Demystifying ChatGPT

A shallow deep-dive into LLMs

SKYLYZE DSA

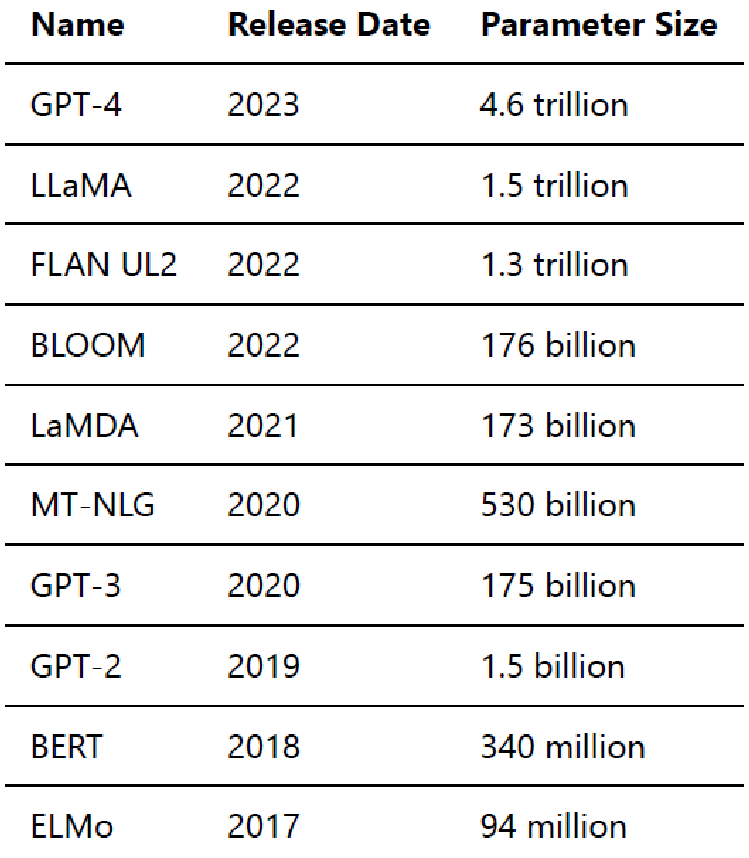
DSA Daten- und Systemtechnik GmbH | SKYLYZE | Roland Stoffel | Pascal Rößner

AGENDA

- Introduction ChatGPT aka LLM (Lecture: 15min)
  - Defintion and Use-Cases
  - Theoretical Background
  - Project Lifecycle
- From Scratch (Hands-On Session: 35-45min)
  - Train your own model
- Challenges of LLMS and Generative AI (Lecture: 10min)
  - Infrastructure and Costs
  - Transfer to Image Generation
- Experimentation (Hands-On Session: 20min)
  - Have Fun with Stable Diffusion

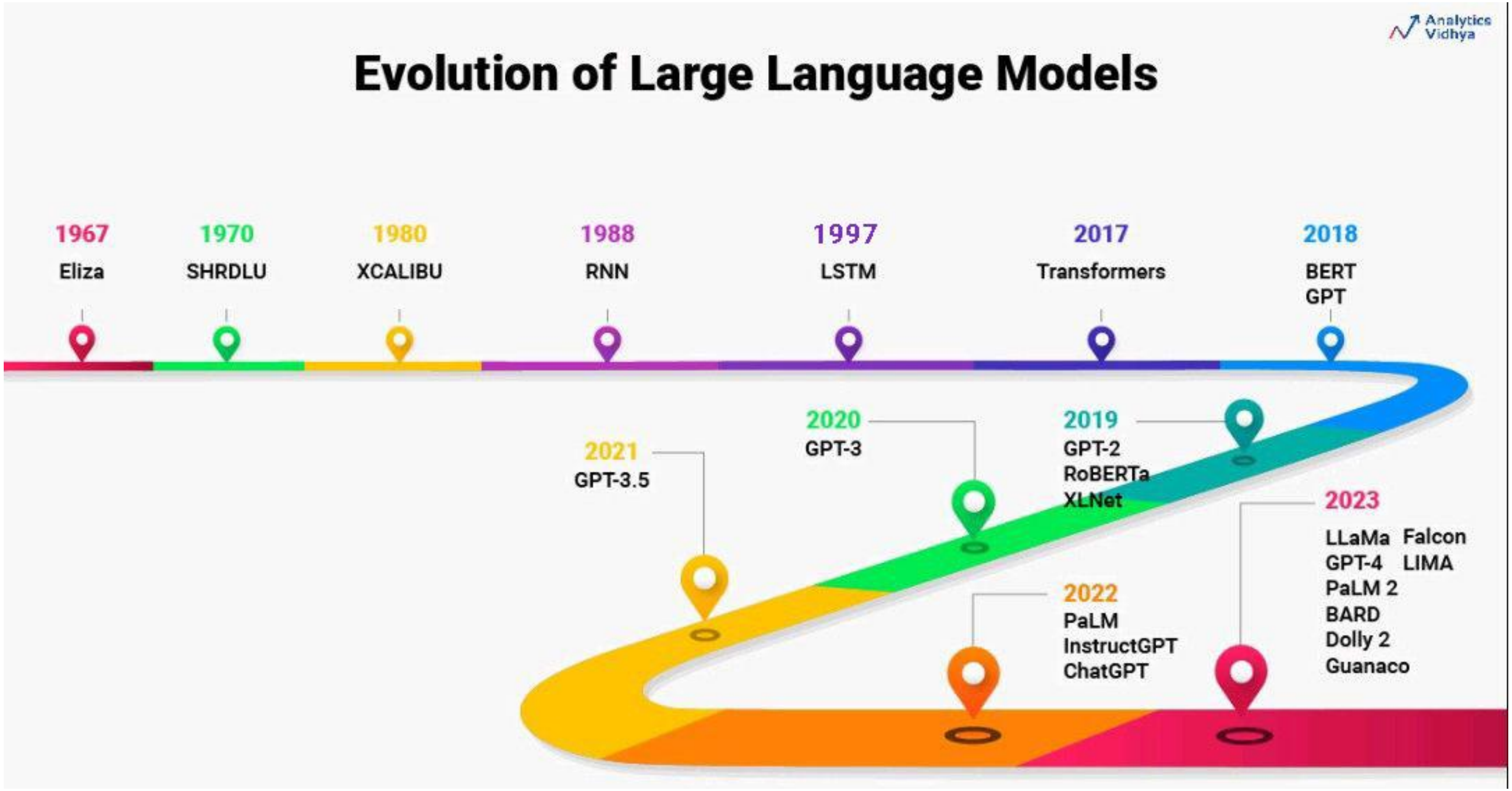
WHAT IS A LLM

- ChatGPT is a so-called Large Language Model (LLM)
- Large Language Models are deep learning models trained on huge datasets to perform NLP tasks.
- Its core objective is to learn and understand human languages precisely.
- Large Language Models enable the machines to interpret languages just like the way we, as humans, interpret them.

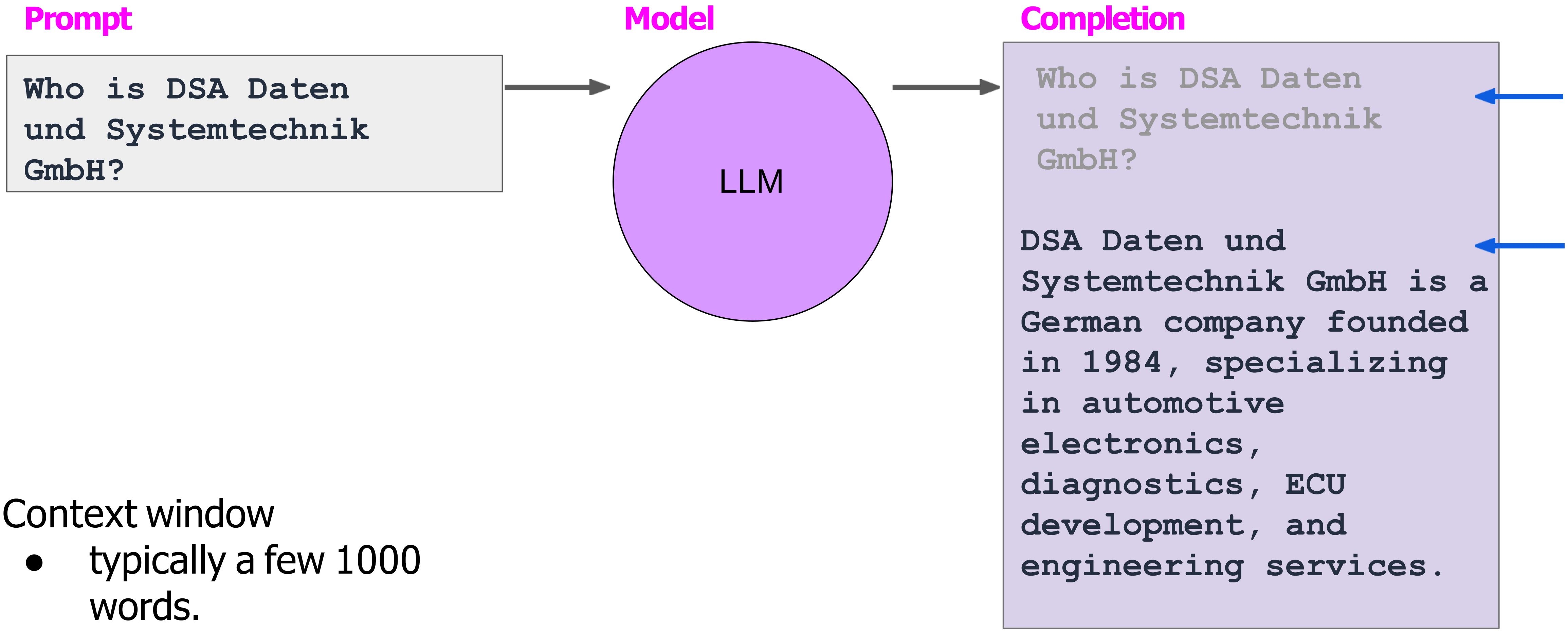


Name	Release Date	Parameter Size
GPT-4	2023	4.6 trillion
LLaMA	2022	1.5 trillion
FLAN UL2	2022	1.3 trillion
BLOOM	2022	176 billion
LaMDA	2021	173 billion
MT-NLG	2020	530 billion
GPT-3	2020	175 billion
GPT-2	2019	1.5 billion
BERT	2018	340 million
ELMo	2017	94 million

HISTORICAL EVOLUTION



PROMPTS AND COMPLETIONS



Prompt

```
Who is DSA Daten
und Systemtechnik
GmbH?
```

Model

LLM

Completion

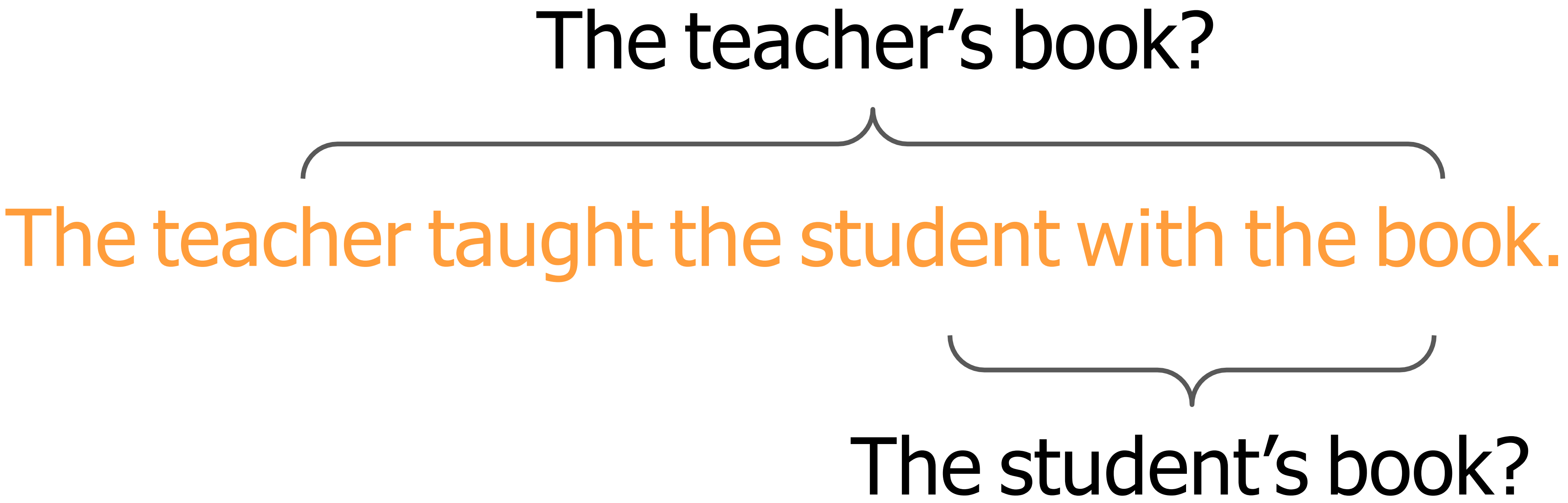
```
Who is DSA Daten
und Systemtechnik
GmbH?

DSA Daten und
Systemtechnik GmbH is a
German company founded
in 1984, specializing
in automotive
electronics,
diagnostics, ECU
development, and
engineering services.
```

Context window

- typically a few 1000 words.

UNDERSTANDING LANGUAGE CAN BE CHALLENGING



The teacher’s book?

The teacher taught the student with the book.

The student’s book?

TRANSFORMER ARCHITECTURE

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

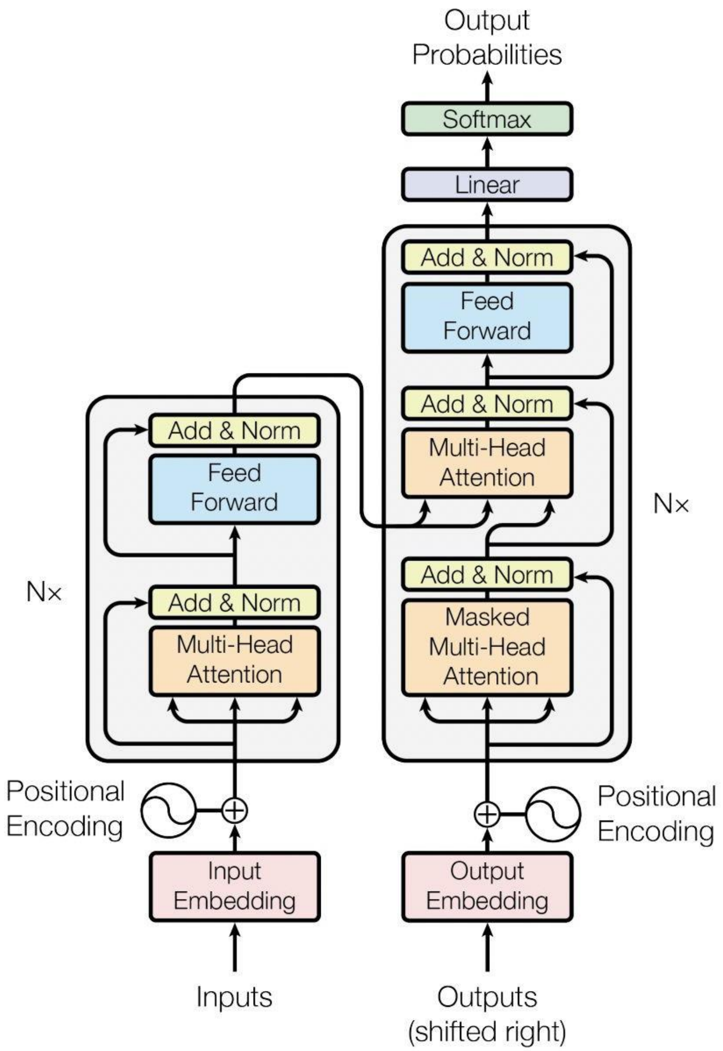
Aidan N. Gomez[*] [†] University of Toronto aidan@cs.toronto.edu

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



TRANSFORMER ARCHITECTURE

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

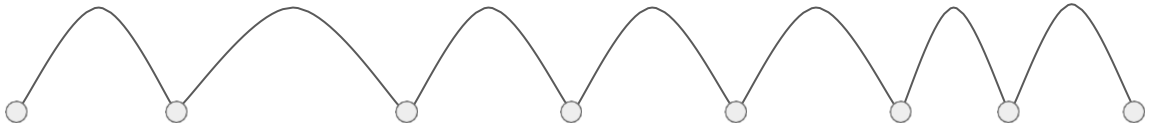
Illia Polosukhin[*] [‡]
illia.polosukhin@gmail.com

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

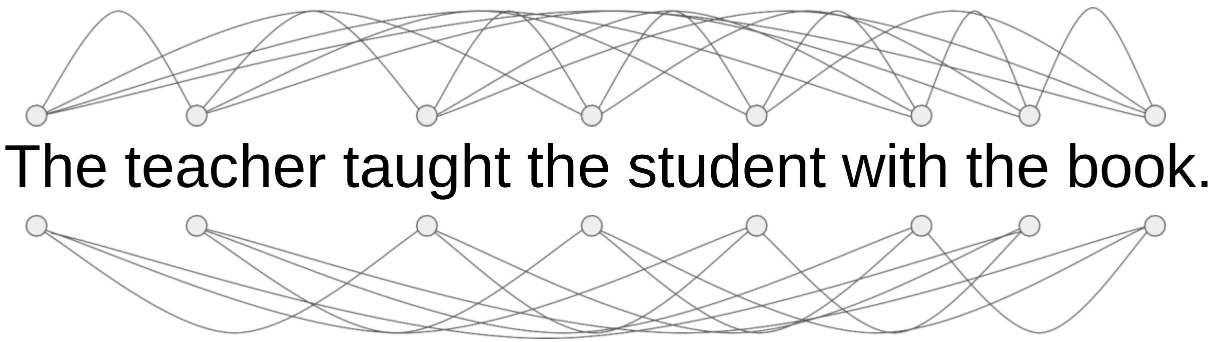
- Scale efficiently
- Parallelprocess
- Attention to input meaning

ATTENTION IS ALL YOU NEED



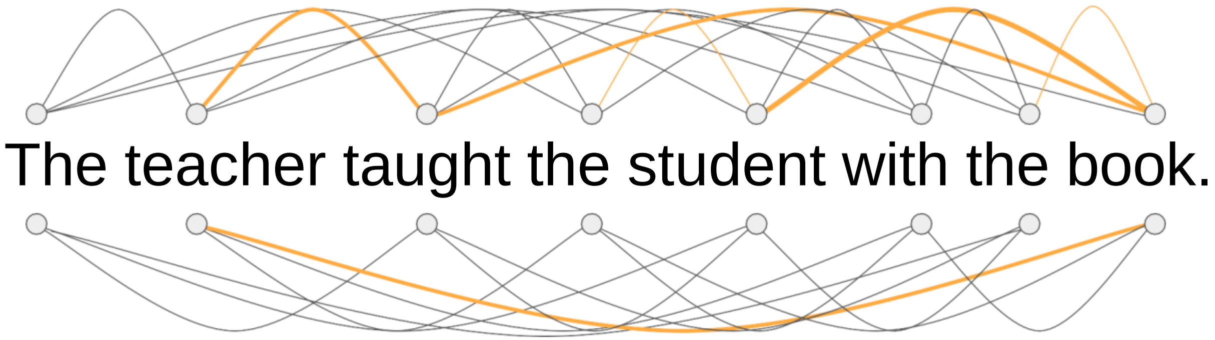
The teacher taught the student with the book.

ATTENTION IS ALL YOU NEED



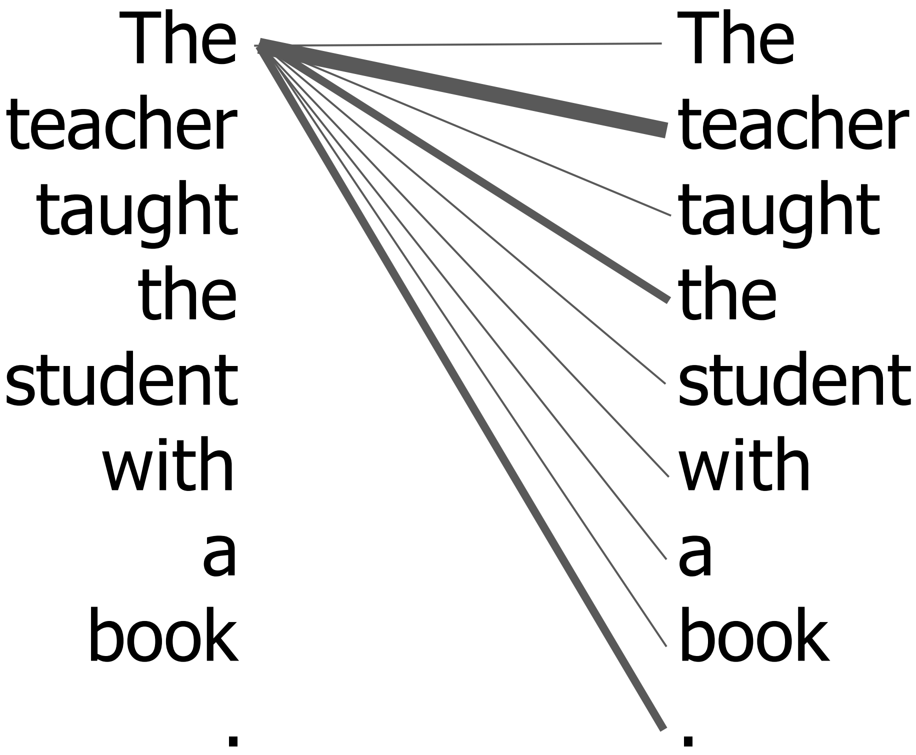
The teacher taught the student with the book.

ATTENTION IS ALL YOU NEED



The teacher taught the student with the book.

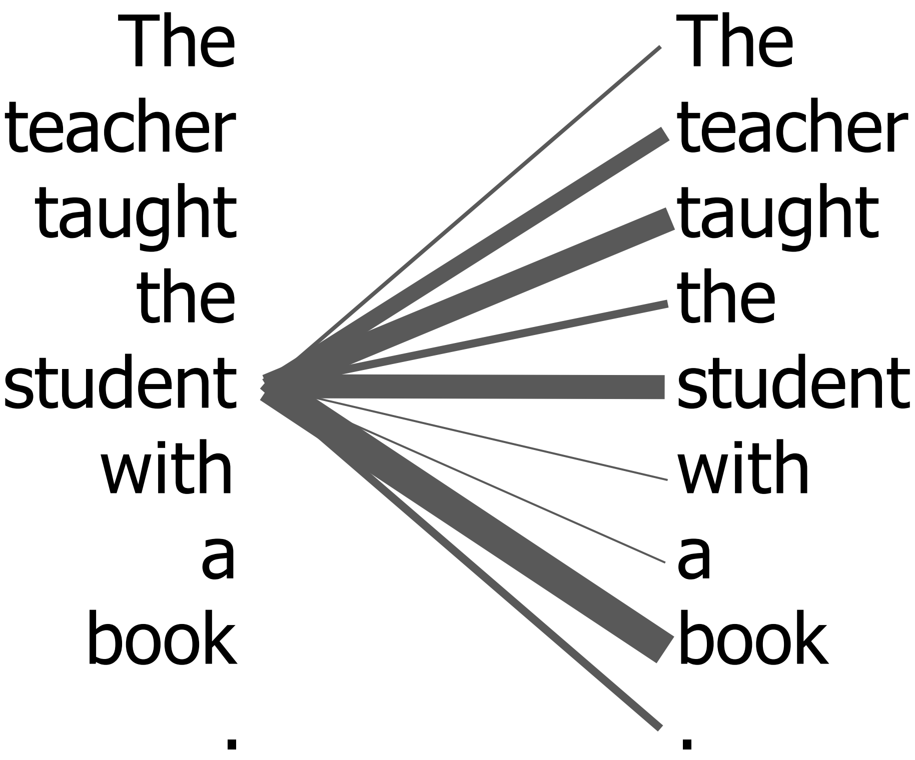
SELF-ATTENTION



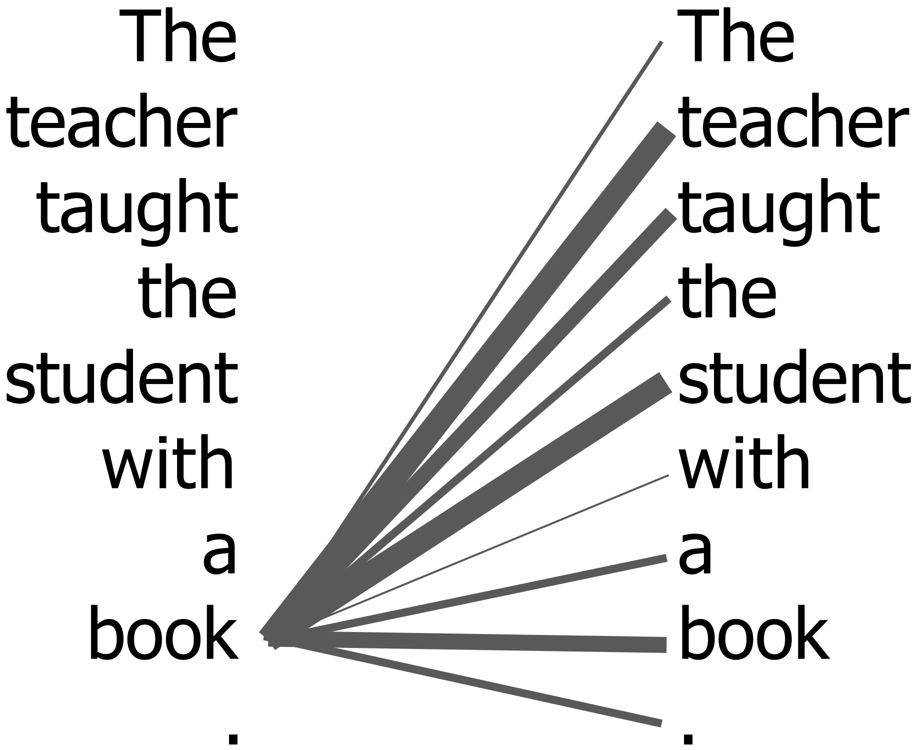
The
teacher
taught
the
student
with
a
book
.

The
teacher
taught
the
student
with
a
book
.

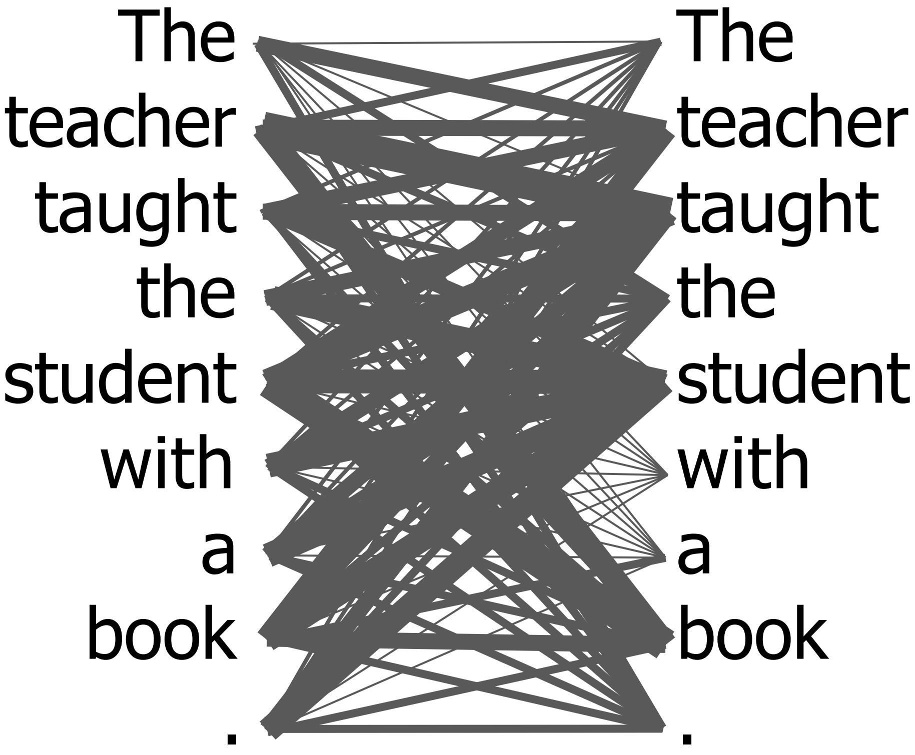
SELF-ATTENTION



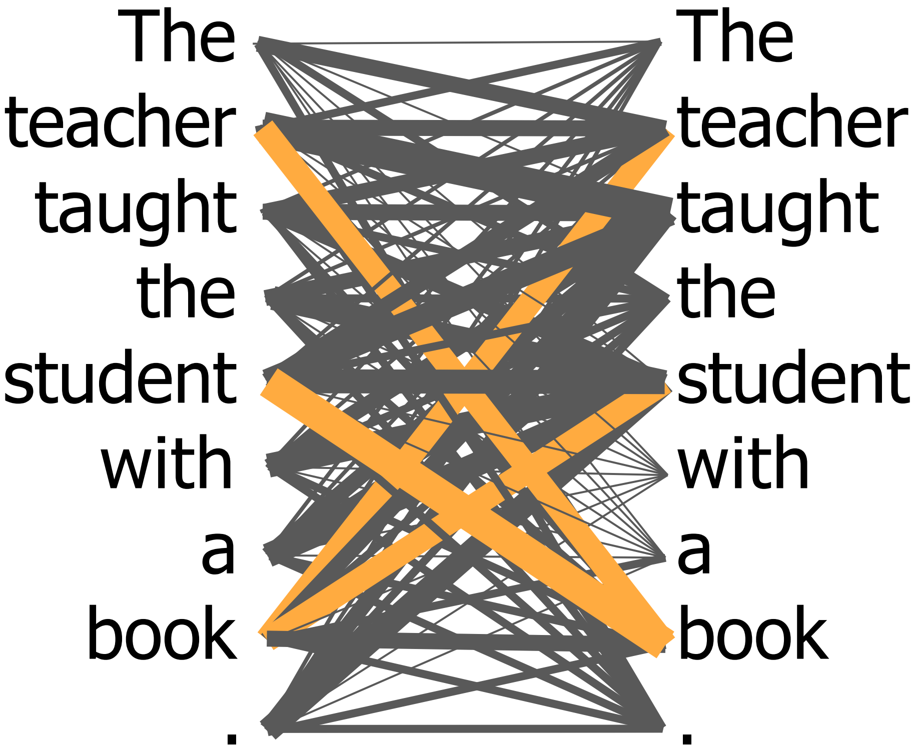
SELF-ATTENTION



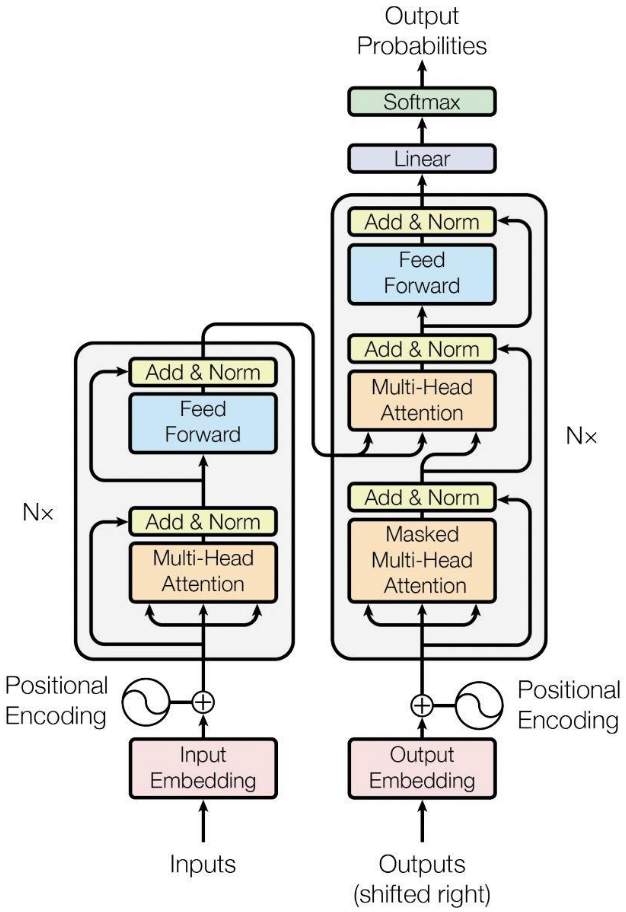
SELF-ATTENTION



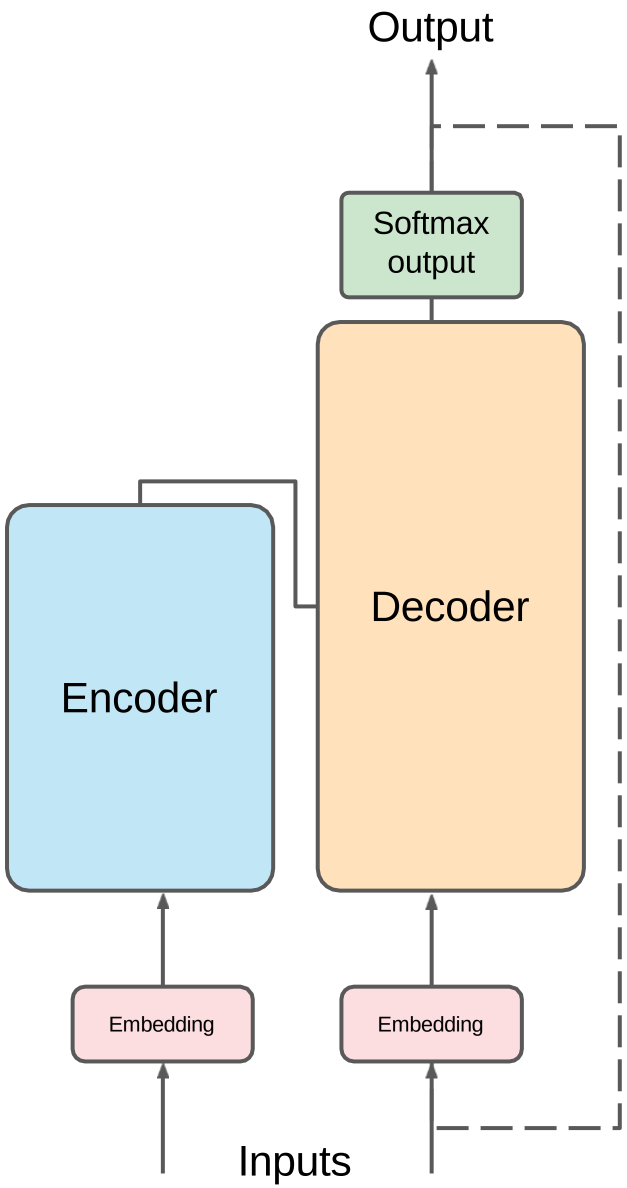
SELF-ATTENTION



TRANSFORMERS

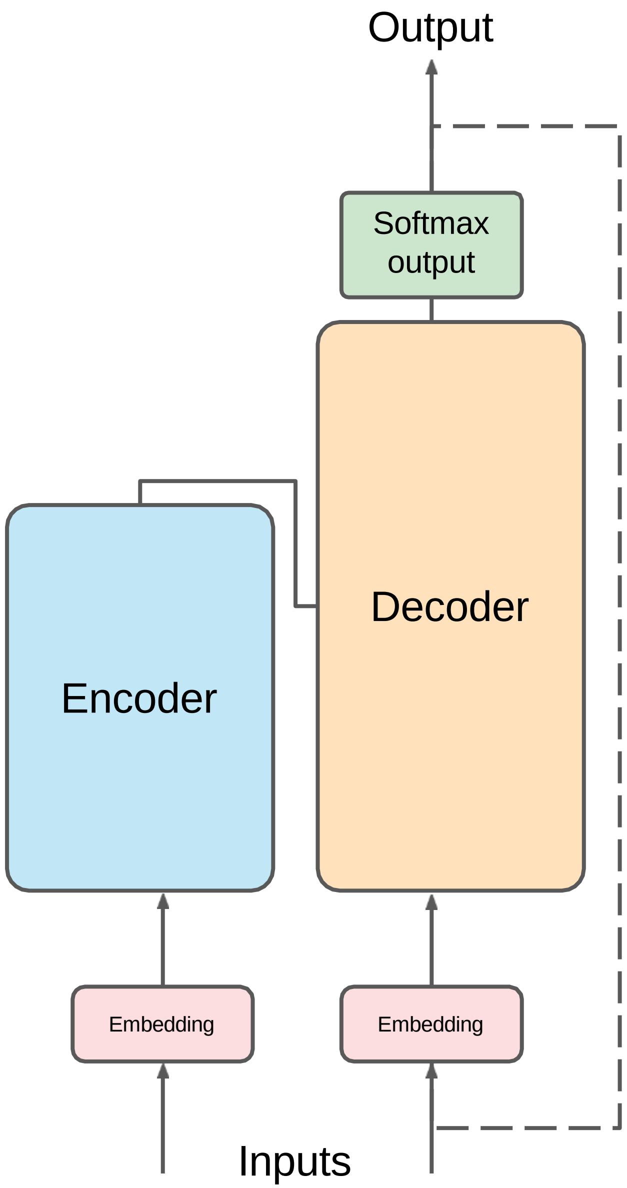


TRANSFORMERS



TRANSFORMERS

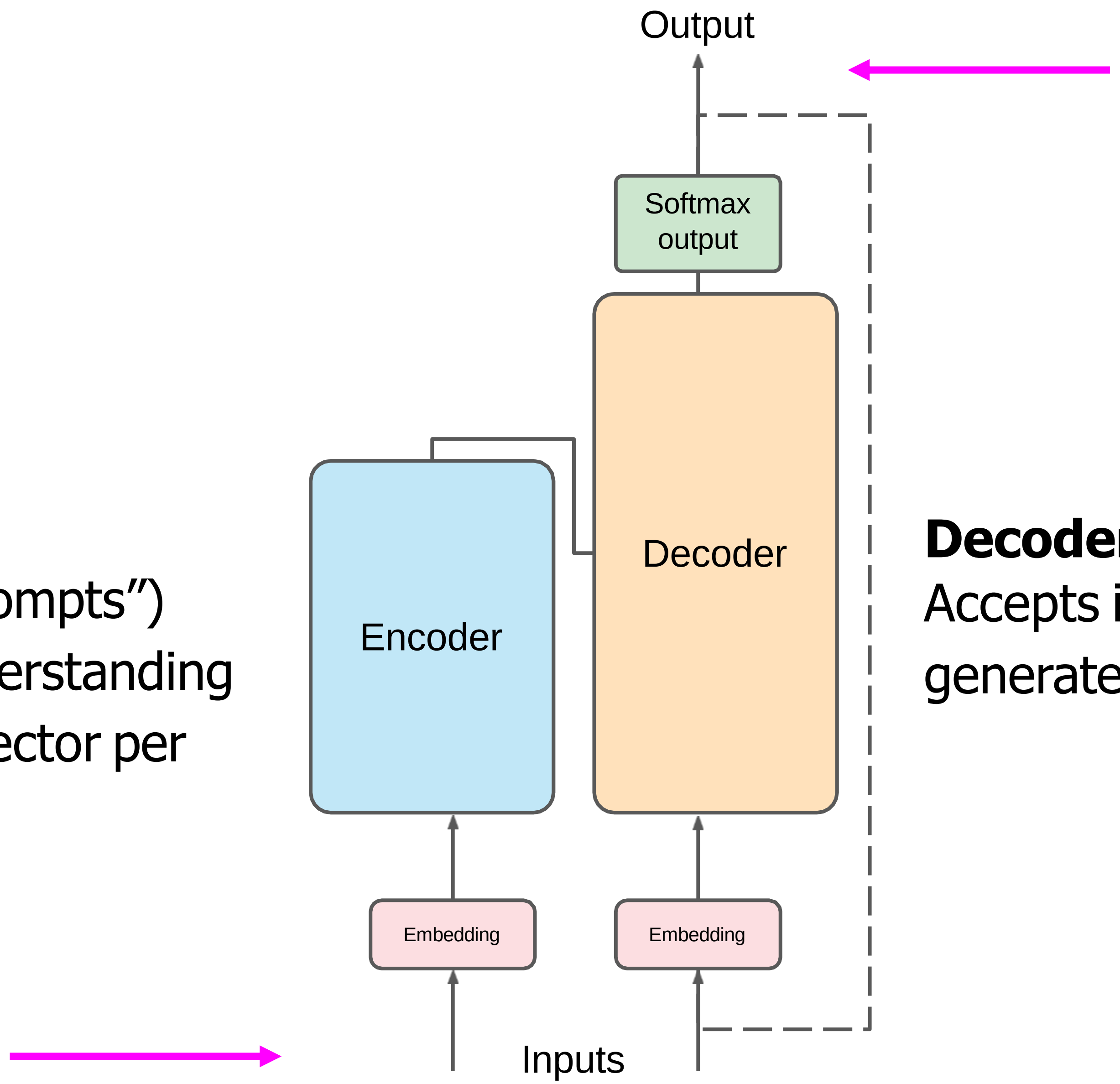
Encoder
Encodes inputs ("prompts") with contextual understanding and produces one vector per input token.



Decoder
Accepts input tokens and generates new tokens.

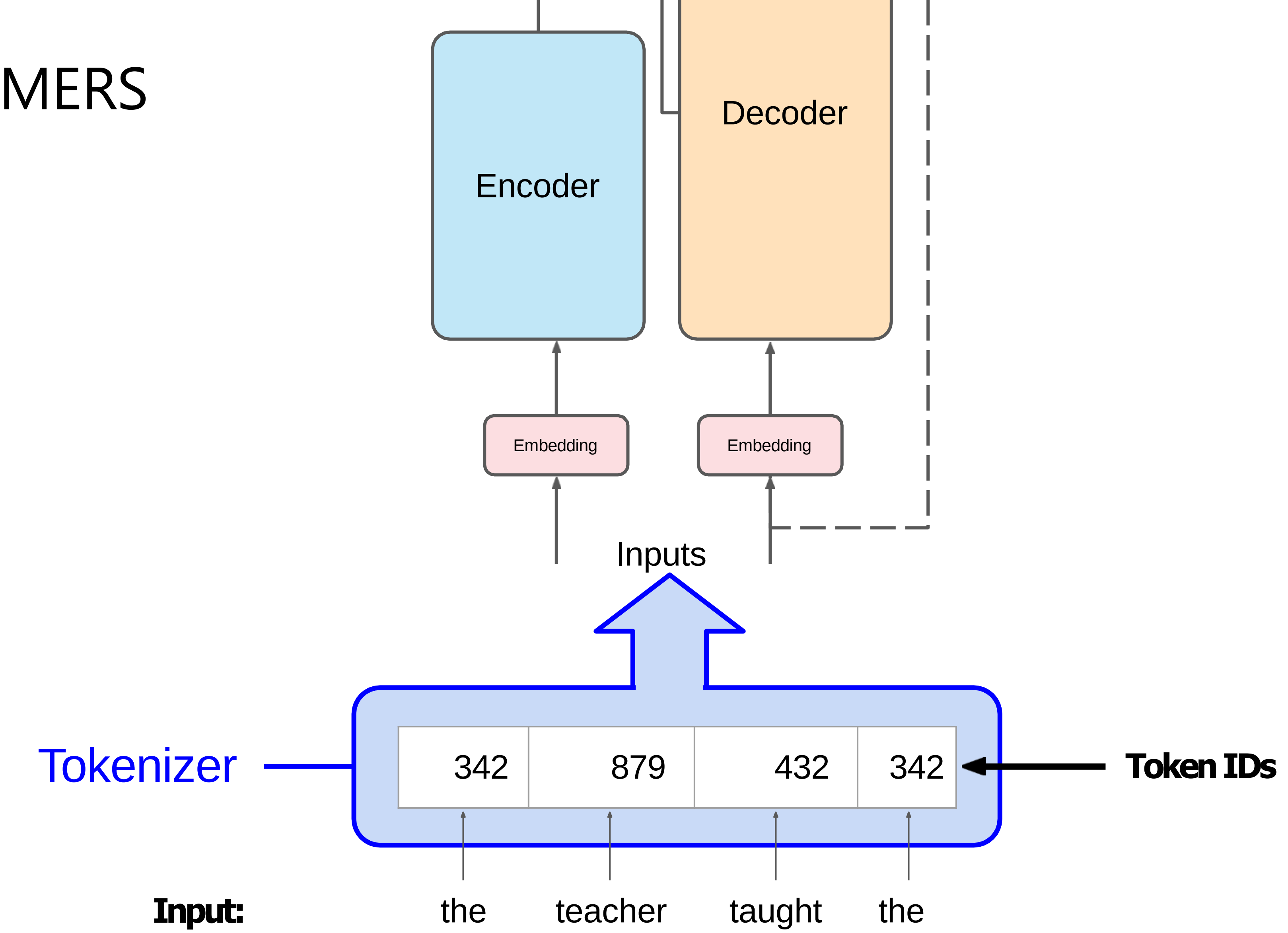
TRANSFORMERS

Encoder
Encodes inputs ("prompts") with contextual understanding and produces one vector per input token.

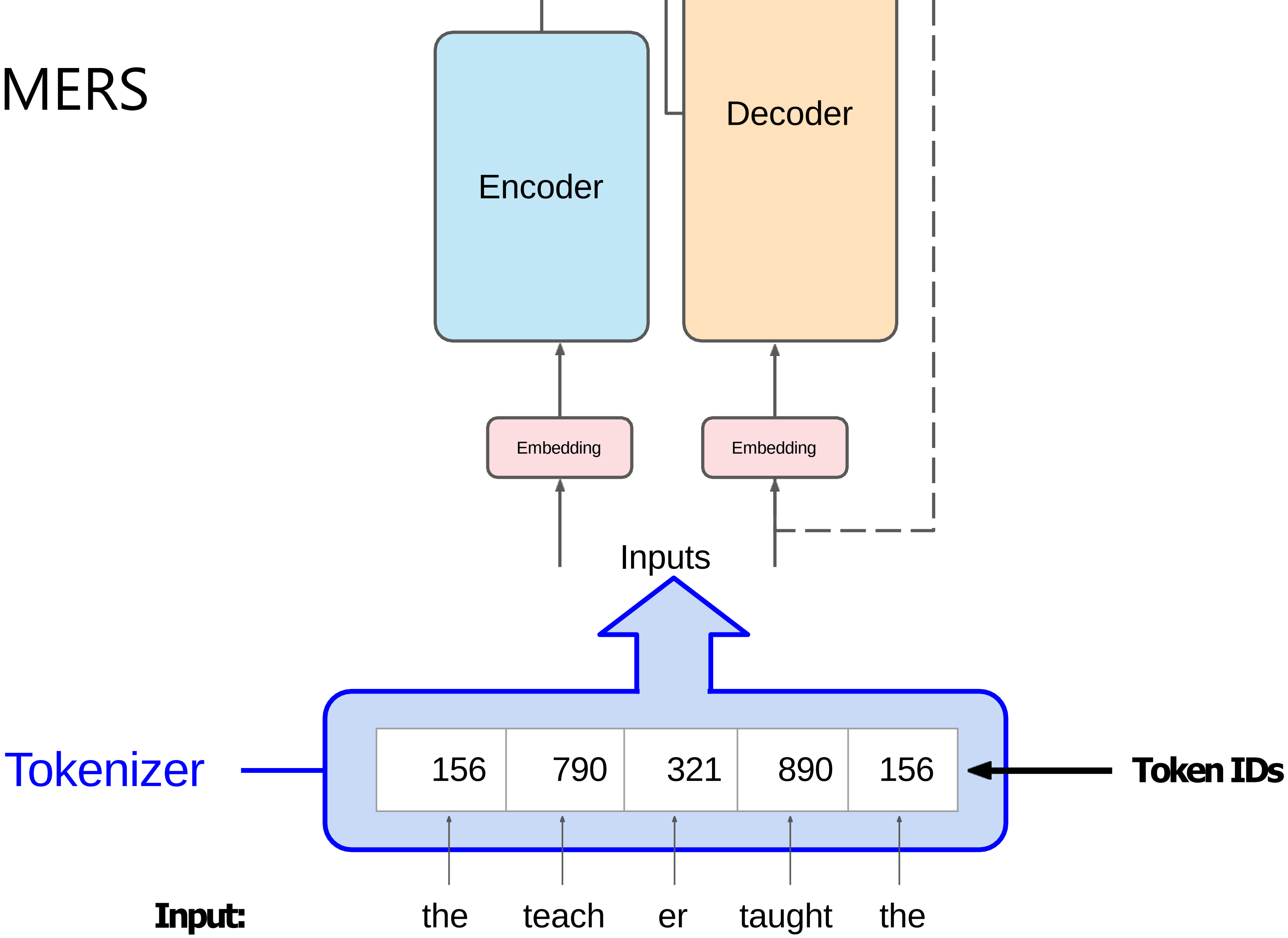


Decoder
Accepts input tokens and generates new tokens.

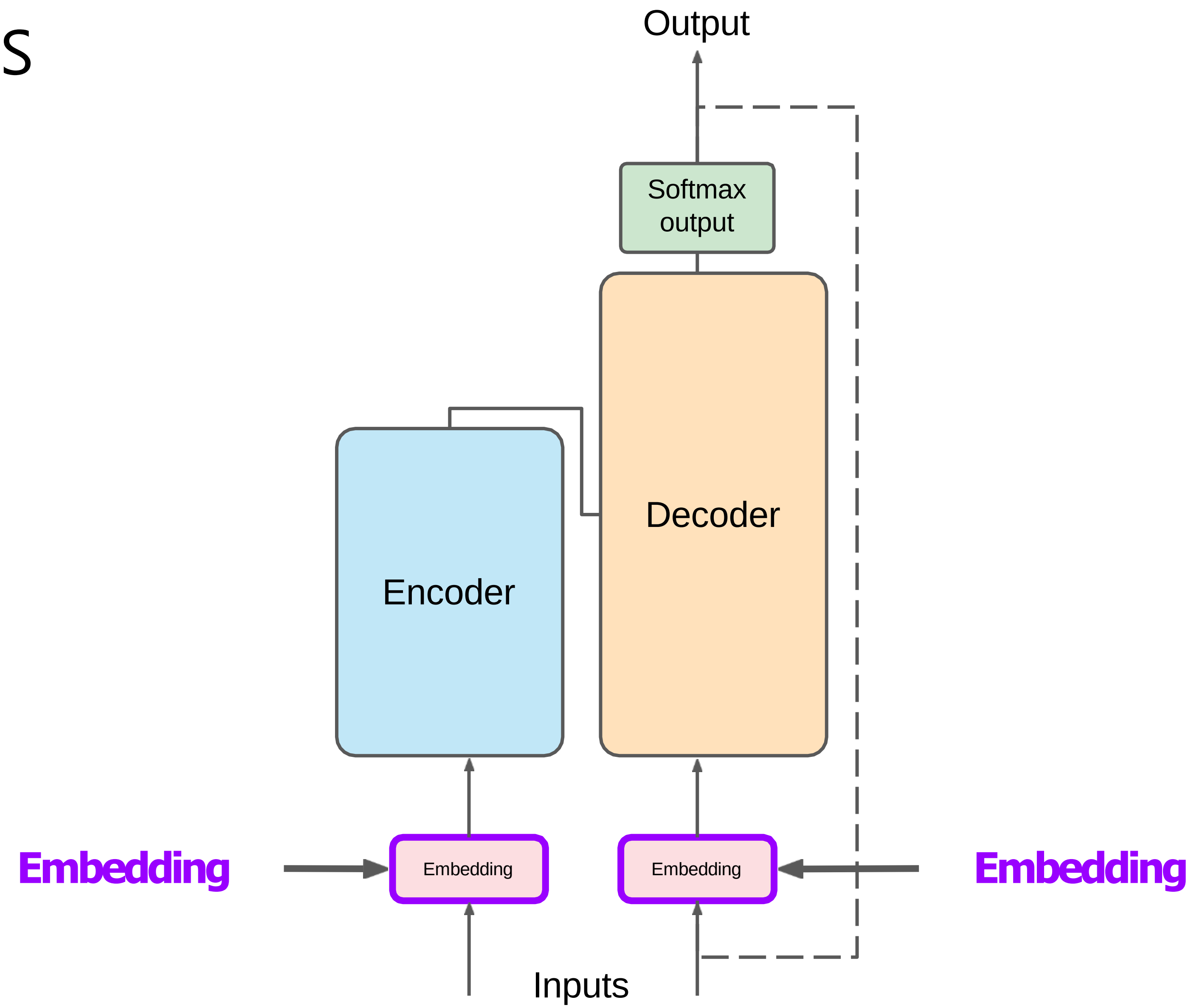
TRANSFORMERS



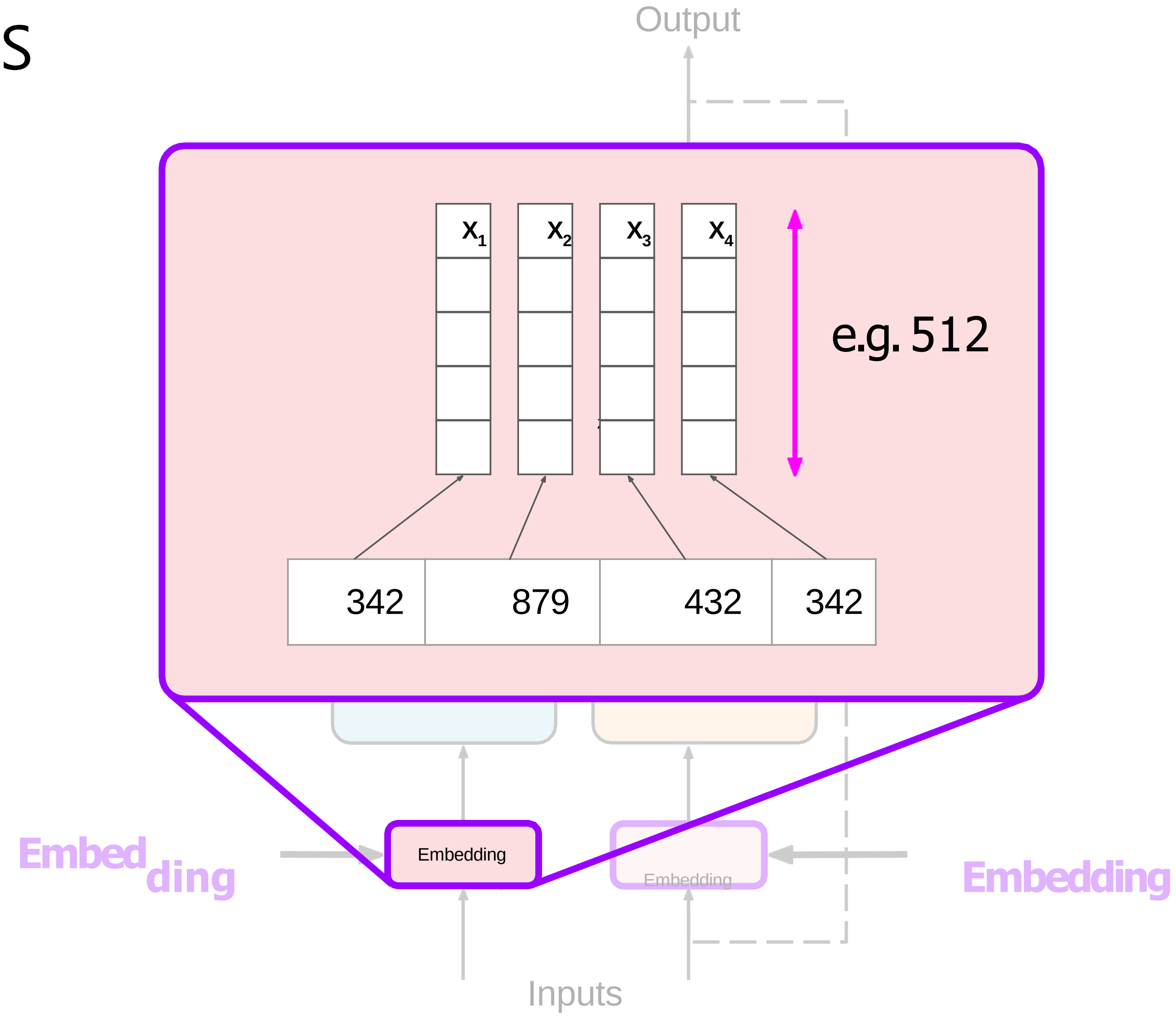
TRANSFORMERS



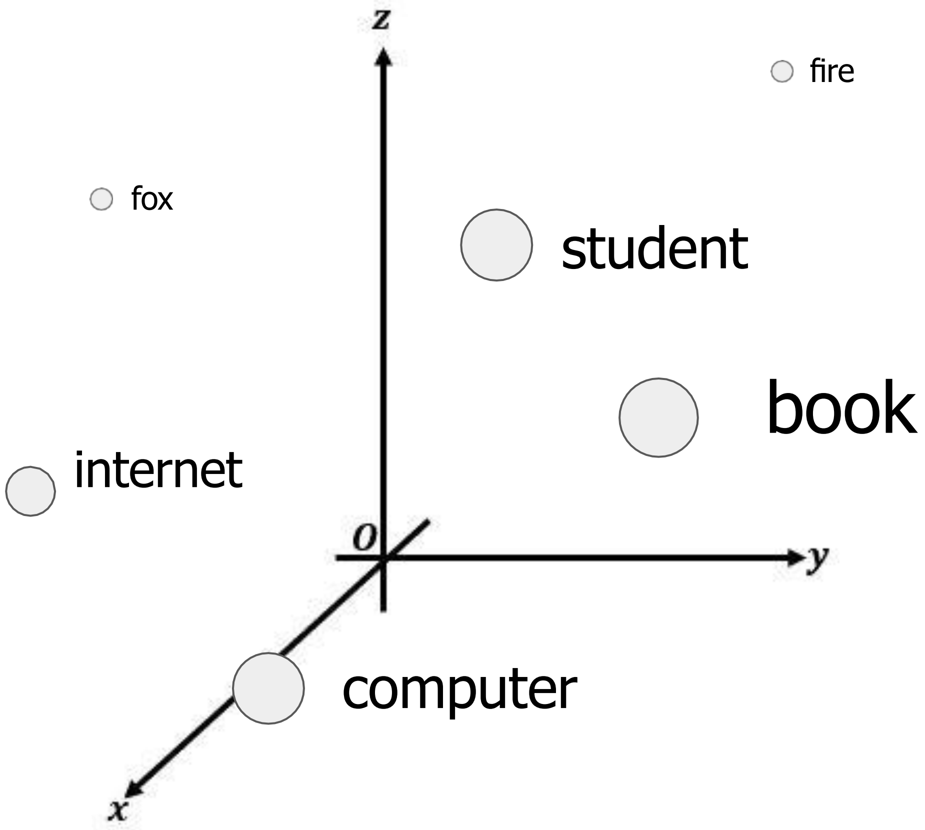
TRANSFORMERS



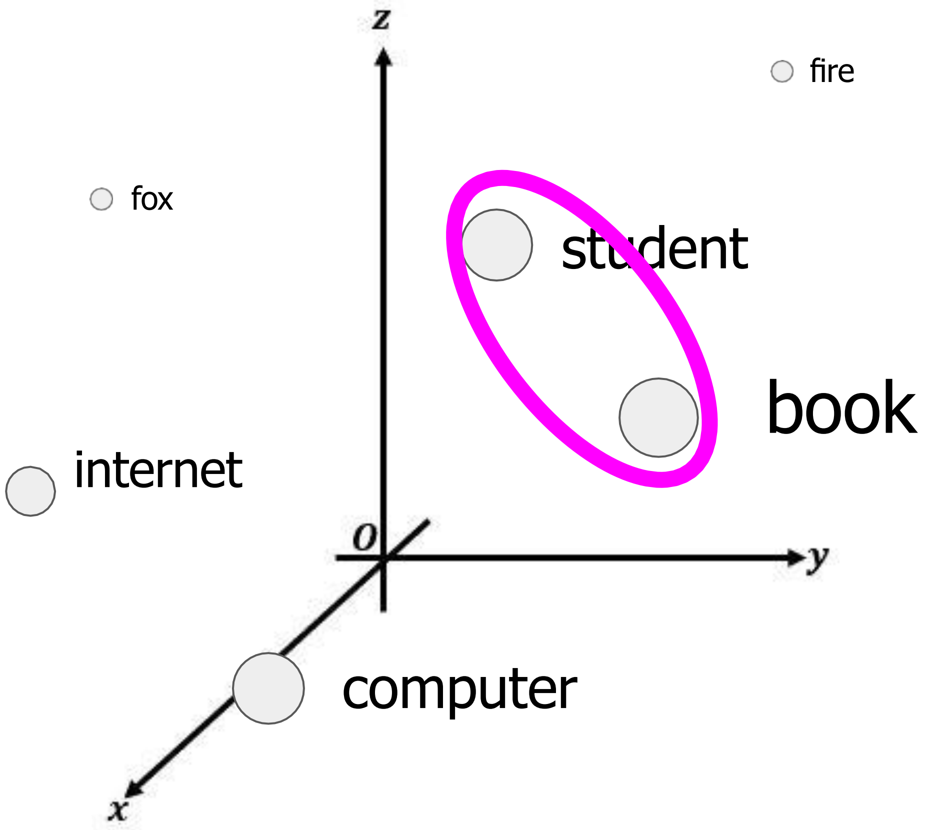
TRANSFORMERS



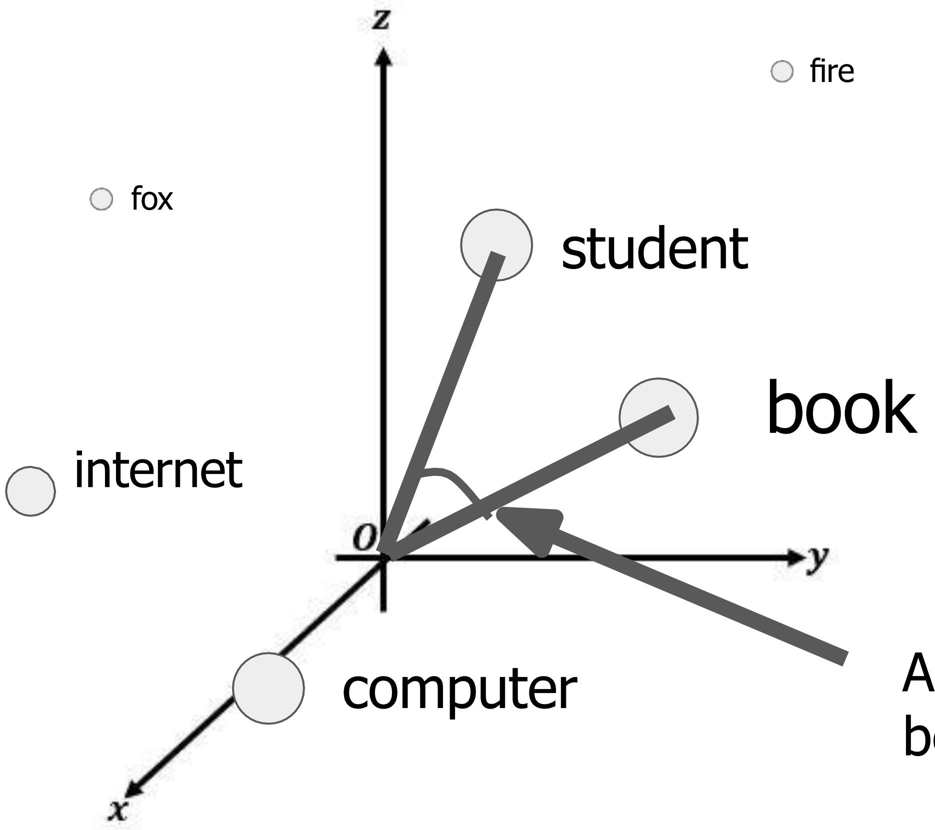
TRANSFORMERS



TRANSFORMERS

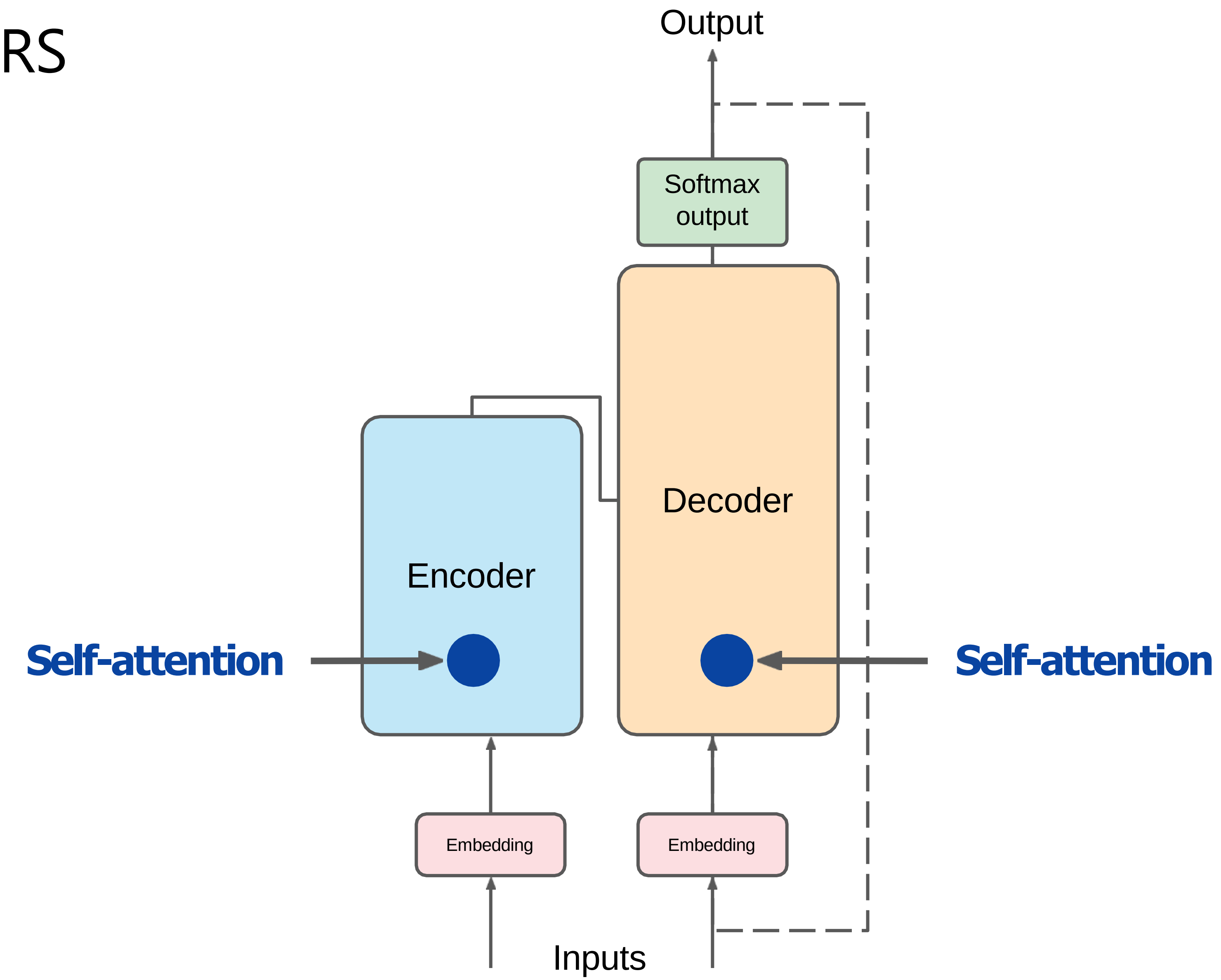


TRANSFORMERS

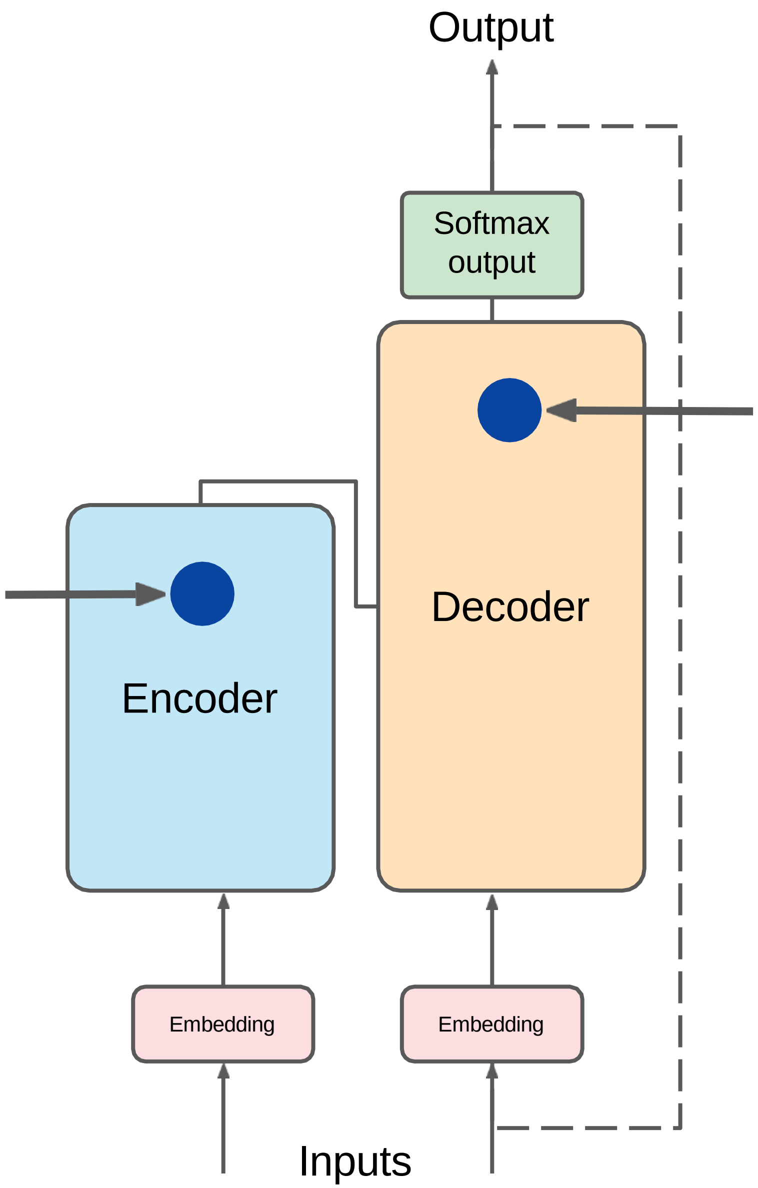


Angle measures distance between words

TRANSFORMERS

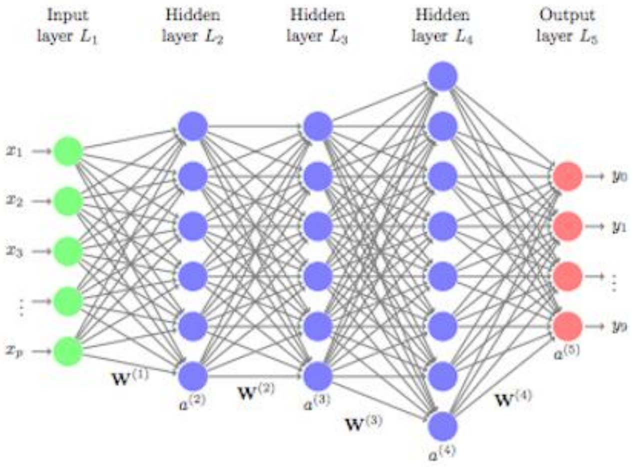


TRANSFORMERS

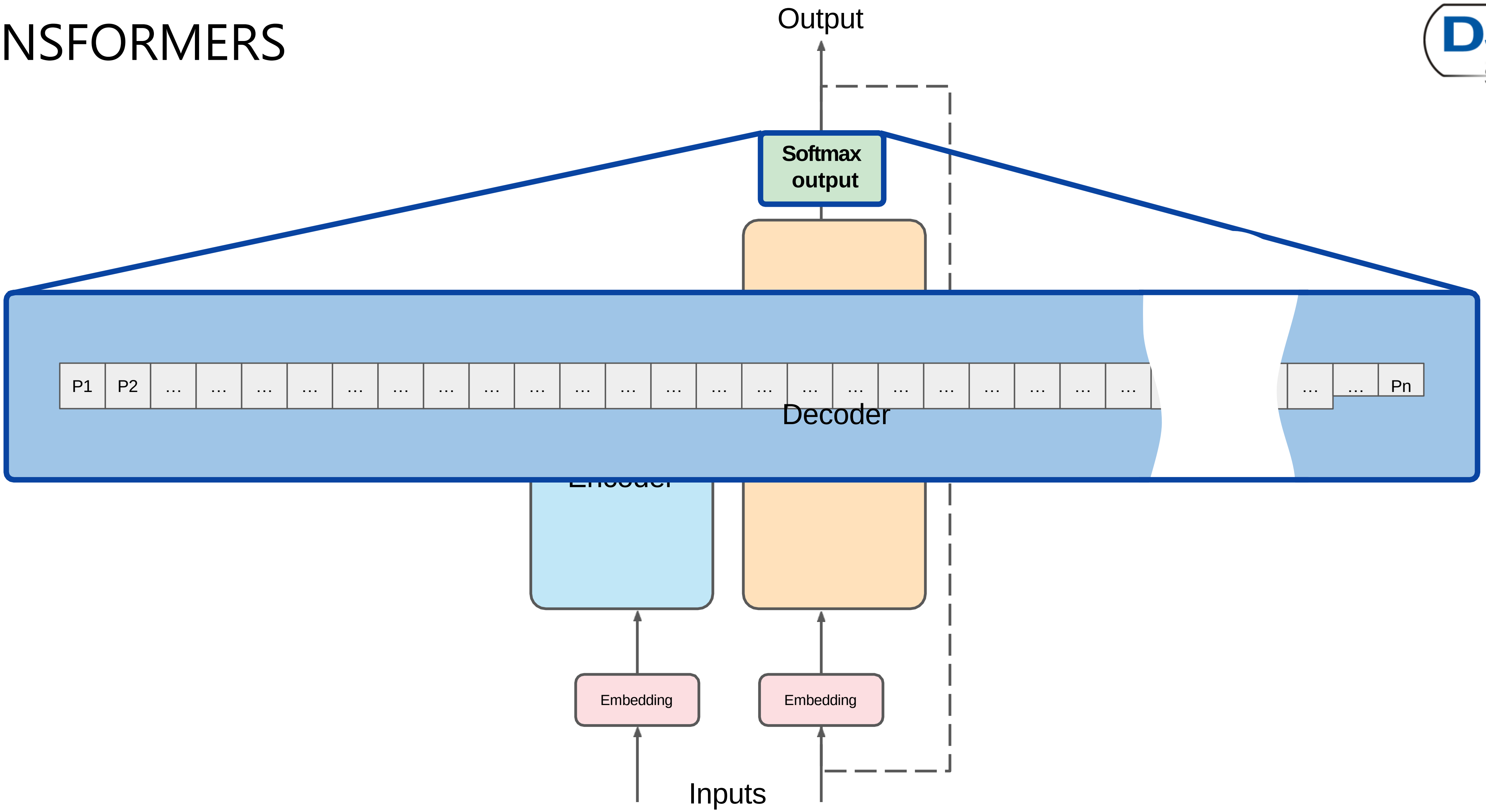


Feed forward network

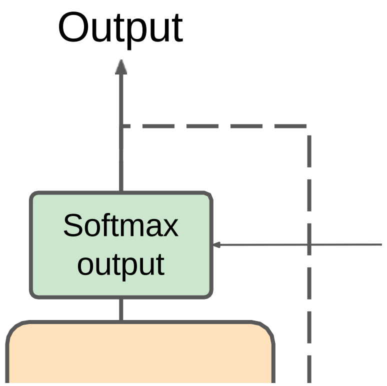
Feed forward network



TRANSFORMERS



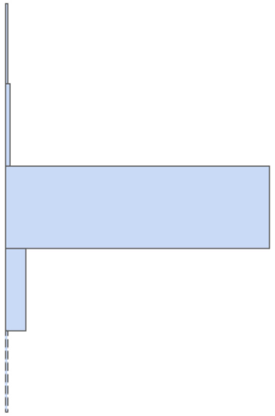
TRANSFORMERS



Softmax output

Cooler temperature (e.g <1)

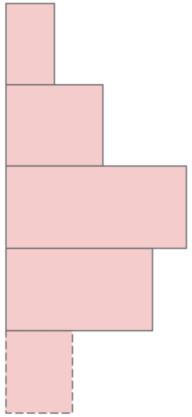
prob	word
0.001	apple
0.002	banana
0.400	cake
0.012	donut
…	…



Strongly peaked probability distribution

Higher temperature (>1)

prob	word
0.040	apple
0.080	banana
0.150	cake
0.120	donut
…	…

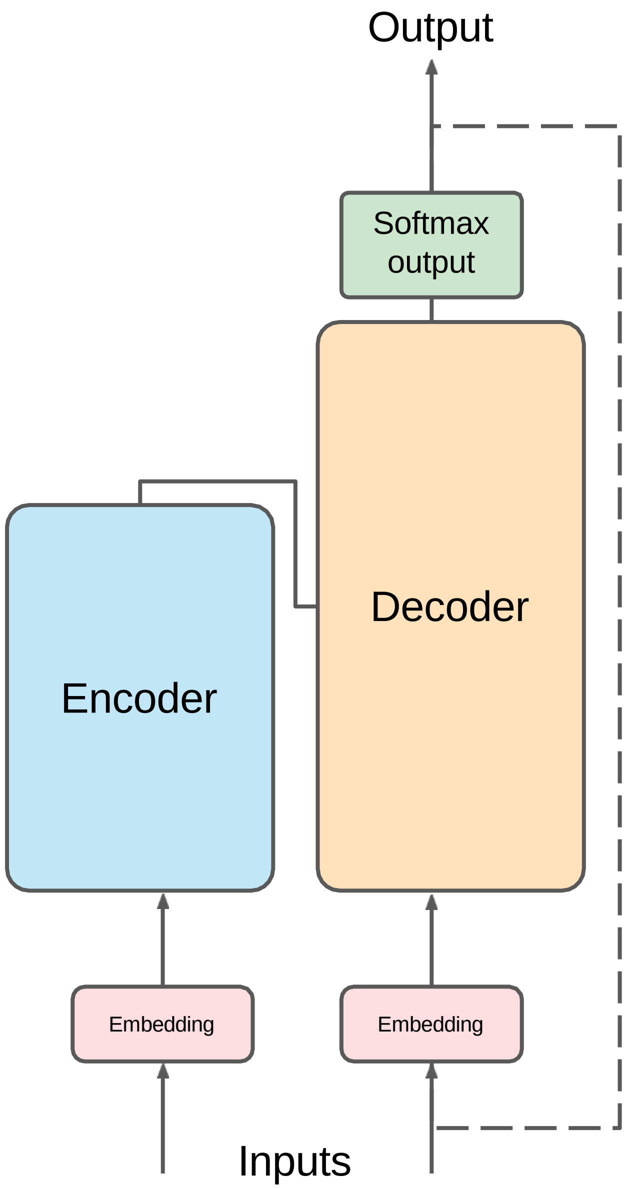


Broader, flatter probability distribution

TRANSFORMERS

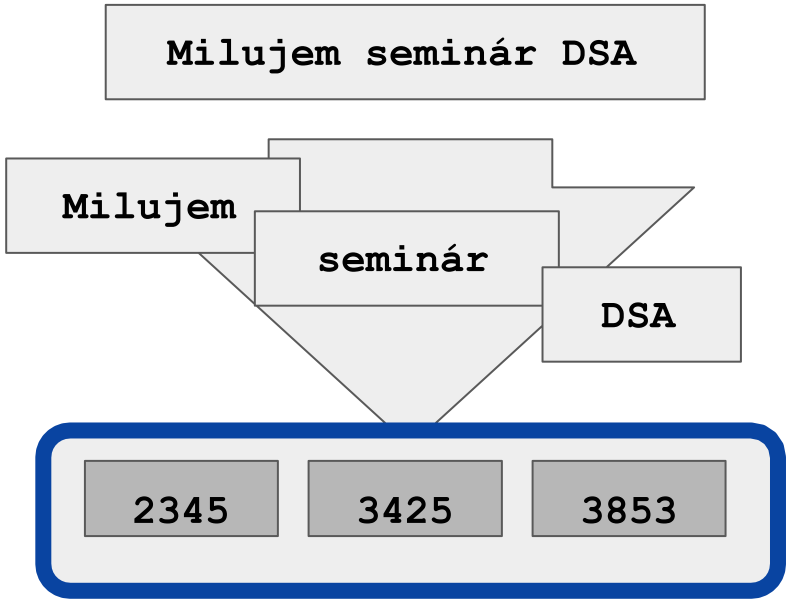
Translation:
sequence-to-sequence task

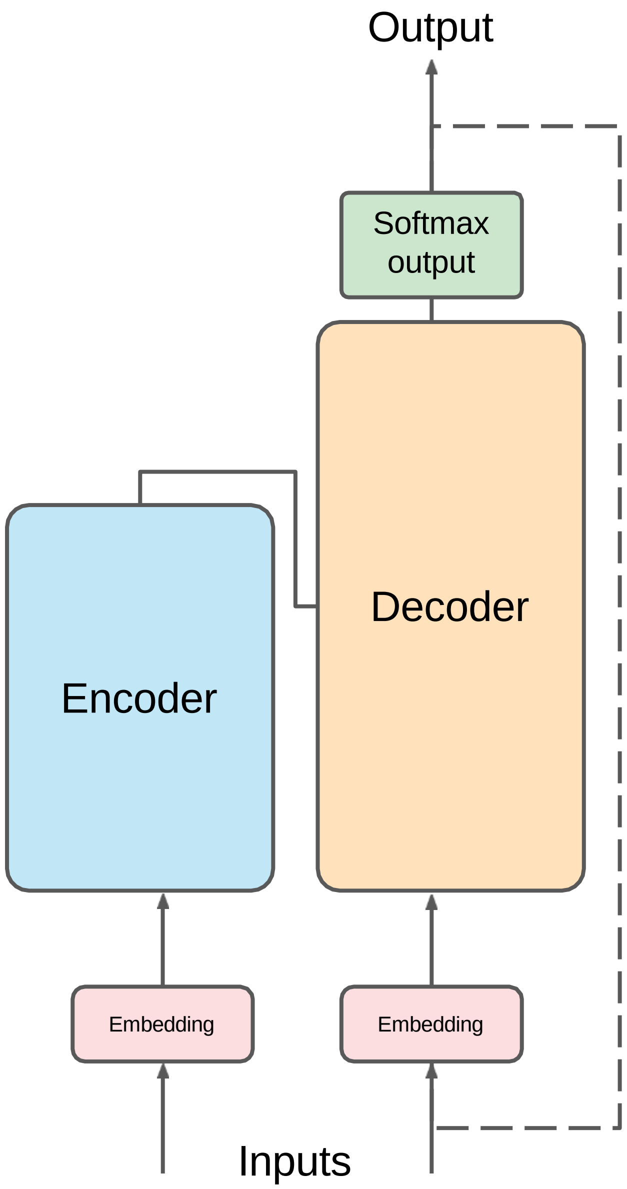
```
Milujem seminár DSA
```



TRANSFORMERS

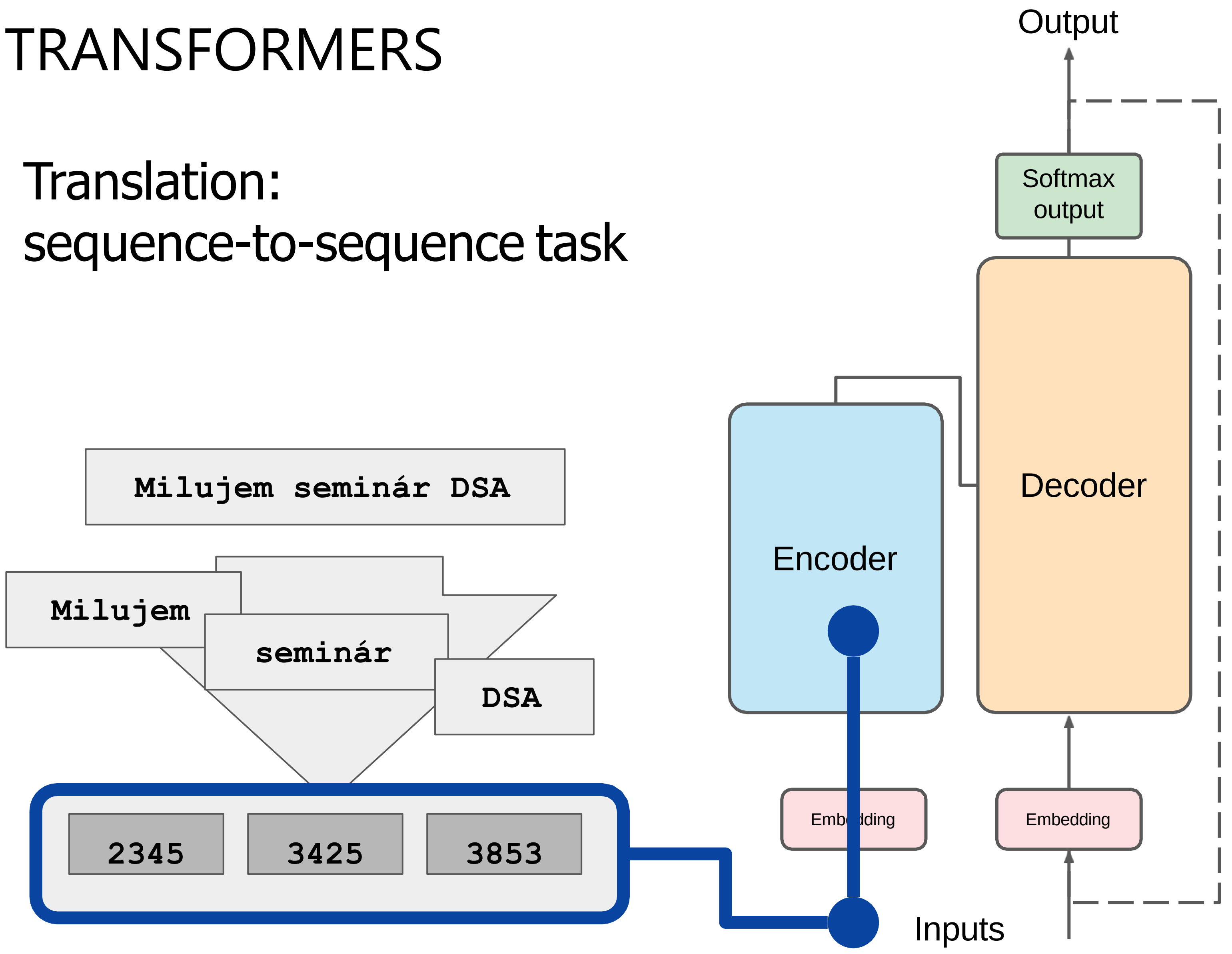
Translation:
sequence-to-sequence task





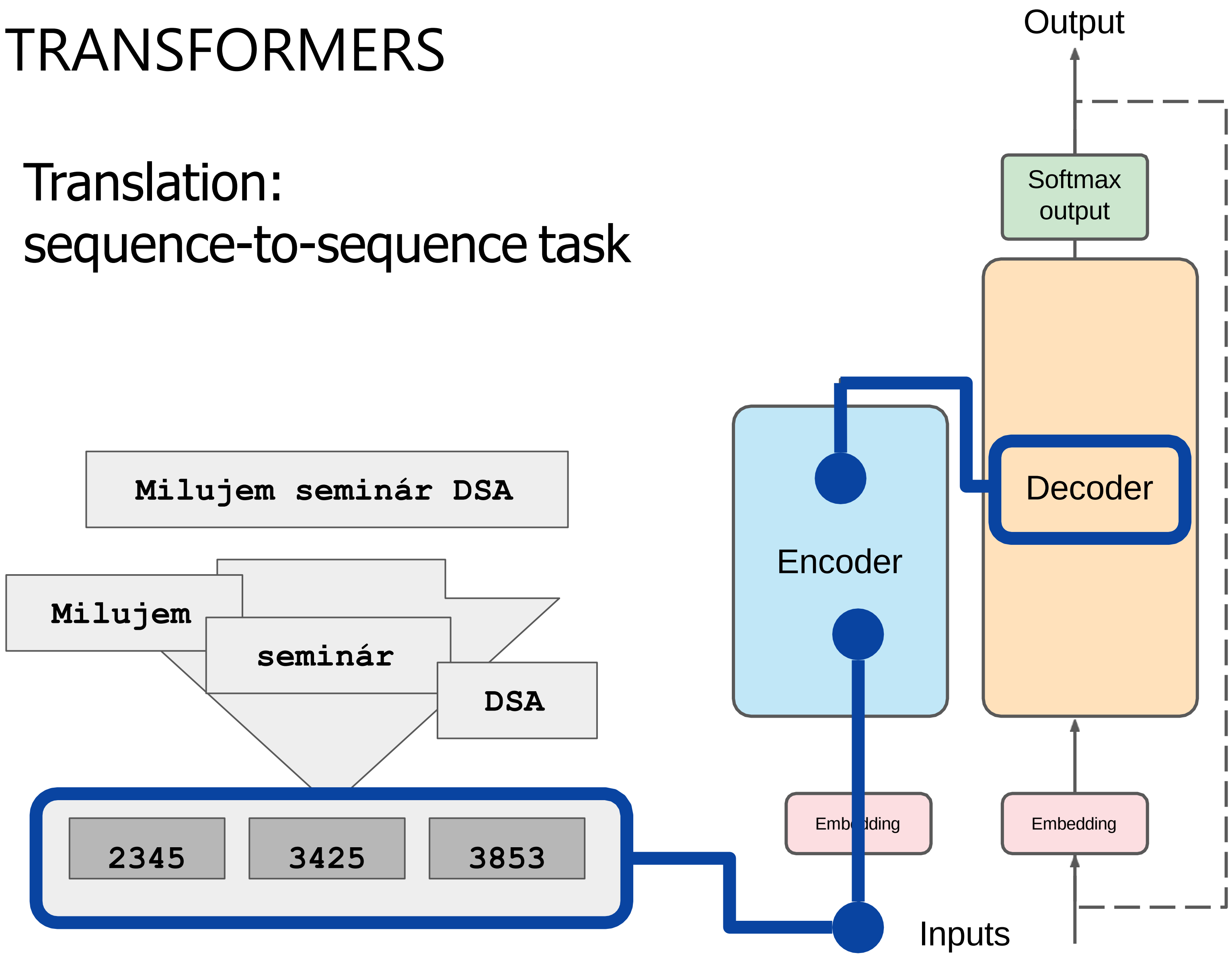
TRANSFORMERS

Translation:
sequence-to-sequence task



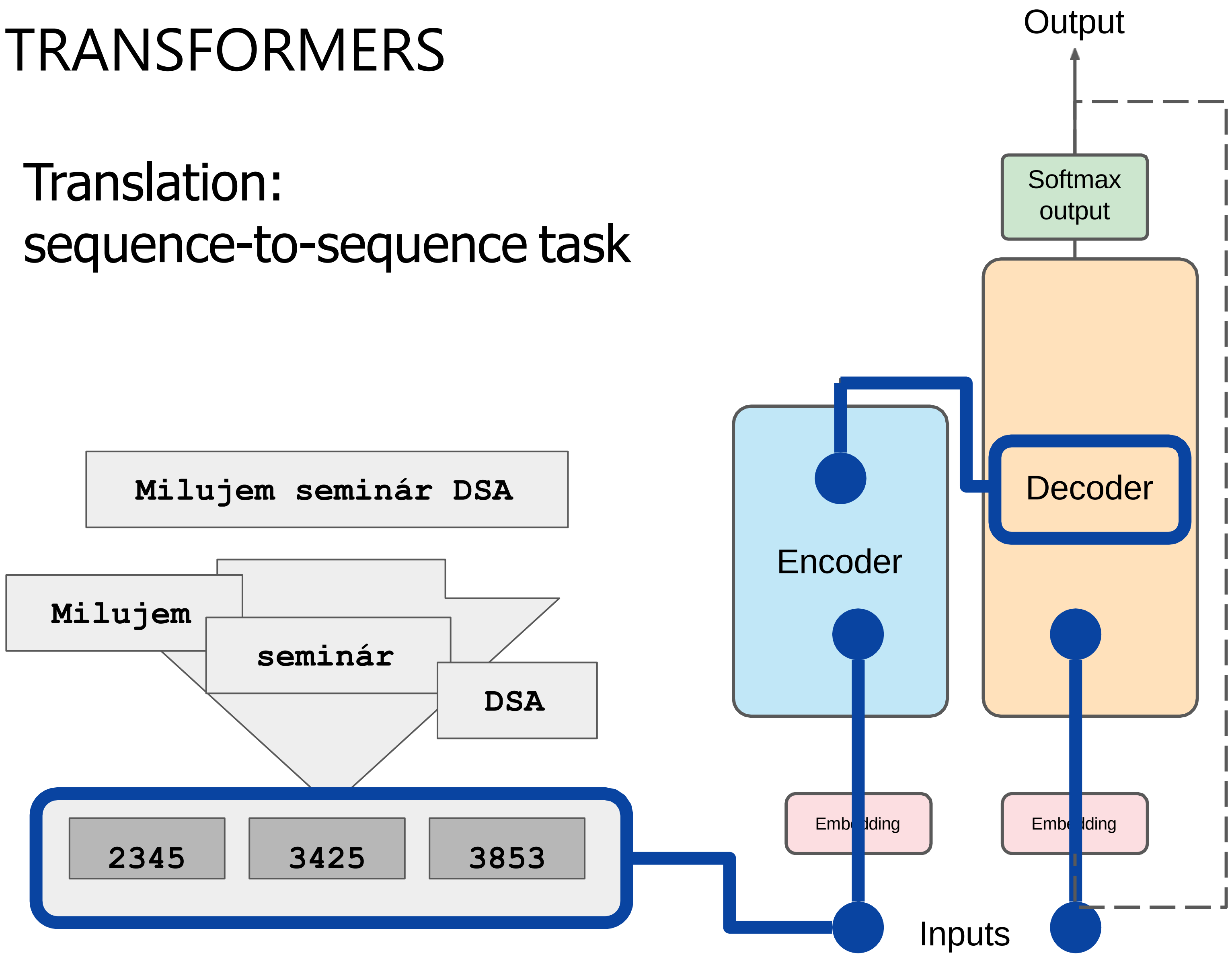
TRANSFORMERS

Translation:
sequence-to-sequence task



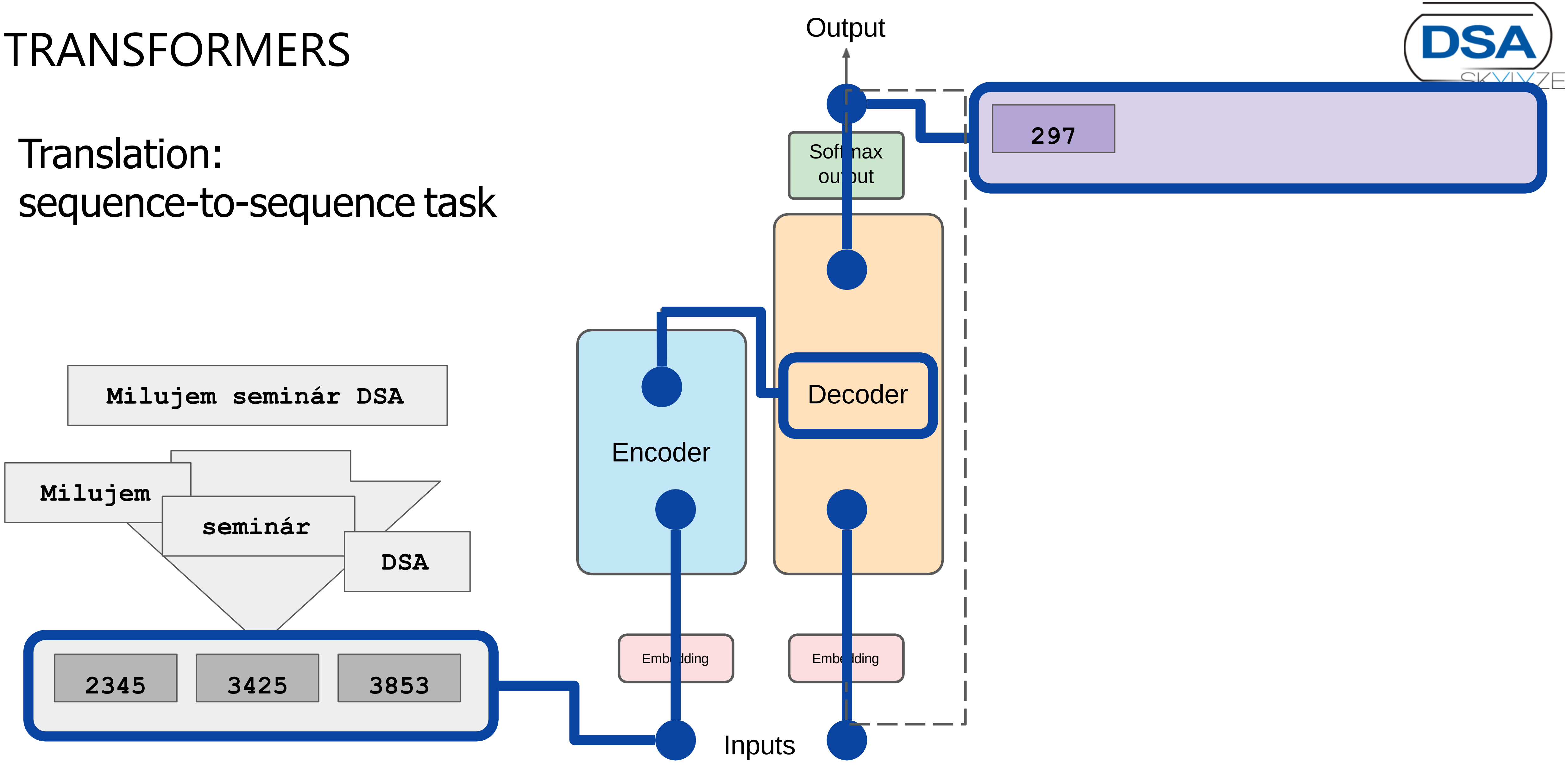
TRANSFORMERS

Translation:
sequence-to-sequence task



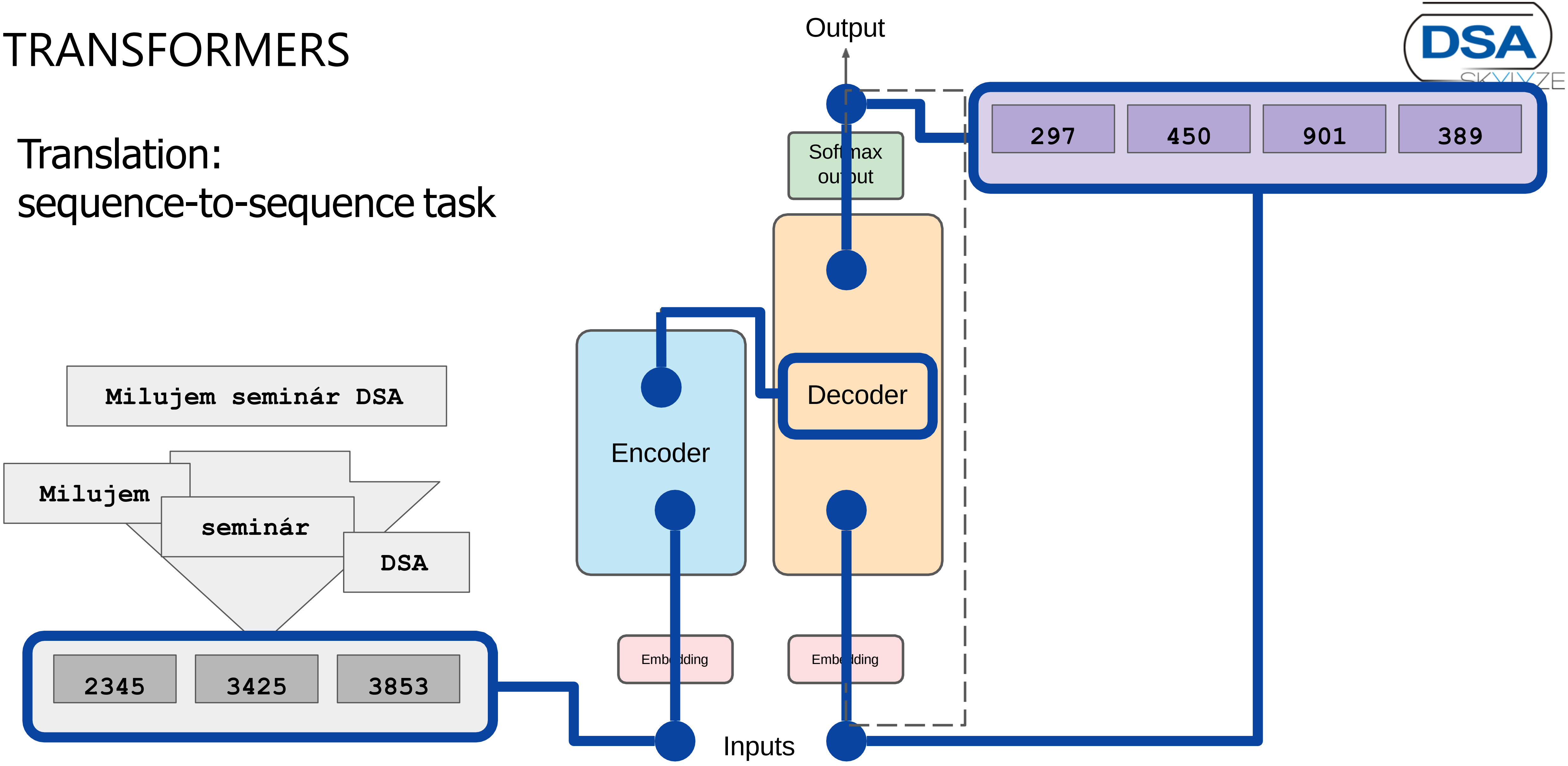
TRANSFORMERS

Translation:
sequence-to-sequence task



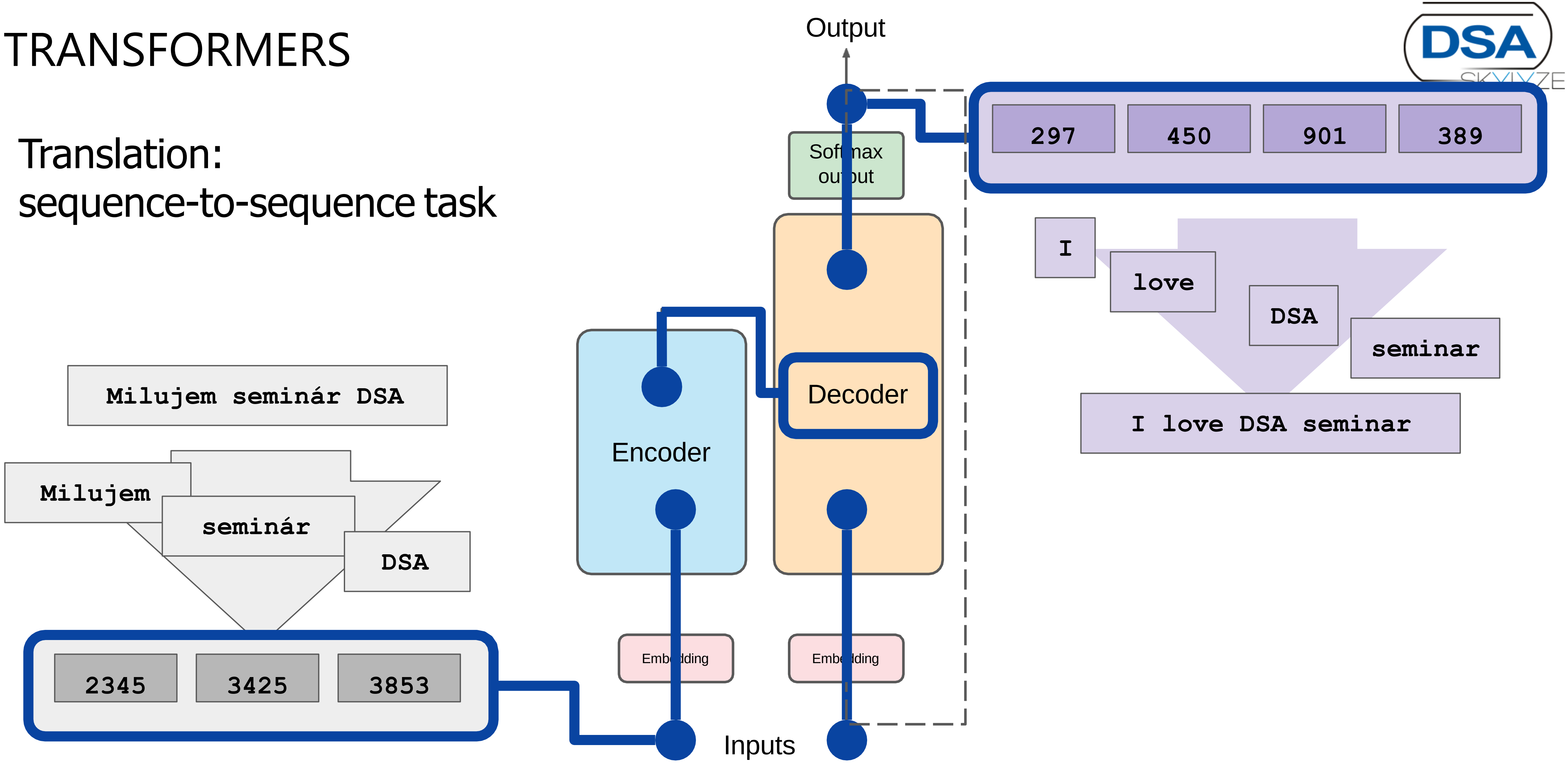
TRANSFORMERS

Translation:
sequence-to-sequence task



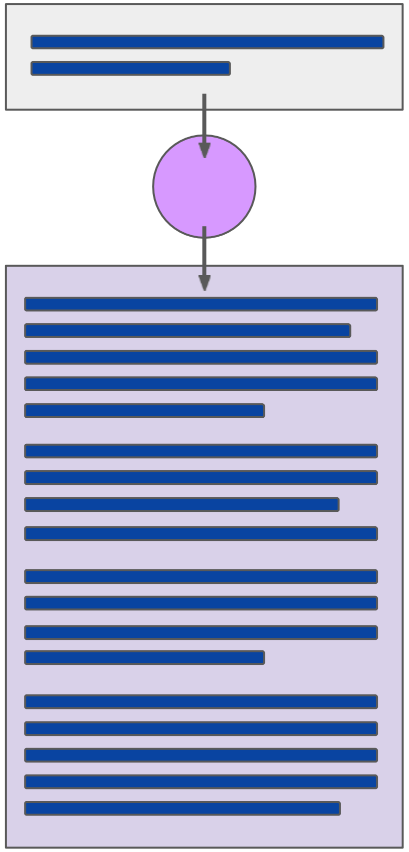
TRANSFORMERS

Translation:
sequence-to-sequence task

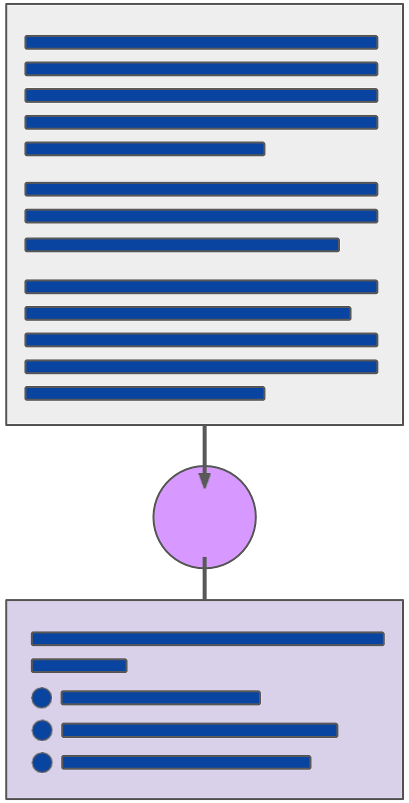


LLM USE CASES & TASKS

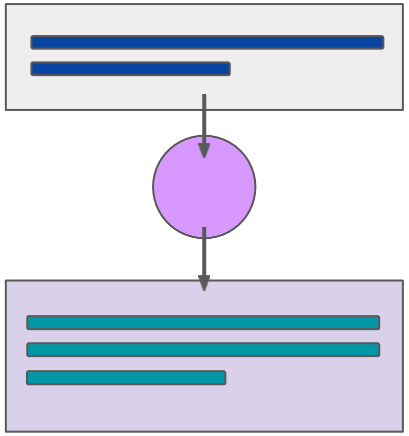
Essay Writing



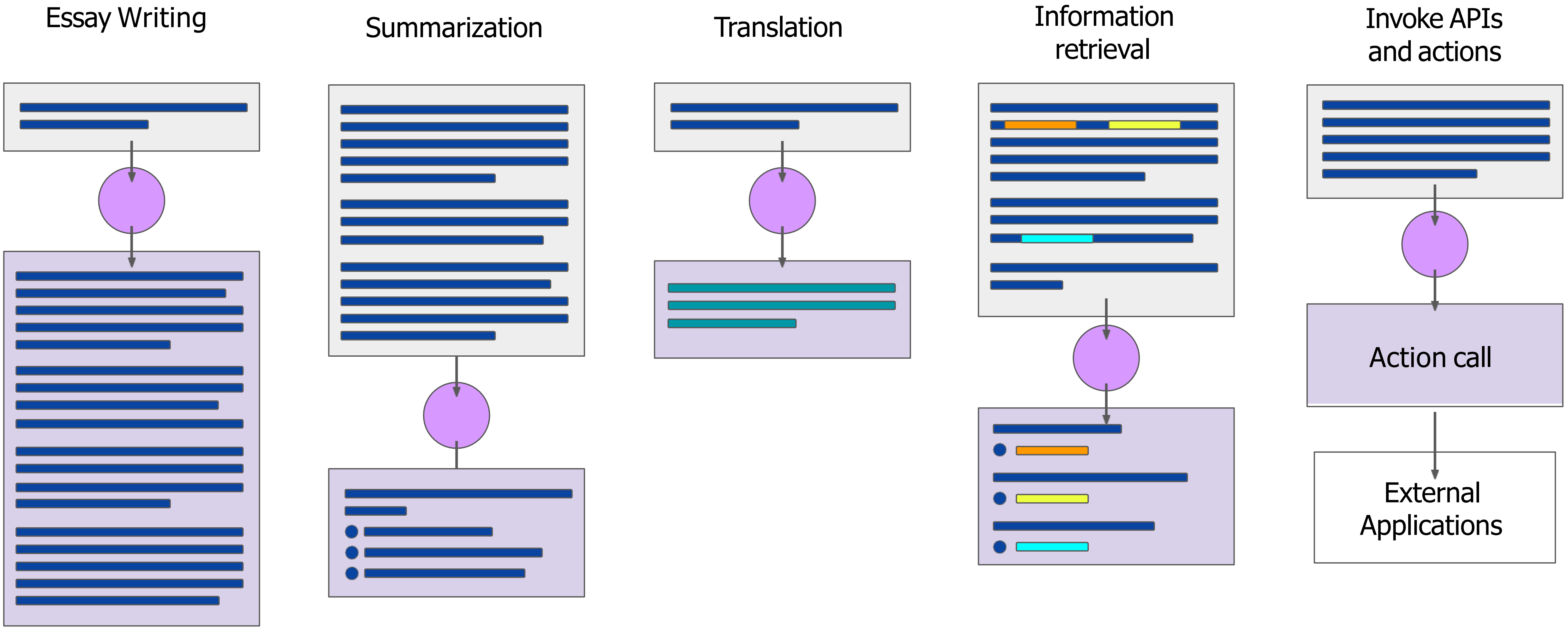
Summarization



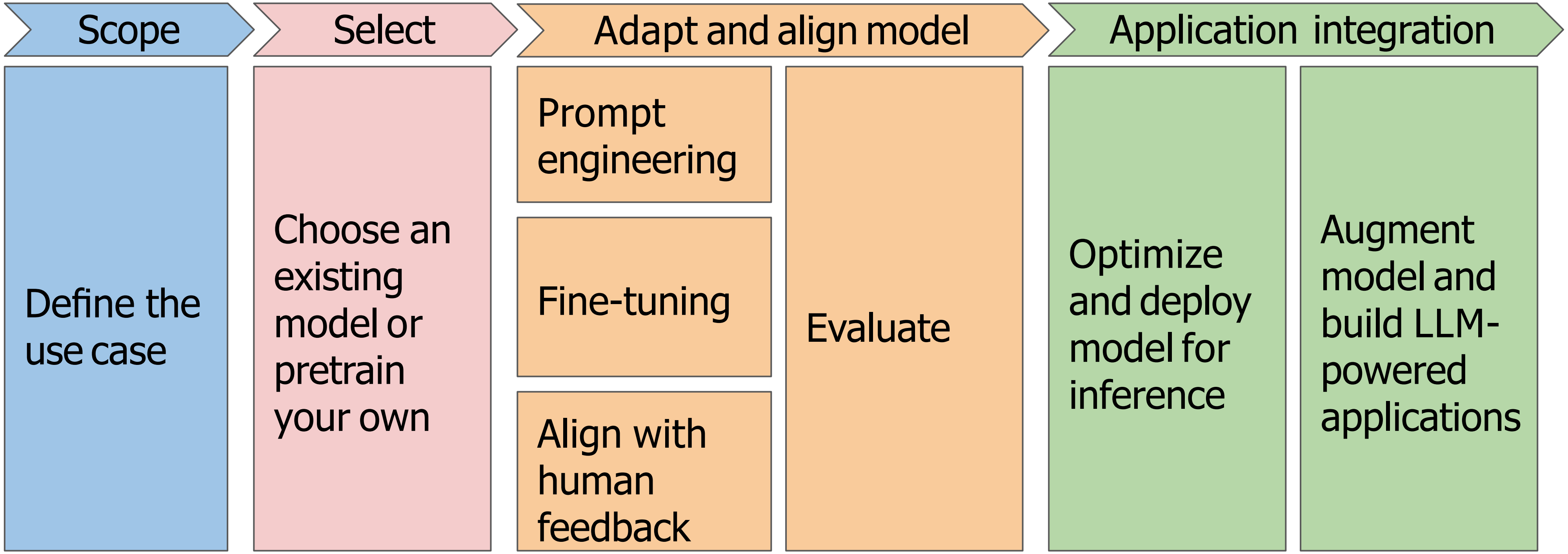
Translation



LLM USE CASES & TASKS



GENERATIVE AI PROJECT LIFECYCLE



Scope

Define the use case

Select

Choose an existing model or pretrain your own

Adapt and align model

Prompt engineering

Fine-tuning

Align with human feedback

Evaluate

Application integration

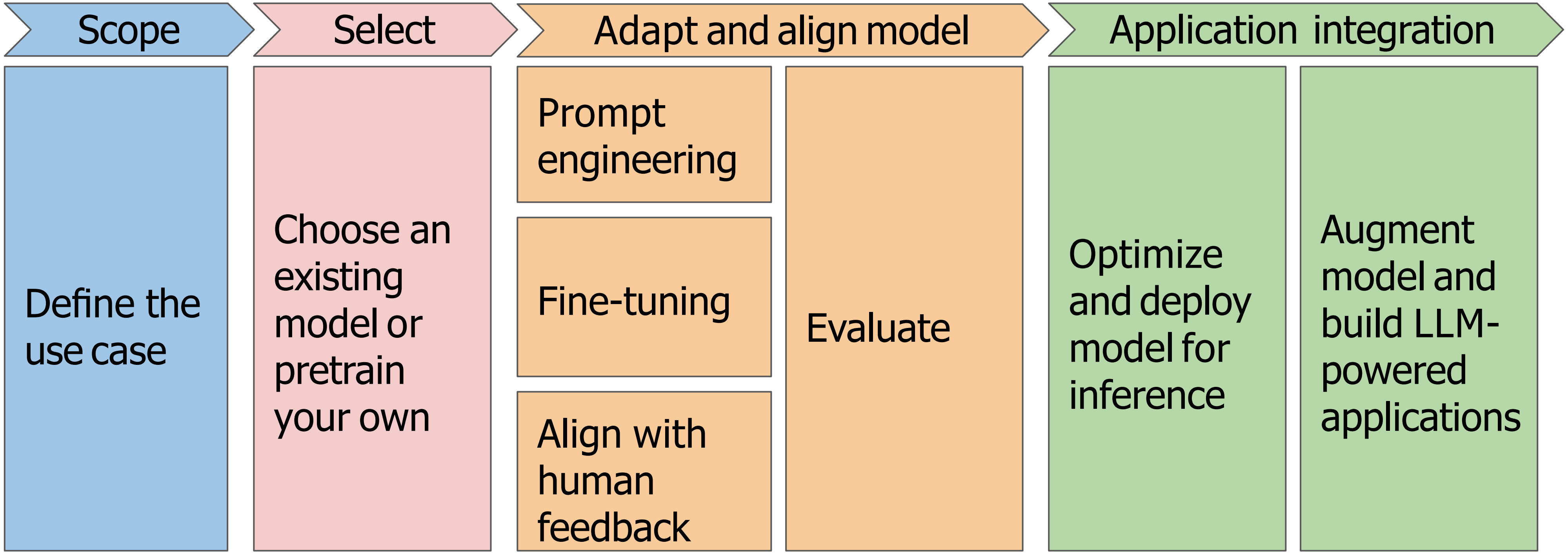
Optimize and deploy model for inference

Augment model and build LLM-powered applications

HANDS-ON

Nächstes Kapitel | Next Topic | Chapitre Prochaine | Capitolo Successivo

GENERATIVE AI PROJECT LIFECYCLE



Scope

Define the use case

Select

Choose an existing model or pretrain your own

Adapt and align model

Prompt engineering

Fine-tuning

Align with human feedback

Evaluate

Application integration

Optimize and deploy model for inference

Augment model and build LLM-powered applications

CHALLENGES OF TRAINING LLMS

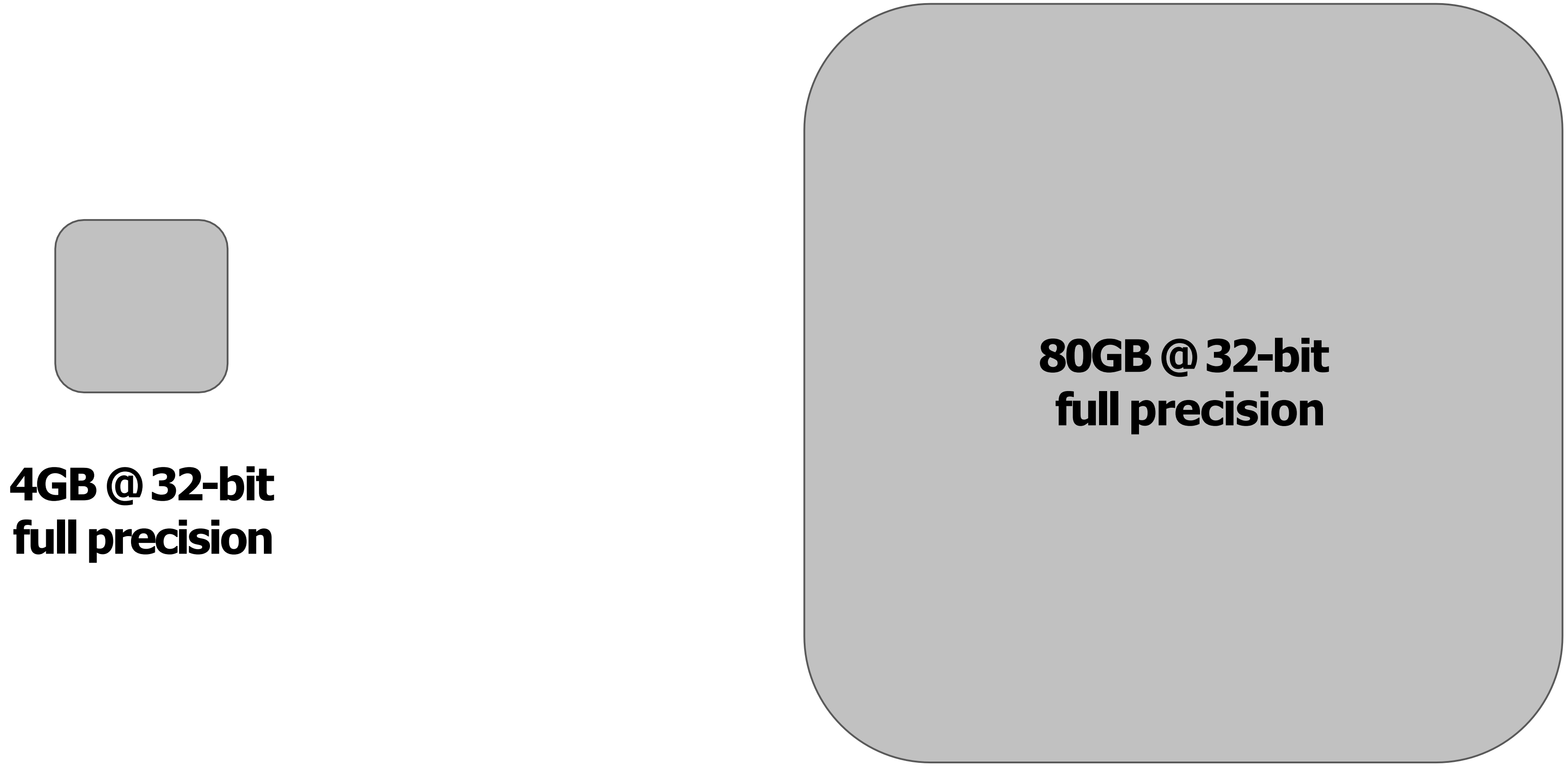
Infrastructure and costs

- Trained on massive corpus of text data
- Billions to trillions of parameter
- GPU-Hardware/Infrastructure needed
- Approximate GPU RAM needed to store 1B parameters
  - 1 parameter = 4 bytes (32-bit float)
  - 1B parameters = 4 x 1e-09 bytes = 4GB
- Training needs ~20 extra bytes per parameter

APPROXIMATE GPU RAM NEEDED TO TRAIN 1B-PARAMS

Memory needed to store model

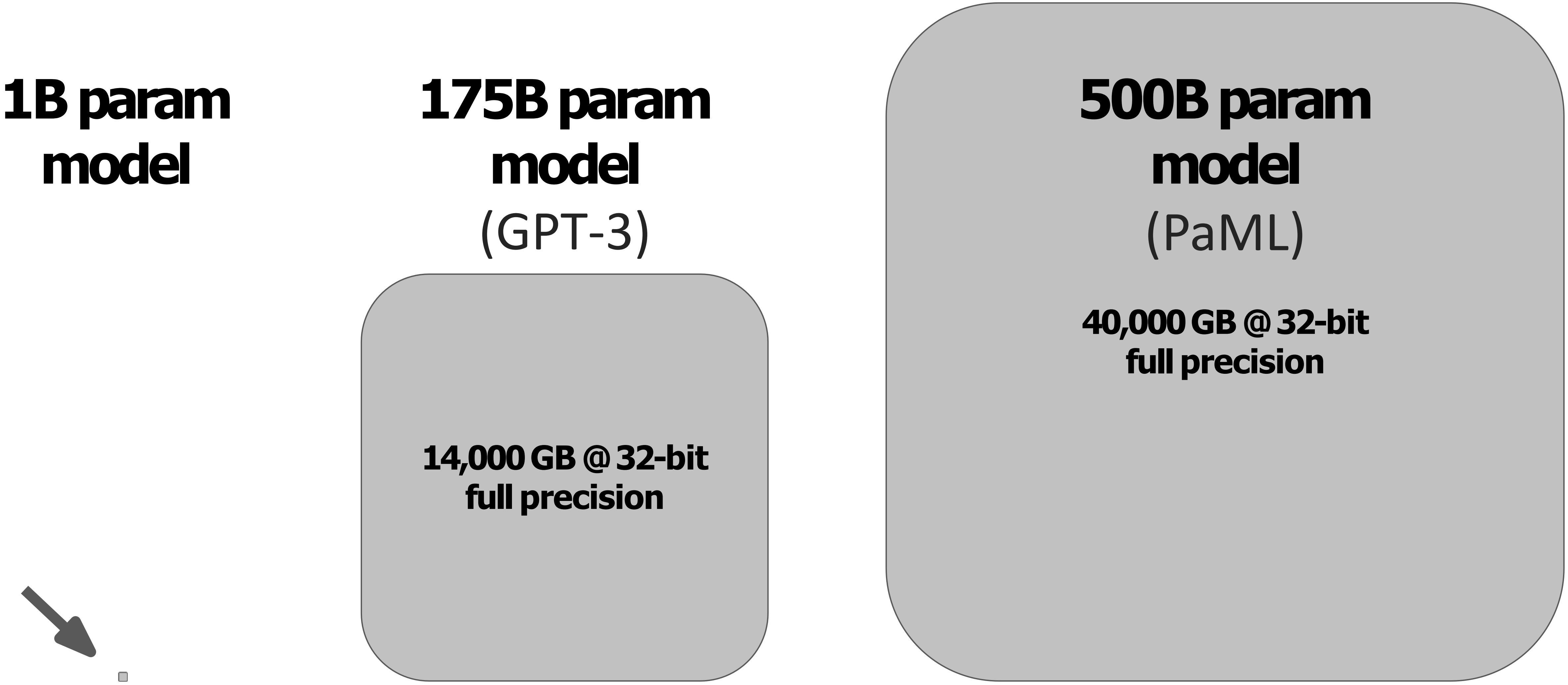
Memory needed to train model



4GB @ 32-bit full precision

80GB @ 32-bit full precision

GPU RAM NEEDED TO TRAIN LARGER MODELS



1B param model

175B param model

(GPT-3)

14,000 GB @ 32-bit full precision

500B param model

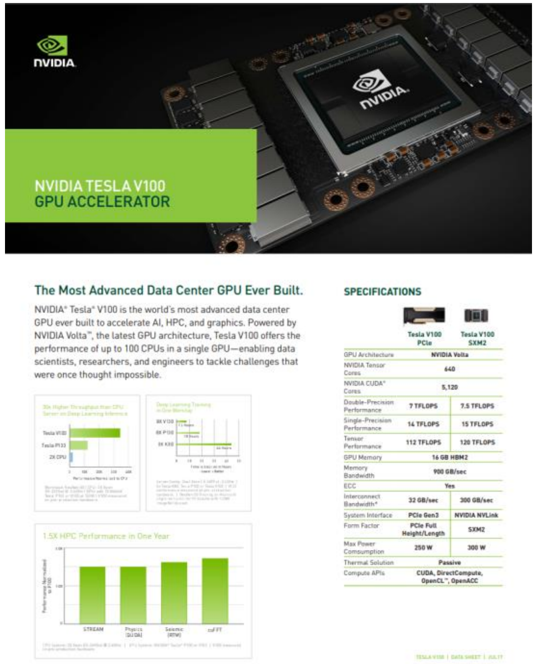
(PaML)

40,000 GB @ 32-bit full precision

CHALLENGES OF TRAINING LLMS

Infrastructure and costs

- It would take 288 years to train GPT-3 (175B) on a single NVIDIA Tesla V100 GPU.
  - 1Y = 8760h
  - 275W * 8760h = 2409000Wh = 2490kWh
  - 2 Person-Household : 2000 – 2800kWh
- Researchers calculated that OpenAI could have trained GPT-3 in as little as 34 days on 1,024x A100 GPUs
- It is estimated that GPT-3 cost around $4.6 million dollars to train from scratch
- Cost of this workshop ~90$ (1$/min)



NVIDIA TESLA V100
GPU ACCELERATOR

The Most Advanced Data Center GPU Ever Built.

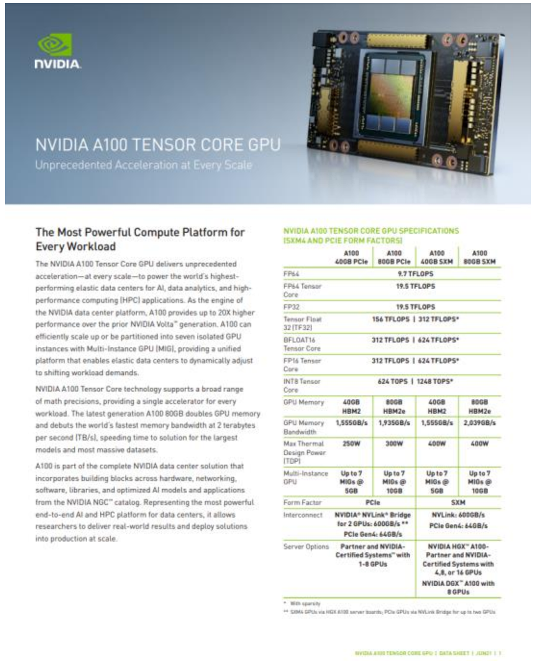
NVIDIA® Tesla® V100 is the world's most advanced data center GPU ever built to accelerate AI, HPC, and graphics. Powered by NVIDIA Volta™, the latest GPU architecture, Tesla V100 offers the performance of up to 100 CPUs in a single GPU—enabling data scientists, researchers, and engineers to tackle challenges that were once thought impossible.

SPECIFICATIONS

	Tesla V100 PCle	Tesla V100 SXM2
GPU Architecture	NVIDIA Volta	
NVIDIA Tensor Cores	640	
NVIDIA CUDA® Cores	5,120	
Double-Precision Performance	7 TFLOPS	7.5 TFLOPS
Single-Precision Performance	14 TFLOPS	15 TFLOPS
Tensor Performance	112 TFLOPS	120 TFLOPS
GPU Memory	16 GB HBM2	
Memory Bandwidth	900 GB/sec	
ECC	Yes	
Interconnect Bandwidth*	32 GB/sec	300 GB/sec
System Interface	PCIe Gen3	NVIDIA NVLink
Form Factor	PCIe Full Height/Length	SXM2
Max Power Consumption	250 W	300 W
Thermal Solution	Passive	
Compute APIs	CUDA, DirectCompute, OpenCL™, OpenACC	

TESLA V100 | DATA SHEET | JUL17

- Jun 2017
- 16GB RAM
- 250-300 Watt



NVIDIA A100 TENSOR CORE GPU
Unprecedented Acceleration at Every Scale

The Most Powerful Compute Platform for Every Workload

The NVIDIA A100 Tensor Core GPU delivers unprecedented acceleration—at every scale—to power the world's highest-performing elastic data centers for AI, data analytics, and high-performance computing (HPC) applications. As the engine of the NVIDIA data center platform, A100 provides up to 20X higher performance over the prior NVIDIA Volta™ generation. A100 can efficiently scale up or be partitioned into seven isolated GPU instances with Multi-Instance GPU (MIG), providing a unified platform that enables elastic data centers to dynamically adjust to shifting workload demands.

NVIDIA A100 Tensor Core technology supports a broad range of math precisions, providing a single accelerator for every workload. The latest generation A100 80GB doubles GPU memory and debuts the world's fastest memory bandwidth at 2 terabytes per second (TB/s), speeding time to solution for the largest models and most massive datasets.

A100 is part of the complete NVIDIA data center solution that incorporates building blocks across hardware, networking, software, libraries, and optimized AI models and applications from the NVIDIA NGC™ catalog. Representing the most powerful end-to-end AI and HPC platform for data centers, it allows researchers to deliver real-world results and deploy solutions into production at scale.

NVIDIA A100 TENSOR CORE GPU SPECIFICATIONS (SXM4 AND PCIE FORM FACTORS)

	A100 40GB PCIe	A100 80GB PCIe	A100 40GB SXM	A100 80GB SXM
FP64	9.7 TFLOPS			
FP64 Tensor Core	19.5 TFLOPS			
FP32	19.5 TFLOPS			
Tensor Float 32 (TF32)	156 TFLOPS | 312 TFLOPS*			
BFLOAT16 Tensor Core	312 TFLOPS | 624 TFLOPS*			
FP16 Tensor Core	312 TFLOPS | 624 TFLOPS*			
INT8 Tensor Core	624 TOPS | 1248 TOPS*			
GPU Memory	40GB HBM2	80GB HBM2e	40GB HBM2	80GB HBM2e
GPU Memory Bandwidth	1,555GB/s	1,935GB/s	1,555GB/s	2,039GB/s
Max Thermal Design Power (TDP)	250W	300W	400W	400W
Multi-Instance GPU	Up to 7 MIGs @ 5GB	Up to 7 MIGs @ 10GB	Up to 7 MIGs @ 5GB	Up to 7 MIGs @ 10GB
Form Factor	PCIe		SXM	
Interconnect	NVIDIA® NVLink® Bridge for 2 GPUs: 600GB/s ** PCIe Gen4: 64GB/s		NVLink: 600GB/s PCIe Gen4: 64GB/s	
Server Options	Partner and NVIDIA-Certified Systems™ with 1-8 GPUs		NVIDIA HGX™ A100-Partner and NVIDIA-Certified Systems with 4,8, or 16 GPUs NVIDIA DGX™ A100 with 8 GPUs	

* With sparsity
** SXM4 GPUs via HGX A100 server boards; PCIe GPUs via NVLink Bridge for up to two GPUs

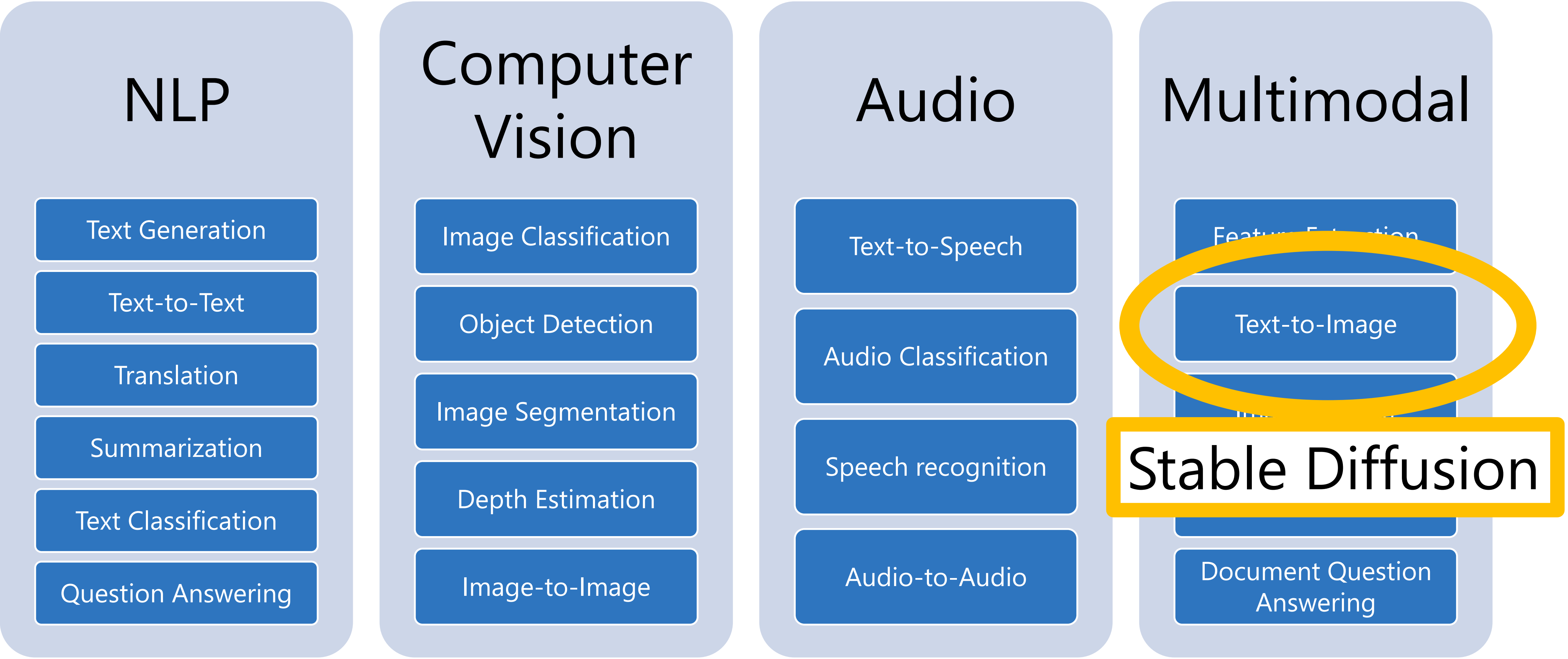
NVIDIA A100 TENSOR CORE GPU | DATA SHEET | JUN21 | 1

- Jun 2021
- 80GB RAM
- 300-400 Watt

GENERATIVE AI

Nächstes Kapitel | Next Topic | Chapitre Prochaine | Capitolo Successivo

GENERATIVE AI



NLP

- Text Generation
- Text-to-Text
- Translation
- Summarization
- Text Classification
- Question Answering

Computer Vision

- Image Classification
- Object Detection
- Image Segmentation
- Depth Estimation
- Image-to-Image

Audio

- Text-to-Speech
- Audio Classification
- Speech recognition
- Audio-to-Audio

Multimodal

- Feature Extraction
- Text-to-Image
- Document Question Answering

Stable Diffusion

STABLE DIFFUSION

Nächstes Kapitel | Next Topic | Chapitre Prochaine | Capitolo Successivo

ACKNOWLEDGEMENT AND COPYRIGHT NOTICE

The presentation is based on the course Generative AI with Large Language Models. The slides are based on the course material from DeepLearning.AI