



Generative AI and Large Language Models for Question Answering and Dialogue Systems

Time: 10:00-12:00; 14:00-16:00, Wednesday, July 26, 2023

Place: Taipei Research and Development Center, Asia University

Address: 16-5, No. 77, Xintai 5th Road, Xizhi District, New Taipei City, Taiwan (FE World Center)

Host: Prof. Wen-Lian Hsu



Min-Yuh Day, Ph.D,
Associate Professor

Institute of Information Management, National Taipei University

<https://web.ntpu.edu.tw/~myday>





戴敏育 博士

Min-Yuh Day, Ph.D.

Associate Professor, Information Management, NTPU

Visiting Scholar, IIS, Academia Sinica

Ph.D., Information Management, NTU

Director, Intelligent Financial Innovation Technology, IFIT Lab, IM, NTPU

Associate Director, Fintech and Green Finance Center, NTPU

**Publications Co-Chairs, IEEE/ACM International Conference on
Advances in Social Networks Analysis and Mining (ASONAM 2013-)**

**Program Co-Chair, IEEE International Workshop on
Empirical Methods for Recognizing Inference in Text (IEEE EM-RITE 2012-)**

**Publications Chair, The IEEE International Conference on
Information Reuse and Integration for Data Science (IEEE IRI 2007-)**



Outline

- **Introduction**
- **Overview of Generative AI**
- **Overview of Large Language Models (LLMs)**
- **Foundation of Transformers: Attention Mechanism**
- **Fine-tuning LLM for Question Answering System**
- **Fine-tuning LLM for Dialogue System**
- **Challenges and Limitations of Generative AI for QA and Dialogue Systems**
- **Q & A**

Introduction

- In this talk, we're embarking on an exciting journey into the fascinating world of Generative AI and Large Language Models (LLMs) and their profound impact on modern NLP, with a special focus on question answering and dialogue systems.
- We're going to dive deep into the core of generative AI, pull back the curtain on the universe of LLMs, and delve into the beating heart of transformers by focusing on the attention mechanism.
- We'll also explore how to fine-tune an LLM for a question answering system, and for a dialogue system.
- Then, we'll tackle the challenges and limitations of using generative AI for QA and dialogue systems.
- Finally, we're going to wrap everything up with a lively Q&A session.

Outline

- Introduction
- **Overview of Generative AI**
- Overview of Large Language Models (LLMs)
- Foundation of Transformers: Attention Mechanism
- Fine-tuning LLM for Question Answering System
- Fine-tuning LLM for Dialogue System
- Challenges and Limitations of Generative AI for QA and Dialogue Systems
- Q & A

Overview of Generative AI

Generative AI
(Gen AI)

AI Generated Content
(AIGC)

Generative AI (Gen AI)

AI Generated Content (AIGC)

Image Generation

Instruction 1:

An astronaut riding a horse in a photorealistic style.

Instruction 2:

Teddy bears working on new AI research on the moon in the 1980s.

 **OpenAI DALL·E 2**

Figure 1



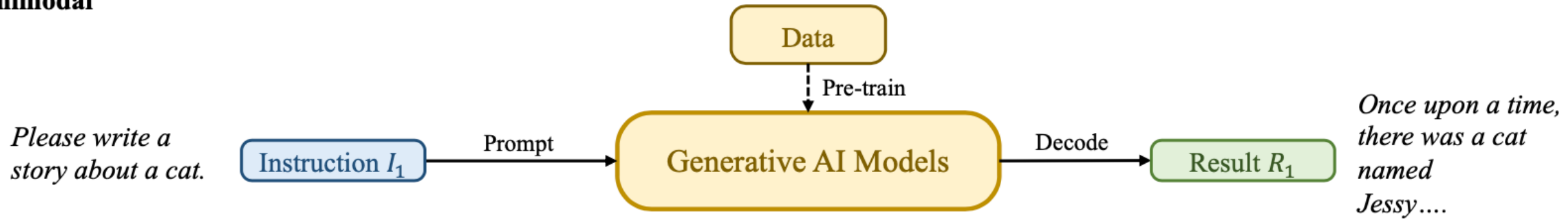
Figure 2



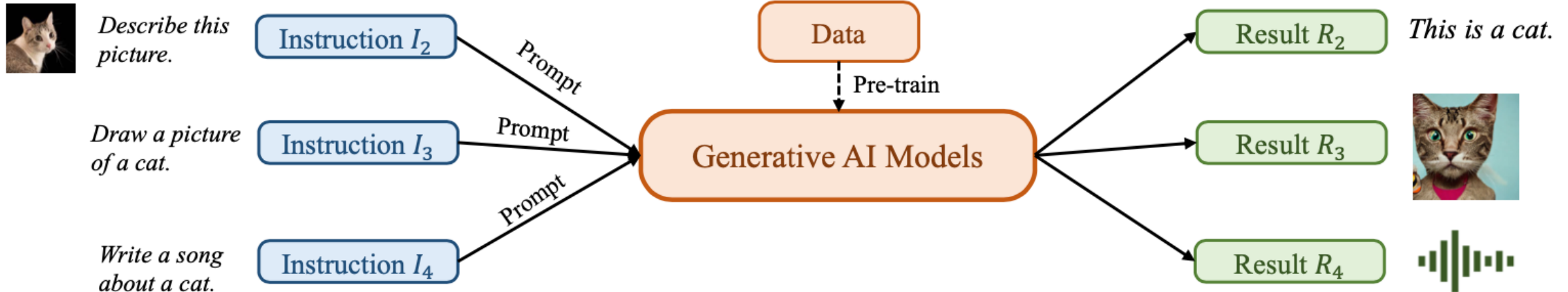
Generative AI (Gen AI)

AI Generated Content (AIGC)

Unimodal

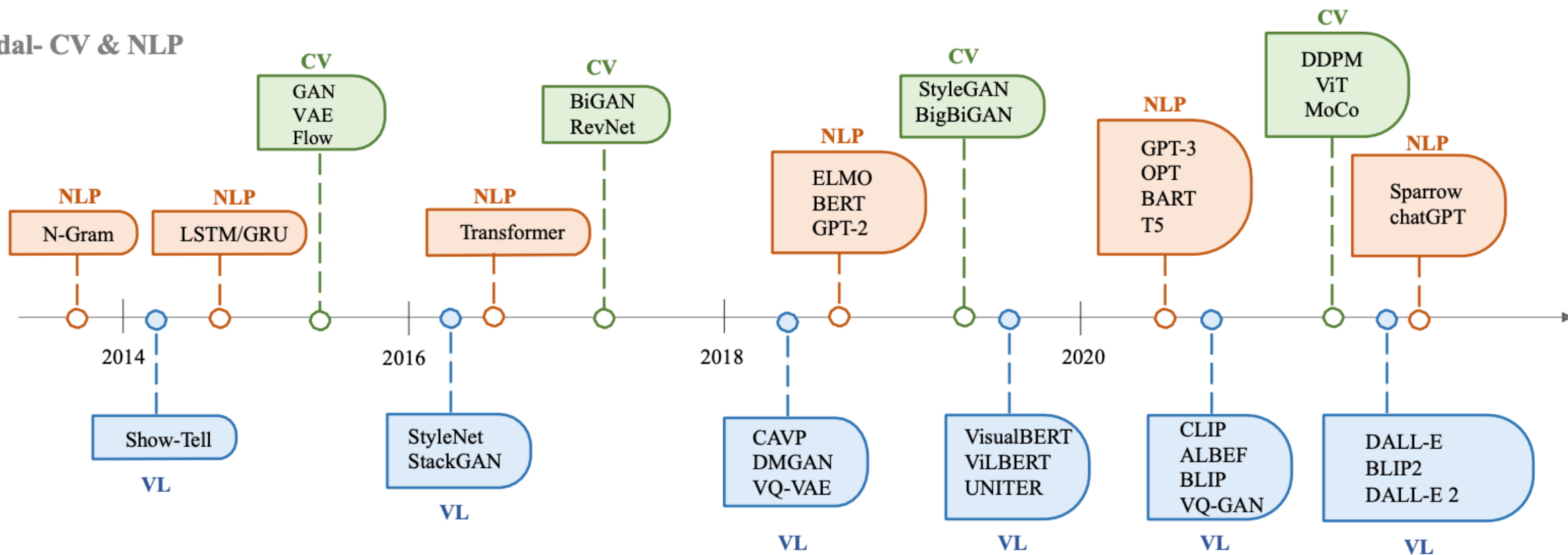


Multimodal



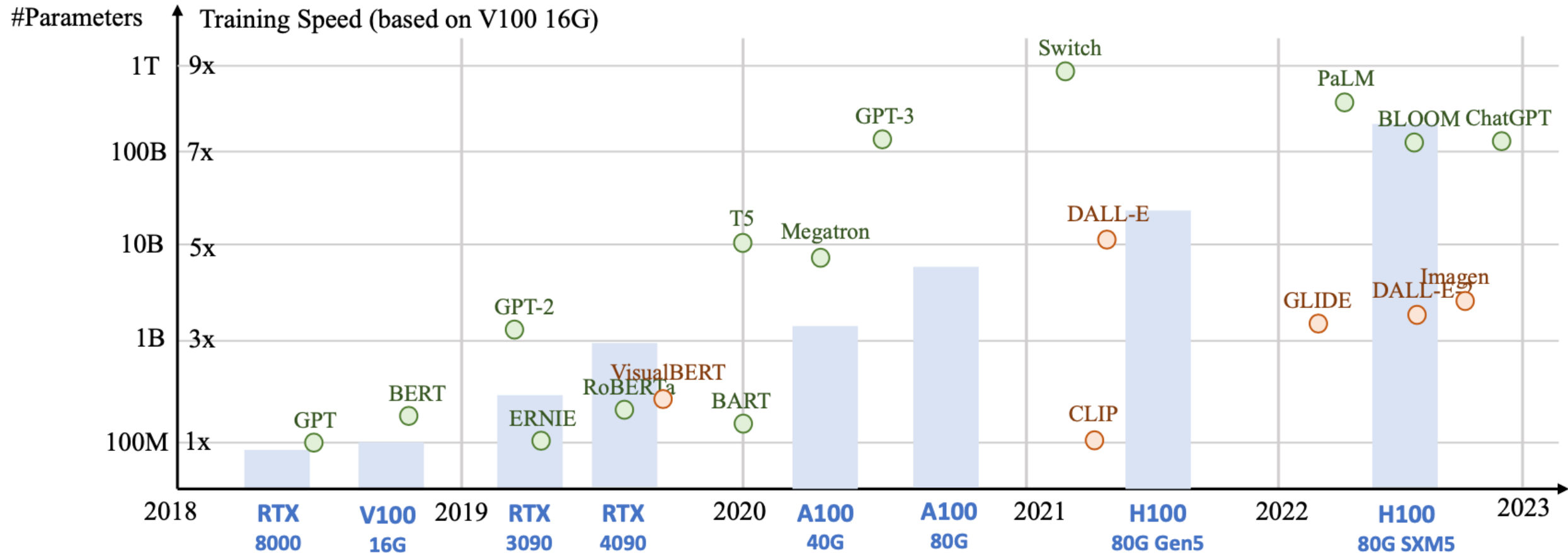
The history of Generative AI in CV, NLP and VL

Unimodal- CV & NLP

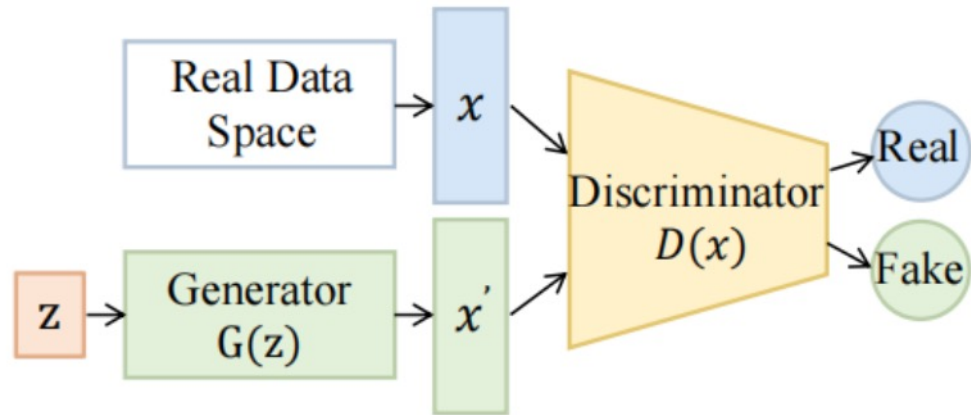


Multimodal – Vision Language

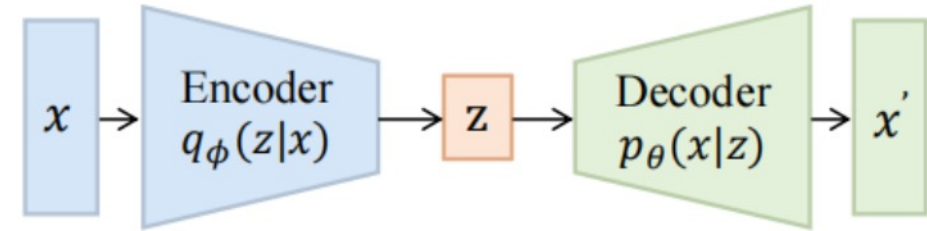
Generative AI Foundation Models



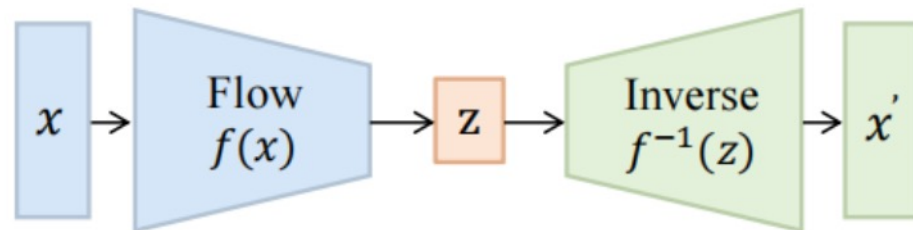
Categories of Vision Generative Models



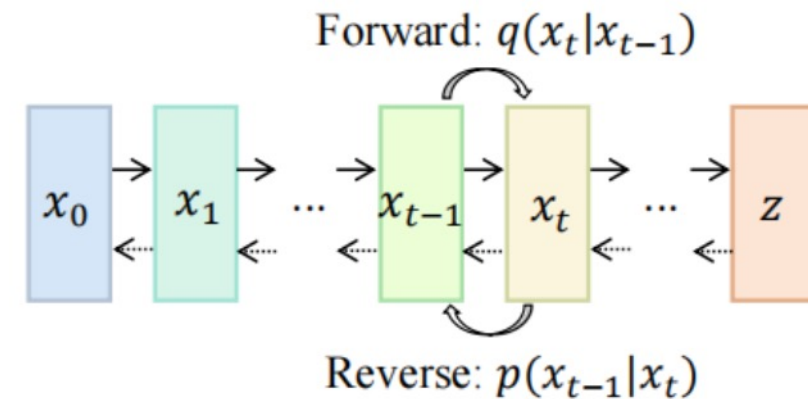
(1) Generative adversarial networks



(2) Variational autoencoders

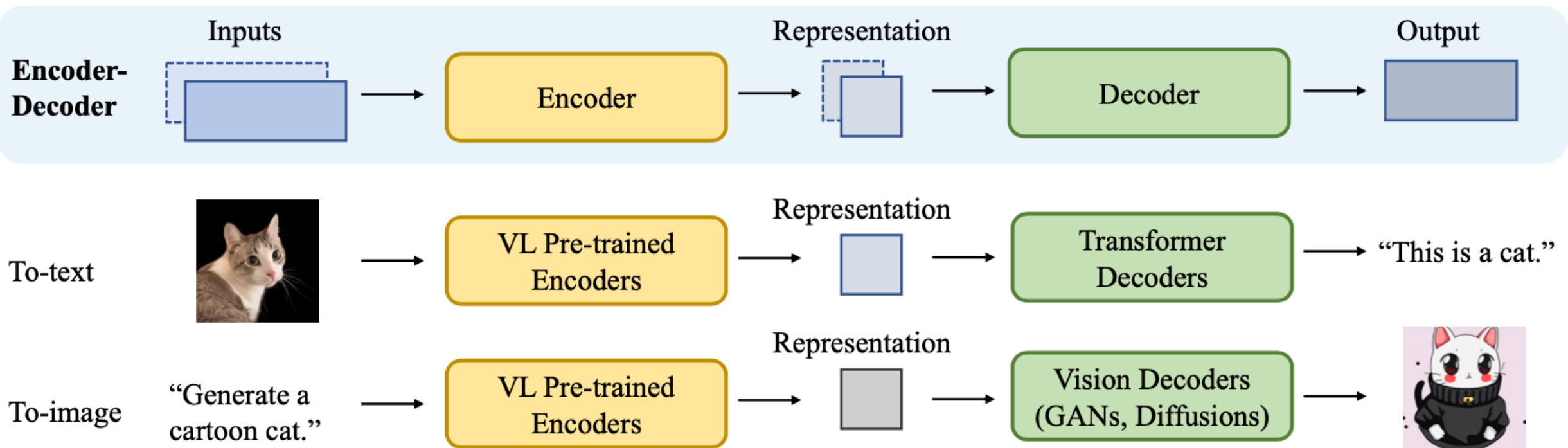


(3) Normalizing flows

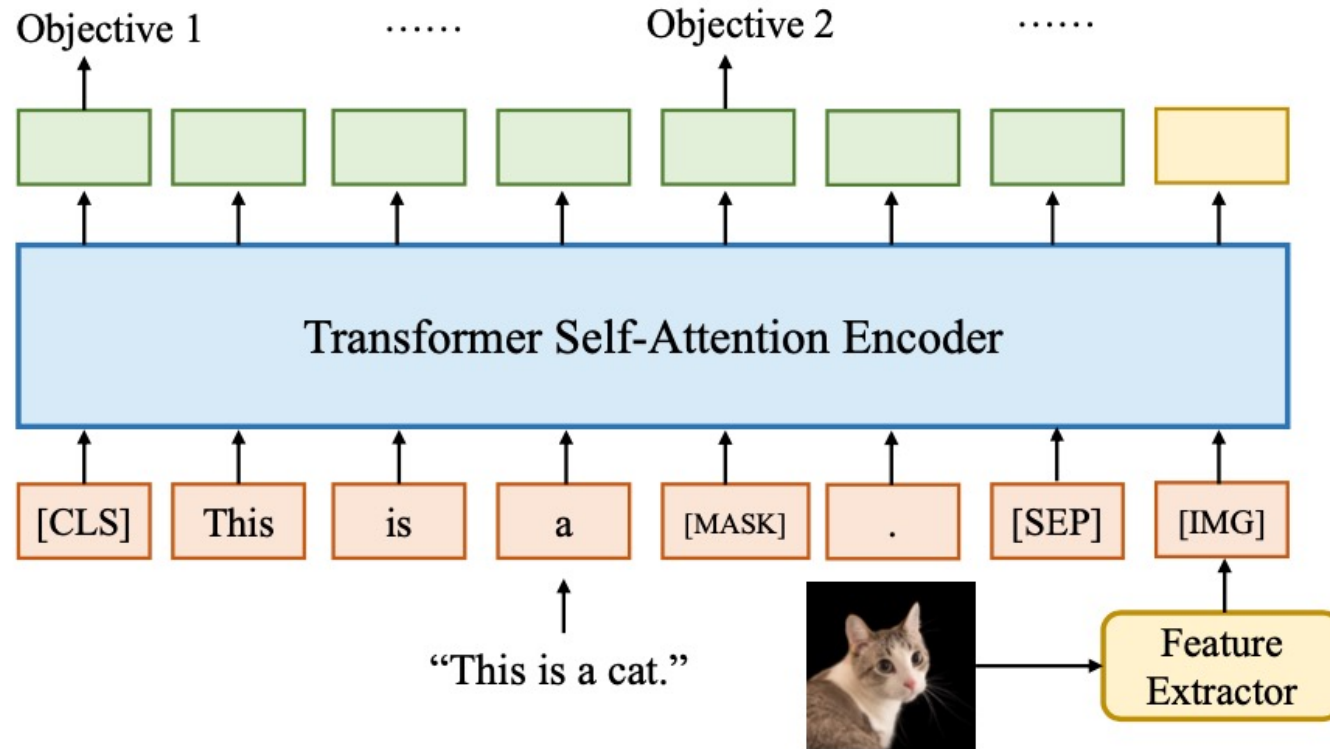


(4) Diffusion models

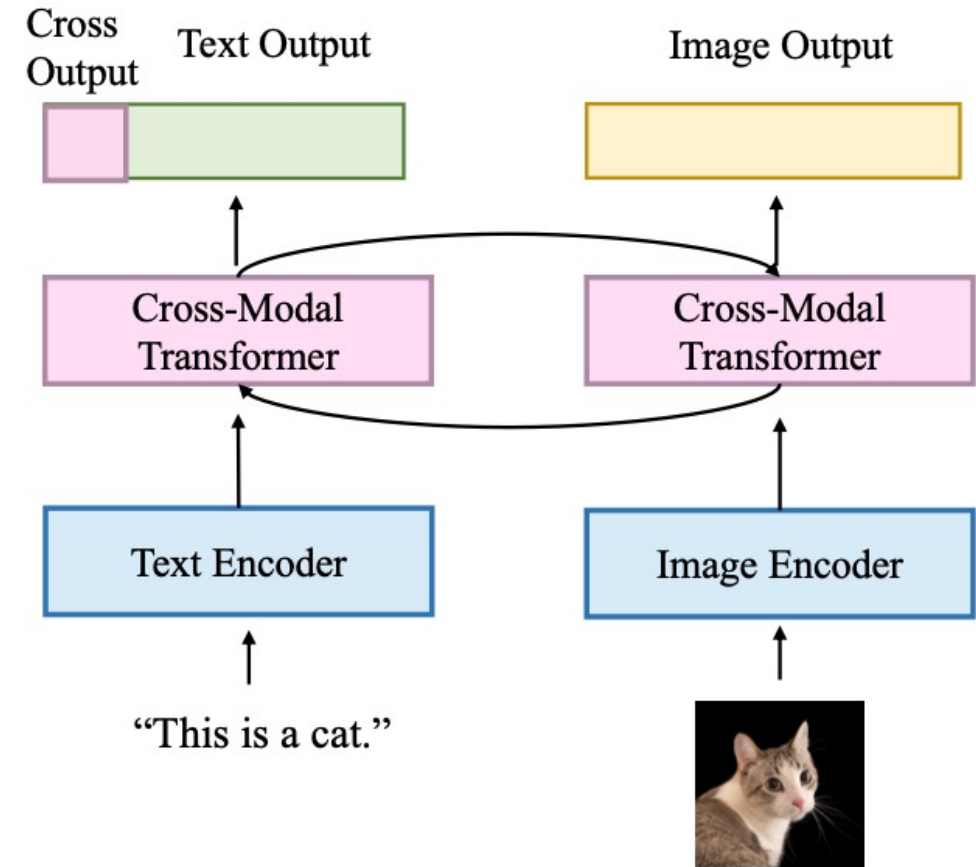
The General Structure of Generative Vision Language



Two Types of Vision Language Encoders: Concatenated Encoders and Cross-aligned Encoders

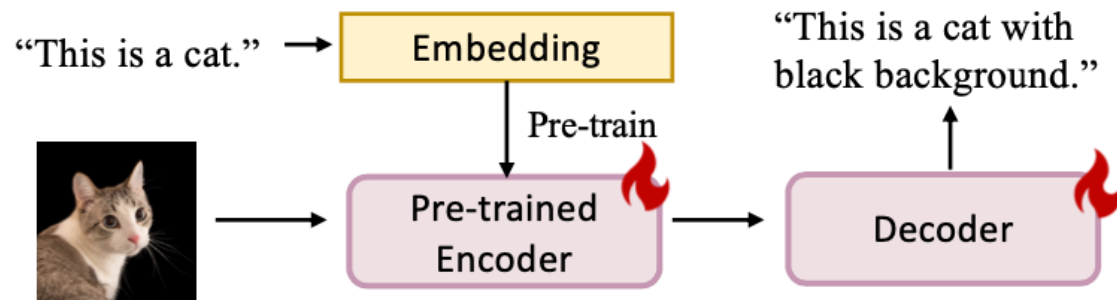


(a) Concatenated Encoder

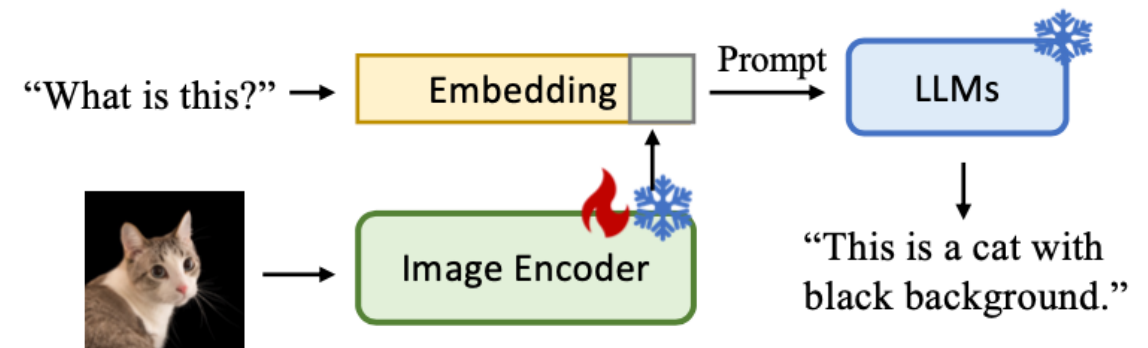


(b) Cross-aligned Encoder

Two Types of to-language Decoder Models: Jointly-trained Models and Frozen Models



(a) Jointly-trained Models

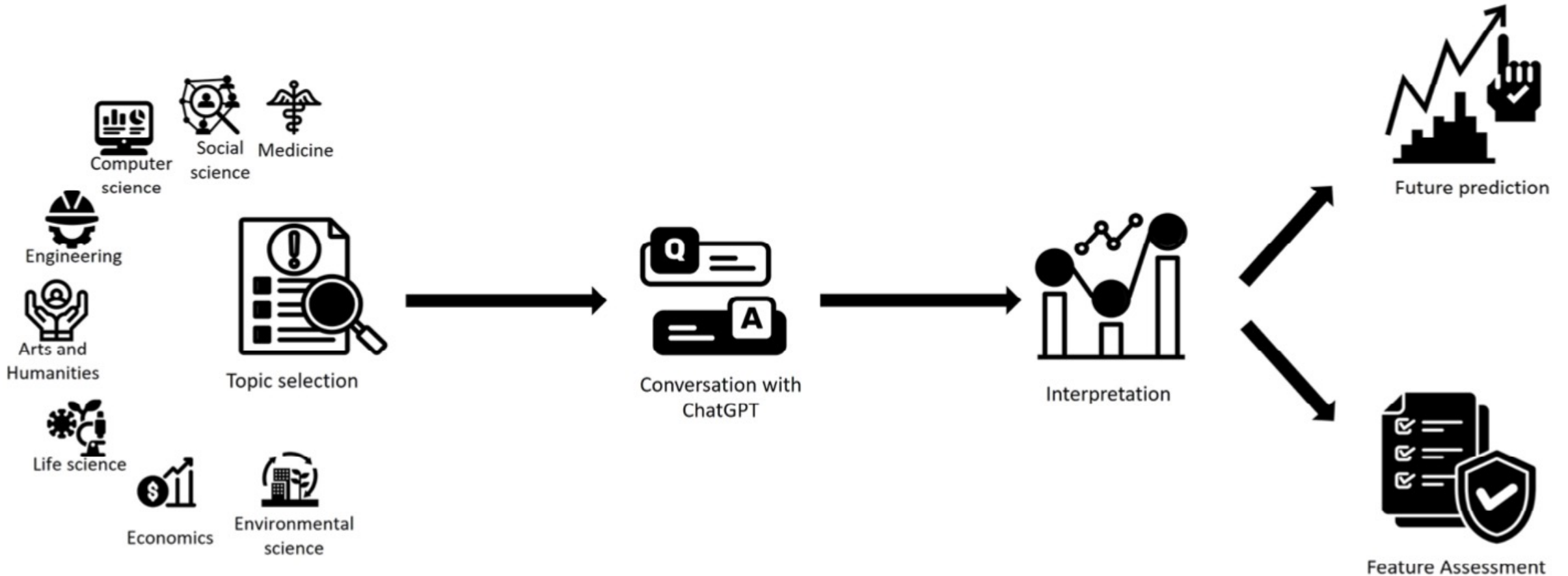


(b) Frozen Models

Generative AI

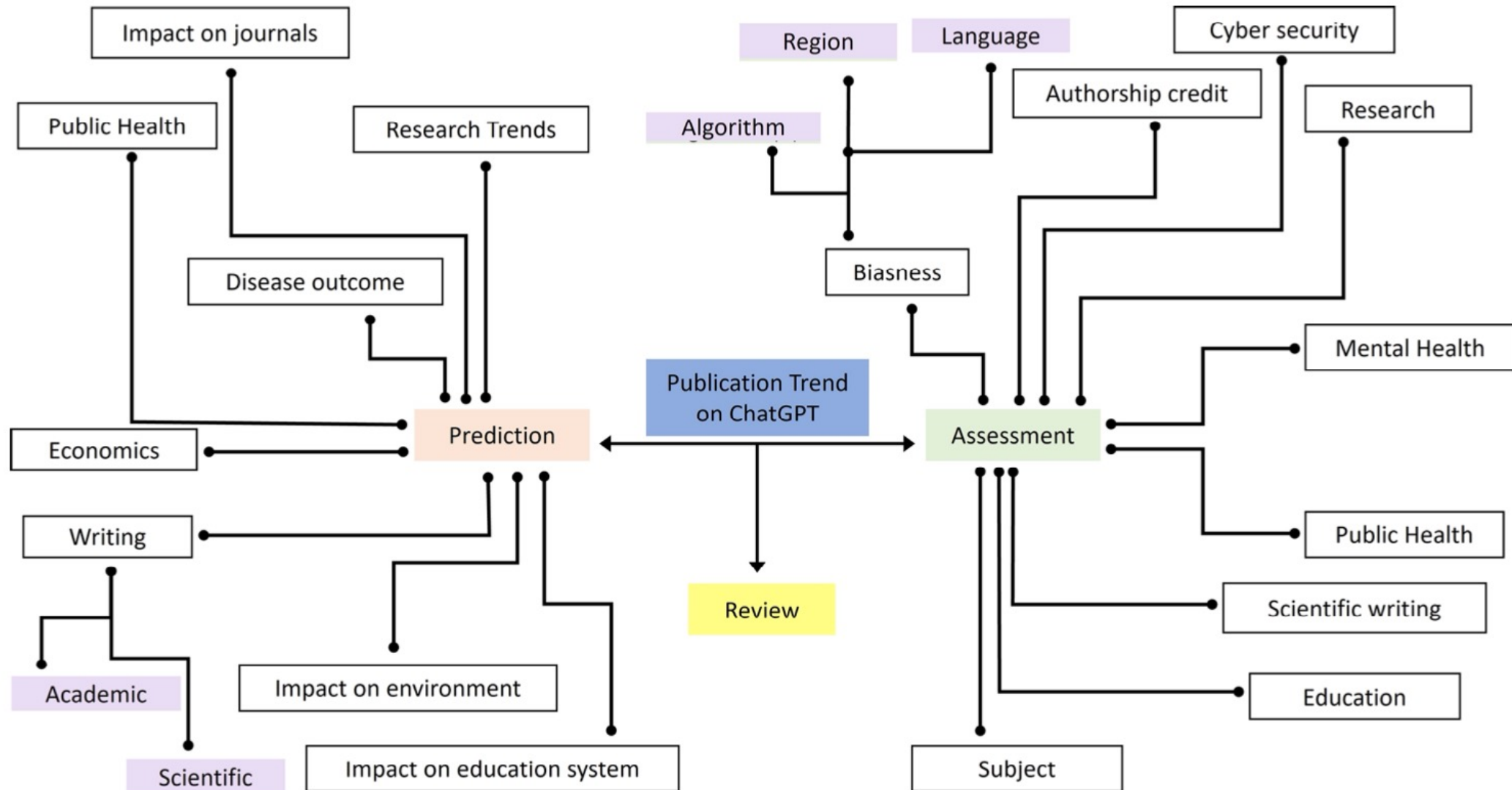
**Text, Image, Video, Audio
Applications**

ChatGPT Research

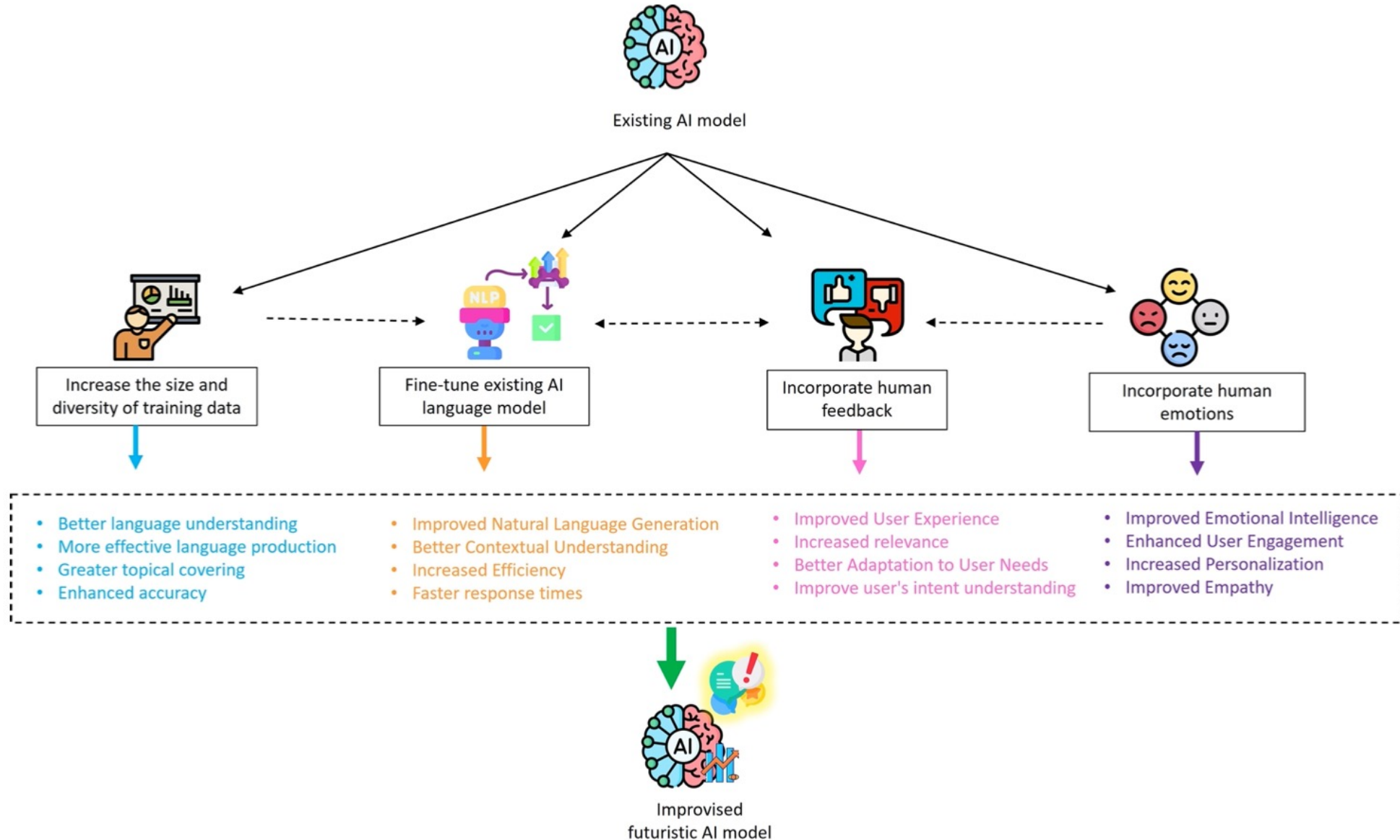


A common trend identified in the reported chatGPT research

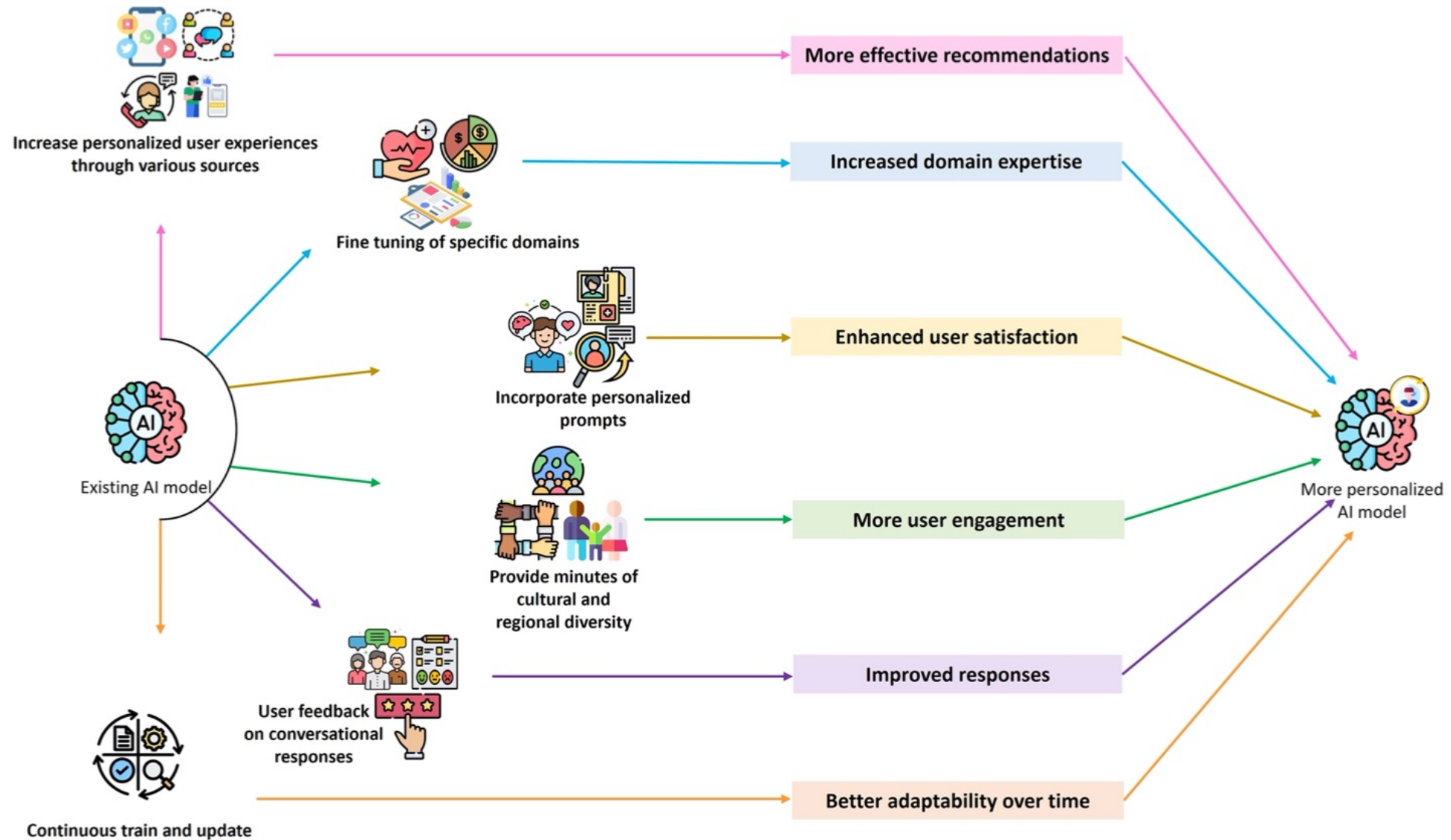
Taxonomy of Literature on ChatGPT



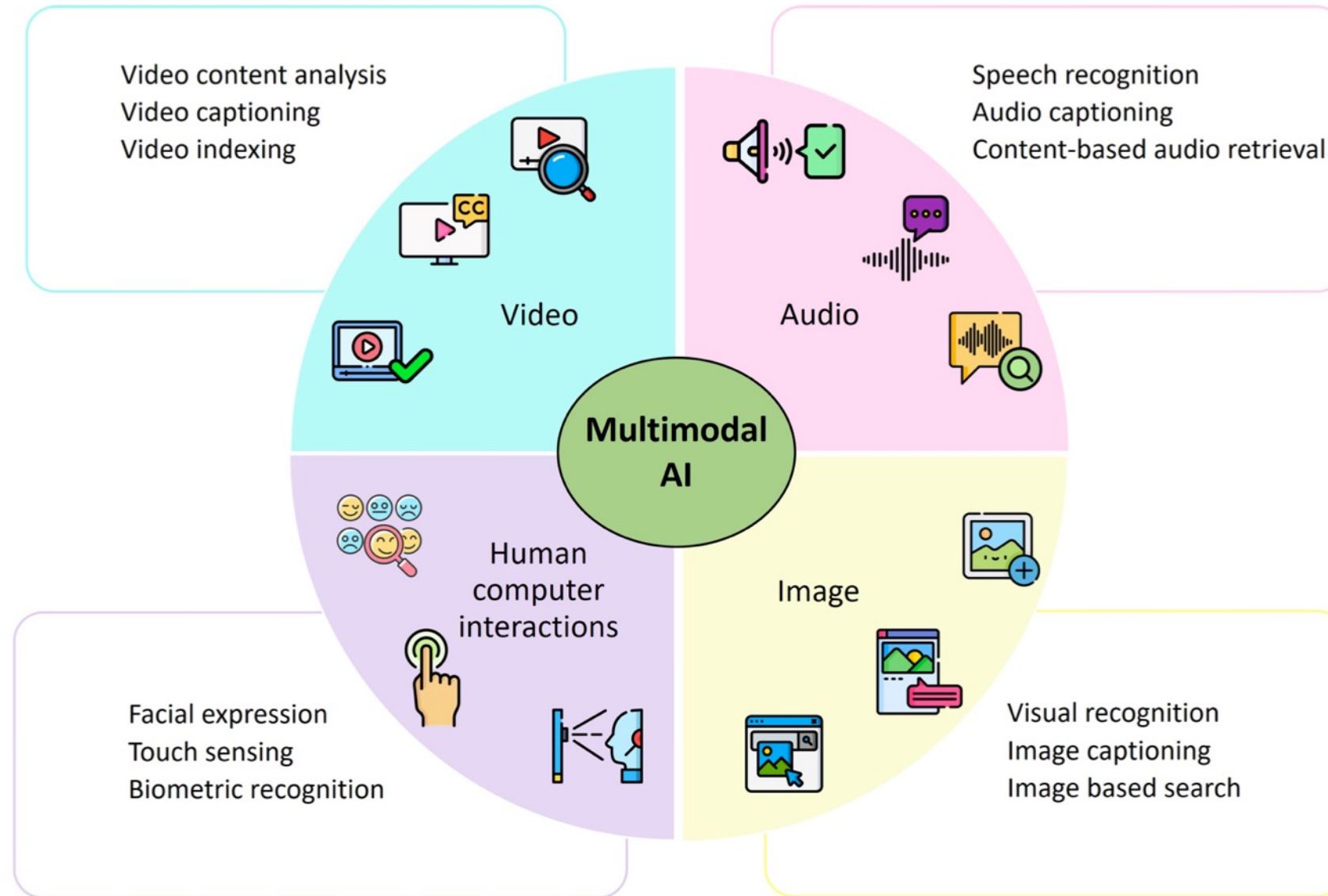
Enhancing the Conversational Ability of ChatGPT



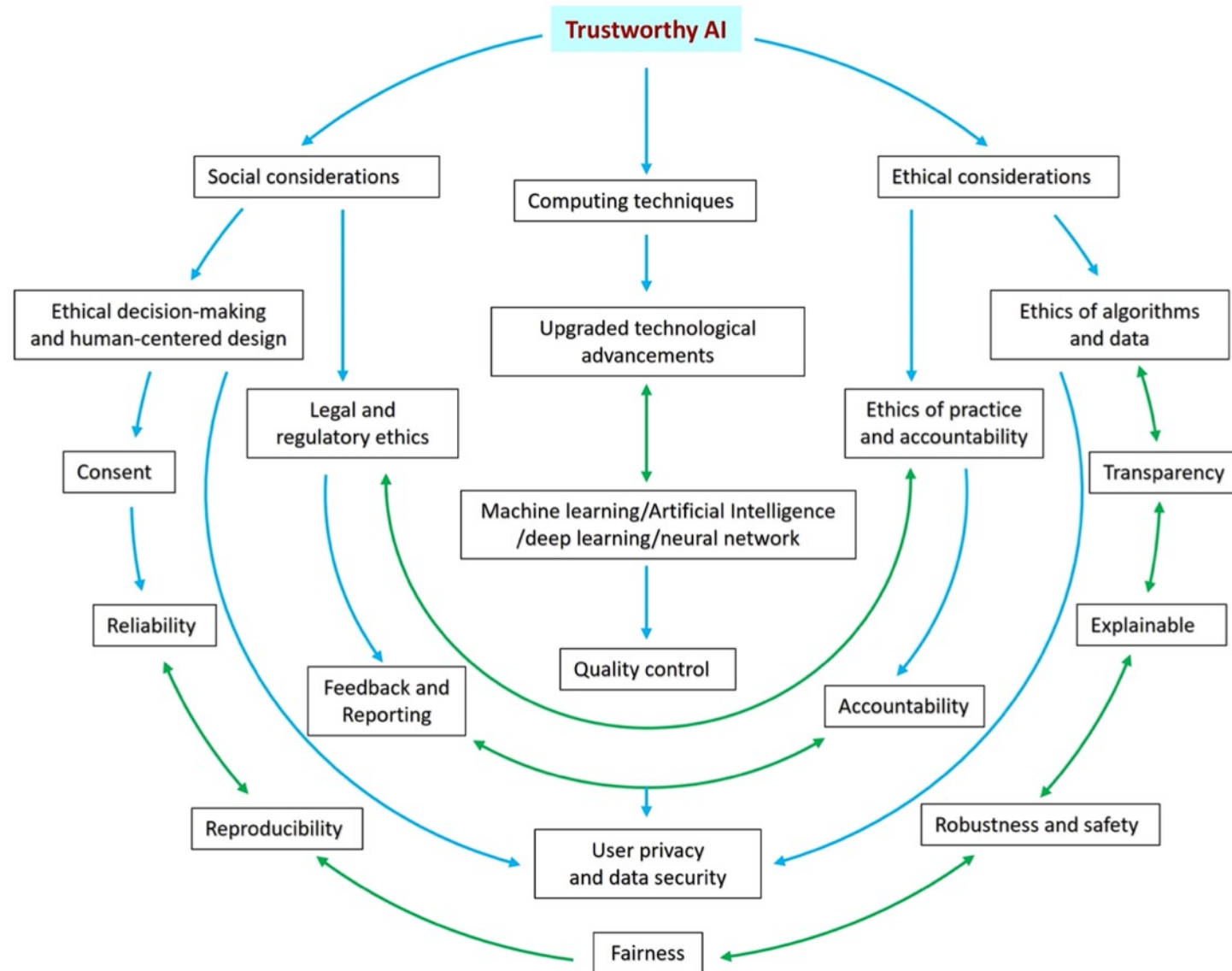
Domain/Person Specific Personalization of ChatGPT



Technological Integration for Multimodal AI



Trustworthy AI: Interplay of Various Factors

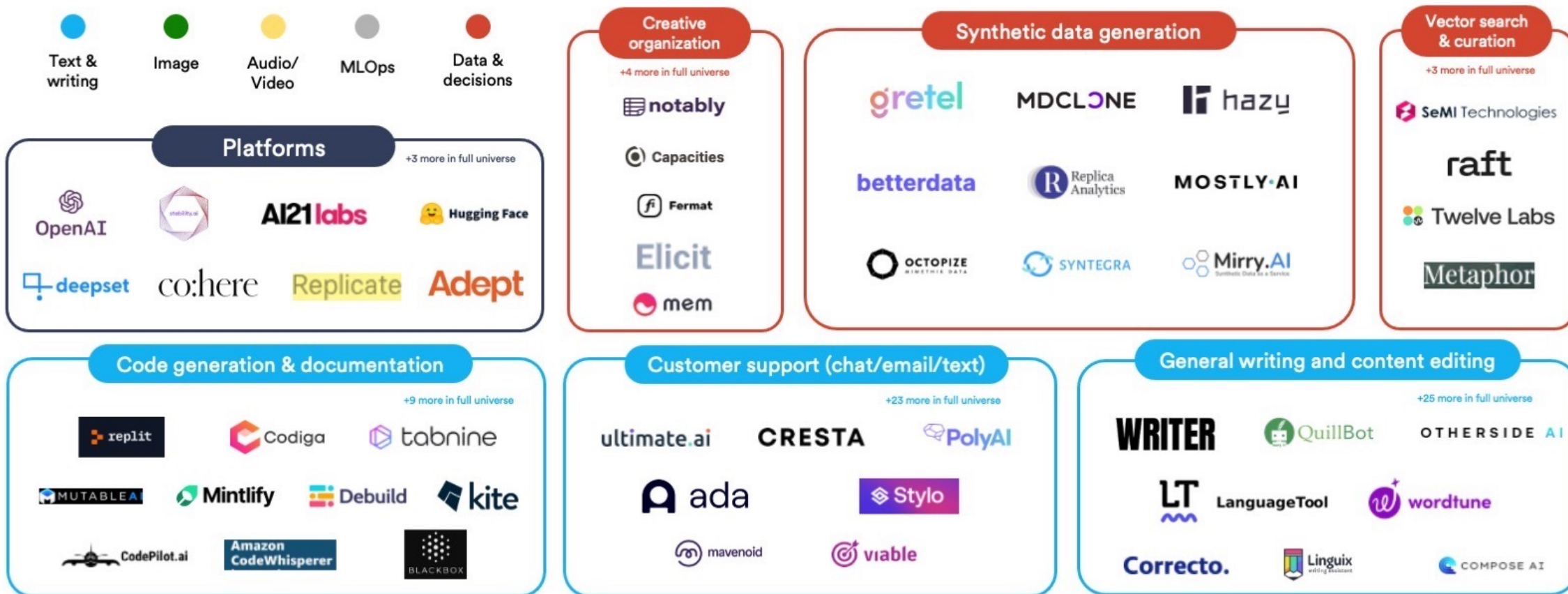


Generative AI

BASE10 TREND MAP: GENERATIVE AI

Companies are grouped based on medium produced and segmented by use case within each medium. Companies that offer products across segments are grouped in the segment of the core product offering.

Base¹⁰



Generative AI

Marketing & sales copy



Text & data summarization

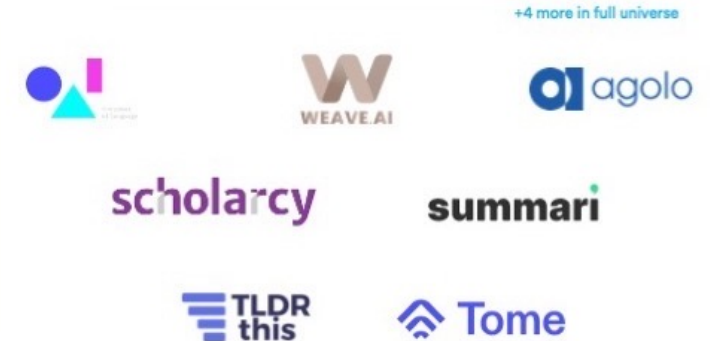


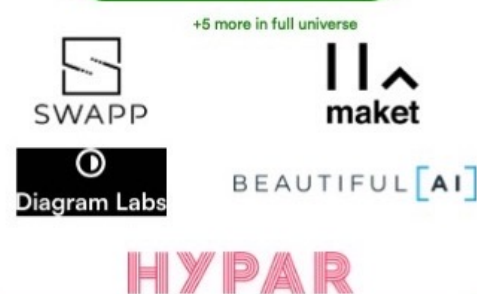
Image editing



Ad collateral



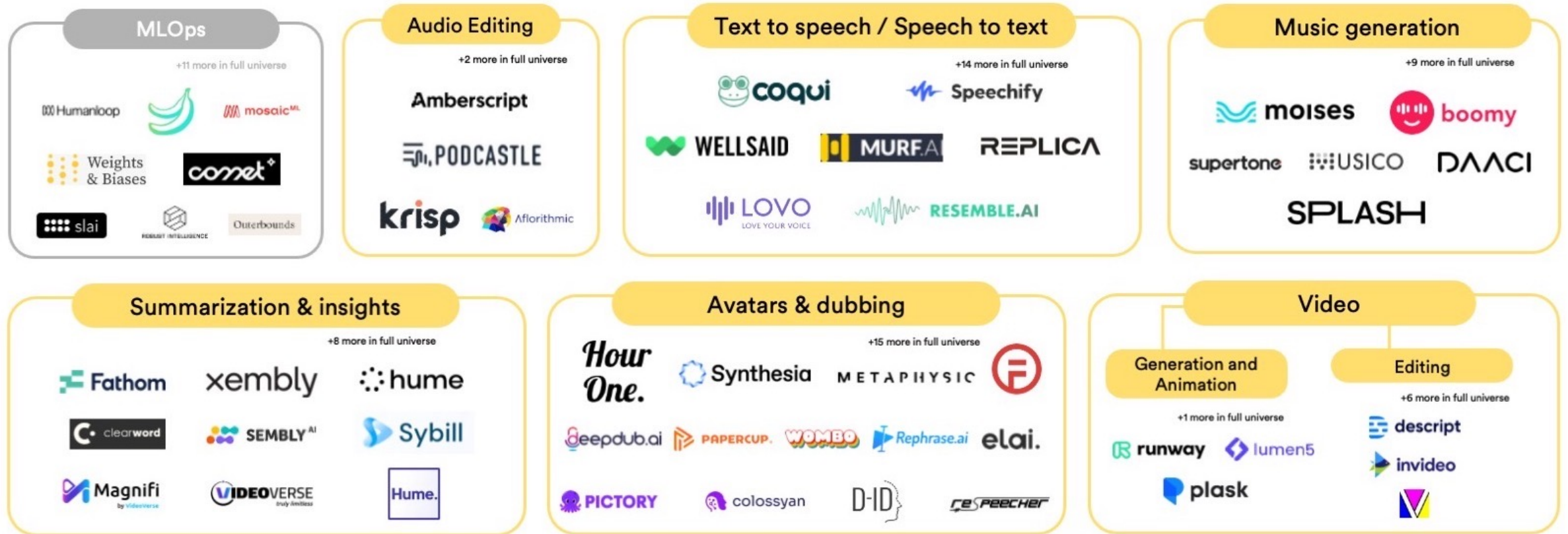
Design



Text to image



Generative AI



DALL·E 2

Create original, realistic images and art from a text description.
It can combine concepts, attributes, and styles.

TEXT DESCRIPTION

An astronaut Teddy bears A
bowl of soup

riding a horse lounging in a
tropical resort in space playing
basketball with cats in space

in a photorealistic style in the
style of Andy Warhol as a pencil
drawing



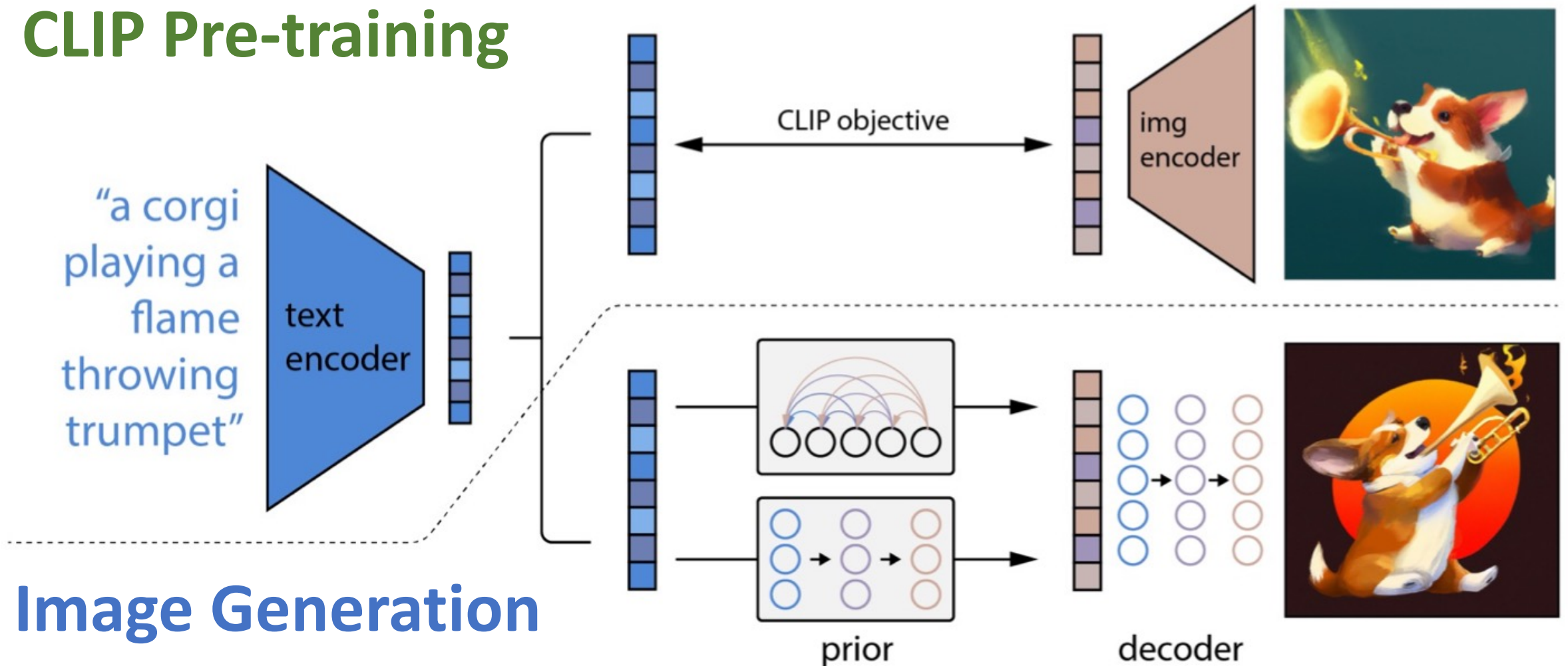
DALL·E 2



<https://openai.com/dall-e-2/>

The Model Structure of DALL-E-2

CLIP Pre-training



Stable Diffusion



Hugging Face

Search models, datasets, users...



Models



Datasets



Spaces



Docs



Solutions

Pricing



Spaces: stabilityai/

stable-diffusion



like 1.89k



Running



App



Files



Community 241



Linked Models

Stable Diffusion Demo

Stable Diffusion is a state of the art text-to-image model that generates images from text.

For faster generation and forthcoming API access you can try [DreamStudio Beta](#)

an insect robot preparing a delicious meal

Generate image



<https://huggingface.co/spaces/stabilityai/stable-diffusion>

Stable Diffusion Colab

woctezuma / stable-diffusion-colab Public

Notifications

Fork 7

Star 31

<> Code Issues Pull requests Actions Projects Wiki Security Insights

main

1 branch 0 tags

Go to file

Code



woctezuma README: add a reference for sampler schedules

37bc02d 24 days ago 18 commits



LICENSE

Initial commit

27 days ago



README.md

README: add a reference for sampler schedules

24 days ago



stable_diffusion.ipynb

Allow to choose the scheduler

25 days ago

README.md

Stable-Diffusion-Colab

The goal of this repository is to provide a Colab notebook to run the text-to-image "Stable Diffusion" model [1].

Usage

- Run `stable_diffusion.ipynb` . [Open in Colab](#)

About

Colab notebook to run Stable Diffusion.

github.com/CompVis/stable-diffusion

deep-learning colab image-generation

text-to-image diffusion text2image

colaboratory google-colab

colab-notebook google-colaboratory

google-colab-notebook

text-to-image-synthesis huggingface

diffusion-models

text-to-image-generation latent-diffusion

stable-diffusion huggingface-diffusers

diffusers stable-diffusion-diffusers

Readme

MIT license

31 stars

2 watching

<https://github.com/woctezuma/stable-diffusion-colab>

Stable Diffusion Reimagine



Clipdrop ▶ Stable diffusion Reimagine
by stability.ai

Apps ▾

API

Blog

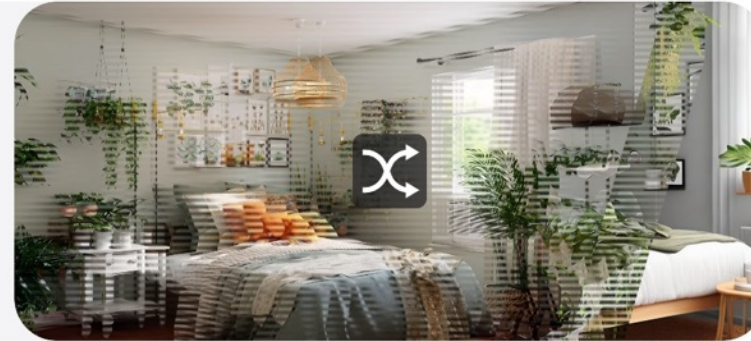
Pricing

Sign-in / Sign-up



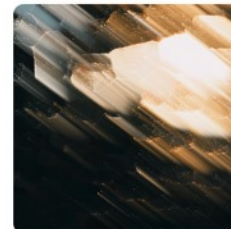
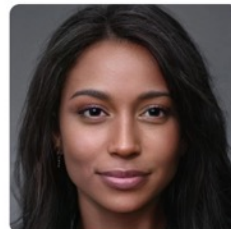
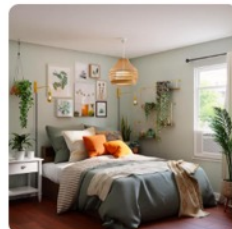
Stable diffusion reimagine

Create multiple variations from a single image.



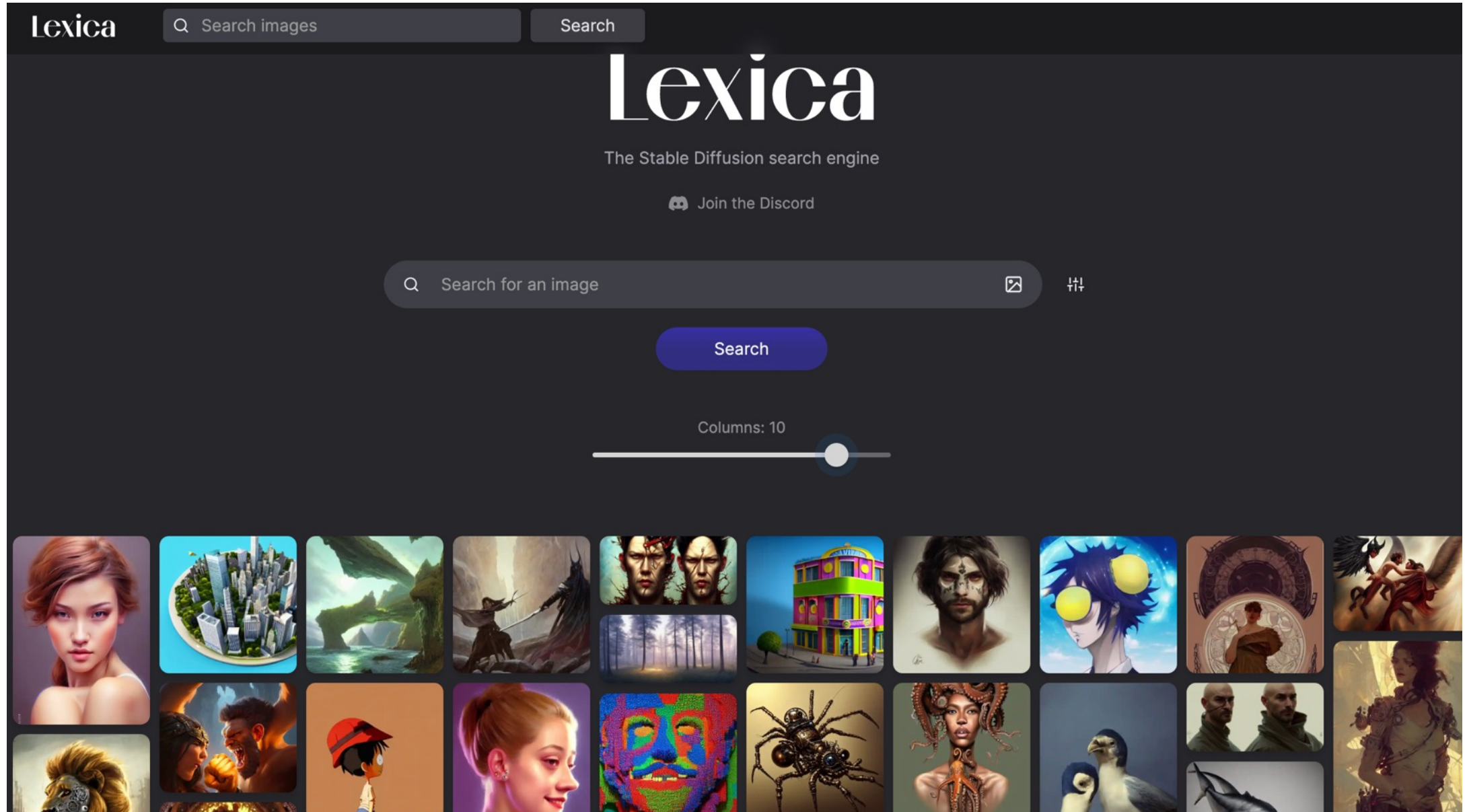
Click, paste, or drop a file here to start.

↓ Or click on an example below



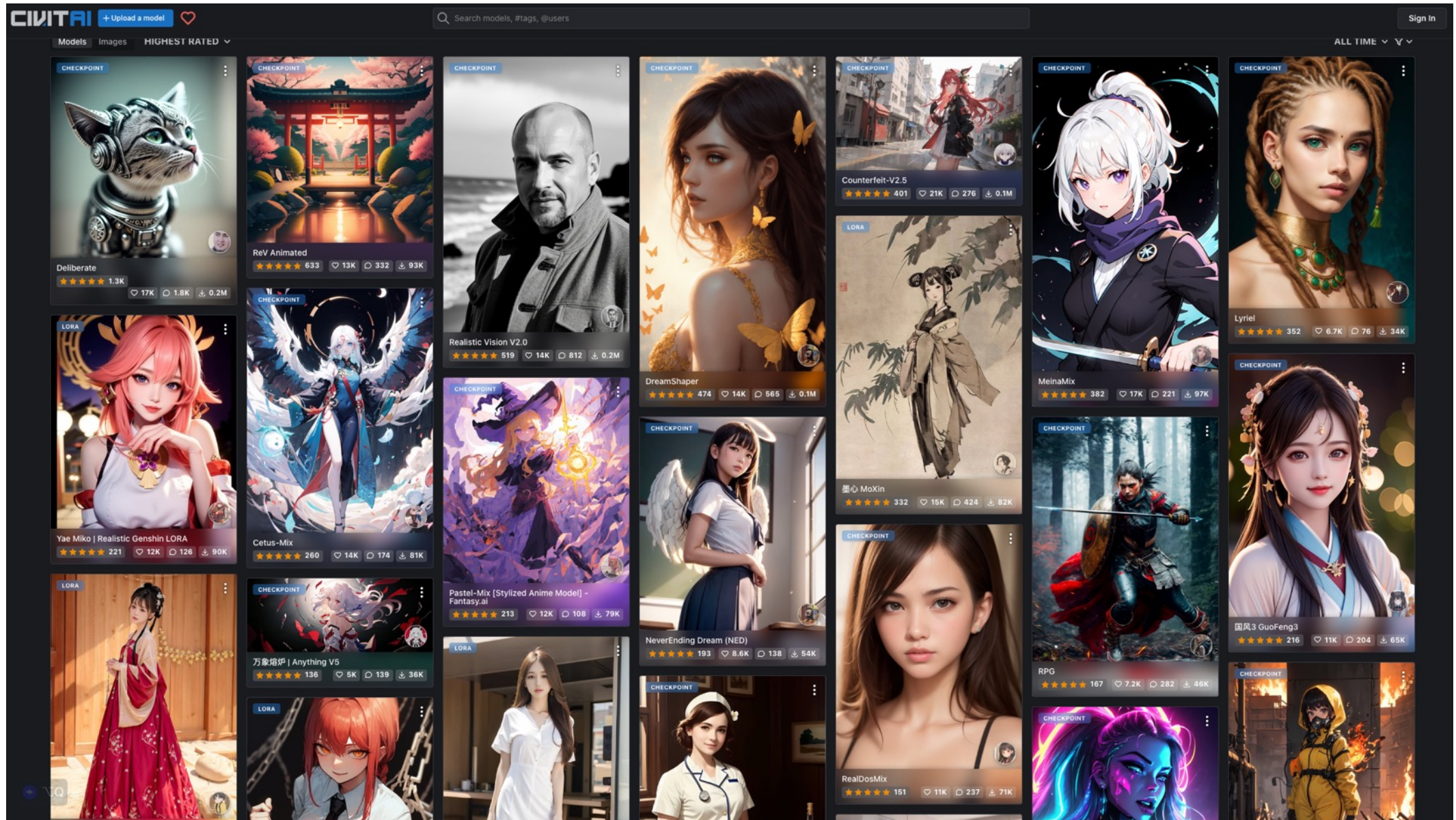
<https://clipdrop.co/stable-diffusion-reimagine>

Lexica Art: Search Stable Diffusion images and prompts



<https://lexica.art/>

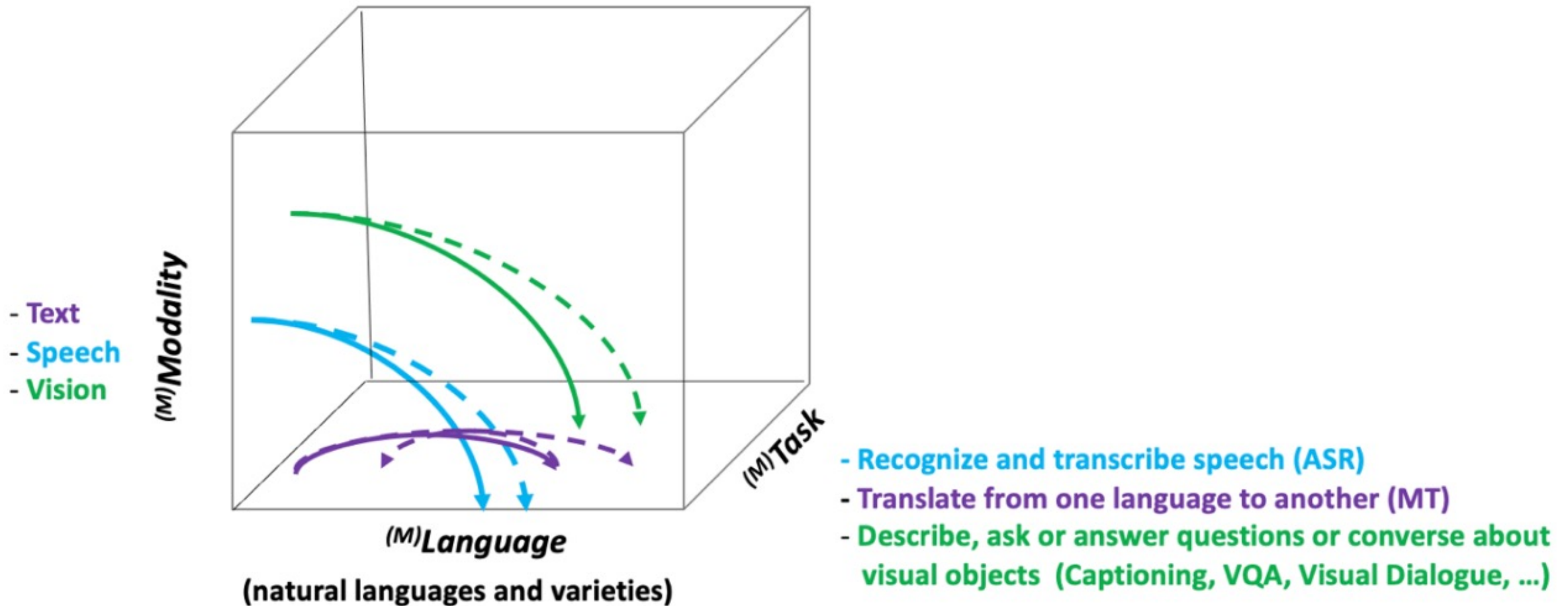
Civitai: Stable Diffusion AI Art Models



<https://civitai.com/>

NLG from a Multilingual, Multimodal and Multi-task perspective

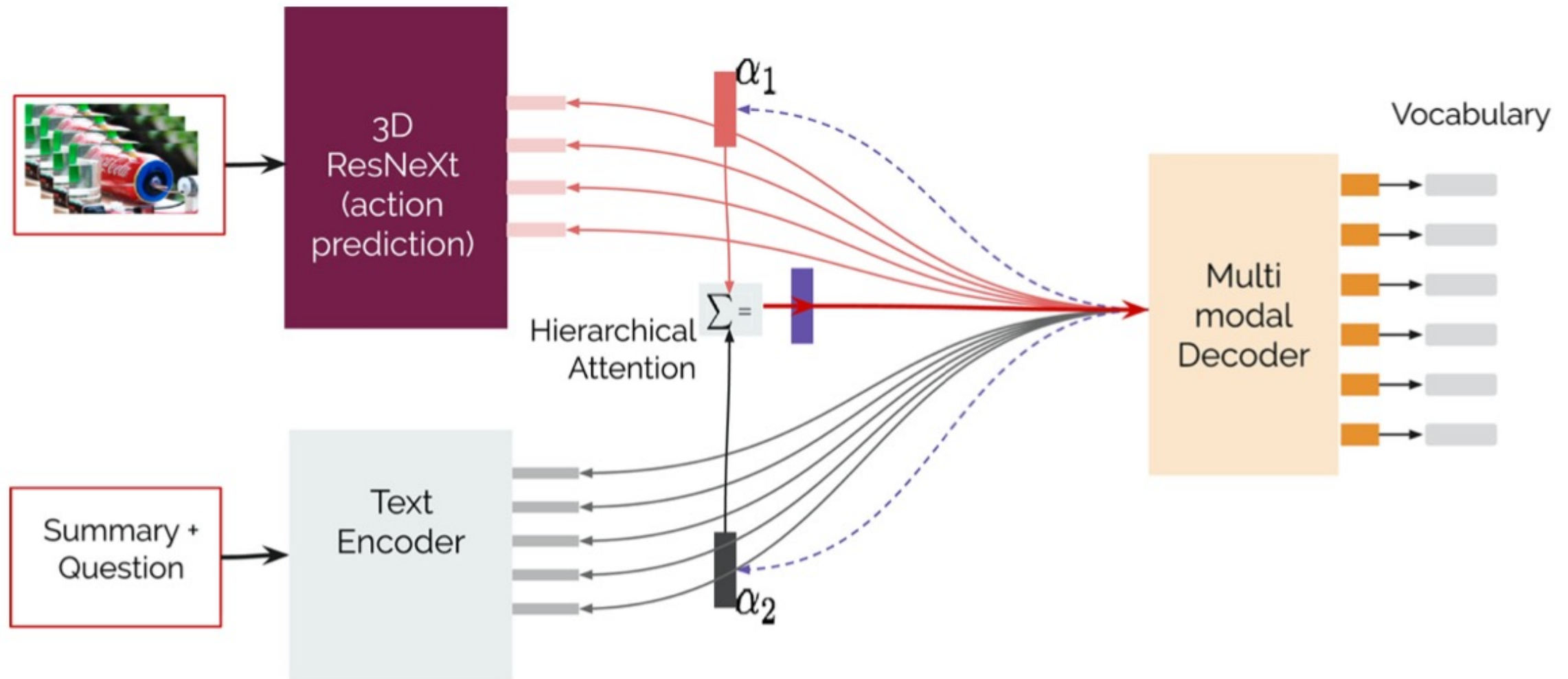
Multi³(Natural Language) Generation



Source: Erdem, Erkut, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara Plank, Andrii Babii et al.

"Neural Natural Language Generation: A Survey on Multilinguality, Multimodality, Controllability and Learning." Journal of Artificial Intelligence Research 73 (2022): 1131-1207.

Text-and-Video Dialog Generation Models with Hierarchical Attention



Source: Erdem, Erkut, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara Plank, Andrii Babii et al.

"Neural Natural Language Generation: A Survey on Multilinguality, Multimodality, Controllability and Learning." Journal of Artificial Intelligence Research 73 (2022): 1131-1207.

Multimodal Few-Shot Learning with Frozen Language Models

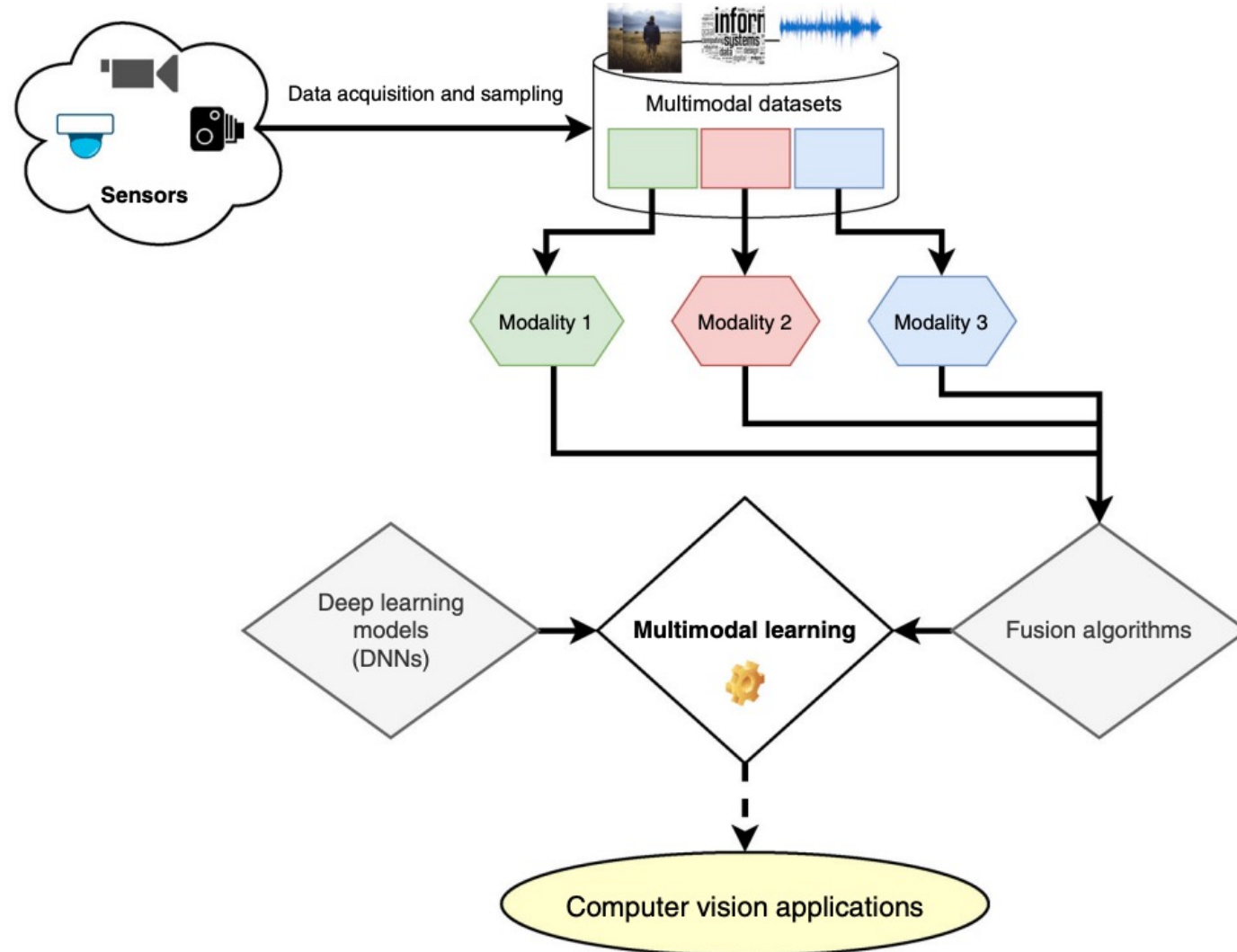


Curated samples with about five seeds required to get past well-known language model failure modes of either repeating text for the prompt or emitting text that does not pertain to the image.

These samples demonstrate the ability to generate open-ended outputs that adapt to both images and text, and to make use of facts that it has learned during language-only pre-training.

Multimodal Pipeline

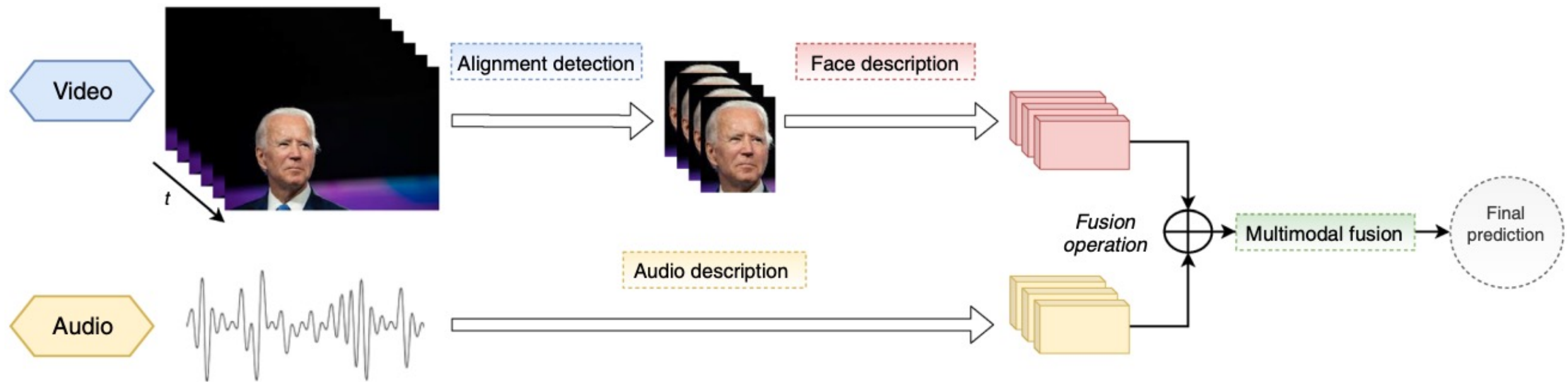
that includes three different modalities (Image, Text, Audio)



Source: Bayoudh, Khaled, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa (2022).

"A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets." The Visual Computer 38, no. 8: 2939-2970.

Video and Audio Multimodal Fusion



Source: Bayoudh, Khaled, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa (2022).

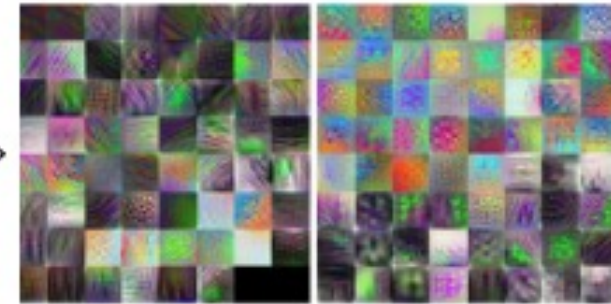
"A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets." The Visual Computer 38, no. 8: 2939-2970.

Visual and Textual Representation

Image



Visual representations (Dense)



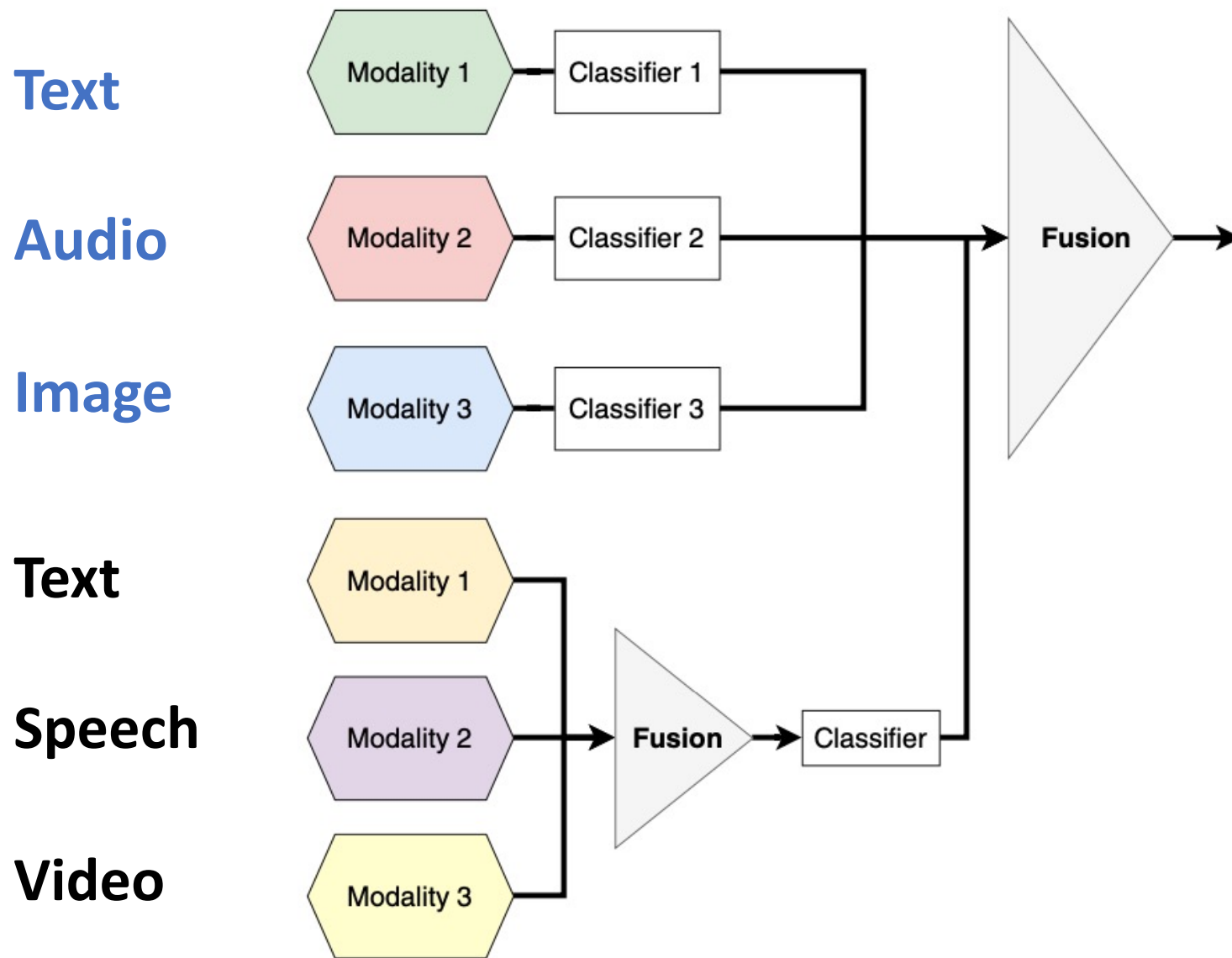
Text

This is the oldest and most important defensive work to have been built along the North African coastline by the Arab conquerors in the early days of Islam. Founded in 796, this building underwent several modifications during the medieval period. Initially, it formed a quadrilateral and then was composed of four buildings giving onto two inner courtyards.

Textual representations (Sparse)



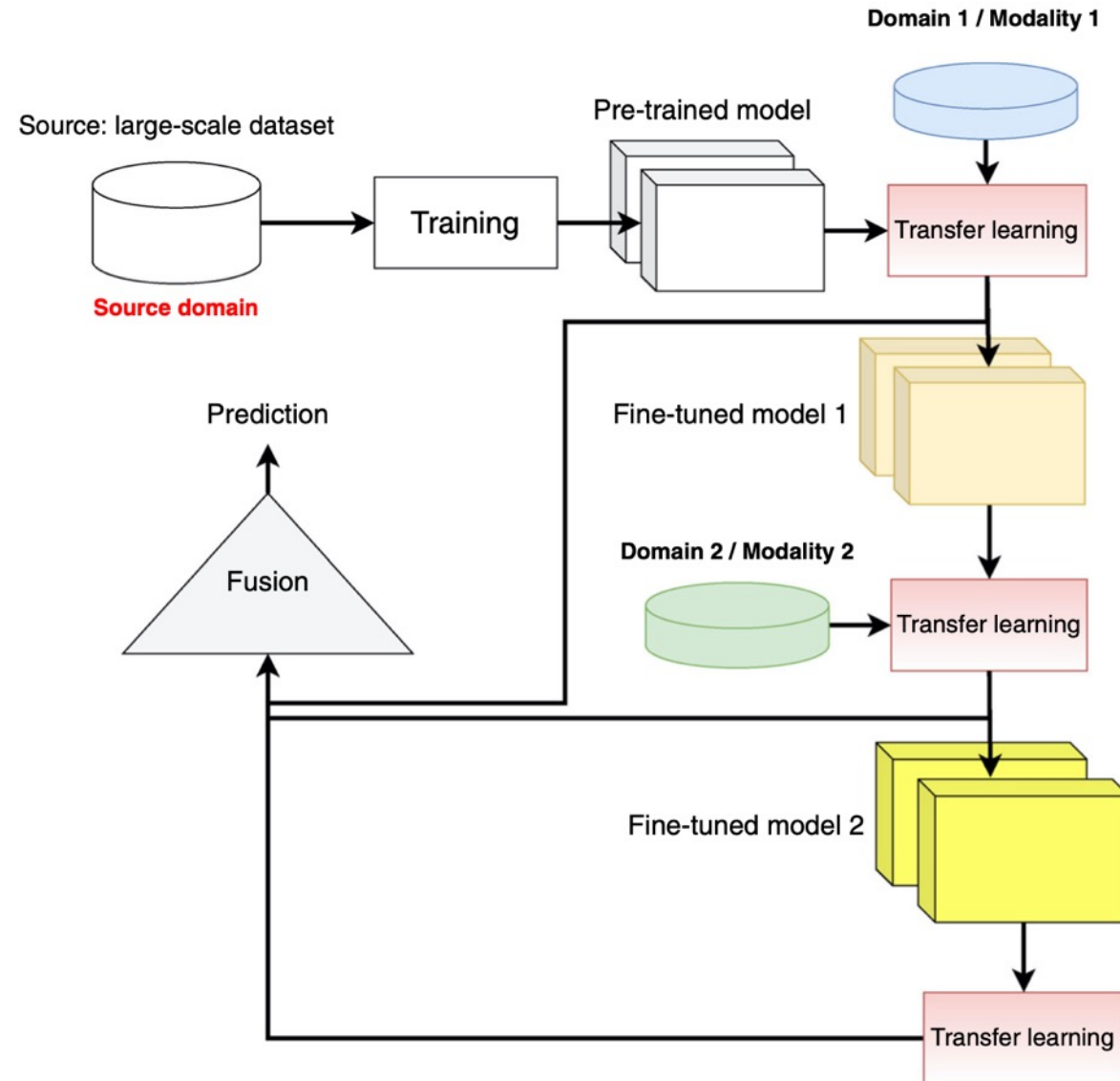
Hybrid Multimodal Data Fusion



Source: Bayoudh, Khaled, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa (2022).

"A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets." The Visual Computer 38, no. 8: 2939-2970.

Multimodal Transfer Learning

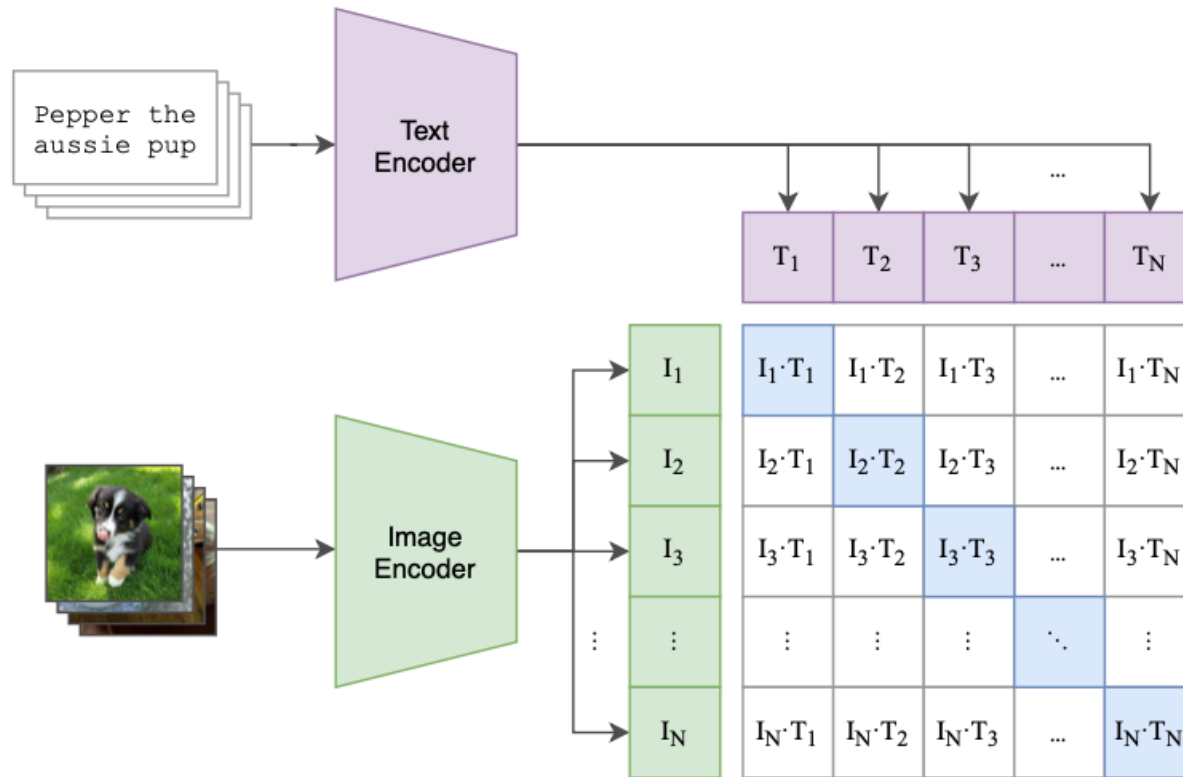


Source: Bayoudh, Khaled, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa (2022).

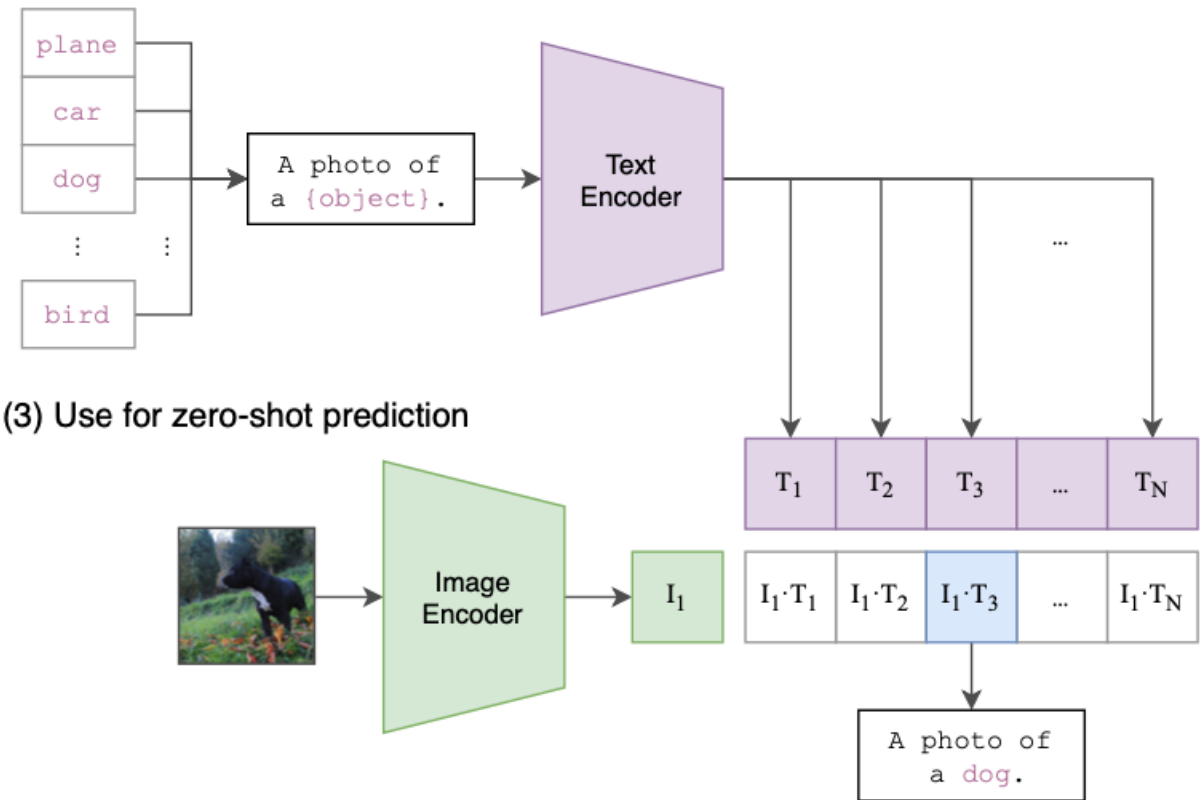
"A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets." The Visual Computer 38, no. 8: 2939-2970.

CLIP: Learning Transferable Visual Models From Natural Language Supervision

(1) Contrastive pre-training

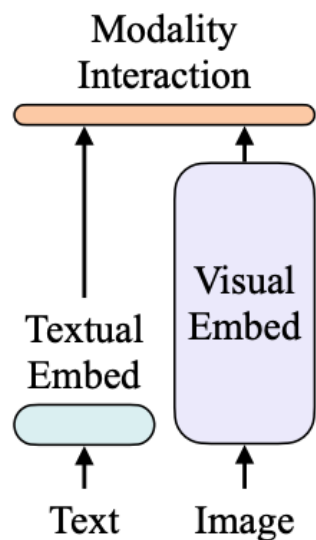


(2) Create dataset classifier from label text

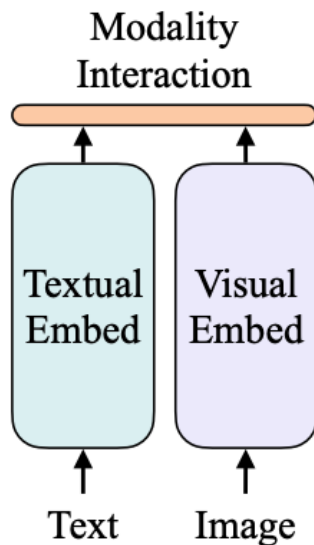


ViLT: Vision-and-Language Transformer

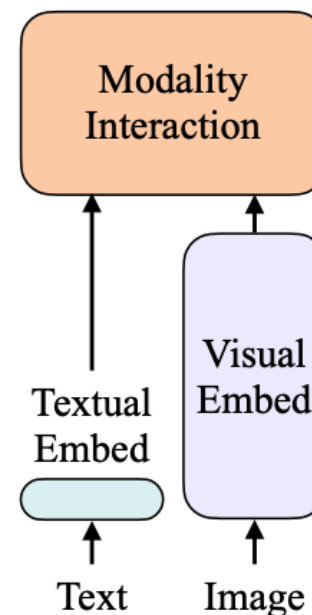
Without Convolution or Region Supervision



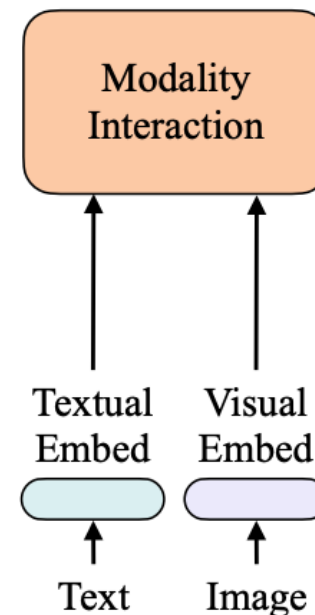
(a) $VE > TE > MI$



(b) $VE = TE > MI$



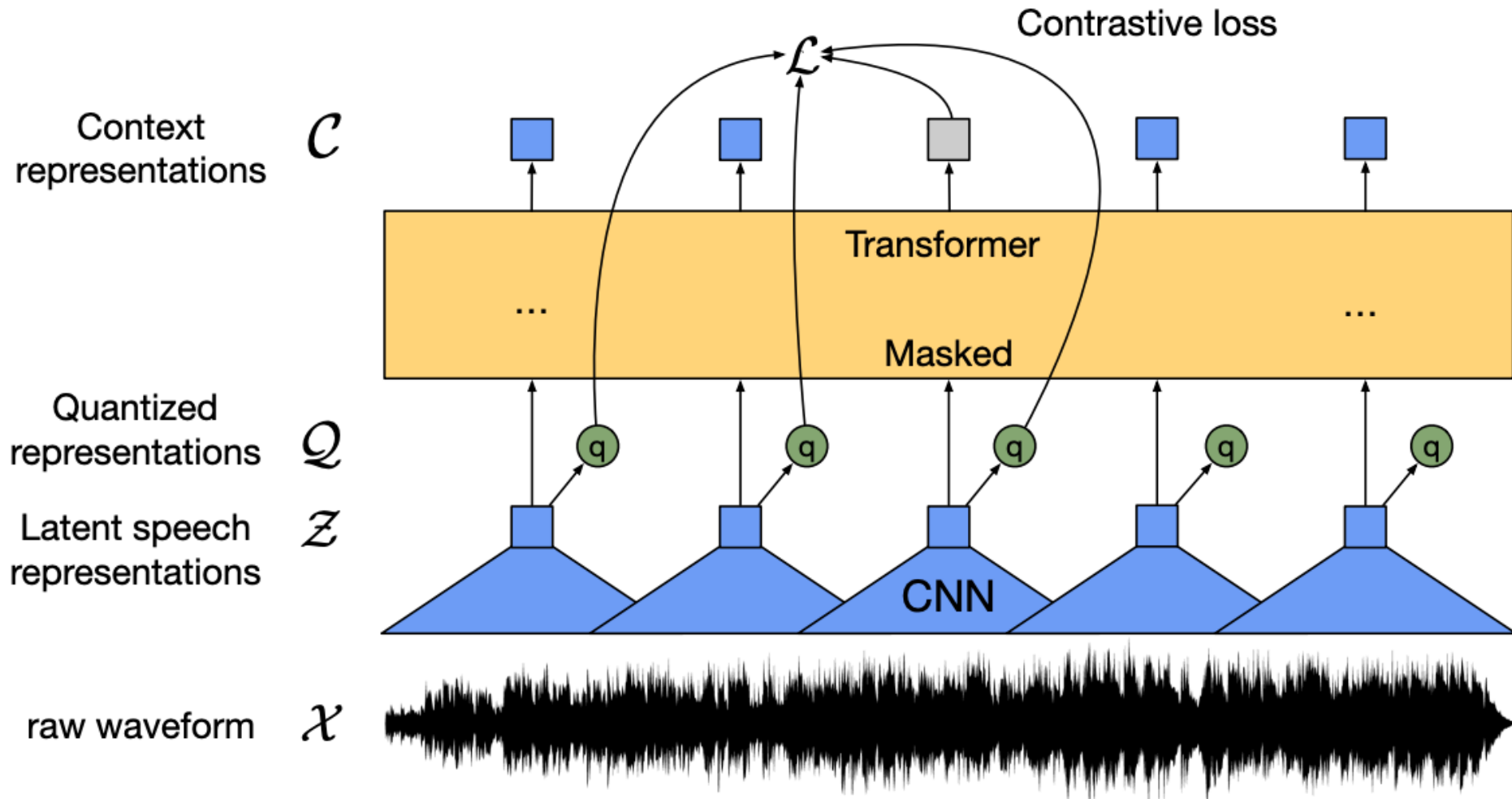
(c) $VE > MI > TE$



(d) $MI > VE = TE$

wav2vec 2.0:

A framework for self-supervised learning of speech representations

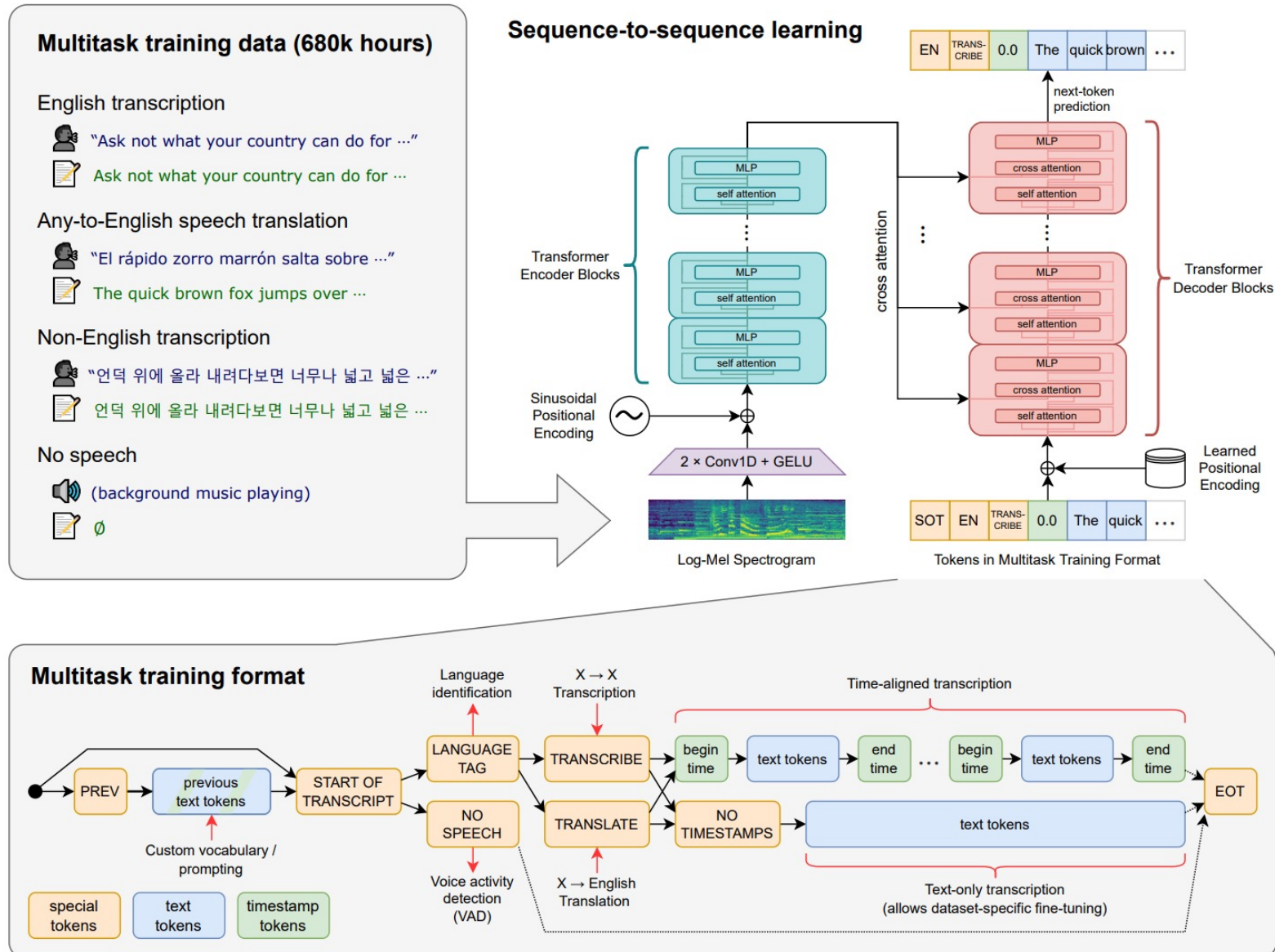


Source: Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli.

"wav2vec 2.0: A framework for self-supervised learning of speech representations." Advances in Neural Information Processing Systems 33 (2020): 12449-12460.

Whisper:

Robust Speech Recognition via Large-Scale Weak Supervision



Microsoft Azure Text to Speech (TTS)

Text SSML

You can replace this text with any text you wish. You can either write in this text box or paste your own text here.

Try different languages and voices. Change the speed and the pitch of the voice. You can even tweak the SSML (Speech Synthesis Markup Language) to control how the different sections of the text sound. Click on SSML above to give it a try!

Enjoy using Text to Speech!

Language

English (United States) ▾

Voice

Jenny (Neural) ▾

Speaking style

General ▾

Speaking speed: 1.00



Pitch: 0.00



Play

Hugging Face



Hugging Face

🔍 Search models, datasets, spaces

📦 Models

📄 Datasets

🏠 Spaces

📄 Docs

📁 Solutions

Pricing



Log In

Sign Up



The AI community building the future.

Build, train and deploy state of the art models powered by
the reference open source in machine learning.

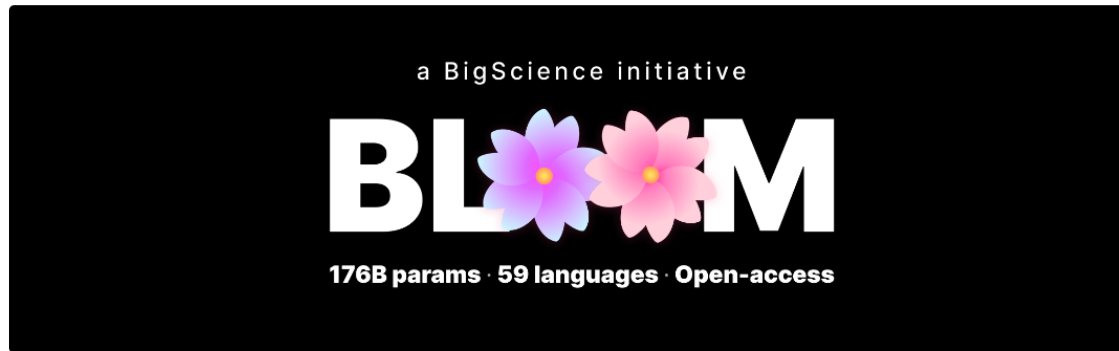
🌟 Star

58,696

<https://huggingface.co/>

BLOOM

BigScience Large Open-science Open-access Multilingual Language Model



BigScience Large Open-science Open-access Multilingual Language Model

Version 1.3 / 6 July 2022

Current Checkpoint: **Training Iteration 95000**

Total seen tokens: **366B**

Downloads last month
12,875



⚡ Hosted inference API ⓘ

📄 Text Generation

Groups ▼

Examples ▼

I love bloom. Super simple, but so effective! I went through a similar process a couple of years ago when I

sampling ☒ greedy

ⓘ [BLOOM prompting tips](#)

Switch to "greedy" for more accurate completion e.g. math/history/translations (but which may be repetitive/less inventive)

Compute

⌘+Enter

1.3

Source: <https://huggingface.co/bigscience/bloom>

OpenAI Whisper



Hugging Face

Models

Datasets

Spaces

Docs

Solutions

Pricing



Spaces: openai/whisper



like 422

Running

App

Files



Community 49

Whisper

Whisper is a general-purpose speech recognition model. It is trained on a large dataset of diverse audio and is also a multi-task model that can perform multilingual speech recognition as well as speech translation and language identification. This demo cuts audio after around 30 secs.

You can skip the queue by using google colab for the space:



Open in Colab



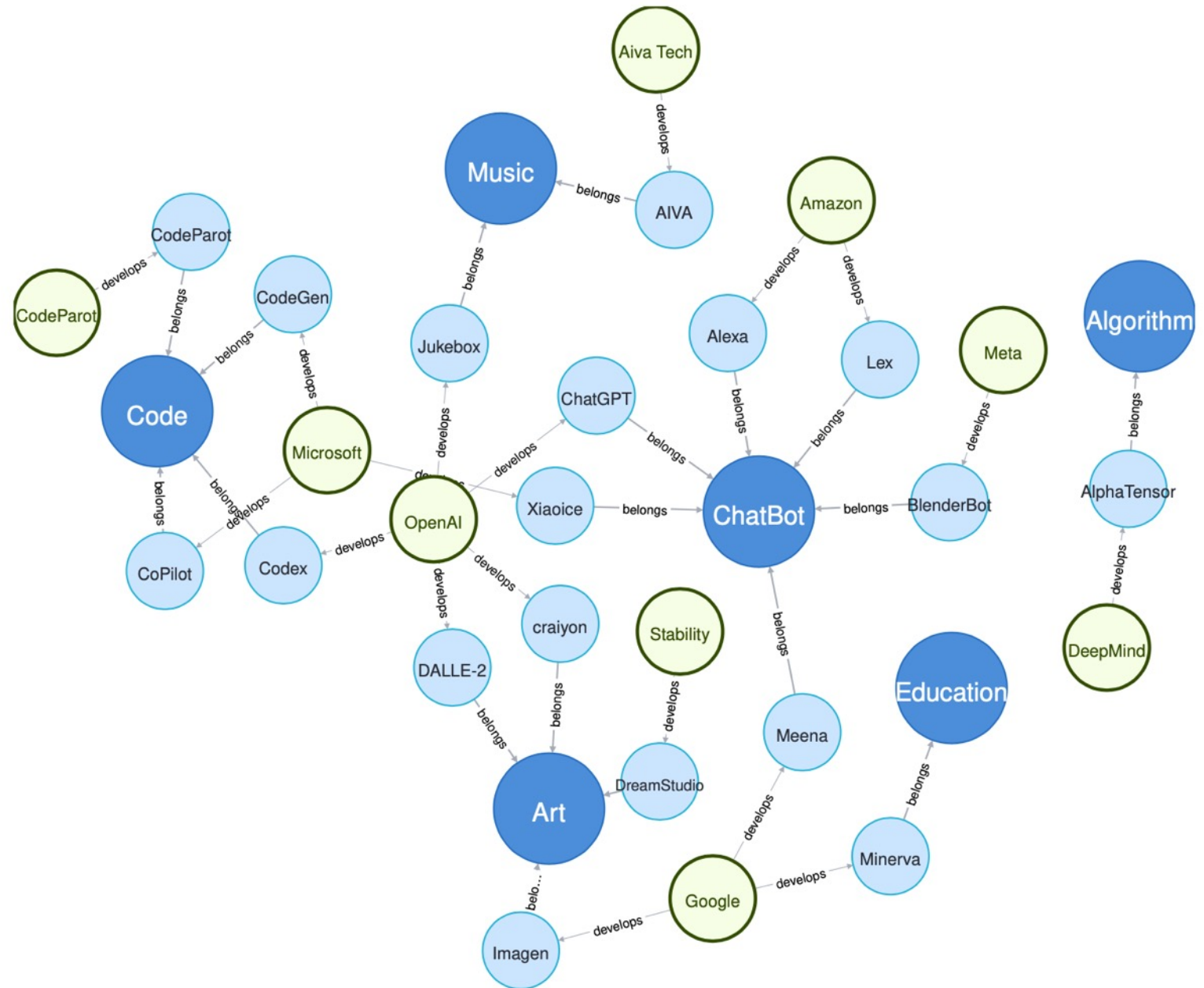
0:05 / 0:05



Transcribe

Source: <https://huggingface.co/spaces/openai/whisper>

Generative AI Research Areas, Applications and Companies



Applications of Generative AI Models

Application	Platform/Software	Company	Year	Papaer	Link
ChatBot	Xiaoice	Microsoft	2018	[200]	Xiaoice
ChatBot	Meena	Google	2020	[201]	Meena Blog
ChatBot	BlenderBot	Meta	2022	[202]	Blenderbot
ChatBot	ChatGPT	OpenAI	2022	[10]	ChatGPT
ChatBot	Alexa	Amazon	2014	-	Amazon Alexa
ChatBot	Lex	Amazon	2017	-	Amazon Lex
Music	AIVA	Aiva Tech	2016	-	AIVA
Music	Jukebox	OpenAI	2020	[203]	Jukebox
Code	CodeGPT	Microsoft	2021	[204]	CodeGPT
Code	CodeParrot	CodeParrot	2022	[205]	CodeParrot
Code	Codex	OpenAI	2021	[206]	Codex blog
Code	CoPilot	Microsoft	2021	[206]	CoPilot
Art	DALL-E-2	OpenAI	2022	[5]	DALL-E-2 Blog
Art	DreamStudio	Stability	2022	[13]	Dreamstudio
Art	craiyon	OpenAI	2021	[1]	Craiyon
Art	Imagen	Google	2022	[152]	Imagen
Education	Minerva	Google	2022	[207]	Minerva Blog
Algorithm	AlphaTensor	DeepMind	2022	[208]	AlphaTensor

Outline

- Introduction
- Overview of Generative AI
- **Overview of Large Language Models (LLMs)**
- Foundation of Transformers: Attention Mechanism
- Fine-tuning LLM for Question Answering System
- Fine-tuning LLM for Dialogue System
- Challenges and Limitations of Generative AI for QA and Dialogue Systems
- Q & A

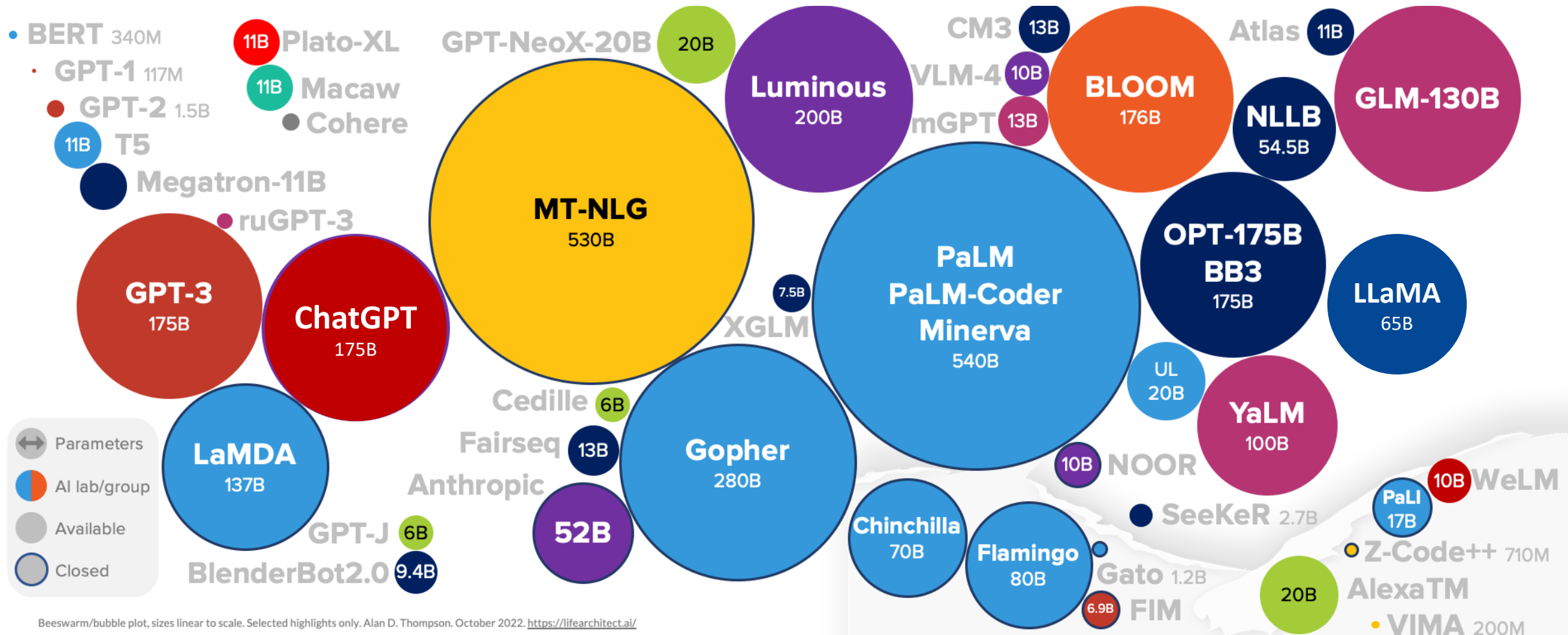
ChatGPT

Large Language Models (LLMs)

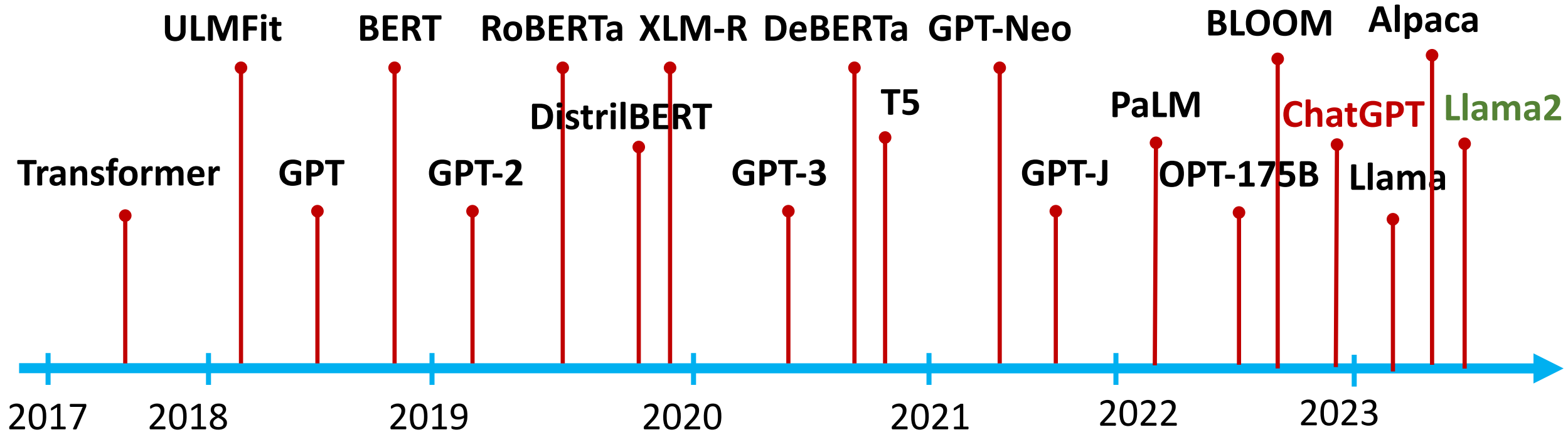
Foundation Models

Large Language Models (LLM)

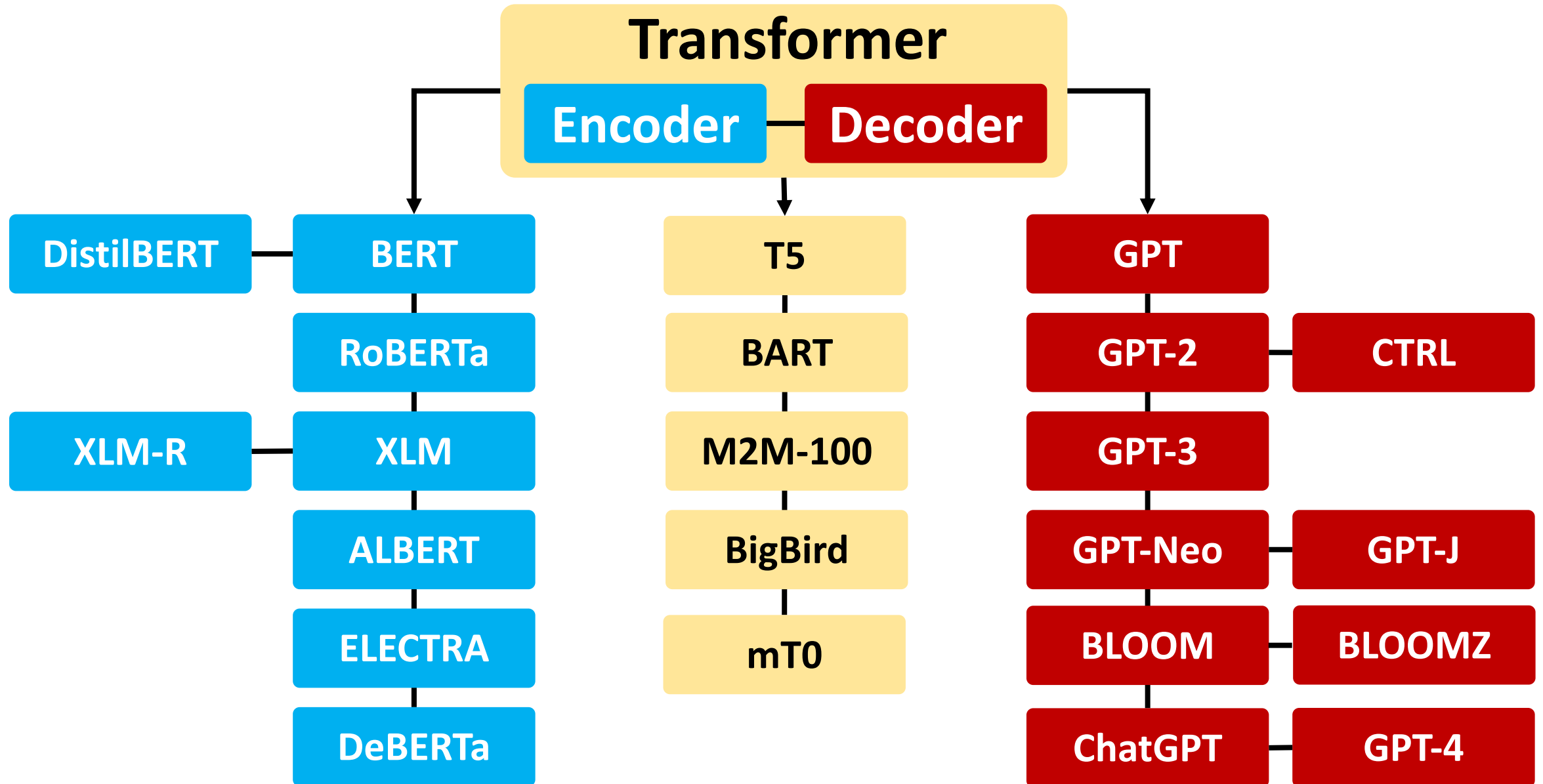
(GPT-3, ChatGPT, PaLM, BLOOM, OPT-175B, LLaMA)



The Transformers Timeline



Transformer Models



OpenAI ChatGPT

[API](#)[RESEARCH](#)[BLOG](#)[ABOUT](#)

ChatGPT: Optimizing Language Models for Dialogue

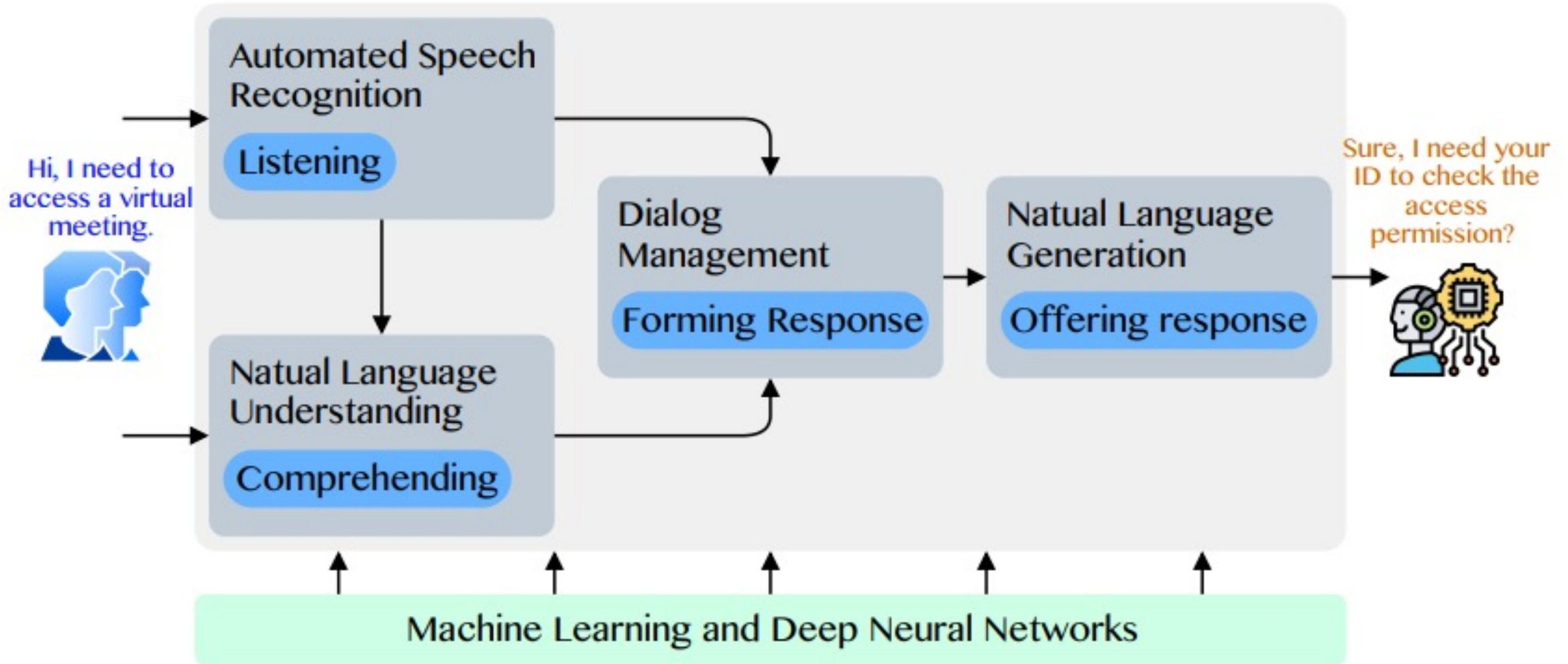
We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.



Source: <https://openai.com/blog/chatgpt/>

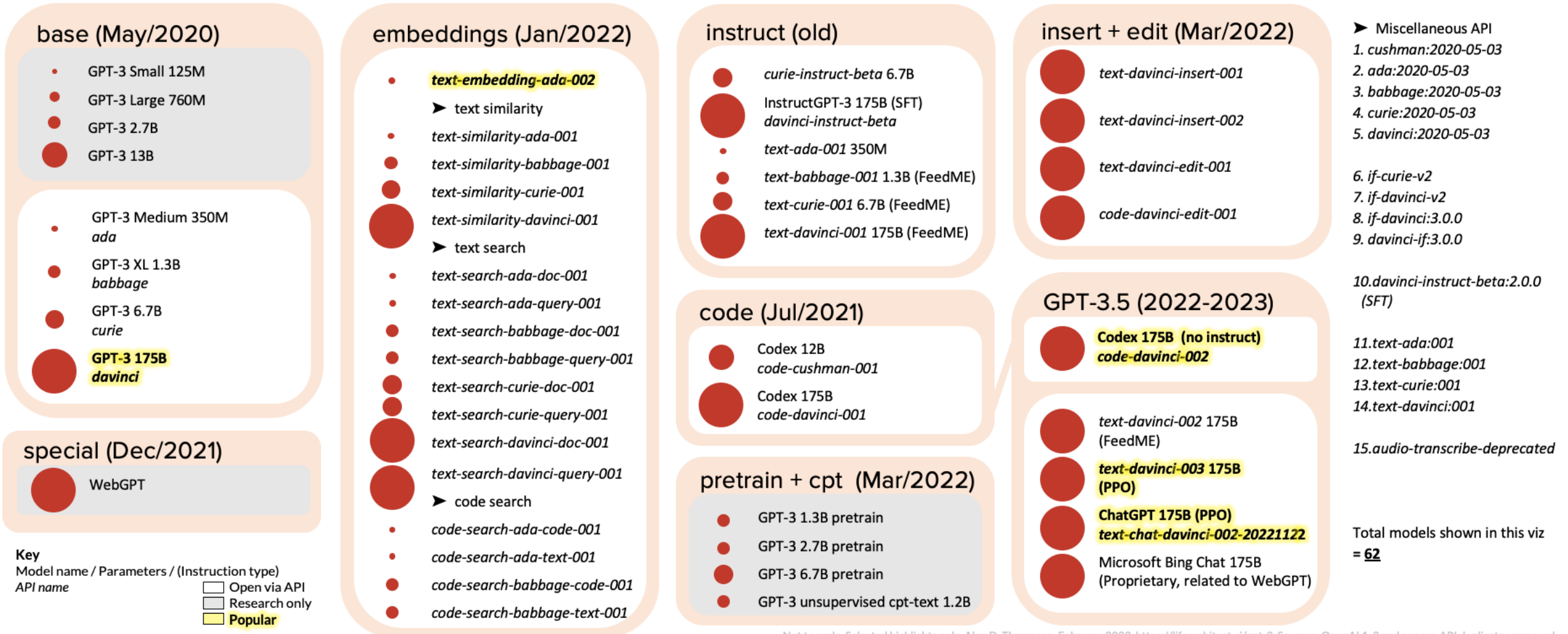
Conversational AI

to deliver contextual and personal experience to users



ChatGPT and GPT-3 Family

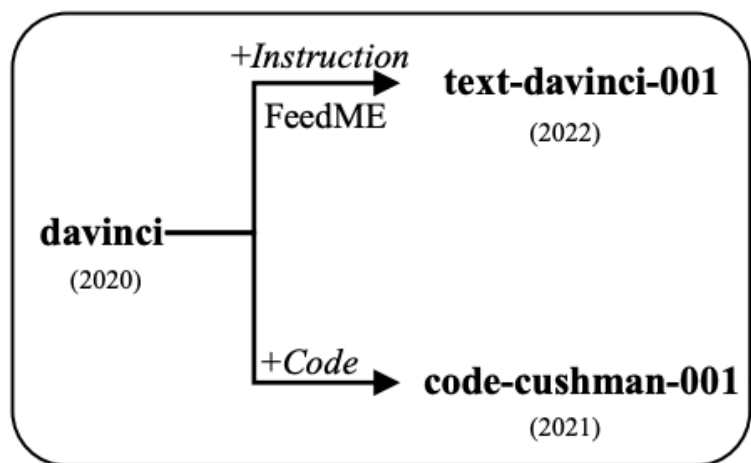
(GPT-3, InstructGPT, GPT-3.5, ChatGPT)



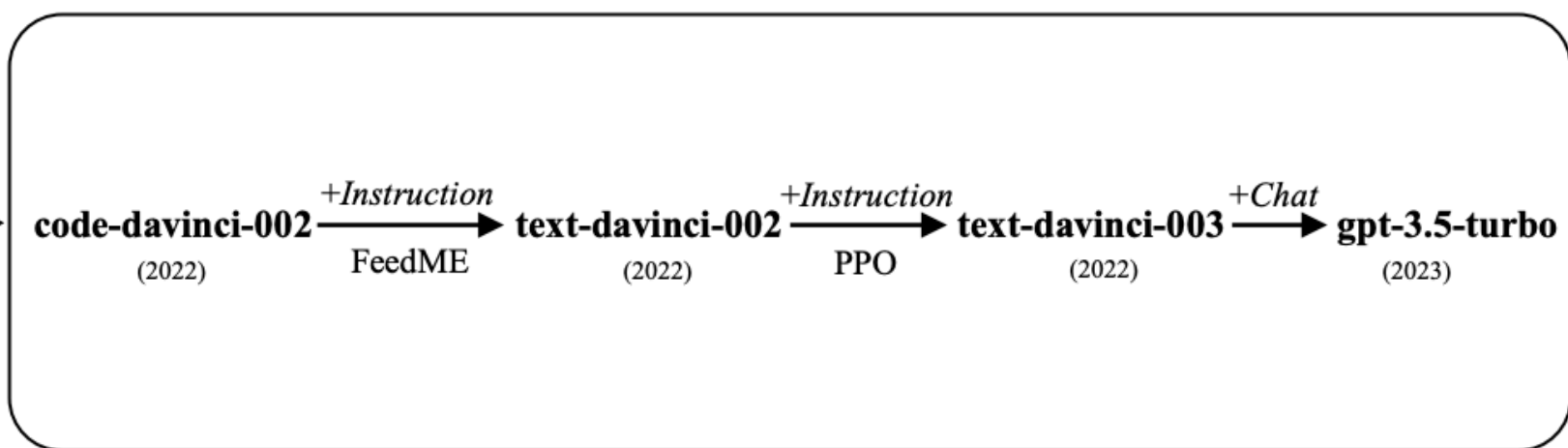
Not to scale. Selected highlights only. Alan D. Thompson. February 2023. <https://lifearchitected.ai/gpt-3/> Sources: OpenAI [1](#), [2](#) and papers, API [duplicates](#) removed.

Evolutionary of ChatGPT Models

GPT-3



GPT-3.5



OpenAI ChatGPT and Open LLM

GPT-4, LLaMA, Alpaca, Dolly, Cerebras-GPT, GPT4All, Vicuna, ColossalChat, Koala, Phoenix

- OpenAI GPT-4
- Deepmind Chinchilla
- Meta OPT (LLaMA)
- Pythia
- **Stanford Alpaca**
- **Databricks Dolly**
- **Cerebras-GPT**
- **GPT4All**
- **Vicuna**
- **ColossalChat**
- **BAIR Koala**

Large Language Models (LLM)

Openness and Training Philosophy

Model	Model architecture	Training data	Model weights	Checkpoints	Compute-optimal training	License
OpenAI GPT-4	Closed	Closed	No	No	Unknown	Not available
Deepmind Chinchilla	Open	Closed	No	No	Yes	Not available
Meta OPT	Open	Open	Researchers Only	Yes	No	Non-commercial
Pythia	Open	Open	Open	Yes	No	Apache 2.0
Cerebras-GPT	Open	Open	Open	Yes	Yes	Apache 2.0

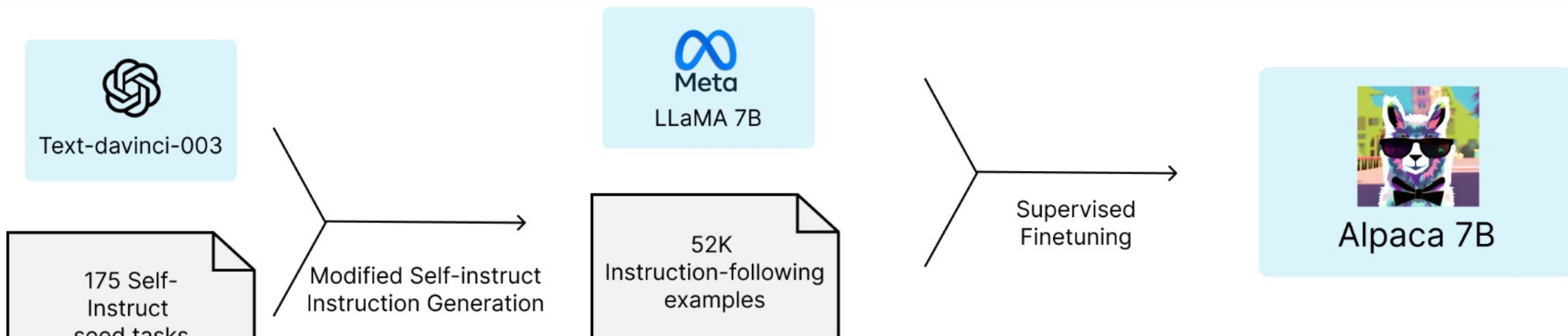
Phoenix: Democratizing ChatGPT across Languages



Model	Backbone	#paras	Open-source		Claimed language	Post-training				Release date
			model	data		instruction data	conversation lang	instruction data	conversation lang	
ChatGPT	unknown	unknown	✗	✗	multi					11/30/22
Wenxin ⁷	unknown	unknown	✗	✗	zh					03/16/23
ChatGLM ⁸	GLM	6B	✓ ¹	✗	en/zh					03/16/23
Tongyi ⁹	unknown	unknown	✗	✗	zh					04/07/23
Shangliang ¹⁰	unknown	unknown	✗	✗	zh					04/10/23
Alpaca [12]	LLaMA	7B	✗	✓	en	52K	en	✗	✗	03/13/23
Dolly ¹¹ ²	GPT-J	6B	✓	✓	en	52k	en	✗	✗	03/24/23
BELLE [6]	BLOOMZ	7B	✓	✓	zh	1.5M	ch	✗	✗	03/26/23
Guanaco ¹²	LLaMA	7B	✓	✓	en/zh/ja/de	534K ³	4 ⁴	✗	✗	03/26/23
Chinese-alpaca [3]	LLaMA	7/13B	✓	✓	en/zh	2M/3M	en/zh	✗	✗	03/28/23
LuoTuo [7]	LLaMA	7B	✓	✓	zh	52k	cn	✗	✗	03/31/23
Vicuna [2]	LLaMA	7/13B	✓	✓ ⁵	en	✗	✗	70K	multi ⁶	03/13/23
Koala ¹³	LLaMA	13B	✓	✓	en	355K	en	117K	en	04/03/23
BAIZE [17]	LLaMA	7/13/30B	✓	✓	en	✗	✗	111.5K	en	04/04/23
Phoenix	BLOOMZ	7B	✓	✓	multi	267K	40+	189K	40+	04/08/23
Latin Phoenix (Chimera)	LLaMA	7B/13B	✓	✓	Latin	267K	40+	189K	40+	04/08/23

Stanford Alpaca:

A Strong, Replicable Instruction-Following Model



Example seed task

Instruction: Brainstorm a list of possible New Year's resolutions.

Output:

- Lose weight
- Exercise more
- Eat healthier

Example Generated task

Instruction: Brainstorm creative ideas for designing a conference room.

Output:

... incorporating flexible components, such as moveable walls and furniture ...

Stanford Alpaca



GPT4All:

Training an Assistant-style Chatbot with Large Scale Data Distillation from GPT-3.5-Turbo

- Demo, data and code to train an assistant-style large language model with ~800k GPT-3.5-Turbo Generations based on LLaMa
- Reproducibility
 - Trained LoRa Weights:
 - gpt4all-lora (four full epochs of training):
 - <https://huggingface.co/nomic-ai/gpt4all-lora>

GPT4All-J

An Apache-2 Licensed Assistant-Style Chatbot

GPT4All-J (GPT4All v2) based on Open Source **GPT-J** model

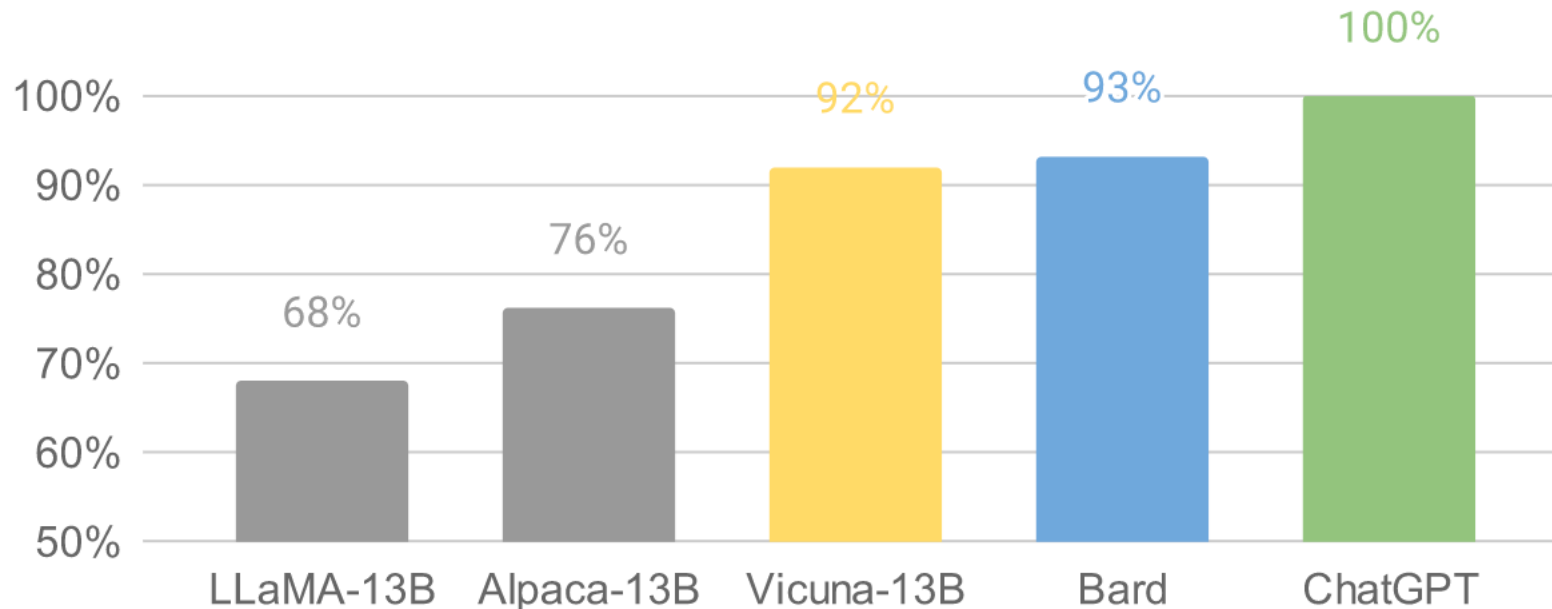
Model	BoolQ	PIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA
GPT4All-J 6.7B	73.4	74.8	63.4	64.7	54.9	36.0	40.2
GPT4All-J Lora 6.7B	68.6	75.8	66.2	63.5	56.4	35.7	40.2
GPT4All LLaMa Lora 7B	73.1	77.6	72.1	67.8	51.1	40.4	40.2
Dolly 6B	68.8	77.3	67.6	63.9	62.9	38.7	41.2
Dolly 12B	56.7	75.4	71.0	62.2	64.6	38.5	40.4
Alpaca 7B	73.9	77.2	73.9	66.1	59.8	43.3	43.4
Alpaca Lora 7B	74.3	79.3	74.0	68.8	56.6	43.9	42.6
GPT-J 6.7B	65.4	76.2	66.2	64.1	62.2	36.6	38.2
LLaMa 7B	73.1	77.4	73.0	66.9	52.5	41.4	42.4
Pythia 6.7B	63.5	76.3	64.0	61.1	61.3	35.2	37.2
Pythia 12B	67.7	76.6	67.3	63.8	63.9	34.8	38

Vicuna: An Open-Source Chatbot

Impressing GPT-4 with 90%* ChatGPT Quality

by the Team with members from UC Berkeley, CMU, Stanford, and UC San Diego

- **Vicuna-13B: an open-source chatbot trained by fine-tuning LLaMA on user-shared conversations collected from ShareGPT.**
- **The cost of training Vicuna-13B is around \$300.**



Source: <https://vicuna.lmsys.org/>

Chinese-Vicuna:

A Chinese Instruction-following LLaMA-based Model

一個中文低資源的 llama+lora 方案

Dataset Preview

Size: 409 MB

</> API

Go to dataset viewer

instruction (string)	input (string)	output (string)
"用一句话描述地球为什么是独一无二的。 \n "	" "	"地球上有适宜生命存在的条件和多样化的生命形式。"
"给出一段对话，要求GPT模型使用合适的语气和回答方式继续对话。 对话： A：你今...	" "	"A： 这太好了！你的新工作听起来很令人兴奋。你对接接下来的日子有什么期望吗？ B： 是啊，我非...
"基于以下提示填写以下句子的空格。 提示： - 提供多种现实世界的场景 - 空格...	" "	"多种形容词可填，以下是其中一些例子： - 愉快的 - 惬意的 - 轻松的 - 安静的 - 美妙的"
"请生成一个新闻标题，描述一场正在发生的大型自然灾害。 \n "	" "	" "强烈飓风肆虐，数百万人疏散！ " "
"为指定的词汇创建一个关于该词汇的简短解释。 "人工智能"	" "	"人工智能是一种科技，通过使用计算机程序和算法实现类似于人类智能的能力，例如学习、推理和自...
"在给定的文本中查找并替换特定的单词。 原始文本： "这个句子中有一个错误的单词..."	" "	"替换后的文本： "这个句子中有一个修改的单词"
"根据给定的年份，计算该年是否为闰年。 \n \n1996\n"	" "	"是闰年。"



Bactrian camel



Wild Bactrian camel



Dromedary



Llama



Alpaca



Guanaco

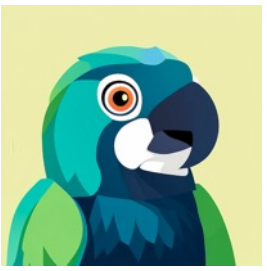


Vicuña

Chinese-Vicuna based on Guanaco Dataset and Belle Dataset

Source: https://huggingface.co/datasets/Chinese-Vicuna/guanaco_belle_merge_v1.0

Source: <https://github.com/Facico/Chinese-Vicuna>



StableLM

Stability AI Language Models

- **StableLM-Alpha models are trained on the new dataset that build on The Pile, which contains 1.5 trillion tokens, roughly 3x the size of The Pile.**
 - **These models will be trained on up to 1.5 trillion tokens.**
 - **The context length for these models is 4096 tokens.**
- **Fine-tuned the model with Stanford Alpaca's procedure using a combination of five recent datasets for conversational agents: Stanford's Alpaca, Nomic-AI's gpt4all, RyokoAI's ShareGPT52K datasets, Databricks labs' Dolly, and Anthropic's HH.**

RedPajama

a project to create leading open-source models,
starts by reproducing LLaMA training dataset of over 1.2 trillion tokens



Dataset	RedPajama	LLaMA*
CommonCrawl	878 billion	852 billion
C4	175 billion	190 billion
Github	59 billion	100 billion
Books	26 billion	25 billion
ArXiv	28 billion	33 billion
Wikipedia	24 billion	25 billion
StackExchange	20 billion	27 billion
Total Tokens	1.2 trillion	1.25 trillion

Chat with Open Large Language Models

chat.lmsys.org

Single Model

Chatbot Arena (battle)

Chatbot Arena (side-by-side)

Leaderboard



Chat with Open Large Language Models

- Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90% ChatGPT Quality. [\[Blog.post\]](#)
- [\[GitHub\]](#) [\[Twitter\]](#) [\[Discord\]](#)

Terms of use

By using this service, users are required to agree to the following terms: The service is a research preview intended for non-commercial use only. It only provides limited safety measures and may generate offensive content. It must not be used for any illegal, harmful, violent, racist, or sexual purposes. **The service collects user dialogue data and reserves the right to distribute it under a Creative Commons Attribution (CC-BY) license.**

Choose a model to chat with

[Vicuna](#): a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS

[WizardLM](#): an instruction-following LLM using evol-instruct by Microsoft

[GPT4All-Snoozy](#): A finetuned LLaMA model on assistant style data by Nomic AI

[Guanaco](#): a model fine-tuned with QLoRA by UW

[Koala](#): a dialogue model for academic research by BAIR

[RWKV-4-Raven](#): an RNN with transformer-level LLM performance

[Alpaca](#): a model fine-tuned from LLaMA on instruction-following demonstrations by Stanford

[ChatGLM](#): an open bilingual dialogue language model by Tsinghua University

[OpenAssistant \(oasst\)](#): an Open Assistant for everyone by LAION

[LLaMA](#): open and efficient foundation language models by Meta

[Dolly](#): an instruction-tuned open large language model by Databricks

[StableLM](#): Stability AI language models

[MPT-Chat](#): a chatbot fine-tuned from MPT-7B by MosaicML

[FastChat-T5](#): a chat assistant fine-tuned from FLAN-T5 by LMSYS

vicuna-13b



<https://chat.lmsys.org/>

Chat with Open Large Language Models: Chatbot Arena

vicuna-13b

Model A

Large Language Models for Data Science

Large Language Models (LLMs) are a type of artificial neural network that have been trained on a massive amount of text data to generate human-like language. These models have achieved state-of-the-art performance on a variety of natural language processing tasks, such as language translation, text generation, and question answering. In the field of data science, LLMs have been used to process and analyze unstructured text data, such as customer reviews, social media posts, and news articles, to extract insights and make predictions.

👉 A is better

👉 B is better

👉 Tie

👉 Both are bad

mpt-7b-chat

Model B

Large Language Models for Data Science

Large Language Models (LLMs) are a type of artificial intelligence (AI) model that are capable of processing and generating human-like natural language. They are based on deep neural networks that are trained on large amounts of text data, such as books, news articles, and social media posts. LLMs have shown promising results in a variety of data science applications, including language translation, text summarization, and sentiment analysis. One of the most popular LLMs is GPT-3 (Generative Pre-trained Transformer 3), developed by OpenAI. GPT-3 has 175 billion parameters, making it one of the largest language models ever created. It can perform a wide range of language tasks, such as text completion, question answering, and text generation. GPT-3 has been used in various industries, including healthcare, finance, and journalism, to automate

<https://chat.lmsys.org/>

Chat with Open Large Language Models: Leaderboard




[Single Model](#)[Chatbot Arena \(battle\)](#)[Chatbot Arena \(side-by-side\)](#)[Leaderboard](#)

Leaderboard

[\[Blog\]](#) [\[GitHub\]](#) [\[Twitter\]](#) [\[Discord\]](#)

We use the Elo rating system to calculate the relative performance of the models. You can view the voting data, basic analyses, and calculation procedure in this [notebook](#). We will periodically release new leaderboards. If you want to see more models, please help us [add them](#).

Last updated: 2023-05-22 09:35:17 PDT

Rank	Model	Elo Rating	Description
1	 gpt-4	1225	ChatGPT-4 by OpenAI
2	 claude-v1	1195	Claude by Anthropic
3	 claude-instant-v1	1153	Claude Instant by Anthropic
4	gpt-3.5-turbo	1143	ChatGPT-3.5 by OpenAI
5	vicuna-13b	1054	a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS
6	palm-2	1042	PaLM 2 for Chat (chat-bison@001) by Google
7	vicuna-7b	1007	a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS
8	koala-13b	980	a dialogue model for academic research by BAIR
9	mpt-7b-chat	952	a chatbot fine-tuned from MPT-7B by MosaicML
10	fastchat-t5-3b	941	a chat assistant fine-tuned from FLAN-T5 by LMSYS
11	alpaca-13b	937	a model fine-tuned from LLaMA on instruction-following demonstrations by Stanford
12	RWKV-4-Raven-14B	928	an RNN with transformer-level LLM performance

<https://chat.lmsys.org/>

ChatPDF

www.chatpdf.com

Chat with any PDF



Join Discord



Post to Twitter



Share on Facebook



Drop PDF here


[Browse my Computer](#)



[From URL](#) [Find a PDF](#)


<https://www.chatpdf.com/>


Perplexity.ai







New Thread  



 Popular

 Your Threads

 Login

Sign Up

 Download

 Follow  Join

About Blog Privacy Policy

Ask anything...

Quick



Popular Now



Turkey election results see Erdogan forced in...

The 2023 Turkish presidential election and parliamentary elections were held on May 15,...

how to make a successful career pivot

Making a successful career pivot requires careful planning and execution. Here are...

did mammoths evolve from dinosaurs

No, mammoths did not evolve from dinosaurs. Mammoths are mammals that evolved...

Jayson Tatum's 51-point game sends Celtic...

Jayson Tatum scored 51 points in the Boston Celtics' 112-88 victory over the Philadelphia...

WhatsApp introduces chat lock feature

WhatsApp has introduced a new feature called "Chat Lock" that adds an extra layer of...

number of calories in a potato

The number of calories in a potato varies depending on the size and type of potato...



Perplexity for iPhone
Available Now



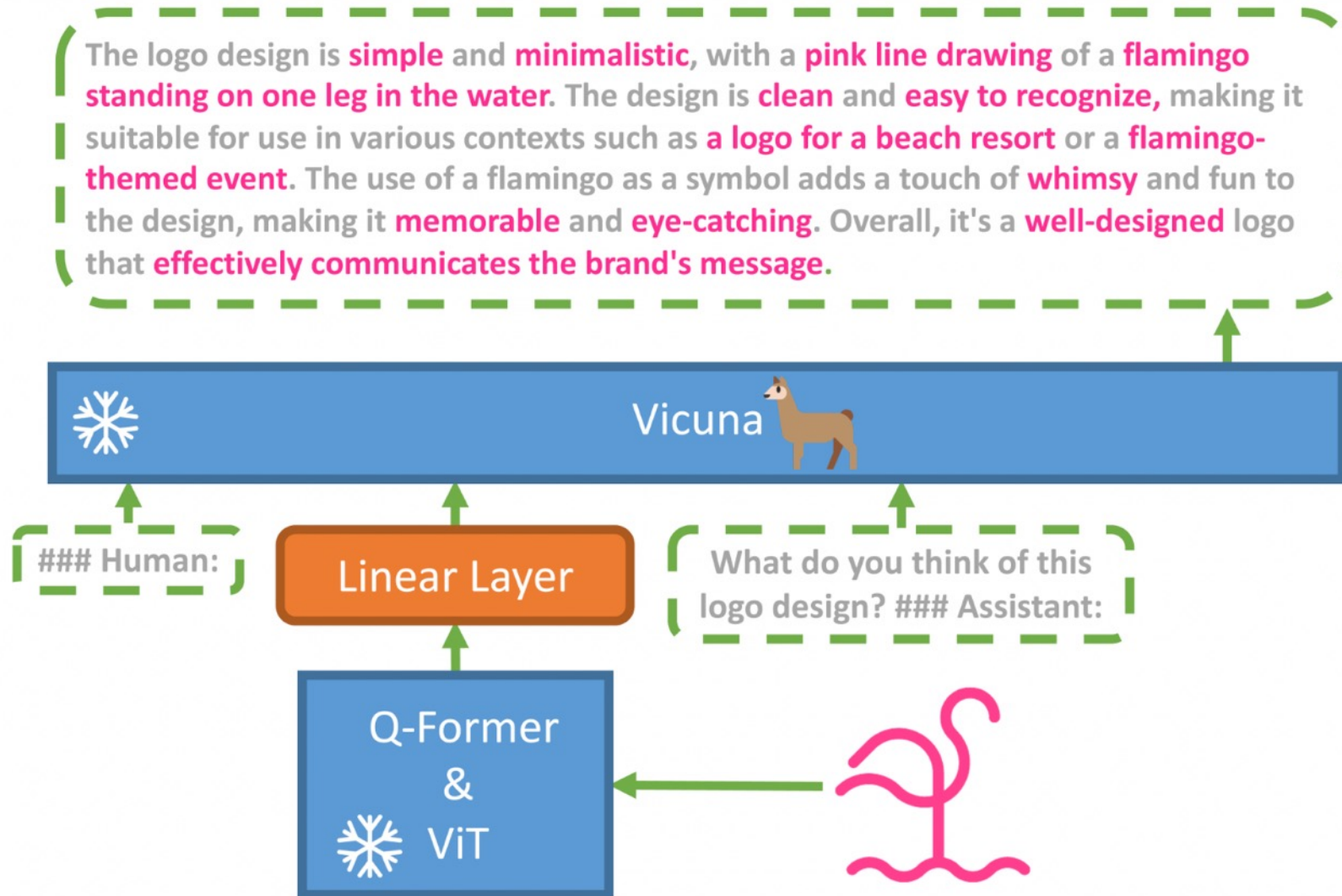
Perplexity for Android
Join the waitlist



<https://www.perplexity.ai/>

MiniGPT-4:

Enhancing Vision-language Understanding with Advanced Large Language Models



LLaVA: Large Language and Vision Assistant

[llava-vl.github.io](https://l1lava-vl.github.io)

LLaVA: Large Language and Vision Assistant

[\[Project Page\]](#) [\[Paper\]](#) [\[Code\]](#) [\[Model\]](#)

LLaVA-13B-v0

Image

Drop Image Here
- or -
Click to Upload

Preprocess

When the image is not square, you may try adjusting this .

☒ Crop

☐ Resize

☐ Pad

Parameters


Terms of use

By using this service, users are required to agree to the following terms:

The service is a research preview intended for non-commercial use only. It only provides limited safety measures and may generate offensive content. It must not be used for any illegal, harmful, violent, racist, or sexual purposes. The service may collect user dialogue data for future research.

LLaVA Chatbot

What are the things I should be cautious about when I visit here?



When visiting this location, one should be cautious about the weather conditions and the state of the wooden pier. In the image, the wooden pier is extended over a large, still lake under cloudy skies, which could suggest the possibility of sudden changes in weather, such as wind, rain, or even high waves. Visitors should pay attention to weather forecasts and remain vigilant while on the pier to ensure their safety. Additionally, they should be mindful of the pier's condition and the surrounding water, as wood can be slippery, especially when wet or in windy conditions. It is crucial to maintain proper footing and follow any posted safety guidelines, as the quiet waters of the lake can be deceptively

Enter text and press ENTER

Submit

Upvote

Downvote

Flag

Regenerate

Visual Instruction Tuning

LLaVA: Large Language and Vision Assistant

University of Wisconsin-Madison, Microsoft Research, Columbia University

Science QA:
New SoTA with
the synergy of
LLaVA with
GPT-4



LLaVA represents a novel end-to-end trained large multimodal model that combines a vision encoder and Vicuna for general-purpose visual and language understanding, achieving impressive chat capabilities mimicking spirits of the multimodal GPT-4 and setting a new state-of-the-art accuracy on Science QA.

Source: <https://llava-vl.github.io/>

MPT-7B:


A New Standard for Open-Source, Commercially Usable LLMs



Source: <https://www.mosaicml.com/blog/mpt-7b>

MPT-7B (MosaicML Pretrained Transformer)

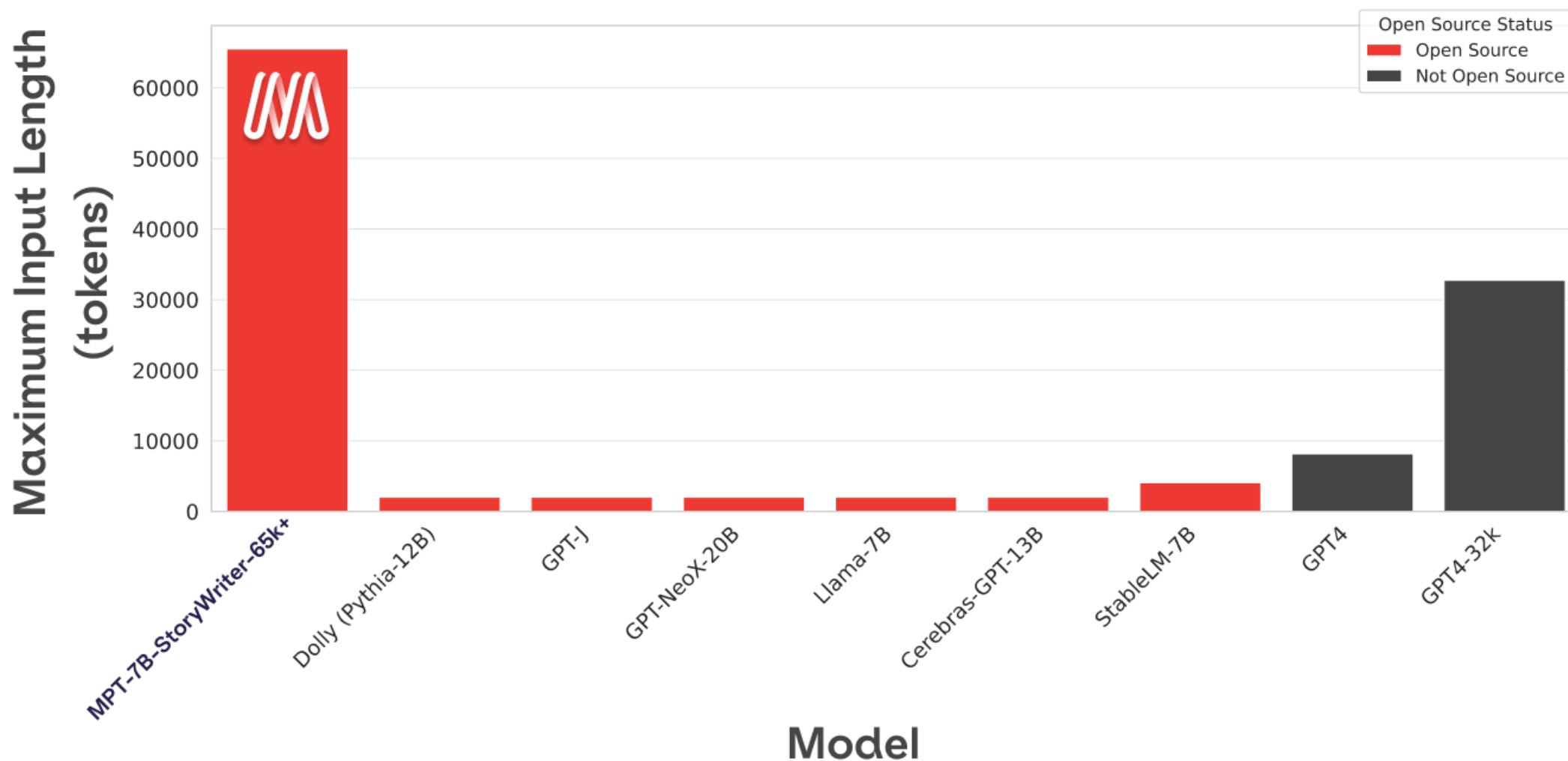
vs. open source models on academic tasks

Model	LAMBADA (OpenAI)	HellaSwag	PIQA	ARC-Easy	ARC- Challenge	BoolQ	COPA	Winograd	Winogrande	TriviaQA	Jeopardy	MMLU
 MPT-7B	0.703	0.761	0.799	0.673	0.394	0.750	0.813	0.878	0.683	0.343	0.308	0.296
LLaMA-7B	0.738	0.751	0.792	0.652	0.411	0.767	0.779	0.807	0.675	0.443	0.334	0.302
StableLM-7B (alpha)	0.533	0.411	0.666	0.435	0.259	0.606	0.672	0.646	0.513	0.049	0.000	0.251
Pythia-7B	0.667	0.636	0.761	0.581	0.325	0.634	0.769	0.786	0.607	0.198	0.022	0.265
Pythia-12B	0.704	0.672	0.768	0.605	0.351	0.675	0.781	0.847	0.627	0.233	0.026	0.253
GPTJ-6B	0.683	0.665	0.762	0.583	0.355	0.648	0.789	0.833	0.641	0.234	0.026	0.261
GPT-NeoX-20B	0.719	0.712	0.780	0.644	0.392	0.691	0.781	0.861	0.665	0.347	0.146	0.269
Cerebras-7B	0.636	0.582	0.744	0.564	0.311	0.625	0.734	0.779	0.603	0.141	0.012	0.259
Cerebras-13B	0.635	0.588	0.740	0.571	0.321	0.611	0.719	0.760	0.602	0.146	0.013	0.258
OPT-7B	0.677	0.676	0.773	0.579	0.329	0.665	0.719	0.840	0.656	0.227	0.020	0.251
OPT-13B	0.692	0.701	0.774	0.586	0.345	0.657	0.805	0.851	0.670	0.282	0.126	0.257

Source: <https://www.mosaicml.com/blog/mpt-7b>

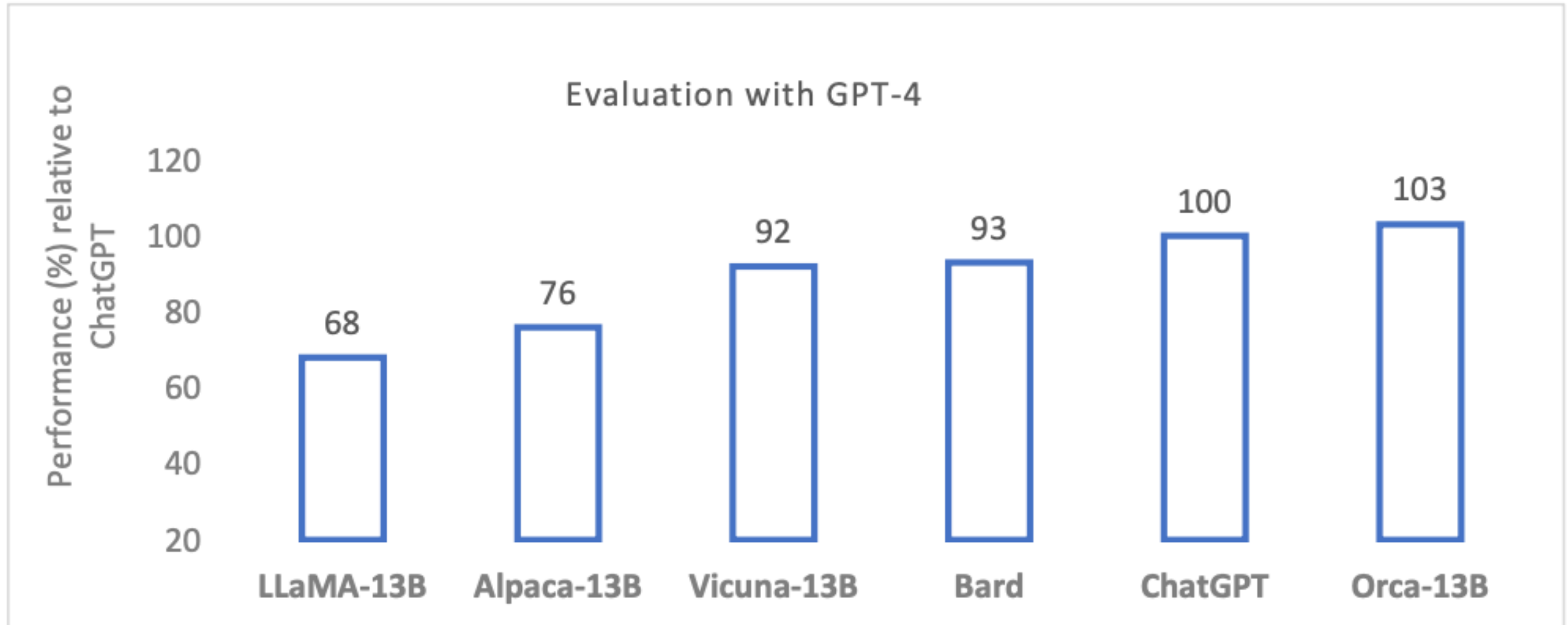
MPT-7B-StoryWriter-65k+

Maximum Input Lengths of Different LLMs

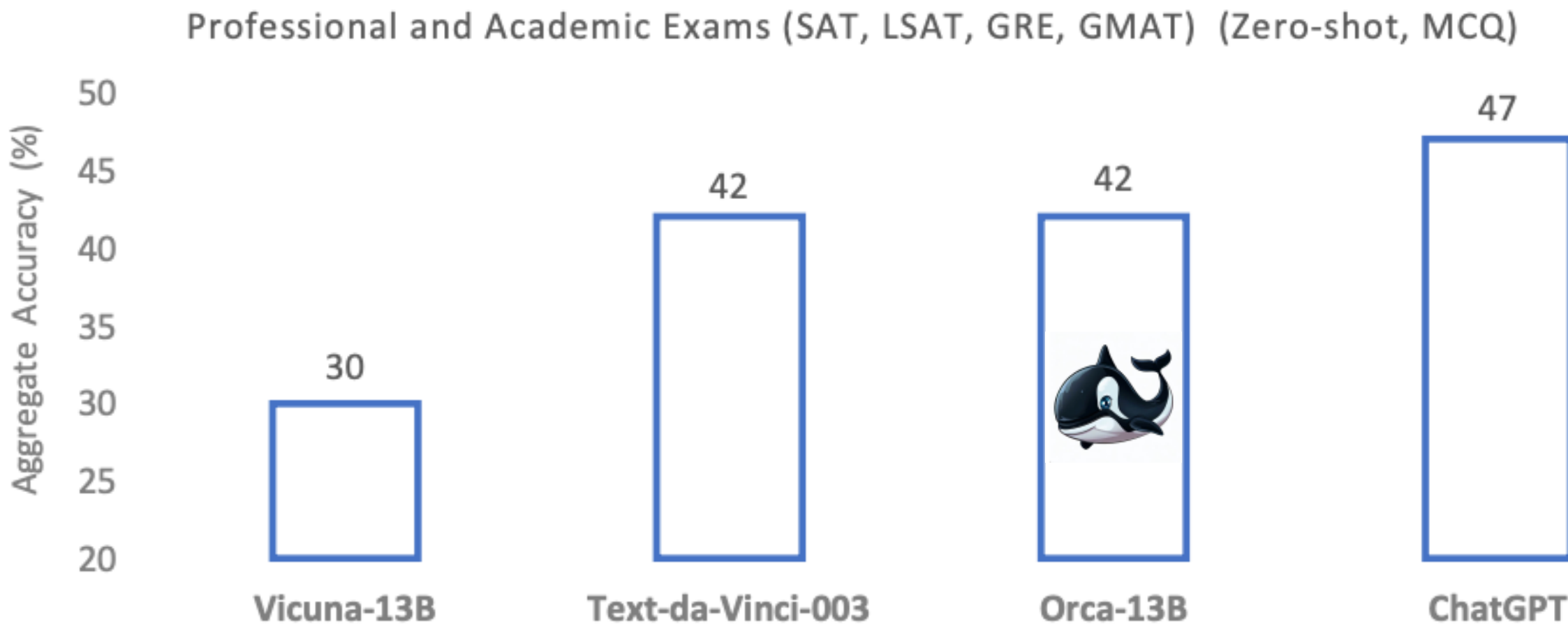




Orca: Progressive Learning from Complex Explanation Traces of GPT-4



Orca-13B and ChatGPT for GRE and GMAT



Meta Llama-2 70B: Best Open Source and Commercial LLM (Llama-2, Falcon, MPT)



Introducing Llama 2

The next generation of our
open source large language model

Llama 2 is available for free for research and commercial use.

[Download the Model](#)

Meta Llama-2 70B: Best Open Source and Commercial LLM (Llama-2, Falcon, MPT)

MODEL SIZE (PARAMETERS)	PRETRAINED	FINE-TUNED FOR CHAT USE CASES
7B	Model architecture:	Data collection for helpfulness and safety:
13B	Pretraining Tokens: 2 Trillion	Supervised fine-tuning: Over 100,000
70B	Context Length: 4096	Human Preferences: Over 1,000,000

Llama 2 pretrained models are trained on 2 trillion tokens, and have double the context length than Llama 1. Its fine-tuned models have been trained on over 1 million human annotations.

**Meta
Llama-2 70B:
Best
Open Source
and
Commercial
LLM
(Llama-2,
Falcon, MPT)**

Benchmark (Higher is better)	MPT (7B)	Falcon (7B)	Llama-2 (7B)	Llama-2 (13B)	MPT (30B)	Falcon (40B)	Llama-1 (65B)	Llama-2 (70B)
MMLU	26.8	26.2	45.3	54.8	46.9	55.4	63.4	68.9
TriviaQA	59.6	56.8	68.9	77.2	71.3	78.6	84.5	85.0
Natural Questions	17.8	18.1	22.7	28.0	23.0	29.5	31.0	33.0
GSM8K	6.8	6.8	14.6	28.7	15.2	19.6	50.9	56.8
HumanEval	18.3	N/A	12.8	18.3	25.0	N/A	23.7	29.9
AGIEval (English tasks only)	23.5	21.2	29.3	39.1	33.8	37.0	47.6	54.2
BoolQ	75.0	67.5	77.4	81.7	79.0	83.1	85.3	85.0

Llama 2 outperforms other open source language models on many external benchmarks, including reasoning, coding, proficiency, and knowledge tests.

Source: <https://ai.meta.com/llama/>

Llama-2: Comparison to closed-source models (GPT-3.5, GPT-4, PaLM) on academic benchmarks

Benchmark (shots)	GPT-3.5	GPT-4	PaLM	PaLM-2-L	LLAMA 2
MMLU (5-shot)	70.0	86.4	69.3	78.3	68.9
TriviaQA (1-shot)	–	–	81.4	86.1	85.0
Natural Questions (1-shot)	–	–	29.3	37.5	33.0
GSM8K (8-shot)	57.1	92.0	56.5	80.7	56.8
HumanEval (0-shot)	48.1	67.0	26.2	–	29.9
BIG-Bench Hard (3-shot)	–	–	52.3	65.7	51.2

Results for GPT-3.5 and GPT-4 are from OpenAI (2023).

Results for the PaLM model are from Chowdhery et al. (2022).

Results for the PaLM-2-L are from Anil et al. (2023).

Llama 2: Open Foundation and Fine-Tuned Chat Models

:2307.09288v2 [cs.CL] 19 Jul 2023

LLAMA 2: Open Foundation and Fine-Tuned Chat Models

Hugo Touvron* Louis Martin† Kevin Stone†

Peter Albert Amjad Almahairi Yasmine Babaei Nikolay Bashlykov Soumya Batra
Prajwal Bhargava Shruti Bhosale Dan Bikel Lukas Blecher Cristian Canton Ferrer Moya Chen
Guillem Cucurull David Esiobu Jude Fernandes Jeremy Fu Wenyin Fu Brian Fuller
Cynthia Gao Vedanuj Goswami Naman Goyal Anthony Hartshorn Saghar Hosseini Rui Hou
Hakan Inan Marcin Kardas Viktor Kerkez Madian Khabsa Isabel Kloumann Artem Korenev
Punit Singh Koura Marie-Anne Lachaux Thibaut Lavril Jenya Lee Diana Liskovich
Yinghai Lu Yuning Mao Xavier Martinet Todor Mihaylov Pushkar Mishra
Igor Molybog Yixin Nie Andrew Poulton Jeremy Reizenstein Rashi Rungta Kalyan Saladi
Alan Schelten Ruan Silva Eric Michael Smith Ranjan Subramanian Xiaoqing Ellen Tan Binh Tang
Ross Taylor Adina Williams Jian Xiang Kuan Puxin Xu Zheng Yan Iliyan Zarov Yuchen Zhang
Angela Fan Melanie Kambadur Sharan Narang Aurelien Rodriguez Robert Stojnic
Sergey Edunov Thomas Scialom*

GenAI, Meta

Abstract

In this work, we develop and release Llama 2, a collection of pretrained and fine-tuned large language models (LLMs) ranging in scale from 7 billion to 70 billion parameters. Our fine-tuned LLMs, called LLAMA 2-CHAT, are optimized for dialogue use cases. Our models outperform open-source chat models on most benchmarks we tested, and based on our human evaluations for helpfulness and safety, may be a suitable substitute for closed-source models. We provide a detailed description of our approach to fine-tuning and safety improvements of LLAMA 2-CHAT in order to enable the community to build on our work and contribute to the responsible development of LLMs.

InstructBLIP

Vision-Language Models with Instruction Tuning

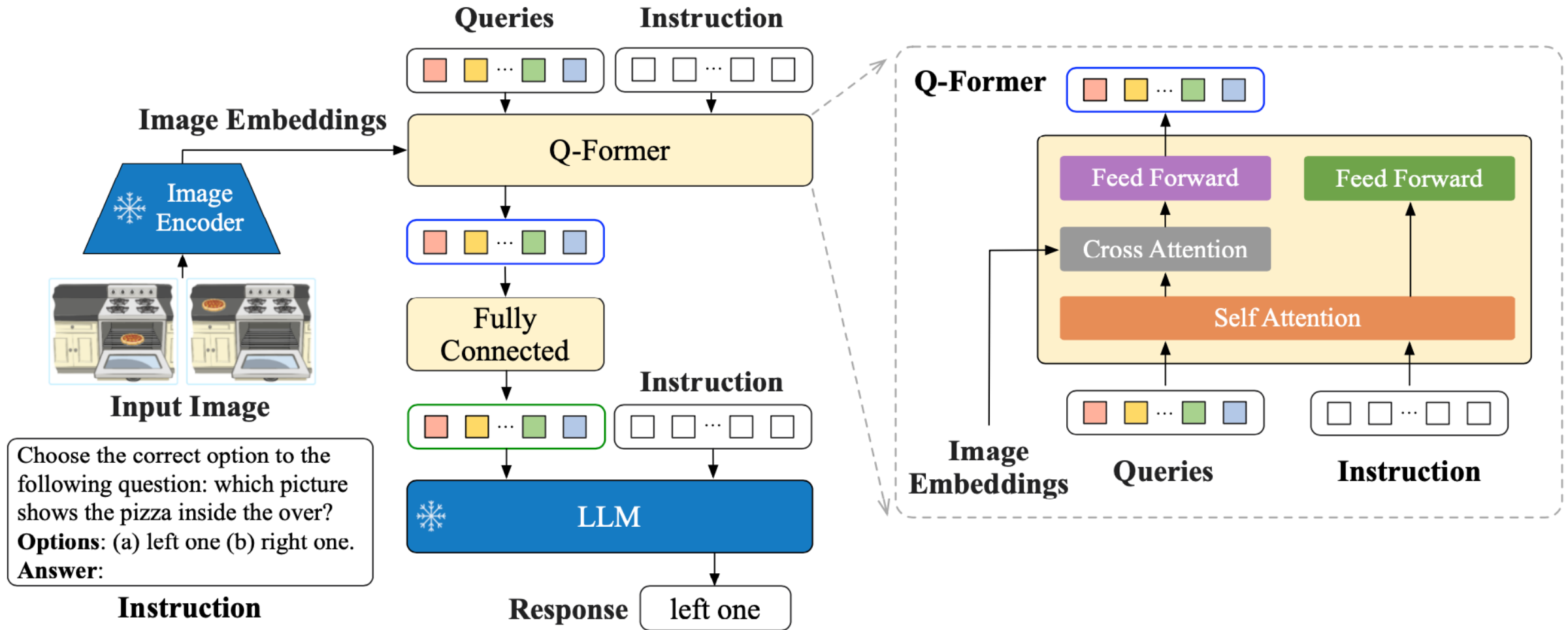
Introduce me this painting
in detail.



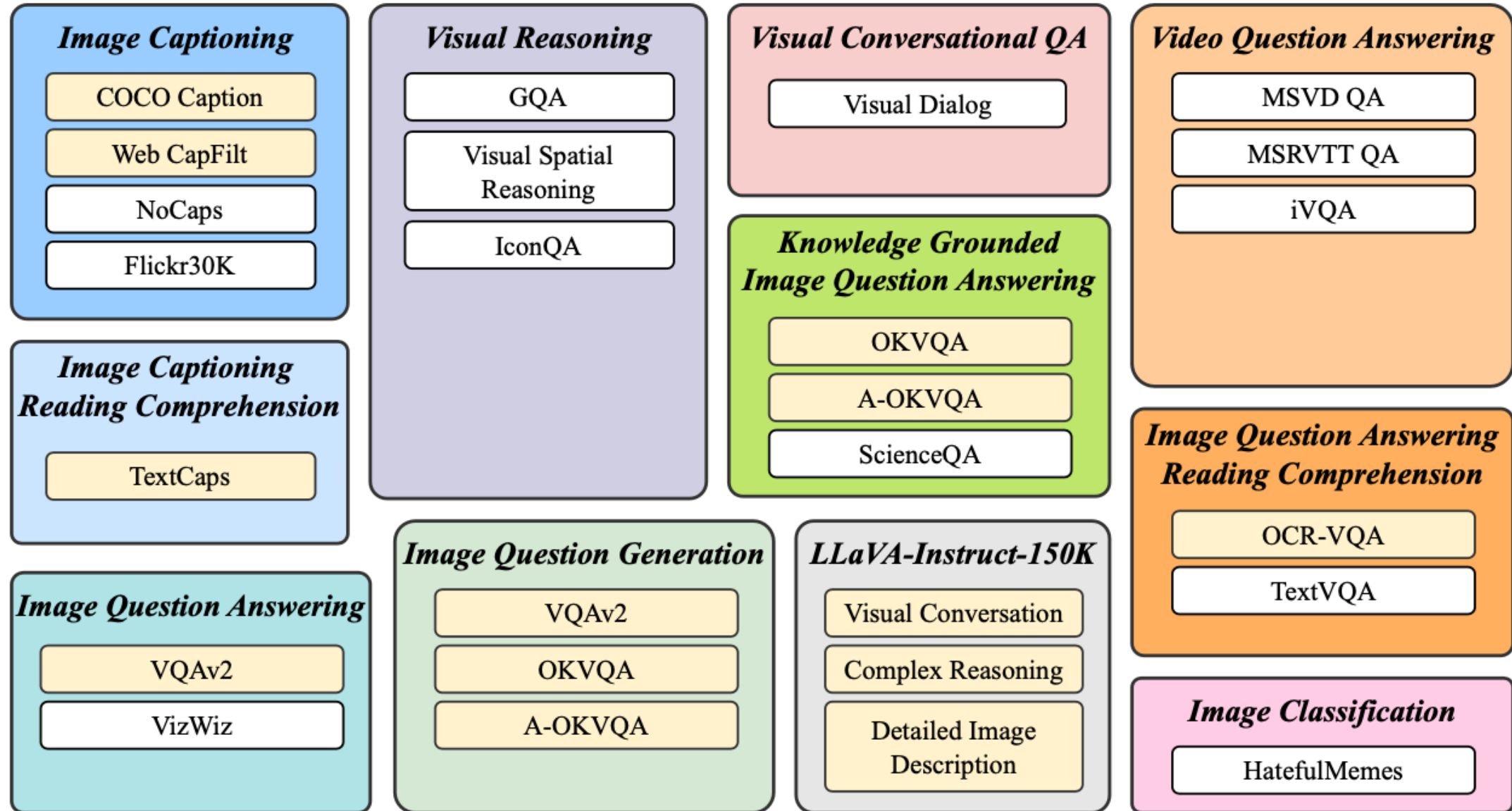
The painting depicts a young girl with a pearl earring, also known as "The Girl with the Pearl Earring" or "Girl with a Pearl Earring". She is wearing a blue dress and has a pearl earring on her left ear. Her hair is pulled back into a bun, and she is looking directly at the viewer. This painting is a famous work of art by Dutch painter Johannes Vermeer, created in 1665. It is considered one of his masterpieces and is widely recognized for its intricate details and realistic portrayal of the young girl's facial expression.

Model Architecture of InstructBLIP

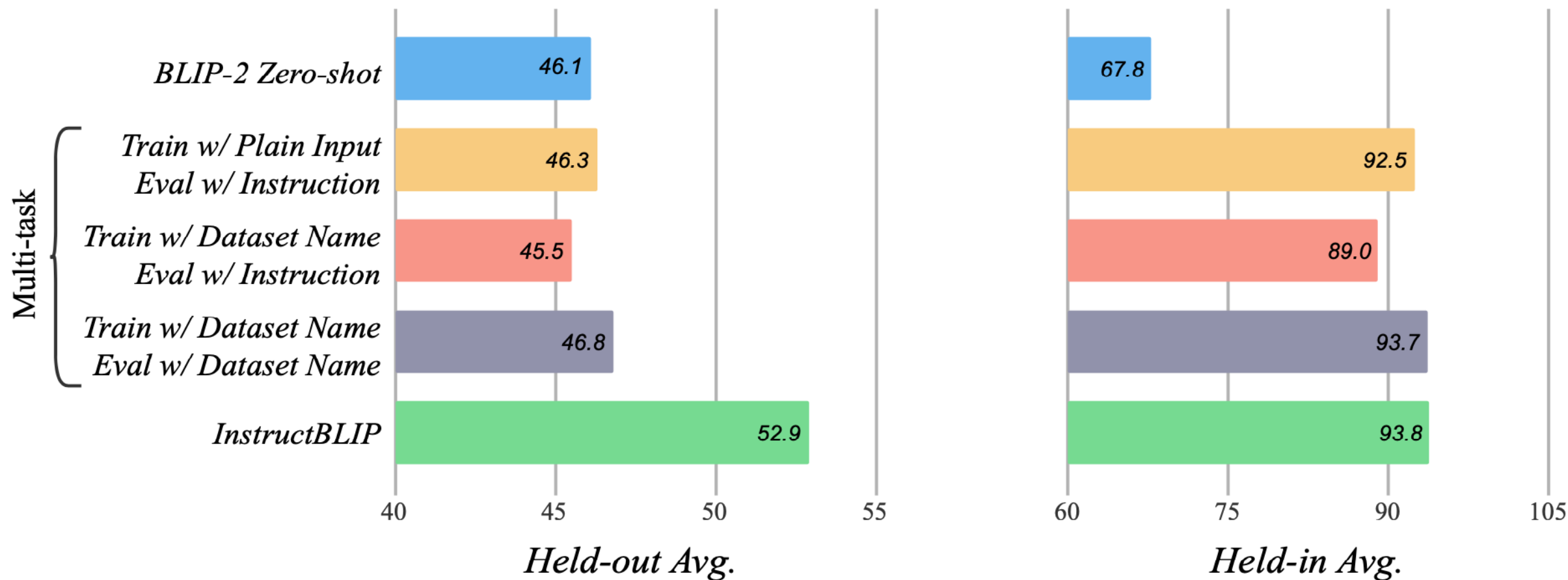
Vision-Language Models with Instruction Tuning



Vision-Language Instruction Tuning: Datasets



Instruction Tuning vs. Multitask Learning based on BLIP-2 FlanT5XL



Finetuning BLIP-2 and InstructBLIP on downstream datasets

	ScienceQA IMG	OCR-VQA	OKVQA	A-OKVQA			
				Direct Val	Answer Test	Multi-choice Val	Multi-choice Test
Previous SOTA	LLaVA [25] 89.0	GIT [42] 70.3	PaLM-E(562B) [9] 66.1	[15] 56.3	[36] 61.6	[15] 73.2	[36] 73.6
BLIP-2 (FlanT5 _{XXL})	89.5	72.7	54.7	57.6	53.7	80.2	76.2
InstructBLIP (FlanT5 _{XXL})	90.7	73.3	55.5	57.1	54.8	81.0	76.7
BLIP-2 (Vicuna-7B)	77.3	69.1	59.3	60.0	58.7	72.1	69.0
InstructBLIP (Vicuna-7B)	79.5	72.8	62.1	64.0	62.1	75.7	73.4

InstructBLIP

Vision-Language Models with Instruction Tuning

LLaVA

Large Language and Vision Assistant

MiniGPT-4

[Input Image]



[Input Instruction] Do you know who drew this painting?

[Model Response]

InstructBLIP 

Leonardo da Vinci.

LLaVA

The painting depicts a woman, commonly believed to be Mona Lisa, the famous artwork by Leonardo da Vinci. It is a portrait painting that showcases the woman's enigmatic smile and has become one of the most famous and iconic art pieces in the world. The original work is displayed in the Louvre Museum in Paris, and it is known for its intricate details, use of oil paint, and the artist's innovative techniques that contributed to its enduring appeal and mystery.

MiniGPT-4

The painting is the Mona Lisa, also known as the Mona Lisa, by Leonardo da Vinci. It is one of the most famous paintings in the world and is known for its enigmatic smile and the mysterious identity of the sitter. The painting is housed in the Louvre Museum in Paris, France.

Instruction Tuning Datasets

Dataset Name	Held-out	Dataset Description
COCO Caption [23]	✗	We use the large-scale COCO dataset for the image captioning task. Specifically, Karpathy split [17] is used, which divides the data into 82K/5K/5K images for the train/val/test sets.
Web CapFilt	✗	14M image-text pairs collected from the web with additional BLIP-generated synthetic captions, used in BLIP [21] and BLIP-2 [20].
NoCaps [3]	✓ (val)	NoCaps contains 15,100 images with 166,100 human-written captions for novel object image captioning.
Flickr30K [50]	✓ (test)	The Flickr30k dataset consists of 31K images collected from Flickr, each image has five ground truth captions. We use the test split as the held-out which contains 1K images.
TextCaps [37]	✗	TextCaps is an image captioning dataset that requires the model to comprehend and reason the text in images. Its train/val/test sets contain 21K/3K/3K images, respectively.
VQAv2 [11]	✗	VQAv2 is dataset for open-ended image question answering. It is split into 82K/40K/81K for train/val/test.
VizWiz [12]	✓ (test-dev)	A dataset contains visual questions asked by people who are blind. 8K images are used for the held-out evaluation.
GQA [16]	✓ (test-dev)	GQA contains image questions for scene understanding and reasoning. We use the balanced test-dev set as held-out.
Visual Spatial Reasoning	✓ (test)	VSR is a collection of image-text pairs, in which the text describes the spatial relation of two objects in the image. Models are required to classify true/false for the description. We use the zero-shot data split given in its official github repository.
IconQA [28]	✓ (test)	IconQA measures the abstract diagram understanding and comprehensive cognitive reasoning abilities of models. We use the test set of its multi-text-choice task for held-out evaluation.
OKVQA [29]	✗	OKVQA contains visual questions that require outside knowledge to answer. It has been split into 9K/5K for train and test.
A-OKVQA [35]	✗	A-OKVQA is a successor of OKVQA with more challenging and diverse questions. It has 17K/1K/6K questions for train/val/test.

Instruction Tuning Datasets

Dataset Name	Held-out	Dataset Description
ScienceQA [27]	✓ (test)	ScienceQA covers diverse science topics with corresponding lectures and explanations. In out settings, we only use the part with image context (IMG).
Visual Dialog [8]	✓ (val)	Visual dialog is a conversational question answering dataset. We use the val split as the held-out, which contains 2,064 images and each has 10 rounds.
OCR-VQA [30]	✗	OCR-VQA contains visual questions that require models to read text in the image. It has 800K/100K/100K for train/val/test, respectively.
TextVQA [38]	✓ (val)	TextVQA requires models to comprehend visual text to answer questions.
HatefulMemes [18]	✓ (val)	A binary classification dataset to justify whether a meme contains hateful content.
LLaVA-Instruct-150K [25]	✗	An instruction tuning dataset which has three parts: detailed caption (23K), reasoning (77K), conversation (58K).
MSVD-QA [46]	✓ (test)	We use the test set (13K video QA pairs) of MSVD-QA for held-out testing.
MSRVTT-QA [46]	✓ (test)	MSRVTT-QA has more complex scenes than MSVD, with 72K video QA pairs as the test set.
iVQA [48]	✓ (test)	iVQA is a video QA dataset with mitigated language biases. It has 6K/2K/2K samples for train/val/test.

Instruction Templates

Image Captioning

VQA

Vision Question Answering

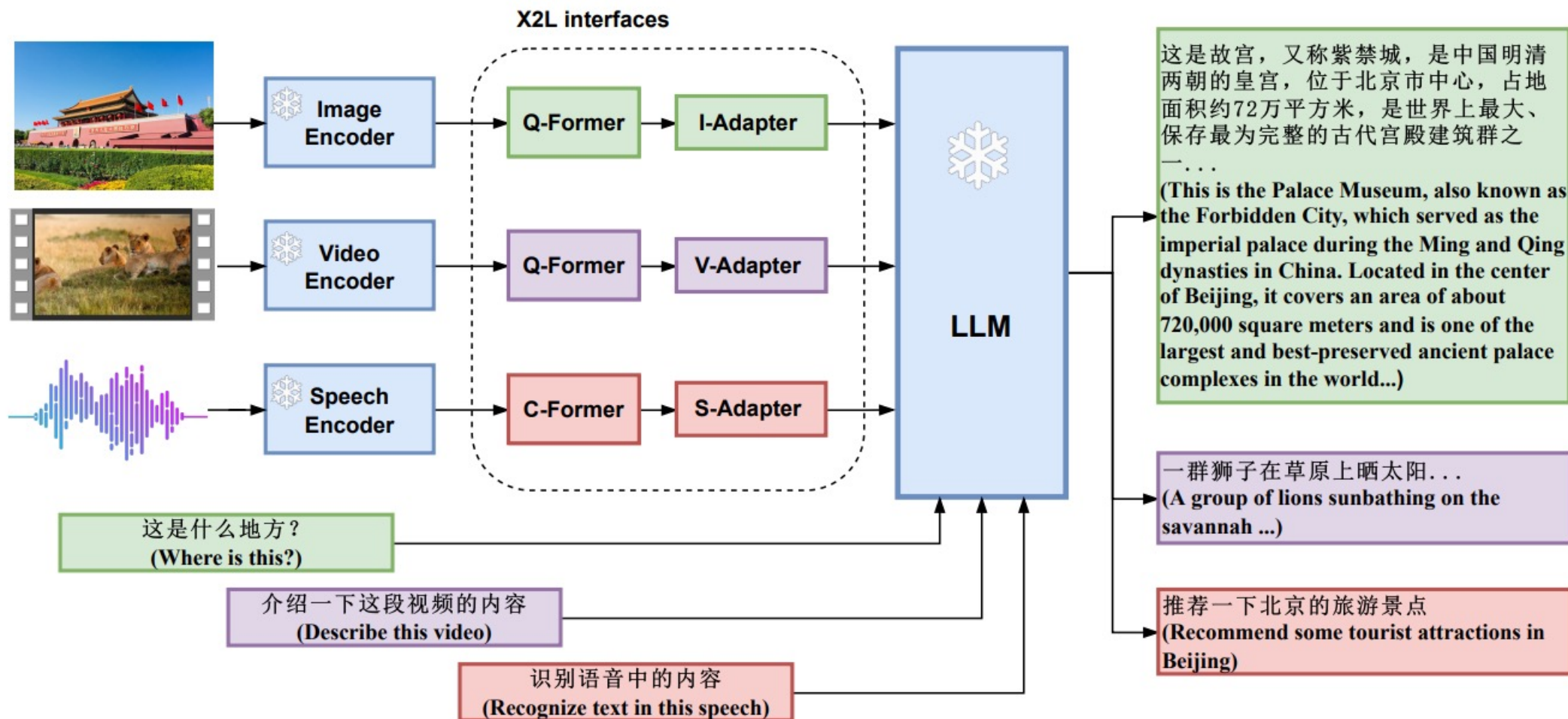
VQG

Vision Question Generation

Task	Instruction Template
Image Captioning	<p><Image>A short image caption: <Image>A short image description: <Image>A photo of <Image>An image that shows <Image>Write a short description for the image. <Image>Write a description for the photo. <Image>Provide a description of what is presented in the photo. <Image>Briefly describe the content of the image. <Image>Can you briefly explain what you see in the image? <Image>Could you use a few words to describe what you perceive in the photo? <Image>Please provide a short depiction of the picture. <Image>Using language, provide a short account of the image. <Image>Use a few words to illustrate what is happening in the picture.</p>
VQA	<p><Image>{Question} <Image>Question: {Question} <Image>{Question} A short answer to the question is <Image>Q: {Question} A: <Image>Question: {Question} Short answer: <Image>Given the image, answer the following question with no more than three words. {Question} <Image>Based on the image, respond to this question with a short answer: {Question}. Answer: <Image>Use the provided image to answer the question: {Question} Provide your answer as short as possible: <Image>What is the answer to the following question? "{Question}" <Image>The question "{Question}" can be answered using the image. A short answer is</p>
VQG	<p><Image>Given the image, generate a question whose answer is: {Answer}. Question: <Image>Based on the image, provide a question with the answer: {Answer}. Question: <Image>Given the visual representation, create a question for which the answer is "{Answer}". <Image>From the image provided, craft a question that leads to the reply: {Answer}. Question: <Image>Considering the picture, come up with a question where the answer is: {Answer}. <Image>Taking the image into account, generate an question that has the answer: {Answer}. Question:</p>

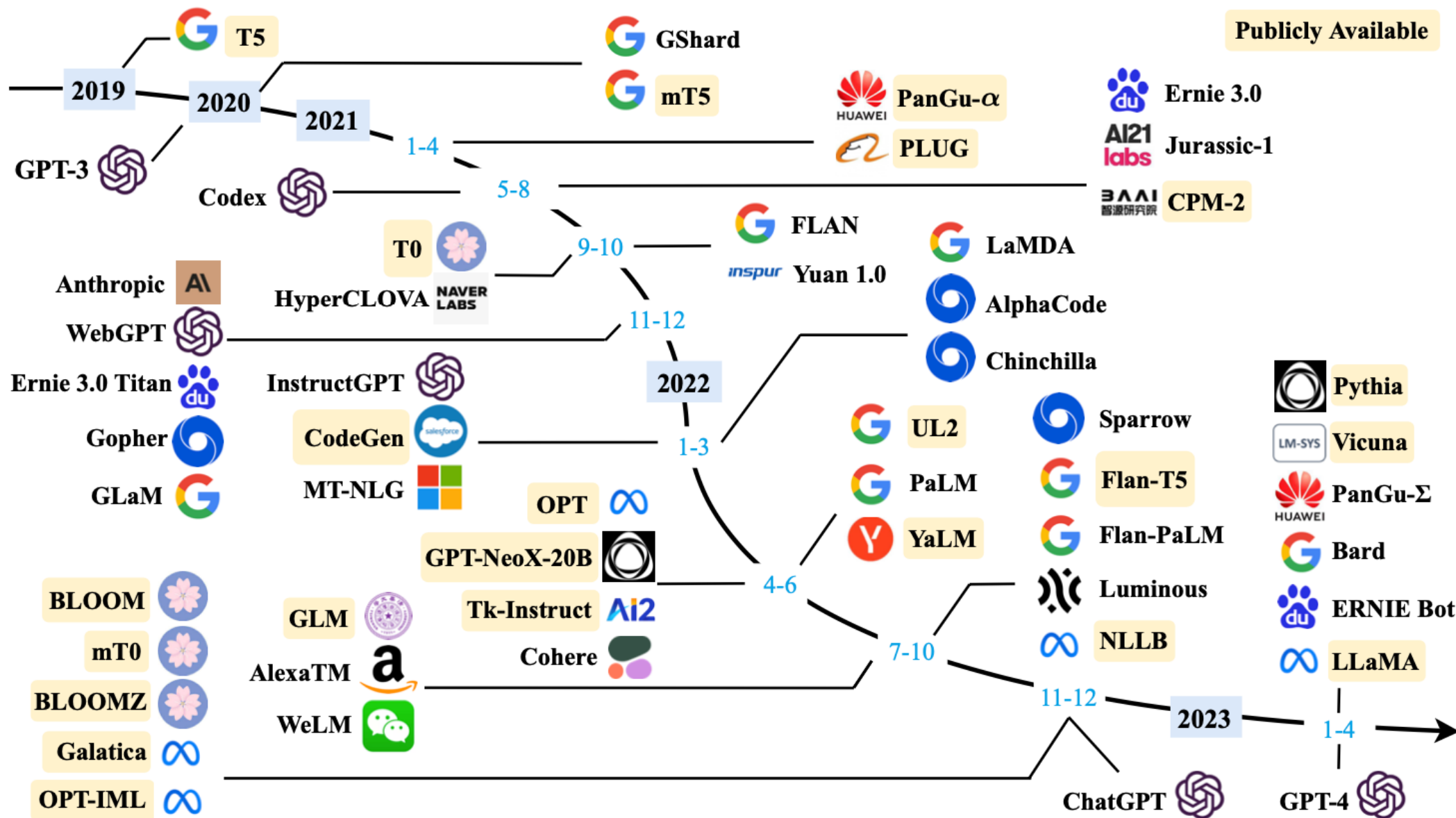
X-LLM:

Bootstrapping Advanced Large Language Models by Treating Multi-Modalities as Foreign Languages



Large Language Models (LLMs) Foundation Models

Large Language Models (LLMs) (larger than 10B)



Large Language Models (LLMs) (larger than 10B)

	Model	Release Time	Size (B)	Base Model	Adaptation		Pre-train Data Scale	Latest Data Timestamp	Hardware (GPUs / TPUs)	Training Time	Evaluation	
					IT	RLHF					ICL	CoT
Publicly Available	T5 [72]	Oct-2019	11	-	-	-	1T tokens	Apr-2019	1024 TPU v3	-	✓	-
	mT5 [73]	Oct-2020	13	-	-	-	1T tokens	-	-	-	✓	-
	PanGu- α [74]	Apr-2021	13*	-	-	-	1.1TB	-	2048 Ascend 910	-	✓	-
	CPM-2 [75]	Jun-2021	198	-	-	-	2.6TB	-	-	-	-	-
	T0 [28]	Oct-2021	11	T5	✓	-	-	-	512 TPU v3	27 h	✓	-
	CodeGen [76]	Mar-2022	16	-	-	-	577B tokens	-	-	-	✓	-
	GPT-NeoX-20B [77]	Apr-2022	20	-	-	-	825GB	-	96 40G A100	-	✓	-
	Tk-Instruct [78]	Apr-2022	11	T5	✓	-	-	-	256 TPU v3	4 h	✓	-
	UL2 [79]	May-2022	20	-	-	-	1T tokens	Apr-2019	512 TPU v4	-	✓	✓
	OPT [80]	May-2022	175	-	-	-	180B tokens	-	992 80G A100	-	✓	-
	NLLB [81]	Jul-2022	54.5	-	-	-	-	-	-	-	✓	-
	GLM [82]	Oct-2022	130	-	-	-	400B tokens	-	768 40G A100	60 d	✓	-
	Flan-T5 [83]	Oct-2022	11	T5	✓	-	-	-	-	-	✓	✓
	BLOOM [68]	Nov-2022	176	-	-	-	366B tokens	-	384 80G A100	105 d	✓	-
	mT0 [84]	Nov-2022	13	mT5	✓	-	-	-	-	-	✓	-
	Galactica [35]	Nov-2022	120	-	-	-	106B tokens	-	-	-	✓	✓
	BLOOMZ [84]	Nov-2022	176	BLOOM	✓	-	-	-	-	-	✓	-
	OPT-IML [85]	Dec-2022	175	OPT	✓	-	-	-	128 40G A100	-	✓	✓
	LLaMA [57]	Feb-2023	65	-	-	-	1.4T tokens	-	2048 80G A100	21 d	✓	-
	Pythia [86]	Apr-2023	12	-	-	-	300B tokens	-	256 40G A100	-	✓	-

Large Language Models (LLMs) (larger than 10B)

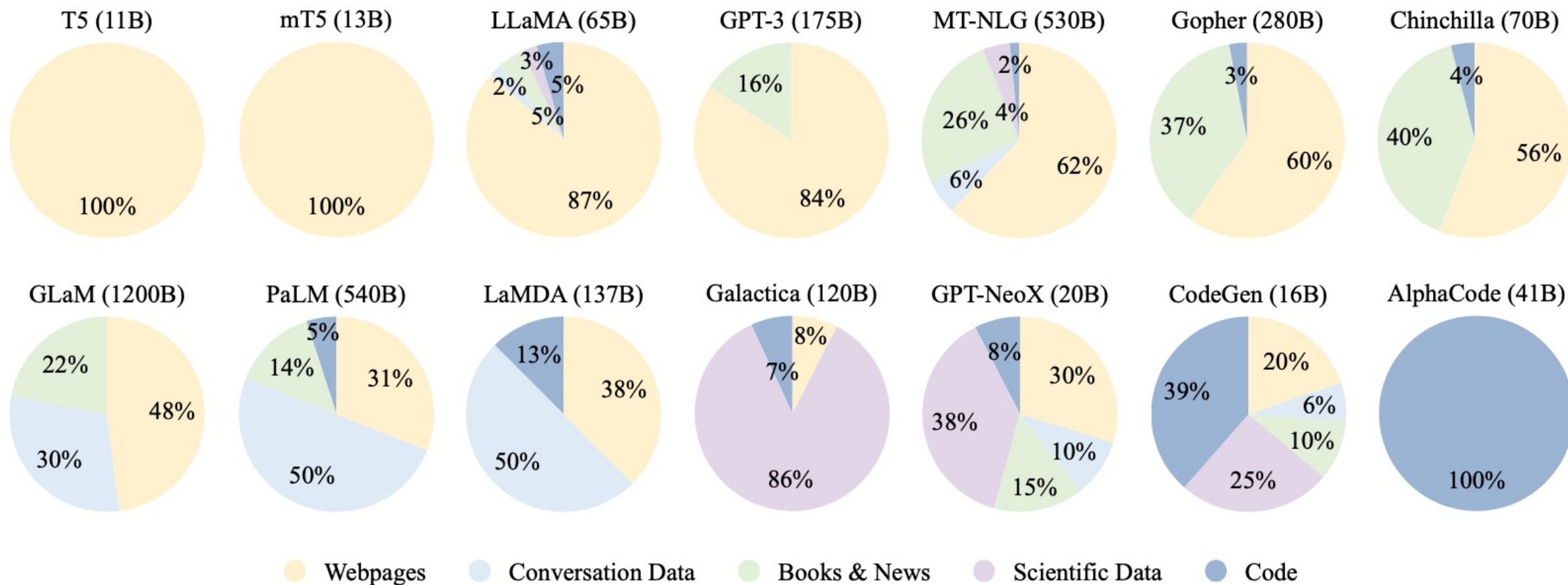
	Model	Release Time	Size (B)	Base Model	Adaptation		Pre-train Data Scale	Latest Data Timestamp	Hardware (GPUs / TPUs)	Training Time	Evaluation	
					IT	RLHF					ICL	CoT
Closed Source	GPT-3 [55]	May-2020	175	-	-	-	300B tokens	-	-	-	✓	-
	GShard [87]	Jun-2020	600	-	-	-	1T tokens	-	2048 TPU v3	4 d	-	-
	Codex [88]	Jul-2021	12	GPT-3	-	-	100B tokens	May-2020	-	-	✓	-
	ERNIE 3.0 [89]	Jul-2021	10	-	-	-	375B tokens	-	384 V100	-	✓	-
	Jurassic-1 [90]	Aug-2021	178	-	-	-	300B tokens	-	800 GPU	-	✓	-
	HyperCLOVA [91]	Sep-2021	82	-	-	-	300B tokens	-	1024 A100	13.4 d	✓	-
	FLAN [62]	Sep-2021	137	LaMDA	✓	-	-	-	128 TPU v3	60 h	✓	-
	Yuan 1.0 [92]	Oct-2021	245	-	-	-	180B tokens	-	2128 GPU	-	✓	-
	Anthropic [93]	Dec-2021	52	-	-	-	400B tokens	-	-	-	✓	-
	WebGPT [71]	Dec-2021	175	GPT-3	-	✓	-	-	-	-	✓	-
	Gopher [59]	Dec-2021	280	-	-	-	300B tokens	-	4096 TPU v3	920 h	✓	-
	ERNIE 3.0 Titan [94]	Dec-2021	260	-	-	-	300B tokens	-	2048 V100	28 d	✓	-
	GLaM [95]	Dec-2021	1200	-	-	-	280B tokens	-	1024 TPU v4	574 h	✓	-
	LaMDA [96]	Jan-2022	137	-	-	-	2.81T tokens	-	1024 TPU v3	57.7 d	-	-
	MT-NLG [97]	Jan-2022	530	-	-	-	270B tokens	-	4480 80G A100	-	✓	-
	AlphaCode [98]	Feb-2022	41	-	-	-	967B tokens	Jul-2021	-	-	-	-
	InstructGPT [61]	Mar-2022	175	GPT-3	✓	✓	-	-	-	-	✓	-
	Chinchilla [34]	Mar-2022	70	-	-	-	1.4T tokens	-	-	-	✓	-
	PaLM [56]	Apr-2022	540	-	-	-	780B tokens	-	6144 TPU v4	-	✓	✓
	AlexaTM [99]	Aug-2022	20	-	-	-	1.3T tokens	-	128 A100	120 d	✓	✓
	Sparrow [100]	Sep-2022	70	-	-	✓	-	-	64 TPU v3	-	✓	-
	WeLM [101]	Sep-2022	10	-	-	-	300B tokens	-	128 A100 40G	24 d	✓	-
	U-PaLM [102]	Oct-2022	540	PaLM	-	-	-	-	512 TPU v4	5 d	✓	✓
	Flan-PaLM [83]	Oct-2022	540	PaLM	✓	-	-	-	512 TPU v4	37 h	✓	✓
	Flan-U-PaLM [83]	Oct-2022	540	U-PaLM	✓	-	-	-	-	-	✓	✓
	GPT-4 [46]	Mar-2023	-	-	✓	✓	-	-	-	-	✓	✓
	PanGu- Σ [103]	Mar-2023	1085	PanGu- α	-	-	329B tokens	-	512 Ascend 910	100 d	✓	-

Statistics of Commonly-used Data Sources for LLMs

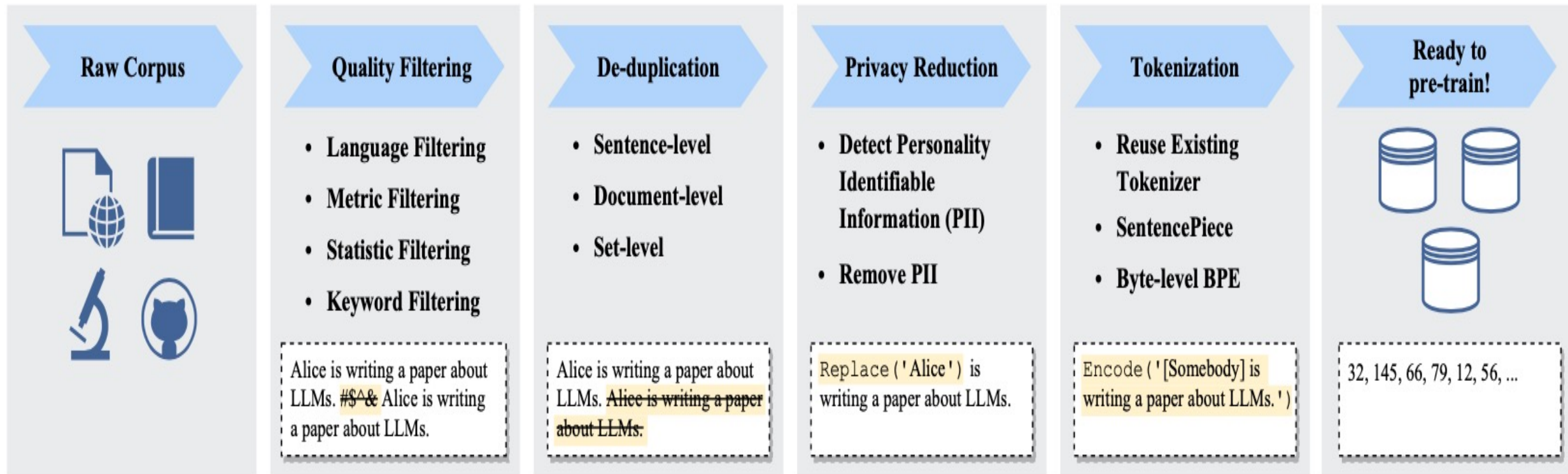
Corpora	Size	Source	Latest Update Time
BookCorpus [109]	5GB	Books	Dec-2015
Gutenberg [110]	-	Books	Dec-2021
C4 [72]	800GB	CommonCrawl	Apr-2019
CC-Stories-R [111]	31GB	CommonCrawl	Sep-2019
CC-NEWS [27]	78GB	CommonCrawl	Feb-2019
REALNEWS [112]	120GB	CommonCrawl	Apr-2019
OpenWebText [113]	38GB	Reddit links	Mar-2023
Pushift.io [114]	-	Reddit links	Mar-2023
Wikipedia [115]	-	Wikipedia	Mar-2023
BigQuery [116]	-	Codes	Mar-2023
the Pile [117]	800GB	Other	Dec-2020
ROOTS [118]	1.6TB	Other	Jun-2022

Source: Wanyin Liu Zhao, Kun Zhao, Junyi Li, Hanyu Tang, Xiaohu Wang, Lupeng Hou, Mingqian Wang et al. (2023). A Survey of Large Language Models. arXiv preprint arXiv:2305.10223.

Ratios of various data sources in the pre-training data for existing LLMs



Typical Data Preprocessing Pipeline for Pre-training Large Language Models (LLMs)



LLMs with Public Configuration Details

Model	Category	Size	Normalization	PE	Activation	Bias	#L	#H	d_{model}	MCL
GPT3 [55]	Causal decoder	175B	Pre Layer Norm	Learned	GeLU	✓	96	96	12288	2048
PanGU- α [74]	Causal decoder	207B	Pre Layer Norm	Learned	GeLU	✓	64	128	16384	1024
OPT [80]	Causal decoder	175B	Pre Layer Norm	Learned	ReLU	✓	96	96	12288	2048
PaLM [56]	Causal decoder	540B	Pre Layer Norm	RoPE	SwiGLU	×	118	48	18432	2048
BLOOM [68]	Causal decoder	176B	Pre Layer Norm	ALiBi	GeLU	✓	70	112	14336	2048
MT-NLG [97]	Causal decoder	530B	-	-	-	-	105	128	20480	2048
Gopher [59]	Causal decoder	280B	Pre RMS Norm	Relative	-	-	80	128	16384	2048
Chinchilla [34]	Causal decoder	70B	Pre RMS Norm	Relative	-	-	80	64	8192	-
Galactica [35]	Causal decoder	120B	Pre Layer Norm	Learned	GeLU	×	96	80	10240	2048
LaMDA [96]	Causal decoder	137B	-	Relative	GeGLU	-	64	128	8192	-
Jurassic-1 [90]	Causal decoder	178B	Pre Layer Norm	Learned	GeLU	✓	76	96	13824	2048
LLaMA [57]	Causal decoder	65B	Pre RMS Norm	RoPE	SwiGLU	✓	80	64	8192	2048
GLM-130B [82]	Prefix decoder	130B	Post Deep Norm	RoPE	GeGLU	✓	70	96	12288	2048
T5 [72]	Encoder-decoder	11B	Pre RMS Norm	Relative	ReLU	×	24	128	1024	512

Note: PE denotes position embedding, #L denotes the number of layers, #H denotes the number of attention heads, d_{model} denotes the size of hidden states, and MCL denotes the maximum context length during training.

Detailed Optimization Settings of LLMs

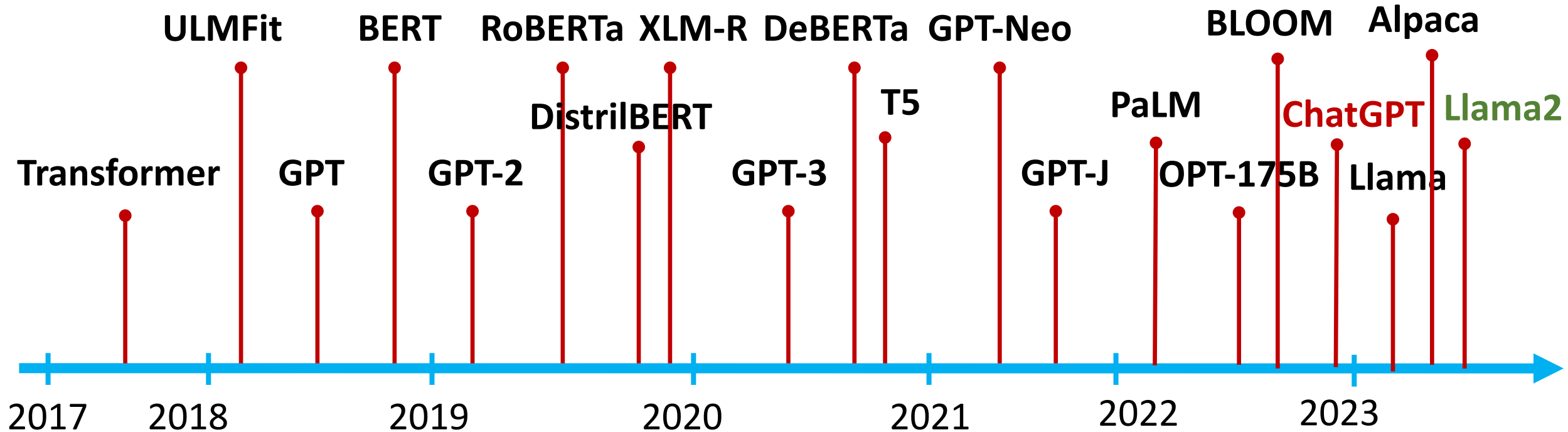
Model	Batch Size (#tokens)	Learning Rate	Warmup	Decay Method	Optimizer	Precision Type	Weight Decay	Grad Clip	Dropout
GPT3 (175B)	32K→3.2M	6×10^{-5}	yes	cosine decay to 10%	Adam	FP16	0.1	1.0	-
PanGu- α (200B)	-	2×10^{-5}	-	-	Adam	-	0.1	-	-
OPT (175B)	2M	1.2×10^{-4}	yes	manual decay	AdamW	FP16	0.1	-	0.1
PaLM (540B)	1M→4M	1×10^{-2}	no	inverse square root	Adafactor	BF16	lr^2	1.0	0.1
BLOOM (176B)	4M	6×10^{-5}	yes	cosine decay to 10%	Adam	BF16	0.1	1.0	0.0
MT-NLG (530B)	64 K→3.75M	5×10^{-5}	yes	cosine decay to 10%	Adam	BF16	0.1	1.0	-
Gopher (280B)	3M→6M	4×10^{-5}	yes	cosine decay to 10%	Adam	BF16	-	1.0	-
Chinchilla (70B)	1.5M→3M	1×10^{-4}	yes	cosine decay to 10%	AdamW	BF16	-	-	-
Galactica (120B)	2M	7×10^{-6}	yes	linear decay to 10%	AdamW	-	0.1	1.0	0.1
LaMDA (137B)	256K	-	-	-	-	BF16	-	-	-
Jurassic-1 (178B)	32 K→3.2M	6×10^{-5}	yes	-	-	-	-	-	-
LLaMA (65B)	4M	1.5×10^{-4}	yes	cosine decay to 10%	AdamW	-	0.1	1.0	-
GLM (130B)	0.4M→8.25M	8×10^{-5}	yes	cosine decay to 10%	AdamW	FP16	0.1	1.0	0.1
T5 (11B)	64K	1×10^{-2}	no	inverse square root	AdaFactor	-	-	-	0.1
ERNIE 3.0 Titan (260B)	-	1×10^{-4}	-	-	Adam	FP16	0.1	1.0	-
PanGu- Σ (1.085T)	0.5M	2×10^{-5}	yes	-	Adam	FP16	-	-	-

Outline

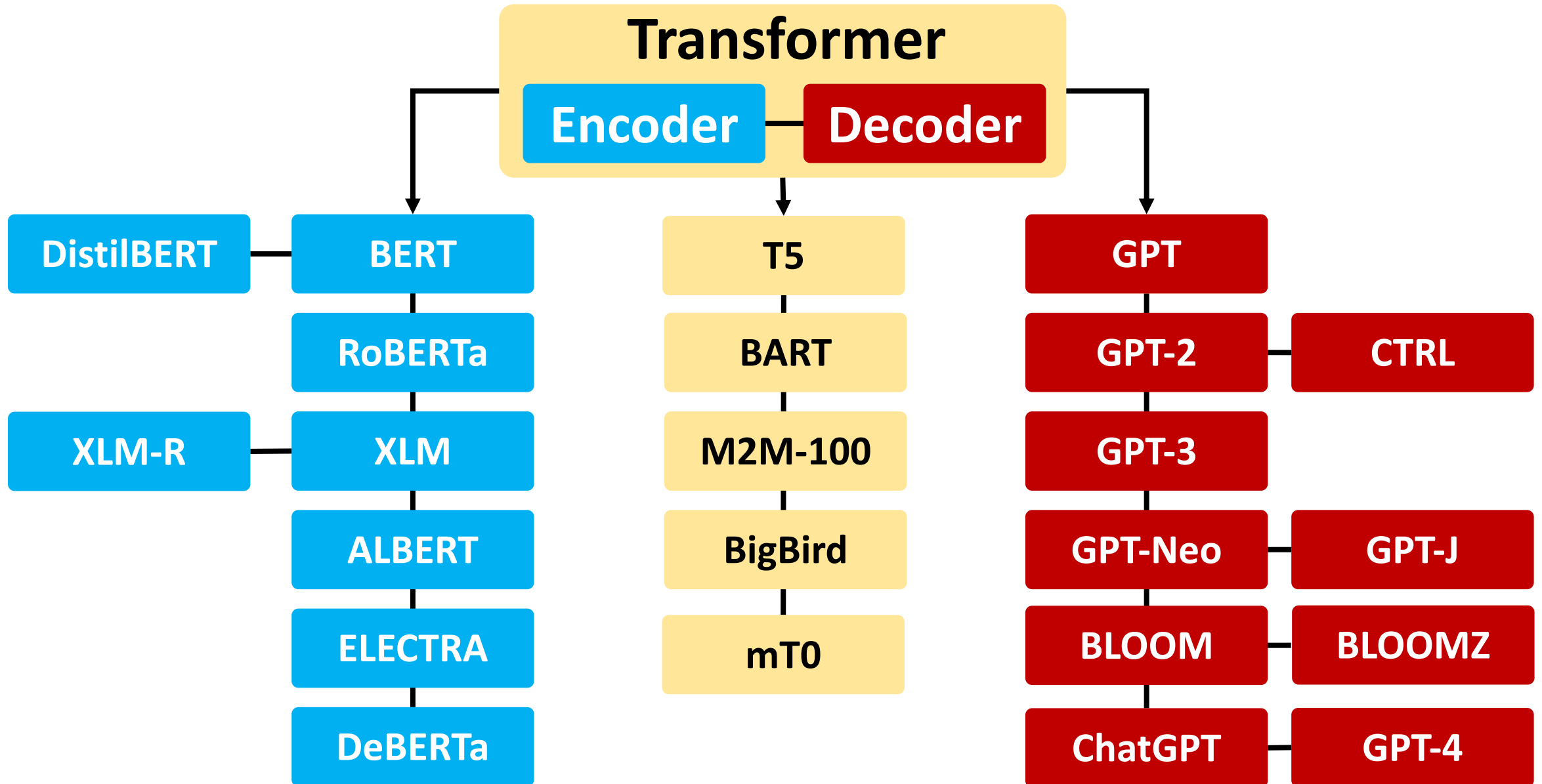
- Introduction
- Overview of Generative AI
- Overview of Large Language Models (LLMs)
- **Foundation of Transformers: Attention Mechanism**
- Fine-tuning LLM for Question Answering System
- Fine-tuning LLM for Dialogue System
- Challenges and Limitations of Generative AI for QA and Dialogue Systems
- Q & A

Foundation of Transformers: Attention Mechanism

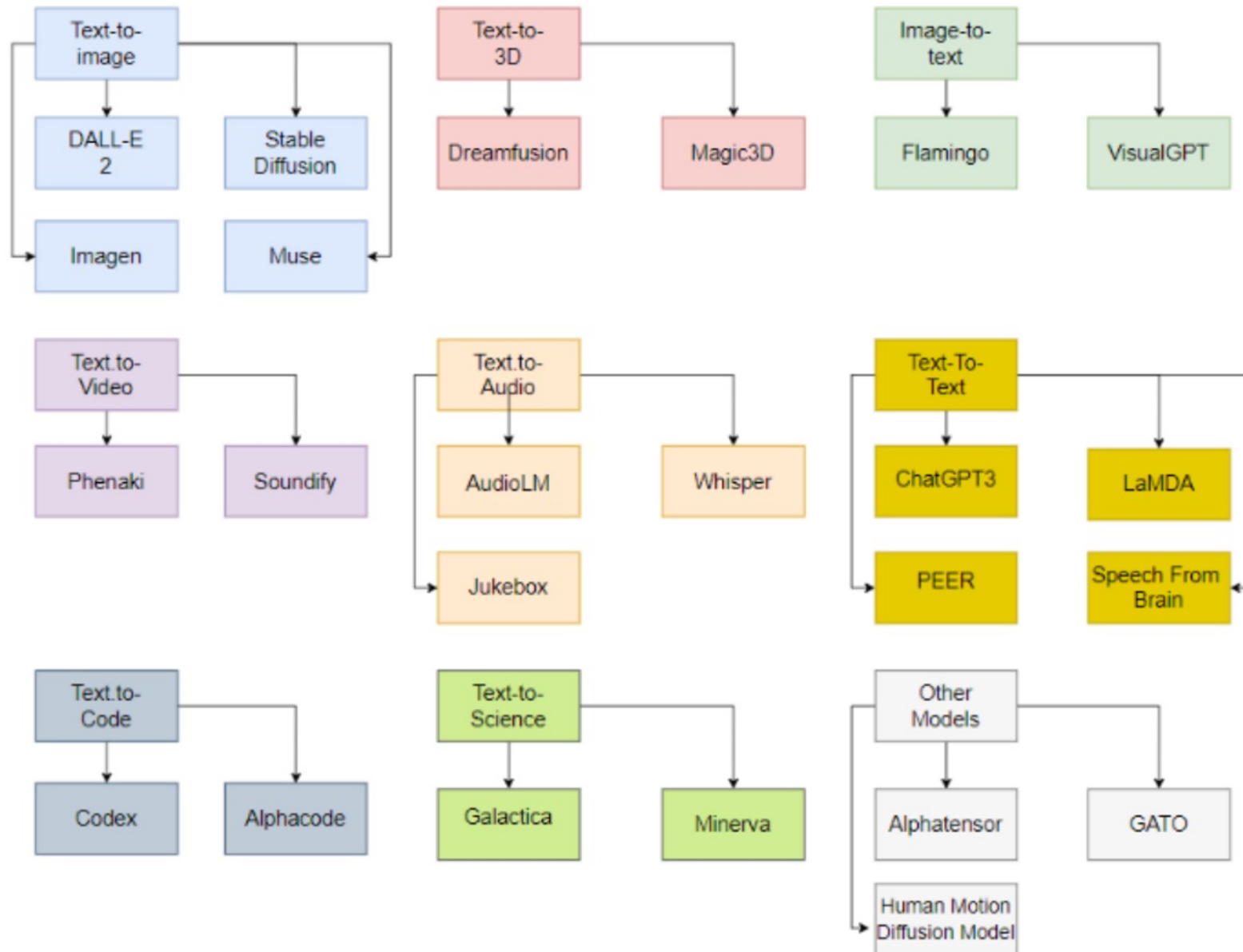
The Transformers Timeline



Transformer Models



Generative AI Models

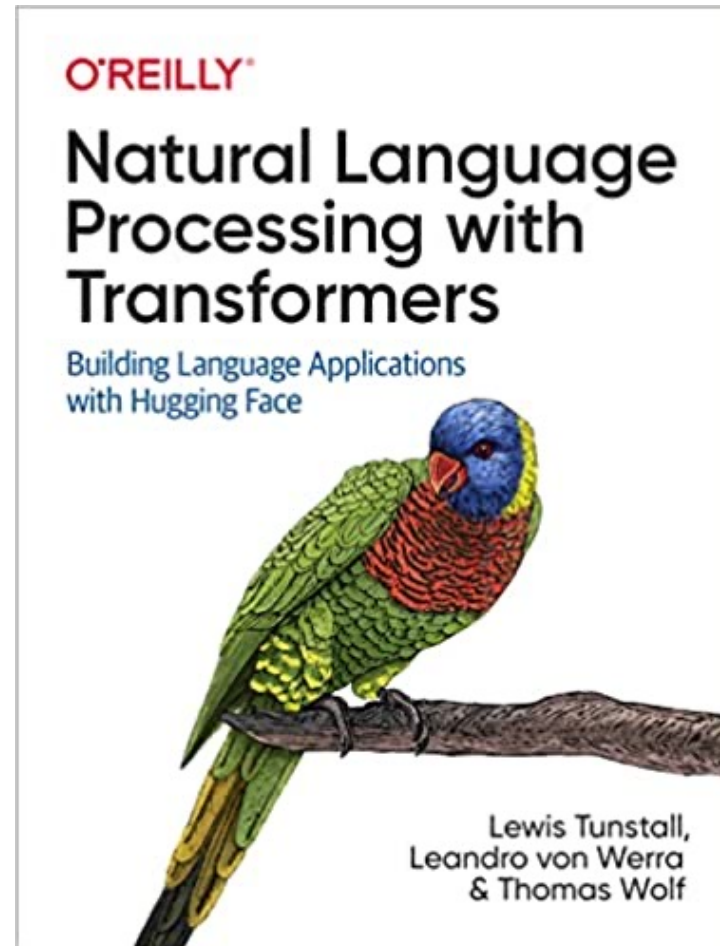


**ChatGPT
is not
all you need**

**Attention
is
all you need**

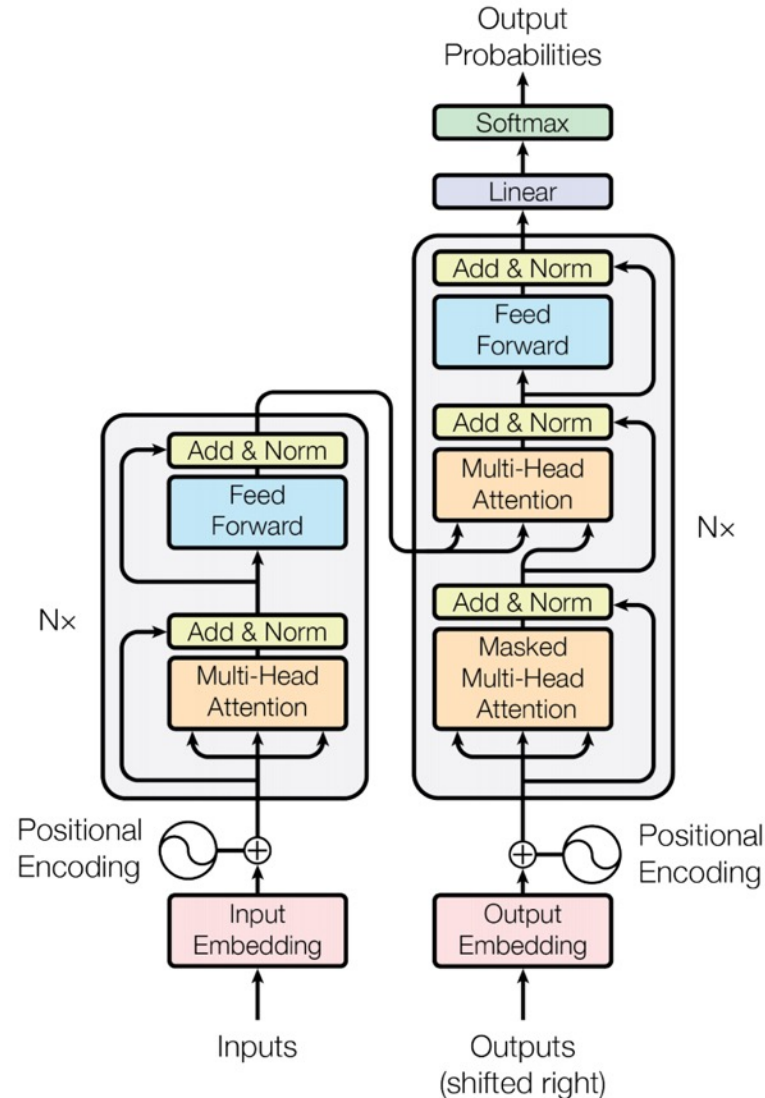
Natural Language Processing with Transformers

Lewis Tunstall, Leandro von Werra, and Thomas Wolf (2022),
Natural Language Processing with Transformers:
Building Language Applications with Hugging Face,
O'Reilly Media.



Transformer (Attention is All You Need)

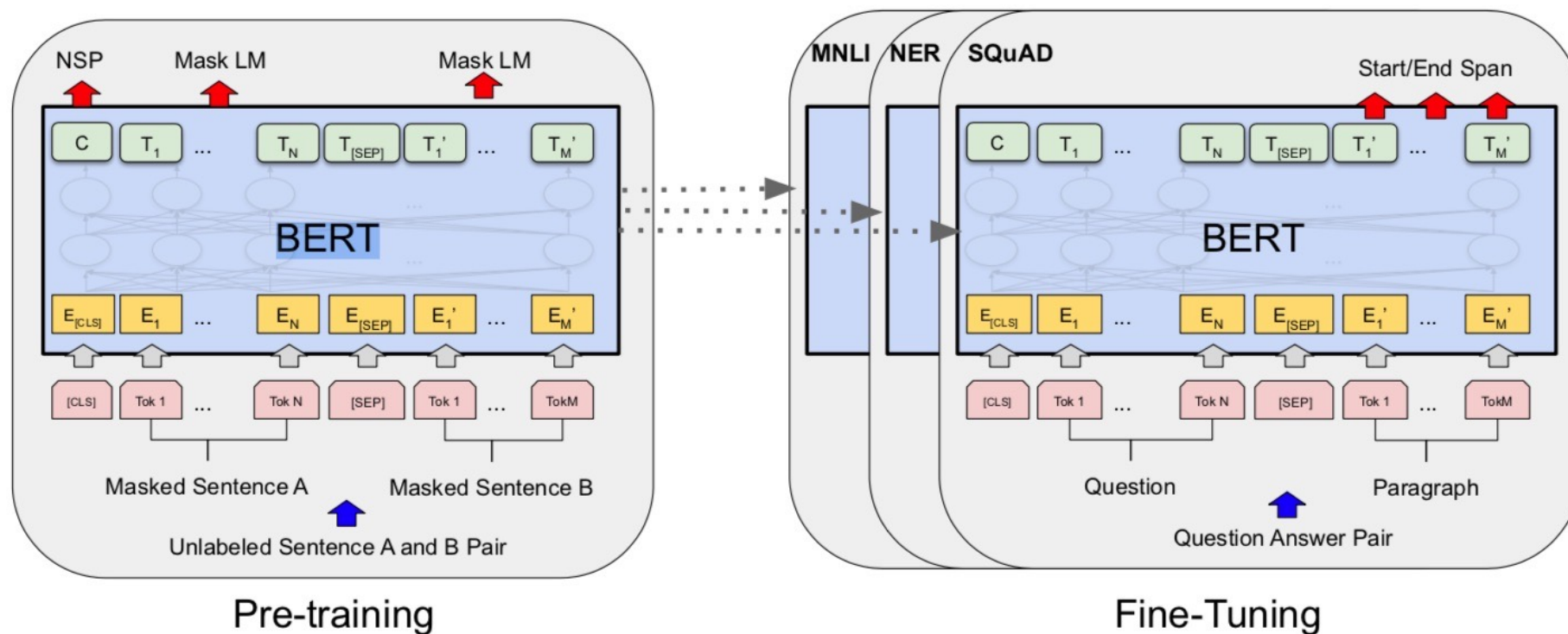
(Vaswani et al., 2017)



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT (Bidirectional Encoder Representations from Transformers)

Overall pre-training and fine-tuning procedures for BERT



BERT:

Pre-training of Deep Bidirectional Transformers for Language Understanding

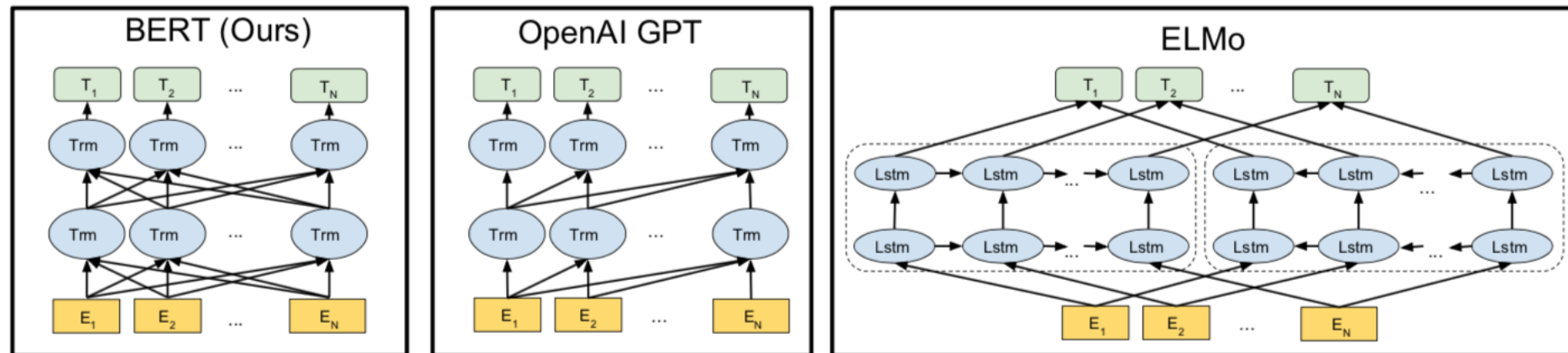
**BERT: Pre-training of Deep Bidirectional Transformers for
Language Understanding**

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language

`{jacobdevlin, mingweichang, kentonl, kristout}@google.com`

BERT

Bidirectional Encoder Representations from Transformers



Pre-training model architectures

BERT uses a bidirectional Transformer.

OpenAI GPT uses a left-to-right Transformer.

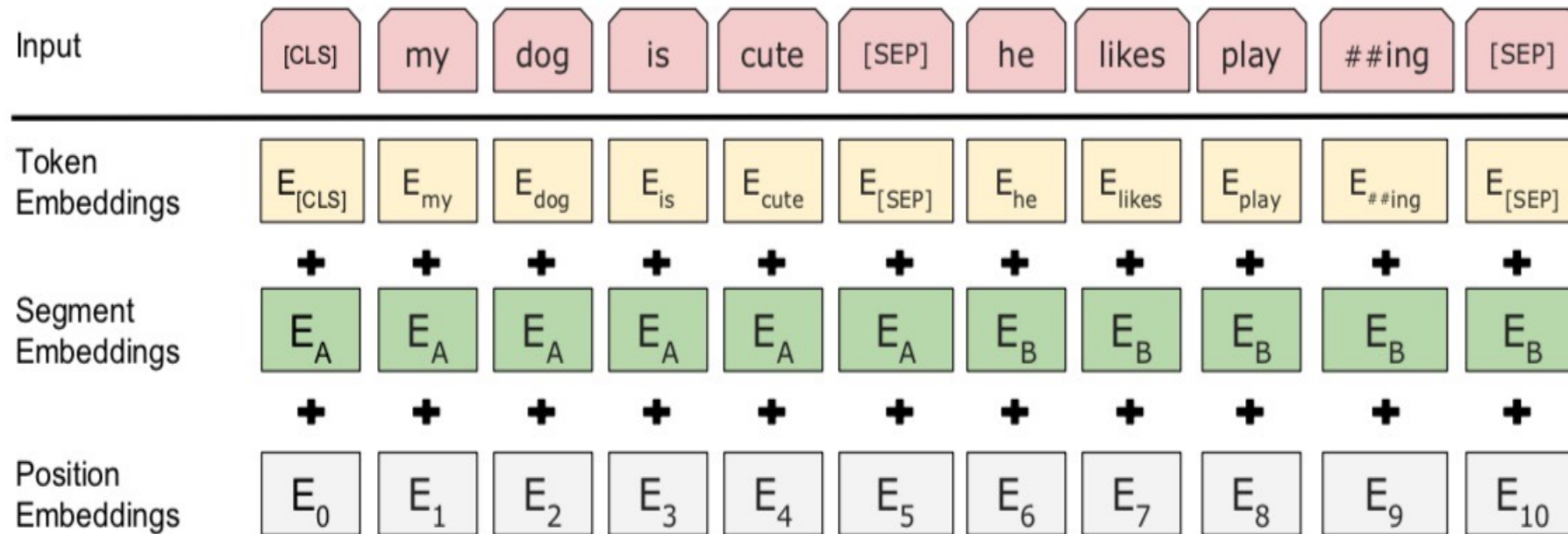
ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTM to generate features for downstream tasks.

Among three, only BERT representations are jointly conditioned on both left and right context in all layers.

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

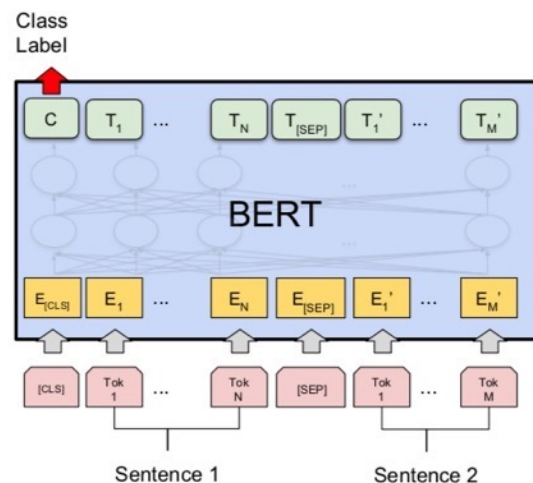
BERT (Bidirectional Encoder Representations from Transformers)

BERT input representation

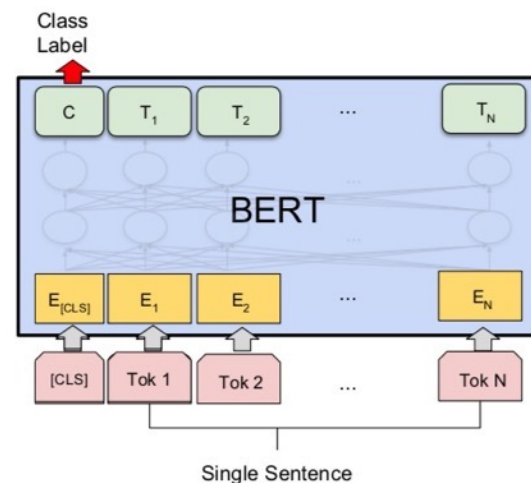


The input embeddings is the sum of the token embeddings, the segmentation embeddings and the position embeddings.

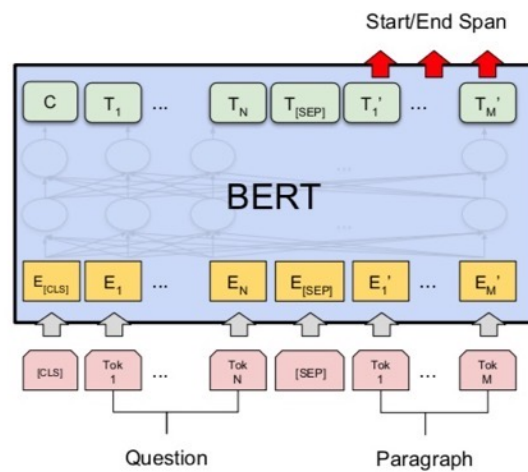
Fine-tuning BERT on NLP Tasks



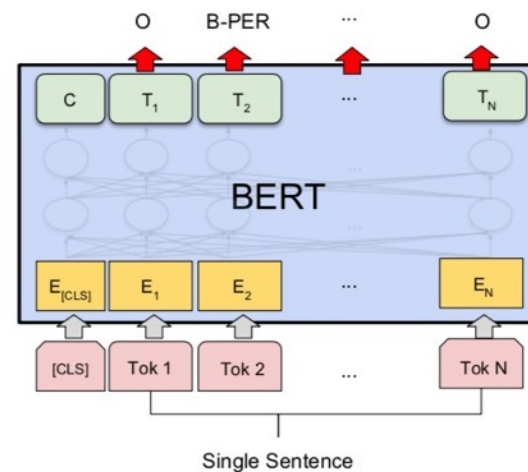
(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG



(b) Single Sentence Classification Tasks:
SST-2, CoLA



(c) Question Answering Tasks:
SQuAD v1.1

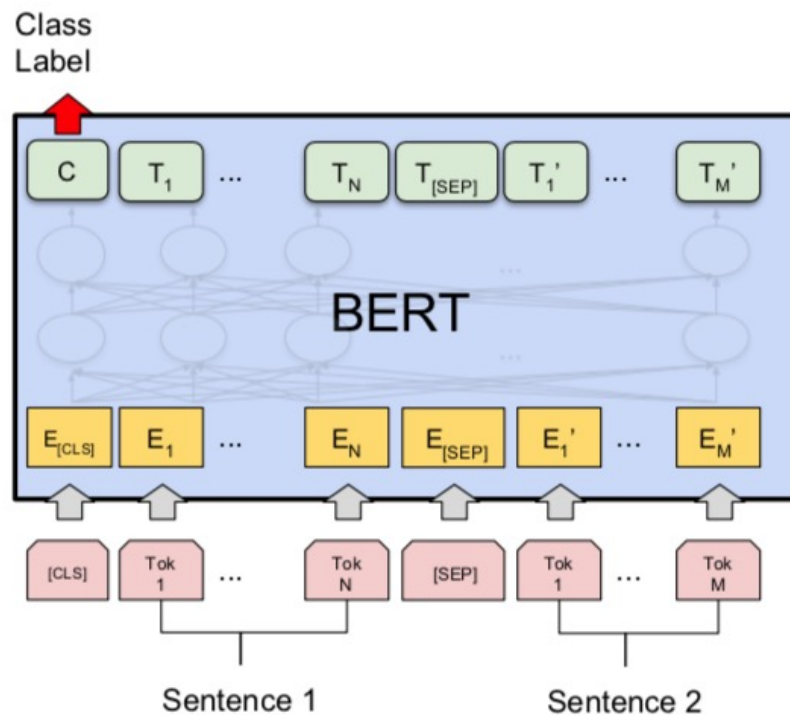


(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

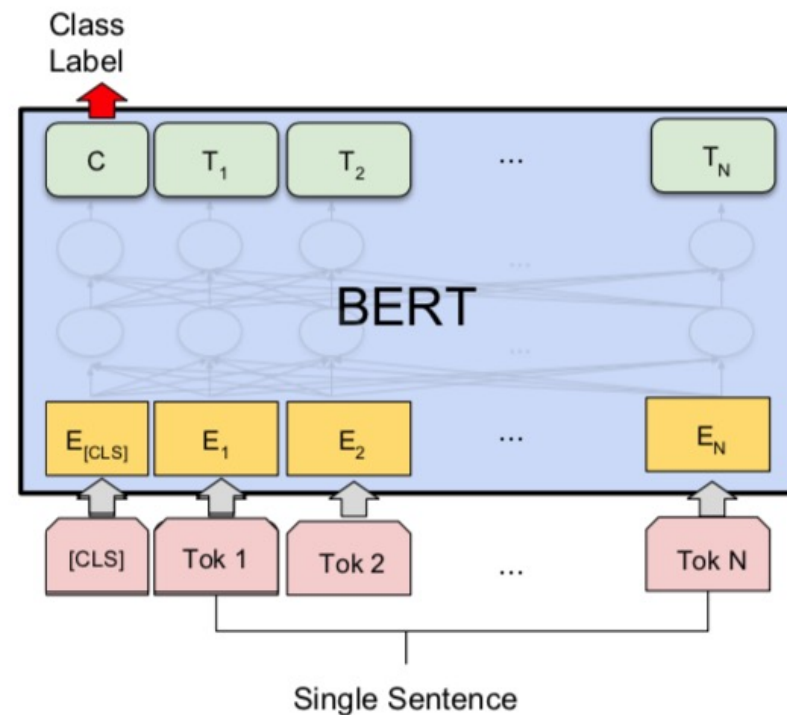
Source: Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018).

"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv preprint arXiv:1810.04805

BERT Sequence-level tasks

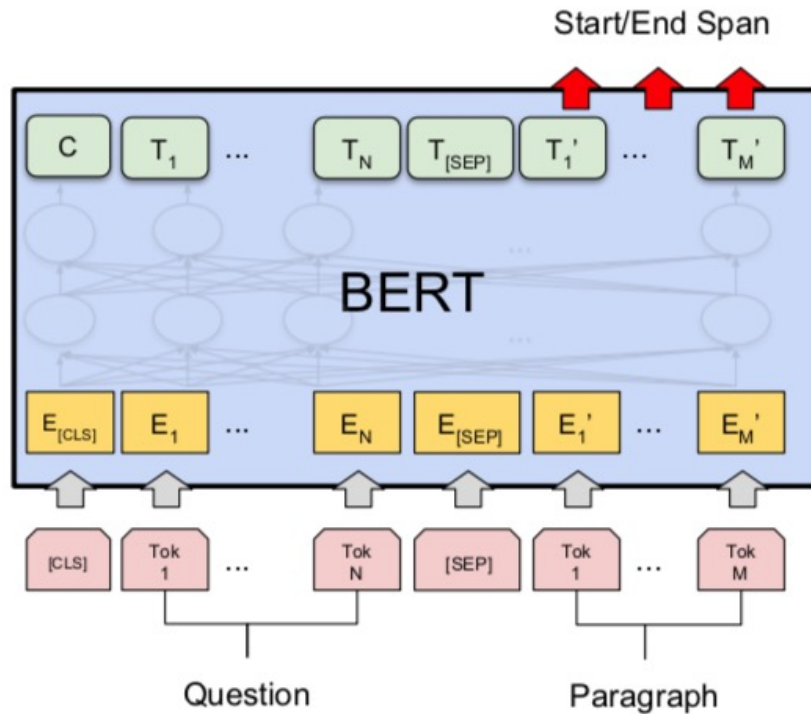


(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG

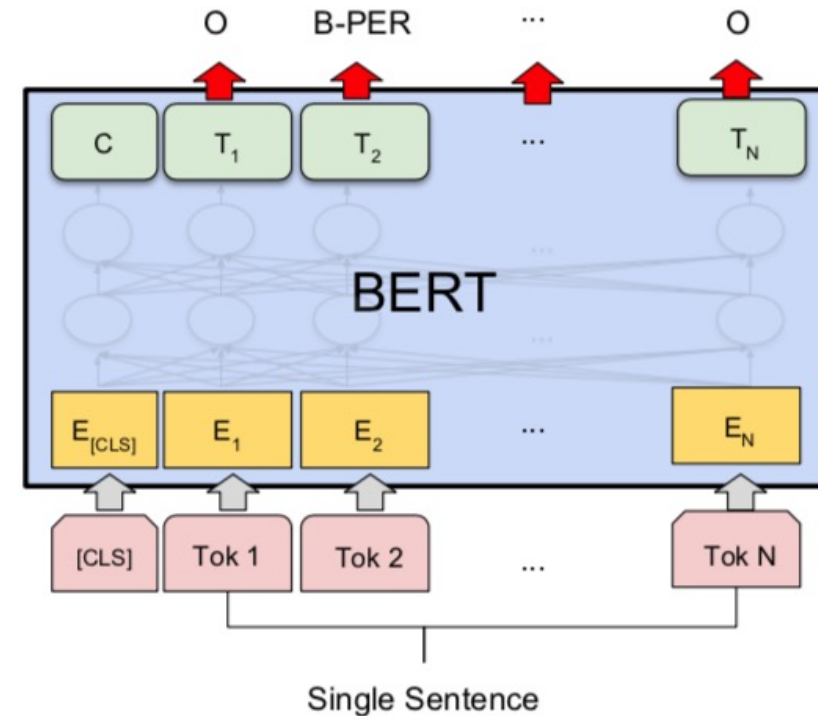


(b) Single Sentence Classification Tasks:
SST-2, CoLA

BERT Token-level tasks



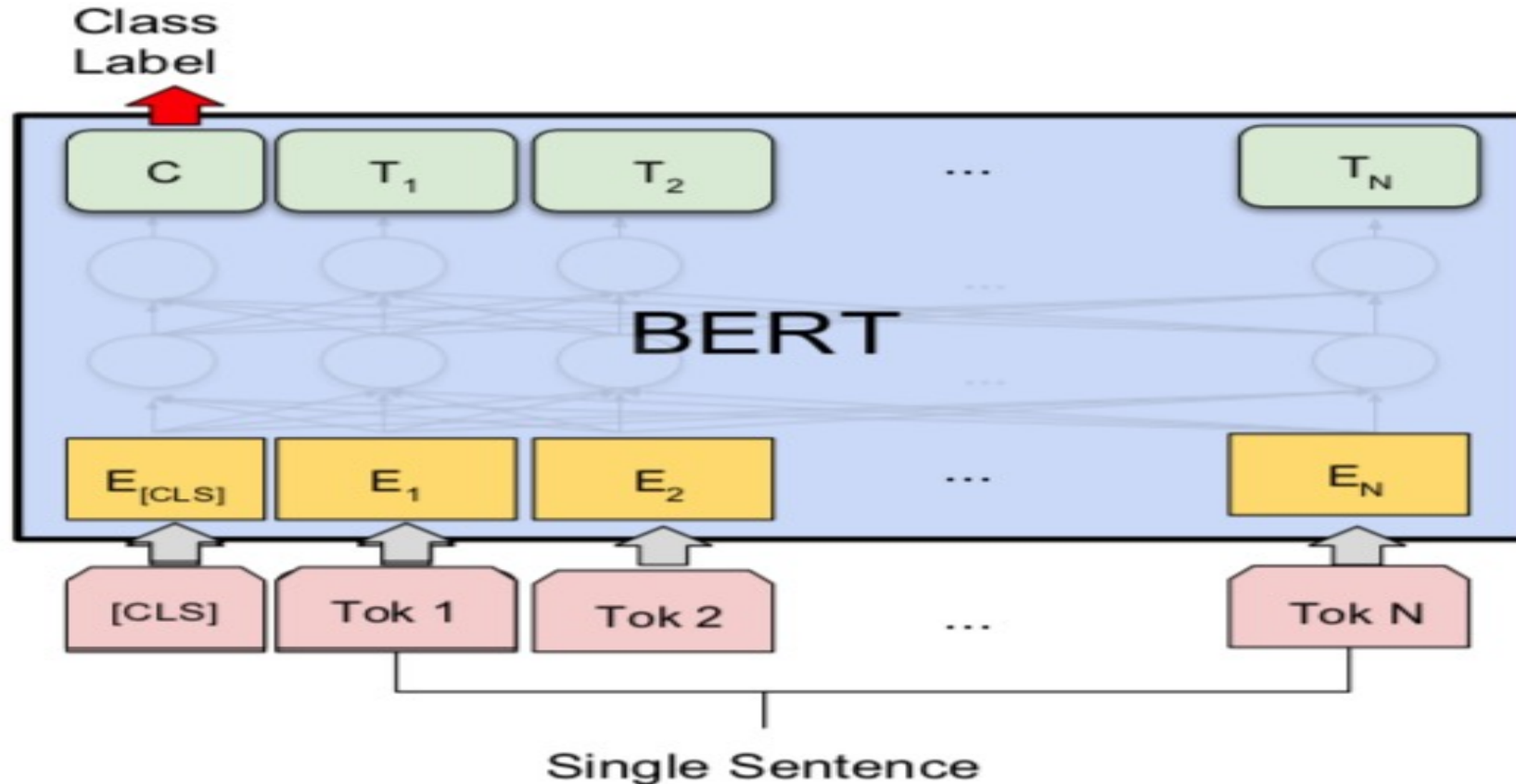
(c) Question Answering Tasks:
SQuAD v1.1



(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

Sentiment Analysis:

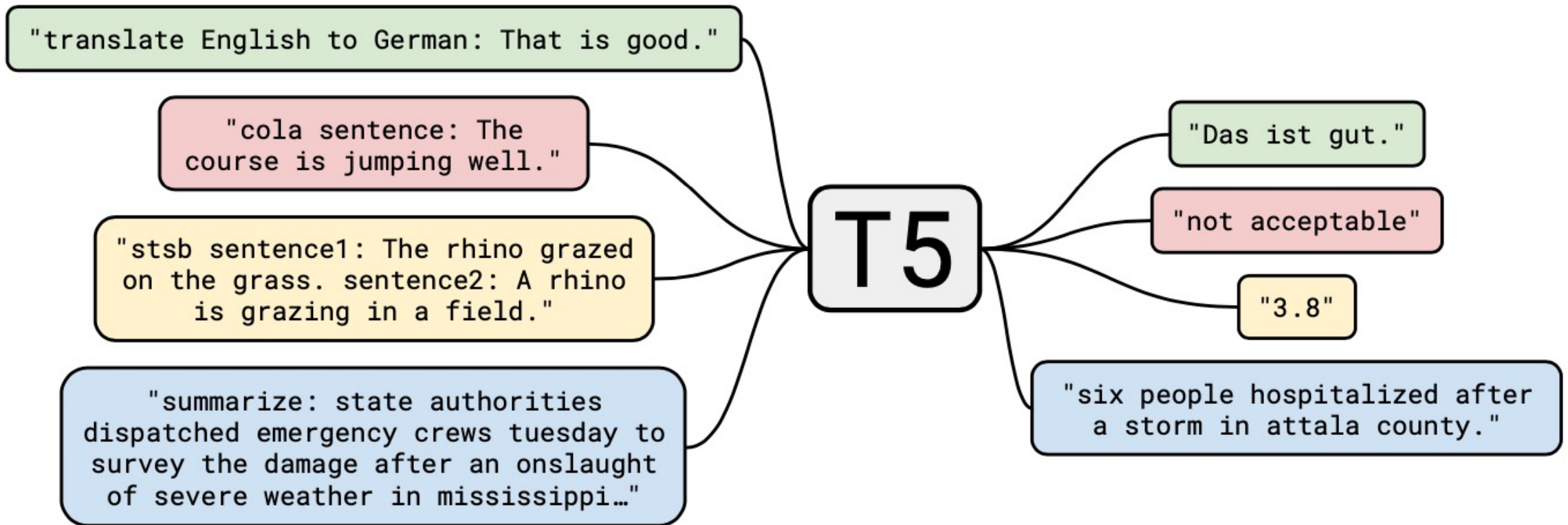
Single Sentence Classification



(b) Single Sentence Classification Tasks:
SST-2, CoLA

T5

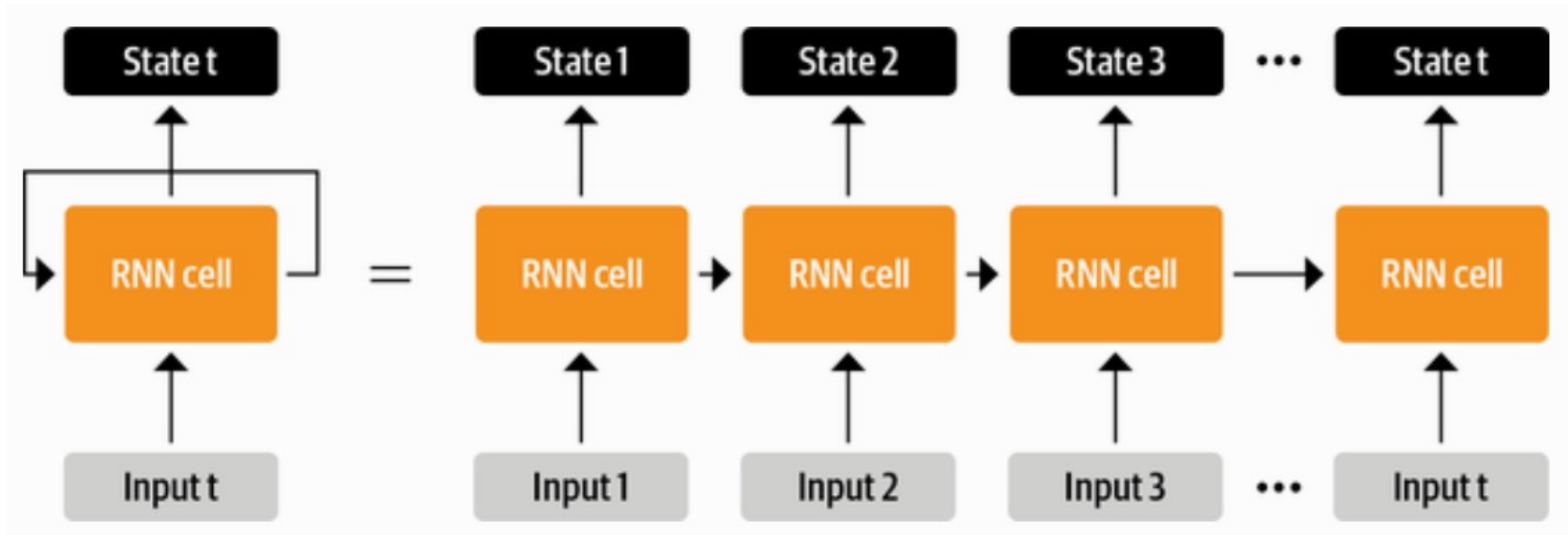
Text-to-Text Transfer Transformer



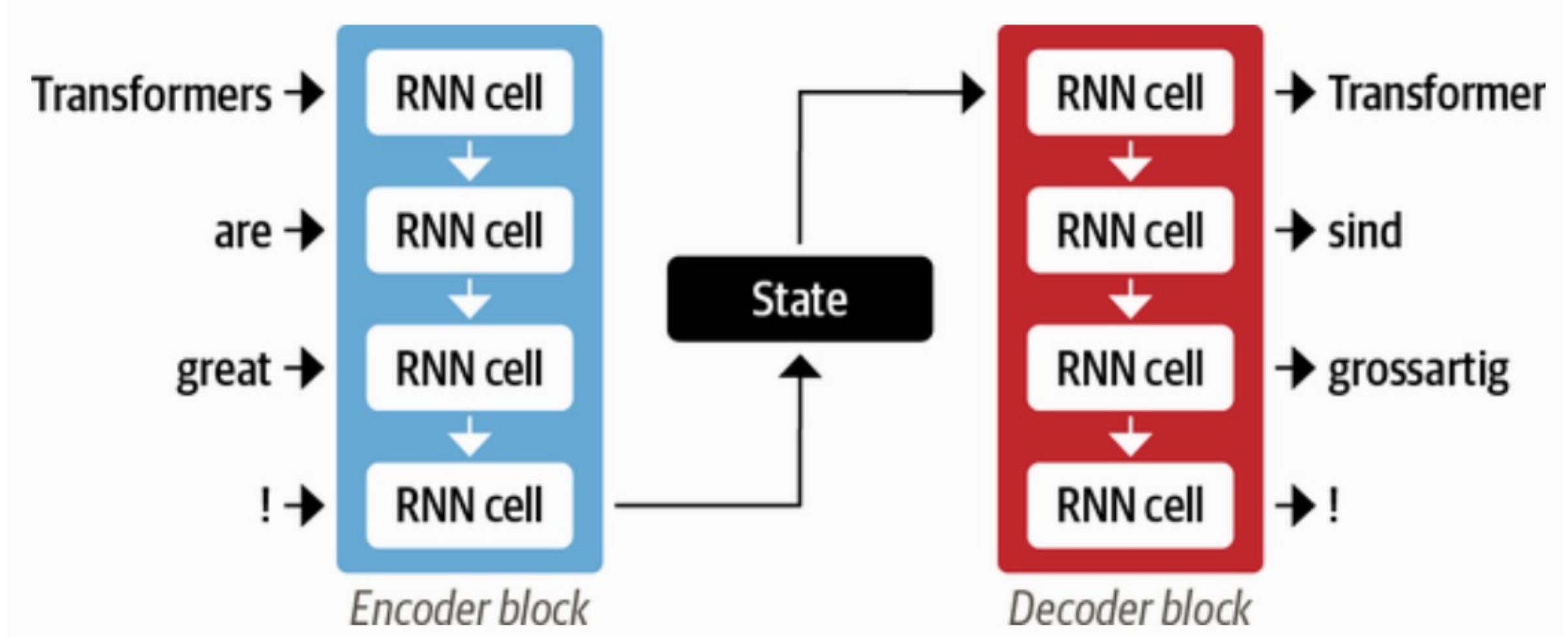
The Encoder-Decoder Framework

- **The encoder-decoder framework**
- **Attention Mechanisms**
- **Transfer Learning in NLP**

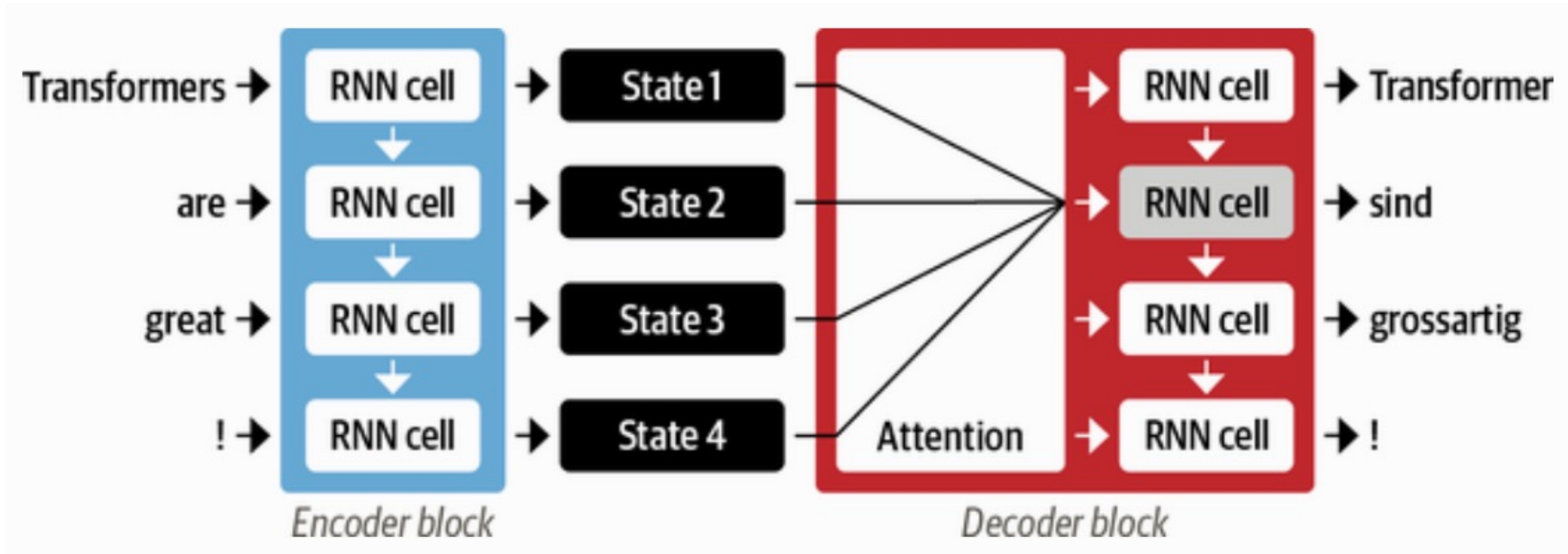
RNN



An encoder-decoder architecture with a pair of RNN



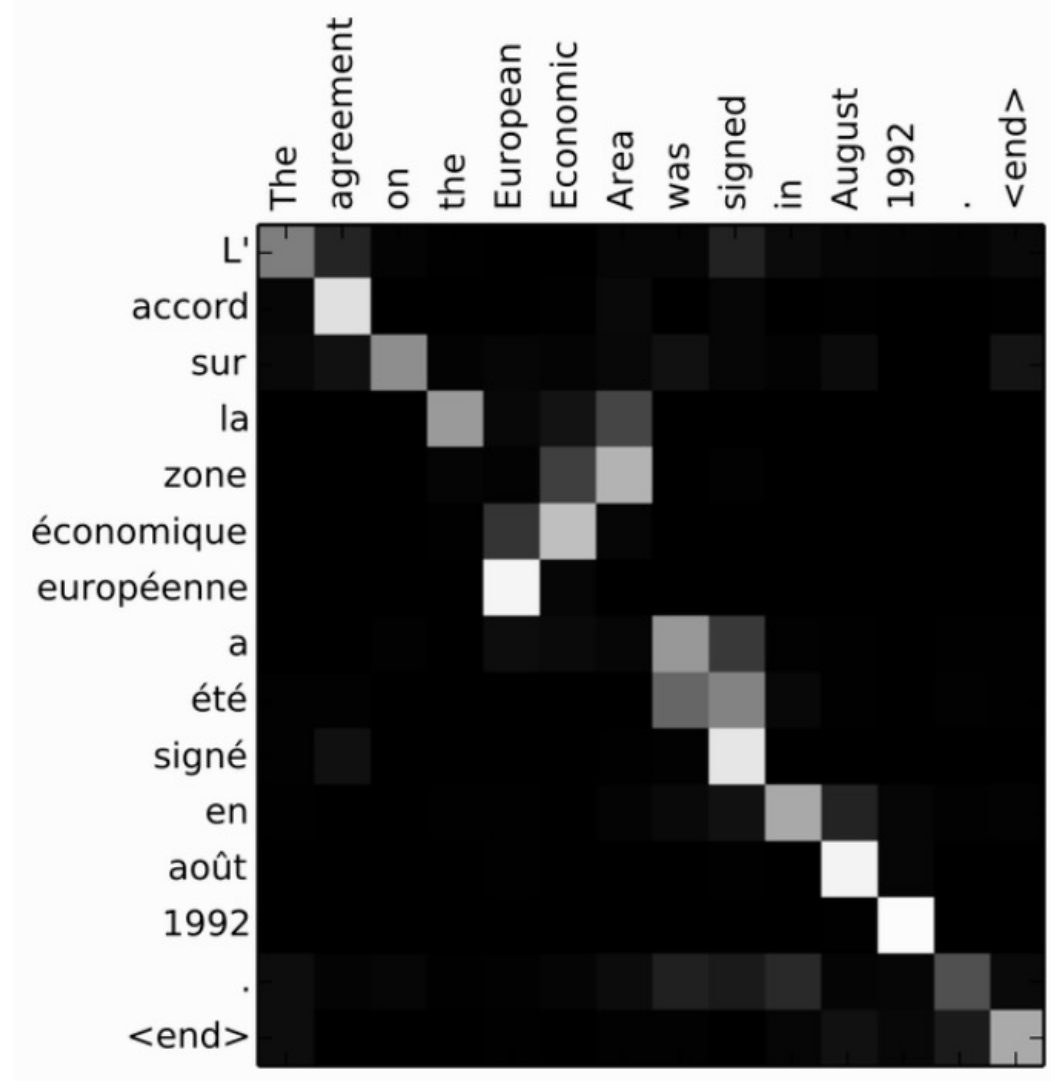
Attention Mechanisms



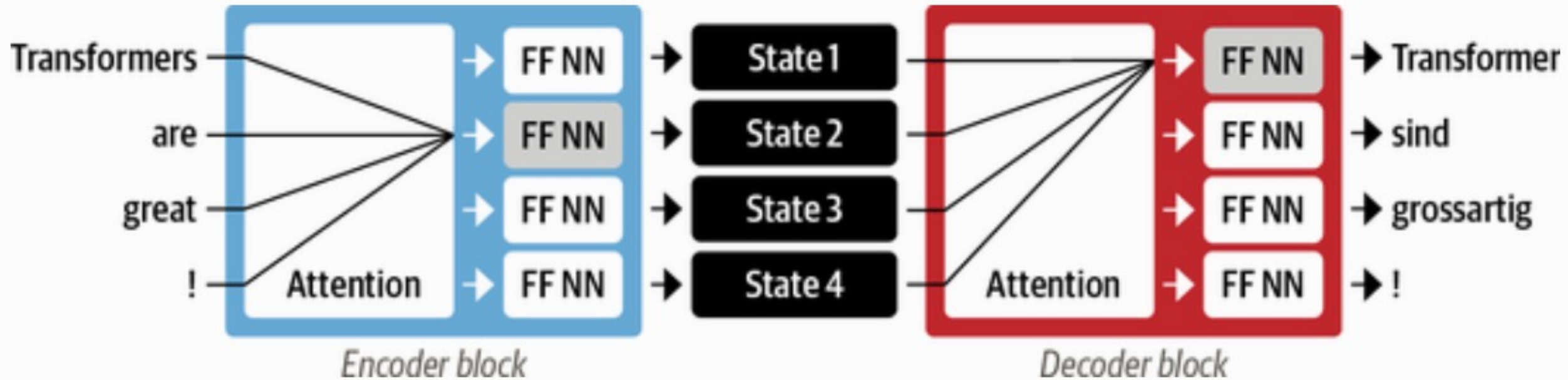
An encoder-decoder architecture with an attention mechanism

RNN Encoder-Decoder

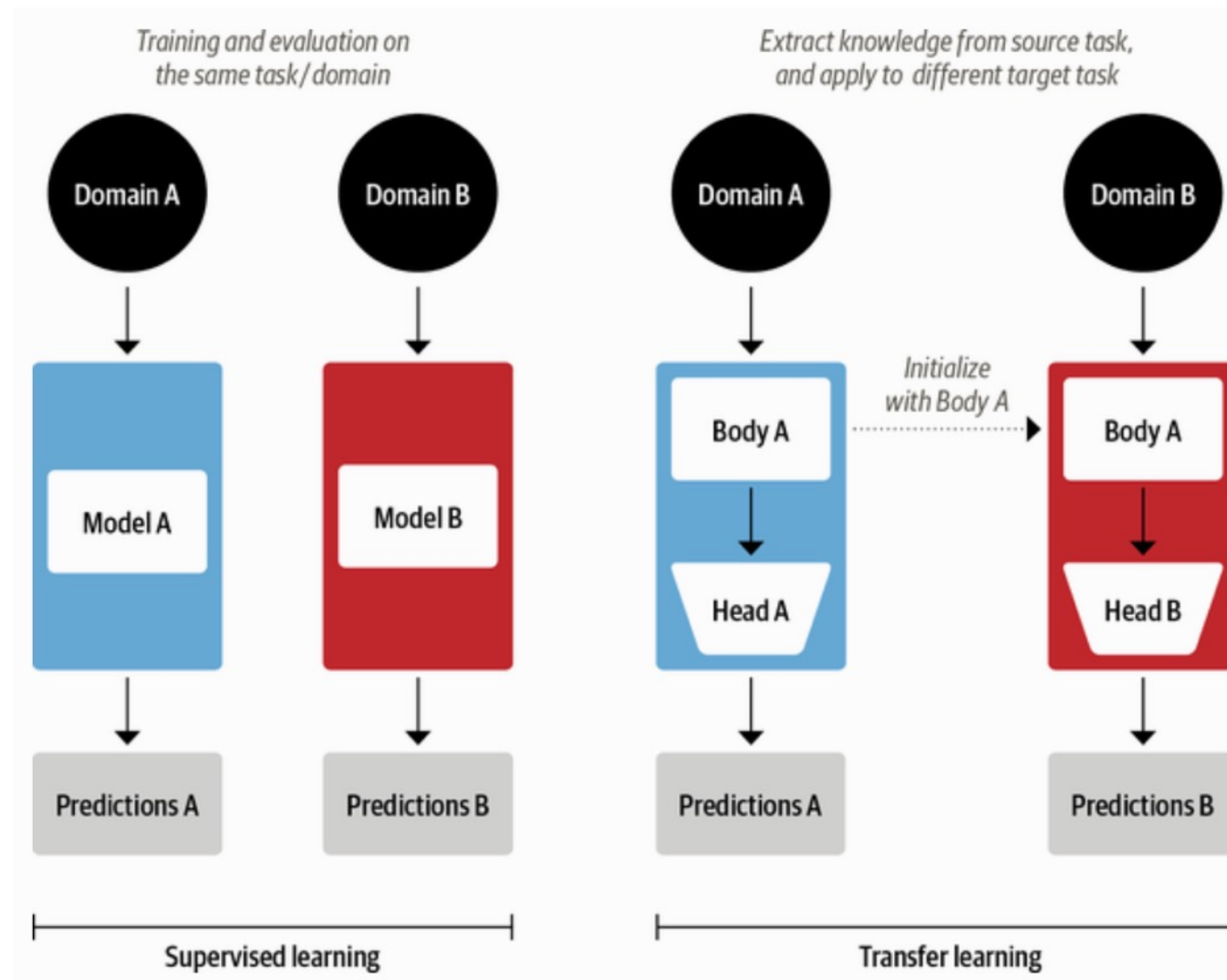
alignment of words in English and the generated translation in French



Encoder-Decoder Architecture of the Original Transformer

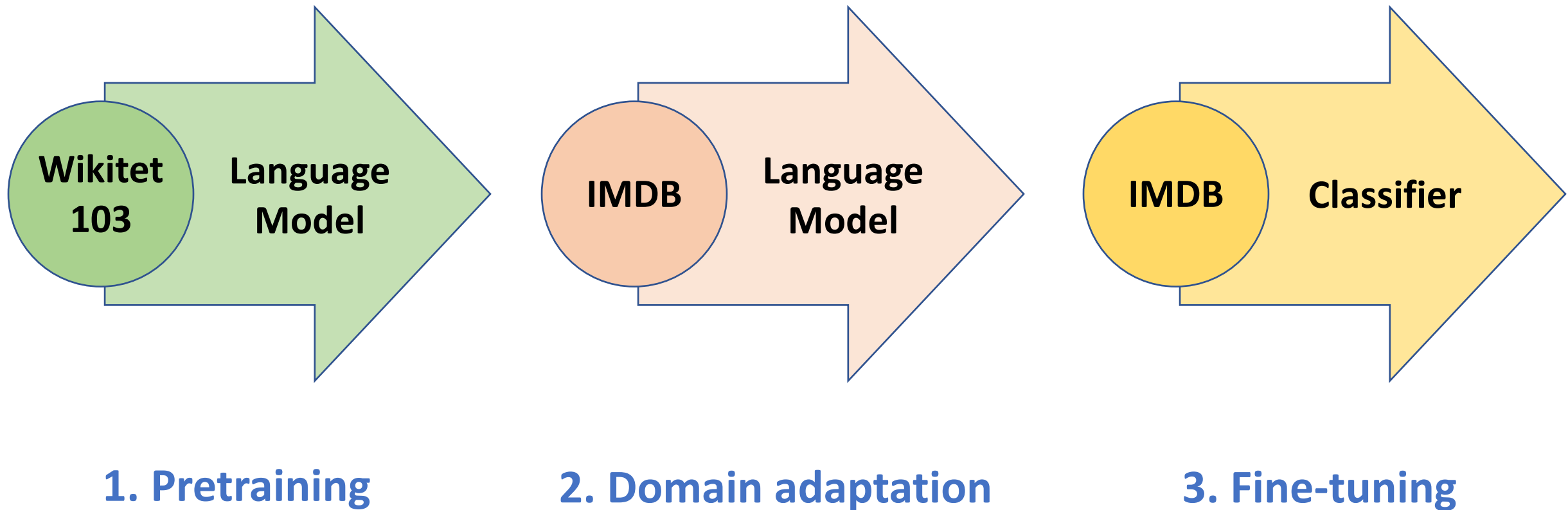


Comparison of Traditional Supervised Learning and Transfer Learning

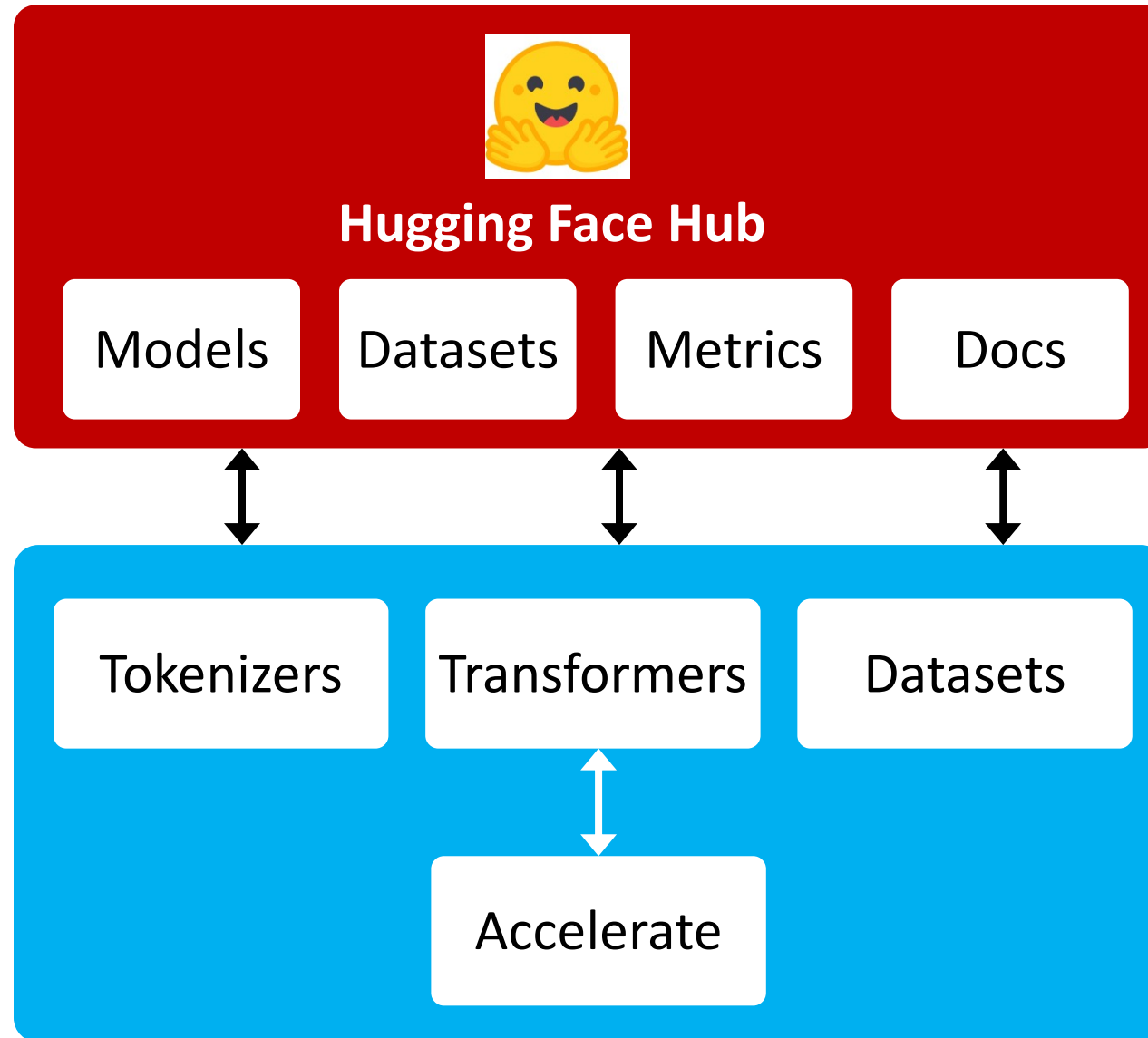


ULMFiT: 3 Steps

Transfer Learning in NLP

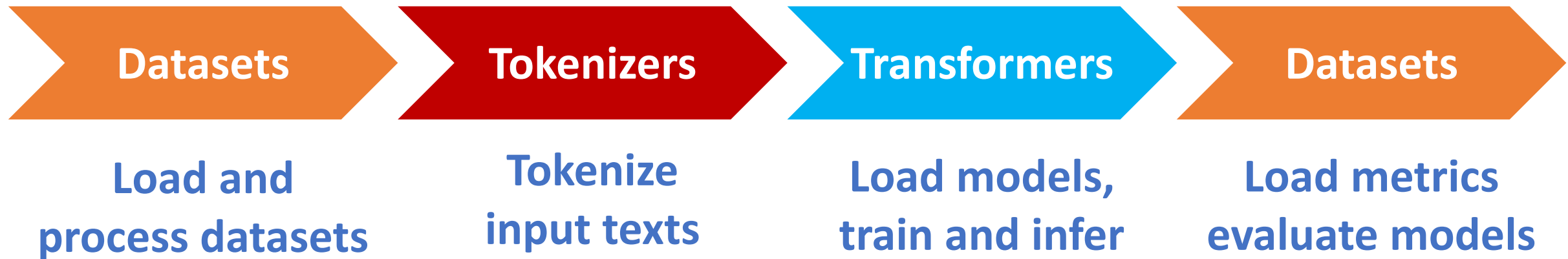


An overview of the Hugging Face Ecosystem



A typical pipeline for training transformer models

with the Datasets, Tokenizers, and Transformers libraries



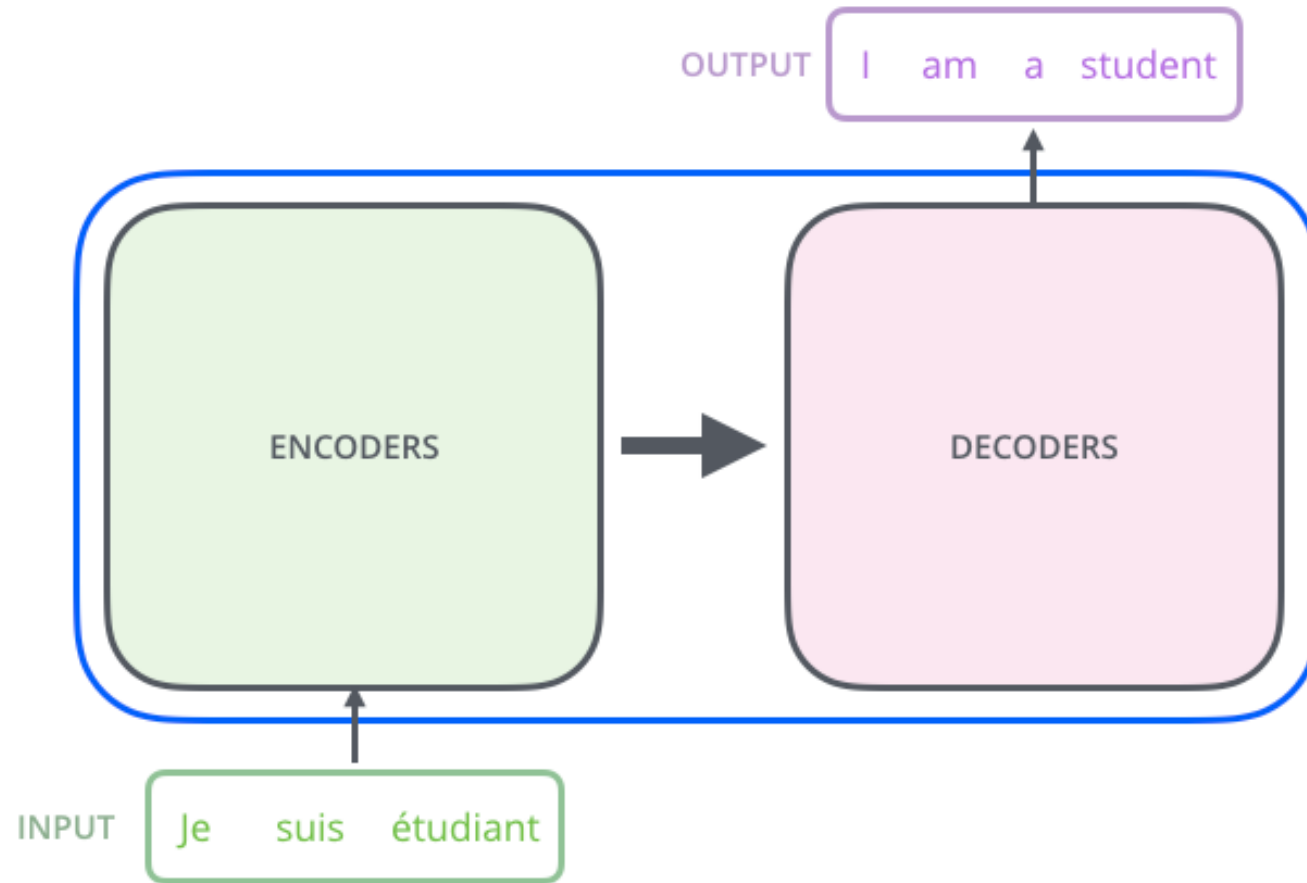
The Illustrated Transformer

Jay Alammar (2018)



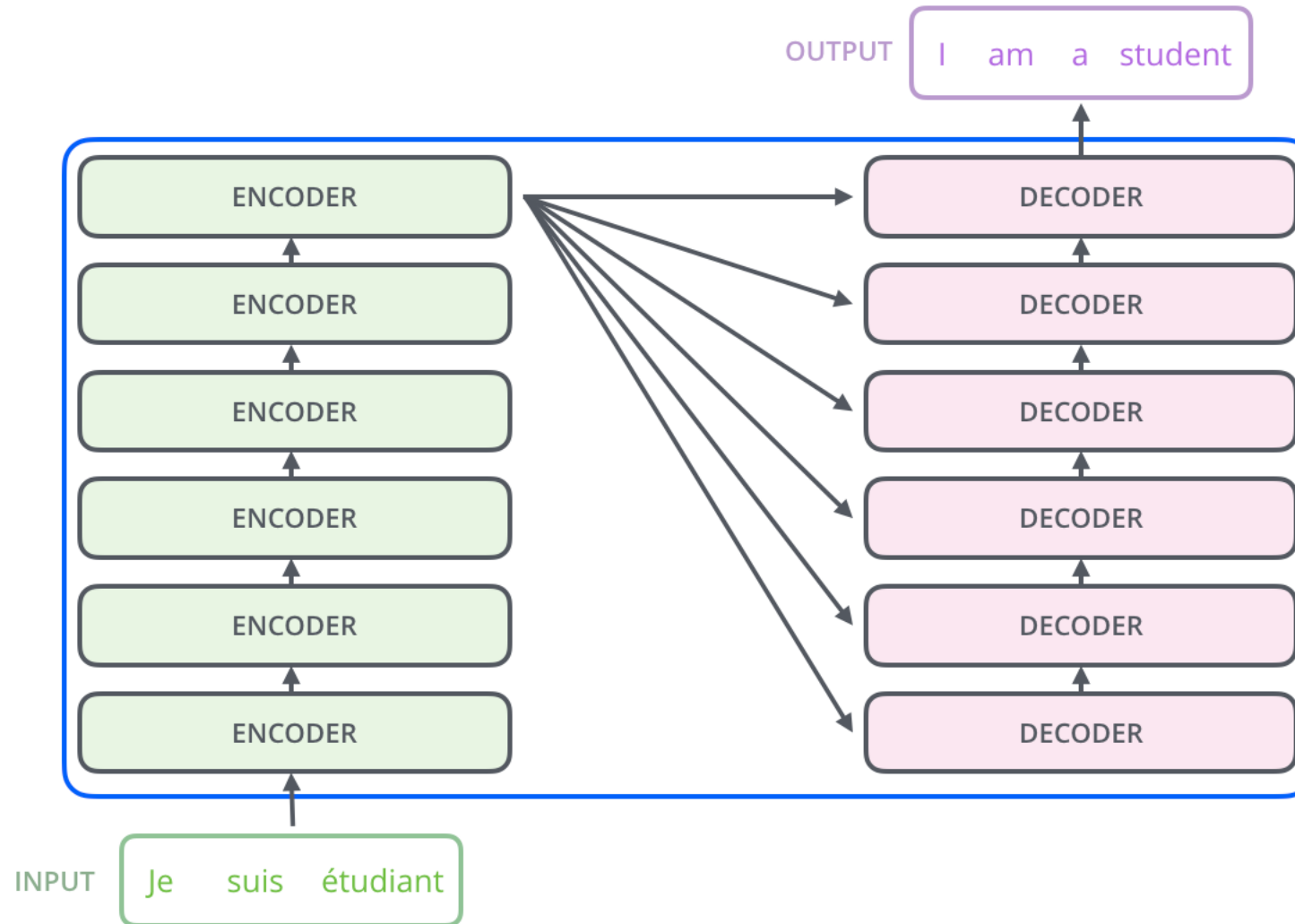
The Illustrated Transformer

Jay Alammar (2018)



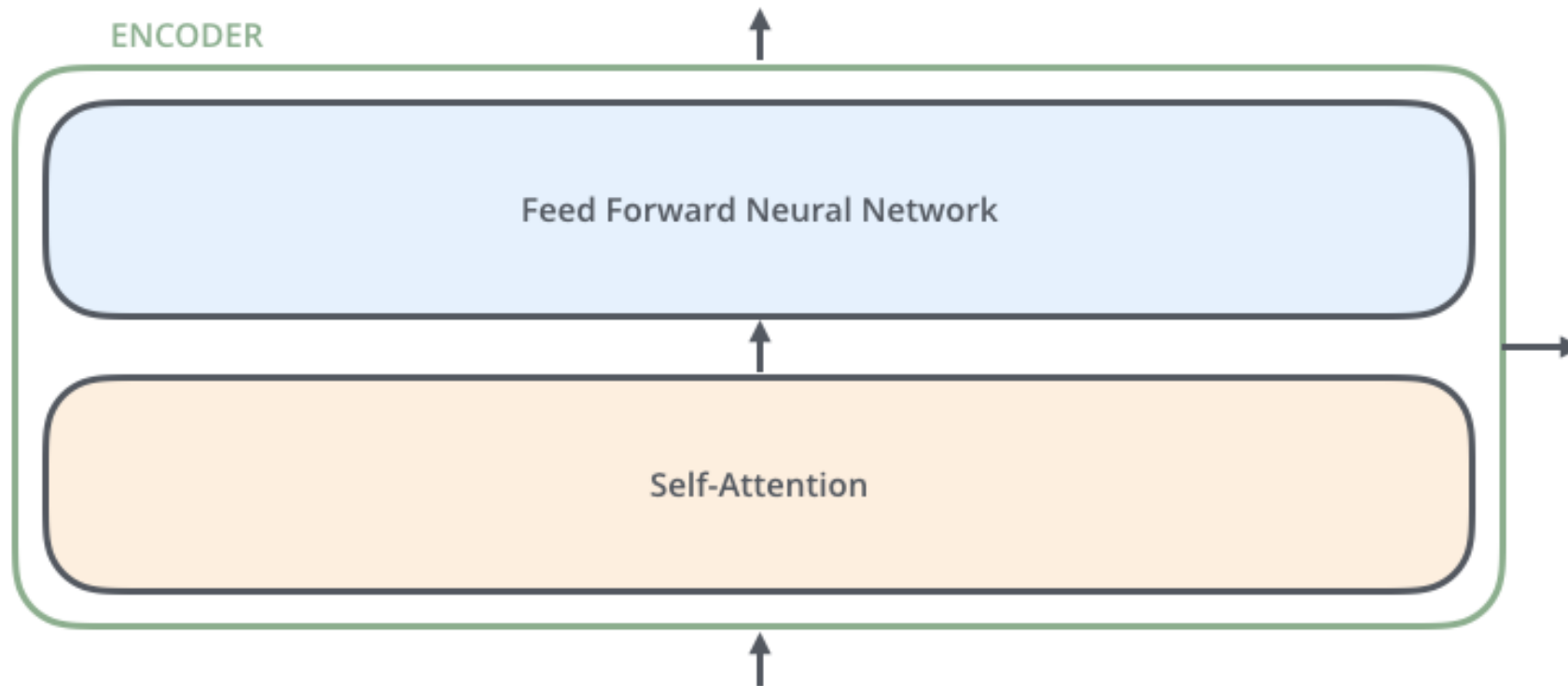
The Illustrated Transformer

Jay Alammar (2018)



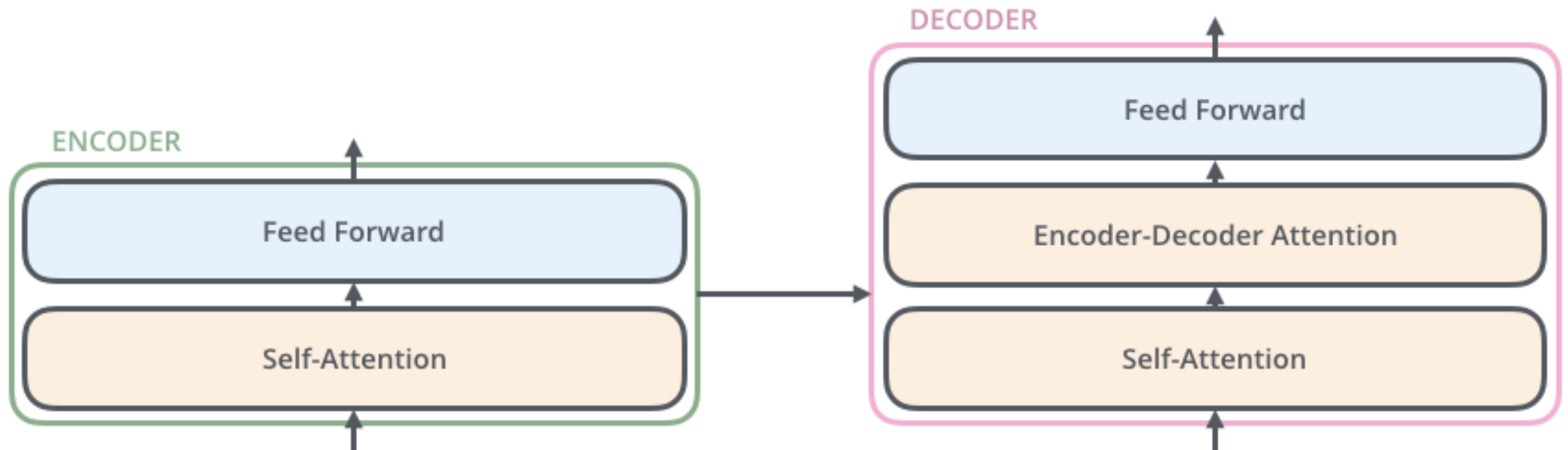
The Illustrated Transformer

Jay Alammar (2018)



The Illustrated Transformer

Jay Alammar (2018)



The Illustrated Transformer

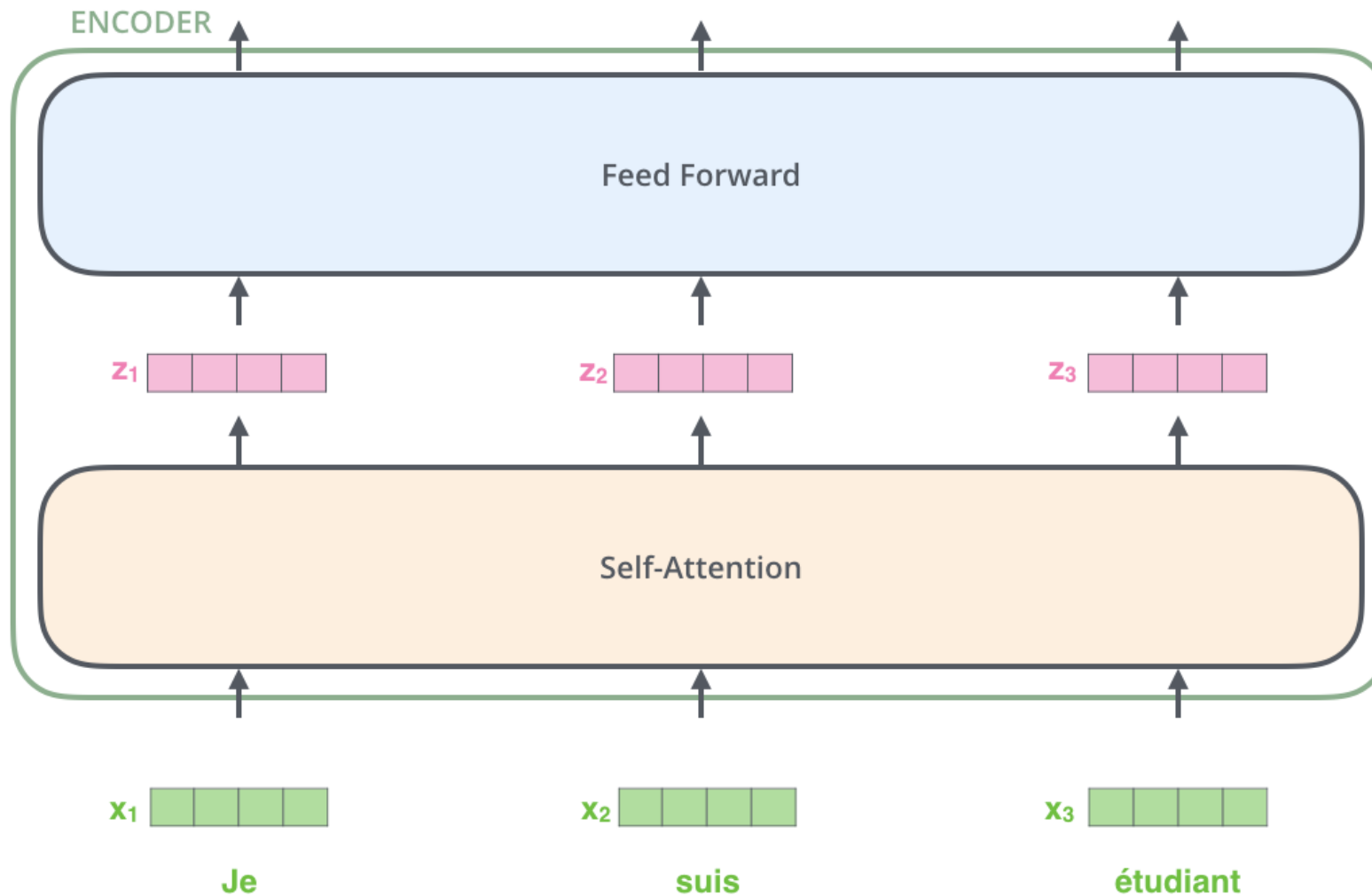
Jay Alammar (2018)



Each word is embedded into a vector of size 512.

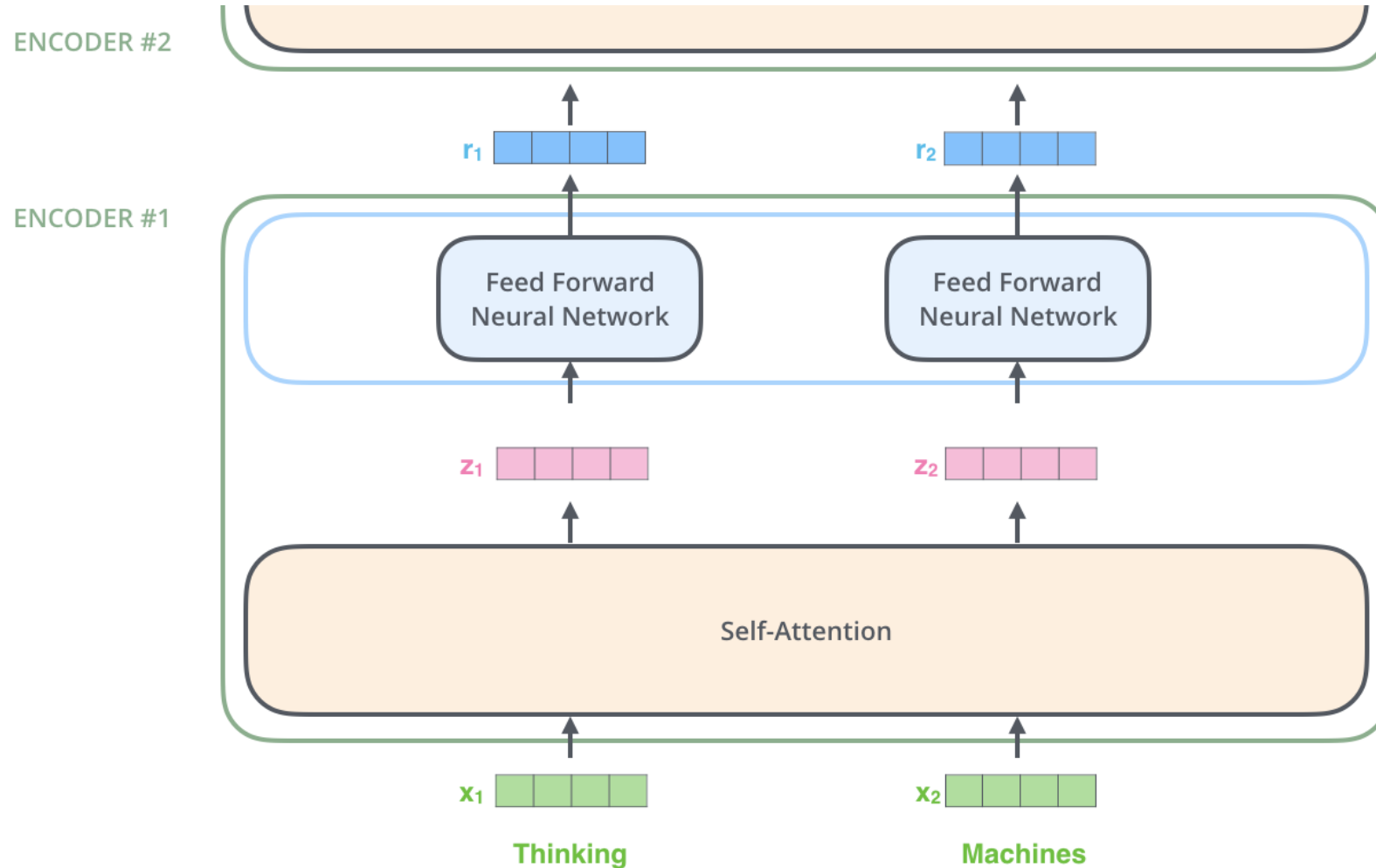
The Illustrated Transformer

Jay Alammar (2018)



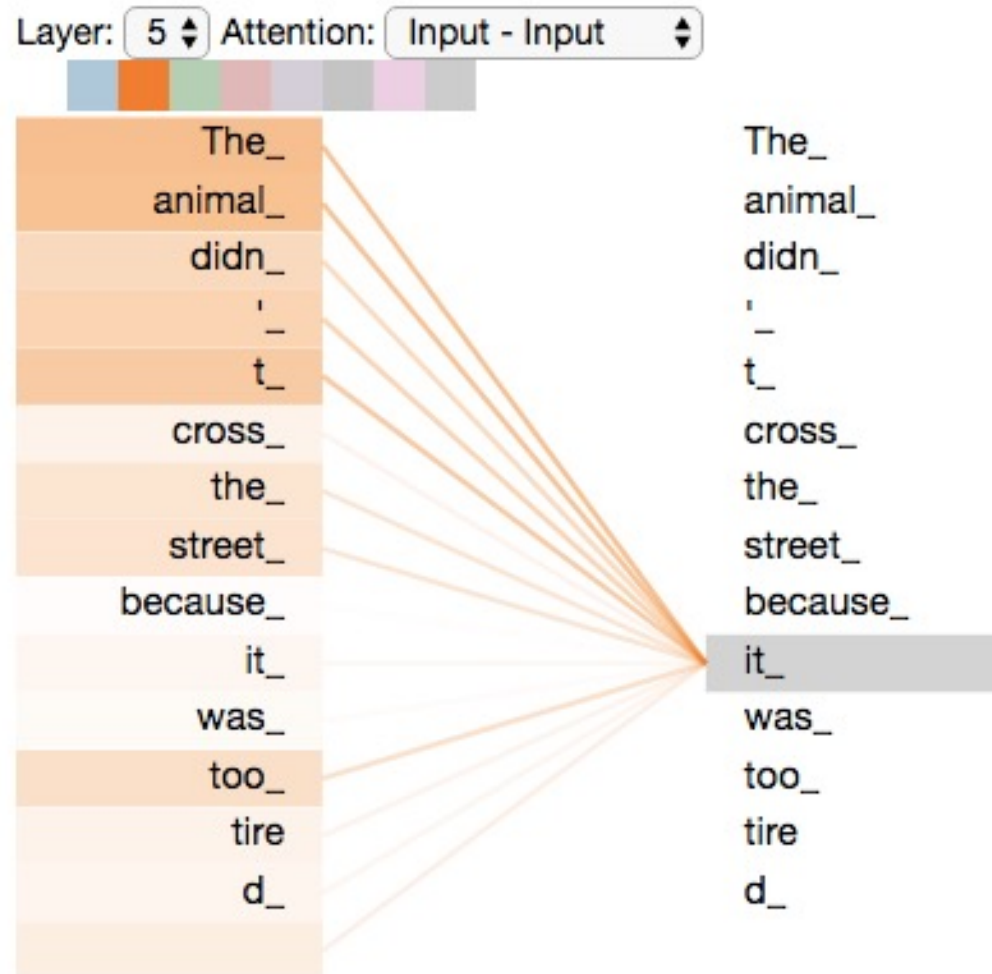
The Illustrated Transformer

Jay Alammar (2018)



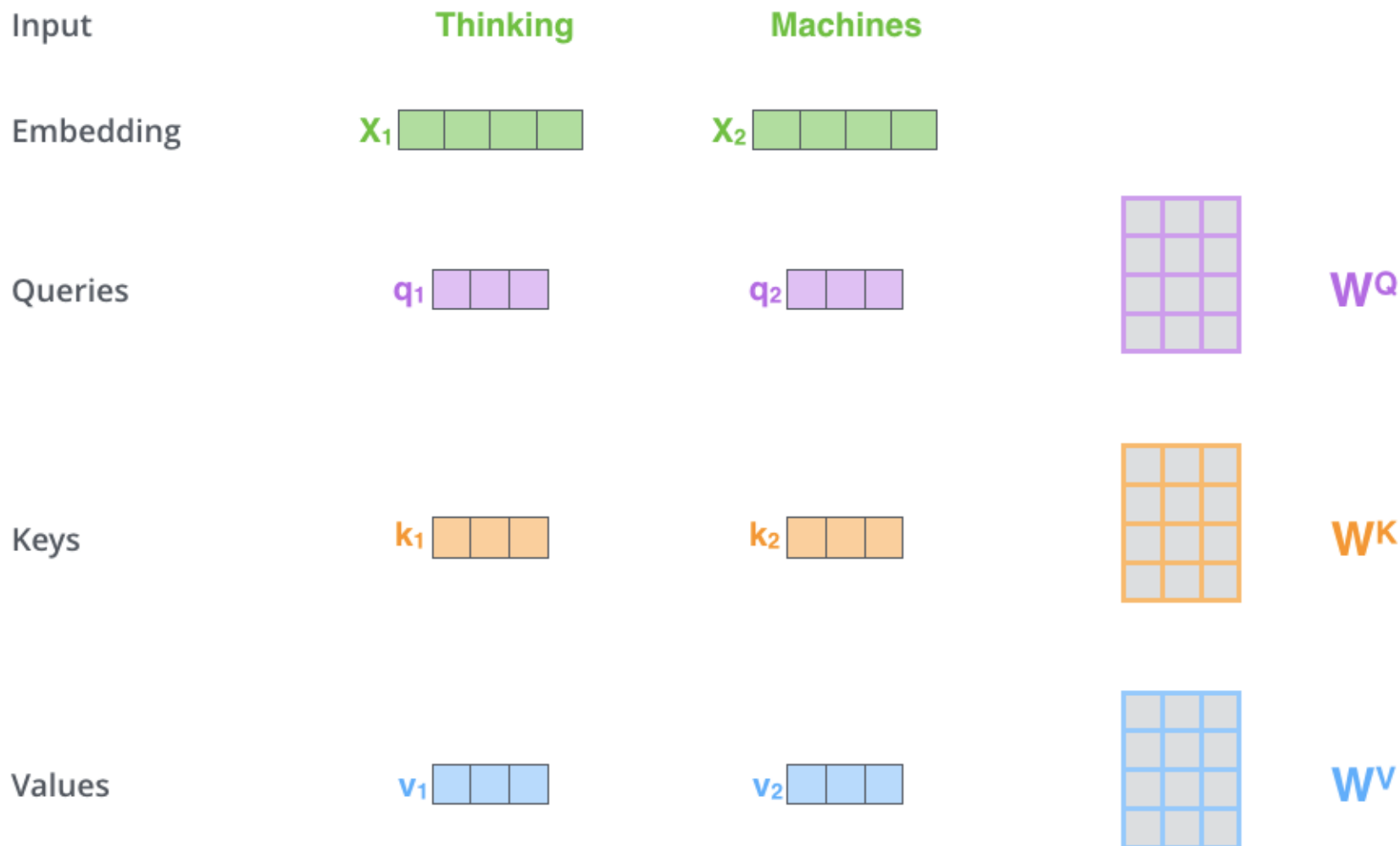
The Illustrated Transformer

Jay Alammar (2018)



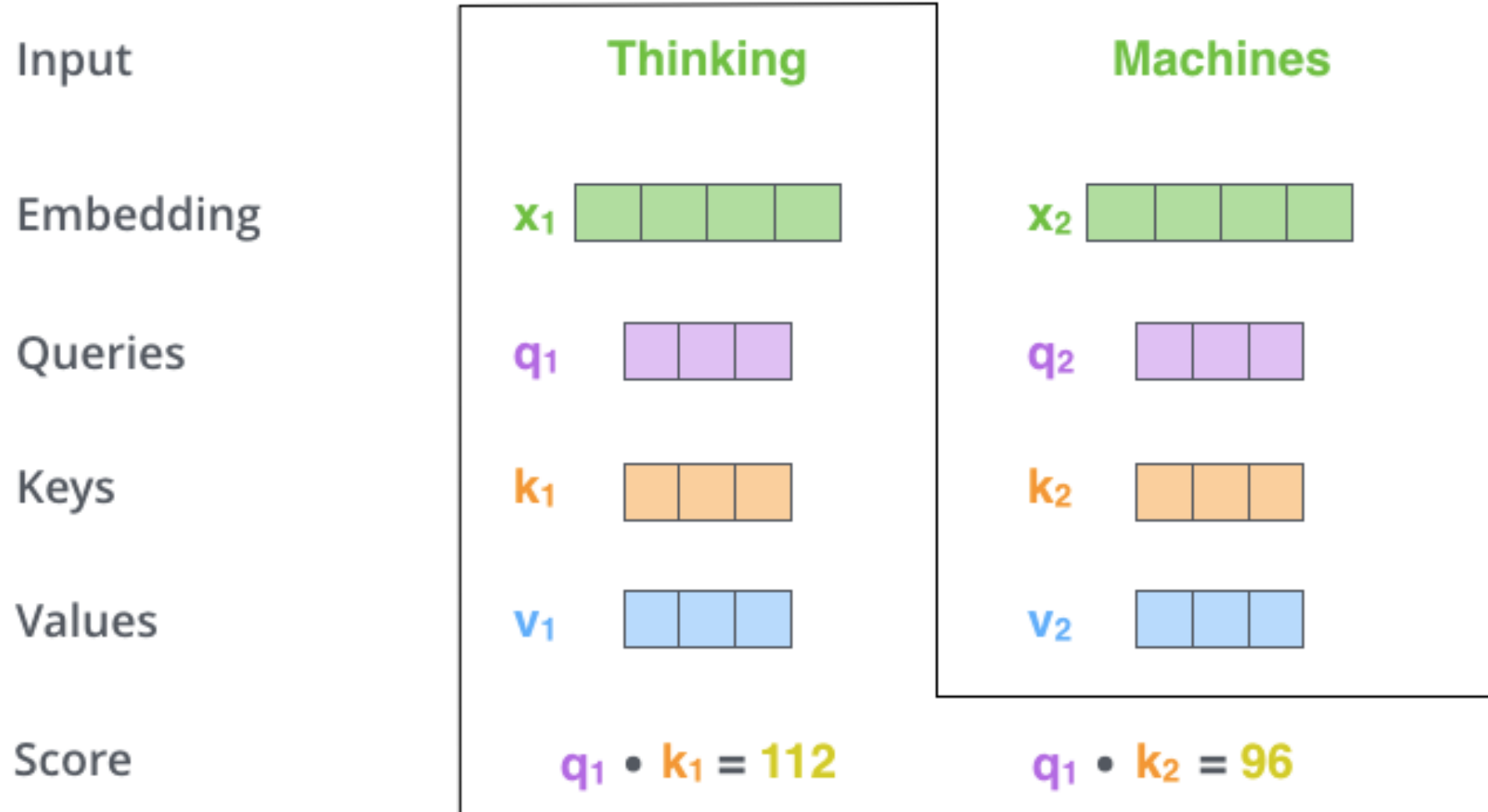
Multiplying x_1 by the W^Q weight matrix produces q_1 , the "query" vector associated with that word.

We end up creating a "query", a "key", and a "value" projection of each word in the input sentence.



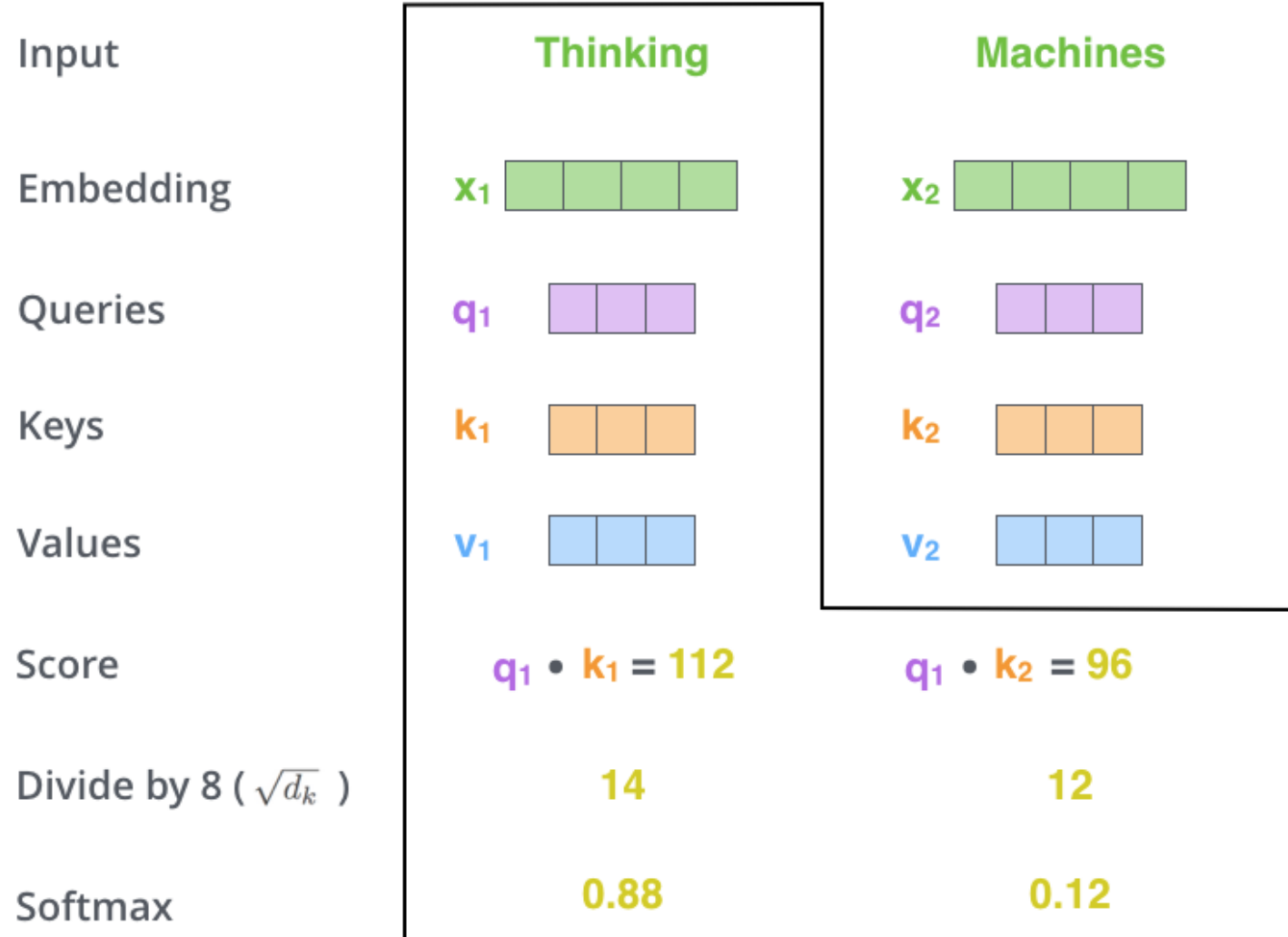
The Illustrated Transformer

Jay Alammar (2018)

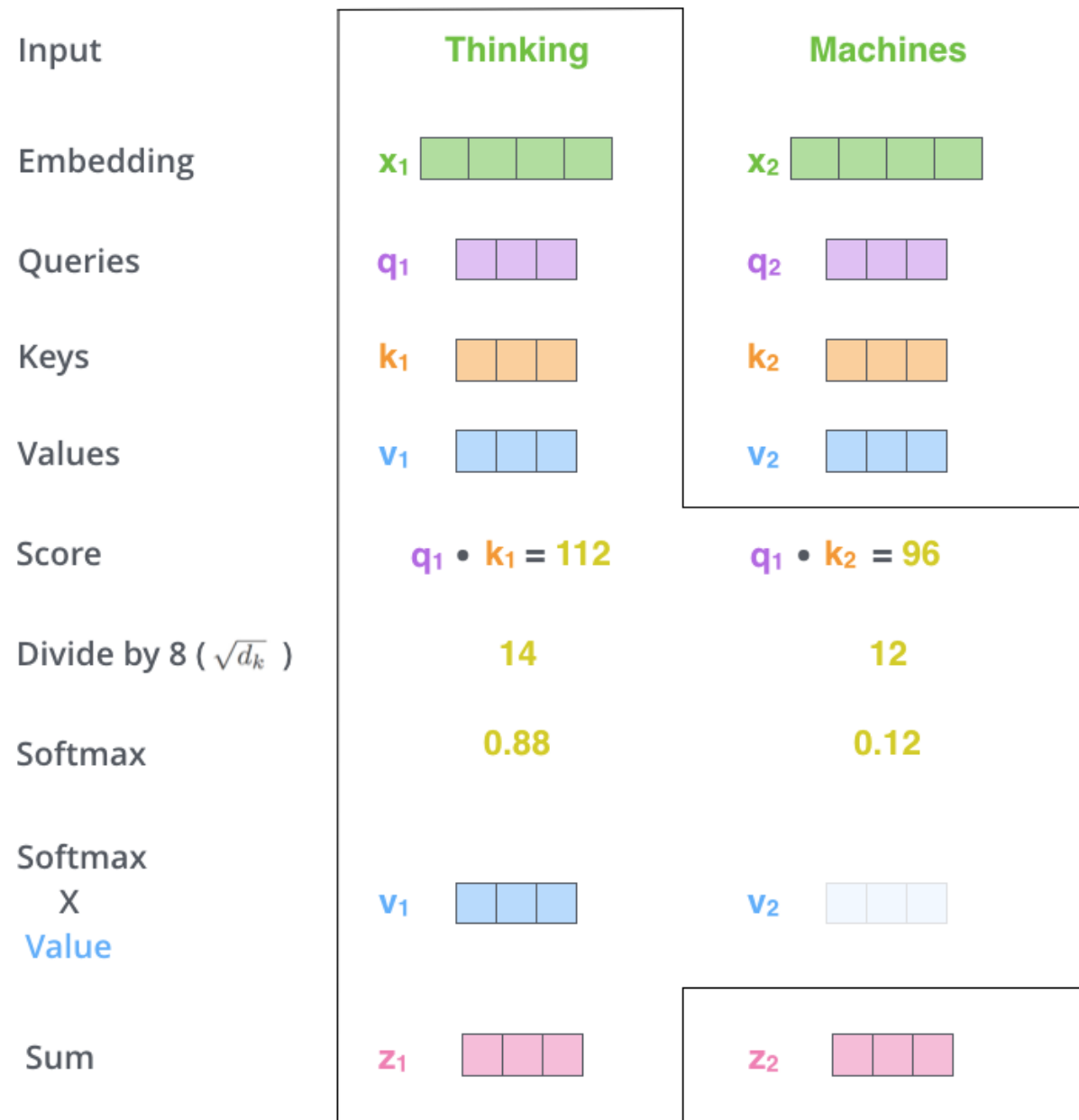


The Illustrated Transformer

Jay Alammar (2018)



Source: Jay Alammar (2018), The Illustrated Transformer,
<http://jalammar.github.io/illustrated-transformer/>



Matrix Calculation of Self-Attention

$$\begin{matrix} \text{X} \\ \begin{array}{|c|c|c|c|} \hline \square & \square & \square & \square \\ \hline \square & \square & \square & \square \\ \hline \end{array} \end{matrix} \times \begin{matrix} \text{W}^{\text{Q}} \\ \begin{array}{|c|c|c|} \hline \square & \square & \square \\ \hline \square & \square & \square \\ \hline \square & \square & \square \\ \hline \square & \square & \square \\ \hline \end{array} \end{matrix} = \begin{matrix} \text{Q} \\ \begin{array}{|c|c|c|} \hline \square & \square & \square \\ \hline \square & \square & \square \\ \hline \end{array} \end{matrix}$$

$$\begin{matrix} \text{X} \\ \begin{array}{|c|c|c|c|} \hline \square & \square & \square & \square \\ \hline \square & \square & \square & \square \\ \hline \end{array} \end{matrix} \times \begin{matrix} \text{W}^{\text{K}} \\ \begin{array}{|c|c|c|} \hline \square & \square & \square \\ \hline \square & \square & \square \\ \hline \square & \square & \square \\ \hline \square & \square & \square \\ \hline \end{array} \end{matrix} = \begin{matrix} \text{K} \\ \begin{array}{|c|c|c|} \hline \square & \square & \square \\ \hline \square & \square & \square \\ \hline \end{array} \end{matrix}$$

$$\begin{matrix} \text{X} \\ \begin{array}{|c|c|c|c|} \hline \square & \square & \square & \square \\ \hline \square & \square & \square & \square \\ \hline \end{array} \end{matrix} \times \begin{matrix} \text{W}^{\text{V}} \\ \begin{array}{|c|c|c|} \hline \square & \square & \square \\ \hline \square & \square & \square \\ \hline \square & \square & \square \\ \hline \square & \square & \square \\ \hline \end{array} \end{matrix} = \begin{matrix} \text{V} \\ \begin{array}{|c|c|c|} \hline \square & \square & \square \\ \hline \square & \square & \square \\ \hline \end{array} \end{matrix}$$

The self-attention calculation in matrix form

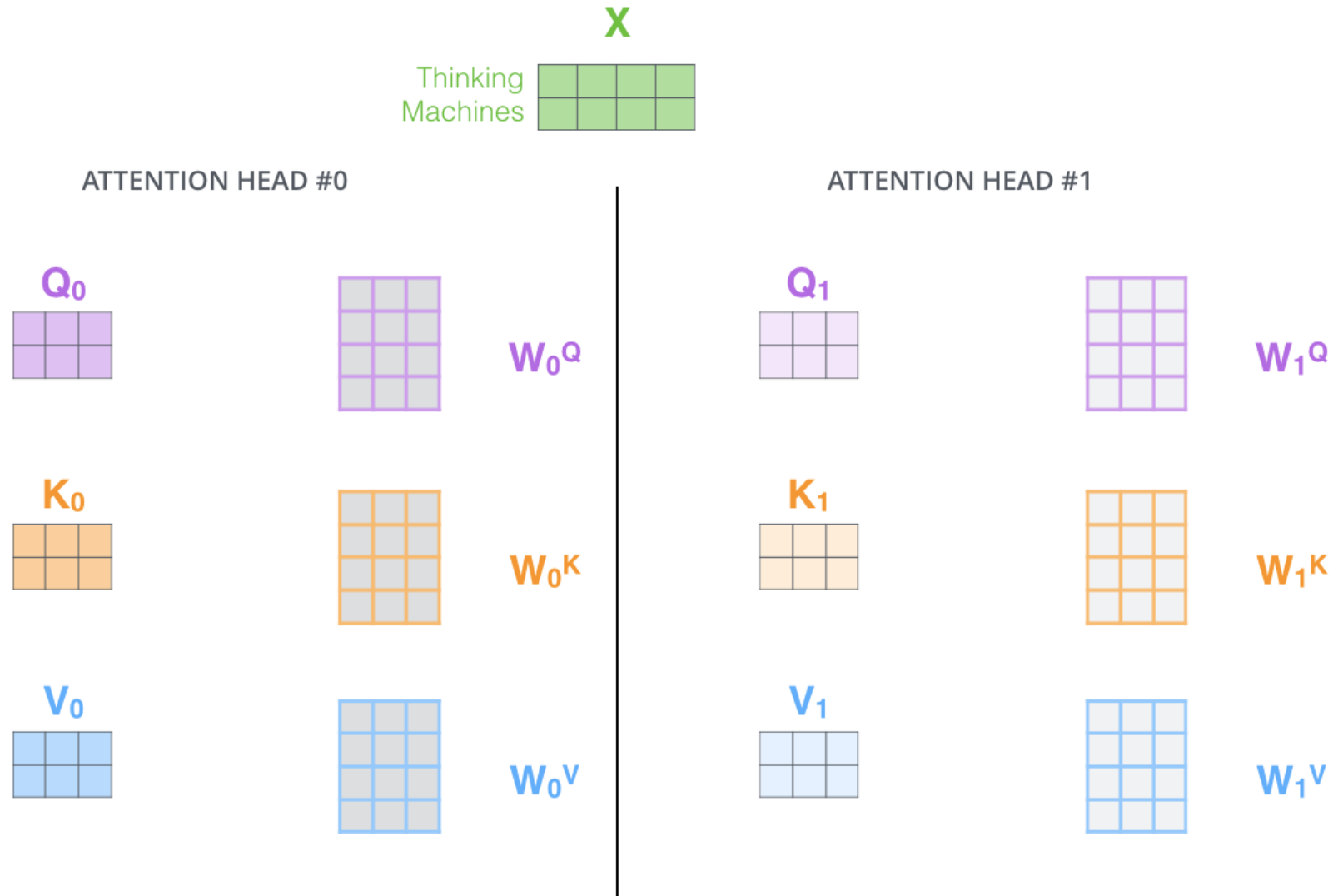
$$\text{softmax} \left(\frac{\begin{matrix} \text{Q} \\ \begin{array}{|c|c|c|} \hline \square & \square & \square \\ \hline \square & \square & \square \\ \hline \end{array} \end{matrix} \times \begin{matrix} \text{K}^T \\ \begin{array}{|c|c|} \hline \square & \square \\ \hline \square & \square \\ \hline \square & \square \\ \hline \end{array} \end{matrix} \right) \begin{matrix} \text{V} \\ \begin{array}{|c|c|c|} \hline \square & \square & \square \\ \hline \square & \square & \square \\ \hline \end{array} \end{matrix}$$

=

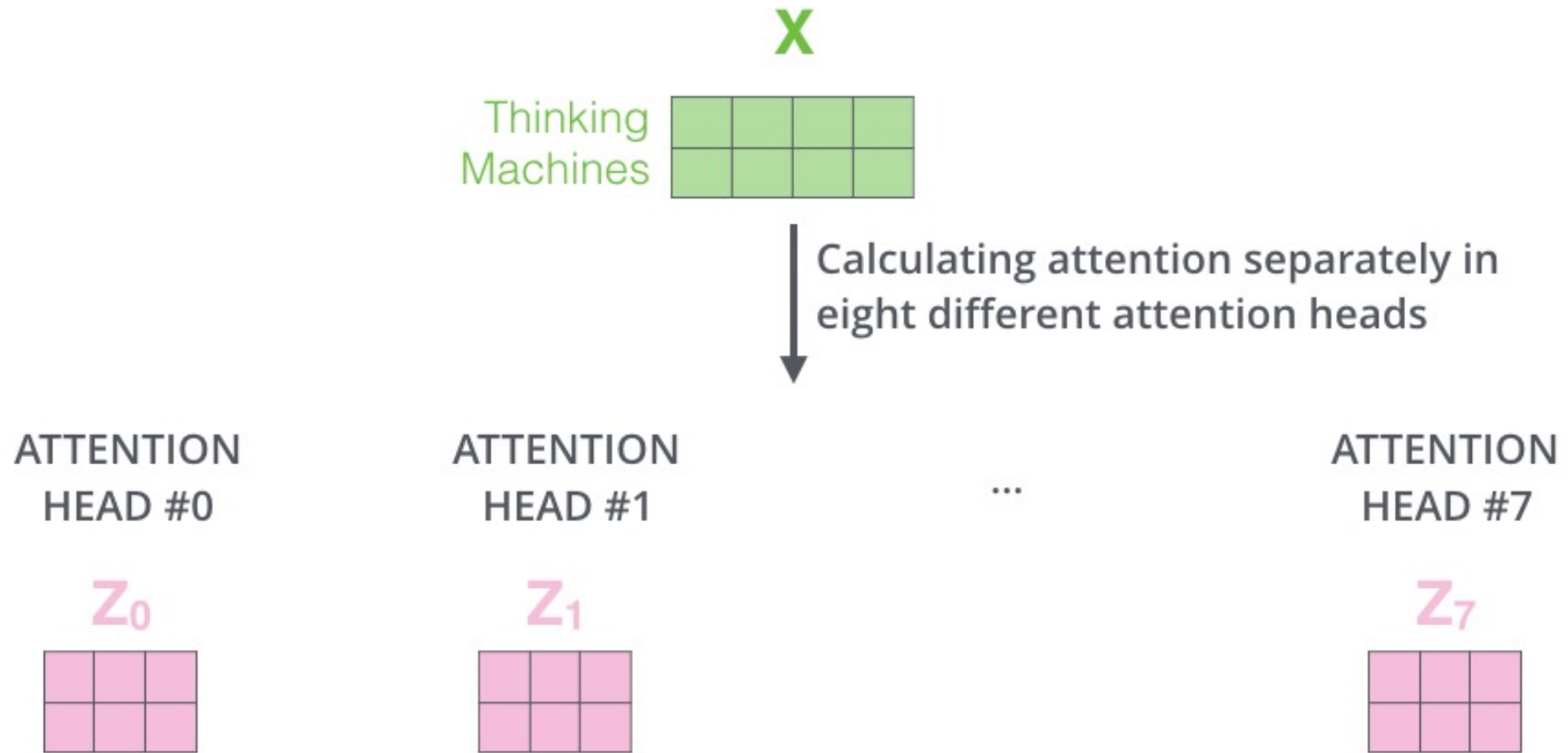
Z

$\begin{array}{|c|c|c|} \hline \square & \square & \square \\ \hline \square & \square & \square \\ \hline \end{array}$

Multi-headed Attention

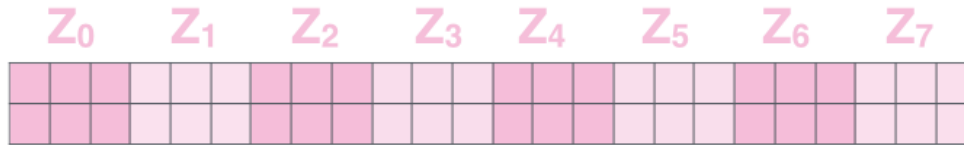


Multi-headed Attention



Multi-headed Attention

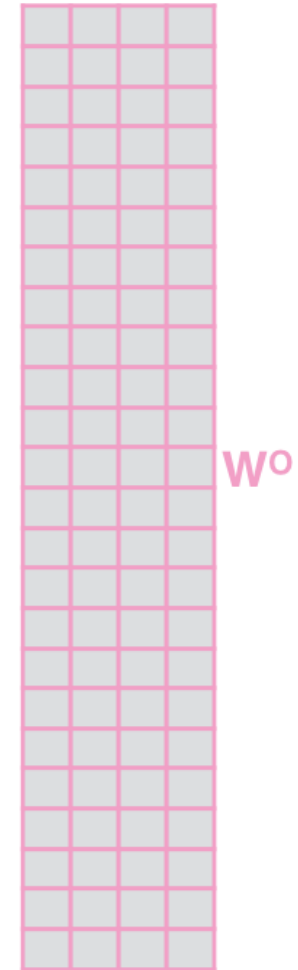
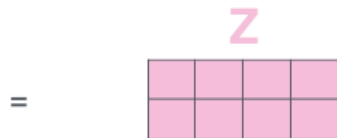
1) Concatenate all the attention heads



2) Multiply with a weight matrix W^O that was trained jointly with the model

X

3) The result would be the Z matrix that captures information from all the attention heads. We can send this forward to the FFNN

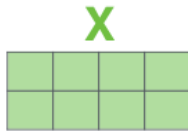


Multi-headed Attention

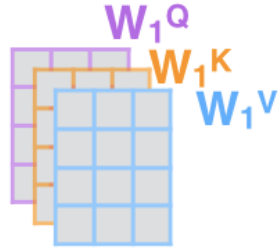
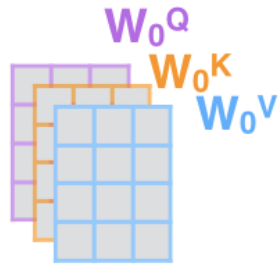
1) This is our input sentence*

Thinking
Machines

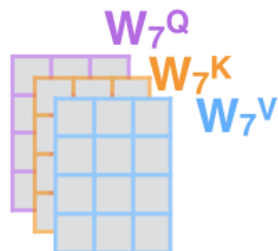
2) We embed each word*



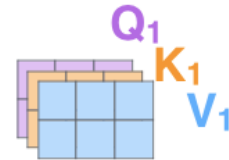
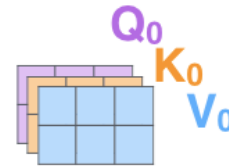
3) Split into 8 heads.
We multiply X or R with weight matrices



...



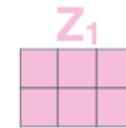
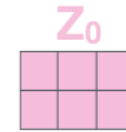
4) Calculate attention using the resulting $Q/K/V$ matrices



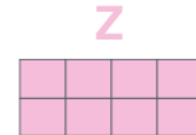
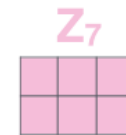
...



5) Concatenate the resulting Z matrices, then multiply with weight matrix W^O to produce the output of the layer



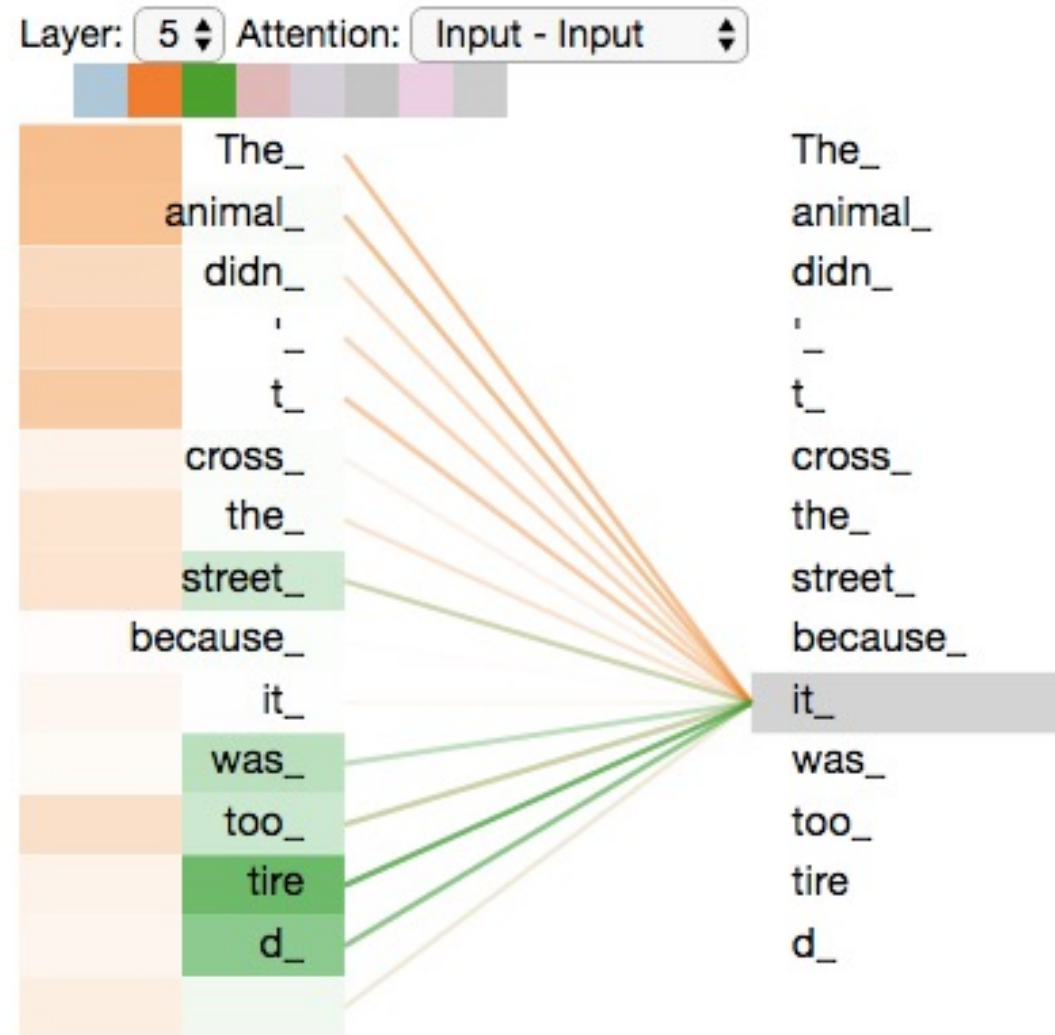
...



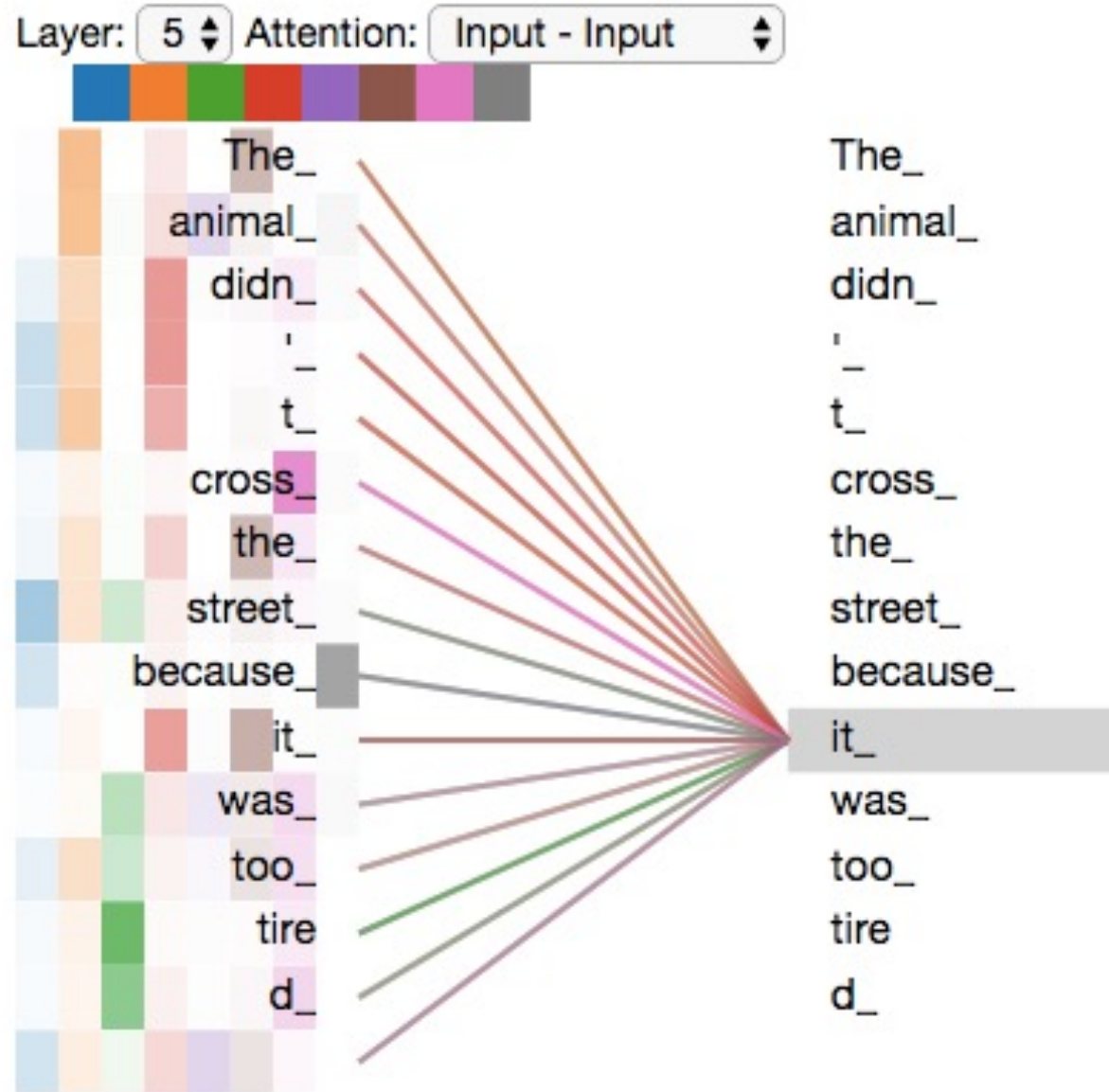
* In all encoders other than #0, we don't need embedding. We start directly with the output of the encoder right below this one



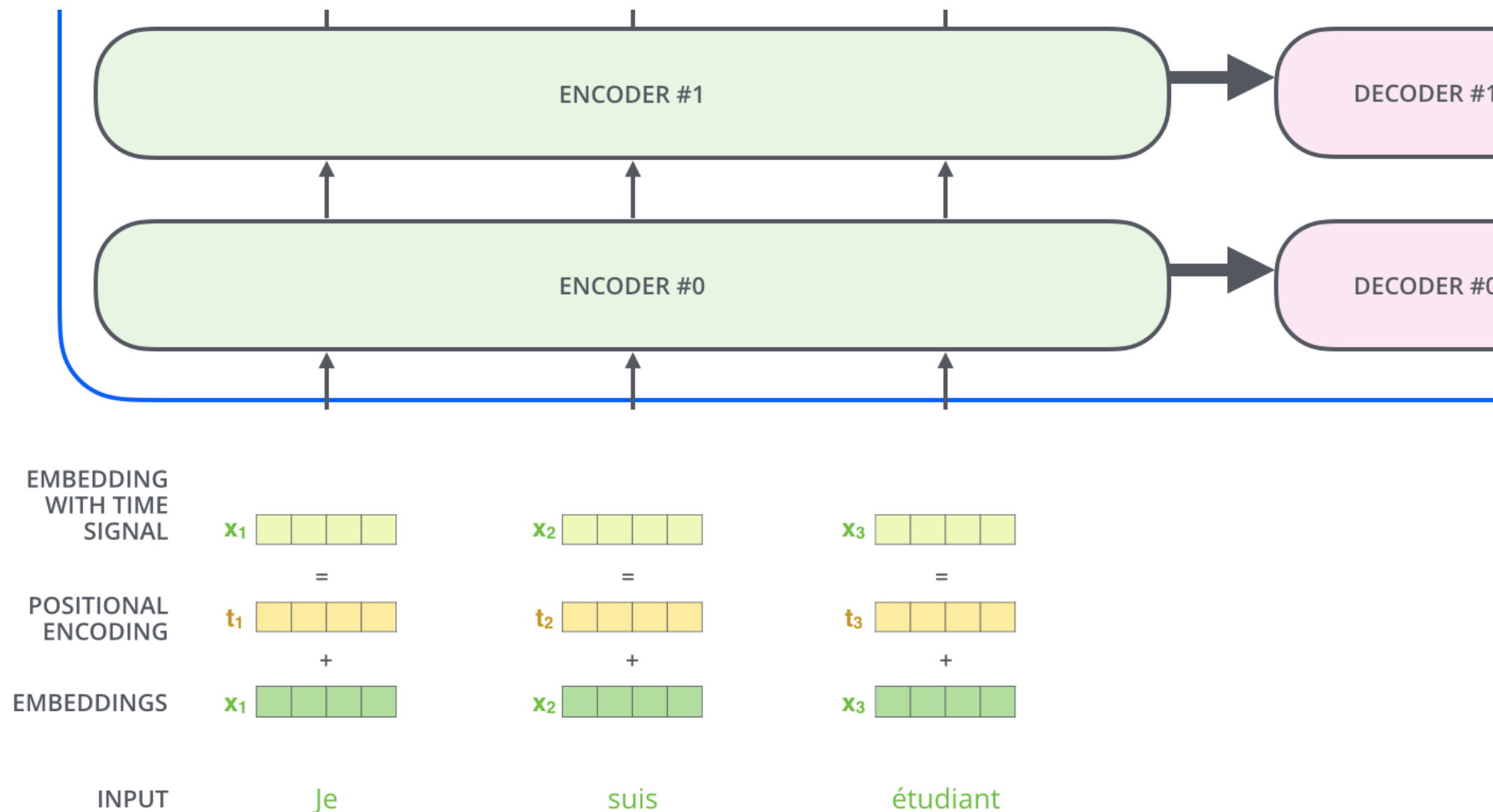
As we encode the word "it", one attention head is focusing most on "the animal", while another is focusing on "tired" -- in a sense, the model's representation of the word "it" bakes in some of the representation of both "animal" and "tired".



Add all the attention heads



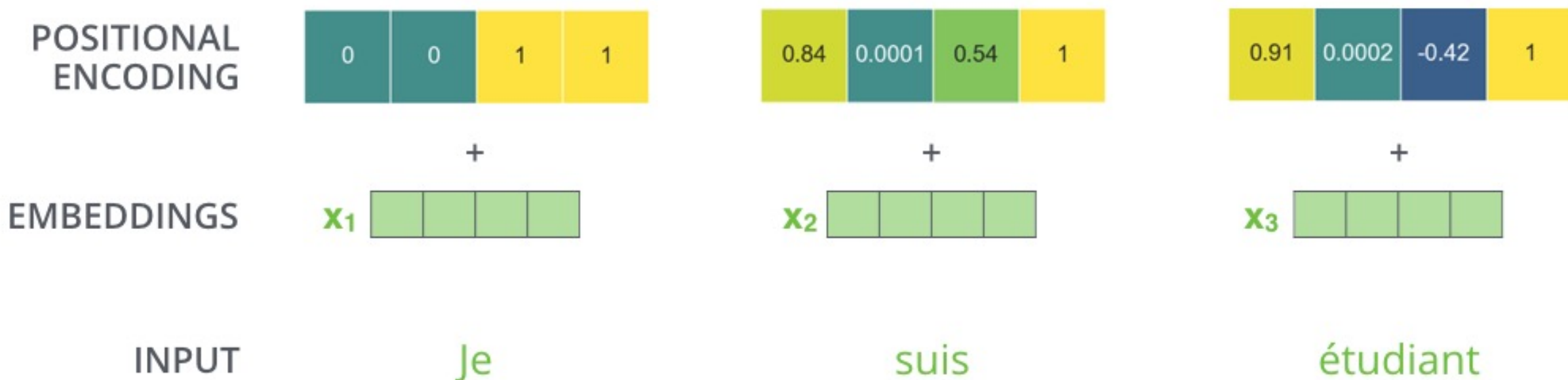
Positional Encoding



To give the model a sense of the order of the words, we add positional encoding vectors -- the values of which follow a specific pattern.

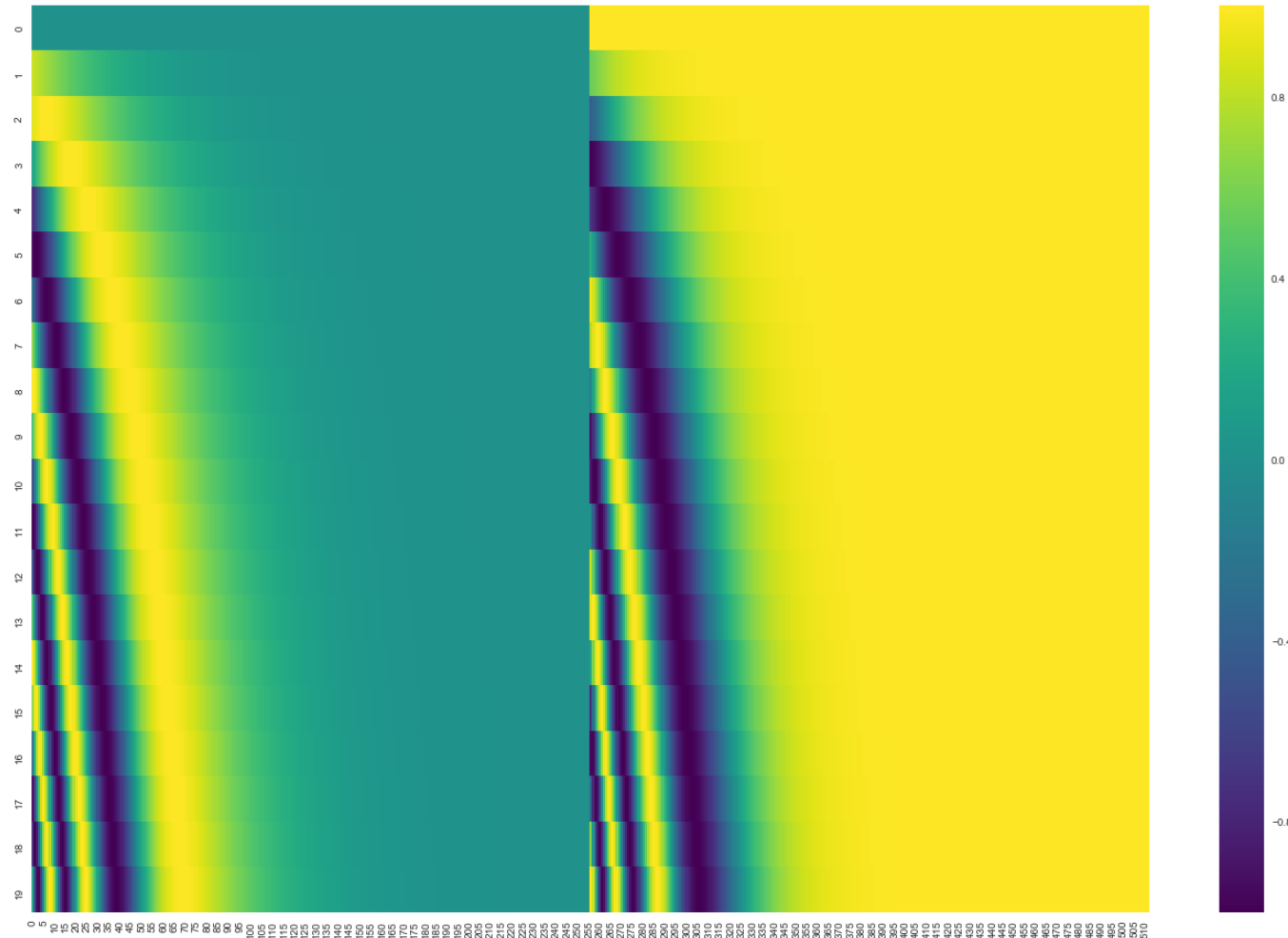
Source: Jay Alammar (2018), The Illustrated Transformer,
<http://jalammar.github.io/illustrated-transformer/>

Positional Encoding



Positional encoding with a toy embedding size of 4

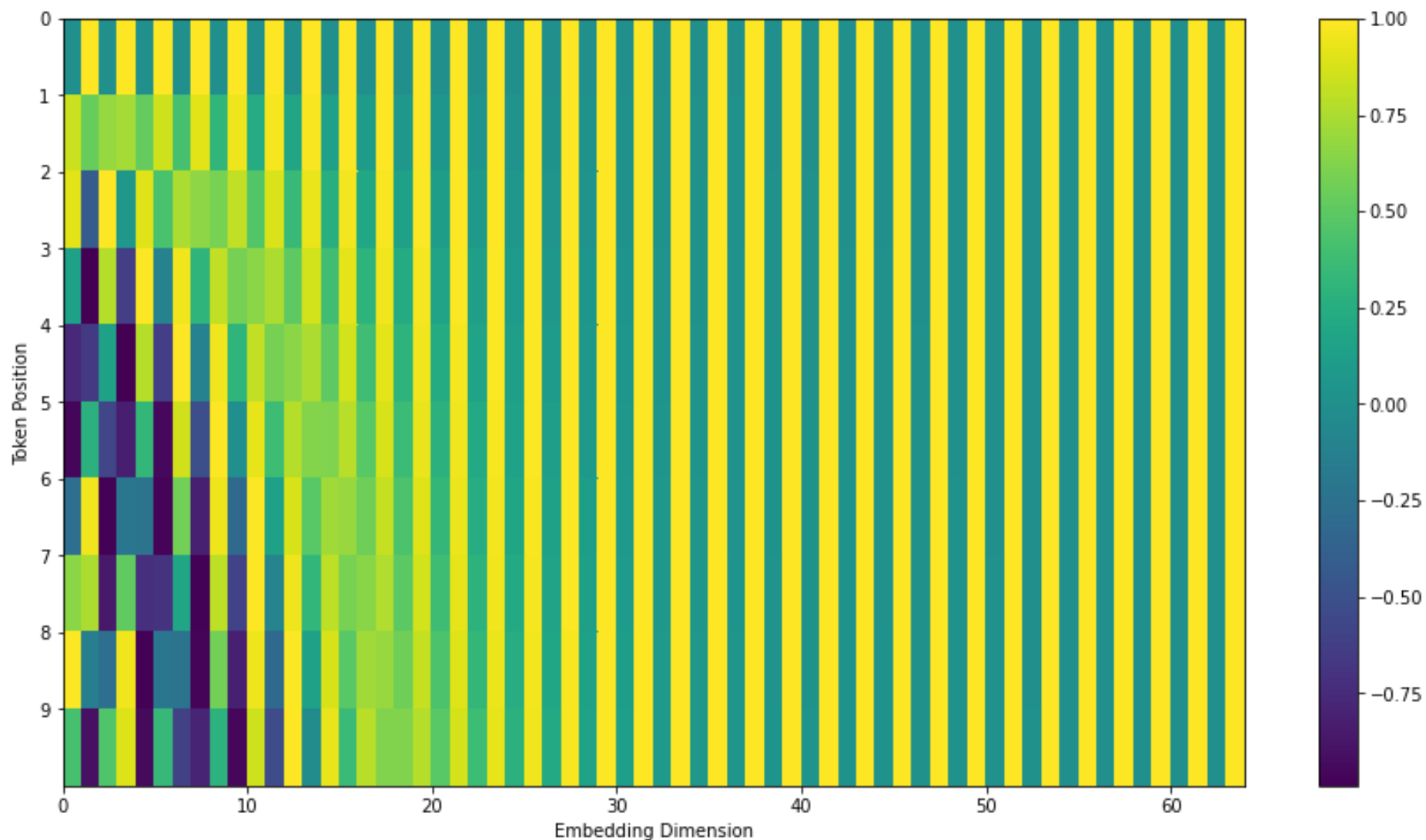
Positional encoding for 20 words (rows) with an embedding size of 512 (columns)



You can see that it appears split in half down the center. That's because the values of the left half are generated by one function (which uses sine), and the right half is generated by another function (which uses cosine). They're then concatenated to form each of the positional encoding vectors.

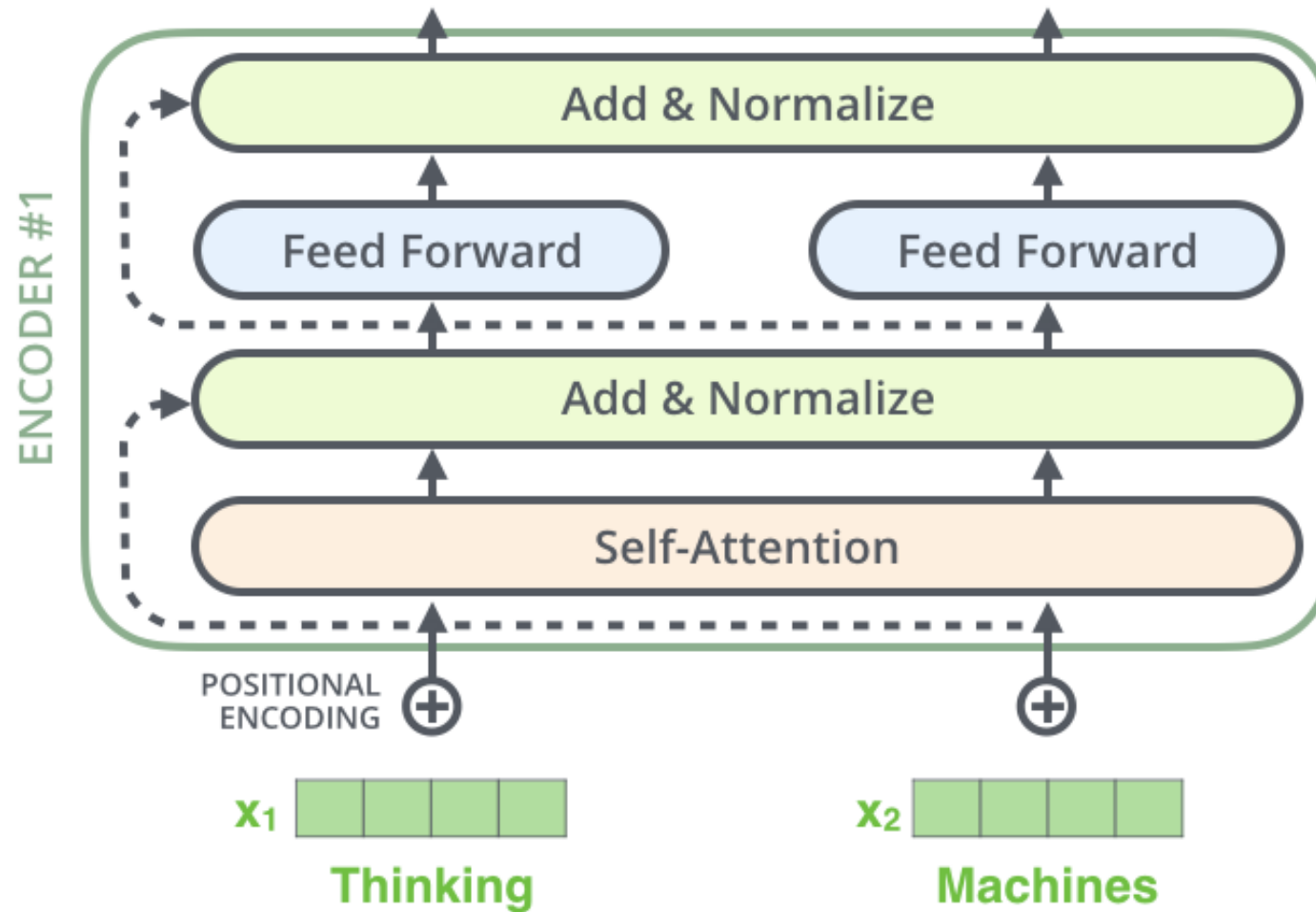
Source: Jay Alammar (2018), The Illustrated Transformer,
<http://jalammar.github.io/illustrated-transformer/>

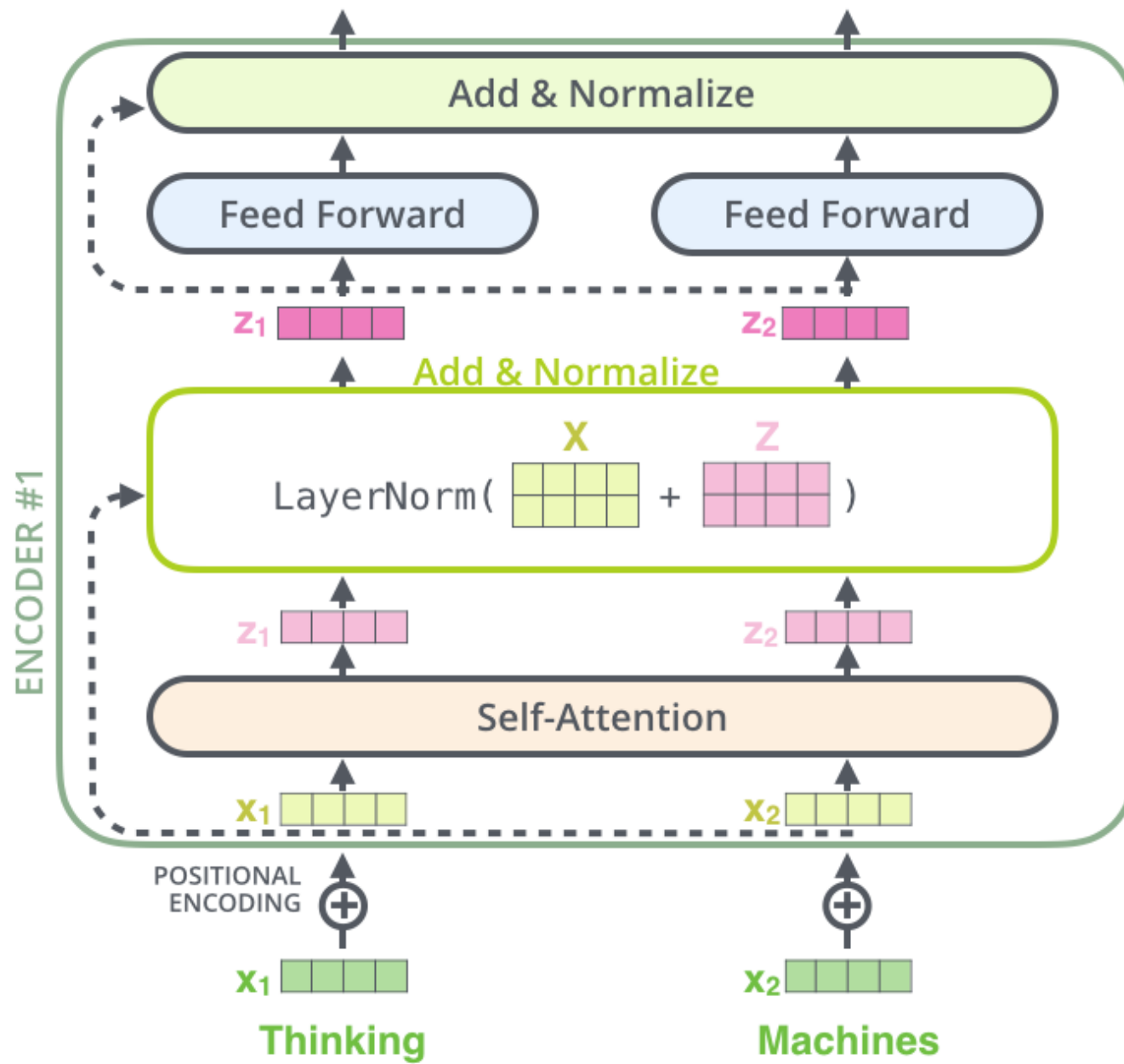
Transformers Positional Encoding

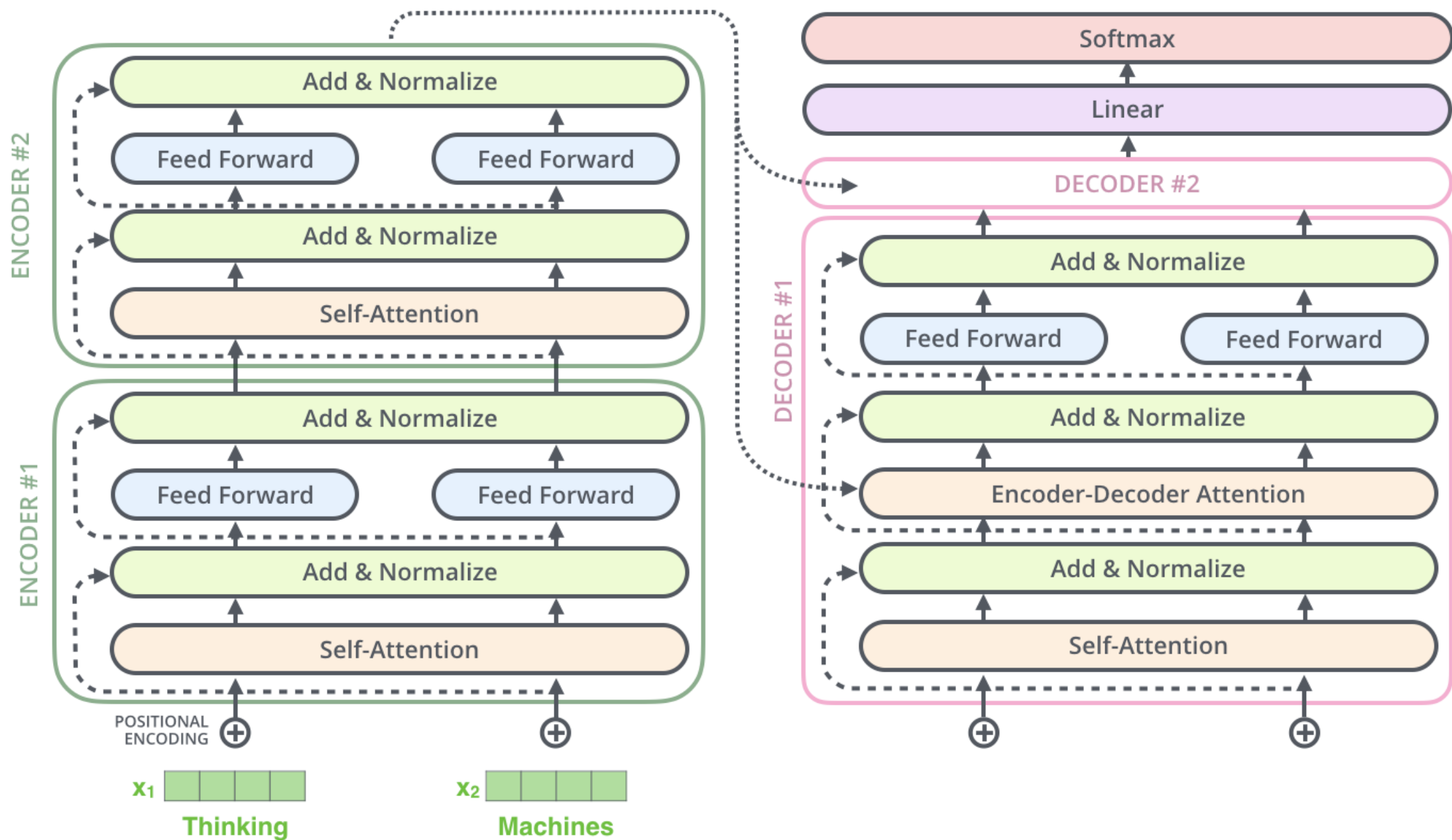


Source: Jay Alammar (2018), The Illustrated Transformer,
<http://jalammar.github.io/illustrated-transformer/>

The Residuals



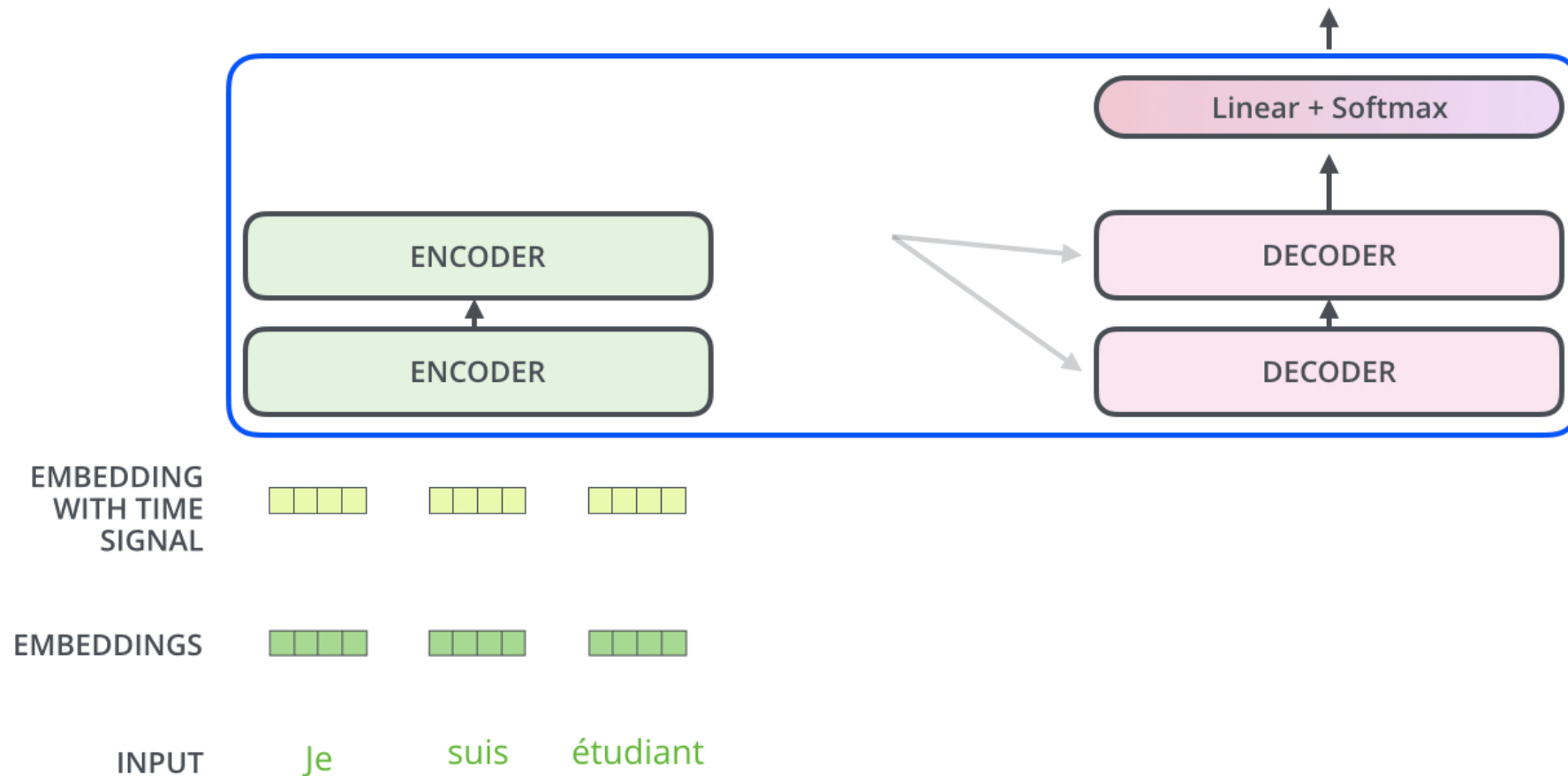




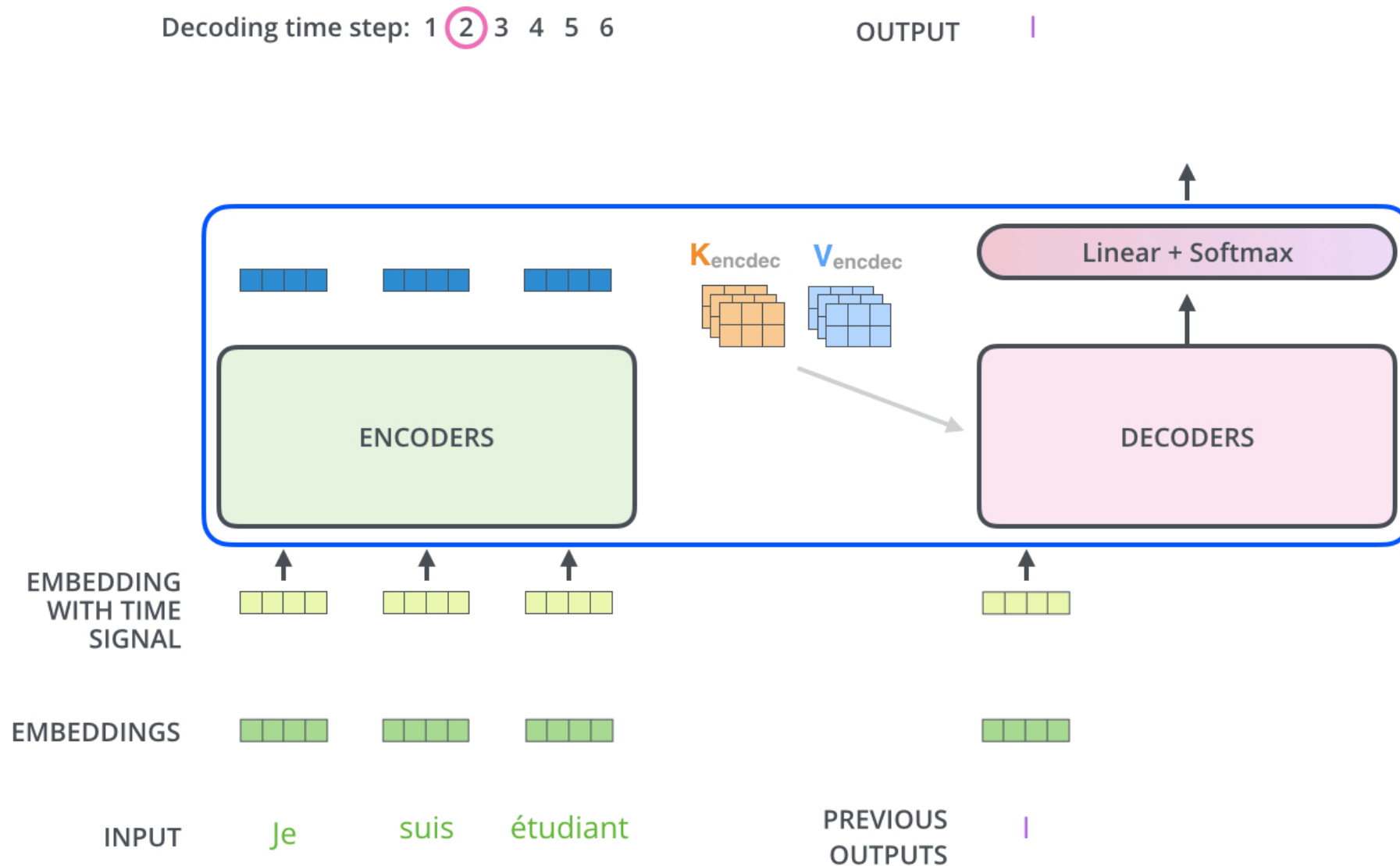
The Decoder Side

Decoding time step: 1 2 3 4 5 6

OUTPUT



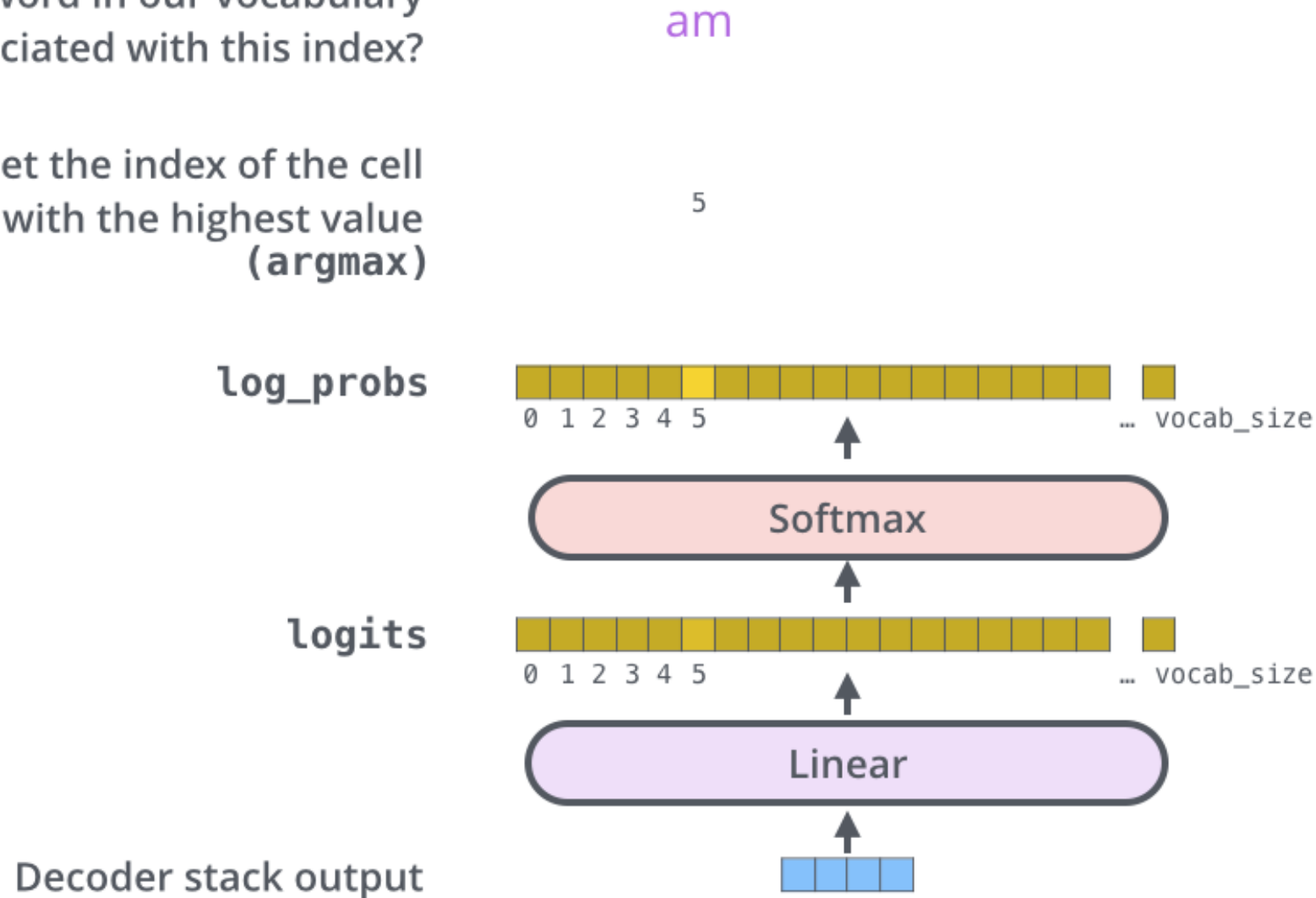
The Decoder Side



The Final Linear and Softmax Layer

Which word in our vocabulary
is associated with this index?

Get the index of the cell
with the highest value
(argmax)



The output vocabulary

Output Vocabulary

WORD	a	am	I	thanks	student	<eos>
INDEX	0	1	2	3	4	5

The output vocabulary of our model is created in the preprocessing phase before we even begin training.

Example: one-hot encoding of output vocabulary

Output Vocabulary

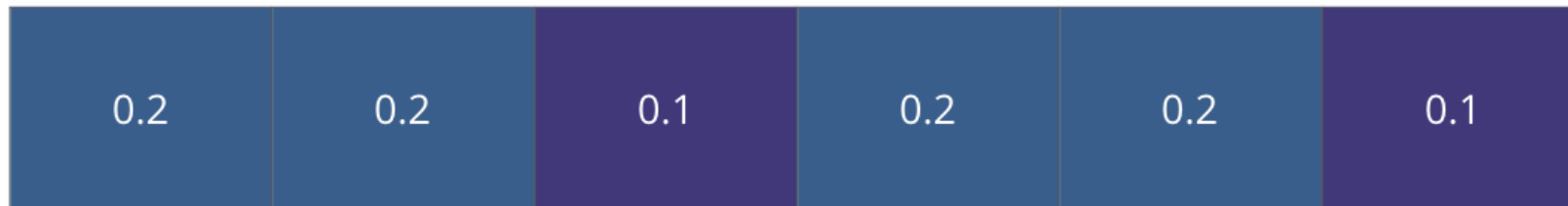
WORD	a	am	I	thanks	student	<eos>
INDEX	0	1	2	3	4	5

One-hot encoding of the word "am"



The Loss Function

Untrained Model Output



Correct and desired output



a

am

I

thanks

student

<eos>

Target Model Outputs

Output Vocabulary: a am I thanks student <eos>

position #1



position #2



position #3



position #4



position #5

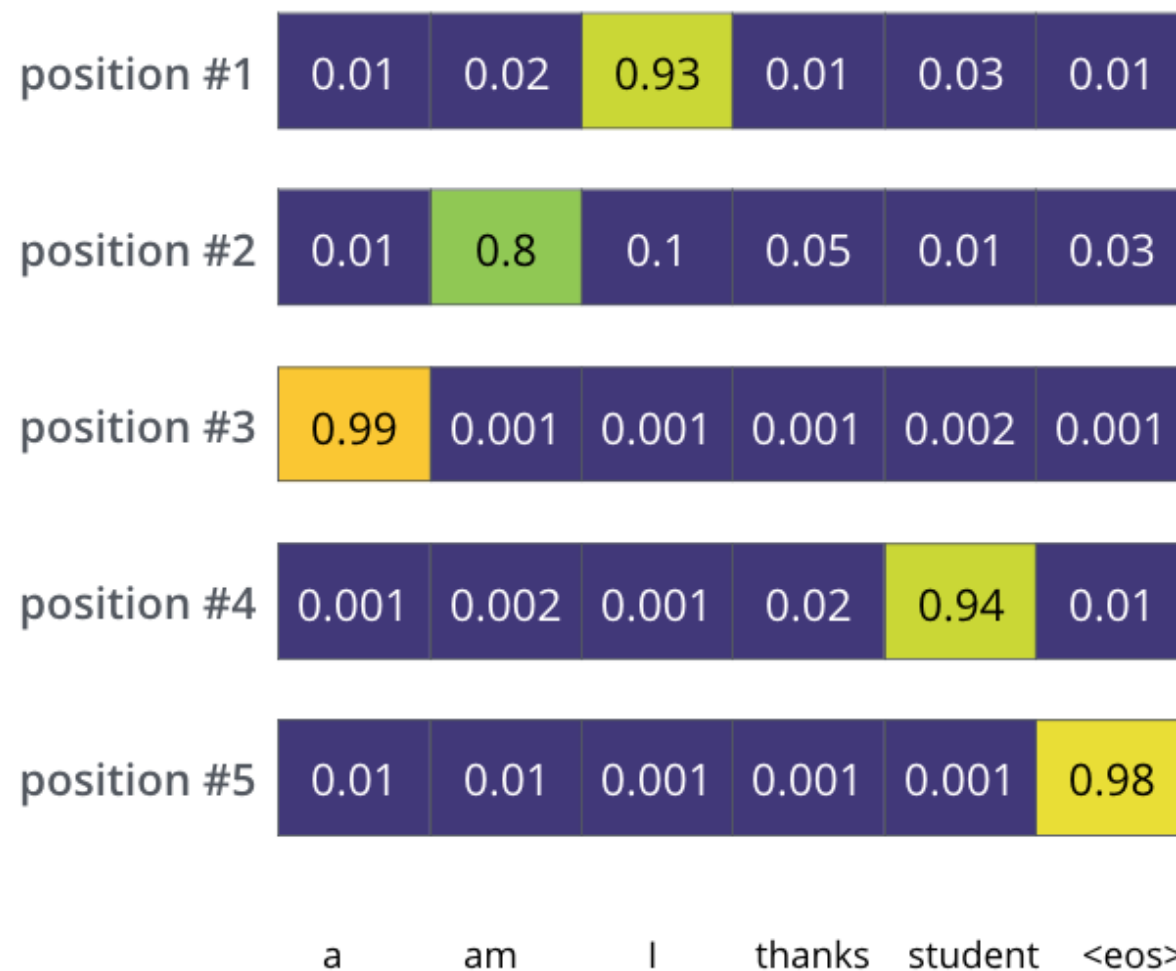


a am I thanks student <eos>



Trained Model Outputs

Output Vocabulary: a am I thanks student <eos>





Transformers Transformers

State-of-the-art Natural Language Processing for TensorFlow 2.0 and PyTorch

- **Transformers**
 - **pytorch-transformers**
 - **pytorch-pretrained-bert**
- **provides state-of-the-art general-purpose architectures**
 - **(BERT, GPT-2, RoBERTa, XLM, DistilBert, XLNet, CTRL...)**
 - **for Natural Language Understanding (NLU) and Natural Language Generation (NLG)**
with over 32+ pretrained models
in 100+ languages
and deep interoperability between
TensorFlow 2.0 and
PyTorch.

Hugging Face



Hugging Face

🔍 Search models, datasets, spaces

📦 Models

📄 Datasets

🏠 Spaces

📄 Docs

👛 Solutions

Pricing



Log In

Sign Up



The AI community building the future.

Build, train and deploy state of the art models powered by
the reference open source in machine learning.




Star


58,696


<https://huggingface.co/>


Hugging Face Transformers


 **Hugging Face**


Search models, datasets, users...

 Models

 Datasets

 Spaces

 Docs

 Solutions

Pricing

Log In

Sign Up

Transformers


Search documentation

V4.16.2

EN

58,697

GET STARTED


 Transformers

Quick tour

Installation

Philosophy

Glossary


USING  TRANSFORMERS

Summary of the tasks


Summary of the models

Preprocessing data


Fine-tuning a pretrained model

Distributed training with 




Accelerate

 **Transformers**


State-of-the-art Machine Learning for Jax, Pytorch and TensorFlow

 Transformers (formerly known as *pytorch-transformers* and *pytorch-pretrained-bert*) provides thousands of pretrained models to perform tasks on different modalities such as text, vision, and audio.

These models can applied on:

-  Text, for tasks like text classification, information extraction, question answering, summarization, translation, text generation, in over 100 languages.
-  Images, for tasks like image classification, object detection, and segmentation.
-  Audio, for tasks like speech recognition and audio classification.

Transformer models can also perform tasks on **several modalities combined**, such as table question answering, optical character recognition, information extraction from scanned documents. video classification. and visual question answering.

 Transformers

If you are looking for custom support from the Hugging Face team

Features

Contents

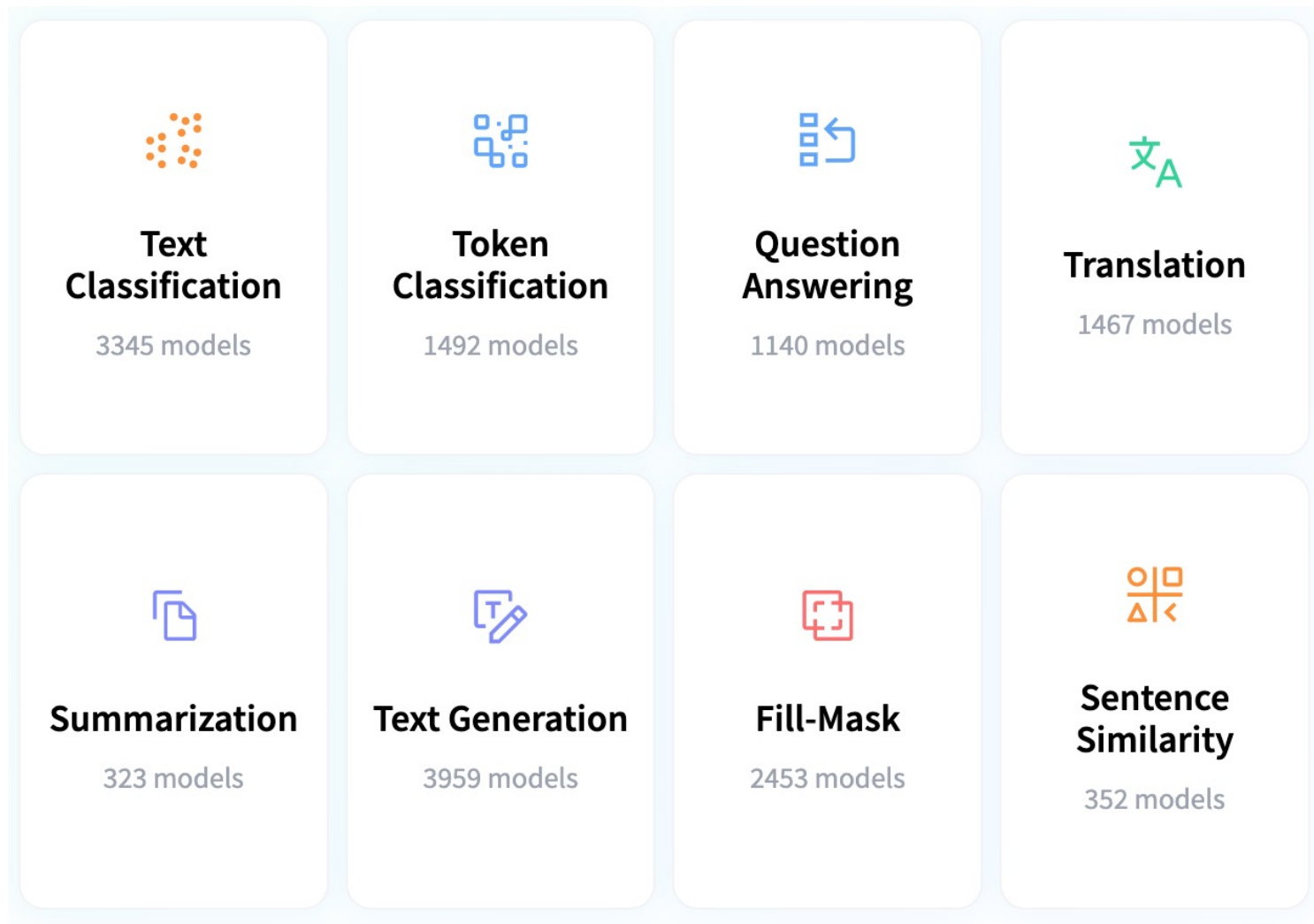
Supported models

Supported frameworks


<https://huggingface.co/docs/transformers/index>

Hugging Face Tasks

Natural Language Processing



NLP with Transformers Github

 Why GitHub? ▾ Team Enterprise Explore ▾ Marketplace Pricing ▾

Search / Sign in Sign up

nlp-with-transformers / notebooks Public

Notifications Fork 170 Star 1.1k ▾

[Code](#) [Issues](#) [Pull requests](#) [Actions](#) [Projects](#) [Wiki](#) [Security](#) [Insights](#)

main ▾ 1 branch 0 tags

Go to file Code ▾

About

Jupyter notebooks for the Natural Language Processing with Transformers book

transformersbook.com/

[Readme](#)

[Apache-2.0 License](#)

[1.1k stars](#)

[33 watching](#)

[170 forks](#)

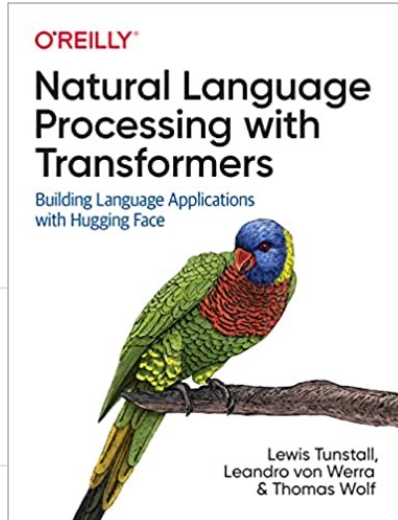
Releases

No releases published

Packages

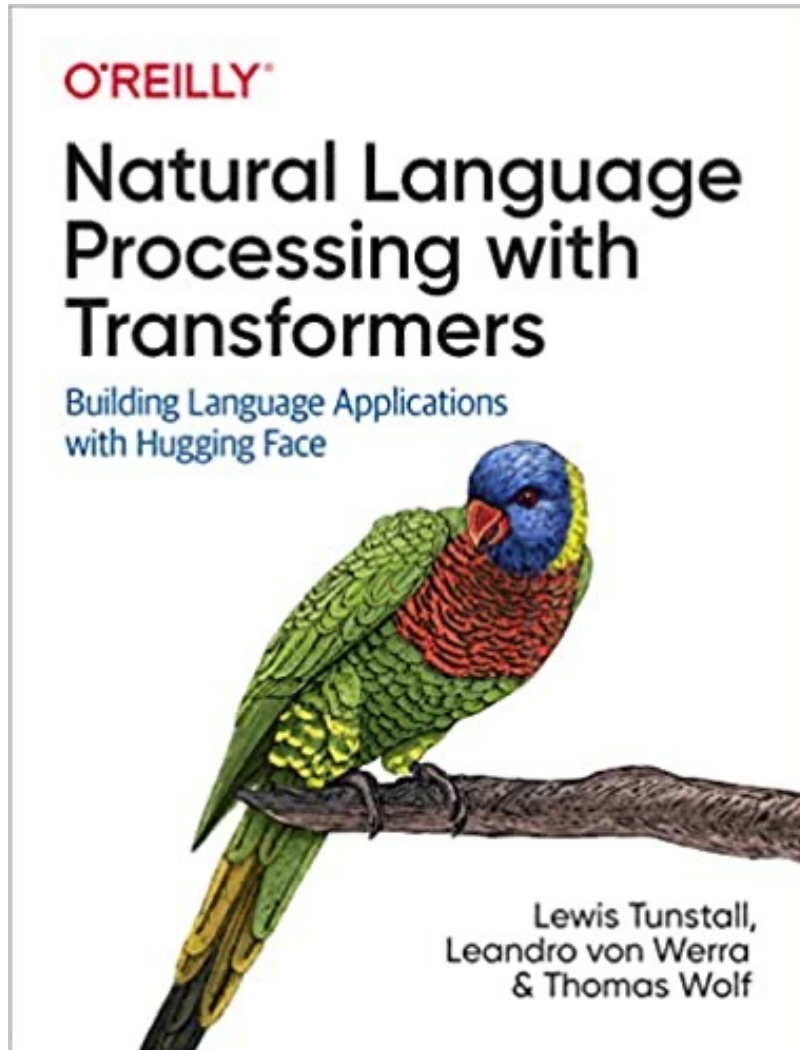
lewtun Merge pull request #21 from JingchaoZhang/patch-3 ... ae5b7c1 15 days ago 71 commits

.github/ISSUE_TEMPLATE	Update issue templates	25 days ago
data	Move dataset to data directory	4 months ago
images	Add README	last month
scripts	Update issue templates	25 days ago
.gitignore	Initial commit	4 months ago
01_introduction.ipynb	Remove Colab badges & fastdoc refs	27 days ago
02_classification.ipynb	Merge pull request #8 from nlp-with-transformers/remove-display-df	26 days ago
03_transformer-anatomy.ipynb	[Transformers Anatomy] Remove cells with figure references	22 days ago
04_multilingual-ner.ipynb	Merge pull request #8 from nlp-with-transformers/remove-display-df	26 days ago
05_text-generation.ipynb	Merge pull request #8 from nlp-with-transformers/remove-display-df	26 days ago















































<https://github.com/nlp-with-transformers/notebooks>

NLP with Transformers Github Notebooks



Running on a cloud platform

To run these notebooks on a cloud platform, just click on one of the badges in the table below:

Chapter	Colab	Kaggle	Gradient	Studio Lab
Introduction	 Open in Colab	 Open in Kaggle	 Run on Gradient	 Open Studio Lab
Text Classification	 Open in Colab	 Open in Kaggle	 Run on Gradient	 Open Studio Lab
Transformer Anatomy	 Open in Colab	 Open in Kaggle	 Run on Gradient	 Open Studio Lab
Multilingual Named Entity Recognition	 Open in Colab	 Open in Kaggle	 Run on Gradient	 Open Studio Lab
Text Generation	 Open in Colab	 Open in Kaggle	 Run on Gradient	 Open Studio Lab
Summarization	 Open in Colab	 Open in Kaggle	 Run on Gradient	 Open Studio Lab
Question Answering	 Open in Colab	 Open in Kaggle	 Run on Gradient	 Open Studio Lab
Making Transformers Efficient in Production	 Open in Colab	 Open in Kaggle	 Run on Gradient	 Open Studio Lab
Dealing with Few to No Labels	 Open in Colab	 Open in Kaggle	 Run on Gradient	 Open Studio Lab
Training Transformers from Scratch	 Open in Colab	 Open in Kaggle	 Run on Gradient	 Open Studio Lab
Future Directions	 Open in Colab	 Open in Kaggle	 Run on Gradient	 Open Studio Lab

Nowadays, the GPUs on Colab tend to be K80s (which have limited memory), so we recommend using [Kaggle](#), [Gradient](#), or [SageMaker Studio Lab](#). These platforms tend to provide more performant GPUs like P100s, all for free!

<https://github.com/nlp-with-transformers/notebooks>

NLP with Transformers

```
!git clone https://github.com/nlp-with-transformers/notebooks.git
%cd notebooks
from install import *
install_requirements()
```

```
from utils import *
setup_chapter()
```

Text Classification

```
text = """Dear Amazon, last week I ordered an Optimus Prime action figure \
from your online store in Germany. Unfortunately, when I opened the package, \
I discovered to my horror that I had been sent an action figure of Megatron \
instead! As a lifelong enemy of the Decepticons, I hope you can understand my \
dilemma. To resolve the issue, I demand an exchange of Megatron for the \
Optimus Prime figure I ordered. Enclosed are copies of my records concerning \
this purchase. I expect to hear from you soon. Sincerely, Bumblebee."""
```

Text Classification

```
text = """Dear Amazon, last week I ordered an Optimus Prime action figure \
from your online store in Germany. Unfortunately, when I opened the package, \
I discovered to my horror that I had been sent an action figure of Megatron \
instead! As a lifelong enemy of the Decepticons, I hope you can understand my \
dilemma. To resolve the issue, I demand an exchange of Megatron for the \
Optimus Prime figure I ordered. Enclosed are copies of my records concerning \
this purchase. I expect to hear from you soon. Sincerely, Bumblebee."""
```

```
from transformers import pipeline
classifier = pipeline("text-classification")
```

```
import pandas as pd
outputs = classifier(text)
pd.DataFrame(outputs)
```

	label	score
0	NEGATIVE	0.901546

Text Classification

```
from transformers import pipeline  
classifier = pipeline("text-classification")
```

```
import pandas as pd  
outputs = classifier(text)  
pd.DataFrame(outputs)
```

	label	score
0	NEGATIVE	0.901546

Named Entity Recognition

```
ner_tagger = pipeline("ner", aggregation_strategy="simple")
outputs = ner_tagger(text)
pd.DataFrame(outputs)
```

	entity_group	score	word	start	end
0	ORG	0.879010	Amazon	5	11
1	MISC	0.990859	Optimus Prime	36	49
2	LOC	0.999755	Germany	90	97
3	MISC	0.556570	Mega	208	212
4	PER	0.590256	##tron	212	216
5	ORG	0.669692	Decept	253	259
6	MISC	0.498349	##icons	259	264
7	MISC	0.775362	Megatron	350	358
8	MISC	0.987854	Optimus Prime	367	380
9	PER	0.812096	Bumblebee	502	511

Question Answering

```
reader = pipeline("question-answering")
question = "What does the customer want?"
outputs = reader(question=question, context=text)
pd.DataFrame([outputs])
```

	score	start	end	answer
0	0.631292	335	358	an exchange of Megatron

Summarization

```
summarizer = pipeline("summarization")  
outputs = summarizer(text, max_length=45, clean_up_tokenization_spaces=True)  
print(outputs[0]['summary_text'])
```

Bumblebee ordered an Optimus Prime action figure from your online store in Germany. Unfortunately, when I opened the package, I discovered to my horror that I had been sent an action figure of Megatron instead.

Translation

```
translator = pipeline("translation_en_to_de",  
                        model="Helsinki-NLP/opus-mt-en-de")  
outputs = translator(text, clean_up_tokenization_spaces=True, min_length=100)  
print(outputs[0]['translation_text'])
```

Sehr geehrter Amazon, letzte Woche habe ich eine Optimus Prime Action Figur aus Ihrem Online-Shop in Deutschland bestellt. Leider, als ich das Paket öffnete, entdeckte ich zu meinem Entsetzen, dass ich stattdessen eine Action Figur von Megatron geschickt worden war! Als lebenslanger Feind der Decepticons, Ich hoffe, Sie können mein Dilemma verstehen. Um das Problem zu lösen, Ich fordere einen Austausch von Megatron für die Optimus Prime Figur habe ich bestellt. Anbei sind Kopien meiner Aufzeichnungen über diesen Kauf. Ich erwarte, bald von Ihnen zu hören. Aufrichtig, Bumblebee.

Text Generation

```
from transformers import set_seed
set_seed(42) # Set the seed to get reproducible results

generator = pipeline("text-generation")
response = "Dear Bumblebee, I am sorry to hear that your order was mixed up."
prompt = text + "\n\nCustomer service response:\n" + response
outputs = generator(prompt, max_length=200)
print(outputs[0]['generated_text'])
```

Customer service response:

Dear Bumblebee, I am sorry to hear that your order was mixed up. The order was completely mislabeled, which is very common in our online store, but I can appreciate it because it was my understanding from this site and our customer service of the previous day that your order was not made correct in our mind and that we are in a process of resolving this matter. We can assure you that your order

Text Generation

Dear Amazon, last week I ordered an Optimus Prime action figure from your online store in Germany. Unfortunately, when I opened the package, I discovered to my horror that I had been sent an action figure of Megatron instead! As a lifelong enemy of the Decepticons, I hope you can understand my dilemma. To resolve the issue, I demand an exchange of Megatron for the Optimus Prime figure I ordered. Enclosed are copies of my records concerning this purchase. I expect to hear from you soon. Sincerely, Bumblebee.

Customer service response:

Dear Bumblebee, I am sorry to hear that your order was mixed up. The order was completely mislabeled, which is very common in our online store, but I can appreciate it because it was my understanding from this site and our customer service of the previous day that your order was not made correct in our mind and that we are in a process of resolving this matter. We can assure you that your order

Python in Google Colab (Python101)

<https://colab.research.google.com/drive/1FEG6DnGvwfUbeo4zJ1zTunjMqf2RkCrT>

python101.ipynb ☆

File Edit View Insert Runtime Tools Help [All changes saved](#)

Table of contents

- Natural Language Processing with Transformers
 - Text Classification
 - Named Entity Recognition
 - Question Answering
 - Summarization
 - Translation
 - Text Generation
- AI in Finance
 - Normative Finance and Financial Theories
 - Uncertainty and Risk
 - Expected Utility Theory (EUT)
 - Mean-Variance Portfolio Theory (MVPT)
 - Capital Asset Pricing Model (CAPM)
 - Arbitrage Pricing Theory (APT)
- Data Driven Finance
 - Financial Econometrics and Regression
 - Data Availability
 - Normative Theories Revisited
 - Mean-Variance Portfolio Theory

+ Code + Text

RAM Disk Editing

Natural Language Processing with Transformers

- Source: Lewis Tunstall, Leandro von Werra, and Thomas Wolf (2022), Natural Language Processing with Transformers: Building Language Applications with Hugging Face, O'Reilly Media.
- Github: <https://github.com/nlp-with-transformers/notebooks>

```
[1] 1 !git clone https://github.com/nlp-with-transformers/notebooks.git
    2 %cd notebooks
    3 from install import *
    4 install_requirements()
```

```
[3] 1 from utils import *
    2 setup_chapter()
```

```
[12] 1 text = """Dear Amazon, last week I ordered an Optimus Prime action figure \
      2 from your online store in Germany. Unfortunately, when I opened the package, \
      3 I discovered to my horror that I had been sent an action figure of Megatron \
      4 instead! As a lifelong enemy of the Decepticons, I hope you can understand my \
      5 dilemma. To resolve the issue, I demand an exchange of Megatron for the \
      6 Optimus Prime figure I ordered. Enclosed are copies of my records concerning \
      7 this purchase. I expect to hear from you soon. Sincerely, Bumblebee."""
```

Text Classification

```
[13] 1 from transformers import pipeline
      2 classifier = pipeline("text-classification")
```

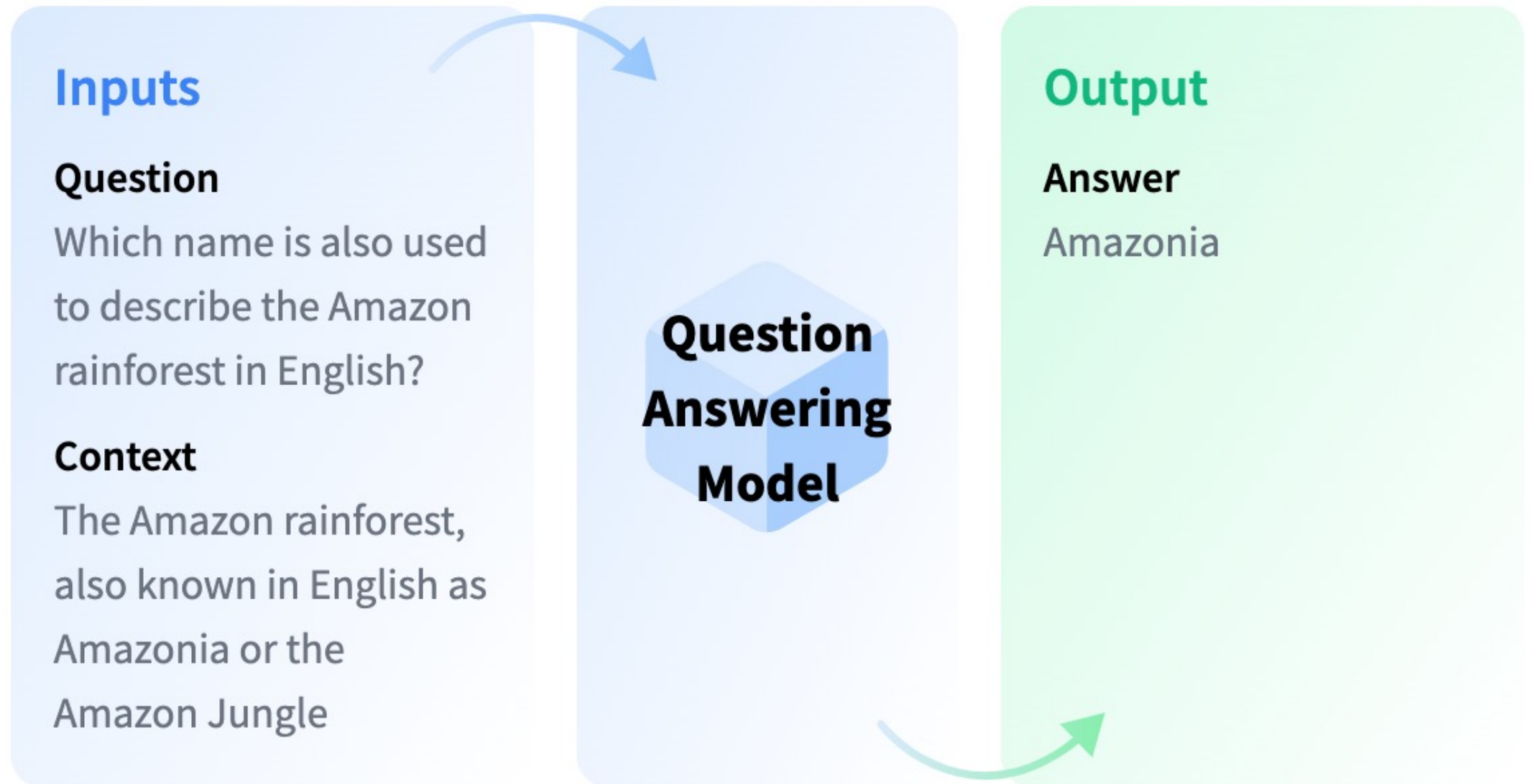
```
[14] 1 import pandas as pd
      2 outputs = classifier(text)
      3 pd.DataFrame(outputs)
```

<https://tinyurl.com/aintpupython101>

Outline

- Introduction
- Overview of Generative AI
- Overview of Large Language Models (LLMs)
- Foundation of Transformers: Attention Mechanism
- **Fine-tuning LLM for Question Answering System**
- Fine-tuning LLM for Dialogue System
- Challenges and Limitations of Generative AI for QA and Dialogue Systems
- Q & A

Question Answering



Question Answering

⚡ Question Answering demo

using [deepset/roberta-base-squad2](#)

📄 Question Answering

Example 2



Where do I live?

Compute

Context

My name is Michael and I live in Taipei.

Computation time on cpu: 0.0492 s

Taipei

0.920

</> JSON Output

🖥 Maximize

Question Answering

```
!pip install transformers
from transformers import pipeline
qamodel = pipeline("question-answering")
question = "Where do I live?"
context = "My name is Michael and I live in Taipei."
qamodel(question = question, context = context)
```

```
{ 'answer': 'Taipei', 'end': 39, 'score': 0.9730741381645203, 'start': 33 }
```

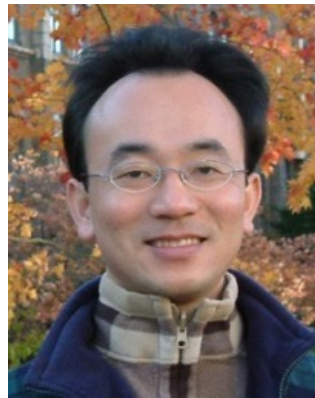
Question Answering

```
from transformers import pipeline
qamodel = pipeline("question-answering", model='deepset/roberta-base-squad2')
question = "Where do I live?"
context = "My name is Michael and I live in Taipei."
output = qamodel(question = question, context = context)
print(output['answer'])
```

Taipei

IMTKU Textual Entailment System for Recognizing Inference in Text at **NTCIR-9** RITE

Department of Information Management
Tamkang University, Taiwan



Min-Yuh Day

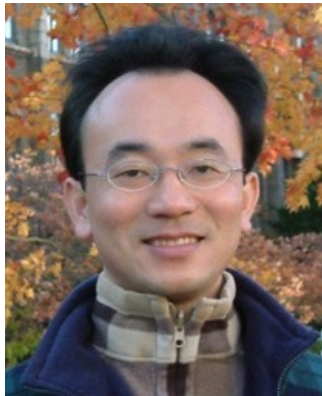
myday@mail.tku.edu.tw



Chun Tu

IMTKU Textual Entailment System for Recognizing Inference in Text at **NTCIR-10** RITE-2

Department of Information Management
Tamkang University, Taiwan



Min-Yuh Day



Chun Tu



Hou-Cheng Vong

myday@mail.tku.edu.tw



Shih-Wei Wu



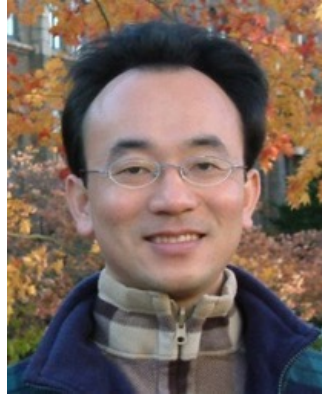
Shih-Jhen Huang

IMTKU Textual Entailment System for Recognizing Inference in Text at NTCIR-11 RITE-VAL

Tamkang University

淡江大學

2014



Min-Yuh Day



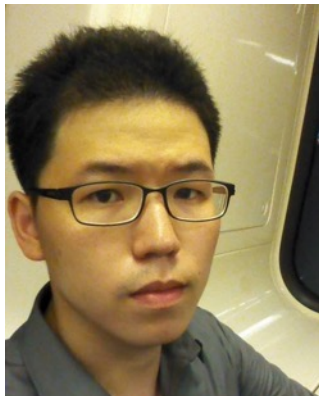
Ya-Jung Wang



Che-Wei Hsu



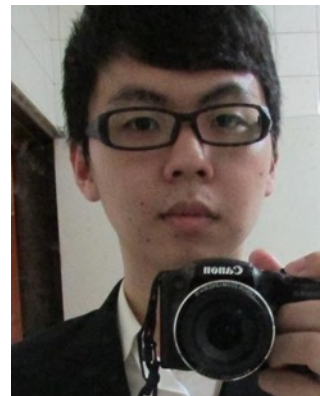
En-Chun Tu



Huai-Wen Hsu



Yu-An Lin



Shang-Yu Wu



Yu-Hsuan Tai

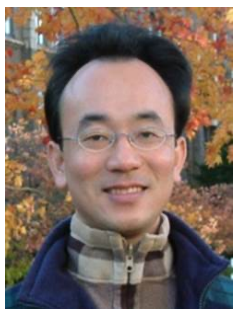


Cheng-Chia Tsai

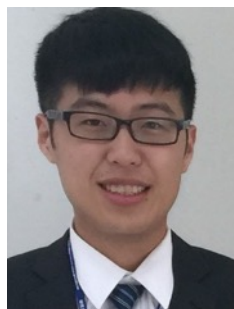
IMTKU Question Answering System for World History Exams at **NTCIR-12** QA Lab2

Department of Information Management
Tamkang University, Taiwan

Sagacity Technology



Min-Yuh Day



Cheng-Chia Tsai



Wei-Chun Chung



Hsiu-Yuan Chang



Tzu-Jui Sun



Yuan-Jie Tsai



Jin-Kun Lin



Cheng-Hung Lee



Yu-Ming Guo



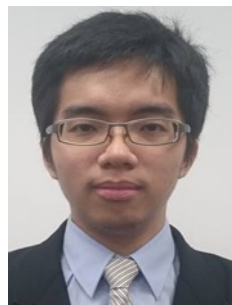
Yue-Da Lin



Wei-Ming Chen



Yun-Da Tsai



Cheng-Jhih Han



Yi-Jing Lin



Yi-Heng Chiang

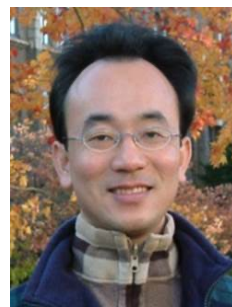


Ching-Yuan Chien

myday@mail.tku.edu.tw

IMTKU Question Answering System for World History Exams at **NTCIR-13** QALab-3

Department of Information Management
Tamkang University, Taiwan



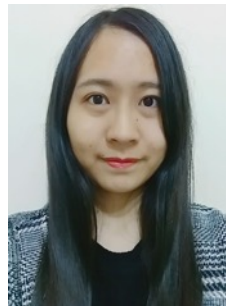
Min-Yuh Day



Chao-Yu Chen



Wanchu Huang



Shi-Ya Zheng



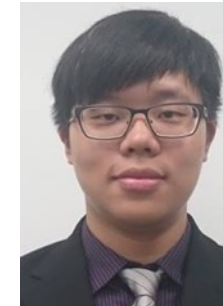
I-Hsuan Huang



Tz-Rung Chen



Min-Chun Kuo



Yue-Da Lin



Yi-Jing Lin

myday@mail.tku.edu.tw

IMTKU Emotional Dialogue System for Short Text Conversation at **NTCIR-14** STC-3 (CECG) Task

Department of Information Management
Tamkang University, Taiwan



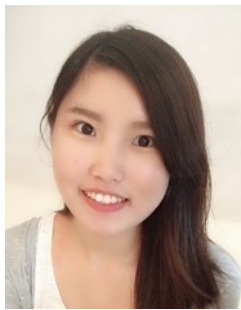
Min-Yuh Day



Chi-Sheng Hung



Yi-Jun Xie



Jhih-Yi Chen



Yu-Ling Kuo



Jian-Ting Lin

IMTKU Multi-Turn Dialogue System Evaluation at the NTCIR-15 DialEval-1 Dialogue Quality and Nugget Detection

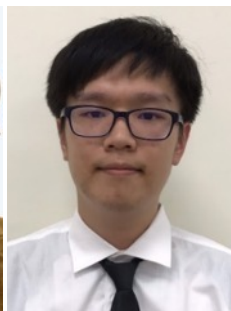
¹ Zeals Co., Ltd. Tokyo, Japan

² Information Management, Tamkang University, Taiwan

³ Information Management, National Taipei University, Taiwan



Mike Tian-Jian Jiang¹



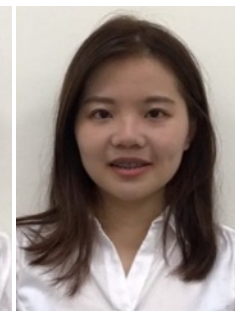
Zhao-Xian Gu²



Cheng-Jhe Chiang²



Yueh-Chia Wu²



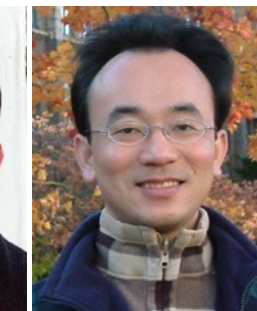
Yu-Chen Huang²



Cheng-Han Chiu²



Sheng-Ru Shaw²



Min-Yuh Day³

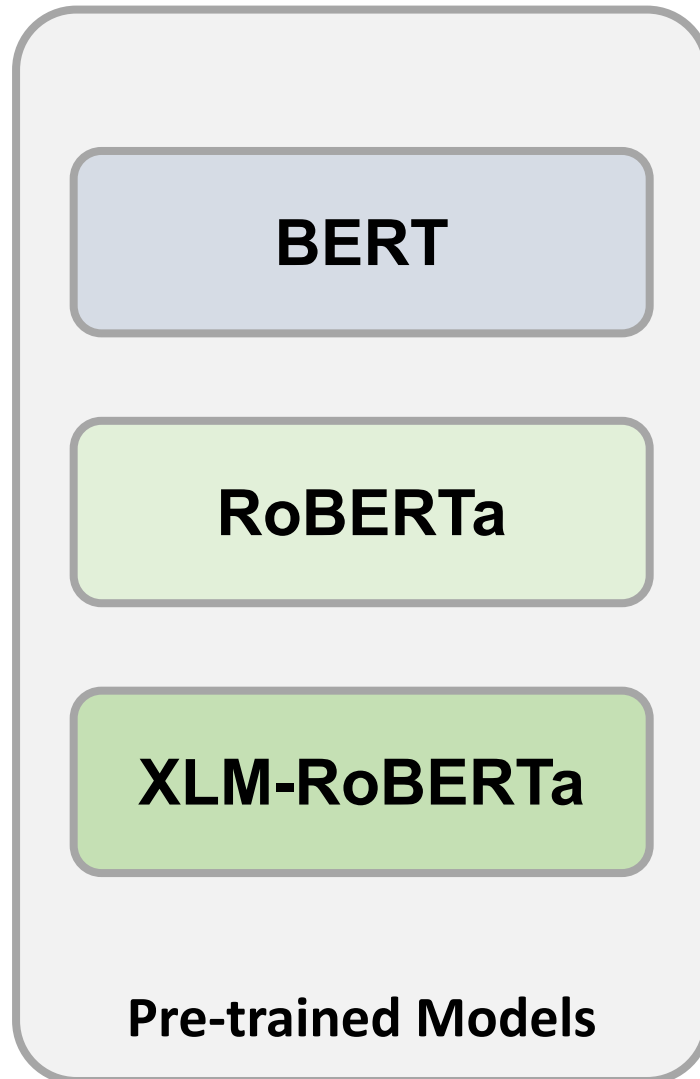
2020 NTCIR-15 Dialogue Evaluation (DialEval-1) Task

Dialogue Quality (DQ) and Nugget Detection (ND)

Chinese Dialogue Quality (S-score) Results (Zeng et al., 2020)

Run	Mean RSNOD	Run	Mean NMD
IMTKU-run2	0.1918	IMTKU-run2	0.1254
IMTKU-run1	0.1964	IMTKU-run0	0.1284
IMTKU-run0	0.1977	IMTKU-run1	0.1290
TUA1-run2	0.2024	TUA1-run2	0.1310
TUA1-run0	0.2053	TUA1-run0	0.1322
NKUST-run1	0.2057	NKUST-run1	0.1363
BL-lstm	0.2088	TUA1-run1	0.1397
WUST-run0	0.2131	BL-popularity	0.1442
RSLNV-run0	0.2141	BL-lstm	0.1455
BL-popularity	0.2288	RSLNV-run0	0.1483
TUA1-run1	0.2302	WUST-run0	0.1540
NKUST-run0	0.2653	NKUST-run0	0.2289
BL-uniform	0.2811	BL-uniform	0.2497

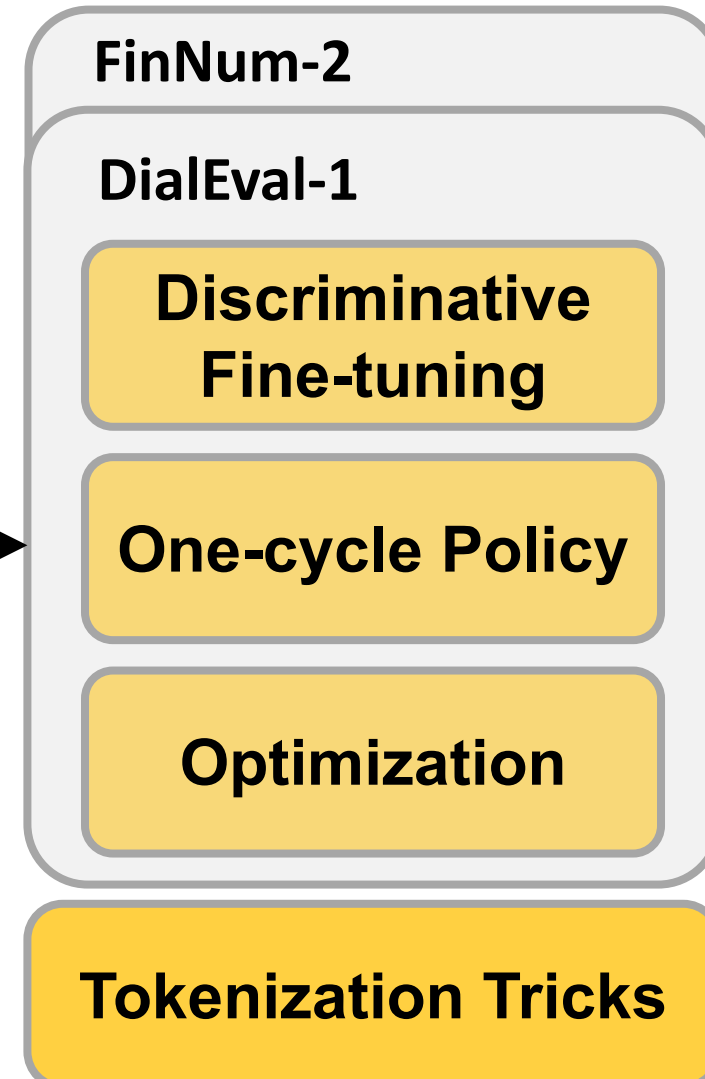
Transformer-based Models Selection



Transfer Learning



Fine-tuning Techniques

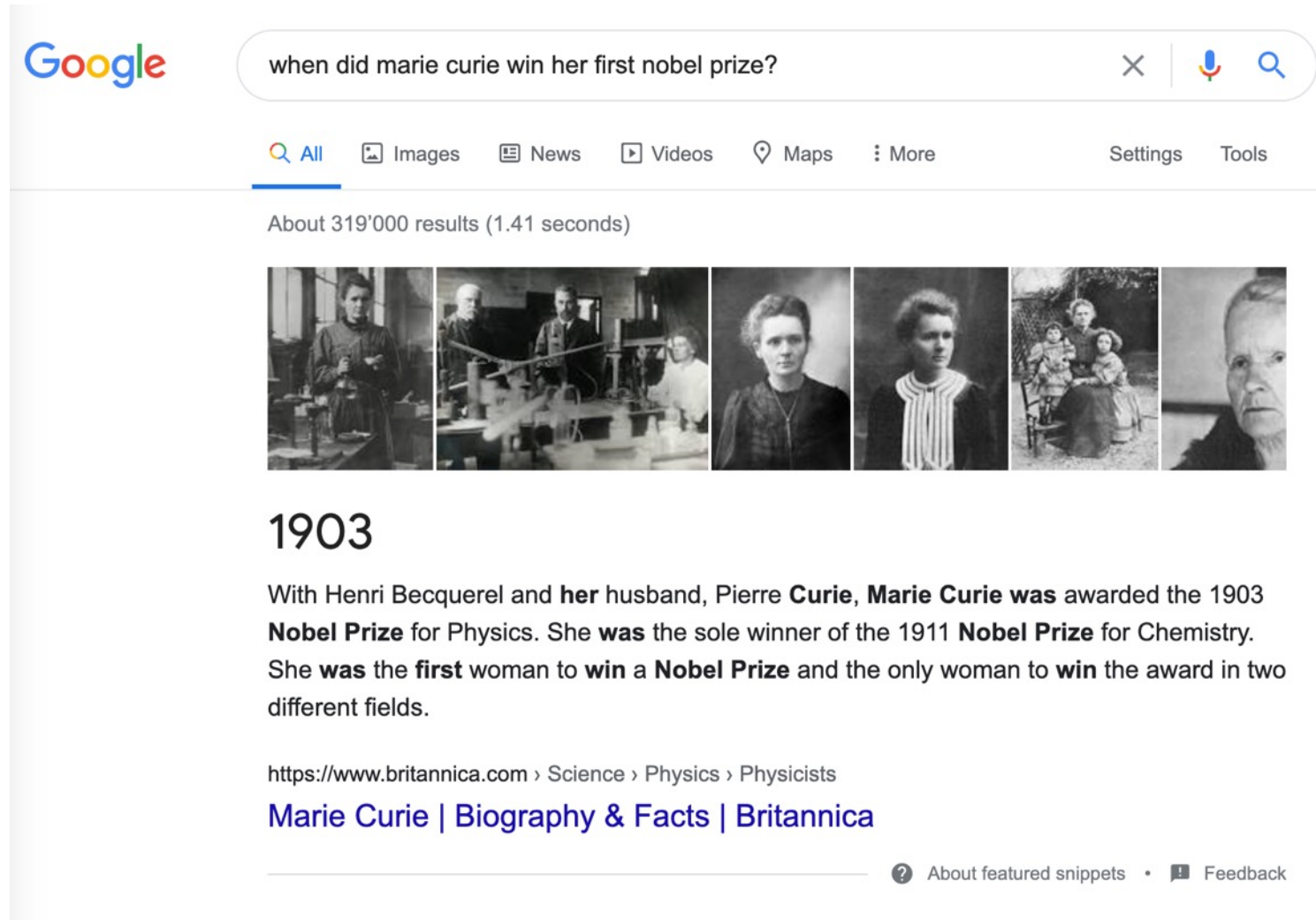


Question Answering

Question Answering

When did Marie Curie win her first Nobel Prize?

1903




The screenshot shows a Google search interface. The search bar contains the text "when did marie curie win her first nobel prize?". Below the search bar, there are tabs for "All", "Images", "News", "Videos", "Maps", and "More". The "All" tab is selected. Below the tabs, it says "About 319'000 results (1.41 seconds)". There is a row of six small images showing Marie Curie in various settings. Below the images, the year "1903" is displayed in large text. Underneath, a text snippet from Britannica reads: "With Henri Becquerel and her husband, Pierre Curie, Marie Curie was awarded the 1903 Nobel Prize for Physics. She was the sole winner of the 1911 Nobel Prize for Chemistry. She was the first woman to win a Nobel Prize and the only woman to win the award in two different fields." Below this snippet, there is a link to "https://www.britannica.com > Science > Physics > Physicists" and a link to "Marie Curie | Biography & Facts | Britannica". At the bottom right, there are links for "About featured snippets" and "Feedback".

Google

when did marie curie win her first nobel prize?

Q All Images News Videos Maps More Settings Tools

About 319'000 results (1.41 seconds)



1903

With Henri Becquerel and her husband, Pierre Curie, Marie Curie was awarded the 1903 Nobel Prize for Physics. She was the sole winner of the 1911 Nobel Prize for Chemistry. She was the first woman to win a Nobel Prize and the only woman to win the award in two different fields.

<https://www.britannica.com > Science > Physics > Physicists>

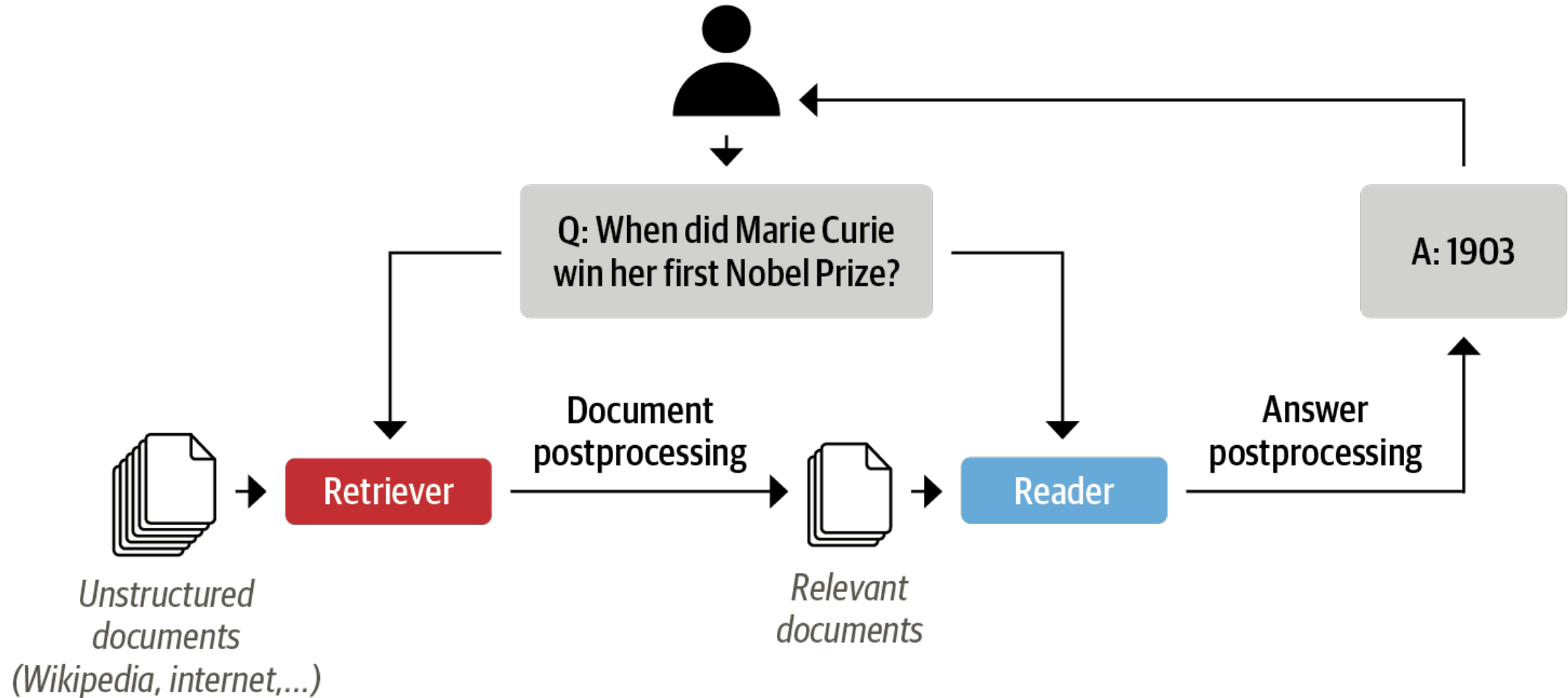
[Marie Curie | Biography & Facts | Britannica](#)

About featured snippets • Feedback

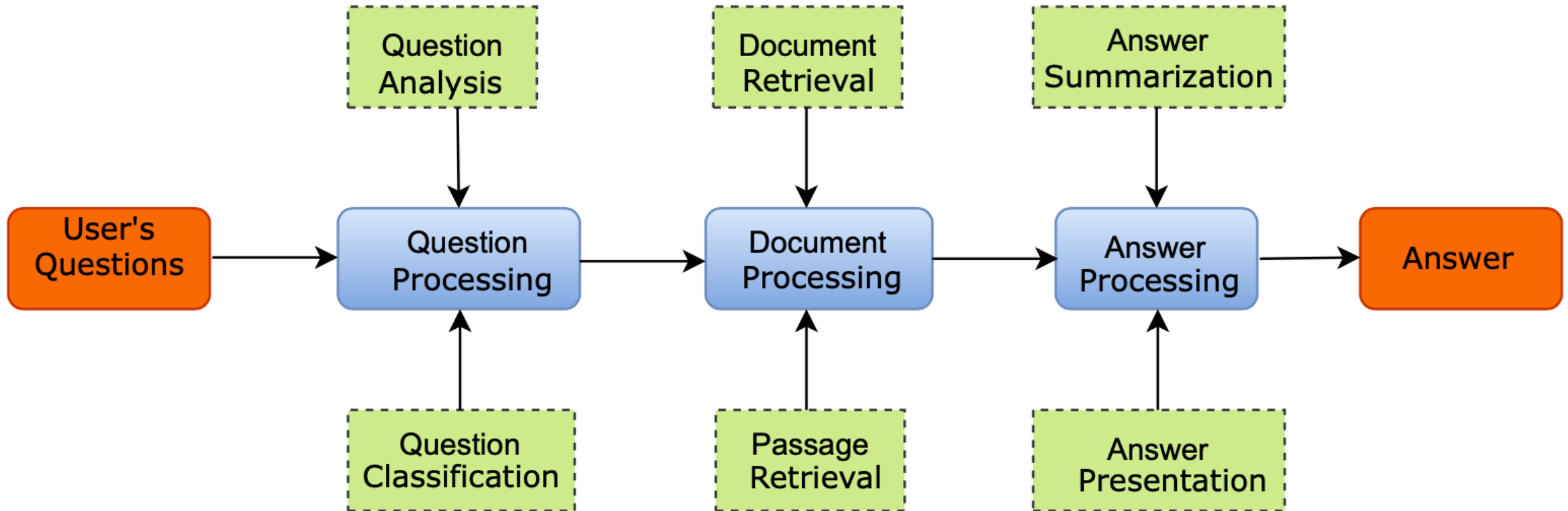
The Retriever-Reader Architecture for Modern QA Systems

When did Marie Curie win her first Nobel Prize?

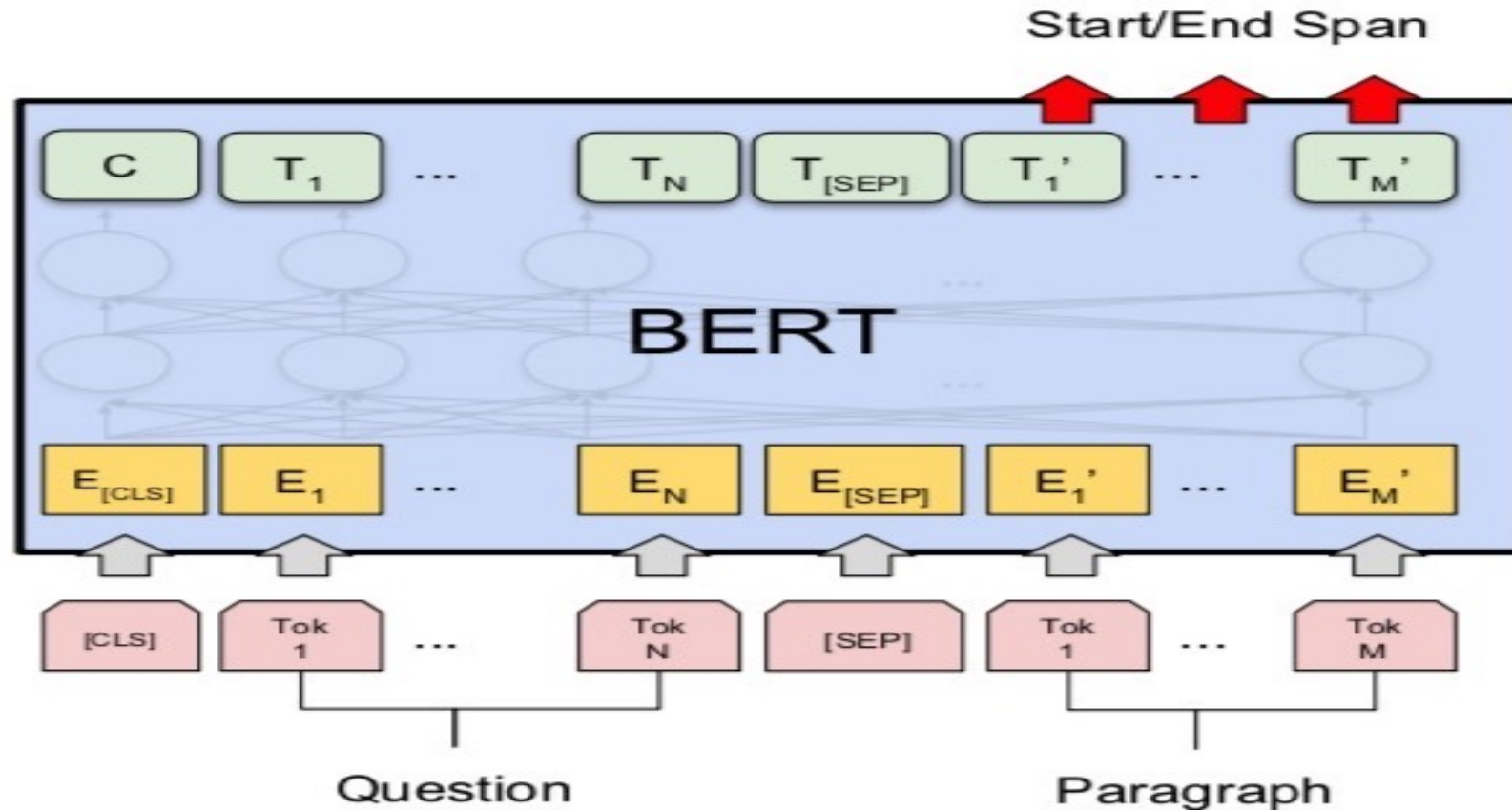
1903



Question Answering System (QAS)

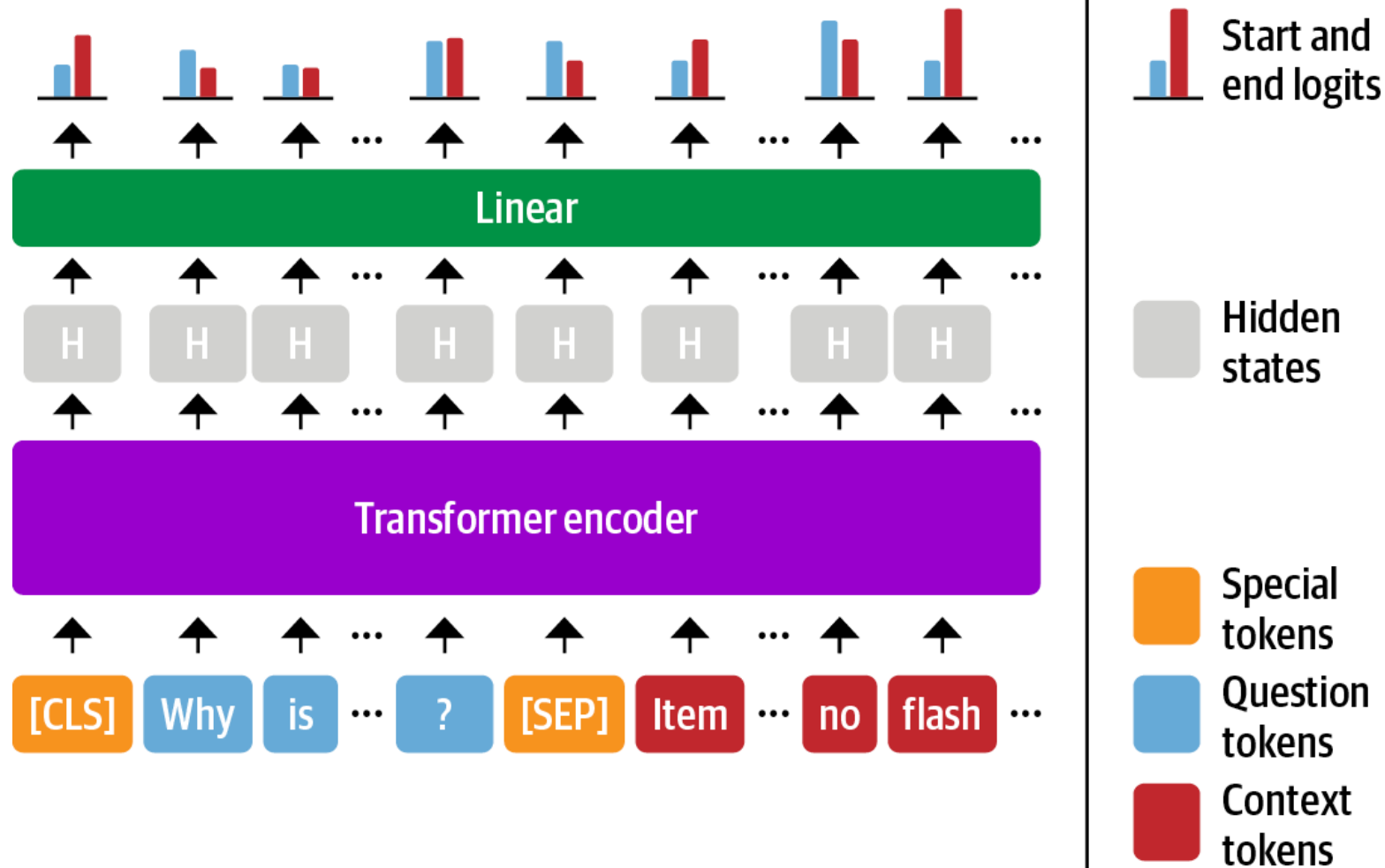


Fine-tuning BERT on Question Answering (QA)



(c) Question Answering Tasks:
SQuAD v1.1

The span classification head for QA tasks



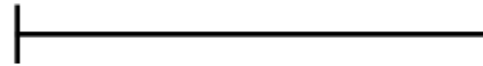
Question answering

Multiple question-context pairs

Why is the camera of poor quality? Item like the picture, fast deliver 3 days well packed, good quality for the price. The camera is decent (as phone cameras go). There is no flash though...



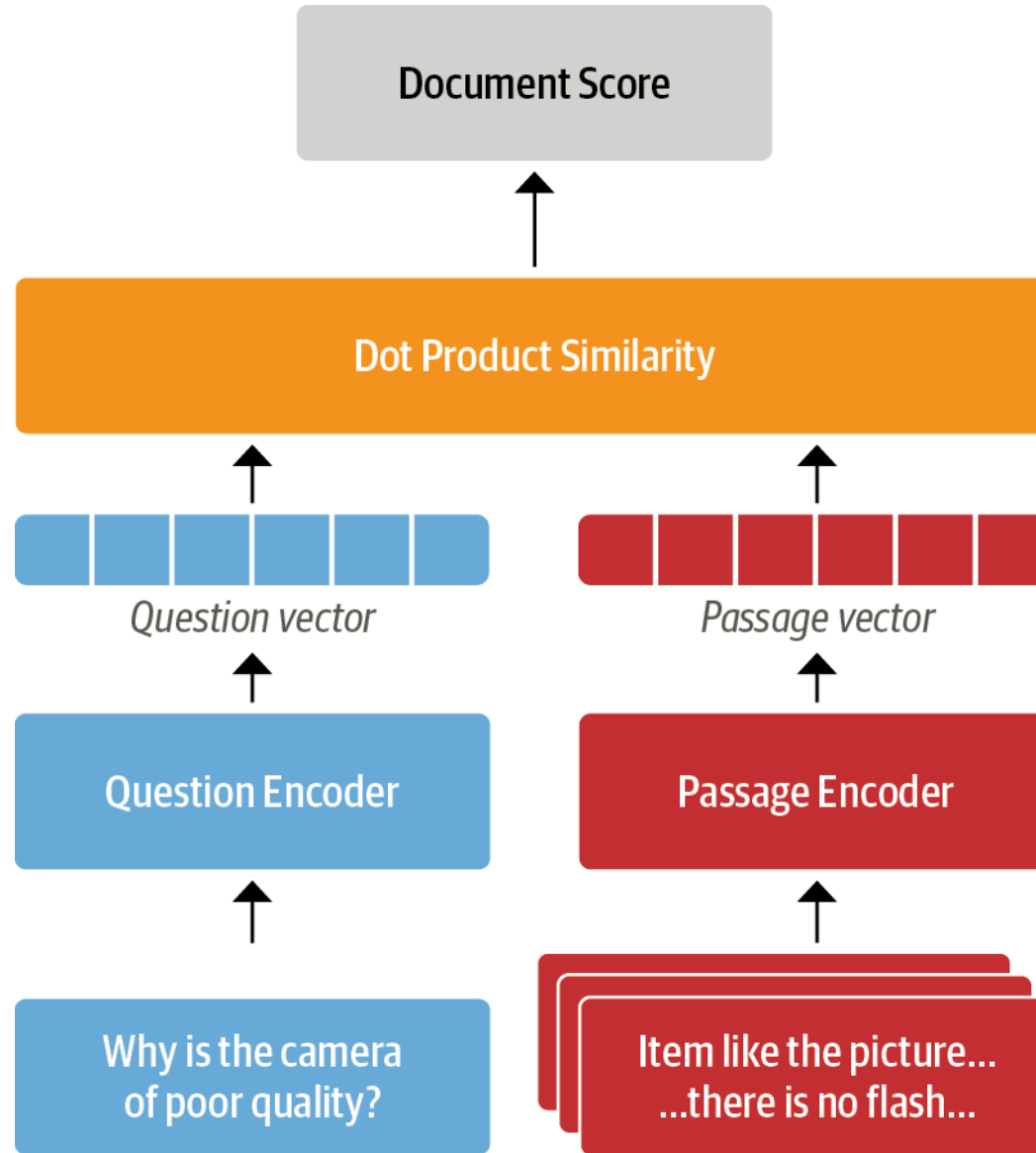
Stride



Why is the camera of poor quality? Item like the picture, fast deliver 3 days well packed, good quality for the price. The camera is decent (as phone cameras go). There is no flash though...



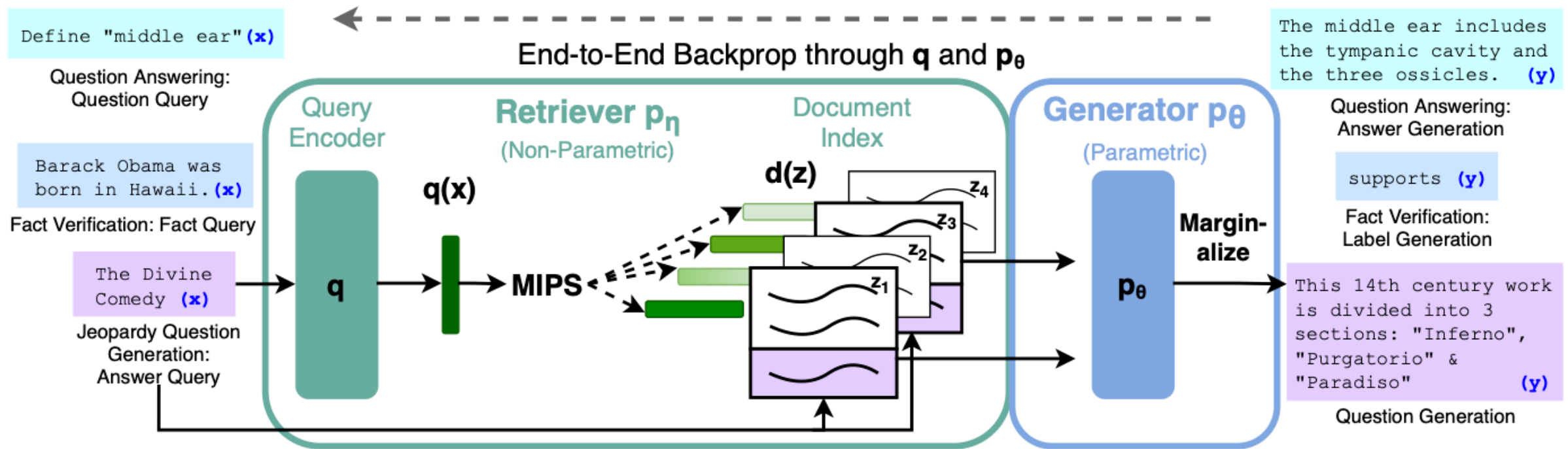
Dense Passage Retrieval (DPR)



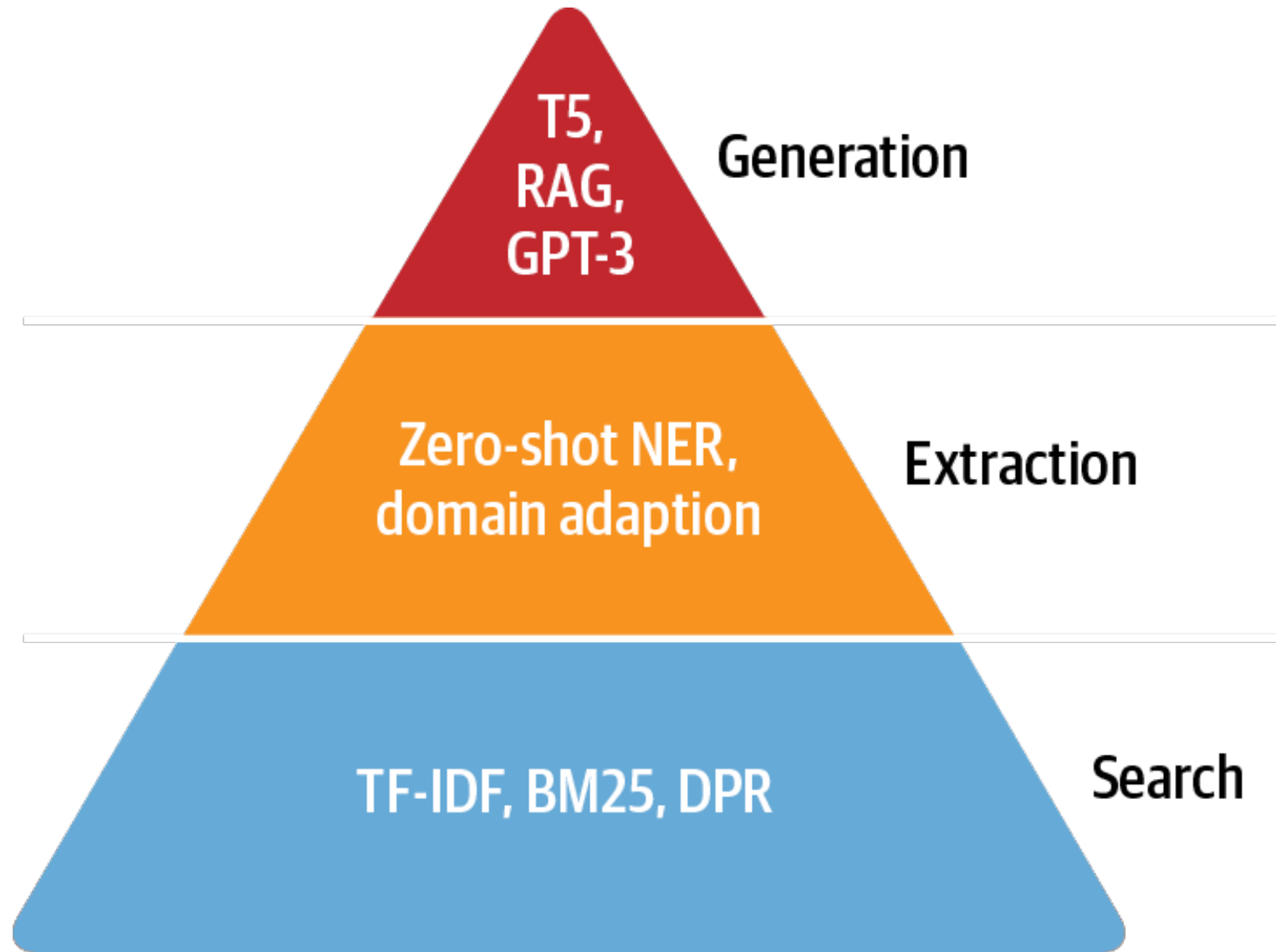
Going Beyond Extractive QA

Retrieval-Augmented Generation (RAG)

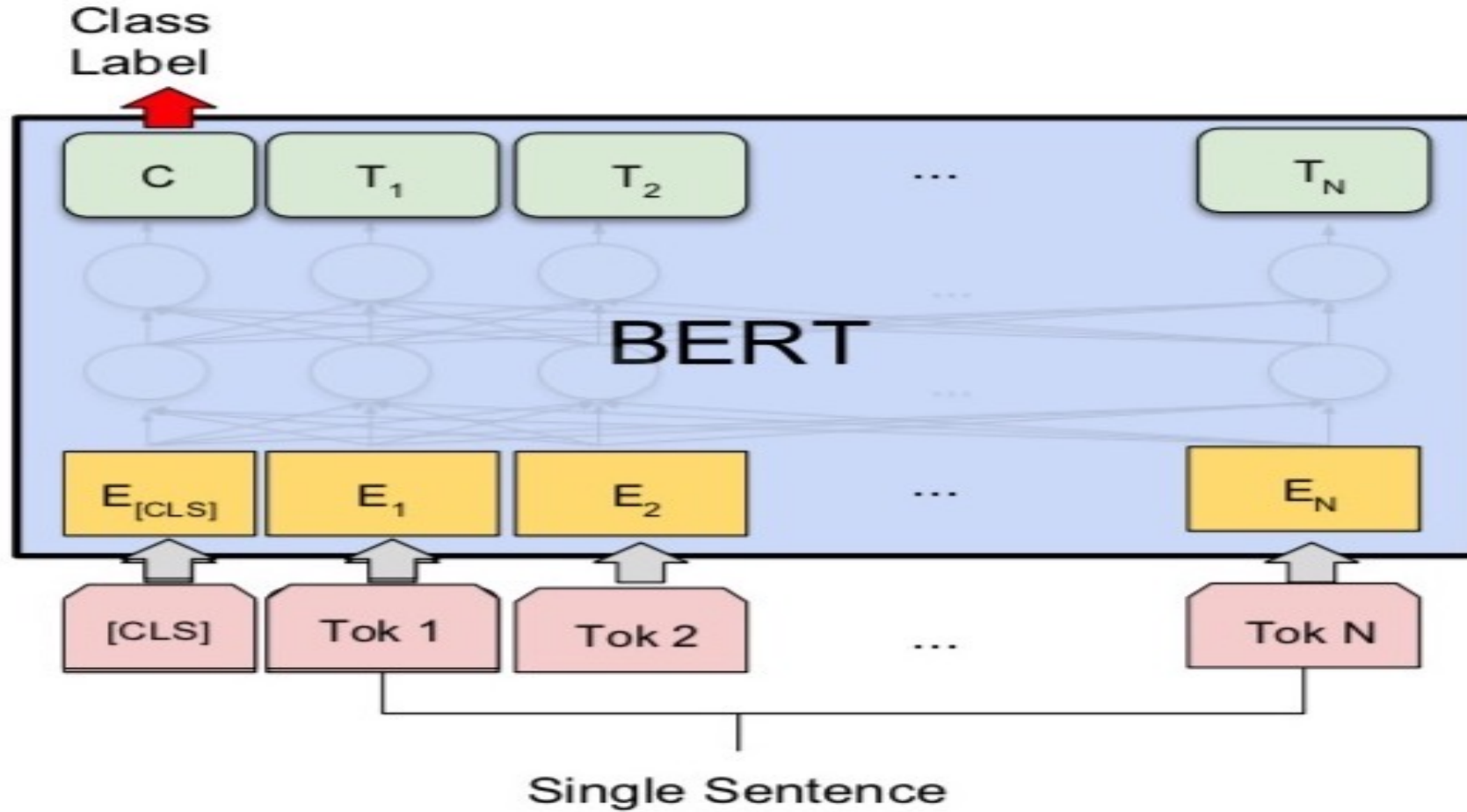
The RAG architecture for fine-tuning a retriever and generator end-to-end



The QA Hierarchy of Needs

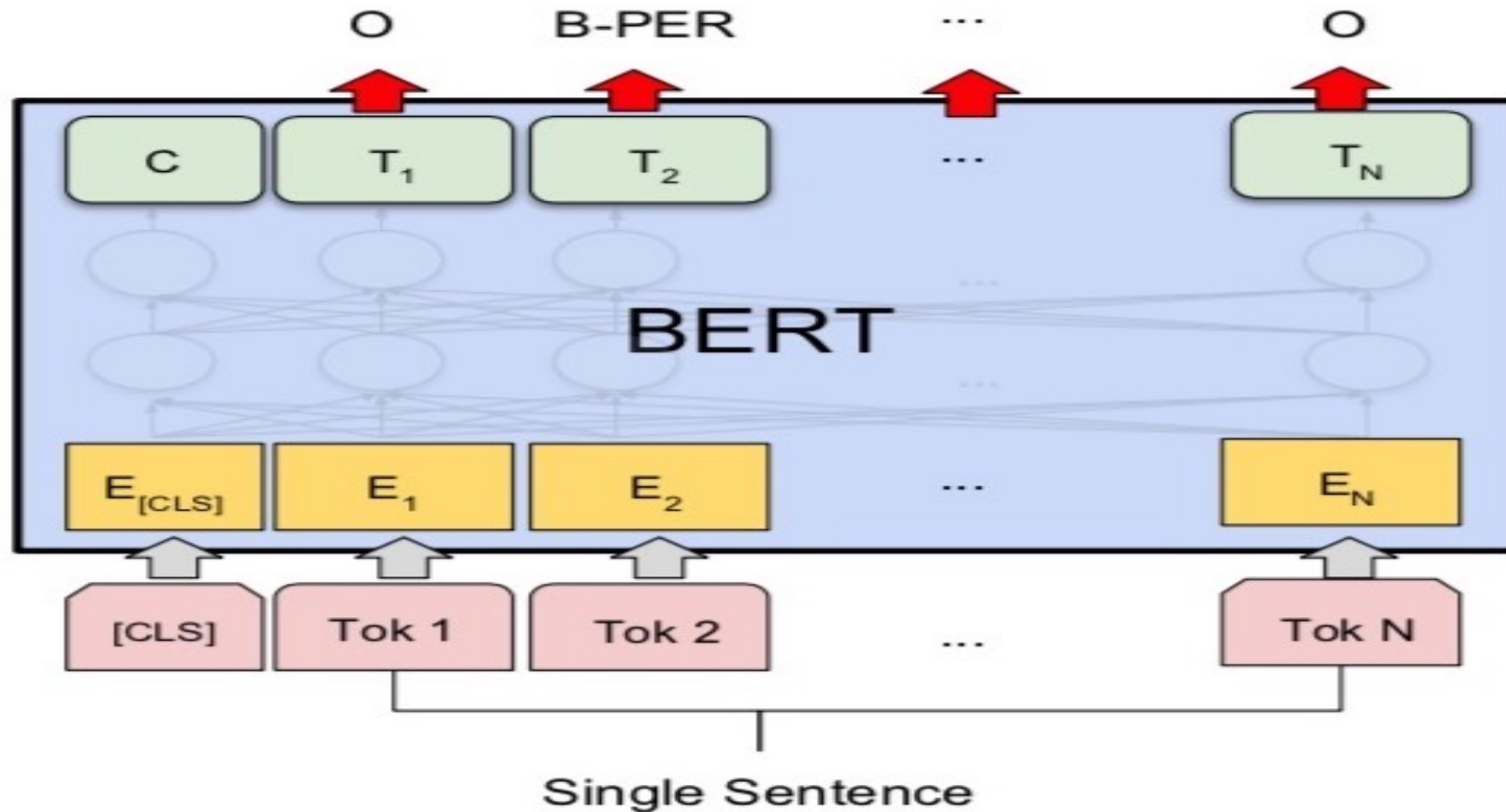


Fine-tuning BERT on Dialogue Intent Detection (ID; Classification)



(b) Single Sentence Classification Tasks:
SST-2, CoLA

Fine-tuning BERT on Dialogue Slot Filling (SF)



(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

Question Answering (QA)

SQuAD

Stanford Question Answering Dataset

SQuAD

SQuAD

Home

Explore 2.0

Explore 1.1

SQuAD2.0

The Stanford Question Answering Dataset

What is SQuAD?

Stanford **Q**uestion **A**nswering **D**ataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable.

SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering.

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Apr 06, 2020	SA-Net on Albert (ensemble) QIANXIN	90.724	93.011
2 May 05, 2020	SA-Net-V2 (ensemble) QIANXIN	90.679	92.948
3	Retro-Reader (ensemble)	90.578	92.978

<https://rajpurkar.github.io/SQuAD-explorer/>

SQuAD

SQuAD: 100,000+ Questions for Machine Comprehension of Text

Pranav Rajpurkar and **Jian Zhang** and **Konstantin Lopyrev** and **Percy Liang**

{pranavs, zjian, klopyrev, pliang}@cs.stanford.edu

Computer Science Department

Stanford University

Abstract

We present the Stanford Question Answering Dataset (SQuAD), a new reading comprehension dataset consisting of 100,000+ questions posed by crowdworkers on a set of Wikipedia articles, where the answer to each question is a segment of text from the corresponding reading passage. We analyze the dataset to understand the types of reasoning required to answer the questions, leaning heavily on dependency and constituency trees. We build a strong logistic regression model, which achieves an F1 score of 51.0%, a significant improvement over a simple baseline (20%). However, human performance (86.8%) is much higher, indicating that the dataset presents a good challenge problem for future research. The dataset is freely available at <https://stanford-qa.com>.

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, **grau-pel** 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?

grau-pel

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

within a cloud

Figure 1: Question-answer pairs for a sample passage in the

Source: Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang.

"Squad: 100,000+ questions for machine comprehension of text." arXiv preprint arXiv:1606.05250 (2016).

SQuAD (Question Answering)

Q: What causes precipitation to fall?

Precipitation

From Wikipedia, the free encyclopedia

For other uses, see [Precipitation \(disambiguation\)](#).

In meteorology, **precipitation** is any product of the condensation of atmospheric water vapor that falls under gravity from clouds.^[2] The main forms of precipitation include drizzle, rain, sleet, snow, ice pellets, graupel and hail. Precipitation occurs when a portion of the atmosphere becomes saturated with water vapor (reaching 100% [relative humidity](#)), so that the water condenses and "precipitates". Thus, fog and mist are not precipitation but suspensions, because the water vapor does not condense sufficiently to precipitate. Two processes, possibly acting together, can lead to air becoming saturated: cooling the air or adding water vapor to the air. 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."**^[3]

SQuAD (Question Answering)

Paragraph

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”.

Q: What causes precipitation to fall?

SQuAD (Question Answering)

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”.

Q: What causes precipitation to fall?

A: gravity

SQuAD (Question Answering)

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”.

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

A: graupel

SQuAD (Question Answering)

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”.

Q: Where do water droplets collide with ice crystals to form precipitation?

A: within a cloud

SQuAD (Question Answering)

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”.

Q: What causes precipitation to fall?

A: gravity

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


A: graupel

Q: Where do water droplets collide with ice crystals to form precipitation?

A: within a cloud

Python in Google Colab (Python101)

<https://colab.research.google.com/drive/1FEG6DnGvwfUbeo4zJ1zTunjMqf2RkCrT>



 python101.ipynb ☆

File Edit View Insert Runtime Tools Help [All changes saved](#)

Comment Share ⚙️ A

+ Code + Text

Question Answering

RAM  Disk  Editing ^

[12]


```
1 from transformers import pipeline
2 qamodel = pipeline("question-answering", model='deepset/roberta-base-squad2')
3 question = "What causes precipitation to fall?"
4 context = """In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravi·
5 output = qamodel(question = question, context = context)
6 print(output['answer'])
```

gravity

[13]

```
1 from transformers import pipeline
2 qamodel = pipeline("question-answering", model='deepset/roberta-base-squad2')
3 question = "What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?"
4 context = """In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravi·
5 output = qamodel(question = question, context = context)
6 print(output['answer'])
```

graupel



```
1 #from transformers import pipeline
2 #qamodel = pipeline("question-answering", model='deepset/roberta-base-squad2')
3 question = "Where do water droplets collide with ice crystals to form precipitation?"
4 context = """In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravi·
5 output = qamodel(question = question, context = context)
6 print(output['answer'])
```

within a cloud

<https://tinyurl.com/aintpupython101>

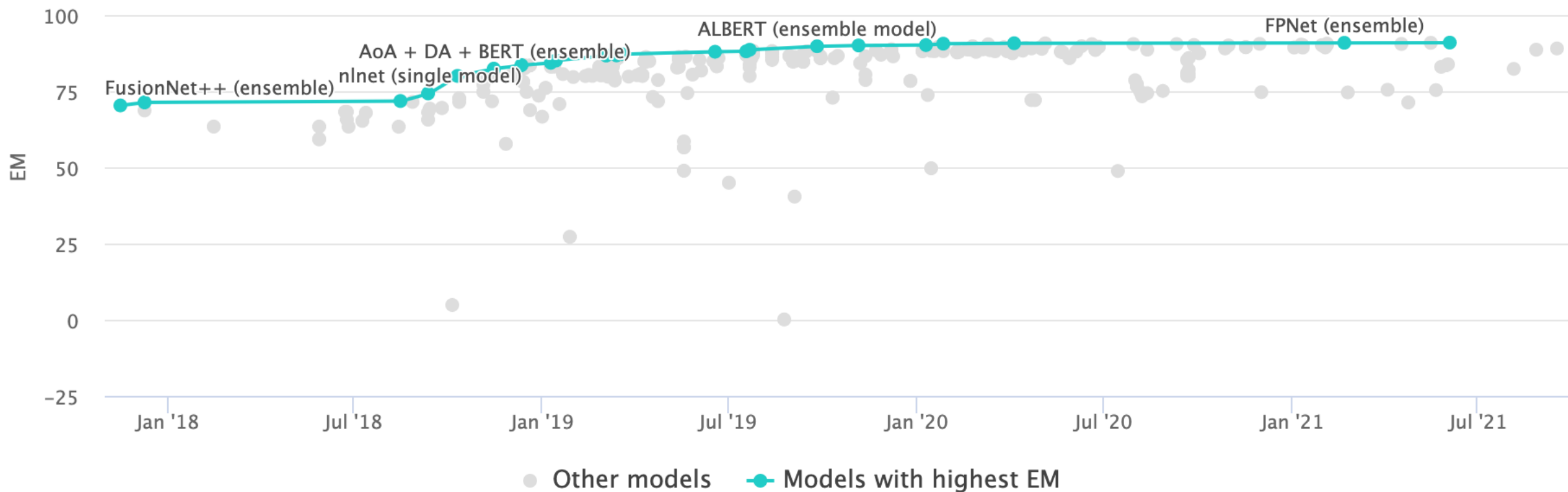
Question Answering

```
from transformers import pipeline
qamodel = pipeline("question-answering", model='deepset/roberta-base-squad2')
question = "What causes precipitation to fall?"
context = """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"."""
output = qamodel(question = question, context = context)
print(output['answer'])
```

gravity

Question Answering on SQuAD2.0

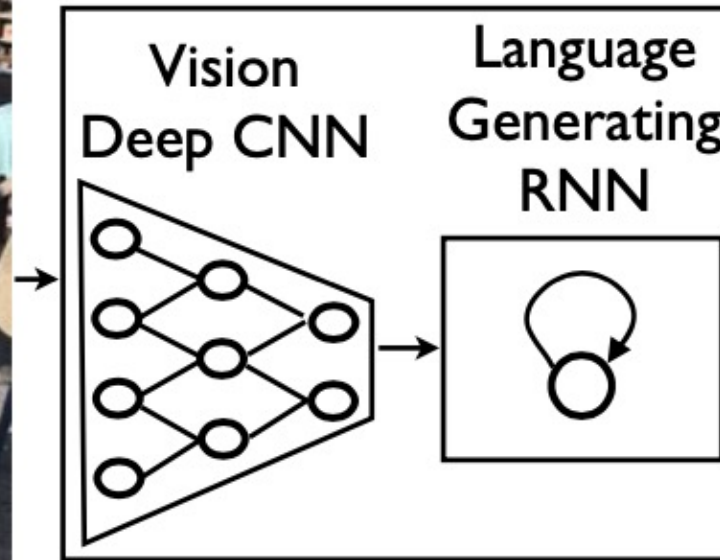
SQuAD 2.0 benchmark (Papers with Code)



<https://paperswithcode.com/sota/question-answering-on-squad20>

Neural Image Captioning (NIC)

image-to-text description generation

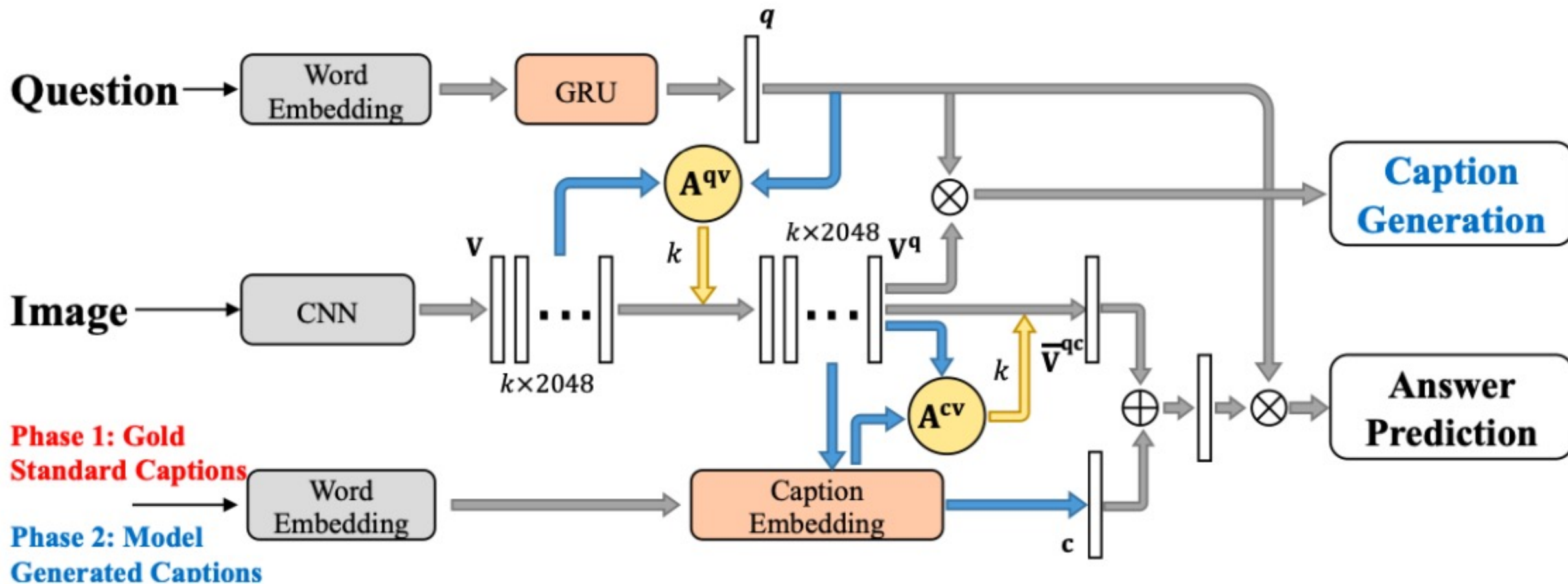


A group of people shopping at an outdoor market.

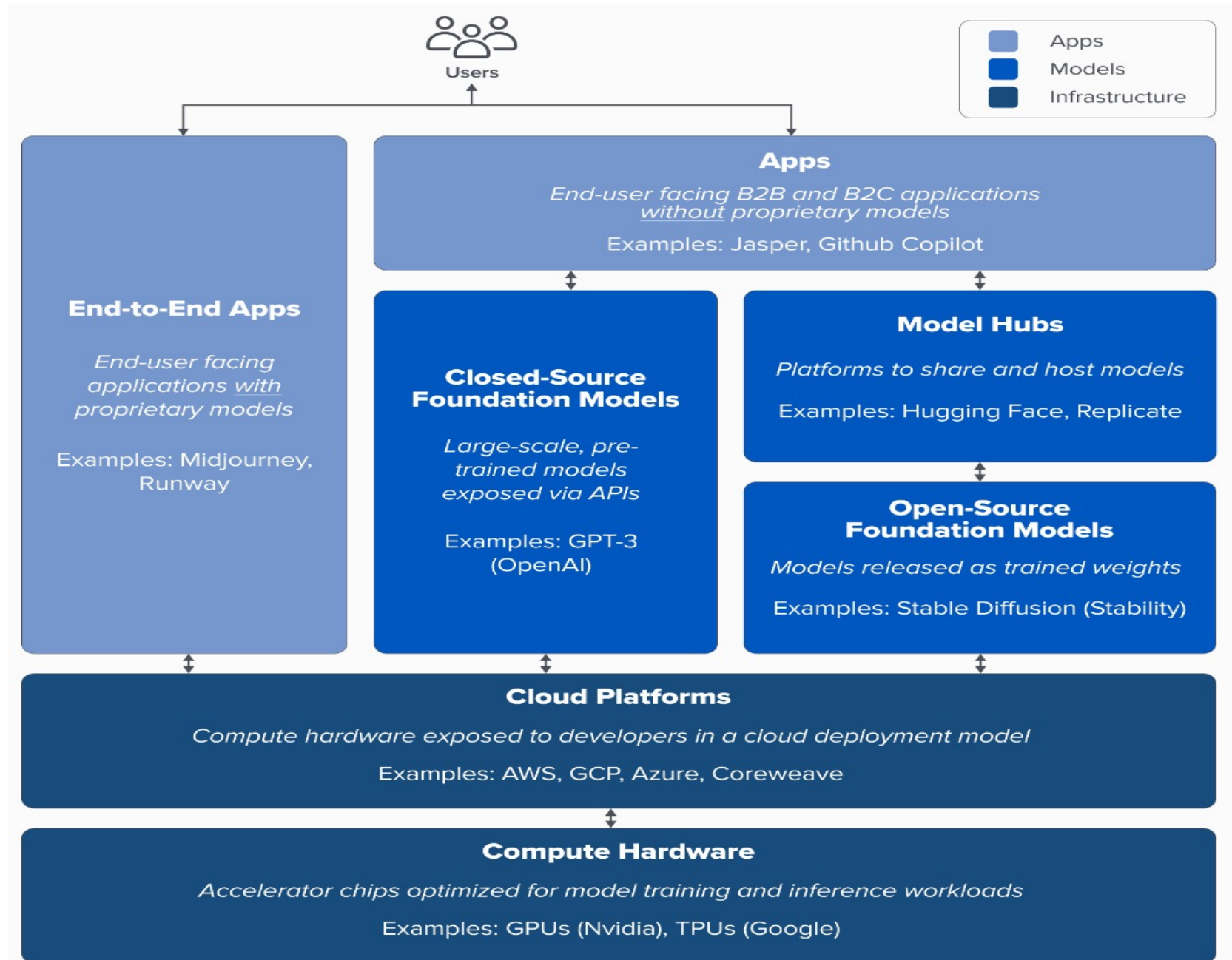
There are many vegetables at the fruit stand.

Visual Question Answering

Neural caption generation is employed to aid answer prediction



Generative AI Tech Stack

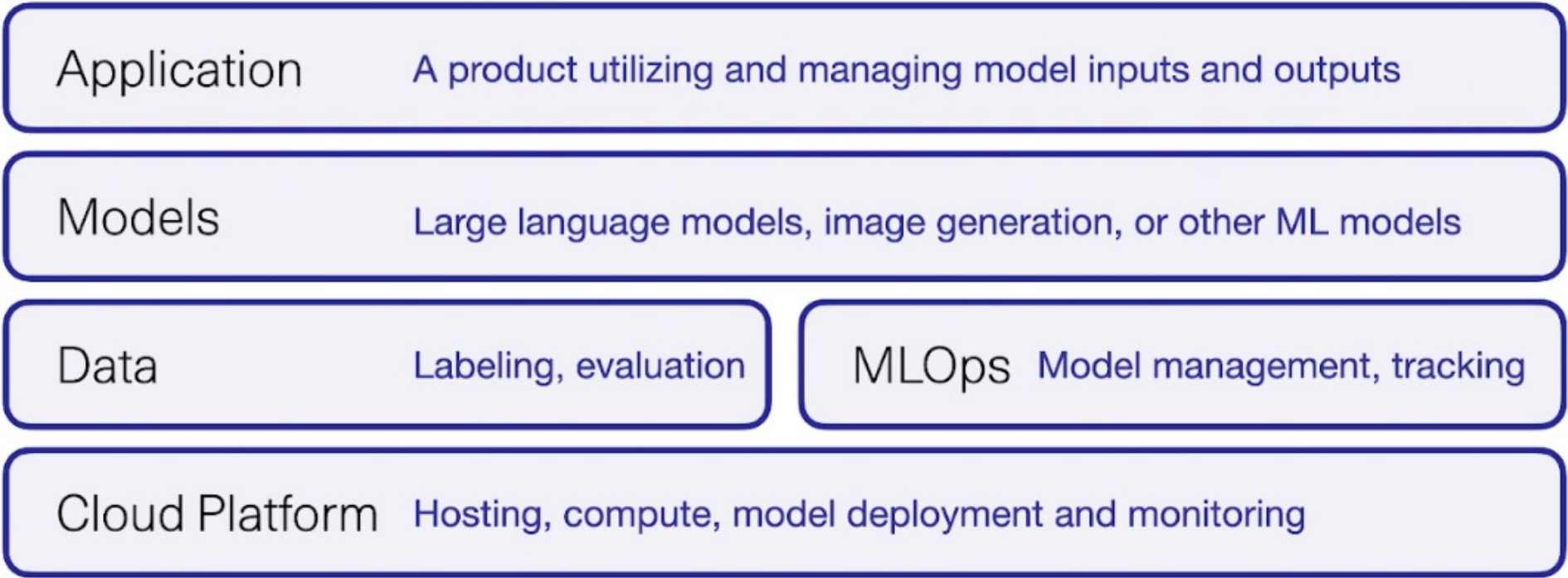


Generative AI Software and Business Factors

Business
Factors



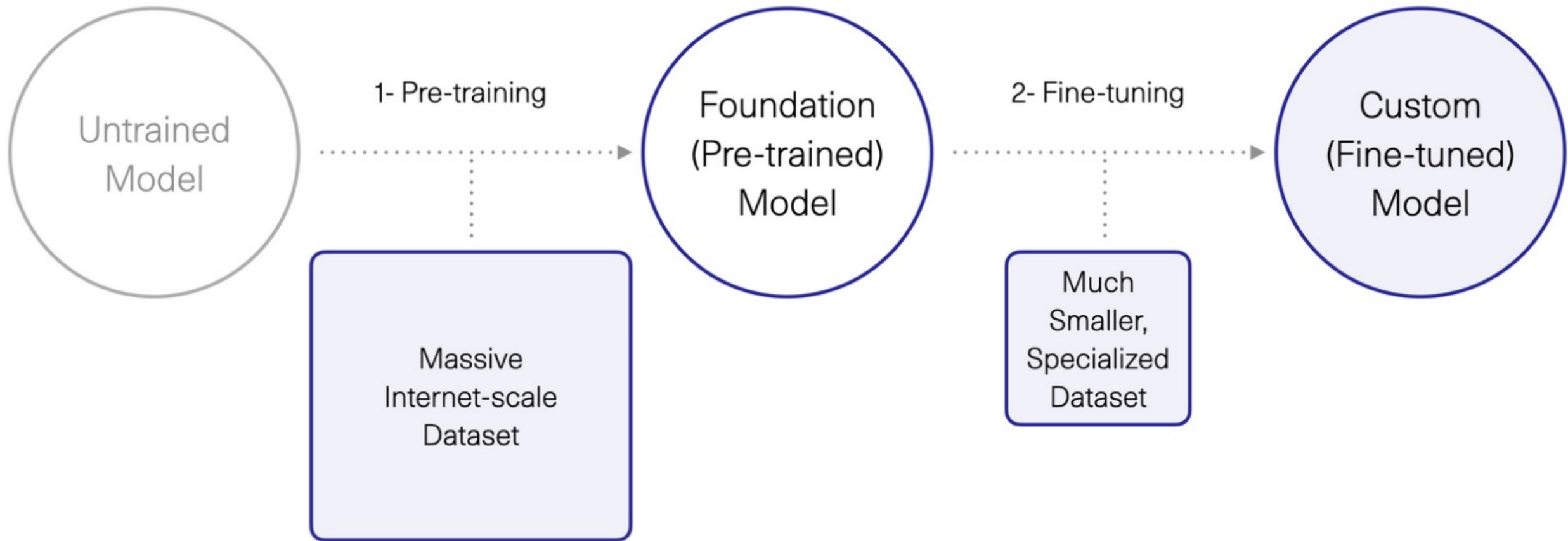
Software



Generative AI

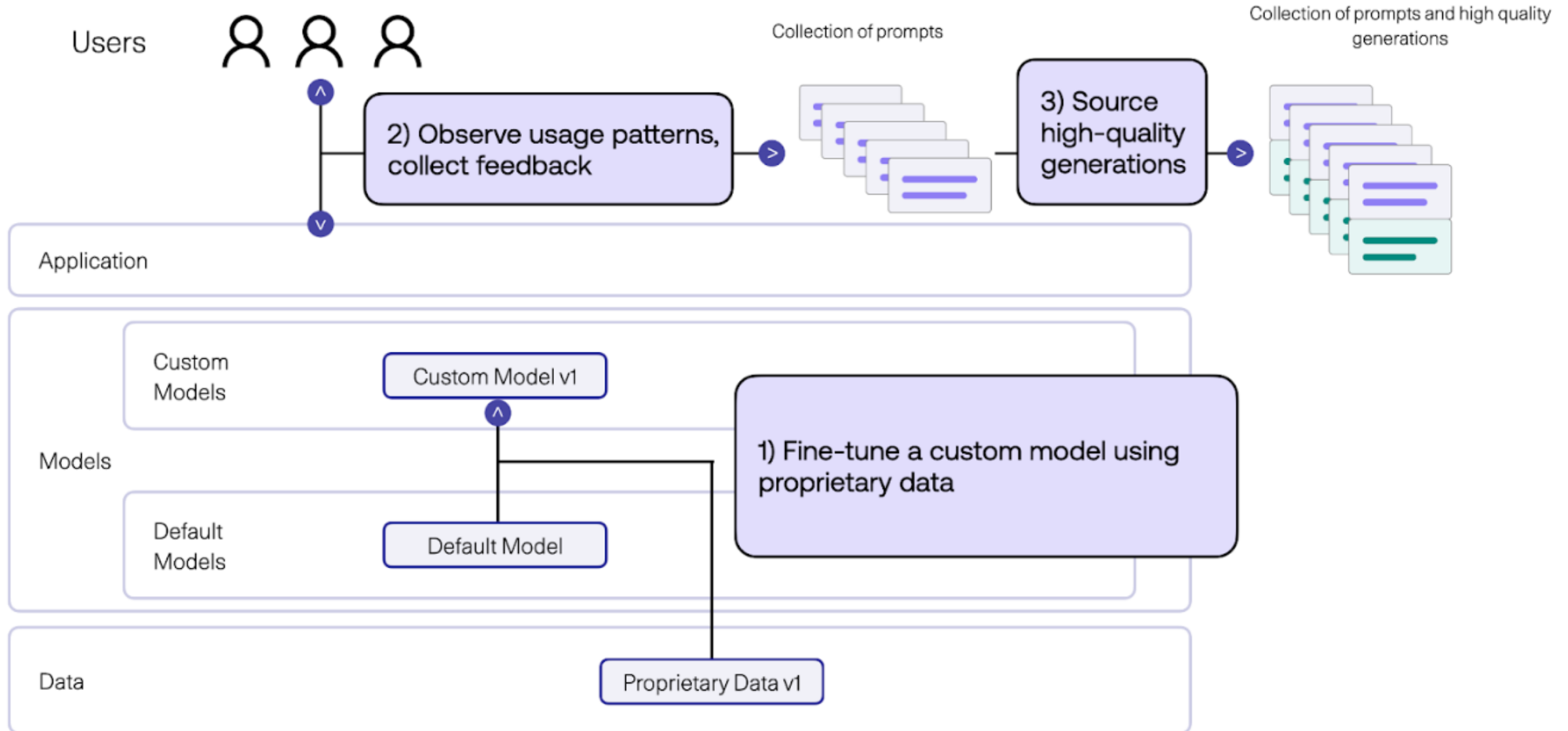
1. Pre-training Foundation (Pre-trained) Model

2. Fine-tuning Custom (Fine-tuned) Model



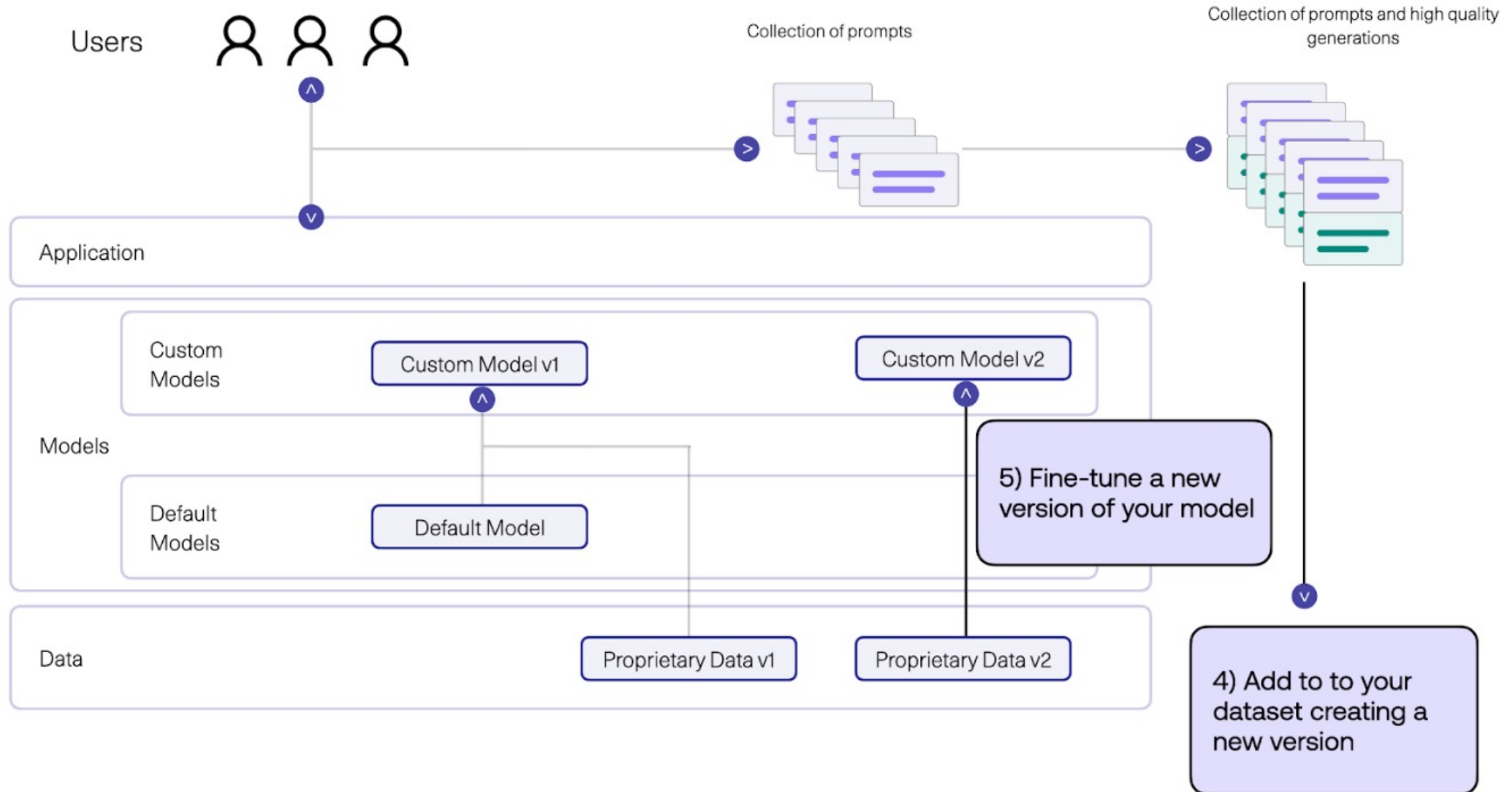
Generative AI

Fine-tune Custom Models using Proprietary Data

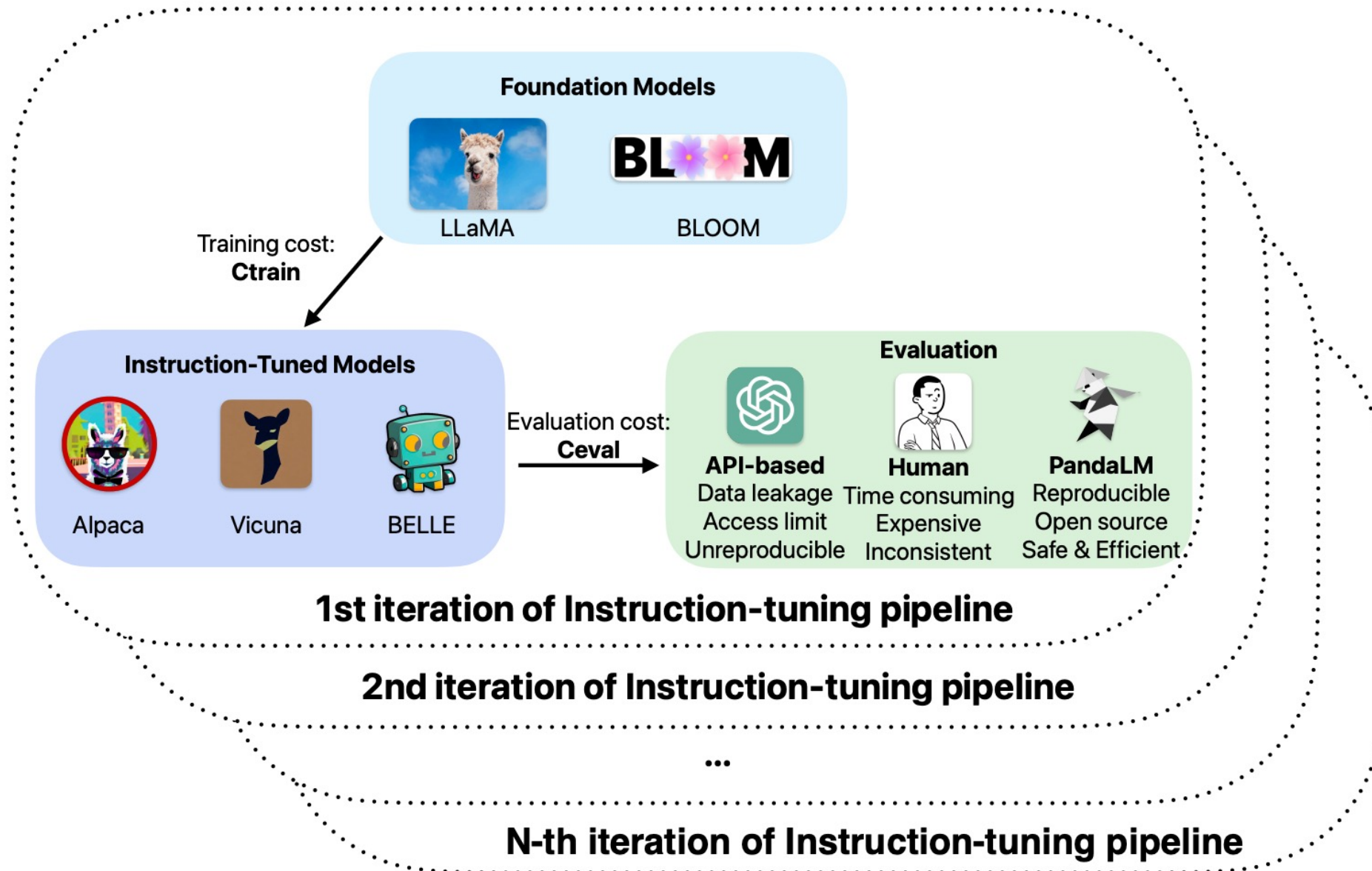


Generative AI

Fine-tune Custom Models using Proprietary Data



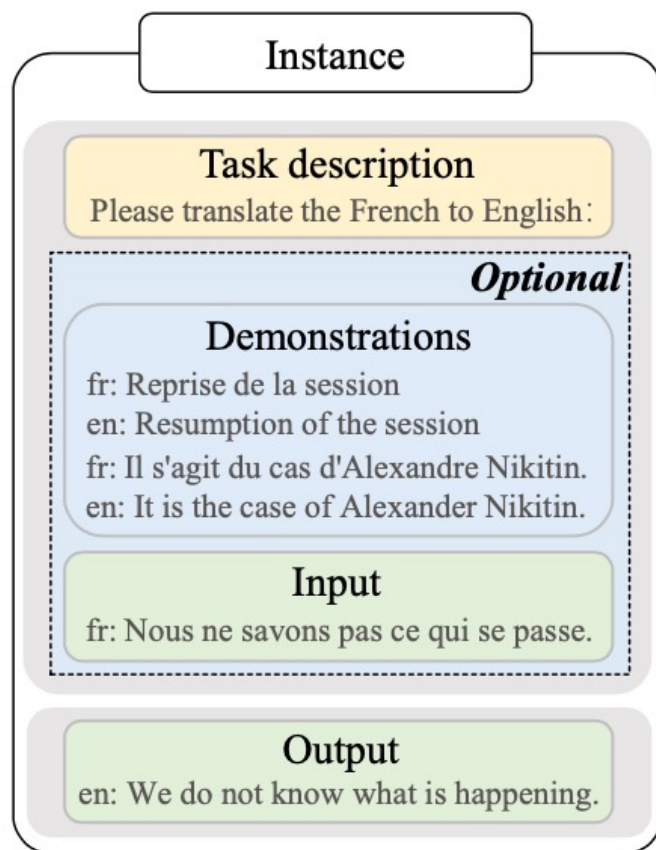
Pipeline of Instruction Tuning LLMs.



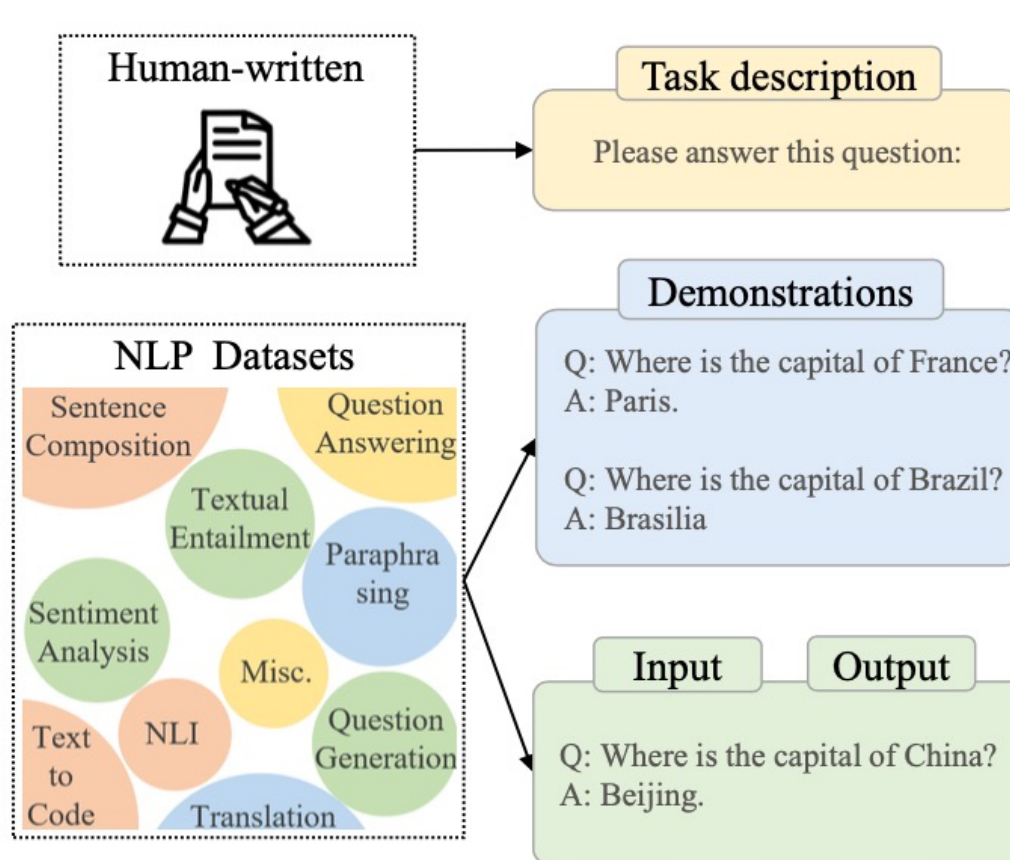
Available Task Collections for Instruction Tuning

Collections	Time	#Task types	#Tasks	#Examples
Nat. Inst. [193]	Apr-2021	6	61	193K
CrossFit [194]	Apr-2021	13	160	7.1M
FLAN [62]	Sep-2021	12	62	4.4M
P3 [195]	Oct-2021	13	267	12.1M
ExMix [196]	Nov-2021	11	107	18M
UnifiedSKG [197]	Jan-2022	6	21	812K
Super Nat. Inst. [78]	Apr-2022	76	1616	5M
MVPCorpus [198]	Jun-2022	11	77	41M
xP3 [84]	Nov-2022	17	85	81M
OIG ¹⁴	Mar-2023	-	-	43M

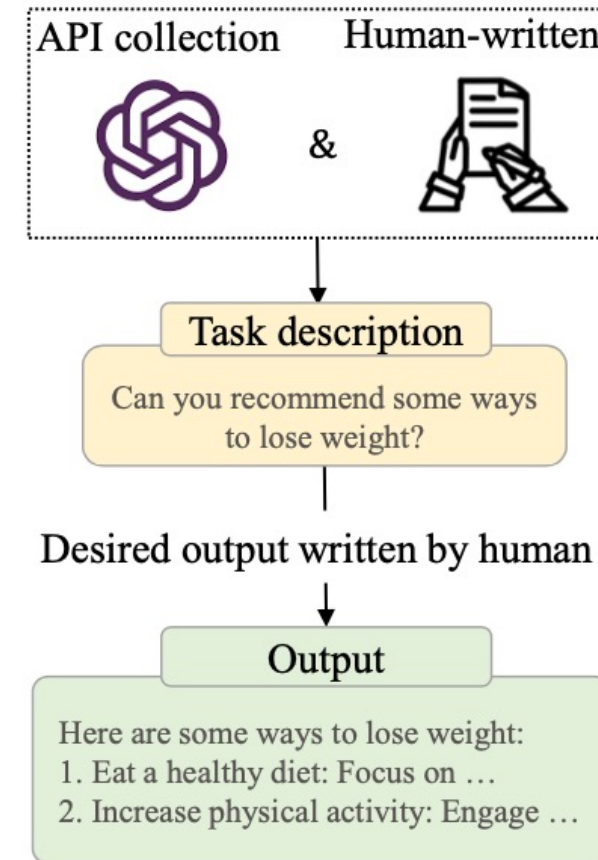
Instance Formatting and Two Different Methods for Constructing the Instruction-formatted Instances



(a) Instance format



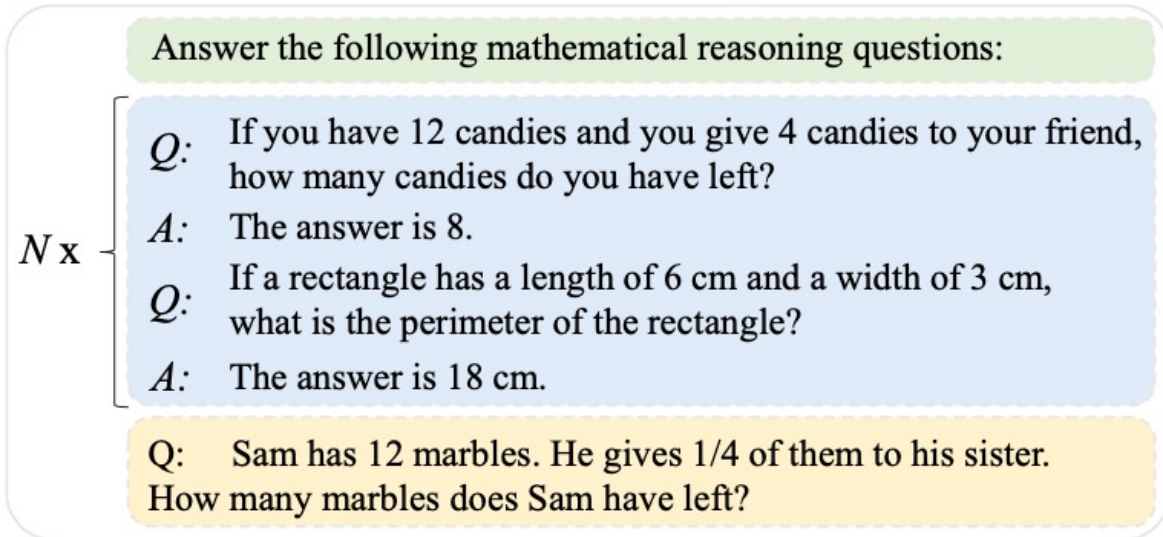
(b) Formatting existing datasets



(c) Formatting human needs

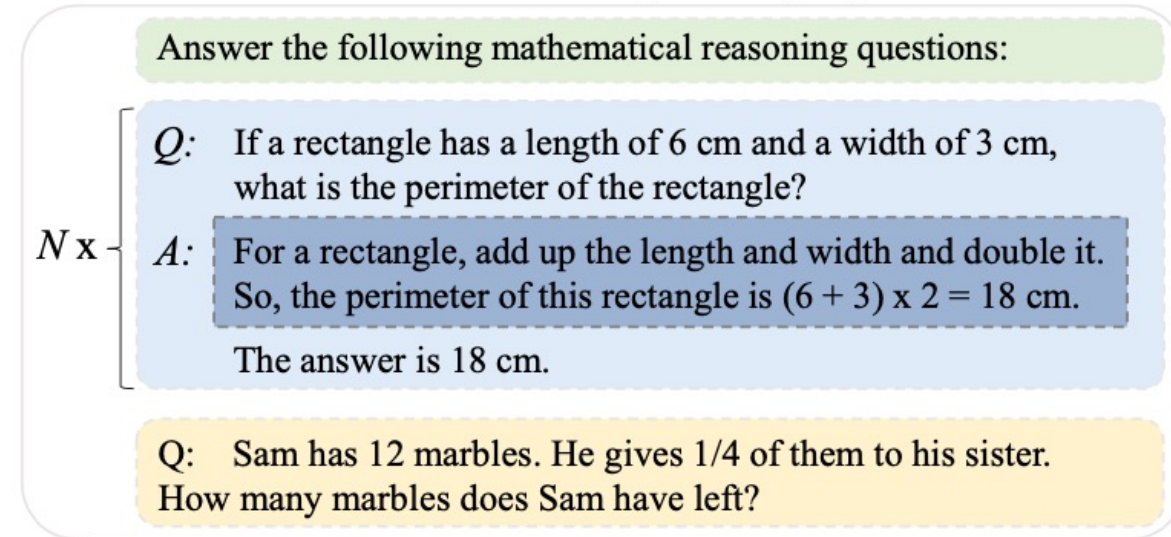
In-context Learning (ICL) and Chain-of-thought (CoT) Prompting

In-Context Learning



A: The answer is 9.

Chain-of-Thought Prompting





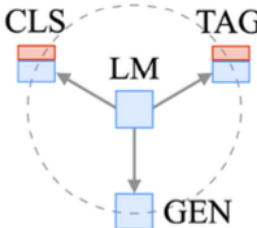
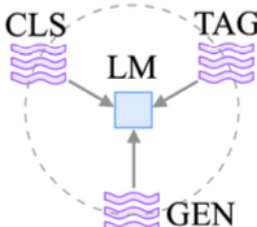
A: He gives $(1 / 4) \times 12 = 3$ marbles. So Sam is left with $12 - 3 = 9$ marbles. The answer is 9.

LLM

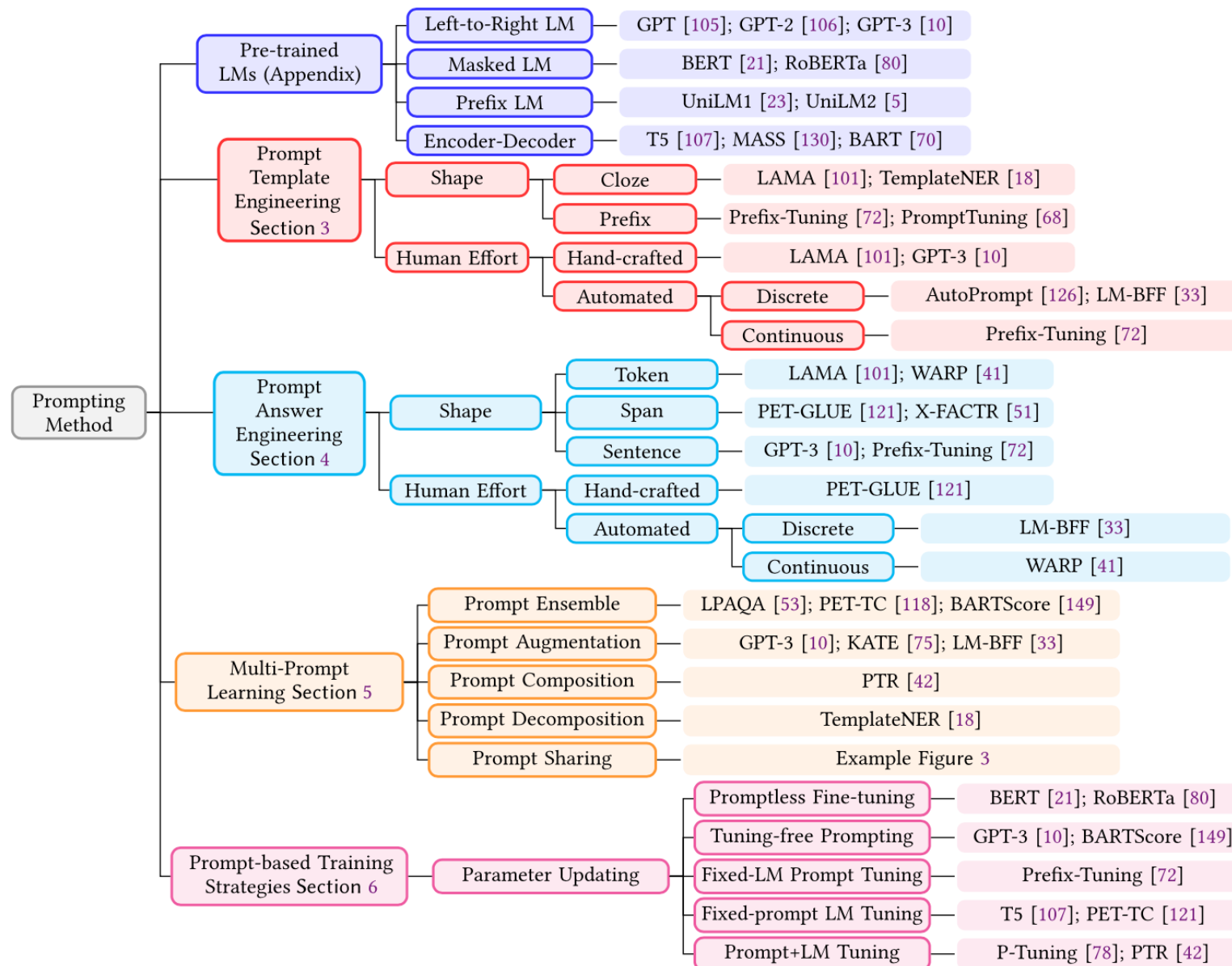
 : Task description  : Demonstration  : Chain-of-Thought  : Query

Pre-train, Prompt, and Predict: Prompting Methods in Natural Language Processing (LLMs)

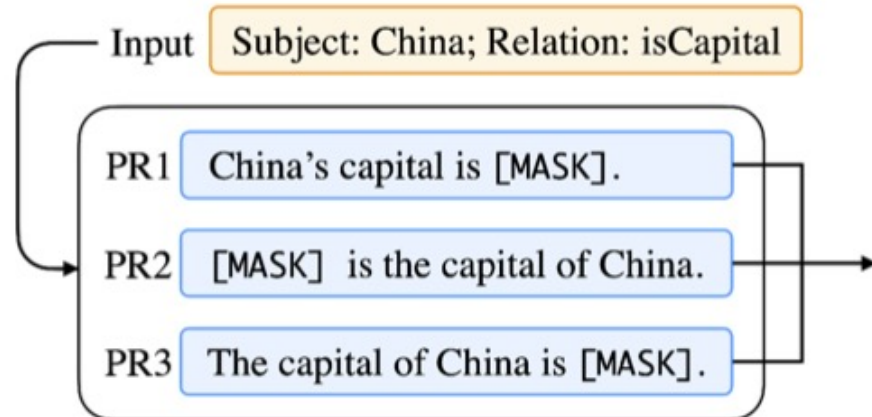
Four Paradigms in NLP

Paradigm	Engineering	Task Relation
a. Fully Supervised Learning (Non-Neural Network)	Feature (e.g. word identity, part-of-speech, sentence length)	
b. Fully Supervised Learning (Neural Network)	Architecture (e.g. convolutional, recurrent, self-attentional)	
c. Pre-train, Fine-tune	Objective (e.g. masked language modeling, next sentence prediction)	
d. Pre-train, Prompt, Predict	Prompt (e.g. cloze, prefix)	

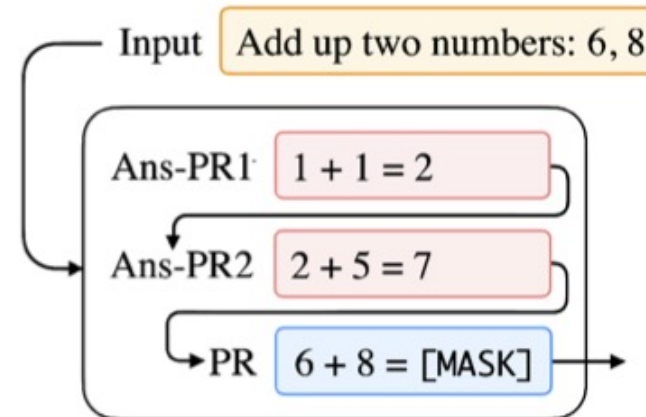
Typology of Prompting Methods



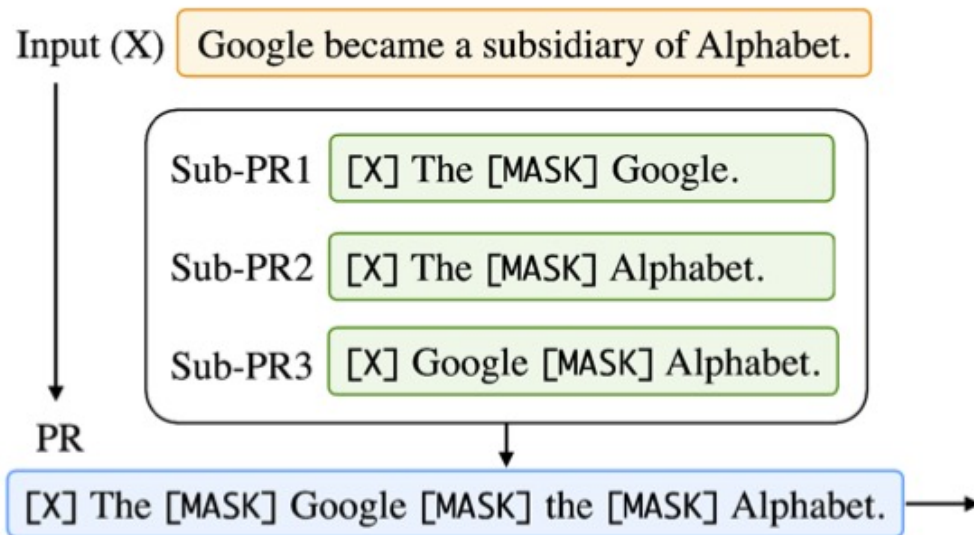
Different Multi-Prompt Learning Strategies



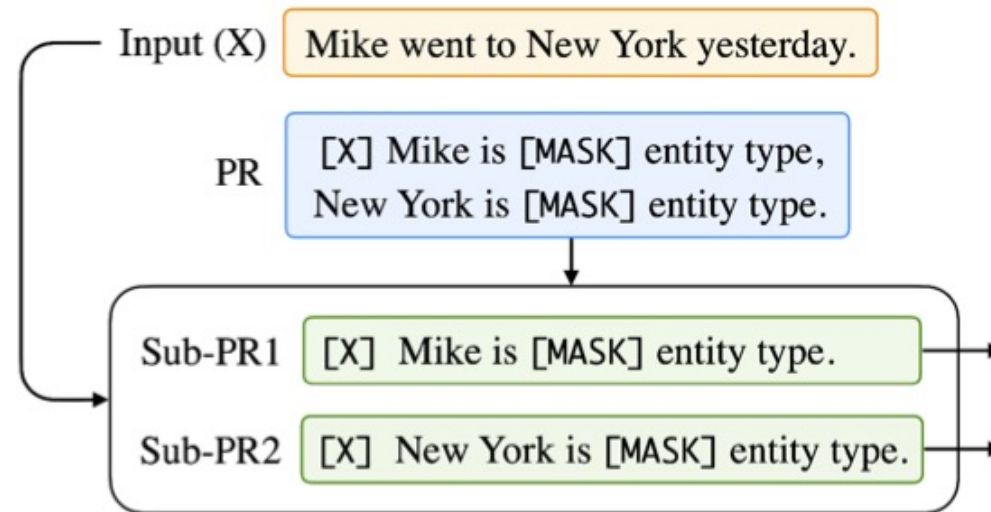
(a) Prompt Ensembling.



(b) Prompt Augmentation.



(c) Prompt Composition.



(d) Prompt Decomposition.

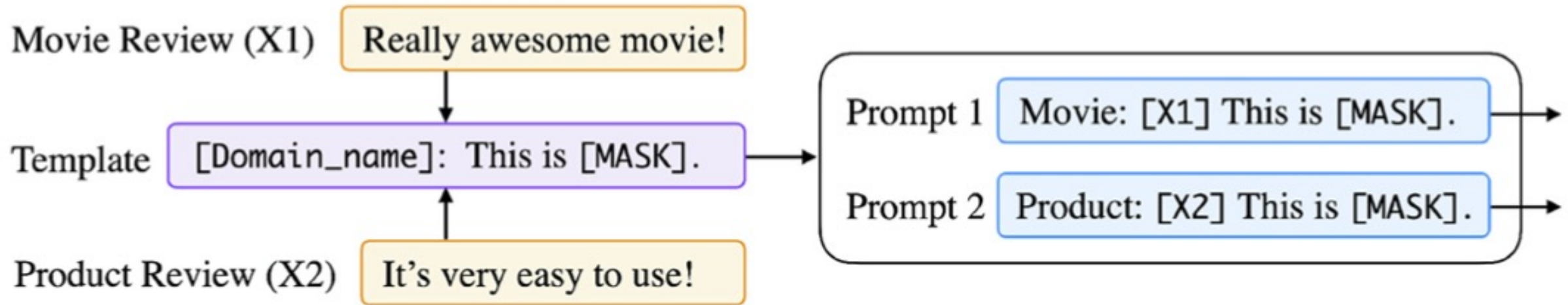
Examples of Input, Template, and Answer for Different Tasks

Type	Task Example	Input ([X])	Template	Answer ([Z])
Text Classification	Sentiment	I love this movie.	[X] The movie is [Z].	great fantastic ...
	Topics	He prompted the LM.	[X] The text is about [Z].	sports science ...
	Intention	What is taxi fare to Denver?	[X] The question is about [Z].	quantity city ...
Text-span Classification	Aspect Sentiment	Poor service but good food.	[X] What about service? [Z].	Bad Terrible ...
Text-pair Classification	Natural Language Inference	[X1]: An old man with ... [X2]: A man walks ...	[X1]? [Z], [X2]	Yes No ...
Tagging	Named Entity Recognition	[X1]: Mike went to Paris. [X2]: Paris	[X1][X2] is a [Z] entity.	organization location ...
Text Generation	Summarization	Las Vegas police ...	[X] TL;DR: [Z]	The victim ... A woman
	Translation	Je vous aime.	French: [X] English: [Z]	I love you. I fancy you. ...
Regression	Textual Similarity	[X1]: A man is smoking. [X2]: A man is skating.	[X1] [Z], [X2]	Yes No ...

Characteristics of Different Tuning Strategies

Strategy	LM Params	Prompt Params		Example
		Additional	Tuned	
Promptless Fine-tuning	Tuned	—		ELMo [97], BERT [20], BART [69]
Tuning-free Prompting	Frozen	✗	✗	GPT-3 [9], AutoPrompt [125], LAMA [100]
Fixed-LM Prompt Tuning	Frozen	✓	Tuned	Prefix-Tuning [71], Prompt-Tuning [67]
Fixed-prompt LM Tuning	Tuned	✗	✗	PET-TC [117], PET-Gen [118], LM-BFF [32]
Prompt+LM Fine-tuning	Tuned	✓	Tuned	PADA [5], P-Tuning [77], PTR [41]

Multi-prompt Learning for Multi-task, Multi-domain, or Multi-lingual Learning



GPT-3.5 Prompt Engineering for Question Answering

- **Find the answer to the question from the given context.**
- **When the question cannot be answered with the given context, say "unanswerable".**
- **Just say the answer without repeating the question.**
- **Context: {context}**
- **Question:{question}**
- **Answer:**

Prompts and QA Inference For FLAN T5

Question Answering

- Prompts and QA Inference For FLAN T5, we follow [41] and use the following prompt:
- Context: {context}\nQuestion: {question}\nAnswer:
- Context: {context}
- Question: {question}
- Answer:

Prompt Engineering

- **Prompts For FLAN models**

- Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., ... & Roberts, A. (2023). The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

- **MNLI, NLI-FEVER, VitaminC:**

- "Premise: {premise}\n\nHypothesis: {hypothesis}\n\nDoes the premise entail the hypothesis?\n\nA yes\nB it is not possible to tell\nC no"

- **ANLI:**

- "{context}\n\nBased on the paragraph above can we conclude that \"{hypothesis}\"?\n\nA Yes\nB It's impossible to say\nC No"

- **SNLI:**

- "If \"{premise}\", does this mean that \"{hypothesis}\"?\n\nA yes\nB it is not possible to tell\nC no"

Outline

- Introduction
- Overview of Generative AI
- Overview of Large Language Models (LLMs)
- Foundation of Transformers: Attention Mechanism
- Fine-tuning LLM for Question Answering System
- **Fine-tuning LLM for Dialogue System**
- Challenges and Limitations of Generative AI for QA and Dialogue Systems
- Q & A

Fine-tuning LLM for Dialogue System

Reinforcement Learning from Human Feedback (RLHF)

**ChatGPT:
Optimizing Language Models for Dialogue**

Dialogue Systems

Chatbot
Dialogue System
Intelligent Agent

Dialogue Subtasks

Browse SoTA > Natural Language Processing > Dialogue

Dialogue subtasks

Dialogue Generation

Dialogue Generation

📊 9 benchmarks

78 papers with code



Dialogue State Tracking

📊 2 benchmarks

51 papers with code

Task-Oriented Dialogue Systems

Task-Oriented Dialogue Systems

📊 2 benchmarks

48 papers with code



Visual Dialog

📊 8 benchmarks

37 papers with code



Goal-Oriented Dialogue

📊 1 benchmark

20 papers with code



Dialogue Management

12 papers with code



Dialogue Understanding

📊 11 benchmarks

8 papers with code



Dialogue Act Classification

📊 2 benchmarks

8 papers with code

Short-Text Conversation

Short-Text Conversation

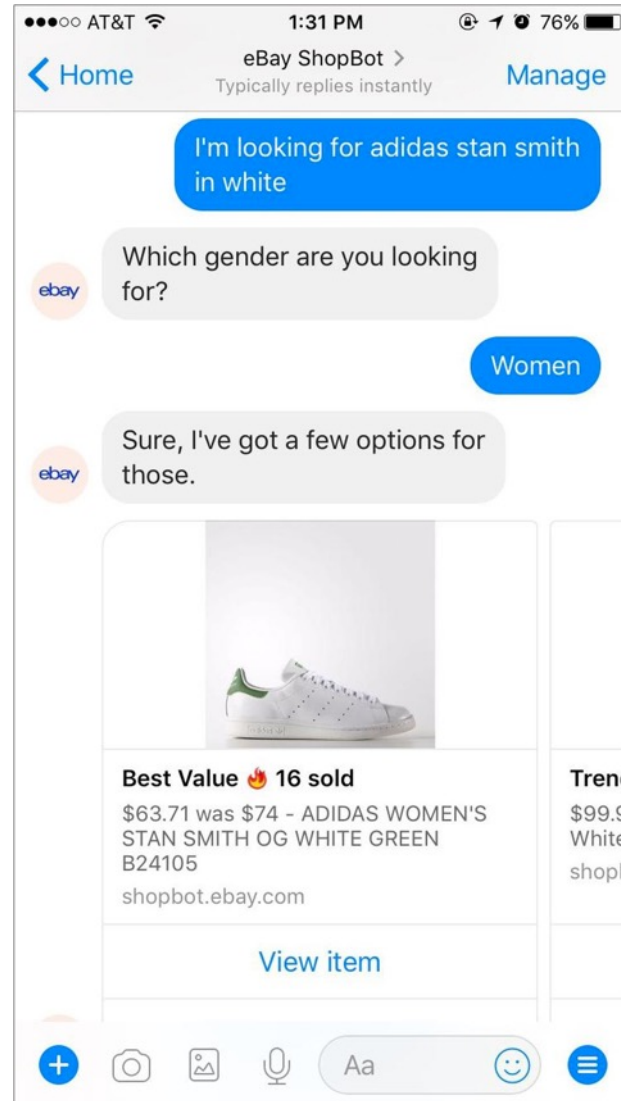
7 papers with code



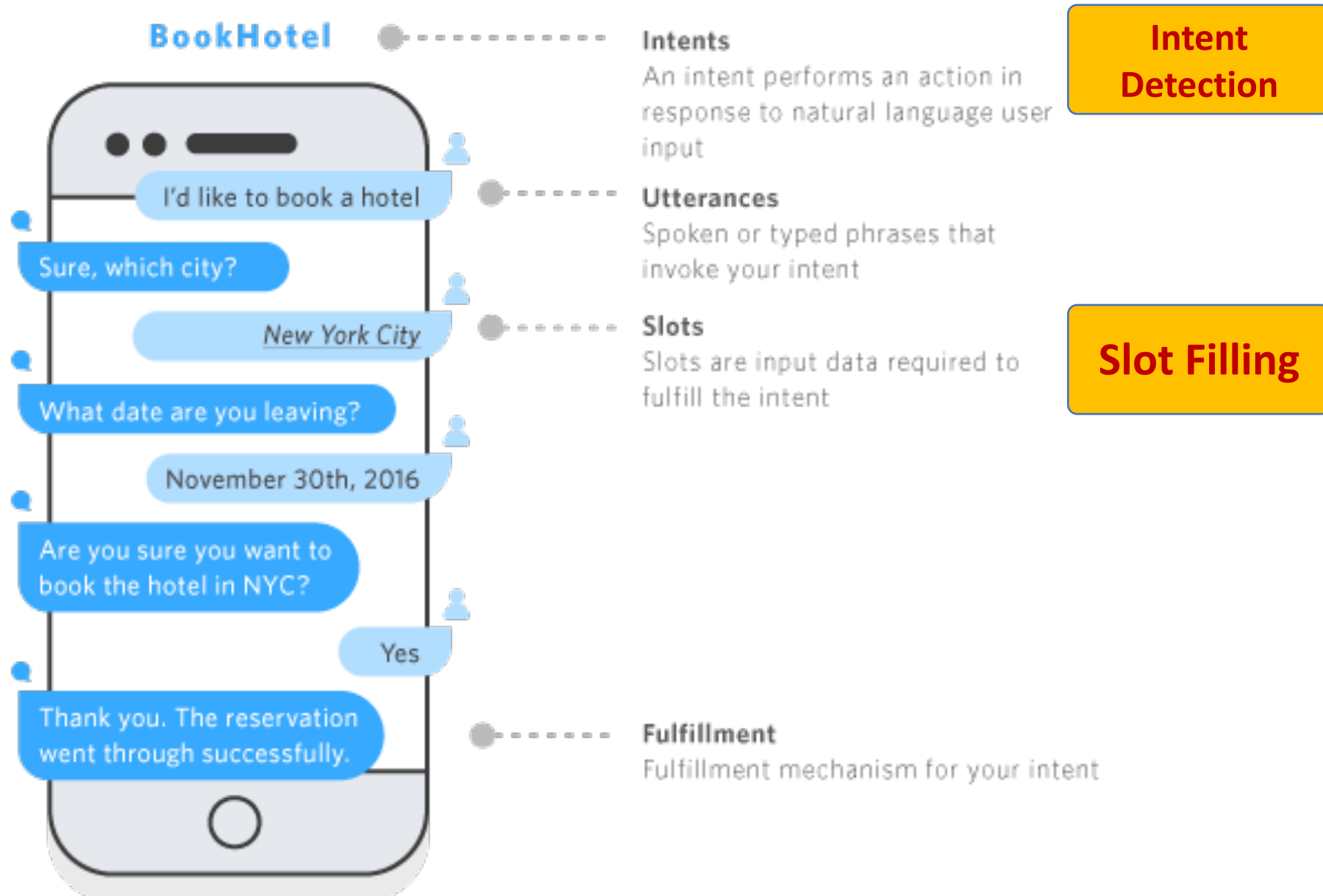
Goal-Oriented Dialogue Systems

7 papers with code

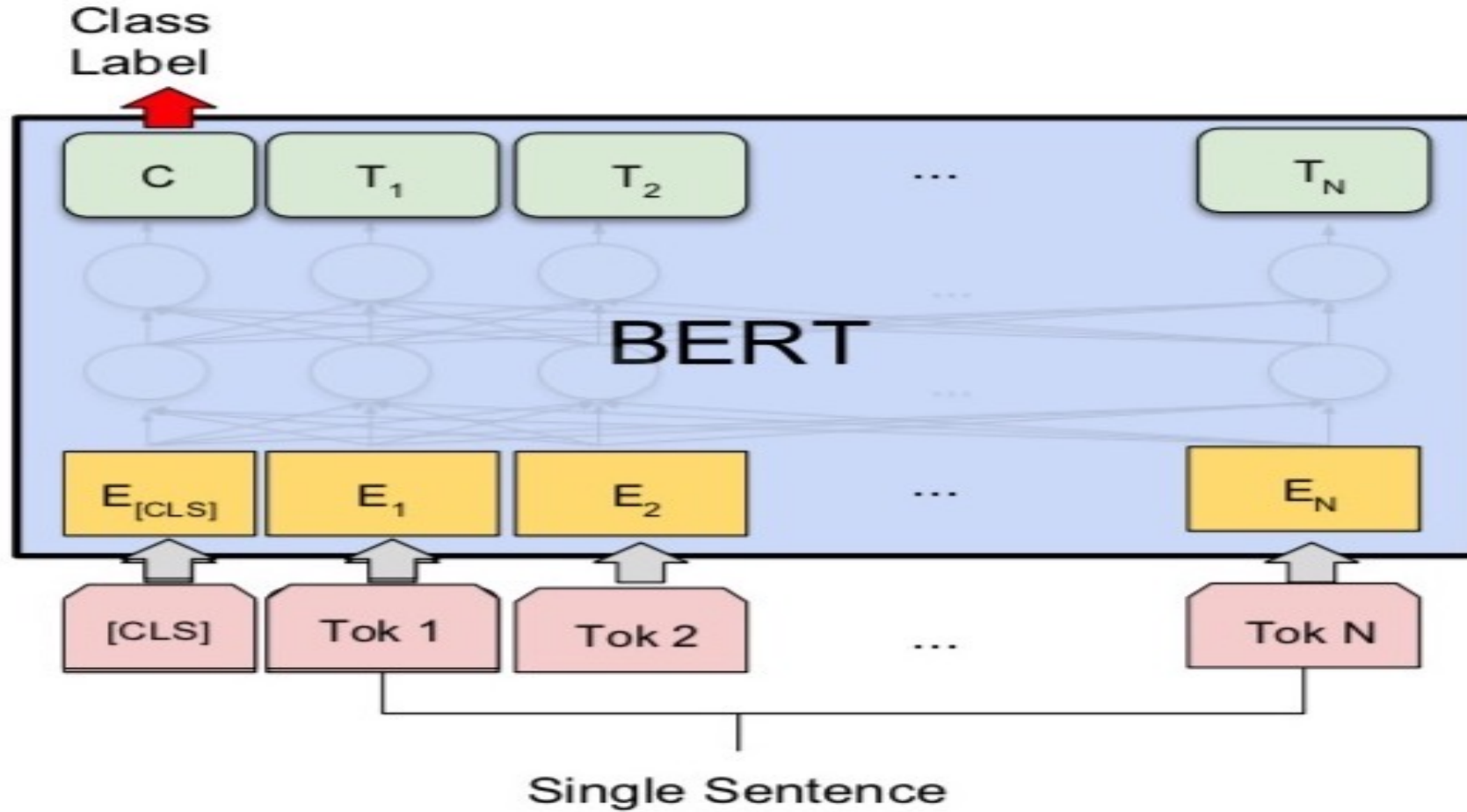
Conversational Commerce: eBay AI Chatbots



Hotel Chatbot

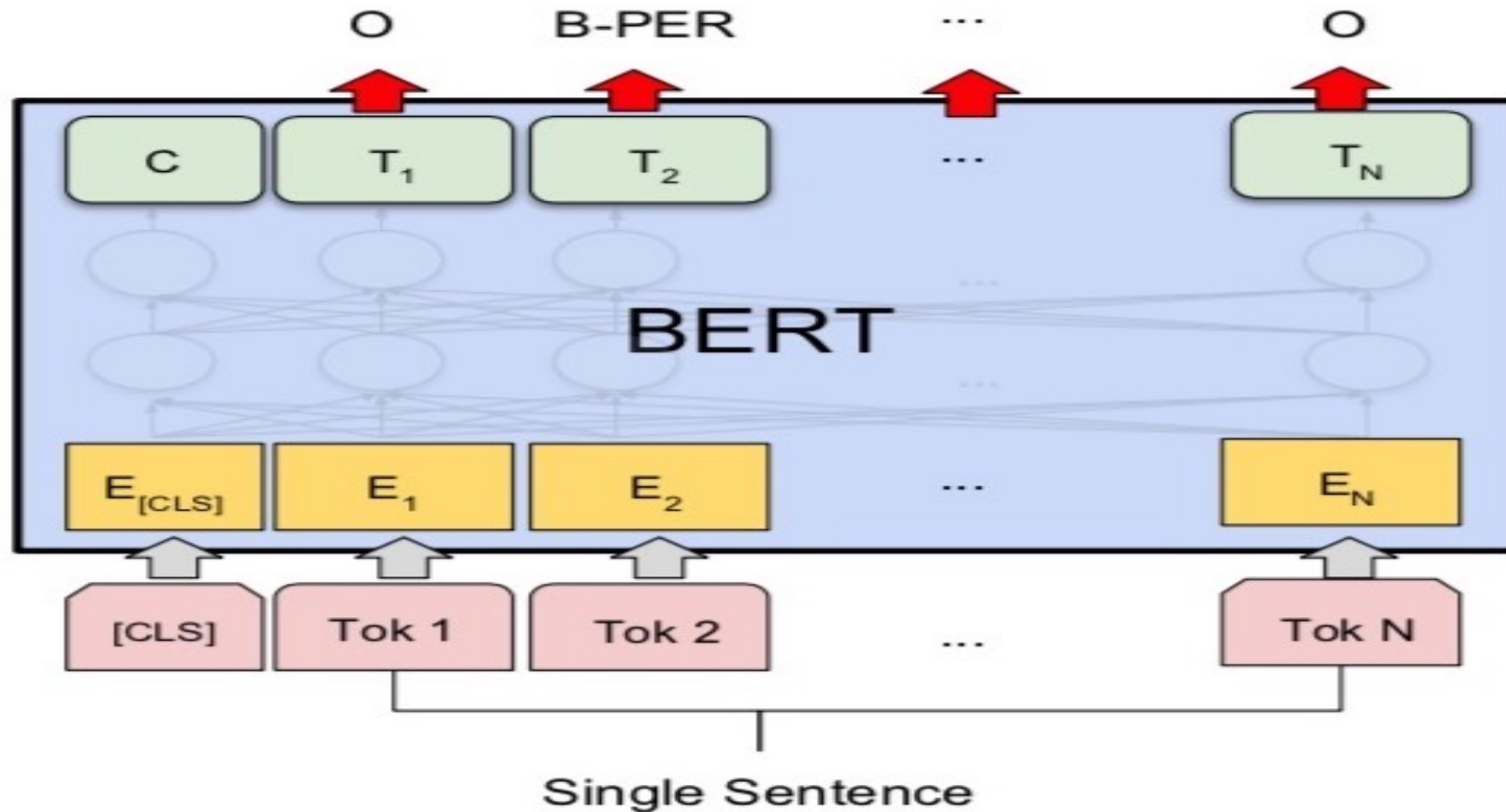


Fine-tuning BERT on Dialogue Intent Detection (ID; Classification)



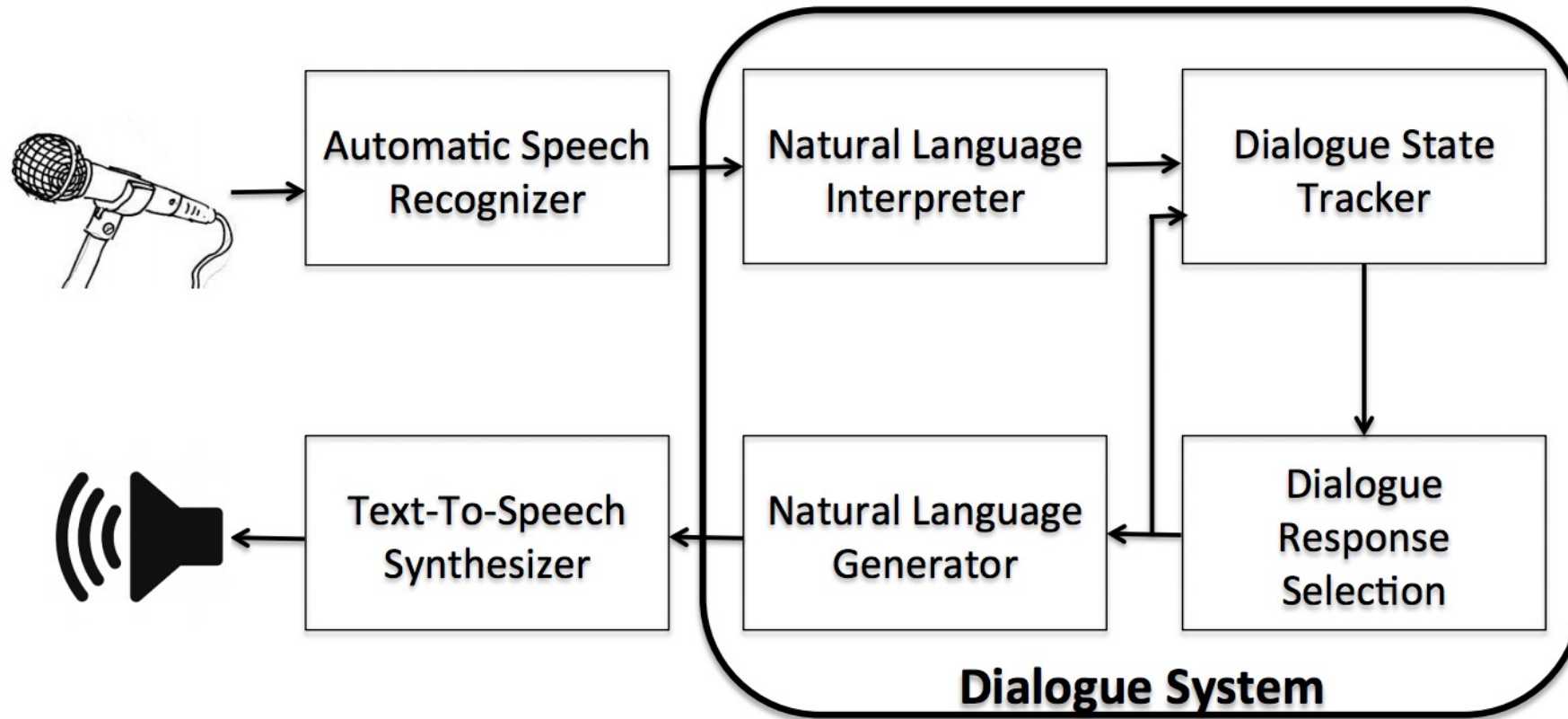
(b) Single Sentence Classification Tasks:
SST-2, CoLA

Fine-tuning BERT on Dialogue Slot Filling (SF)

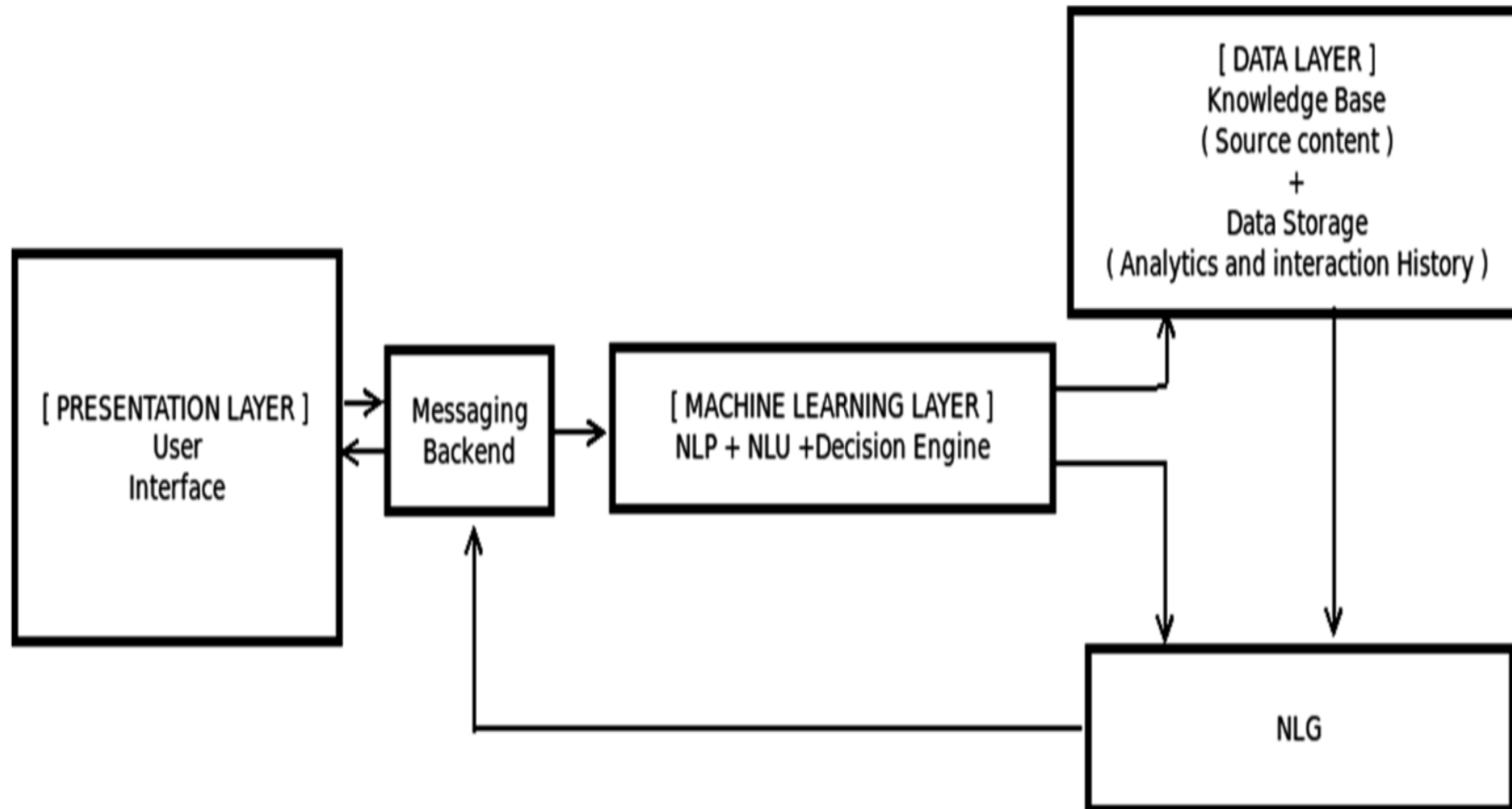


(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

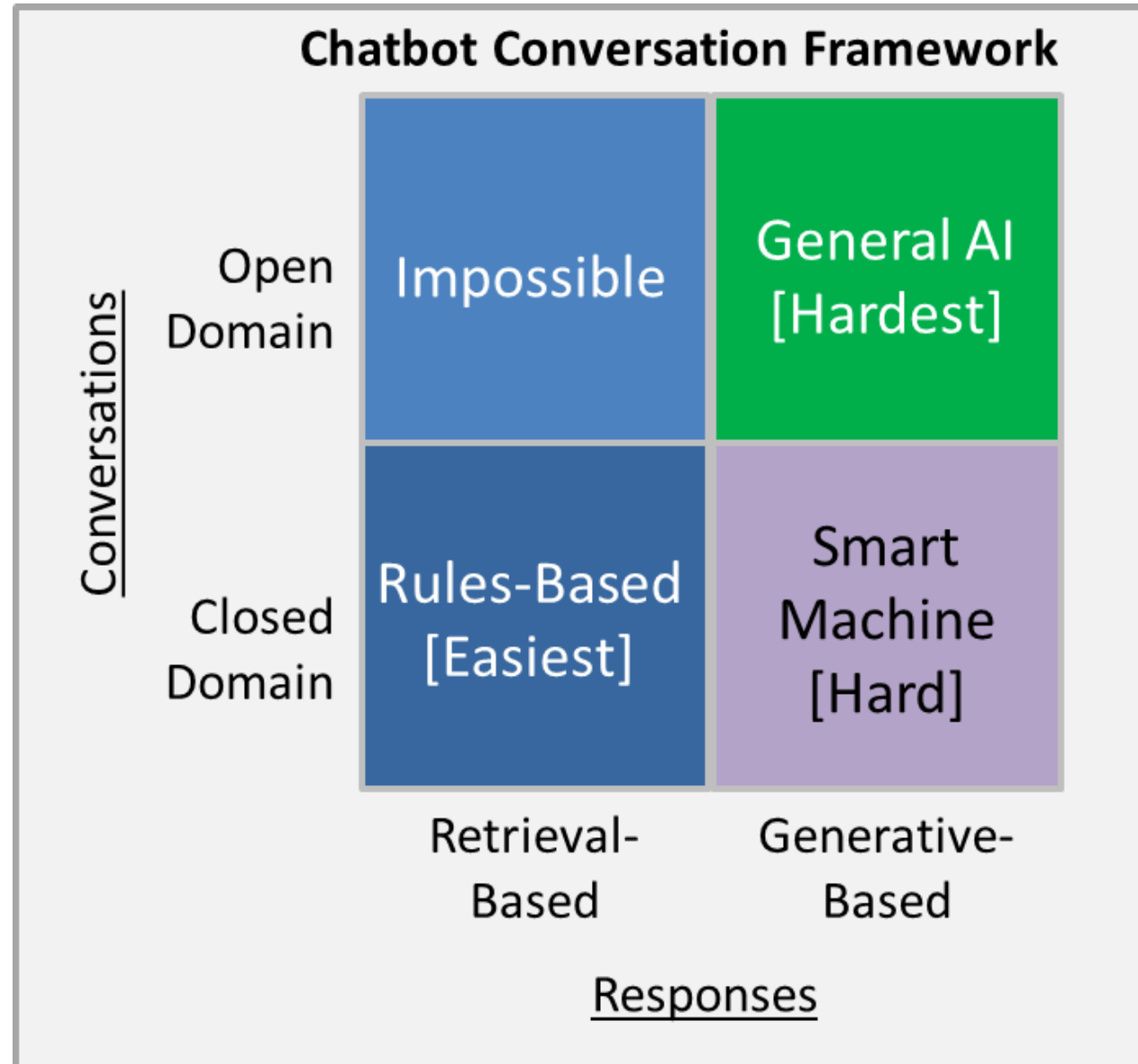
Dialogue System



Overall Architecture of Intelligent Chatbot



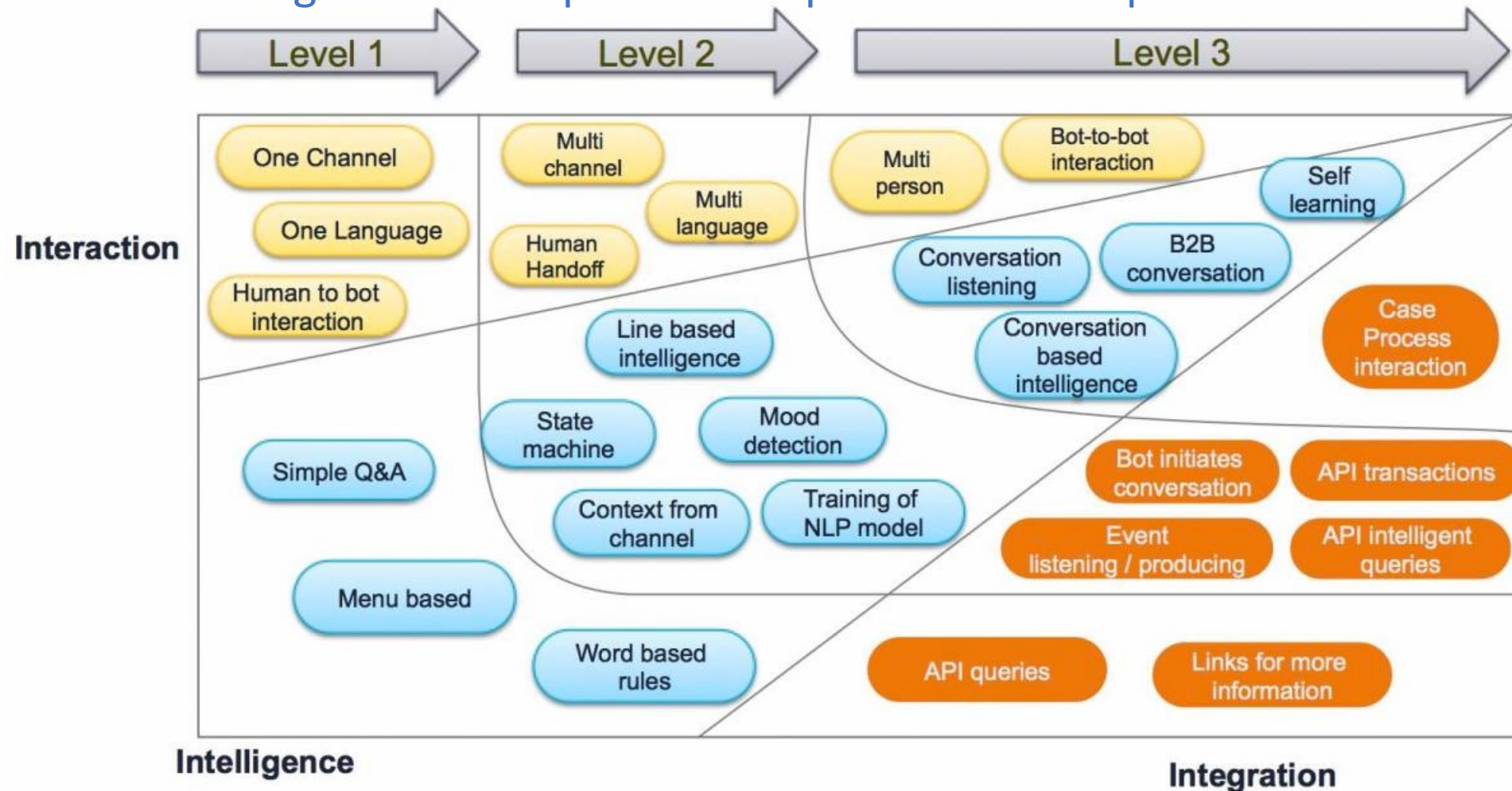
Chatbot Conversation Framework



Chatbots

Bot Maturity Model

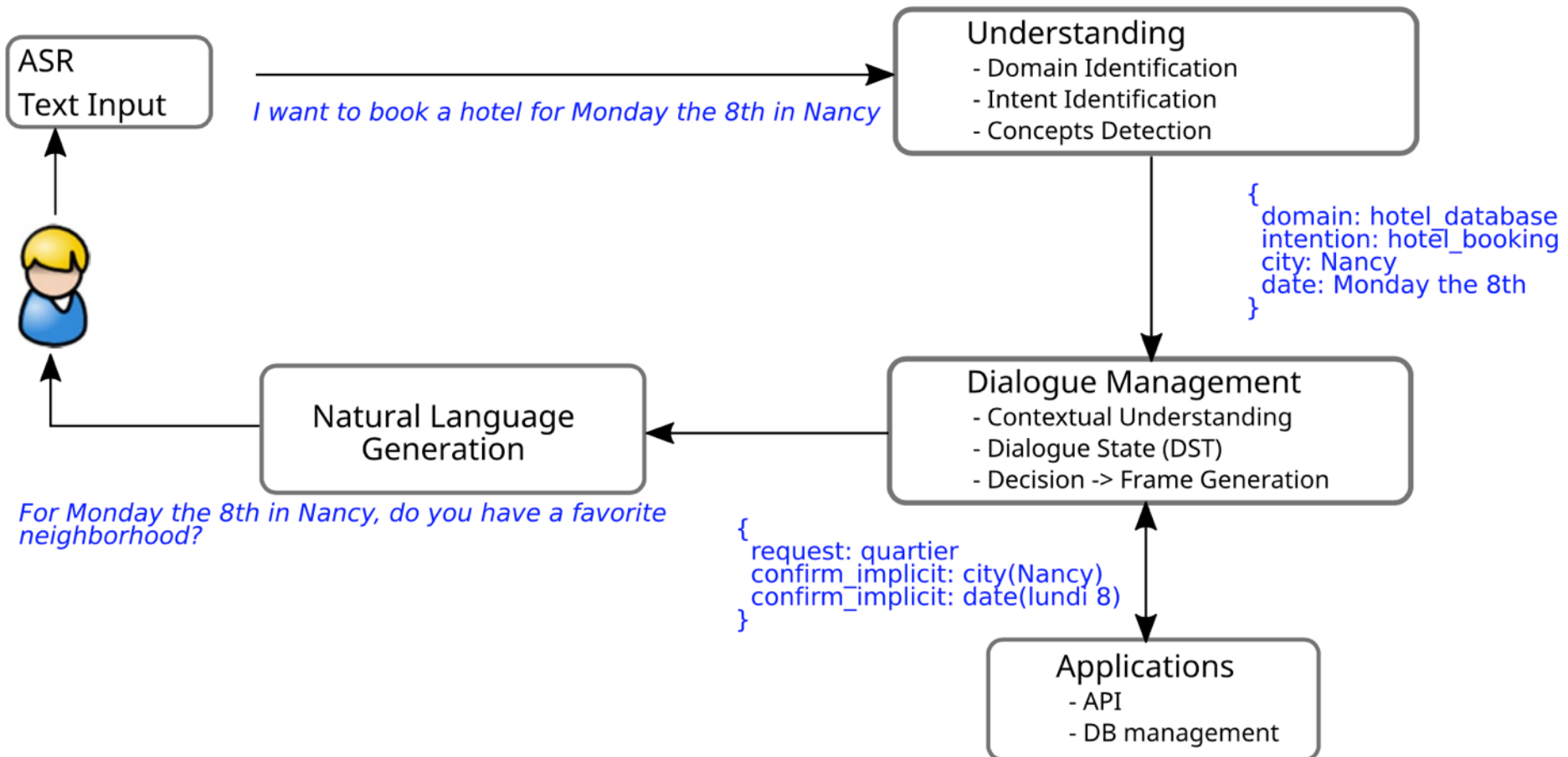
Customers want to have simpler means to interact with businesses and get faster response to a question or complaint.



Task-Oriented Dialogue System

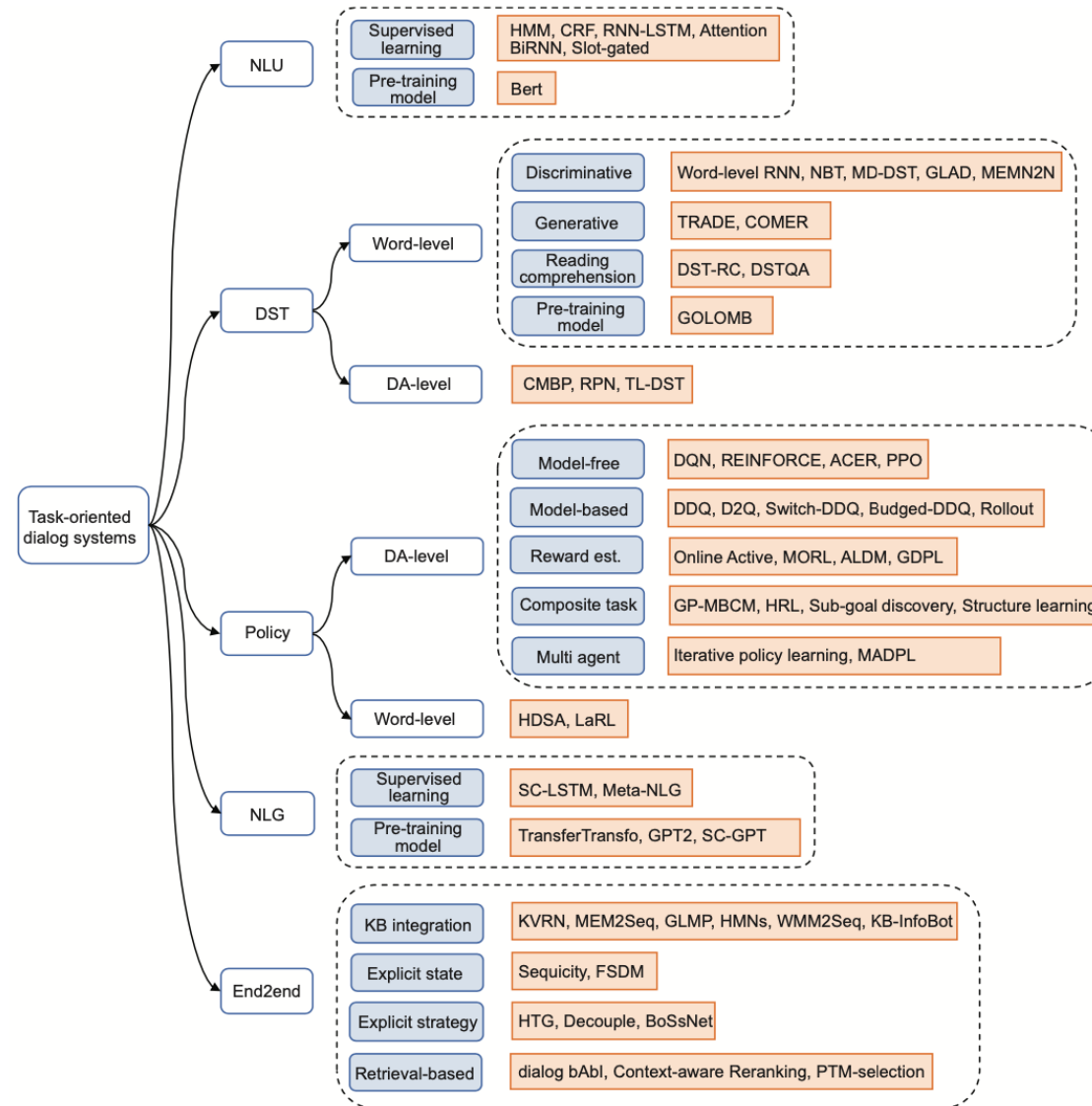
Task-Oriented Dialogue System

(Deriu et al., 2021)



Task-Oriented Dialogue Systems

(Zhang et al., 2020)



Dialog State Tracker (DST)

Dialog state tracker inputs			Dialog state tracker outputs	
System action/ user response	ASR output	SLU output	State	Score
How can I help you? <i>welcome()</i>	Cheap restaurant	inform(price=cheap)	price=cheap	<div></div>
	Restaurant	inform(food=italian)	food=Italian	<div></div>
	Italian		food=italian,price=cheap	<div></div>
An Italian restaurant			[none]	<div></div>
What price did you want? <i>request(price)</i>	East Area	inform(area=east)	price=cheap	<div></div>
	Italian	inform(food=italian)	food=Italian	<div></div>
	Yeah	affirm()	food=italian,price=cheap	<div></div>
Uh, Italian			area=east	<div></div>
			food=italian, area=east	<div></div>
			price=cheap, area=east	<div></div>
			food=italian,price=cheap,area=east	<div></div>
			[none]	<div></div>

Source: Deriu, Jan, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak (2021). "Survey on evaluation methods for dialogue systems." Artificial Intelligence Review 54, no. 1 (2021): 755-810.

Dialogue Acts

(Young et al., 2010)

Dialogue act	Description
hello($a = x, b = y, \dots$)	Open a dialogue and give info $a = x, b = y, \dots$
inform($a = x, b = y, \dots$)	Give information $a = x, b = y, \dots$
request($a, b = x, \dots$)	Request value for a given $b = x, \dots$
reqalts($a = x, \dots$)	Request alternative with $a = x, \dots$
confirm($a = x, b = y, \dots$)	Explicitly confirm $a = x, b = y, \dots$
confreq($a = x, \dots, d$)	Implicitly confirm $a = x, \dots$ and request value of d
select($a = x, a = y$)	Select either $a = x$ or $a = y$
affirm($a = x, b = y$)	Affirm and give further info $a = x, b = y, \dots$
negate($a = x$)	Negate and give corrected value $a = x$
deny($a = x$)	Deny that $a = x$
bye()	Close a dialogue

Sample Dialogue Acts

Utterance	Dialogue Act
U: Hi, I am looking for somewhere to eat	hello(task = find,type=restaurant)
S: You are looking for a restaurant. What type of food?	confreq(type = restaurant,food)
U: I'd like an Italian somewhere near the museum.	inform(food = Italian,near=museum)
S: Roma is a nice Italian restaurant near the museum.	inform(name = "Roma", type = restaurant, food = Italian, near = museum)
U: Is it reasonably priced?	confirm(pricerange = moderate)
S: Yes, Roma is in the moderate price range.	affirm(name = "Roma", pricerange = moderate)
U: What is the phone number?	request(phone)
S: The number of Roma is 385456.	inform(name = "Roma", phone = "385456")
U: Ok, thank you goodbye.	bye()

Dialogue on Airline Travel Information System (ATIS)

The ATIS (Airline Travel Information System) Dataset

<https://www.kaggle.com/siddhadev/atis-dataset-from-ms-cntk>

Sentence	what	flights	leave	from	phoenix
Slots	O	O	O	O	B-fromloc
Intent	atis_flight				

Training samples: 4978

Testing samples: 893

Vocab size: 943

Slot count: 129

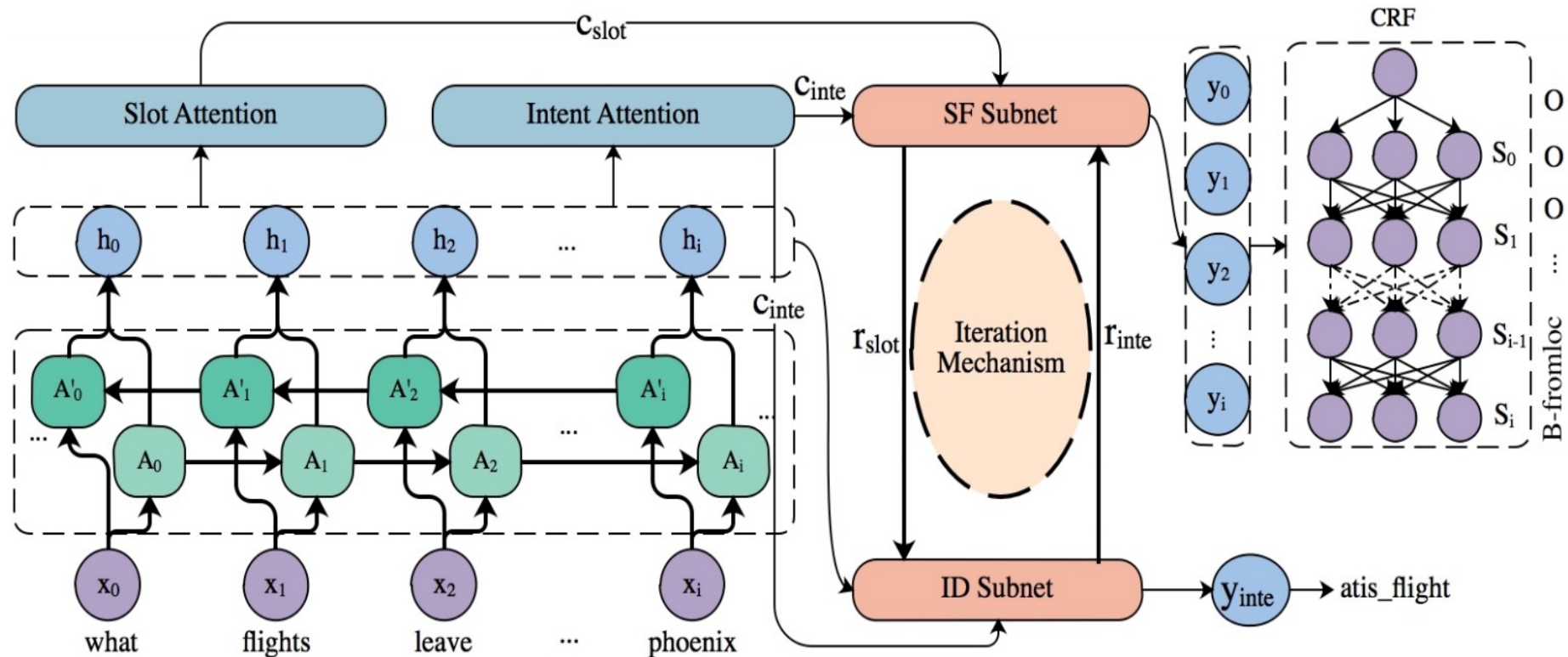
Intent count: 26

SF-ID Network (E et al., 2019)

Slot Filling (SF)

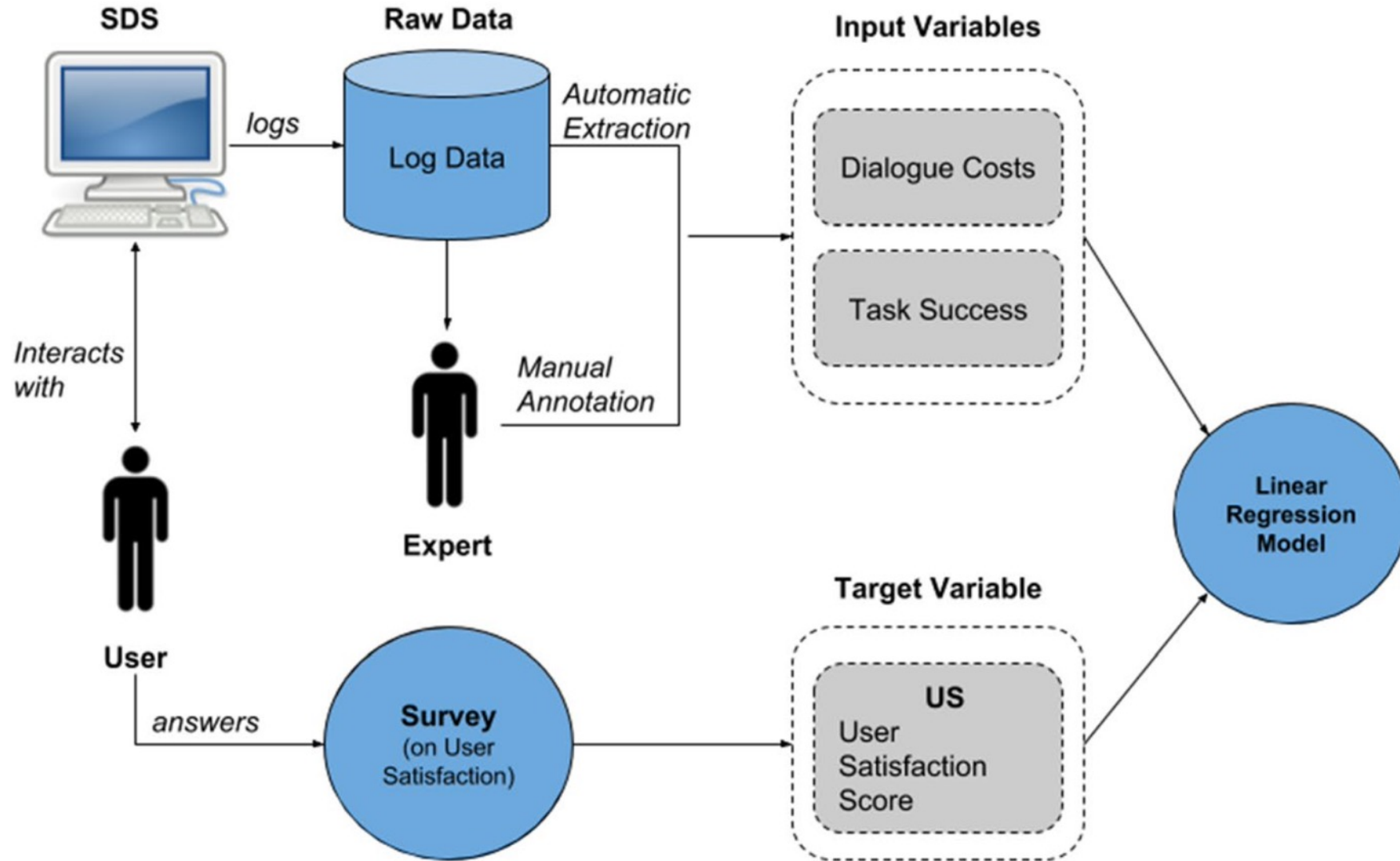
Intent Detection (ID)

A Novel Bi-directional Interrelated Model for Joint Intent Detection and Slot Filling



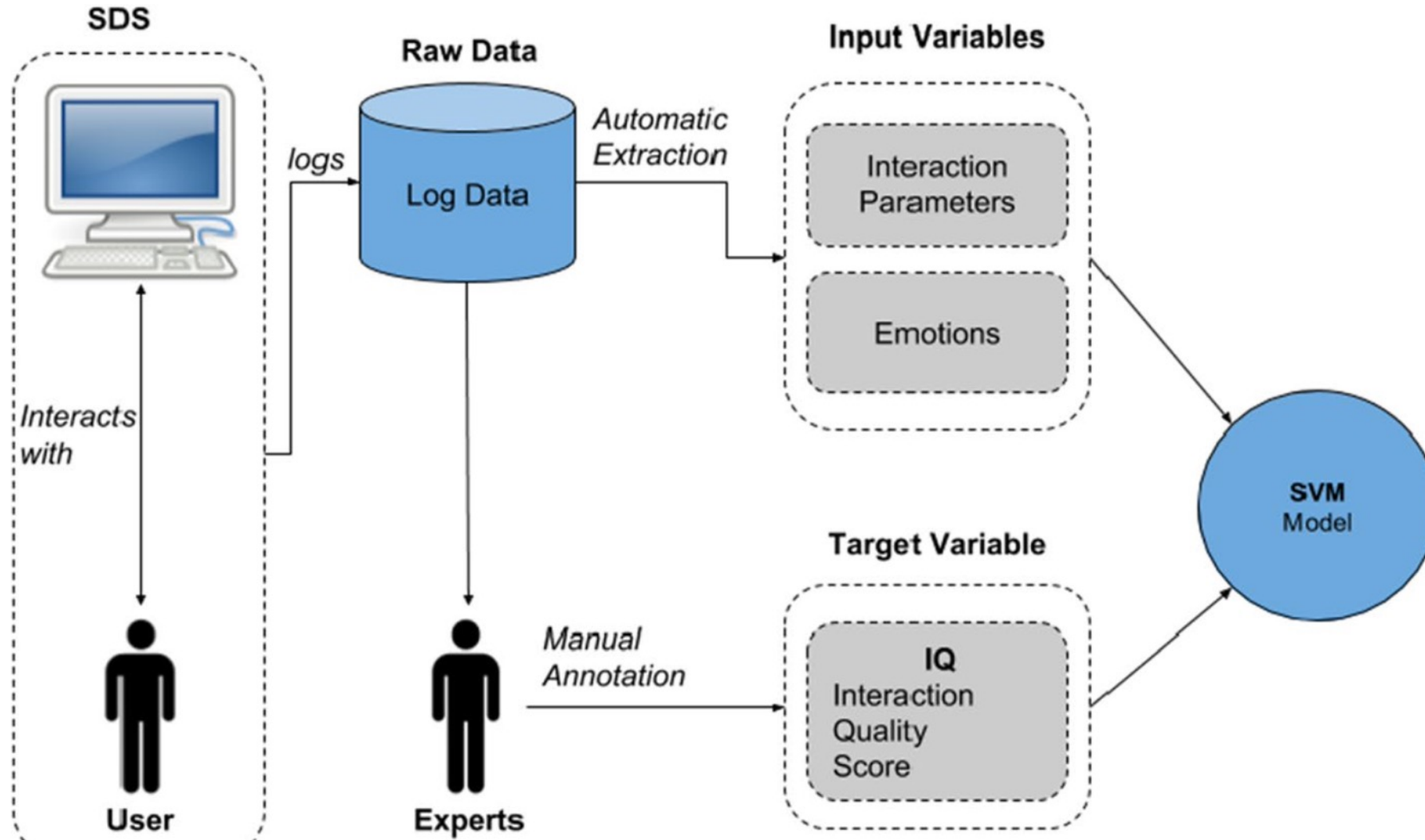
PARAdigm for Dialog System Evaluation

PARADISE Framework (Walker et al. 1997)



Interaction Quality procedure

(Schmitt and Ultes, 2015)



Datasets for task-oriented dialogue systems

Name	Topics	# dialogues	Reference
DSTC1	Bus schedules	15,000	(Williams et al. 2013)
DSTC2	Restaurants	3000	(Henderson et al. 2014)
DSTC3	Tourist information	2265	(Henderson et al. 2013a)
DSTC4 & DSTC5	Tourist information	35	(Kim et al. 2016)
DSTC6	Restaurant reservation	–	(Perez et al. 2017)
DSTC7 (Flex Data)	Student guiding	500	(Gunasekara et al. 2019)
DSTC8 (MetaLWOz)	47 domains	37,884	(Lee et al. 2019)
DSTC8 (Schema-Guided)	20 domains	22,825	(Rastogi et al. 2019)
MultiWOZ	Tourist information	10,438	(Budzianowski et al. 2018)
Taskmaster-1	6 domains	13,215	(Byrne et al. 2019)
MultiDoGo	6 domains	86,698	(Peskov et al. 2019)

Restaurants Dialogue Datasets

- **MIT Restaurant Corpus**
 - <https://groups.csail.mit.edu/sls/downloads/restaurant/>
- **CamRest676**
(Cambridge restaurant dialogue domain dataset)
 - <https://www.repository.cam.ac.uk/handle/1810/260970>
- **DSTC2 (Dialog State Tracking Challenge 2 & 3)**
 - <http://camdial.org/~mh521/dstc/>

CrossWOZ:

A Large-Scale Chinese Cross-Domain Task-Oriented Dialogue Dataset

MultiWOZ	CrossWOZ
usr: I'm looking for a college type attraction. ... usr: I would like to visit <u>in town centre</u> please. ... usr: Can you find an Indian restaurant for me that is also in the town centre ?	usr: 你好，可以帮我推荐一个评分是4.5分以上的景点吗？ Hello, could you recommend an attraction with a rating of 4.5 or higher? sys: 天安门城楼，簋街小吃和 <u>北京欢乐谷</u> 都是很不错的地方呢。 Tiananmen, Gui Street, and <u>Beijing Happy Valley</u> are very nice places. usr: 我喜欢 <u>北京欢乐谷</u> ，你知道这个景点周边的酒店都是什么吗？ I like <u>Beijing Happy Valley</u> . What hotels are around this attraction? sys: 那可多了，有A酒店，B酒店，C酒店。 There are many, such as hotel A, hotel B, and hotel C. usr: 太好了，我正打算在 <u>景点附近</u> 找个酒店住宿呢，知道哪家评分是4分以上，提供叫醒服务的？ Great! I am planning to find a hotel to stay near the attraction . Which one has a rating of 4 or higher and offers wake-up call service?
Schema	
usr: I want a hotel in <u>San Diego</u> and I want to check out on Thursday next week. ... usr: I need a one way flight to go there .	

CrossWOZ:

A Large-Scale Chinese Cross-Domain Task-Oriented Dialogue Dataset

Type	Single-domain goal					Multi-domain goal		
Dataset	DSTC2	WOZ 2.0	Frames	KVRET	M2M	MultiWOZ	Schema	CrossWOZ
Language	EN	EN	EN	EN	EN	EN	EN	CN
Speakers	H2M	H2H	H2H	H2H	M2M	H2H	M2M	H2H
# Domains	1	1	1	3	2	7	16	5
# Dialogues	1,612	600	1,369	2,425	1,500	8,438	16,142	5,012
# Turns	23,354	4,472	19,986	12,732	14,796	115,424	329,964	84,692
Avg. domains	1	1	1	1	1	1.80	1.84	3.24
Avg. turns	14.5	7.5	14.6	5.3	9.9	13.7	20.4	16.9
# Slots	8	4	61	13	14	25	214	72
# Values	212	99	3,871	1363	138	4,510	14,139	7,871

Task-Oriented Dialogue

Initial user state (=user goal)

id=1(Attraction): fee=free,
name=?, nearby hotels=?

id=2(Hotel): **name=near (id=1)**,
wake-up call=yes, rating=?

id=3(Taxi): **from=(id=1)**, **to=(id=2)**,
car type=? plate number=?

...

Final user state

id=1 (Attraction): name=Tiananmen Square,
fee=free, nearby hotels=[Beijing Capital
Hotel, Guidu Hotel Beijing]

id=2 (Hotel): **name=Beijing Capital Hotel**,
wake-up call=yes, rating=4.6

id=3 (Taxi): **from=Tiananmen Square**,
to=Beijing Capital Hotel,
car type=#CX, plate number=#CP



Source: Zhu, Qi, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. "Crosswoz: A large-scale chinese cross-domain task-oriented dialogue dataset." arXiv preprint arXiv:2002.11893 (2020).

An example dialog from the test set for MultiWOZ

(en→zh) sub-task

Speaker	Utterance	Dialog State Update
User	Hello! I am looking for a local guesthouse in the centre. 你好！我在市中心找一家本地宾馆。	hotel: {area: centre, type: guesthouse}
System	OK. I am glad to recommend Alexander Bed and Breakfast to you. 好的。这边很高兴向您推荐亚历山大住宿加早餐旅馆。	旅馆: {区域: 中心, 类型: 宾馆}
User	Where is it? 它在哪里？	hotel: {name: alexander bed and breakfast}
System	It is at 56 Saint Barnabas Road. 圣巴纳巴斯路56号。	旅馆: {名称: 亚历山大住宿加早餐旅馆}
User	I also wish to have a meal in a local European restaurant in the centre. 我还想在市中心的一家本地欧洲餐厅吃饭。	restaurant: {food: european, area: centre}
System	You can choose Eraina. 您可以选择伊莱娜。	餐厅: {食物: 欧洲的, 区域: 中心}
User	Please give me its address. 那请给我它的地址。	restaurant: {name: eraina}
System	It is in Free School Lane City Centre. 市中心自由校园巷。	餐厅: {名称: 伊莱娜}
User	Ok. I'll go there. I need to book a taxi from Alexander Bed and Breakfast to Eraina after 07:00. 好。那我去那里。我还要预订一辆7:00时从亚历山大住宿加早餐旅馆到伊莱娜的出租车。	taxi: {leaveAt: 07:00, destination: eraina, departure: alexander bed and breakfast}
System	Well. I find a yellow Skoda. 好的。是一辆黄色的斯柯达。	出租车: {出发时间: 07:00, 目的地: 伊莱娜, 出发地: 亚历山大住宿加早餐旅馆}
User	How about its phone number? 它的电话号码是多少？	No update
System	It is 78519675253. 78519675253。	
User	Thank you for your help. Bye! 谢谢你帮忙。再见！	No update
System	A pleasure. Bye bye! 我很乐意。再见！	

Reinforcement Learning from Human Feedback (RLHF)

ChatGPT: Optimizing Language Models for Dialogue

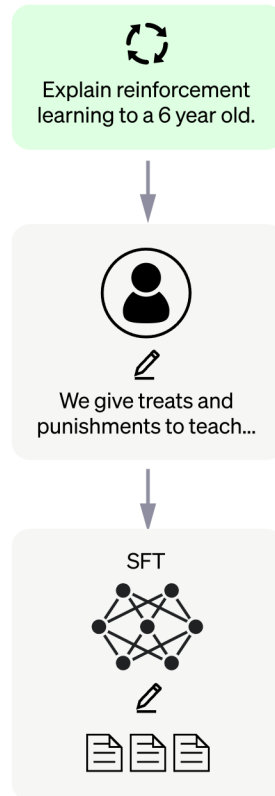
Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.



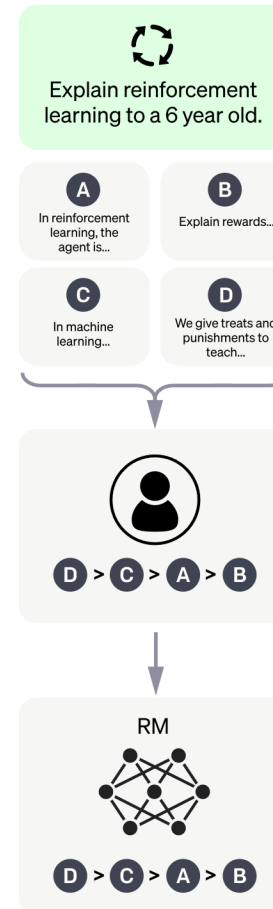
Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

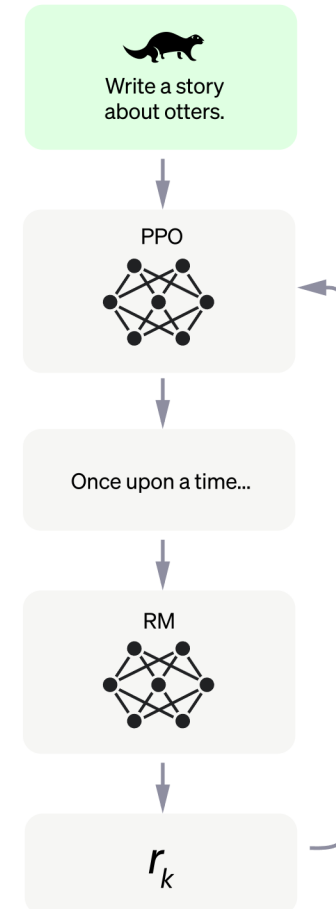
A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



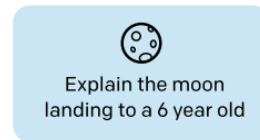
Training language models to follow instructions with human feedback

InstructGPT and GPT 3.5

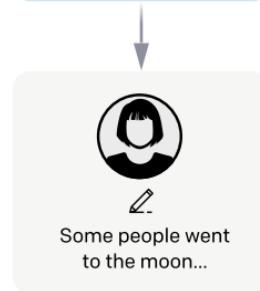
Step 1

**Collect demonstration data,
and train a supervised policy.**

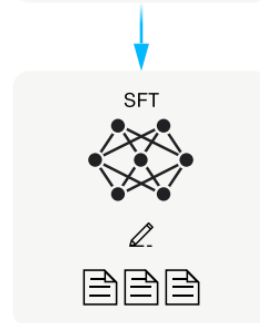
A prompt is
sampled from our
prompt dataset.



A labeler
demonstrates the
desired output
behavior.



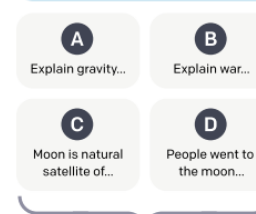
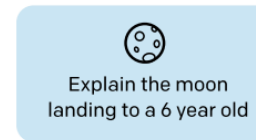
This data is used
to fine-tune GPT-3
with supervised
learning.



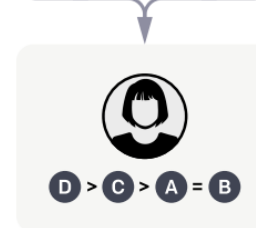
Step 2

**Collect comparison data,
and train a reward model.**

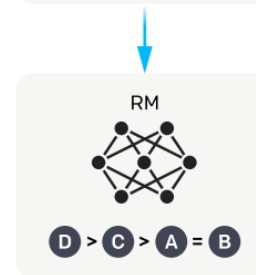
A prompt and
several model
outputs are
sampled.



A labeler ranks
the outputs from
best to worst.



This data is used
to train our
reward model.



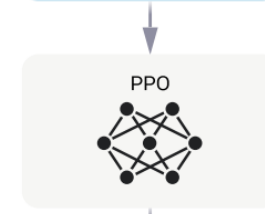
Step 3

**Optimize a policy against
the reward model using
reinforcement learning.**

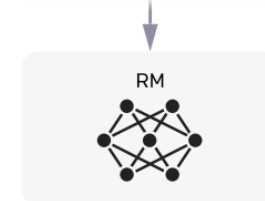
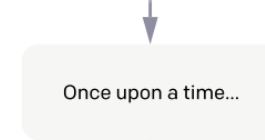
A new prompt
is sampled from
the dataset.



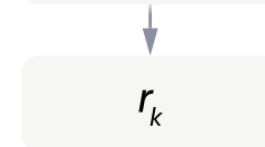
The policy
generates
an output.



The reward model
calculates a
reward for
the output.



The reward is
used to update
the policy
using PPO.

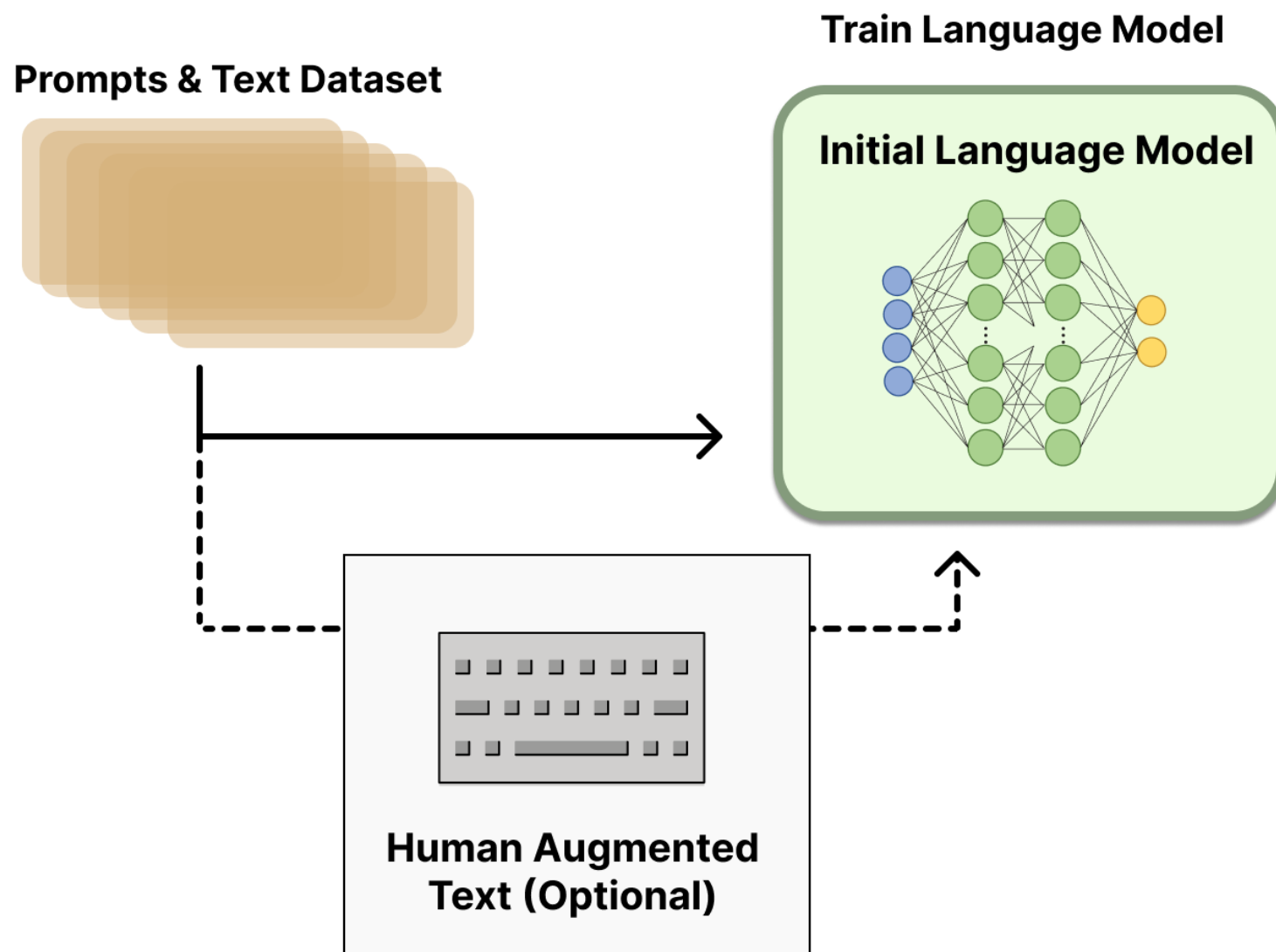


Reinforcement Learning from Human Feedback (RLHF)

1. **Pretraining a Language Model (LM)**
2. **Gathering Data and Training a Reward Model**
3. **Fine-tuning the LM with Reinforcement Learning**

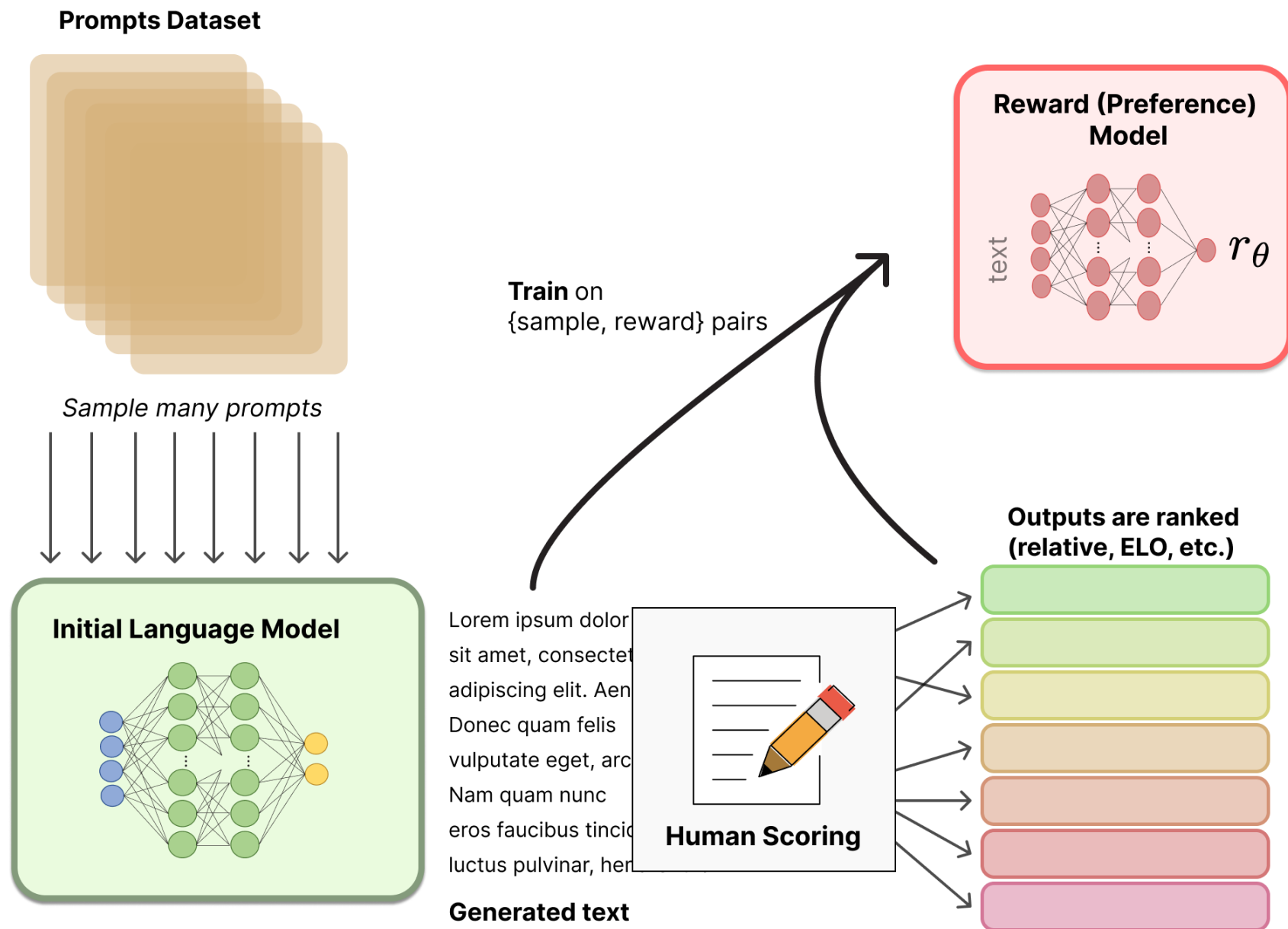
Reinforcement Learning from Human Feedback (RLHF)

Step 1. Pretraining a Language Model (LM)



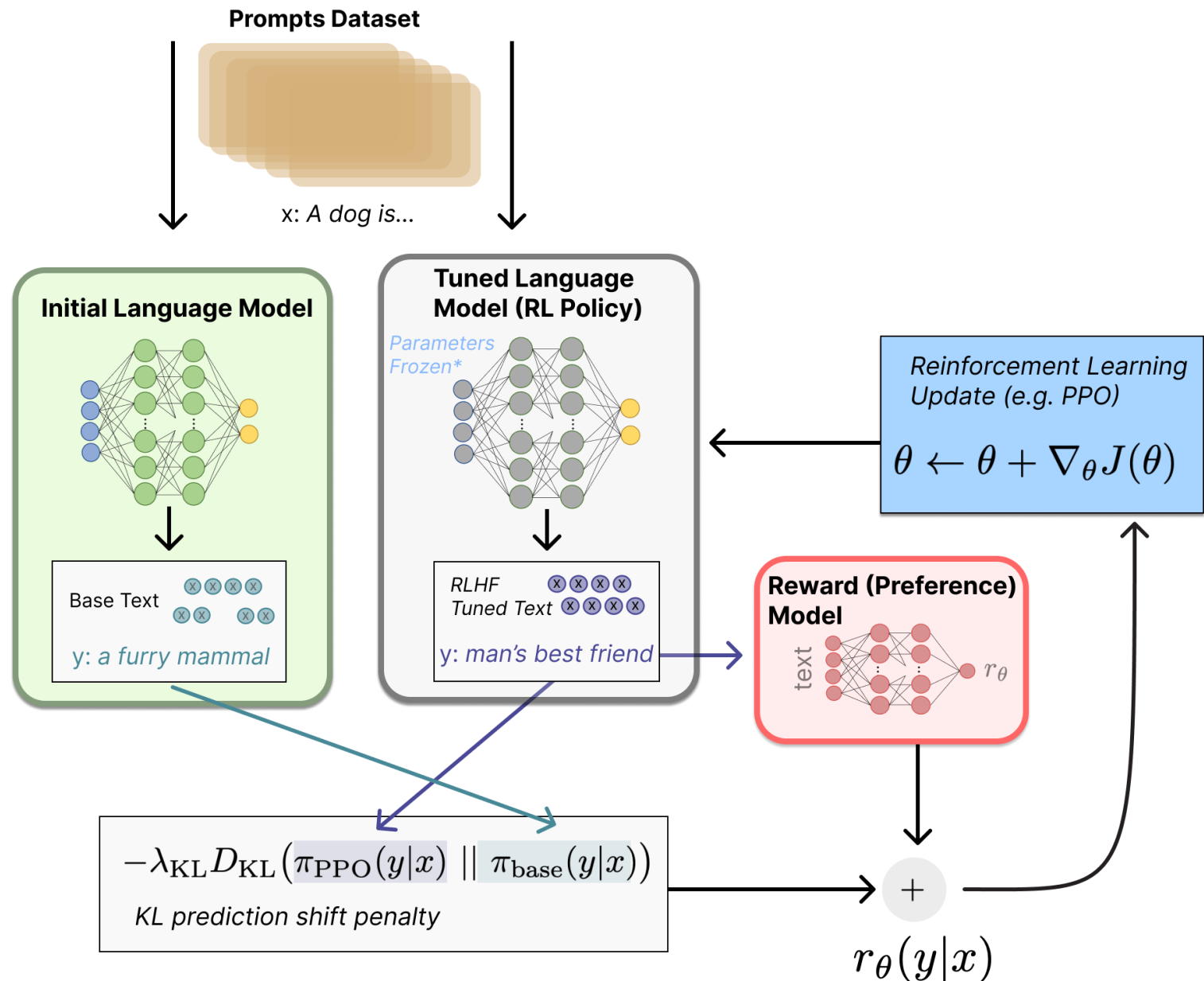
Reinforcement Learning from Human Feedback (RLHF)

Step 2. Gathering Data and Training a Reward Model

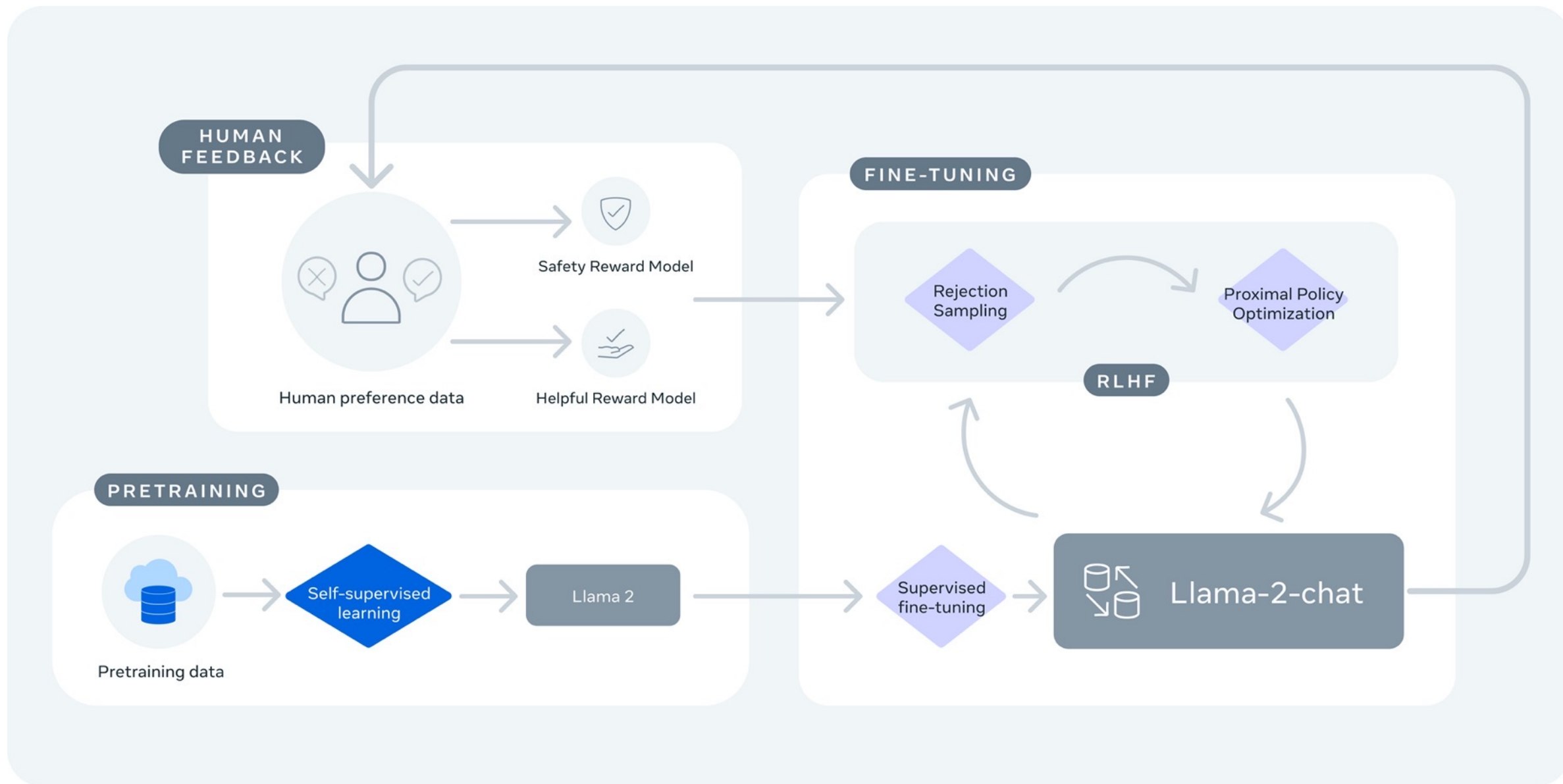


Reinforcement Learning from Human Feedback (RLHF)

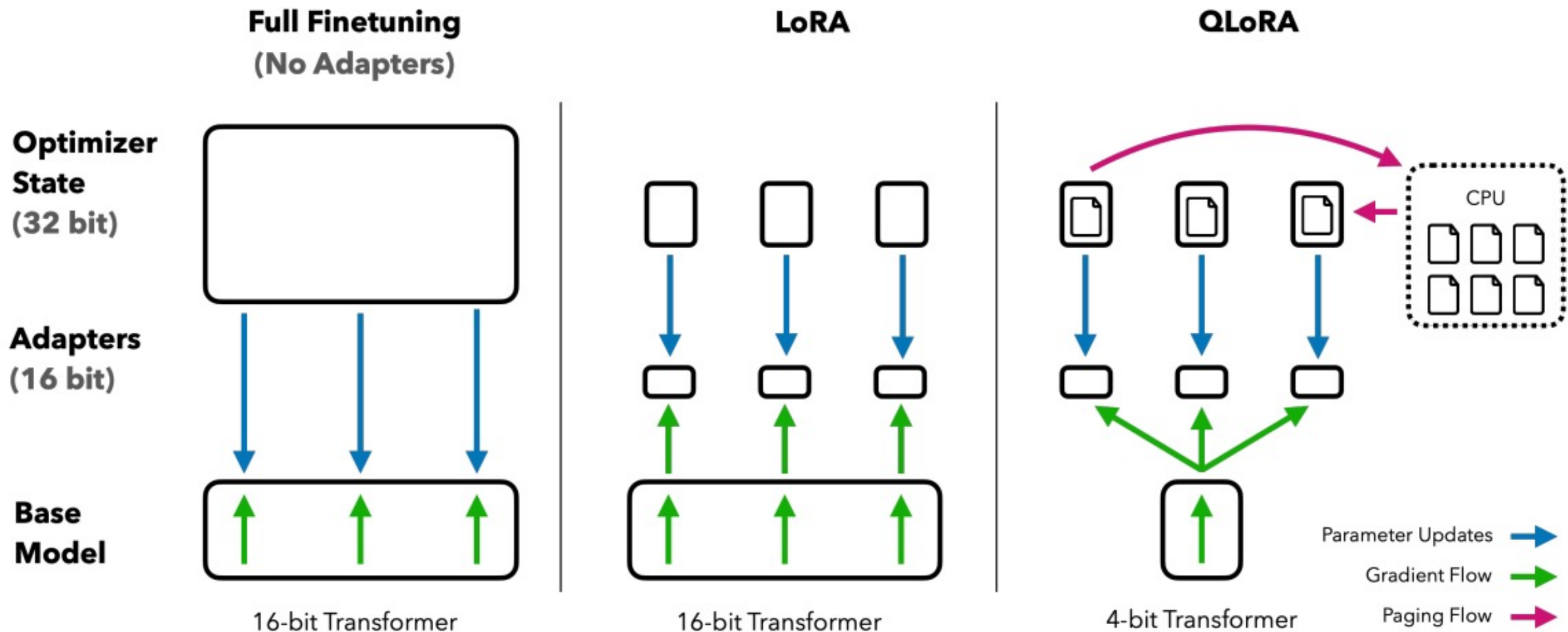
Step 3. Fine-tuning the LM with Reinforcement Learning



Llama-2-chat uses RLHF to ensure safety and helpfulness



QLoRA: Efficient Finetuning of Quantized LLMs



QLoRA: Efficient Finetuning of Quantized LLMs

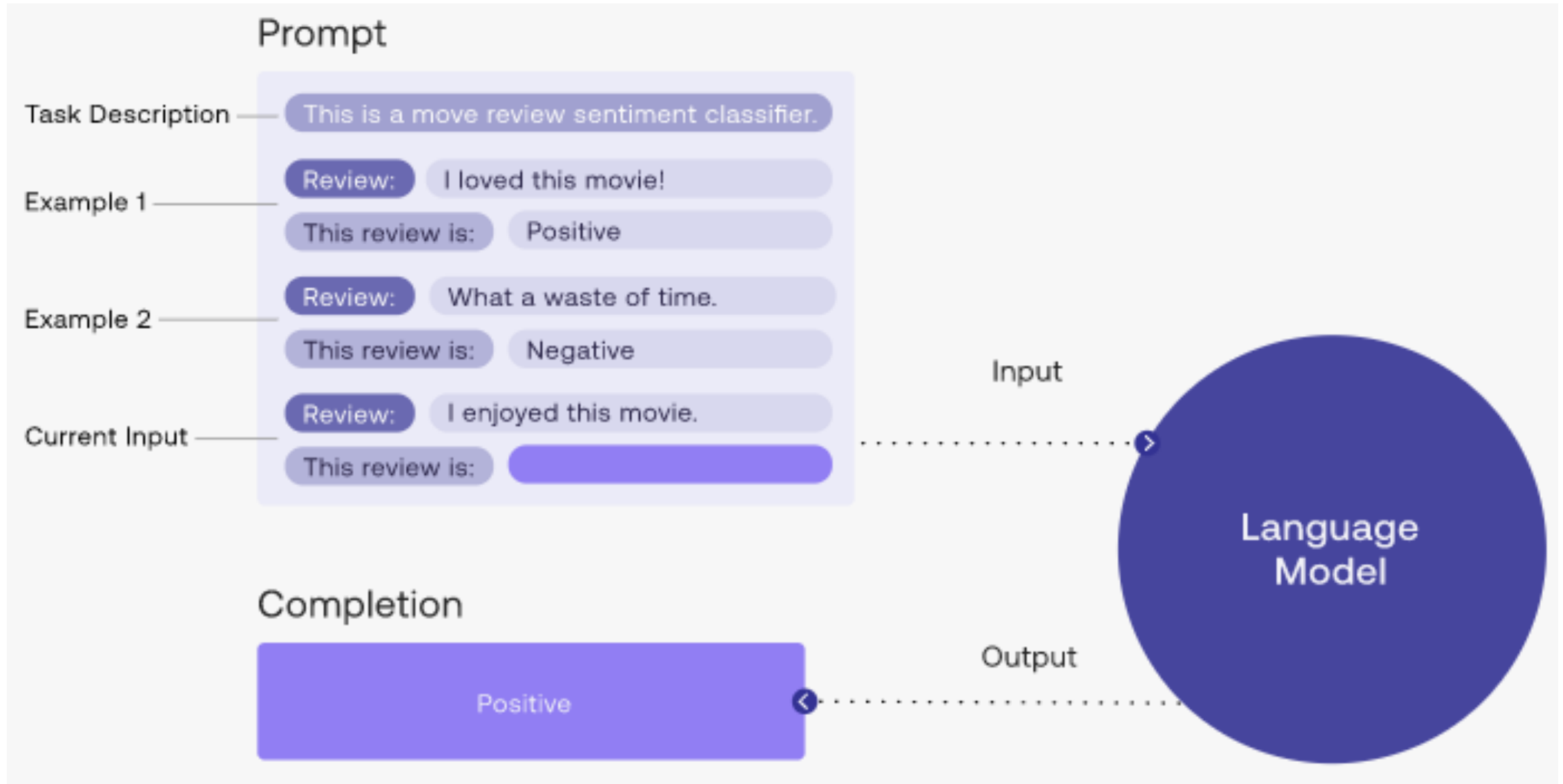
QLoRA reduces the average memory requirements of finetuning a **65B** parameter model from >780GB of **GPU memory** to **<48GB**

Model	Size	Elo
GPT-4	-	1348 \pm 1
Guanaco 65B	41 GB	1022 \pm 1
Guanaco 33B	21 GB	992 \pm 1
Vicuna 13B	26 GB	974 \pm 1
ChatGPT	-	966 \pm 1
Guanaco 13B	10 GB	916 \pm 1
Bard	-	902 \pm 1
Guanaco 7B	6 GB	879 \pm 1

QLoRA: Efficient Finetuning of Quantized LLMs

Model / Dataset	Params	Model bits	Memory	ChatGPT vs Sys	Sys vs ChatGPT	Mean	95% CI
GPT-4	-	-	-	119.4%	110.1%	114.5%	2.6%
Bard	-	-	-	93.2%	96.4%	94.8%	4.1%
Guanaco	65B	4-bit	41 GB	96.7%	101.9%	99.3%	4.4%
Alpaca	65B	4-bit	41 GB	63.0%	77.9%	70.7%	4.3%
FLAN v2	65B	4-bit	41 GB	37.0%	59.6%	48.4%	4.6%
Guanaco	33B	4-bit	21 GB	96.5%	99.2%	97.8%	4.4%
Open Assistant	33B	16-bit	66 GB	91.2%	98.7%	94.9%	4.5%
Alpaca	33B	4-bit	21 GB	67.2%	79.7%	73.6%	4.2%
FLAN v2	33B	4-bit	21 GB	26.3%	49.7%	38.0%	3.9%
Vicuna	13B	16-bit	26 GB	91.2%	98.7%	94.9%	4.5%
Guanaco	13B	4-bit	10 GB	87.3%	93.4%	90.4%	5.2%
Alpaca	13B	4-bit	10 GB	63.8%	76.7%	69.4%	4.2%
HH-RLHF	13B	4-bit	10 GB	55.5%	69.1%	62.5%	4.7%
Unnatural Instr.	13B	4-bit	10 GB	50.6%	69.8%	60.5%	4.2%
Chip2	13B	4-bit	10 GB	49.2%	69.3%	59.5%	4.7%
Longform	13B	4-bit	10 GB	44.9%	62.0%	53.6%	5.2%
Self-Instruct	13B	4-bit	10 GB	38.0%	60.5%	49.1%	4.6%
FLAN v2	13B	4-bit	10 GB	32.4%	61.2%	47.0%	3.6%
Guanaco	7B	4-bit	5 GB	84.1%	89.8%	87.0%	5.4%
Alpaca	7B	4-bit	5 GB	57.3%	71.2%	64.4%	5.0%
FLAN v2	7B	4-bit	5 GB	33.3%	56.1%	44.8%	4.0%

Prompt Engineering with ChatGPT for NLP



Outline

- Introduction
- Overview of Generative AI
- Overview of Large Language Models (LLMs)
- Foundation of Transformers: Attention Mechanism
- Fine-tuning LLM for Question Answering System
- Fine-tuning LLM for Dialogue System
- **Challenges and Limitations of Generative AI for QA and Dialogue Systems**
- Q & A

Challenges and Limitations of Generative AI for QA and Dialogue Systems

- **Understanding and Contextualization**
- **Consistency**
- **Data Dependence and Bias**
- **Sensitivity to Input Phrasing**
- **Ethical and Privacy Concerns**

Understanding and Contextualization

- **Generative AI models like ChatGPT can understand and generate human-like text, but they struggle with maintaining context and a deep understanding of complex subjects over a long conversation.**
- **LLMs might not grasp the underlying connotations, cultural references, or subtle humor in a conversation.**
- **LLMs are also limited by the information they were trained on and cannot learn or acquire new information**

Consistency

- **Despite significant improvements in generating relevant responses, LLMs can still be inconsistent.**
- **LLMs might provide different answers to slight rephrasings of the same question or forget previously mentioned information in a conversation.**
- **This is because LLMs lack a persistent memory and are unable to reason about facts in the way humans do.**

Data Dependence and Bias

- **The quality of generative AI's responses largely depends on the data it was trained on.**
- **If the training data contains biases, misinformation, or low-quality information, the model can reproduce those.**
- **As a result, LLMs might propagate harmful stereotypes, misinformation, or inappropriate responses.**

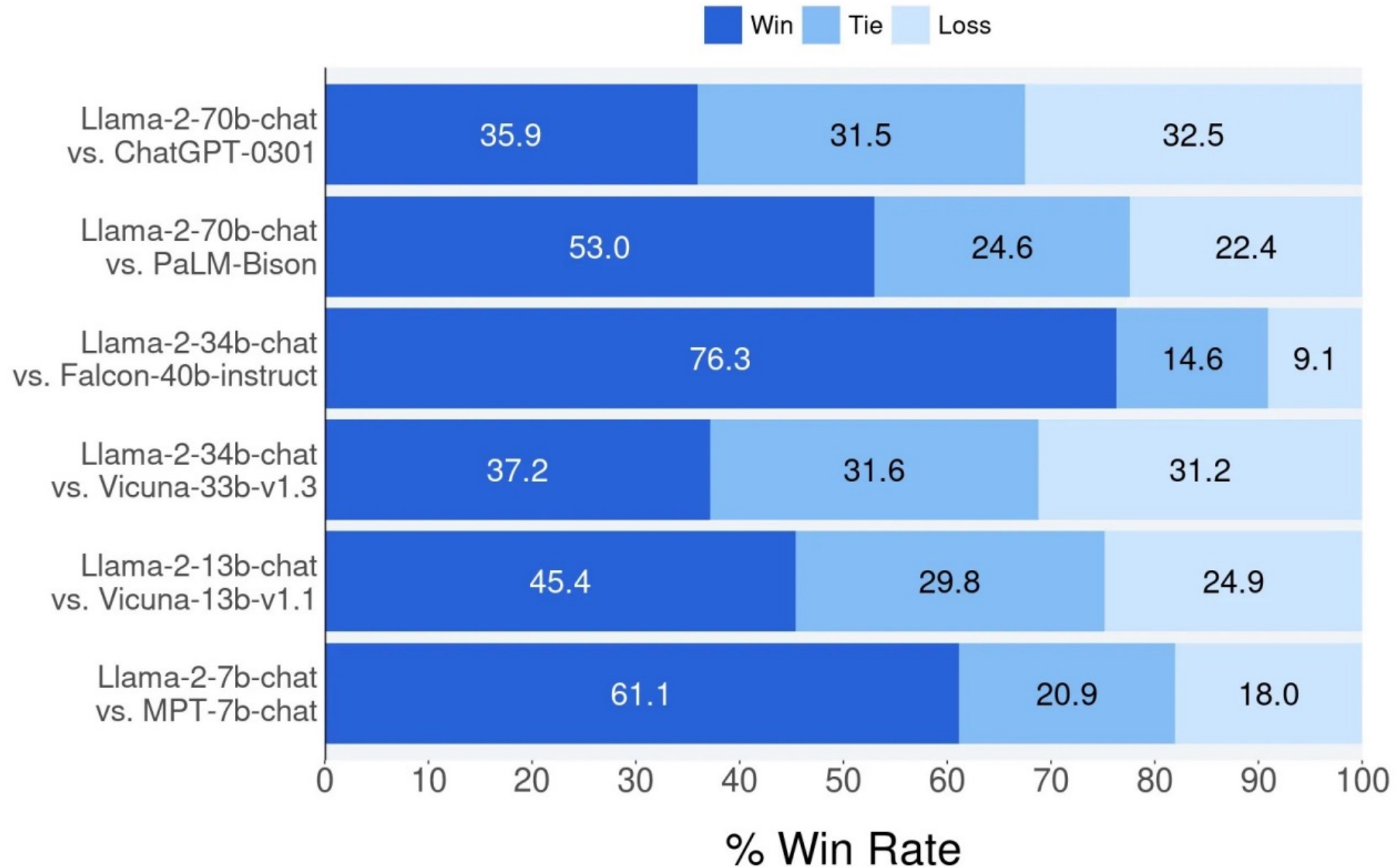
Sensitivity to Input Phrasing

- **Generative AI model's responses can be very sensitive to how a question or statement is phrased.**
- **Slight changes in input phrasing can lead to substantially different responses.**
- **This can make the interaction with Generative AI models unpredictable and unreliable.**

Ethical and Privacy Concerns

- **With the ability to generate human-like text, there's a risk that these systems could be used maliciously, like generating deepfakes or disinformation.**
- **Generative AI models might inadvertently generate sensitive or personal information which they learned during training, even though they are designed to avoid doing so.**
- **Balancing the benefits of these systems with the need to prevent misuse is a significant challenge.**

Llama-2 Chat: Helpfulness Human Evaluation



Acknowledgments: Research Projects

1. **Applying AI technology to construct knowledge graphs of cryptocurrency anti-money laundering: a few-shot learning model**
 - MOST, 110-2410-H-305-013-MY2, 2021/08/01~2023/07/31
2. **Fintech Green Finance for Carbon Market Index, Corporate Finance, and Environmental Policies. Carbon Emission Sentiment Index with AI Text Analytics**
 - NTPU, 112-NTPU_ORDA-F-003 , 2023/01/01~2024/12/31
3. **Research on speech processing, synthesis, recognition, and sentence construction of people with language disabilities. Multimodal Cross-lingual Task-Oriented Dialogue System**
 - NTPU, 112-NTPU_ORDA-F-004, 2023/01/01~2025/12/31
4. **Use deep learning to identify commercially dental implant systems - observational study**
 - USTP-NTPU-TMU, USTP-NTPU-TMU-112-01, 2023/01/01~2023/12/31
5. **Metaverse Avatar Automatic Metadata Generation Module**
 - FormosaVerse x NTPU, NTPU-111A413E01, 2022/12/01~2023/11/30
6. **Establishment and Implement of Smart Assistive Technology for Dementia Care and Its Socio-Economic Impacts. Intelligent, individualized and precise care with smart AT and system integration**
 - MOST, 111-2627-M-038-001-, 2022/08/01~2023/07/31

Summary

- **Introduction**
- **Overview of Generative AI**
- **Overview of Large Language Models (LLMs)**
- **Foundation of Transformers: Attention Mechanism**
- **Fine-tuning LLM for Question Answering System**
- **Fine-tuning LLM for Dialogue System**
- **Challenges and Limitations of Generative AI for QA and Dialogue Systems**
- **Q & A**

References

- Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun (2023). "A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT." arXiv preprint arXiv:2303.04226.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. (2023) "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing." ACM Computing Surveys 55, no. 9 (2023): 1-35.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min et al. (2023) "A Survey of Large Language Models." arXiv preprint arXiv:2303.18223.
- Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov et al. (2023) "Llama 2: Open Foundation and Fine-Tuned Chat Models." arXiv preprint arXiv:2307.09288 (2023).
- Junliang Wang, Chuqiao Xu, Jie Zhang, and Ray Zhong (2022). "Big data analytics for intelligent manufacturing systems: A review." Journal of Manufacturing Systems 62 (2022): 738-752.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155.
- Gozalo-Brizuela, Roberto, and Eduardo C. Garrido-Merchan (2023). "ChatGPT is not all you need. A State of the Art Review of large Generative AI models." arXiv preprint arXiv:2301.04655 (2023).
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. (2023) "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning." arXiv preprint arXiv:2305.06500 (2023).
- Shahab Saquib Sohail, Faiza Farhat, Yassine Himeur, Mohammad Nadeem, Dag Øivind Madsen, Yashbir Singh, Shadi Atalla, and Wathiq Mansoor (2023). "The Future of GPT: A Taxonomy of Existing ChatGPT Research, Current Challenges, and Possible Future Directions." Current Challenges, and Possible Future Directions (April 8, 2023) (2023).
- Longbing Cao (2022). "Decentralized ai: Edge intelligence and smart blockchain, metaverse, web3, and desc." IEEE Intelligent Systems 37, no. 3: 6-19.
- Qinglin Yang, Yetong Zhao, Huawei Huang, Zehui Xiong, Jiawen Kang, and Zibin Zheng (2022). "Fusing blockchain and AI with metaverse: A survey." IEEE Open Journal of the Computer Society 3 : 122-136.
- Russell Belk, Mariam Humayun, and Myriam Brouard (2022). "Money, possessions, and ownership in the Metaverse: NFTs, cryptocurrencies, Web3 and Wild Markets." Journal of Business Research 153: 198-205.
- Thien Huynh-The, Quoc-Viet Pham, Xuan-Quy Pham, Thanh Thi Nguyen, Zhu Han, and Dong-Seong Kim (2022). "Artificial Intelligence for the Metaverse: A Survey." arXiv preprint arXiv:2202.10336.
- Thippa Reddy Gadekallu, Thien Huynh-The, Weizheng Wang, Gokul Yenduri, Pasika Ranaweera, Quoc-Viet Pham, Daniel Benevides da Costa, and Madhusanka Liyanage (2022). "Blockchain for the Metaverse: A Review." arXiv preprint arXiv:2203.09738.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155.



Q & A

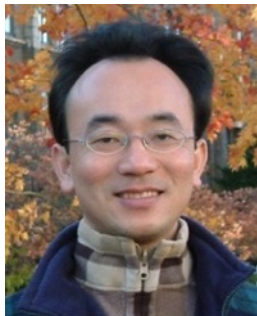
Generative AI and Large Language Models for Question Answering and Dialogue Systems

Time: 10:00-12:00; 14:00-16:00, Wednesday, July 26, 2023

Place: Taipei Research and Development Center, Asia University

Address: 16-5, No. 77, Xintai 5th Road, Xizhi District, New Taipei City, Taiwan (FE World Center)

Host: Prof. Wen-Lian Hsu



Min-Yuh Day, Ph.D,
Associate Professor

Institute of Information Management, National Taipei University

<https://web.ntpu.edu.tw/~myday>

