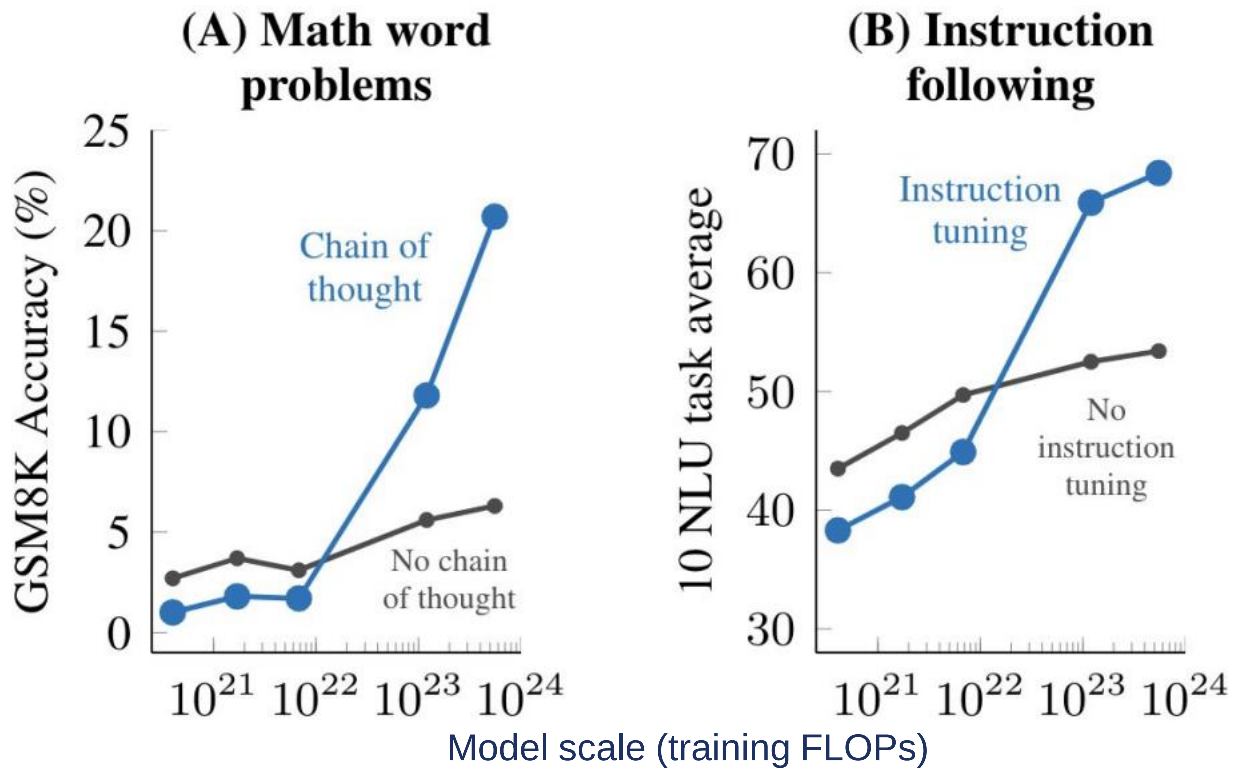


Image credit: Jay Alammar

- Emergent properties in LLMs



(Wei et al., 2022)

Wei, Jason, et al. "Emergent abilities of large language models.", 2022.

Part 2: Deep Learning for NLP

Agenda

- **Background**
  - Introduction to NLP
  - Language modeling
- **Model architectures**
  - Transformer architecture
  - Encoder, decoder, encoder-decoder
  - Pre-training and fine-tuning
- **Large language models (LLMs)**
  - Scaling LMs to LLMs
  - Prompt engineering
  - In context learning
  - Instruction tuning

Prompt Engineering

- Practice of developing and optimizing prompts to efficiently use LLMs for a variety of applications.

Prompt

Analyze the following bar chart in one paragraph.
Chart Data: Characteristics| Public| Private & 2021| 149| 209| & 2020 | 146| 202| 2019| 154 | 250 & …
Chart title: Number of public and private hospitals in Malaysia from 2017 to 2021

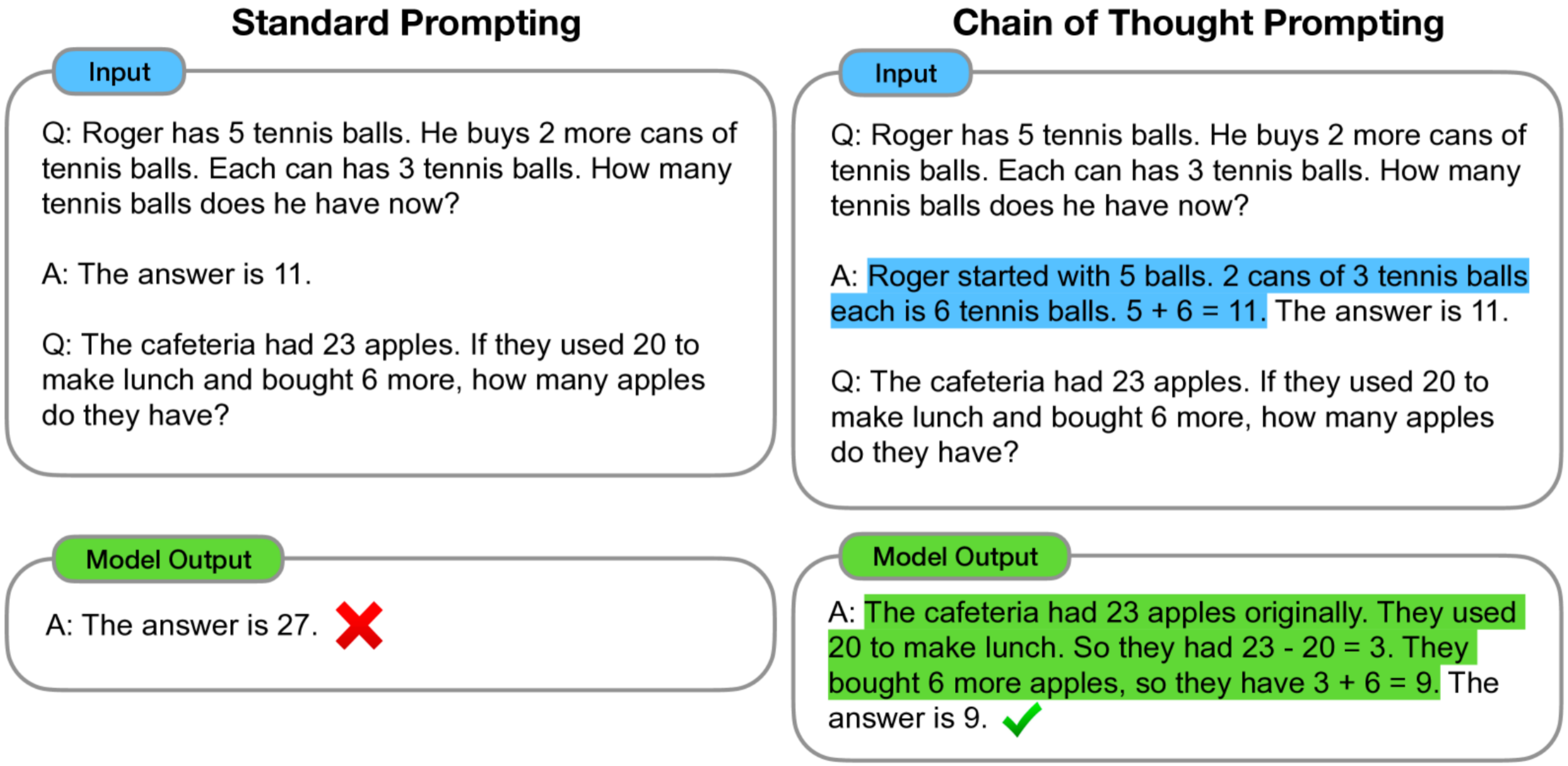
LLM

Model's response

In 2021, there were around 146 government hospitals and 209 private licensed hospitals in Malaysia. During the COVID-19 pandemic, the Malaysian hospitals were prepared by the government to accommodate infected patients by increasing bed numbers…

Chain-of-Thought Prompting

- Models are better at getting the right answer when they first output text that explains the reason for the answer.



Wei, Jason, et al. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models ", 2022.

Part 2: Deep Learning for NLP

Agenda

- **Background**
    - Introduction to NLP
    - Language modeling
- **Model architectures**
    - Transformer architecture
    - Encoder, decoder, encoder-decoder
    - Pre-training and fine-tuning
- **Large language models (LLMs)**
    - Scaling LMs to LLMs
    - Prompt engineering
    - In context learning
    - Instruction tuning

In-Context Learning: An Emergent Ability of LLMs

- LLMs perform a task just by conditioning on input-output examples, with no fine-tuning.

Sentiment analysis task

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. //

Language Model → Positive

Document classification

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. //

Language Model → Finance

In-Context Learning: Few-shot Inference

Prompt (Zero Shot)

```
Classify this review:
I loved this movie!
Sentiment:
```

Context Window
(few thousand words)

Prompt (One Shot)

```
Classify this review:
I loved this movie!
Sentiment: Positive

Classify this review:
I don't like this
chair.
Sentiment:
```

Prompt (Few Shot)

```
Classify this review:
I loved this movie!
Sentiment: Positive

Classify this review:
I don't like this
chair.
Sentiment: Negative

Classify this review:
Who would use this
product?
Sentiment:
```

Part 2: Deep Learning for NLP

Agenda

- **Background**
  - Introduction to NLP
  - Language modeling
- **Model architectures**
  - Transformer architecture
  - Encoder, decoder, encoder-decoder
  - Pre-training and fine-tuning
- **Large language models (LLMs)**
  - Scaling LMs to LLMs
  - Prompt engineering
  - In context learning
  - Instruction tuning

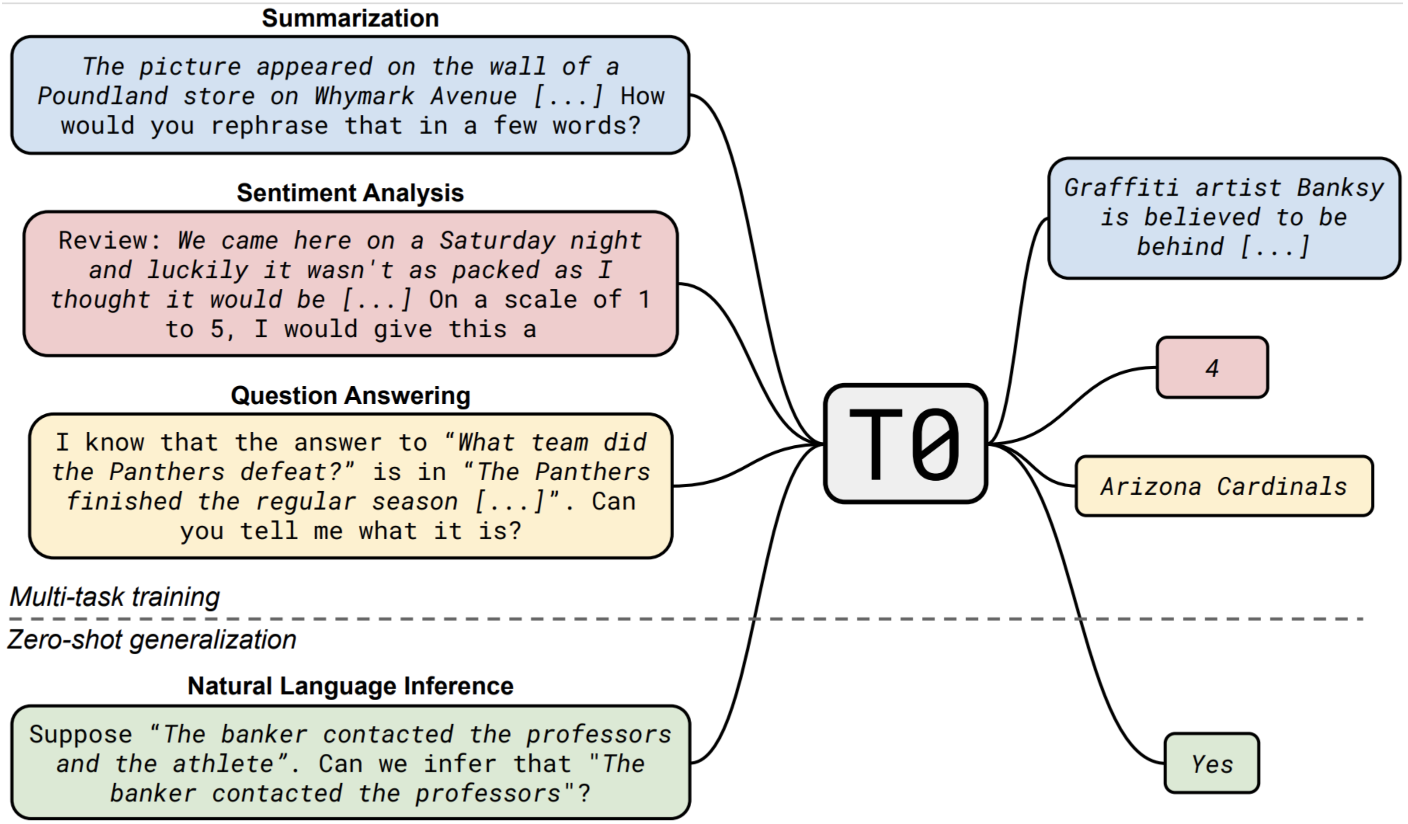
Instruction Tuning

Why?

- LLMs (e.g., GPT-3) as a few-shot learner
  - Perform in-context learning
  - Prompts (examples) trigger few-shot capabilities
- But there’s a mismatch between self-supervision and inference-time use
  - Self-supervision is cheap but lacks form to meaning (grounding)
  - Self-supervision doesn’t teach to follow instructions
  - Tasks generally need more direct supervision
  - Full-scale fine-tuning of LLMs on each task can be expensive and infeasible

Instruction Tuning

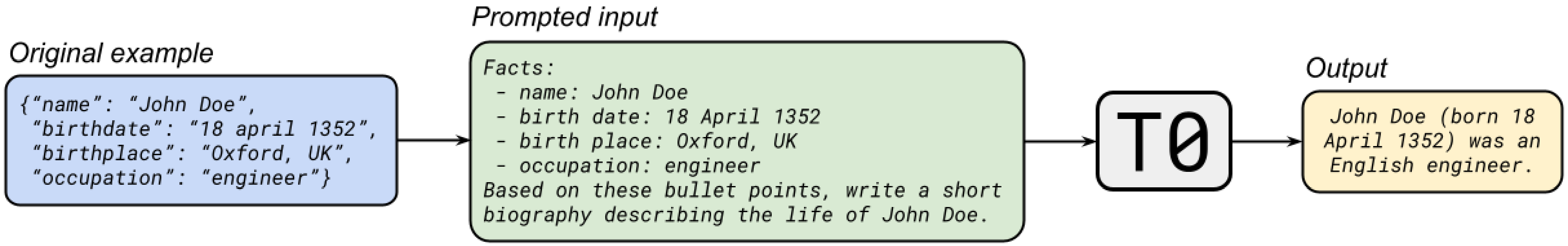
- Train LLMs to follow task instructions
  - ex: T0, FLAN, FLAN-T5, InstructGPT, ChatGPT
- "if we explicitly train a LLM on a massive mixture of diverse NLP tasks, would it generalize to unseen NLP tasks?"
  - Yes!



Sanh et al. Multitask Prompted Training Enables Zero-shot Task Generalization ", 2022.

Instruction Tuning

- The key is to reformulate any task into a text-to-text format as if we are asking another person for the answer to the task



Sanh et al. Multitask Prompted Training Enables Zero-shot Task Generalization ", 2022.

Instructional Tuning

- How/why does it work?
  - Instructions provide a way to map pure textual forms to physical meaning/intention (grounding)
  - Provide more general and direct information than examples
    - Words like summarize, translate, convert, answer provide intents
- A way to unlock different abilities that are already there
- Adjusts LLMs towards different skill sets:
  - Answer questions, generate code, chat
  - Become more honest, helpful and harmless
- Aligns LLMs to human instructions

Instructional Tuning

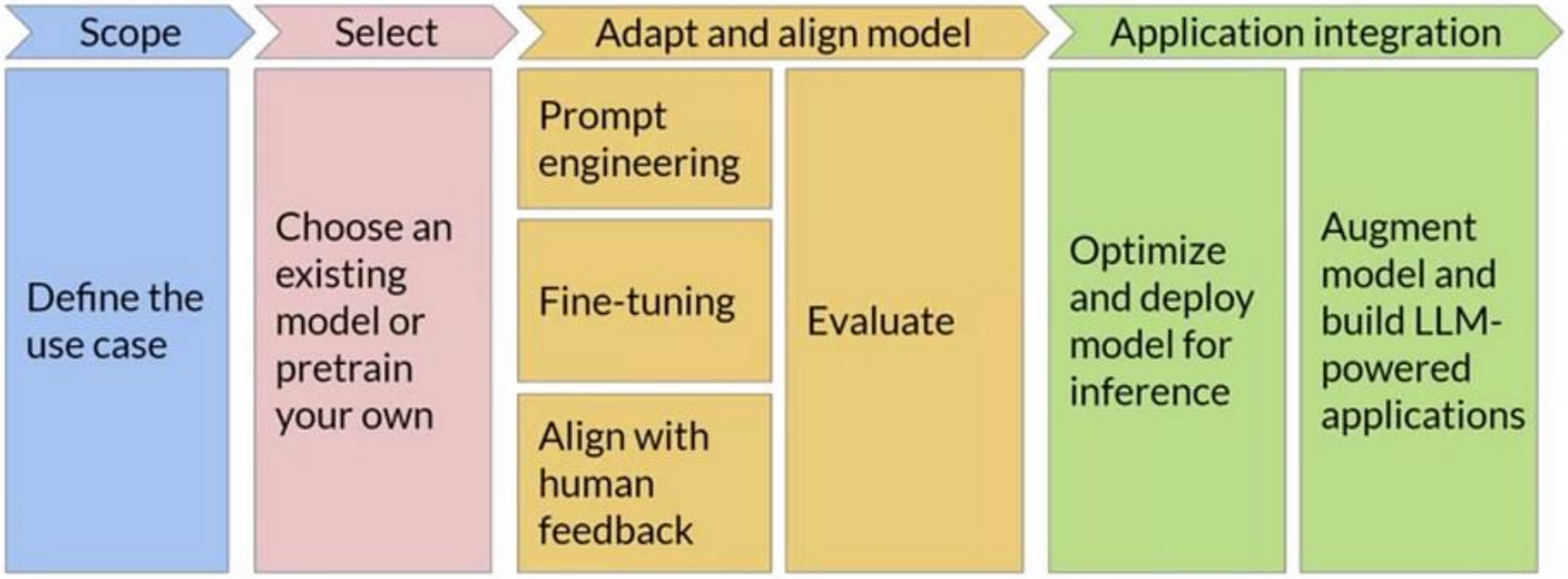
- What is crucial for generalization?
  - Diverse set of tasks
  - Diverse set of prompts per task

- Examples
  - T0 collected 2,000 prompts for 170 English datasets (8-20 prompts per task)
  - FLAN-T5 used 1800 tasks (including Chain-of-Thought)
  - InstructGPT 77K prompts in different stages
  - ChatGPT $\rightarrow$??

Part 2: Deep Learning for NLP

- Agenda
  - Introduction to NLP
  - Language modeling
  - Model Architectures
    - Transformer architecture
    - Encoder, decoder, encoder-decoder
    - Pre-training and fine-tuning
  - Large language models (LLMs)
    - Scaling LMs to LLMs
    - Prompt engineering
    - In context learning
    - Instruction tuning

Generative AI Project Life Cycle



Tutorial Overview

- **Part 1**: Introduction [15 mins]
  - Why NLP + Vis?
  - An overview of NLP + Vis Research
  - An overview of the tutorial
- **Part 2**: Deep Learning for NLP [50 mins]
  - Background
  - Large language models (LLMs)
- **Part 3**: NLP4Vis applications [50 mins]
- **Part 4**: Future challenges and research opportunities [25 mins]