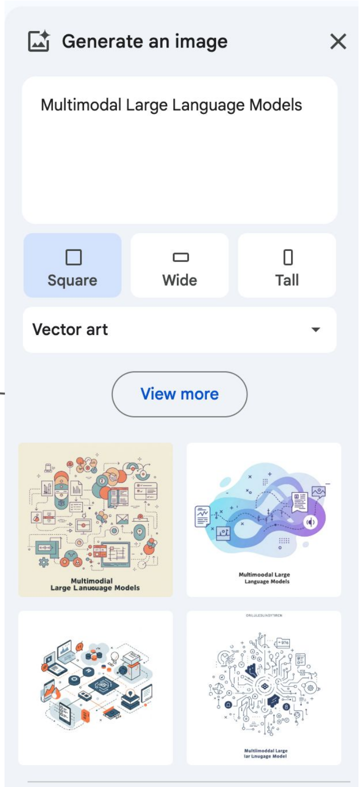
Multimodal LLMs

Chulalongkorn University (Online)
April 21, 2025

Soravit “Beer” Changpinyo
Google DeepMind

Opinions are my own
Compiled from multiple sources with credit given in slides

Multimodal LLMs



Chulalongkorn University (Online)
April 21, 2025

Soravit "Beer" Changpinyo
Google DeepMind

LLMs

Multimodal LLMs

Multimodal LLMs possess

the ability to **process** and **generate** information across different data modalities (text, images, audio, video, etc.)

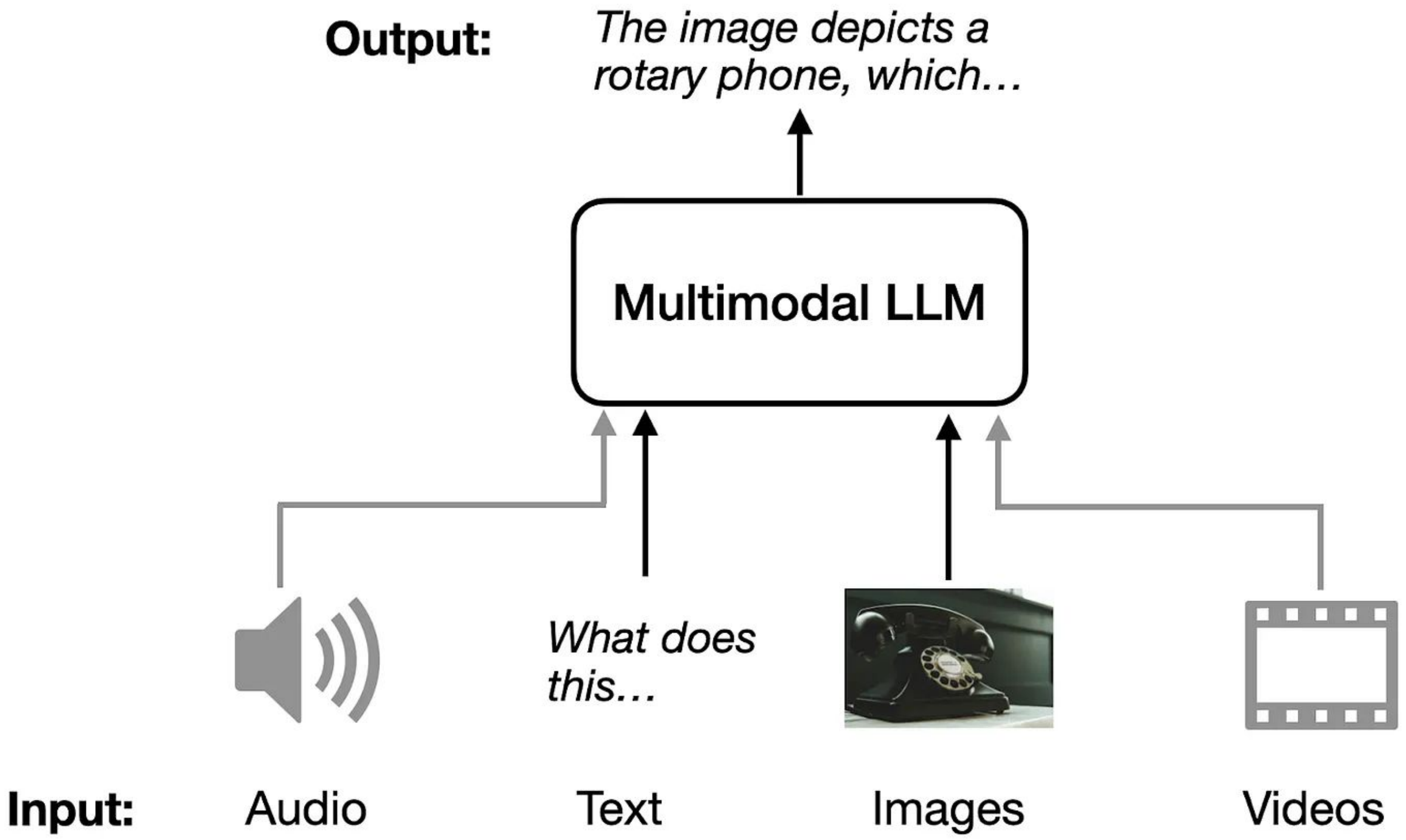


Image credit: Sebastian Raschka

https://updf.com/ocr/thai-ocr-software/

OCR

ASR

Image credit: https://nordicapis.com/5-best-speech-to-text-apis/

Image Captioning & Visual QA

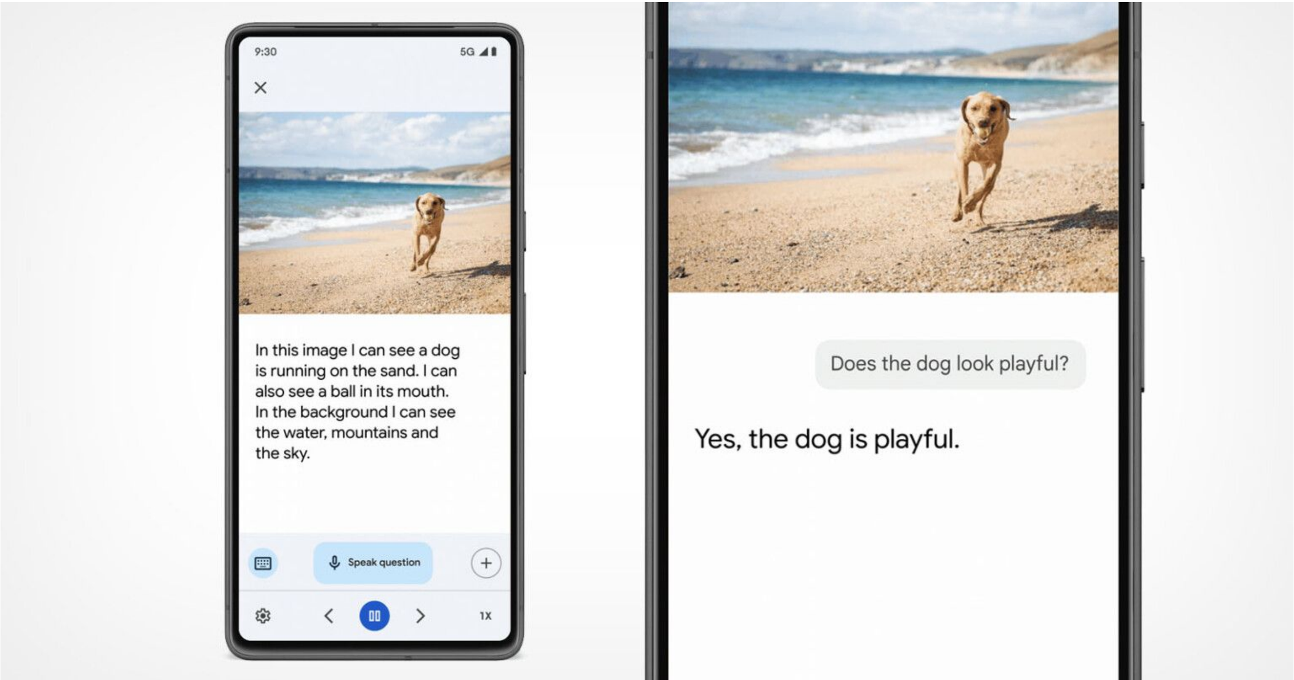
10-09-19

Chrome's new AI feature solves one of the web's eternal problems

To help blind and low-vision users, Google is using machine learning to generate descriptions for millions of images.

Google Can Now Describe and Answer Questions About Images

MAY 18, 2023 JARON SCHNEIDER



ASL

https://www.lifeprint.com/asl101/

ASL

https://www.lifeprint.com/asl101/

Multimodal LLMs possess

the ability to **process** and **generate** information across different data modalities (text, images, audio, video, etc.)

Generate an image of a dessert that looks like hair

Generated by Gemini (Imagen 3)

Image Generation

<text> →

Image Generation

Generate an image of the Songkran festival with people throwing mud instead of water →

Image Generation

<text>

→

Image Generation

Make this happen in
Paris, France instead

→

Ⓐ Problem (input)

You are given two strings s and t, both consisting of lowercase English letters. You are going to type the string s character by character, from the first character to the last one.

When typing a character, instead of pressing the button corresponding to it, you can press the "Backspace" button. It deletes the last character you have typed among those that aren't deleted yet (or does nothing if there are no characters in the current string). For example, if s is "abcbd" and you press Backspace instead of typing the first and the fourth characters, you will get the string "bd" (the first press of Backspace deletes no character, and the second press deletes the character 'c'). Another example, if s is "abcaa" and you press Backspace instead of the last two letters, then the resulting text is "a".

Your task is to determine whether you can obtain the string t, if you type the string s and press "Backspace" instead of typing several (maybe zero) characters of t.

Input

The first line contains a single integer q ($1 \le q \le 10^5$) — the number of test cases.

The first line of each test case contains the string s ($1 \le |s| \le 10^5$). Each character of s is a lowercase English letter.

The second line of each test case contains the string t ($1 \le |t| \le 10^5$). Each character of t is a lowercase English letter.

It is guaranteed that the total number of characters in the strings over all test cases does not exceed $2 \cdot 10^5$.

Output

For each test case, print "YES" if you can obtain the string t by typing the string s and replacing some characters with presses of "Backspace" button, or "NO" if you cannot.

You may print each letter in any case (YES, yes, Yes will all be recognized as positive answer, NO, no and nO will all be recognized as negative answer).

Example

Input	Output
4 ababa ba ababa bb aaa aaaa aababa ababa	YES NO NO YES

Note

Consider the example test from the statement.

In order to obtain "ba" from "ababa", you may press Backspace instead of typing the first and the fourth characters.

There's no way to obtain "bb" while typing "ababa".

There's no way to obtain "aaaa" while typing "aaa".

In order to obtain "ababa" while typing "aababa", you have to press Backspace instead of typing the first character, then type all the remaining characters.

AlphaCode

Ⓑ Solution (output)

```
t=int(input())
for i in range(t):
    s=input()
    t=input()
    a=[]
    b=[]
    for j in s:
        a.append(j)
    for j in t:
        b.append(j)
    a.reverse()
    b.reverse()
    c=[]
    while len(b)!=0 and len(a)!=0:
        if a[0]==b[0]:
            c.append(b.pop(0))
            a.pop(0)
        elif a[0]!=b[0] and len(a)!=1:
            a.pop(0)
            a.pop(0)
        elif a[0]!=b[0] and len(a)==1:
            a.pop(0)
    if len(b)==0:
        print("YES")
    else:
        print("NO")
```

First the solution reads the two phrases.

If the letters at the end of both phrases don't match, the last letter must be deleted. If they do match we can move onto the second last letter and repeat.

If we've matched every letter, it's possible and we output that.

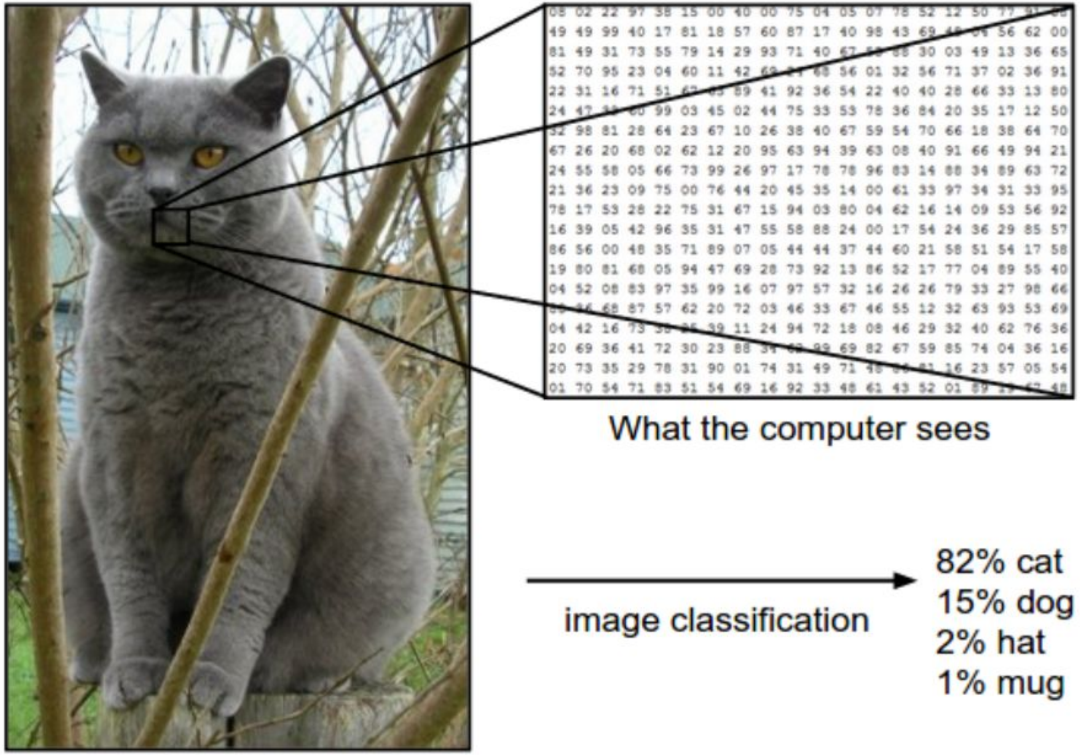
Backspace deletes two letters. The letter you press backspace instead of, and the letter before it.

AlphaFold

This Lecture:
image understanding focus
(where much attention is)
(what I am more aware of)

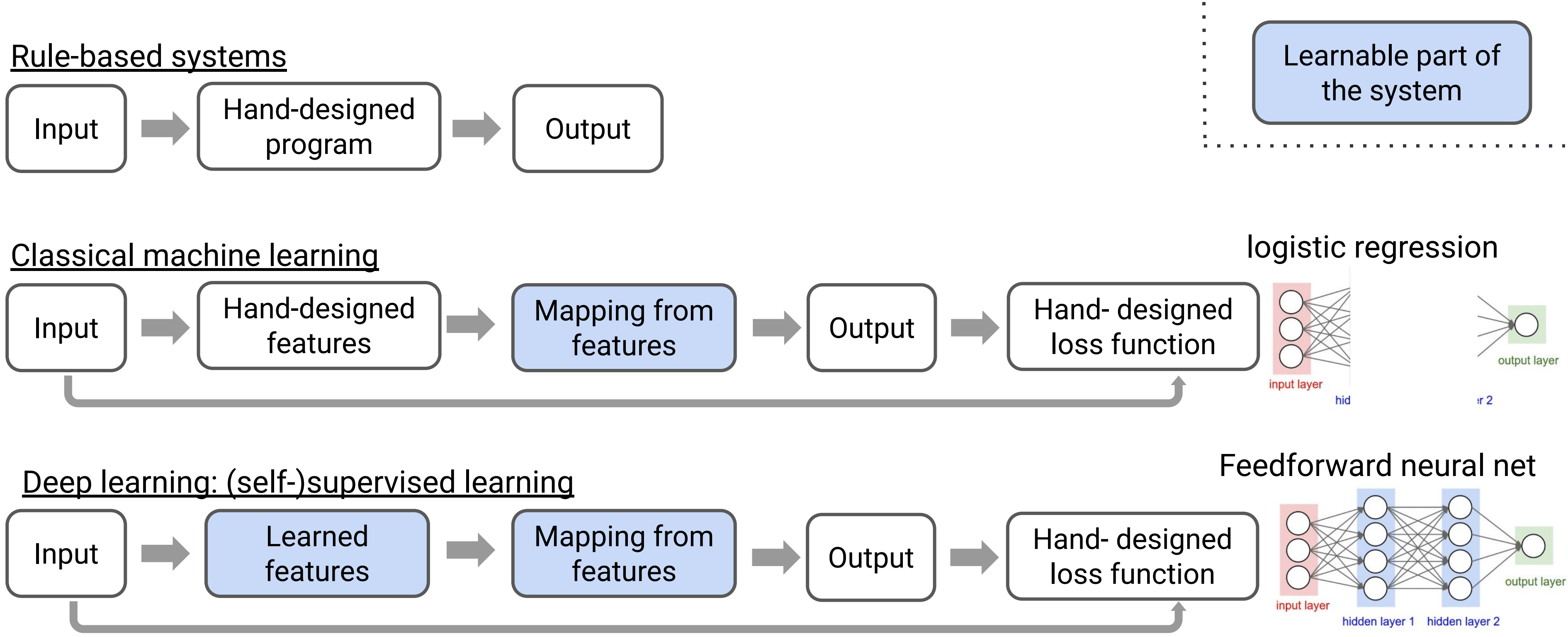
Outline

- Fundamentals & Evolution
- Multimodal LLMs: Modeling, Data, Evaluation
- Example: Gemini



[Image credit: Andrej Karpathy]

From https://www.cs.toronto.edu/~rgrosse/courses/csc311_f20/slides/lec01.pdf



Slide Credit: [Hyung Won Chung]

[PDF] Learning image attributes using the Indian Buffet Process

[PDF] brown.edu

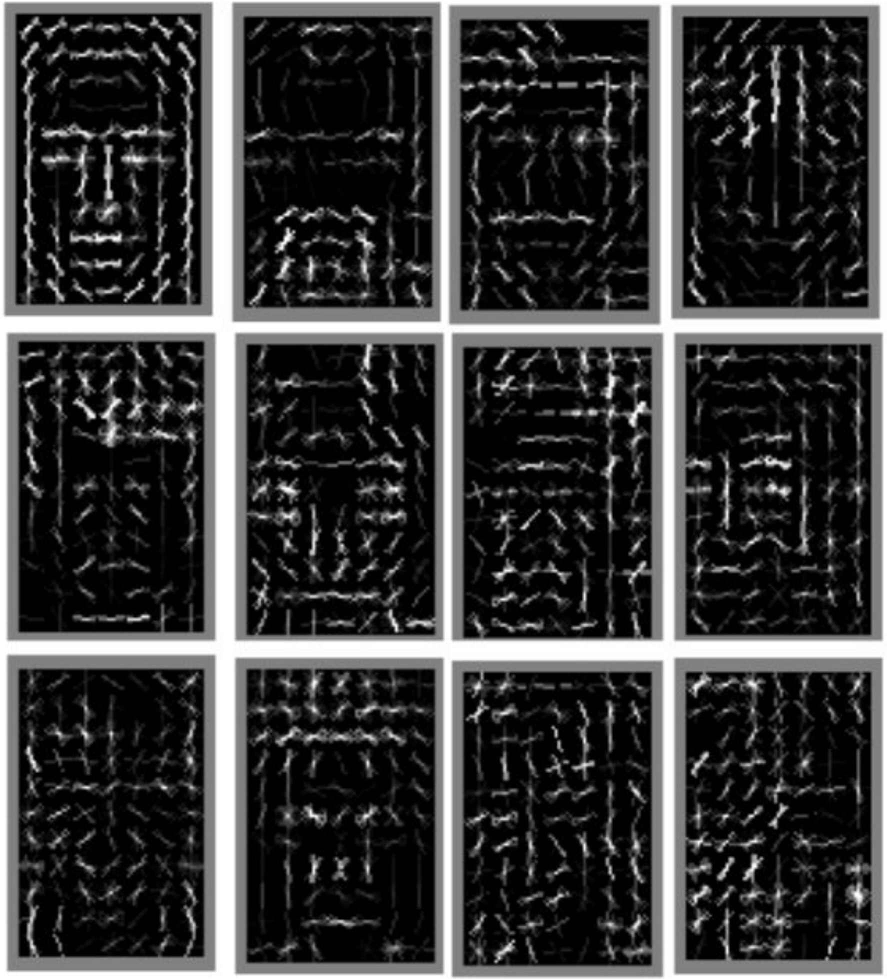
S Changpinyo, E Sudderth

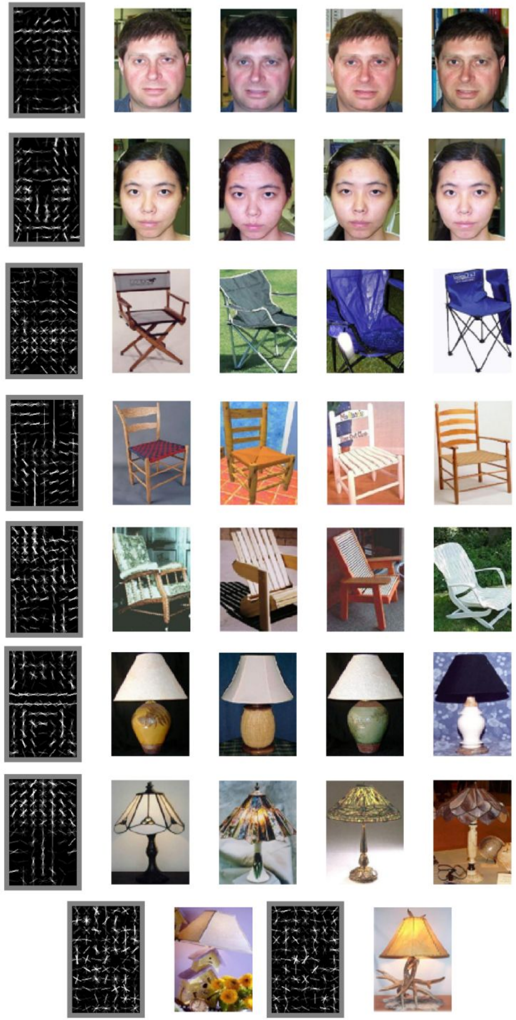
2012 • static.cs.brown.edu

Abstract

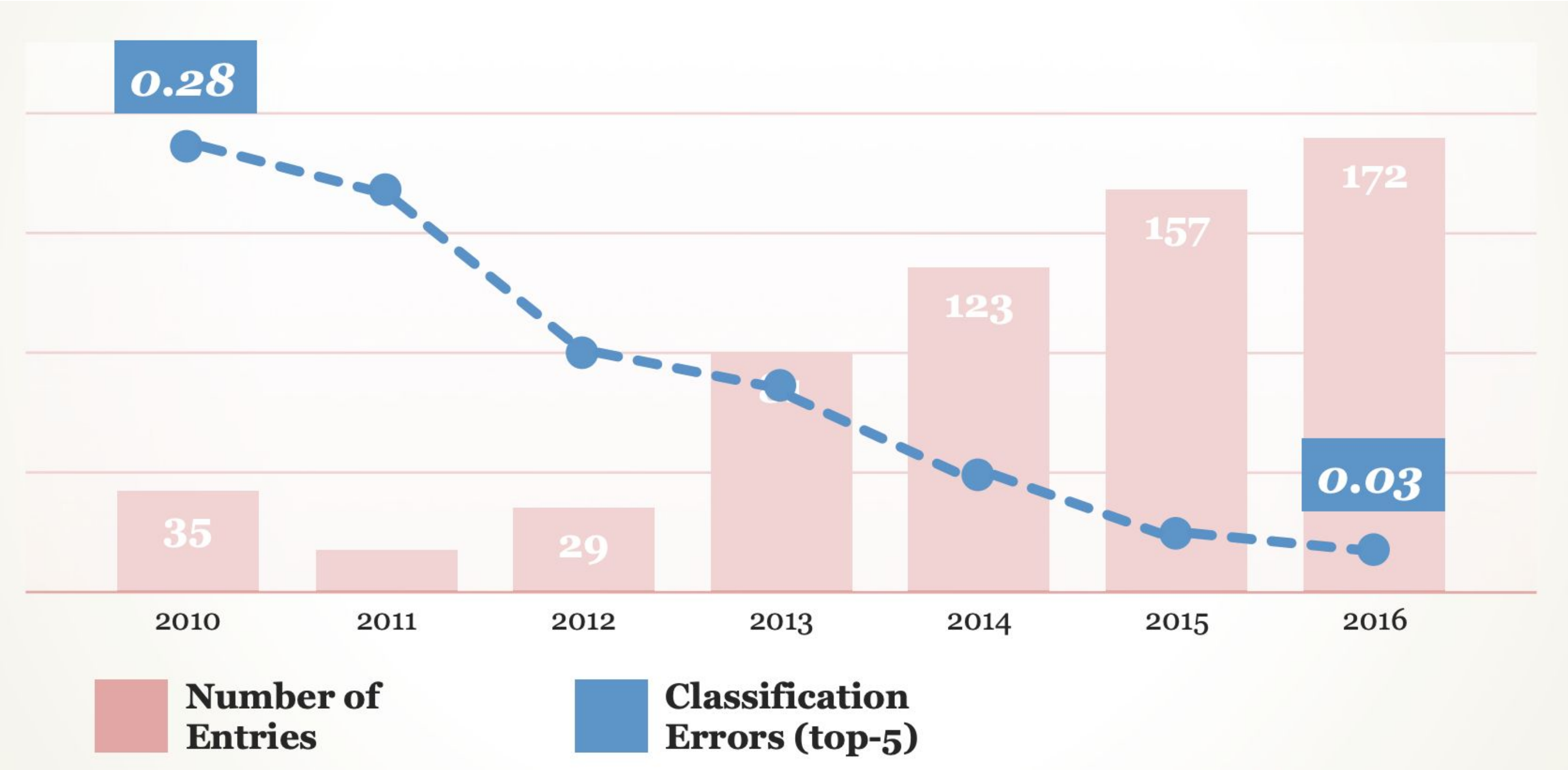
In the domain of object recognition and image classification, a recent trend is to use image properties or attributes to represent the images. Most of the proposed models in the past require that the number of attributes and attribute semantics be specified in advance. In this paper, we propose a generative model for image attributes that combine attribute-based vision models and feature-based nonparamatric models. We learn the model using Gibbs sampling. Qualitatively, we demonstrate the learned attributes of images in three categories. Quantitatively, we show that our model outperforms simple baseline methods in image retrieval and transfer learning tasks.

static.cs.brown.edu





2012



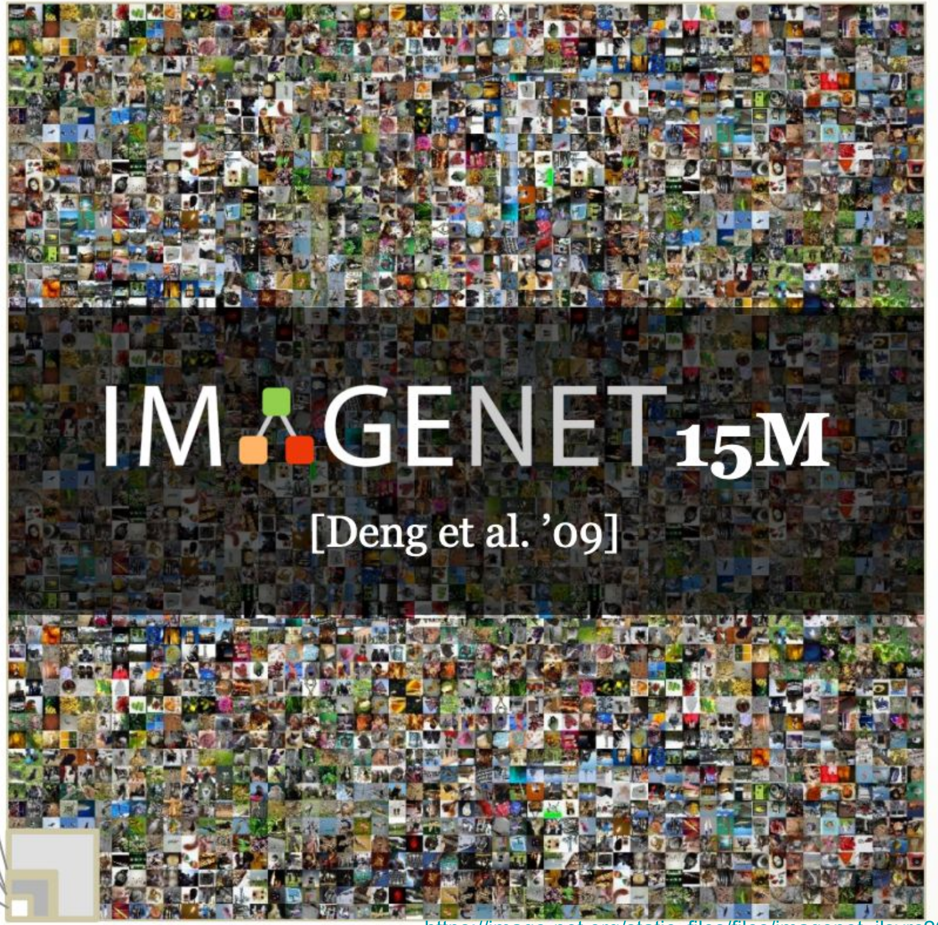
https://image-net.org/static_files/files/imagenet_ilsvrc2017_v1.0.pdf

SUN, 131K
[Xiao et al. '10]

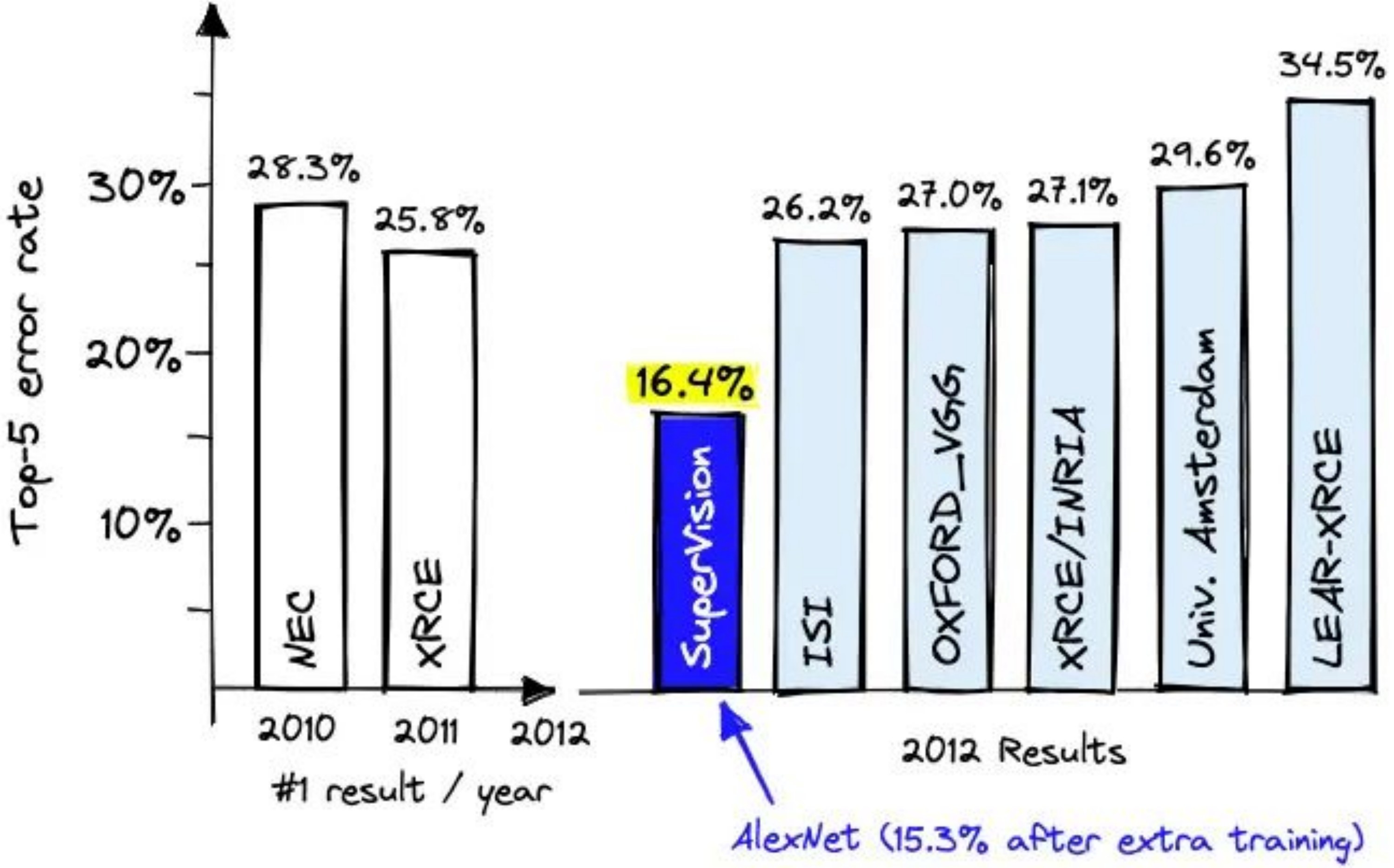
LabelMe, 37K
[Russell et al. '07]

PASCAL VOC, 30K
[Everingham et al. '06-'12]

Caltech101, 9K
[Fei-Fei, Fergus, Perona, '03]

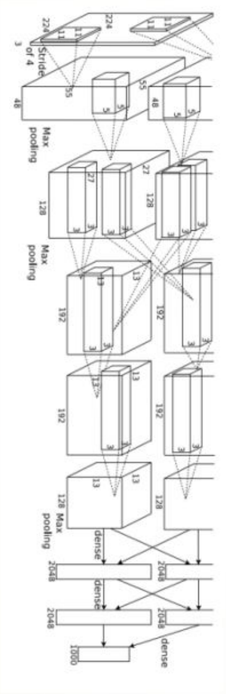


https://image-net.org/static_files/files/imagenet_ilsvrc2017_v1.0.pdf



https://www.pinecone.io/learn/series/image-search/imagenet/

“AlexNet”



[Krizhevsky et al. NIPS 2012]

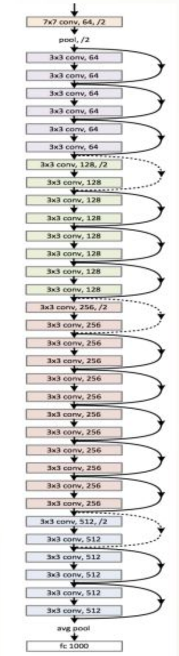
“GoogLeNet”

[Szegedy et al. CVPR 2015]

“VGG Net”

[Simonyan & Zisserman, ICLR 2015]

“ResNet”



[He et al. CVPR 2016]

https://image-net.org/static_files/files/imagenet_ilsvrc2017_v1.0.pdf

Imagenet classification with **deep convolutional** neural **networks**

A Krizhevsky, I Sutskever… - Advances in neural …, 2012 - proceedings.neurips.cc

… We trained a large, **deep convolutional** neural **network** to **classify** the 1.2 million high-resolution images in the **ImageNet** LSVRC-2010 contest into the 1000 different classes. On the test …

☆ Save 99 Cite Cited by 142407 Related articles All 89 versions »

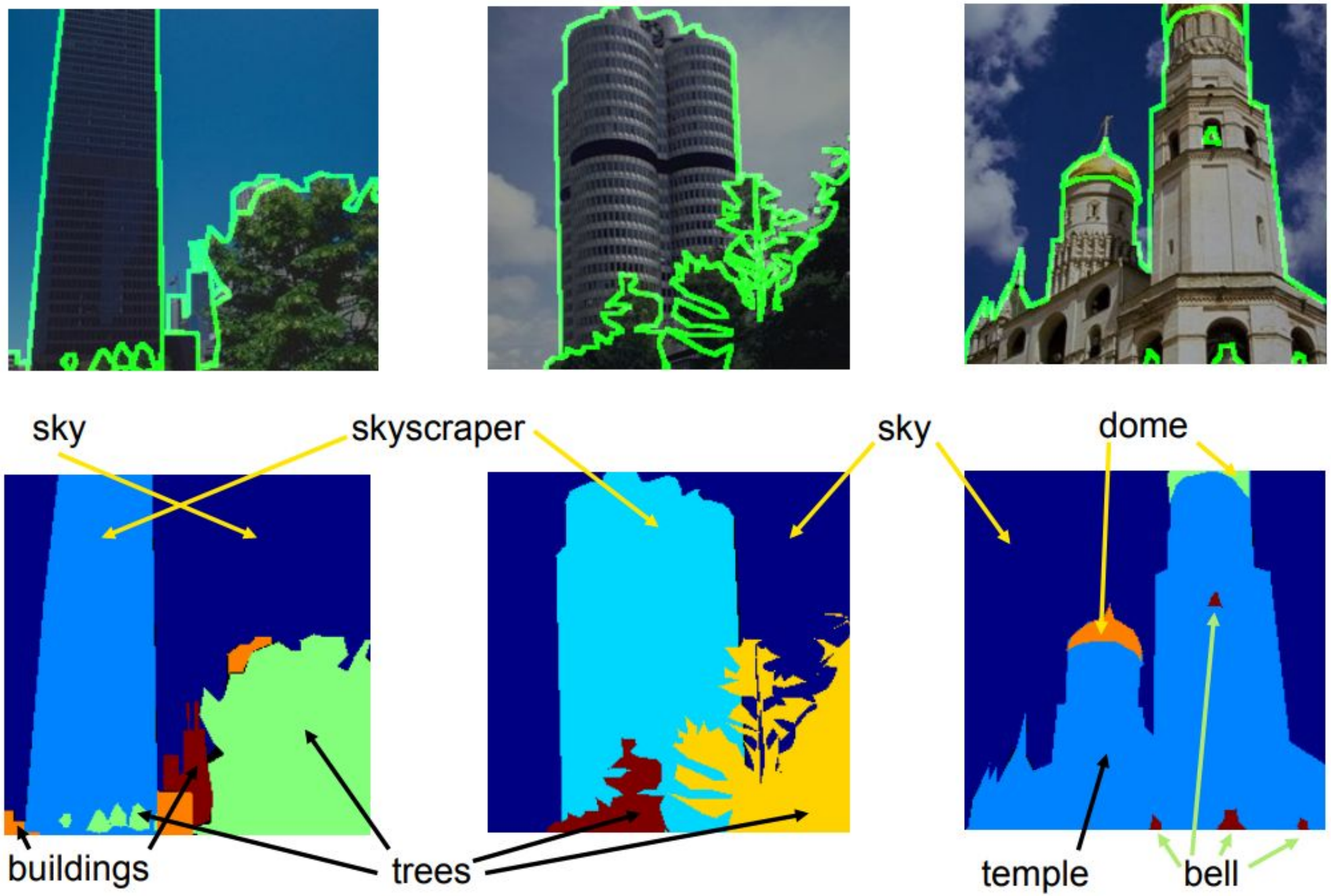
Deep residual learning for **image recognition**

K He, X Zhang, S Ren, J Sun - … and **pattern recognition**, 2016 - openaccess.thecvf.com

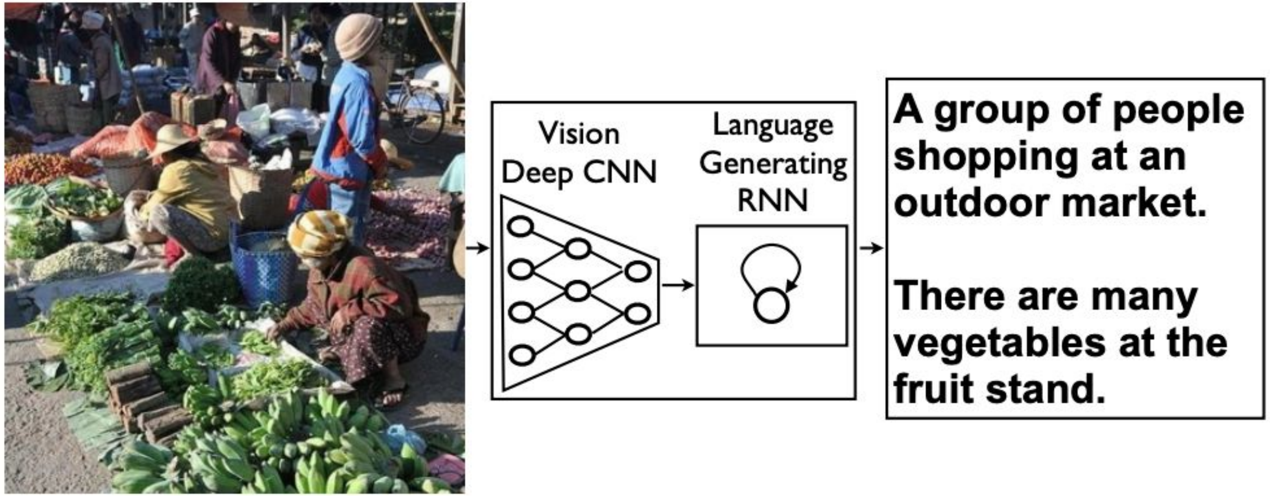
… **Deeper** neural **networks** are more difficult to train. We present a **residual learning** framework to ease the training of **networks** that are substantially **deeper** than those used previously. …

☆ Save 99 Cite Cited by 265023 Related articles All 53 versions »

From https://cs.brown.edu/courses/cs195-5/spring2011/lectures/2011-01-27_overview.pdf



From https://cs.brown.edu/courses/cs195-5/spring2011/lectures/2011-01-27_overview.pdf



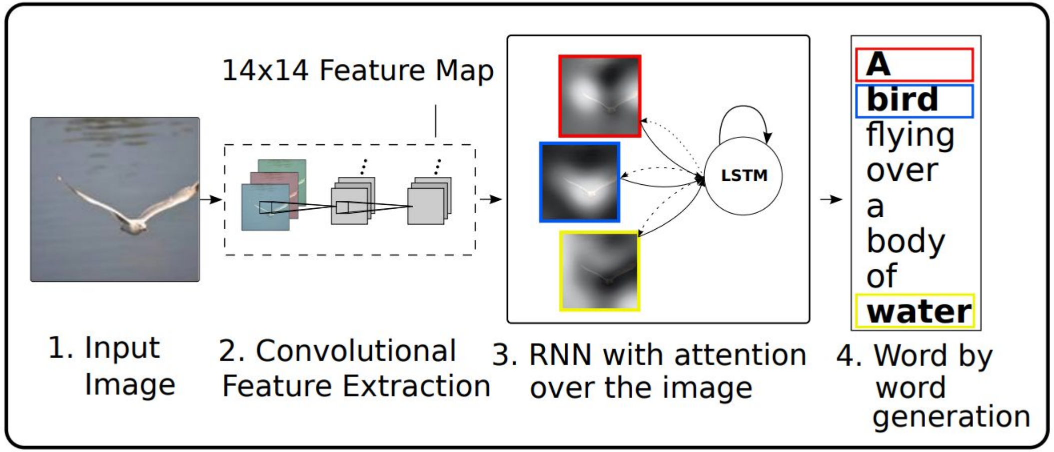
Show and **tell**: A neural image caption generator

O Vinyals, A Toshev, S Bengio… - Proceedings of the IEEE …, 2015 - cv-foundation.org

Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. In this paper, we …

☆ Save 99 Cite Cited by 8225 Related articles All 26 versions »

Figure 1. Our model learns a words/image alignment. The visualized attentional maps (3) are explained in Sections 3.1 & 5.4



Show, attend and tell: Neural image caption generation with visual attention

K Xu, J Ba, R Kiros, K Cho, A Courville… - International …, 2015 - proceedings.mlr.press

… to the task at hand, **and** we **show** how learning to **attend** at different locations in order to … attention mechanism **and** a "soft" deterministic attention mechanism. We also **show** how one …

Aside: My paper in 2013

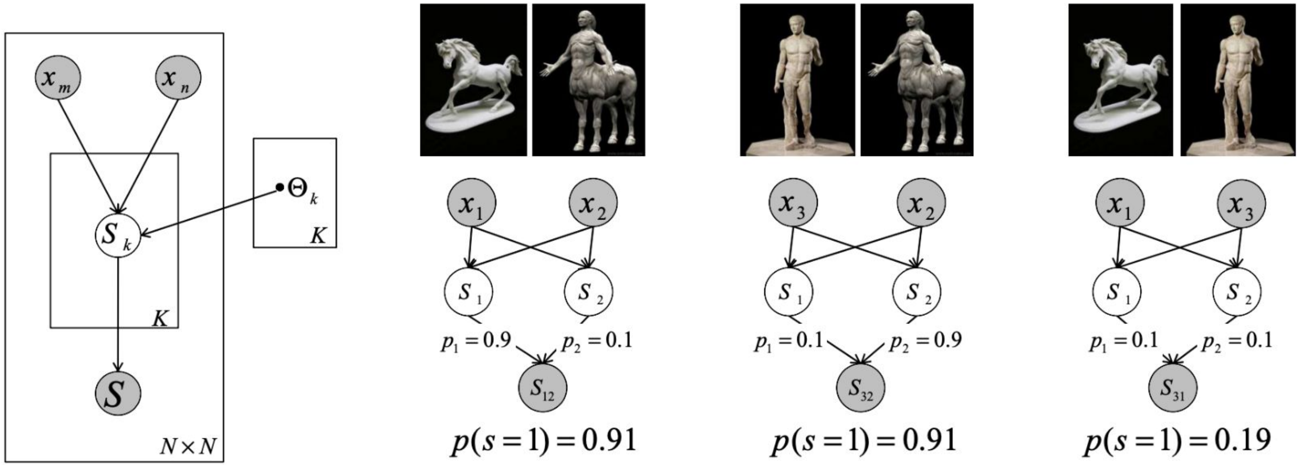


Figure 1: Similarity Component Analysis and its application to the example of CENTAUR, MAN and HORSE. SCA has K latent components which give rise to local similarity values s_k conditioned on a pair of data $\boldsymbol{x}_m$ and $\boldsymbol{x}_n$. The model's output s is a combination of all local values through an OR model (straightforward to extend to a noisy-OR model). $\boldsymbol{\Theta}_k$ is the parameter vector for $p(s_k|\boldsymbol{x}_m, \boldsymbol{x}_n)$. See texts for details.

Aside: My paper in 2016

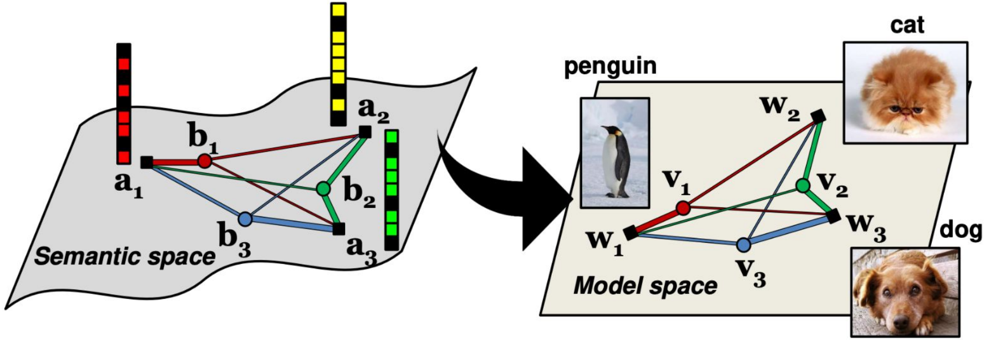


Figure 1: Illustration of our method for zero-shot learning. Object classes live in two spaces. They are characterized in the semantic space with semantic embeddings ($\boldsymbol{a}$s) such as attributes and word vectors of their names. They are also represented as models for visual recognition ($\boldsymbol{w}$s) in the model space. In both spaces, those classes form weighted graphs. The main idea behind our approach is that these two spaces should be aligned. In particular, the coordinates in the model space should be the projection of the graph vertices from the semantic space to the model space — preserving class relatedness encoded i[...] phantom classes ($\boldsymbol{b}$ and $\boldsymbol{v}$) to connect seen and unseen classes — classifiers for the phantom [...] classifiers for real classes. In particular, the synthesis takes the form of convex combination.

Advantage of deep features It is also clear from Table 4 that, across all methods, deep features significantly boost the performance based on shallow features. We use GoogLeNet and AlexNet (numbers in parentheses) and GoogLeNet generally outperforms AlexNet. It is worthwhile to point out that the reported results under

Aside: My paper in 2017

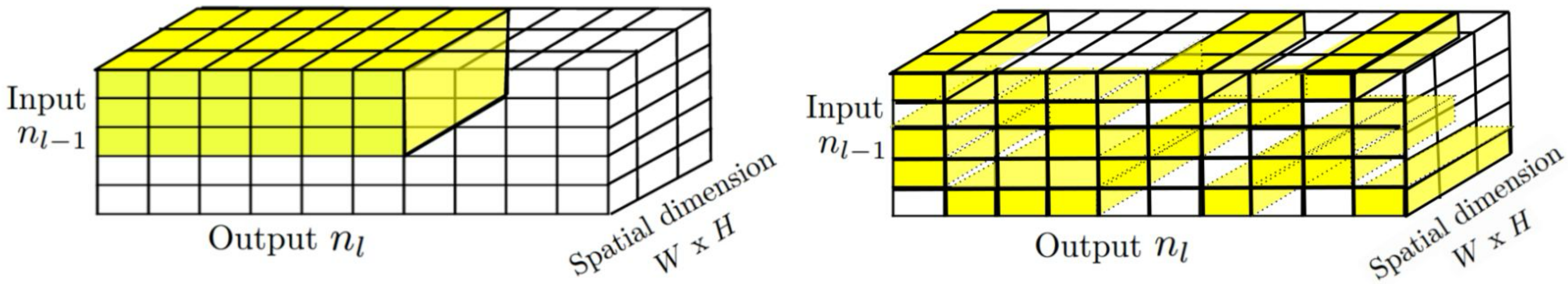
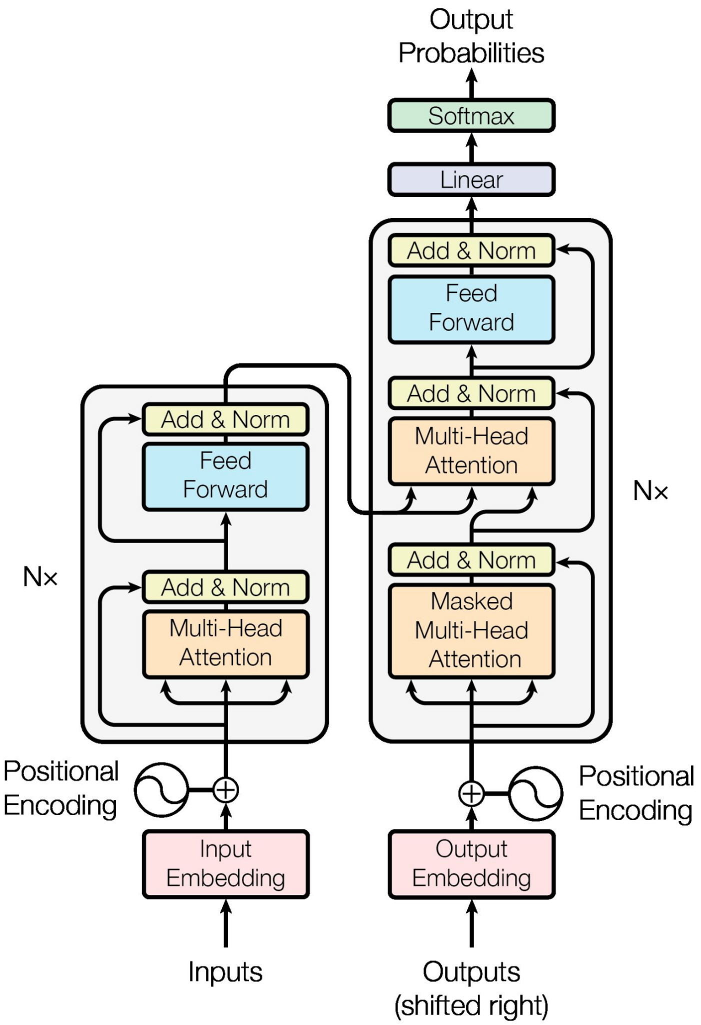


Figure 1: Connection tensors of depth multiplier (left) and sparse random (right) approaches for $n_{l-1} = 5$ and $n_l = 10$. Yellow denotes active connections. For both approaches, the connection pattern is the same across spatial dimension and fixed before training. However, in the sparse random approach, each output channel is connected to a (possibly) different subset of input channels, and vice versa.

2017

What's the most significant event in the AI community in 2017?


Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Add & Norm
Masked
Multi-Head
Attention
Nx
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Positional
Encoding
Positional
Encoding
Input
Embedding
Output
Embedding
Inputs
Outputs
(shifted right)

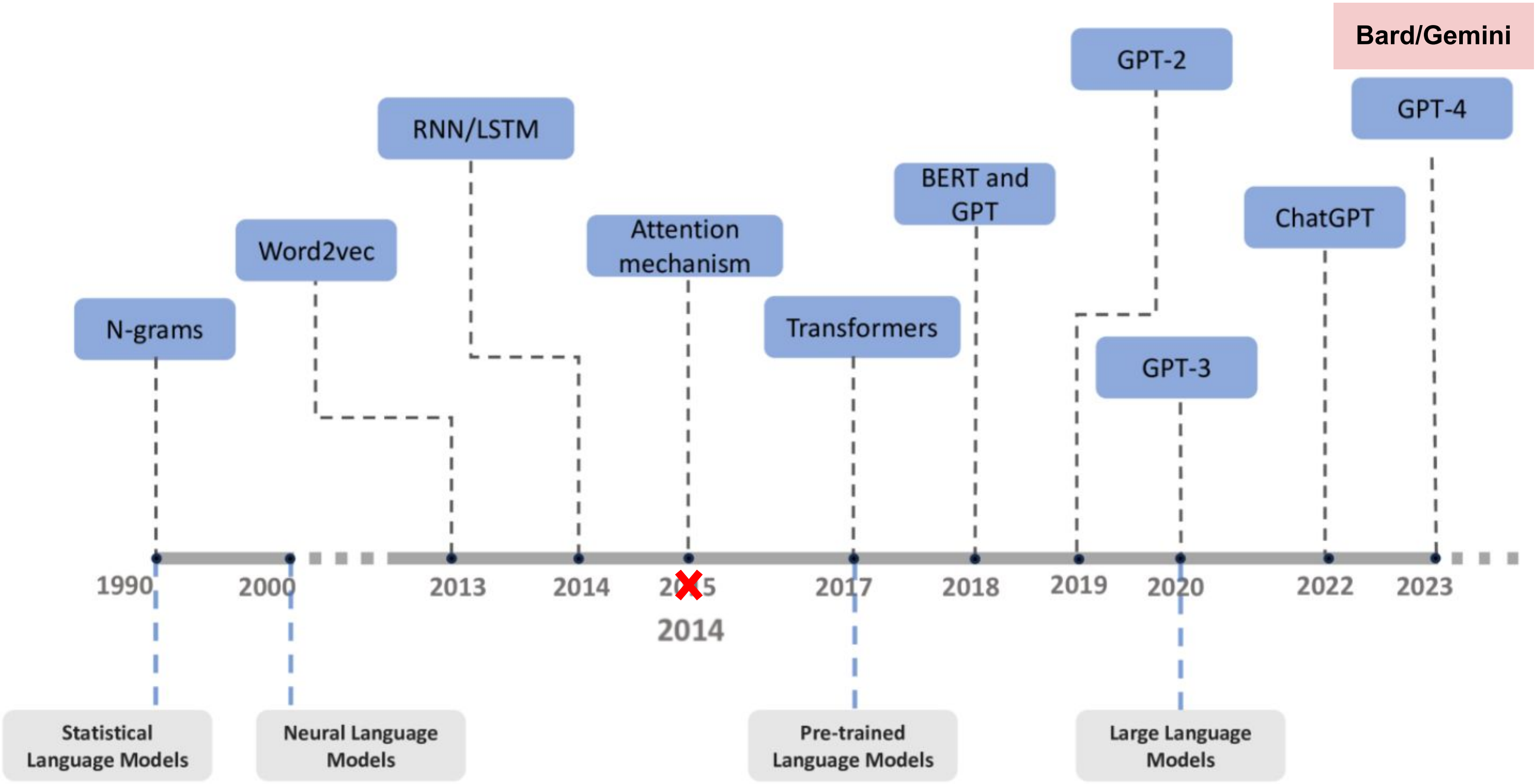
Attention is **all you need**

A Vaswani, N Shazeer, N Parmar… - Advances in neural …, 2017 - proceedings.neurips.cc

… to attend to **all** positions in the decoder up to and including that position. **We need** to prevent … **We** implement this inside of scaled dot-product **attention** by masking out (setting to −∞) …

Abstract

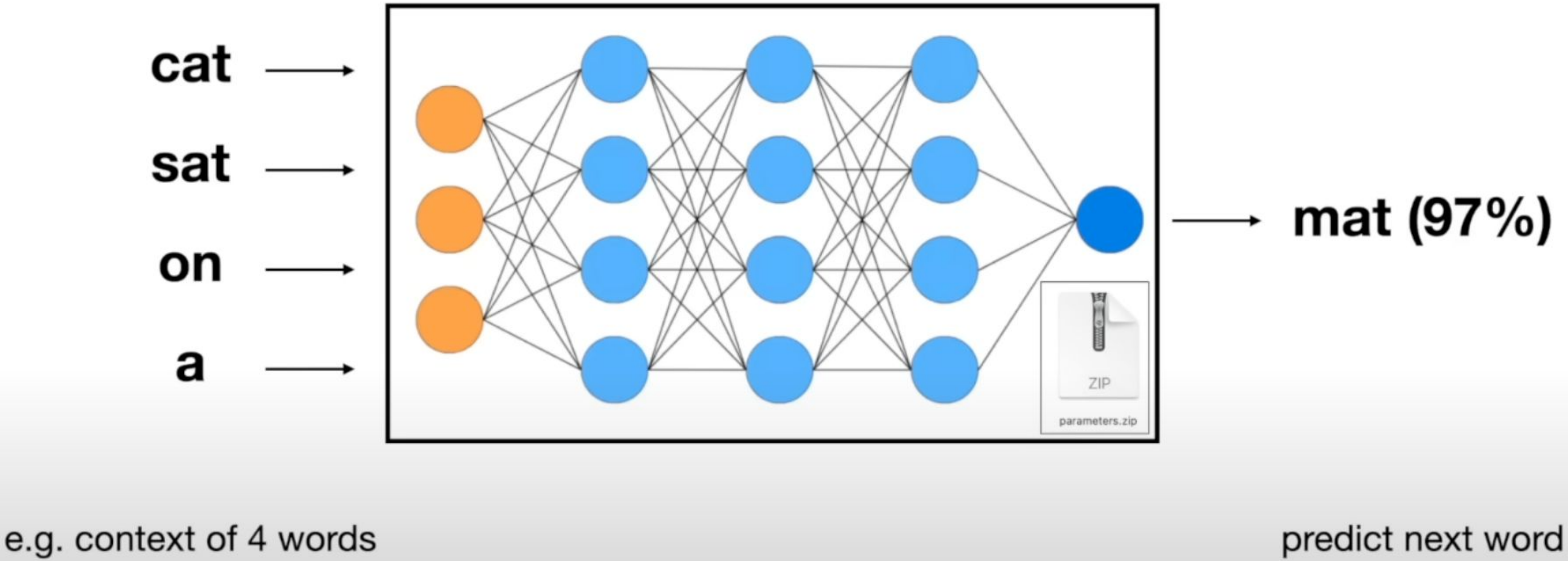
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.



Wang et al. 2024 https://arxiv.org/abs/2402.06853

Neural Network

Predicts the next word in the sequence.



Slide Credit: [Andrej Karpathy]

Next word prediction forces the neural network to learn a lot about the world:

Ruth Marianna Handler (née **Mosko**; November 4, 1916 – April 27, 2002) was an American businesswoman and inventor. She is best known for inventing the Barbie doll in 1959,[2] and being co-founder of toy manufacturer Mattel with her husband Elliot, as well as serving as the company's first president from 1945 to 1975.[3]

The Handlers were forced to resign from Mattel in 1975 after the Securities and Exchange Commission investigated the company for falsifying financial documents.[3][4]

Early life [edit]

Ruth Marianna Mosko[5][2][3] was born on November 4, 1916, in Denver, Colorado, to Polish-Jewish immigrants Jacob Moskowicz, a blacksmith, and Ida Moskowicz, née Rubenstein.[6]

She married her high school boyfriend, Elliot Handler, and moved to Los Angeles in 1938, where she found work at Paramount.[7]

Ruth Handler

Handler in 1961

Born	Ruth Marianna Mosko November 4, 1916 Denver, Colorado, U.S.
Died	April 27, 2002 (aged 85)[1] Los Angeles, California, U.S.

Slide Credit: Andrej Karpathy

Process	Shape
“Many words don't map to one token: indivisible.”	[ ]

Slide Credit: Hyung Won Chung

Process	**Shape**
"Many words don't map to one token: indivisible."	[ ]
⬇ Tokenization	⬇
`Unicode characters like emojis may be split.`	`[length]`
`[7085, 2456, 836, 470, 3975, 284, 530, 11241, 25, 773, 452, 12843, 13]`	

Slide Credit: Hyung Won Chung

Process

"Many words don't map to one token: indivisible."

⬇ Tokenization

`Unicode characters like emojis may be split.`

`[7085, 2456, 836, 470, 3975, 284, 530, 11241, 25, 773, 452, 12843, 13]`

⬇ Embedding

2.3	-3.2	8.3	5.4	2.1	3.9	-8.9	3.8	3.9	3.3
4.5	5.9	4.5	7.1	1.0	5.3	5.0	3.1	0.7	5.0
…	…	…	…	…	…	…	…	…	…
3.8	1.2	3.8	9.0	9.3	3.1	4.2	0.8	9.2	5.8

Shape

`[ ]`

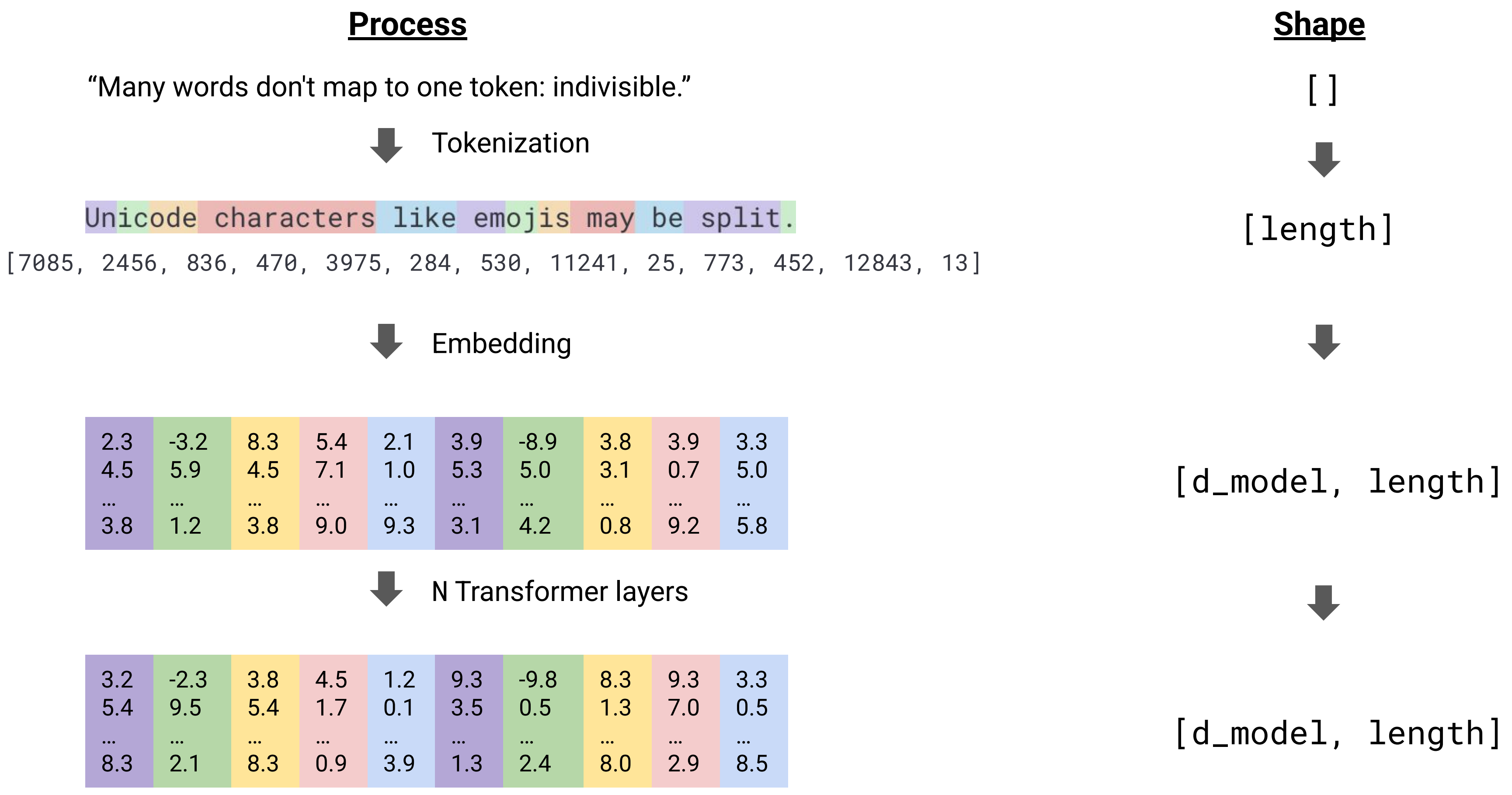
⬇

`[length]`

⬇

`[d_model, length]`

Slide Credit: Hyung Won Chung



Slide Credit: [Hyung Won Chung]()

Process	**Shape**
“Many words don't map to one token: indivisible.”	[ ]
⬇ Tokenization	⬇
`Unicode characters like emojis may be split.` `[7085, 2456, 836, 470, 3975, 284, 530, 11241, 25, 773, 452, 12843, 13]`	`[length]`
⬇ Embedding	⬇

2.3	-3.2	8.3	5.4	2.1	3.9	-8.9	3.8	3.9	3.3
4.5	5.9	4.5	7.1	1.0	5.3	5.0	3.1	0.7	5.0
…	…	…	…	…	…	…	…	…	…
3.8	1.2	3.8	9.0	9.3	3.1	4.2	0.8	9.2	5.8

`[d_model, length]`

⬇ N Transformer layers ⬇

3.2	-2.3	3.8	4.5	1.2	9.3	-9.8	8.3	9.3	3.3
5.4	9.5	5.4	1.7	0.1	3.5	0.5	1.3	7.0	0.5
…	…	…	…	…	…	…	…	…	…
8.3	2.1	8.3	0.9	3.9	1.3	2.4	8.0	2.9	8.5

`[d_model, length]`

⬇ Loss function (predict next token given previous) ⬇

`2.6`

`[ ]`

Slide Credit: Hyung Won Chung

Batched Process

Many words don't map to one token: indivisible.

↓ Tokenization

Many words don't map to one token: indivisible.
Unicode characters like emojis may be split.

[[7085, 2456, 836, 470, 3975, 284, 530, 11241, 25, 773, 452, 12843, 13]
 [3118, 291, 1098, 3435, 588, 795, 13210, 271, 743, 307, 6626]]

↓ Embedding

8.2 2.0 6.9 9.1 8.1 3.1 5.1 4.4 3.7 0.1
2.3 -3.2 8.3 5.4 2.1 3.9 -8.9 3.8 3.9 3.3 0
4.5 5.9 4.5 7.1 1.0 5.3 5.0 3.1 0.7 5.0
… … … … … … … … … … 8
3.8 1.2 3.8 9.0 9.3 3.1 4.2 0.8 9.2 5.8

↓ N Transformer layers

8.2 2.0 6.9 9.1 8.1 3.1 5.1 4.4 3.7 0.1
3.2 -2.3 3.8 4.5 1.2 9.3 -9.8 8.3 9.3 3.3 0
5.4 9.5 5.4 1.7 0.1 3.5 0.5 1.3 7.0 0.5
… … … … … … … … … … 8
8.3 2.1 8.3 0.9 3.9 1.3 2.4 8.0 2.9 8.5

↓ Loss function (predict next token given previous)

2.6

Batched Shape

[batch]

↓

[batch, length]

↓

[batch, d_model, length]

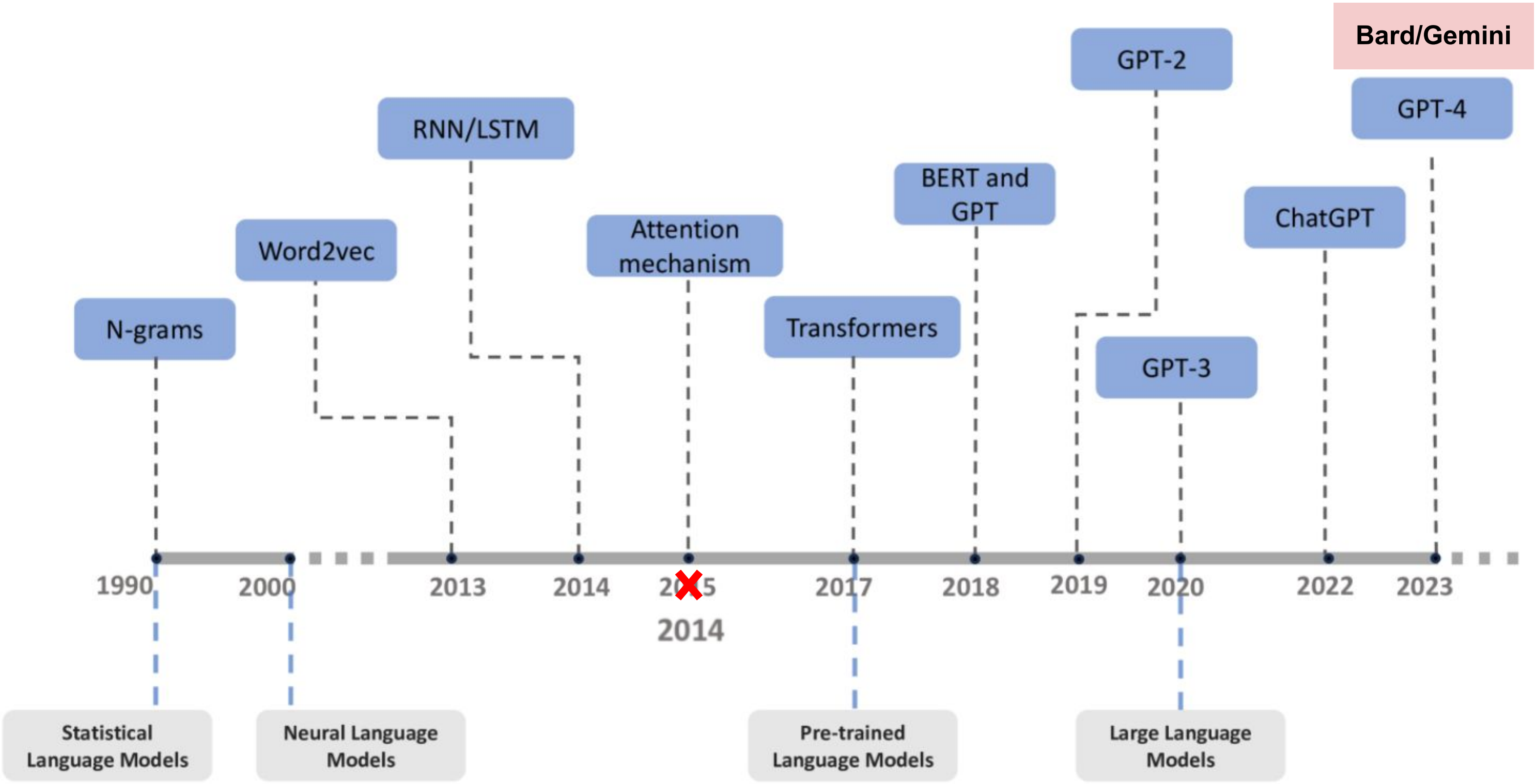
↓

[batch, d_model, length]

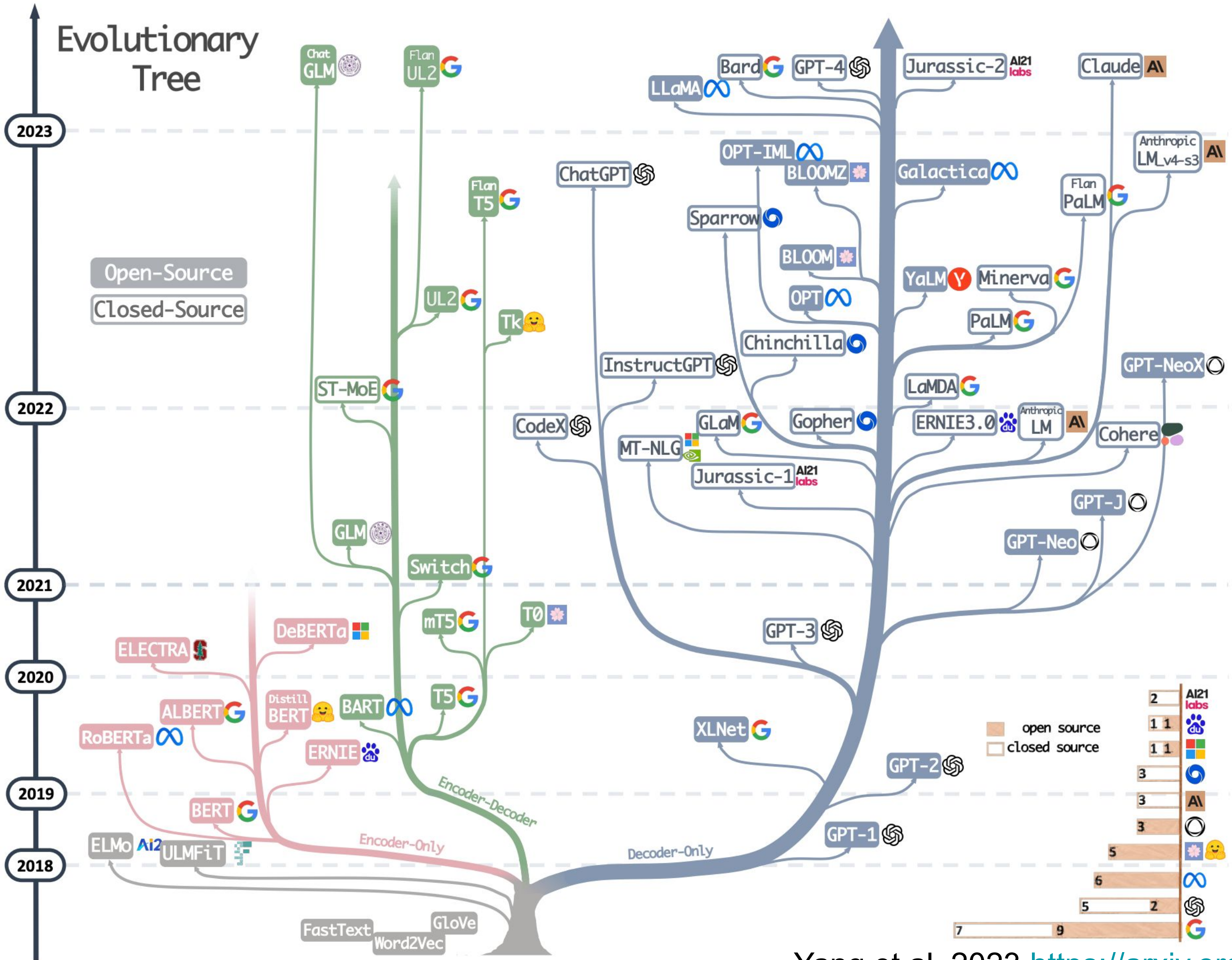
↓

[]

Slide Credit: Hyung Won Chung

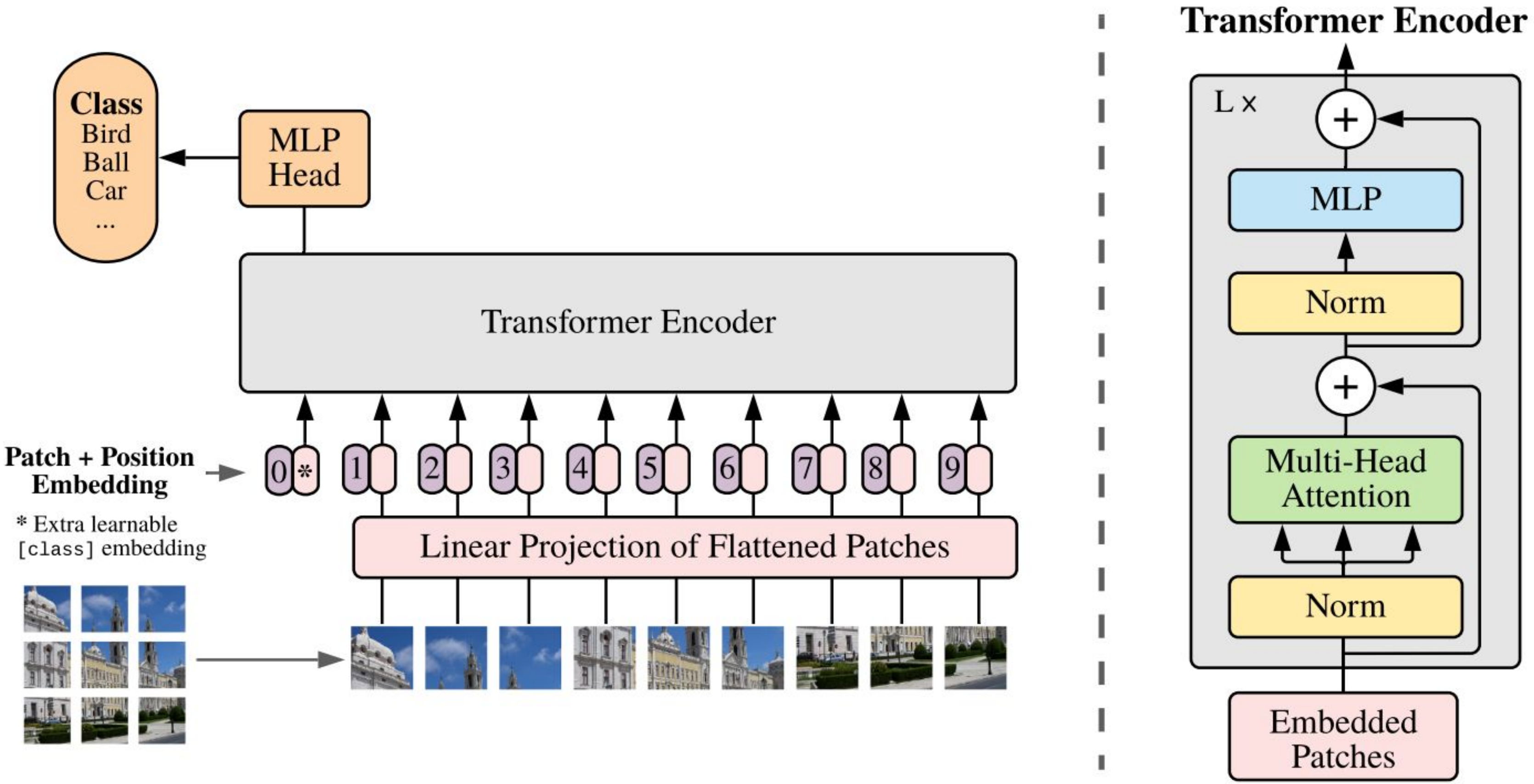


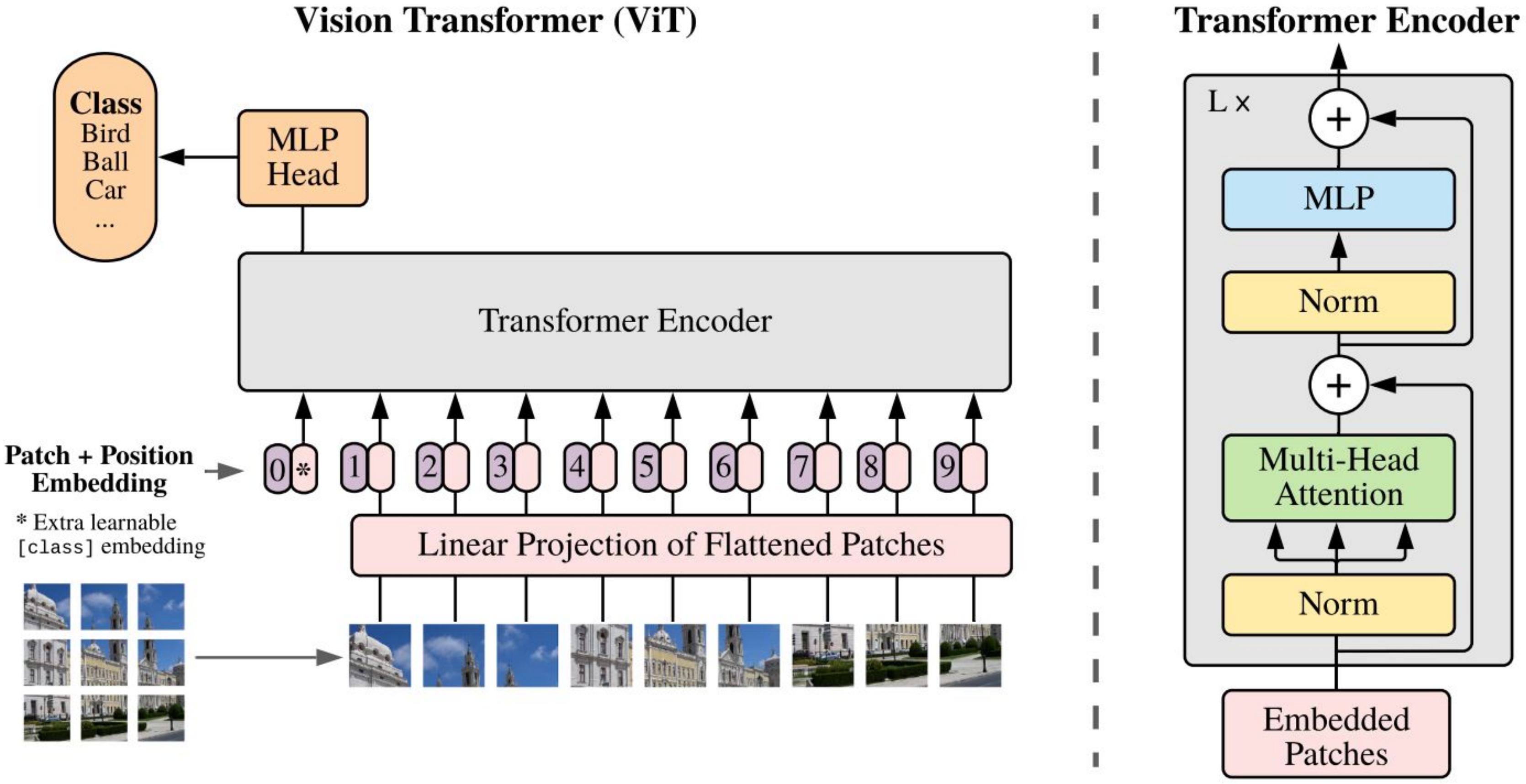
Wang et al. 2024 https://arxiv.org/abs/2402.06853



Yang et al. 2023 https://arxiv.org/abs/2304.13712

2020


Class
Bird
Ball
Car
...
MLP
Head
Transformer Encoder
Patch + Position
Embedding
* Extra learnable
[class] embedding
0 *
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches
Transformer Encoder
L ×
MLP
Norm
Multi-Head
Attention
Norm
Embedded
Patches


Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0 *
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches
Transformer Encoder
L ×
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy[*,†], **Lucas Beyer**[*], **Alexander Kolesnikov**[*], **Dirk Weissenborn**[*], **Xiaohua Zhai**[*], **Thomas Unterthiner**, **Mostafa Dehghani**, **Matthias Minderer**, **Georg Heigold**, **Sylvain Gelly**, **Jakob Uszkoreit**, **Neil Houlsby**[*,†]

[*]equal technical contribution, [†]equal advising

Google Research, Brain Team

{adosovitskiy, neilhoulsby}@google.com

ABSTRACT

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.[1]

[PDF] An image is worth **16x16** words: **Transformers** for image recognition at scale

A Dosovitskiy, L Beyer, A Kolesnikov… - arXiv preprint arXiv …, 2020 - arxiv.org

… To begin to understand how the **Vision Transformer** processes image data, we analyze its internal representations. The first layer of the **Vision Transformer** linearly projects the flattened …

☆ Save 99 Cite Cited by 60237 Related articles All 21 versions »

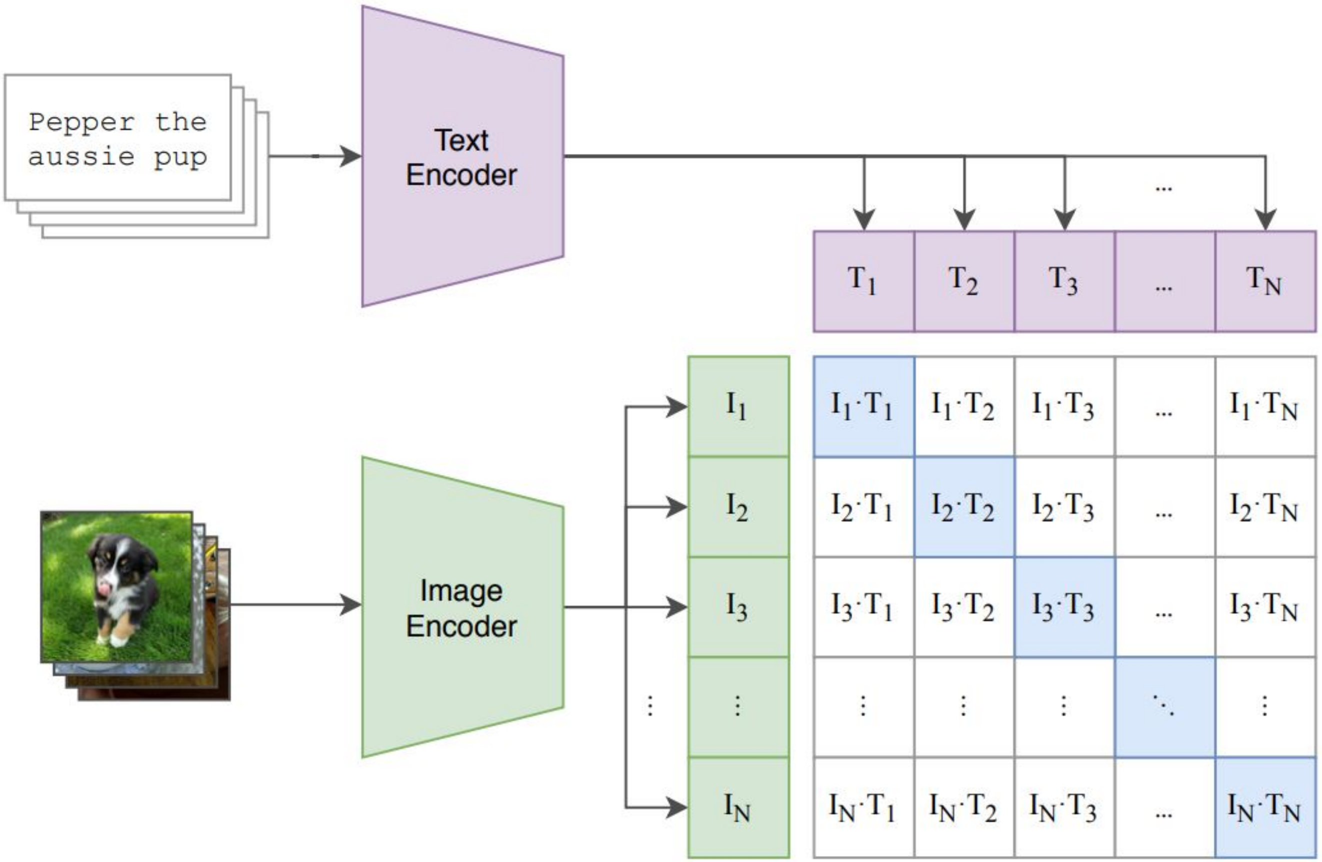
Learning transferable visual models from **natural language supervision**

A Radford, JW Kim, C Hallacy… - … machine **learning**, 2021 - proceedings.mlr.press

… We speculate this is due to **natural language** providing wider **supervision** for **visual** concepts involving verbs, compared to the noun-centric object **supervision** in ImageNet. …

☆ Save 99 Cite Cited by 33015 Related articles All 22 versions »

(1) Contrastive pre-training



Summary

- **Vision: CNNs** significantly advanced state-of-the-art **image recognition** in **2012**, then going on to improve other computer vision related tasks
- **Vision**+**Language**: **Attention mechanisms** developed in **2014-2015**
- **Language: Transformers** started to replace **RNNs** for language tasks since **2017**, starting with **machine translation**
- **Vision:** Transformers as image encoders, **VITs**, popularized in **2020**
- **Vision**+**Language** took whatever worked best **at the moment**

Multimodal LLMs
Modeling & Data & Evaluation

The Revolution of Multimodal Large Language Models: A Survey

Davide Caffagni[1*], Federico Cocchi[1,2*], Luca Barsellotti[1*], Nicholas Moratelli[1*], Sara Sarto[1*], Lorenzo Baraldi[2*], Lorenzo Baraldi[1], Marcella Cornia[1], and Rita Cucchiara[1,3]

[1]University of Modena and Reggio Emilia, Italy

[2]University of Pisa, Italy

[3]IIT-CNR, Italy

[1]{name.surname}@unimore.it [2]{name.surname}@phd.unipi.it

Model	LLM	Visual Encoder	V2L Adapter	VInstr. Tuning	Main Tasks & Capabilities
BLIP-2 (Li et al., 2023g)	FlanT5-XXL-11B★	EVA ViT-g	Q-Former	✗	Visual Dialogue, VQA, Captioning, Retrieval
FROMAGe (Koh et al., 2023b)	OPT-6.7B★	CLIP ViT-L	Linear	✗	Visual Dialogue, Captioning, Retrieval
Kosmos-1 (Huang et al., 2023b)	Magneto-1.3B◊	CLIP ViT-L	Q-Former*	✗	Visual Dialogue, VQA, Captioning
LLaMA-Adapter V2 (Gao et al., 2023)	LLaMA-7B▲	CLIP ViT-L	Linear	✗	VQA, Captioning
OpenFlamingo (Awadalla et al., 2023)	MPT-7B★	CLIP ViT-L	XAttn LLM	✗	VQA, Captioning
Flamingo (Alayrac et al., 2022)	Chinchilla-70B★	NFNet-F6	XAttn LLM	✗	Visual Dialogue, VQA, Captioning
PaLI (Chen et al., 2023j)	mT5-XXL-13B♦	ViT-e	XAttn LLM	✗	Multilingual, VQA, Captioning, Retrieval
PaLI-X (Chen et al., 2023h)	UL2-32B♦	ViT-22B	XAttn LLM	✗	Multilingual, VQA, Captioning
LLaVA (Liu et al., 2023e)	Vicuna-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning
MiniGPT-4 (Zhu et al., 2023a)	Vicuna-13B★	EVA ViT-g	Linear	✓	VQA, Captioning
mPLUG-Owl (Ye et al., 2023c)	LLaMA-7B▲	CLIP ViT-L	Q-Former*	✓	Visual Dialogue, VQA
InstructBLIP (Dai et al., 2023)	Vicuna-13B★	EVA ViT-g	Q-Former	✓	Visual Dialogue, VQA, Captioning
MultiModal-GPT (Gong et al., 2023)	LLaMA-7B▲	CLIP ViT-L	XAttn LLM	✓	Visual Dialogue, VQA, Captioning
LaVIN (Luo et al., 2023)	LLaMA-13B▲	CLIP ViT-L	MLP	✓	Visual Dialogue, VQA, Captioning
Otter (Li et al., 2023b)	LLaMA-7B★	CLIP ViT-L	XAttn LLM	✓	VQA, Captioning
Kosmos-2 (Peng et al., 2023)	Magneto-1.3B◊	CLIP ViT-L	Q-Former*	✓	Visual Dialogue, VQA, Captioning, Referring, REC
Shikra (Chen et al., 2023f)	Vicuna-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
Clever Flamingo (Chen et al., 2023b)	LLaMA-7B▲	CLIP ViT-L	XAttn LLM	✓	Visual Dialogue, VQA, Captioning
SVIT (Zhao et al., 2023a)	Vicuna-13B♦	CLIP ViT-L	MLP	✓	Visual Dialogue, VQA, Captioning
BLIVA (Hu et al., 2024)	Vicuna-7B★	EVA ViT-g	Q-Former+Linear	✓	Visual Dialogue, VQA, Captioning
IDEFICS (Laurençon et al., 2024)	LLaMA-65B★	OpenCLIP ViT-H	XAttn LLM	✓	Visual Dialogue, VQA, Captioning
Qwen-VL (Bai et al., 2023b)	Qwen-7B♦	OpenCLIP ViT-bigG	Q-Former*	✓	Visual Dialogue, Multilingual, VQA, Captioning, REC
StableLLaVA (Li et al., 2023i)	Vicuna-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning
Ferret (You et al., 2023)	Vicuna-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, Captioning, Referring, REC, GroundCap
LLaVA-1.5 (Liu et al., 2023d)	Vicuna-13B♦	CLIP ViT-L	MLP	✓	Visual Dialogue, VQA, Captioning
MiniGPT-v2 (Chen et al., 2023e)	LLaMA-2-7B▲	EVA ViT-g	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
Pink (Xuan et al., 2023)	Vicuna-7B▲	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
CogVLM (Wang et al., 2023c)	Vicuna-7B♦	EVA ViT-E	MLP	✓	Visual Dialogue, VQA, Captioning, REC
DRESS (Chen et al., 2023l)	Vicuna-13B▲	EVA ViT-g	Linear	✓	Visual Dialogue, VQA, Captioning
LION (Chen et al., 2023d)	FlanT5-XXL-11B★	EVA ViT-g	Q-Former+MLP	✓	Visual Dialogue, VQA, Captioning, REC
mPLUG-Owl2 (Ye et al., 2023d)	LLaMA-2-7B♦	CLIP ViT-L	Q-Former*	✓	Visual Dialogue, VQA, Captioning
SPHINX (Lin et al., 2023b)	LLaMA-2-13B♦	Mixture	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
Honeybee (Cha et al., 2023)	Vicuna-13B♦	CLIP ViT-L	ResNet blocks	✓	Visual Dialogue, VQA, Captioning
VILA (Lin et al., 2023a)	LLaMA-2-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning
SPHINX-X (Gao et al., 2024)	Mixtral-8×7B♦	Mixture	Linear	✓	Visual Dialogue, Multilingual, VQA, Captioning, Referring, REC

Table 1: Summary of generalist MLLMs for vision-to-language tasks. For each model, we indicate the LLM used in its best configuration as shown in the original paper (◊: LLM training from scratch; ♦: LLM fine-tuning; ▲: LLM fine-tuning with PEFT techniques; ★: frozen LLM). The * marker indicates variants to the reported vision-to-language adapter, while gray color indicates models not publicly available.

Caffagni et al. 2024 https://arxiv.org/abs/2402.12451

Model	LLM	Visual Encoder	V2L Adapter	VInstr. Tuning
BLIP-2 (Li et al., 2023g)	FlanT5-XXL-11B★	EVA ViT-g	Q-Former	✗
FROMAGe (Koh et al., 2023b)	OPT-6.7B★	CLIP ViT-L	Linear	✗
Kosmos-1 (Huang et al., 2023b)	Magneto-1.3B◊	CLIP ViT-L	Q-Former*	✗
LLaMA-Adapter V2 (Gao et al., 2023)	LLaMA-7B▲	CLIP ViT-L	Linear	✗
OpenFlamingo (Awadalla et al., 2023)	MPT-7B★	CLIP ViT-L	XAttn LLM	✗
Flamingo (Alayrac et al., 2022)	Chinchilla-70B★	NFNet-F6	XAttn LLM	✗
PaLI (Chen et al., 2023j)	mT5-XXL-13B♦	ViT-e	XAttn LLM	✗
PaLI-X (Chen et al., 2023h)	UL2-32B♦	ViT-22B	XAttn LLM	✗

Caffagni et al. 2024 https://arxiv.org/abs/2402.12451

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Colin Raffel*	CRAFFEL@GMAIL.COM
Noam Shazeer*	NOAM@GOOGLE.COM
Adam Roberts*	ADAROB@GOOGLE.COM
Katherine Lee*	KATHERINELEE@GOOGLE.COM
Sharan Narang	SHARANNARANG@GOOGLE.COM
Michael Matena	MMATENA@GOOGLE.COM
Yanqi Zhou	YANQIZ@GOOGLE.COM
Wei Li	MWEILI@GOOGLE.COM
Peter J. Liu	PETERJLIU@GOOGLE.COM

Google, Mountain View, CA

mT5

mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer

Linting Xue* **Noah Constant*** **Adam Roberts***
Mihir Kale **Rami Al-Rfou** **Aditya Siddhant** **Aditya Barua** **Colin Raffel**
Google Research

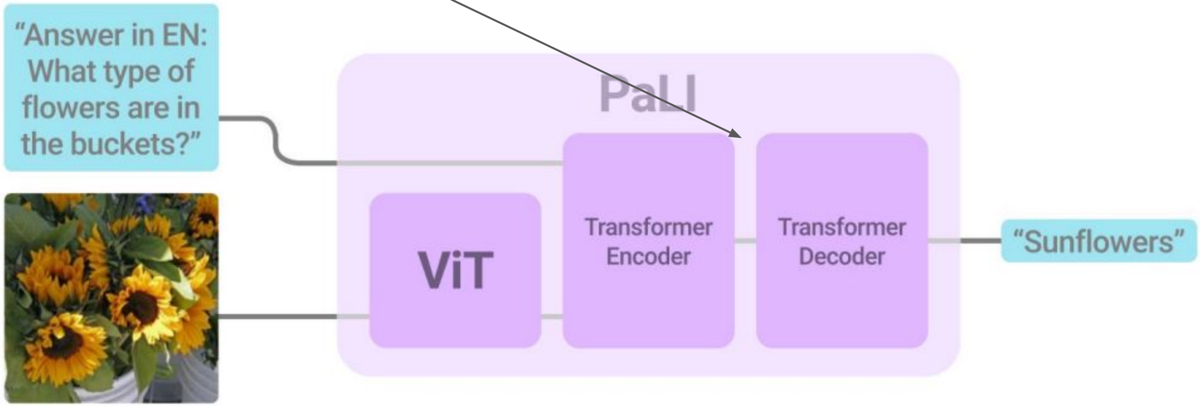
ViT

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy*,†, Lucas Beyer*, Alexander Kolesnikov*, Dirk Weissenborn*, Xiaohua Zhai*, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby*,†
*equal technical contribution, †equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulsby}@google.com

PaLI: A Jointly-Scaled Multilingual Language-Image Model

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, Radu Soricut
Google Research*



UL2

UL2: Unifying Language Learning Paradigms

Yi Tay[*] Mostafa Dehghani[*]

Vinh Q. Tran[♯] Xavier Garcia[♯] Jason Wei[♯] Xuezhi Wang[♯] Hyung Won Chung[♯]

Siamak Shakeri[♯] Dara Bahri[♭] Tal Schuster[♭] Huaixiu Steven Zheng[△]

Denny Zhou[△] Neil Houlsby[△] Donald Metzler[△]

Google Brain

ViT-22B

Scaling Vision Tra lion Parameters

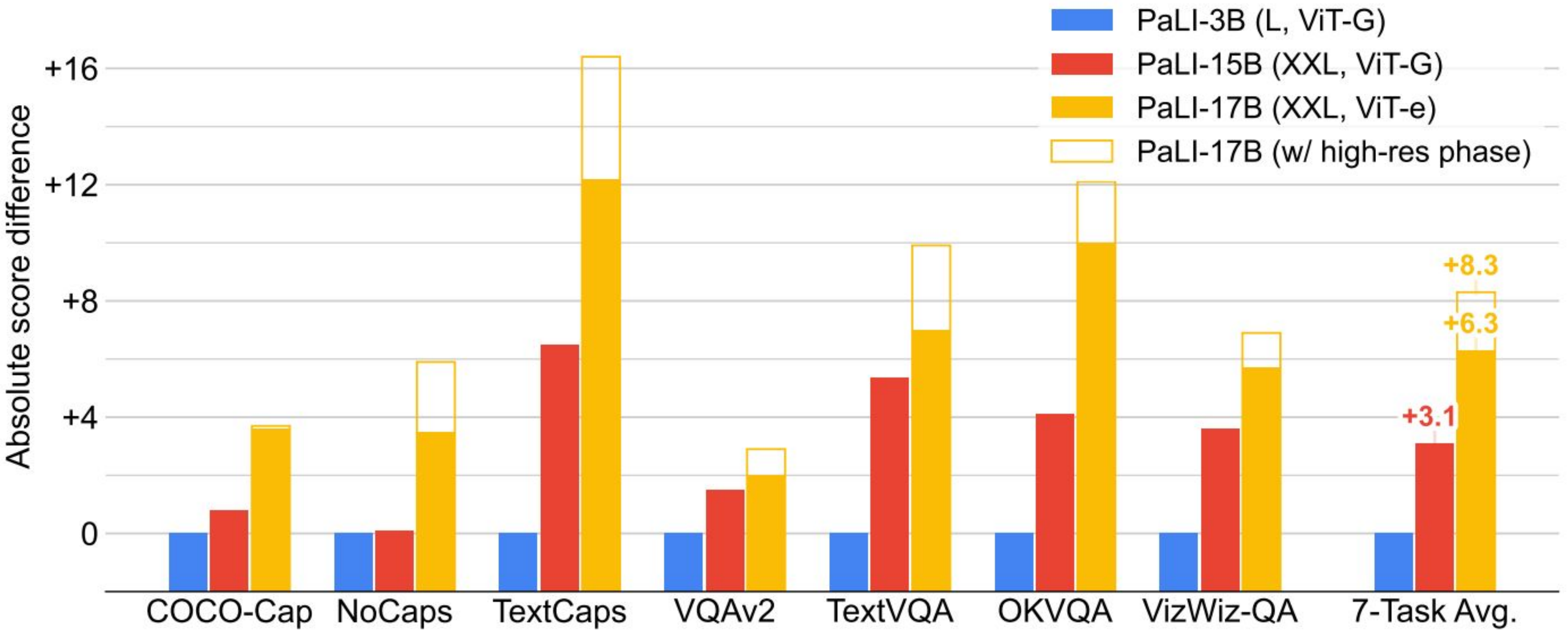
Mostafa Dehghani[*] Josip Djolonga[*] Basil Mustafa[*] Piotr Padlewski[*] Jonathan Heek[*]
Justin Gilmer Andreas Steiner Mathilde Caron Robert Geirhos Ibrahim Alabdulmohsin
Rodolphe Jenatton Lucas Beyer Michael Tschannen Anurag Arnab Xiao Wang
Carlos Riquelme Matthias Minderer Joan Puigcerver Utku Evci Manoj Kumar
Sjoerd van Steenkiste Gamaleldin F. Elsayed Aravindh Mahendran Fisher Yu
Avital Oliver Fantine Huot Jasmijn Bastings Mark Patrick Collier Alexey A. Gritsenko
Vighnesh Birodkar Cristina Vasconcelos Yi Tay Thomas Mensink Alexander Kolesnikov
Filip Pavetić Dustin Tran Thomas Kipf Mario Lučić Xiaohua Zhai Daniel Keysers
Jeremiah Harmsen Neil Houlsby[*]
Google Research

PaLI-X: On Scaling up a Multilingual Vision and Language Model

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, AJ Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, Radu Soricut

Google Research
pali-communications@google.com

Example: PaLI



Chen et al. 2023 https://arxiv.org/abs/2209.06794

Example: PaLI-X

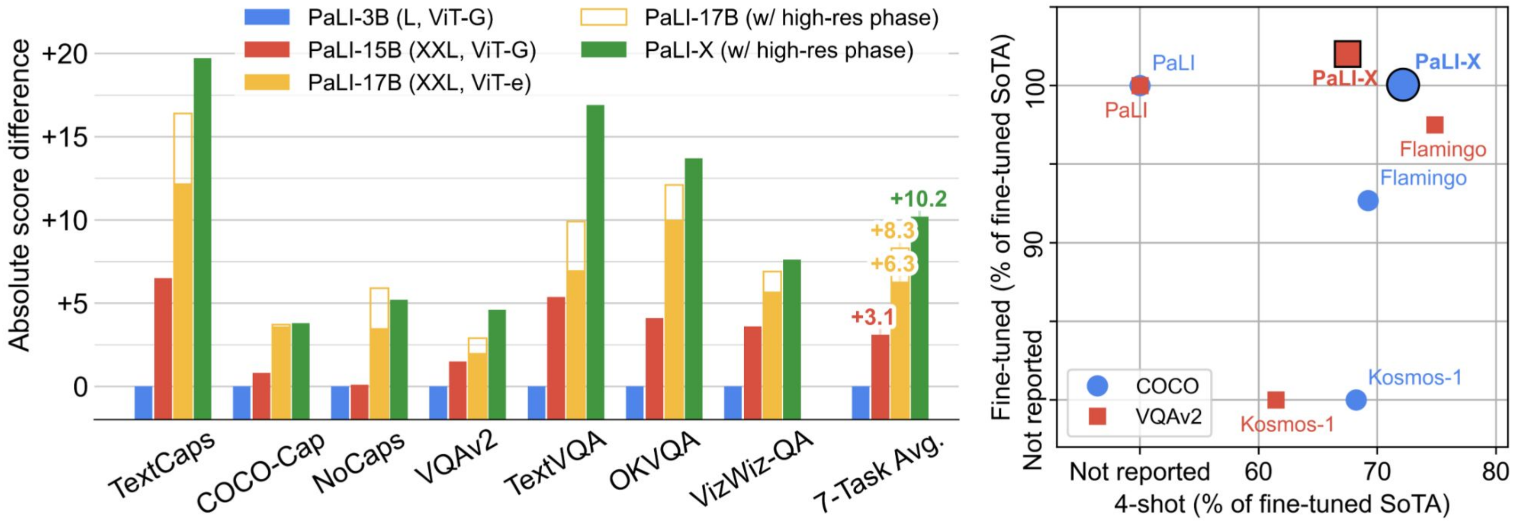


Figure 1: [Left] Comparing PaLI-X against PaLI on image-captioning and VQA benchmarks. [Right] The Pareto frontier between few-shot and fine-tuned performance, comparing PaLI-X with PaLI [5], Flamingo [10], and Kosmos-1 [11].

Chen et al. 2024 https://arxiv.org/abs/2305.18565

Pre-Training For Scene-Text Understanding

Main idea

Scene-text modeling by transferring from off-the-shelf OCR

SplitOCR

PreSTU: Pre-Training for Scene-Text Understanding

Jihyung Kil[1]* Soravit Changpinyo[2]
Xi Chen[2] Hexiang Hu[2] Sebastian Goodman[2] Wei-Lun Chao[1] Radu Soricut[2]
[1]The Ohio State University [2]Google Research

ICCV 2023

PaLI (ICLR 2023) → PaLI-X (CVPR 2024)

Training mixture To accommodate diverse tasks in the image-language space, we train PaLI using a mixture of eight pre-training tasks. This mixture is designed to span a range of general capabilities useful for downstream tasks. **Span corruption on text-only data** uses the same technique described by Xue et al. (2021)

OCR on WebLI OCR-text data

the pre-training obj

into two parts, ⟨cap

with the alt-text string in language ⟨lang⟩ as the target, based on the Conceptual Captions (Sharma et al., 2018) training data and machine translated alt-texts. **OCR on WebLI OCR-text data** uses the concatenation of the annotated OCR texts in language ⟨lang⟩ (Kil et al., 2022) produced by publicly available automatic service for the input image. **English and Cross-Lingual VQA** is VQ^{2}A-CC3M (Changpinyo et al., 2022a), translated in the same way as CC3M-35L.

answers in all instances here, as the English-native answers for VQA are oft

errors to perform out-of-context automatic translation. **English and Cross-**

generation (VQG) is also based on native and translated VQ^{2}A-CC3M-35L

we use only English answers here. **English-only Object-Aware (OA) VQA** is based on VQA triplets derived from automatically-produced, non-exhaustive object labels, inspired by Piergiovanni et al. (2022a). The QA pairs include listing all the objects in the image and whether a subset of objects are in the image. To create these examples, we require object-level annotations, for which we use Open Images (Kuznetsova et al., 2020). **Object detection** is a generative object-detection task inspired by Chen et al. (2021; 2022).

OCR → SplitOCR

3.2 Pretraining Data and Mixture

The main pretraining data for our model is based on WebLI [5], consisting of roughly one billion images with alt-texts from the web and OCR annotations (using the GCP Vision API), covering over 100 languages. In addition to WebLI ⟨image, text⟩ pairs, we introduce here *Episodic WebLI* data, where each episode corresponds to a set of such pairs. We aim to have each episode contain loosely related images (i.e., they are clustered according to their URL field), so as to encourage attention

nd 400M

(iv) split-ocr [24] on WebLI OCR annotations;

on text-only data (15% of tokens); (ii) split-captioning on WebLI alt-text data [21, 5]; (iii) captioning on CC3M [22] on native and translated alt-text data (over the same 35 languages covered by XM3600 [23]); (iv) split-ocr [24] on WebLI OCR annotations; (v) visual-question-answering objective over ⟨image, question, answer⟩ pairs generated using the VQ^{2}A method [25] over the CC3M

e and translated text (same 35 language pairs); (vi) visual-question-generation

me pairs as above; (vii) visual-question-answering objective over ⟨image,

s using the Object-Aware method [26] (English only); (viii) captioning on

ples (target alt-text predicted from the remaining alt-text and images); (ix)

ng on 4-pair examples (resembling Episodic WebLI and using VQ^{2}A-CC3M pairs), with the answer target conditioned on the other pairs of ⟨image, question, answer⟩ data. (x) pix2struct objective, introduced in [27], targeting page layout and structure using screenshot images paired with DOM-tree representations of html pages. (xi) Captioning on short video data, using the VTP data [10] (using four frames per video). (xii) object-detection objective on WebLI data, whereby an OWL-ViT model [28] (L/14) is used to annotate WebLI images, resulting in hundreds of pseudo object labels and bounding boxes per image. (xiii) image-token prediction objective, whereby we tokenize WebLI images (256×256 resolution) using a ViT-VQGAN [29] model with patch size 16×16 (256 tokens per image); this objective is framed as a 2D masked-token task (i.e., fill-in the missing grid pieces, with the corresponding image pixels also masked). Note that the image-token prediction objective is added mainly as a condition to check whether it adversarially impacts the performance on language-output tasks; our ablation experiments show that is does not.

Model	LLM	Visual Encoder	V2L Adapter	VInstr. Tuning
BLIP-2 (Li et al., 2023g)	FlanT5-XXL-11B★	EVA ViT-g	Q-Former	✗
FROMAGe (Koh et al., 2023b)	OPT-6.7B★	CLIP ViT-L	Linear	✗
Kosmos-1 (Huang et al., 2023b)	Magneto-1.3B◇	CLIP ViT-L	Q-Former*	✗
LLaMA-Adapter V2 (Gao et al., 2023)	LLaMA-7B▲	CLIP ViT-L	Linear	✗
OpenFlamingo (Awadalla et al., 2023)	MPT-7B★	CLIP ViT-L	XAttn LLM	✗
Flamingo (Alayrac et al., 2022)	Chinchilla-70B★	NFNet-F6	XAttn LLM	✗
PaLI (Chen et al., 2023j)	mT5-XXL-13B♦	ViT-e	XAttn LLM	✗
PaLI-X (Chen et al., 2023h)	UL2-32B♦	ViT-22B	XAttn LLM	✗

Caffagni et al. 2024 https://arxiv.org/abs/2402.12451

Adaptors

- Linear projection (Linear) & Multi-layer perceptrons (MLP)
- Querying Transformers (Q-Former)
- Cross Attention Mechanisms (XAttn)

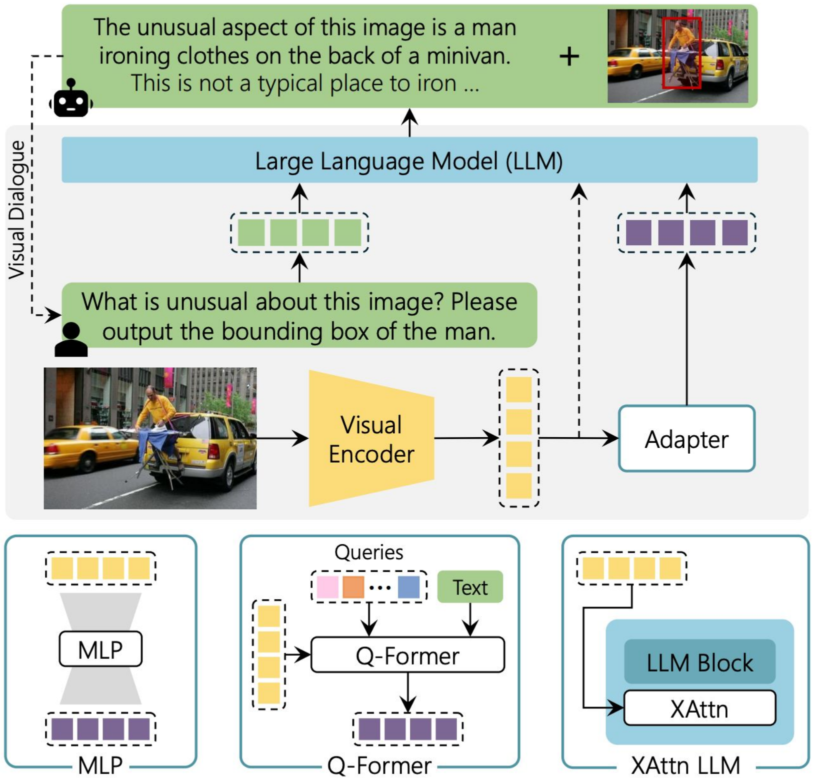
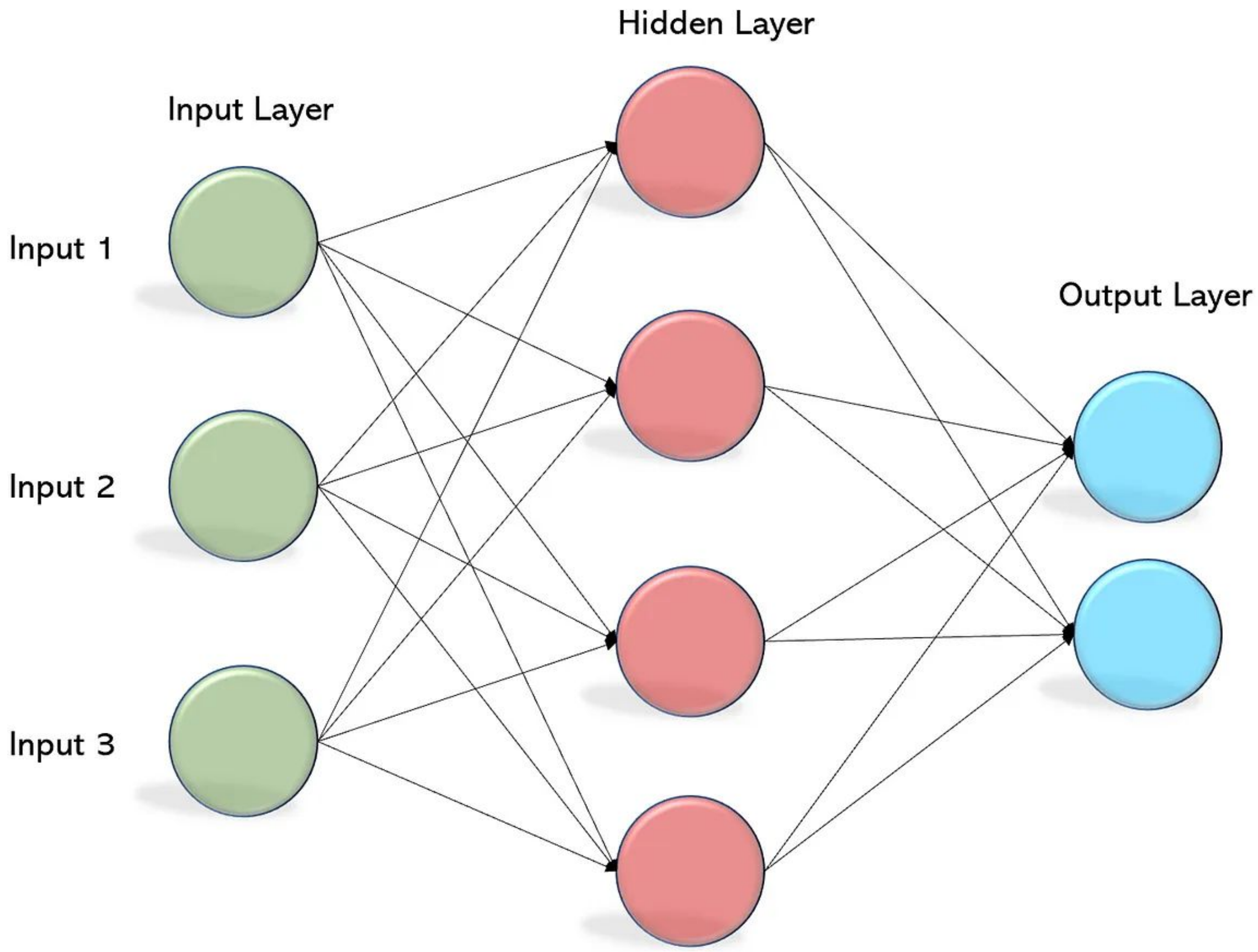
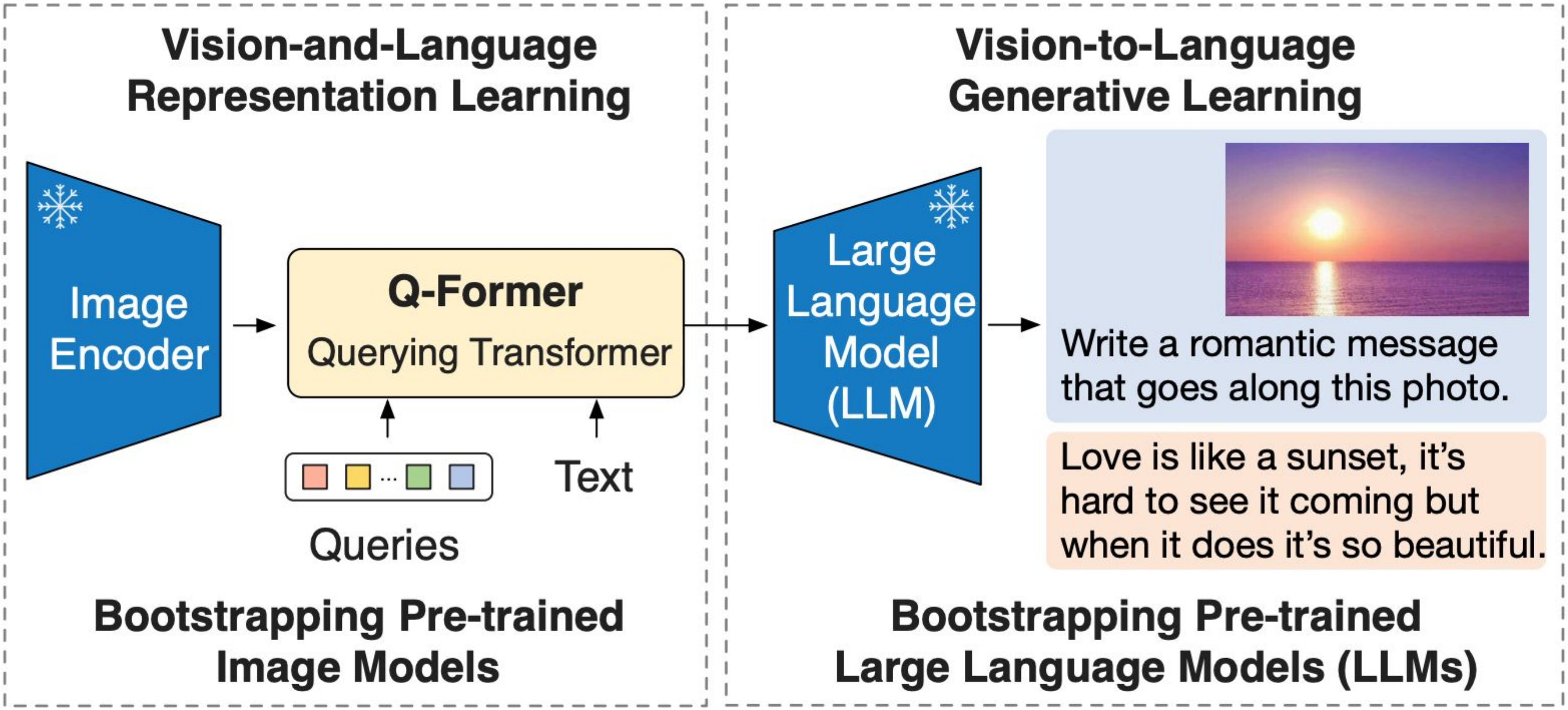


Figure 1: General architecture of Multimodal Large Language Models (MLLMs), composed of a visual encoder, a language model, and an adapter module that connects visual inputs to the textual space.


Hidden Layer
Input Layer
Input 1
Output Layer
Input 2
Input 3



BLIP-2: Li et al. 2023 https://arxiv.org/abs/2301.12597

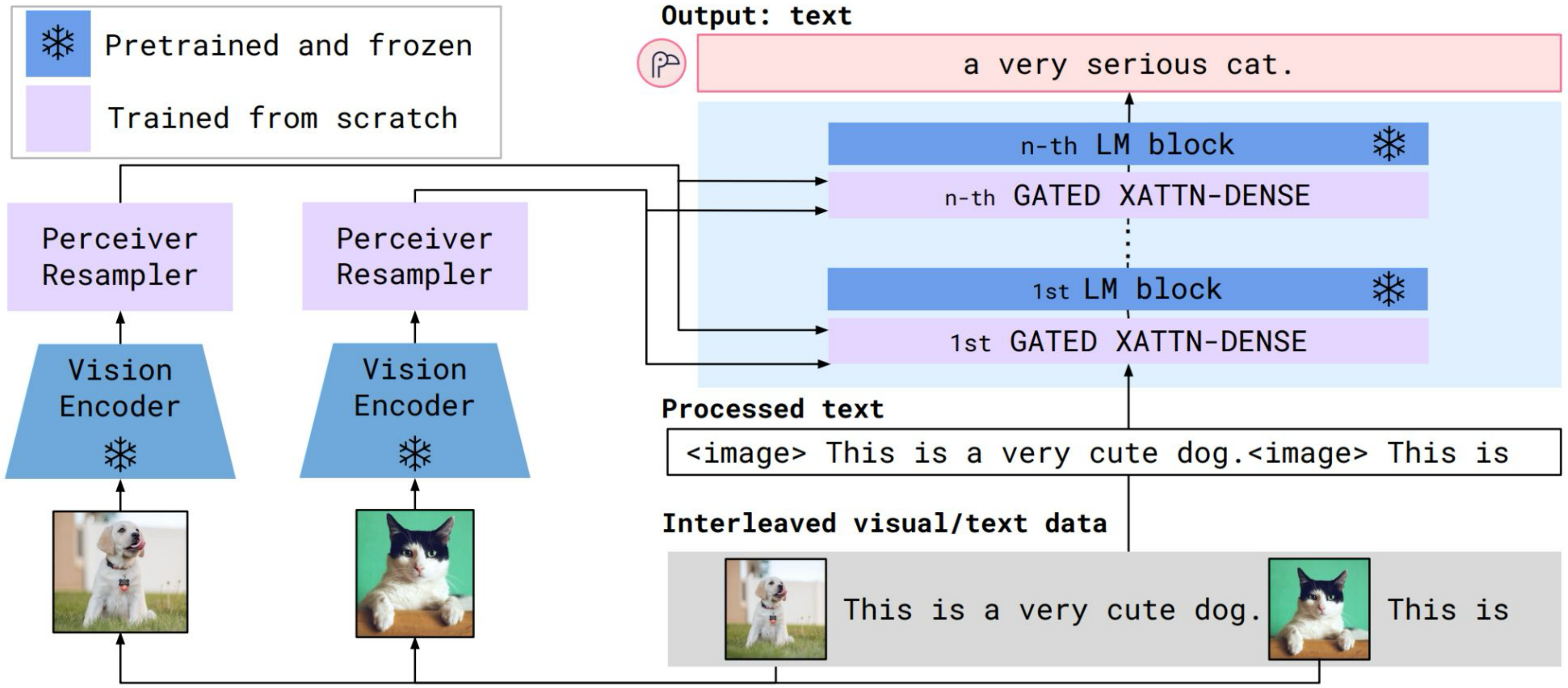
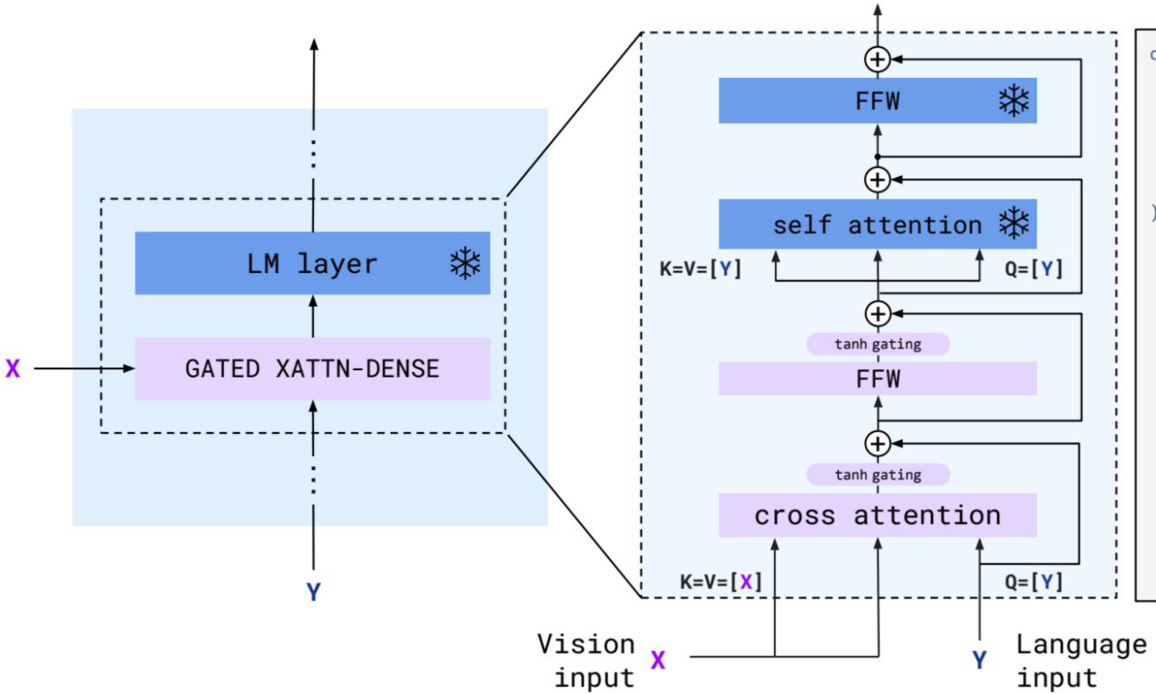


Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

Flamingo: Alayrac et al. 2023 https://arxiv.org/abs/2204.14198




```python
def gated_xattn_dense(
    y,  # input language features
    x,  # input visual features
    alpha_xattn, # xattn gating parameter – init at 0.
    alpha_dense, # ffw gating parameter – init at 0.
):
  """Applies a GATED XATTN-DENSE layer."""

  # 1. Gated Cross Attention
  y = y + tanh(alpha_xattn) * attention(q=y, kv=x)
  # 2. Gated Feed Forward (dense) Layer
  y = y + tanh(alpha_dense) * ffw(y)

  # Regular self-attention + FFW on language
  y = y + frozen_attention(q=y, kv=y)
  y = y + frozen_ffw(y)
  return y  # output visually informed language features
```

Figure 4: **GATED XATTN-DENSE layers.** To condition the LM on visual inputs, we insert new cross-attention layers between existing pretrained and frozen LM layers. The keys and values in these layers are obtained from the vision features while the queries are derived from the language inputs. They are followed by dense feed-forward layers. These layers are *gated* so that the LM is kept intact at initialization for improved stability

Flamingo: Alayrac et al. 2023 https://arxiv.org/abs/2204.14198

Data

- Pre-Training: Web (Later: Books, Video, Synthetic, etc.)
- Post-Training: Mostly human-annotated

Pre-Training

Large-Scale Image+(Alt)text data

Conceptual Captions
[Sharma et al. 18]

“trees in a winter snowstorm”

“a cartoon illustration of a bear waving and smiling”

“the scenic route through mountain range includes these unbelievably coloured mountains”

“facade of an old shop”

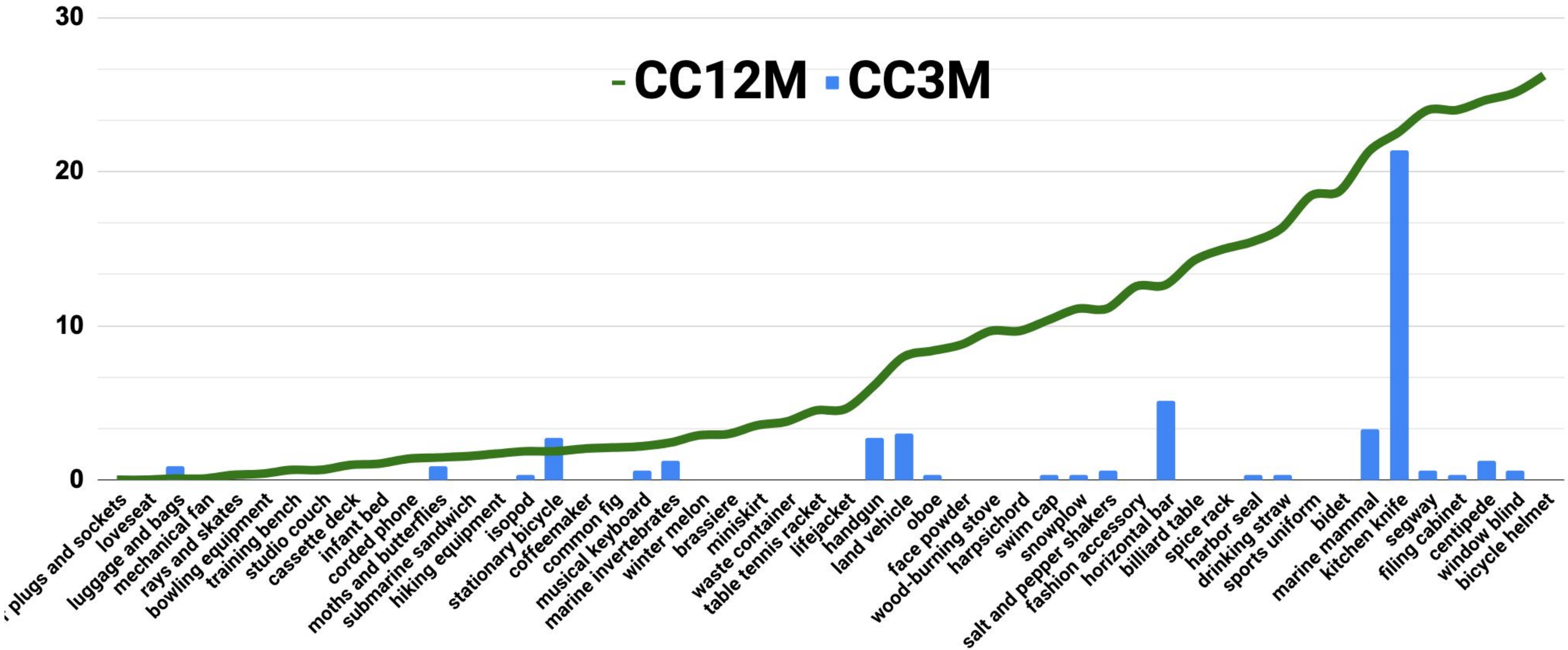
Conceptual 12M
[Changpinyo et al. 21]

<PERSON> was the first US president to attend a tournament in sumo's hallowed Ryogoku Kokugikan arena. (AFP photo)

Hand holding a fresh mangosteen

#jellyfish #blue #ocean #pretty Sea Turtle Wallpaper, Aquarius Aesthetic, Blue Aesthetic Pastel, The Adventure Zone, Capricorn And <PERSON>, Life Aquatic, Ocean Life, Jellyfish, Marine Life

Long-Tail Concepts (*Out-of-Domain* Classes from nocaps)

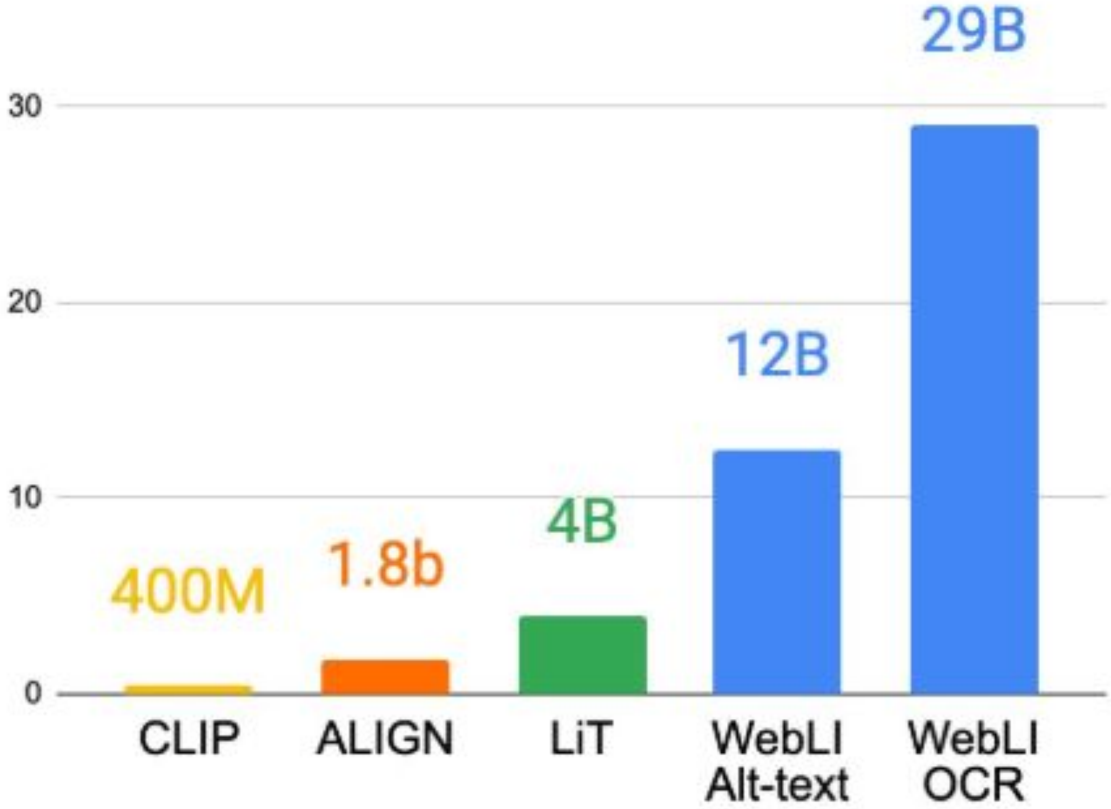


Changpinyo et al. 2021

Nocaps Example

No Pre-Training	A man in a military uniform holding a microphone.
CC3M Pre-Training	A man in a military uniform holds a sword.
CC12M Pre-Training	A man in a kilt playing the bagpipes.

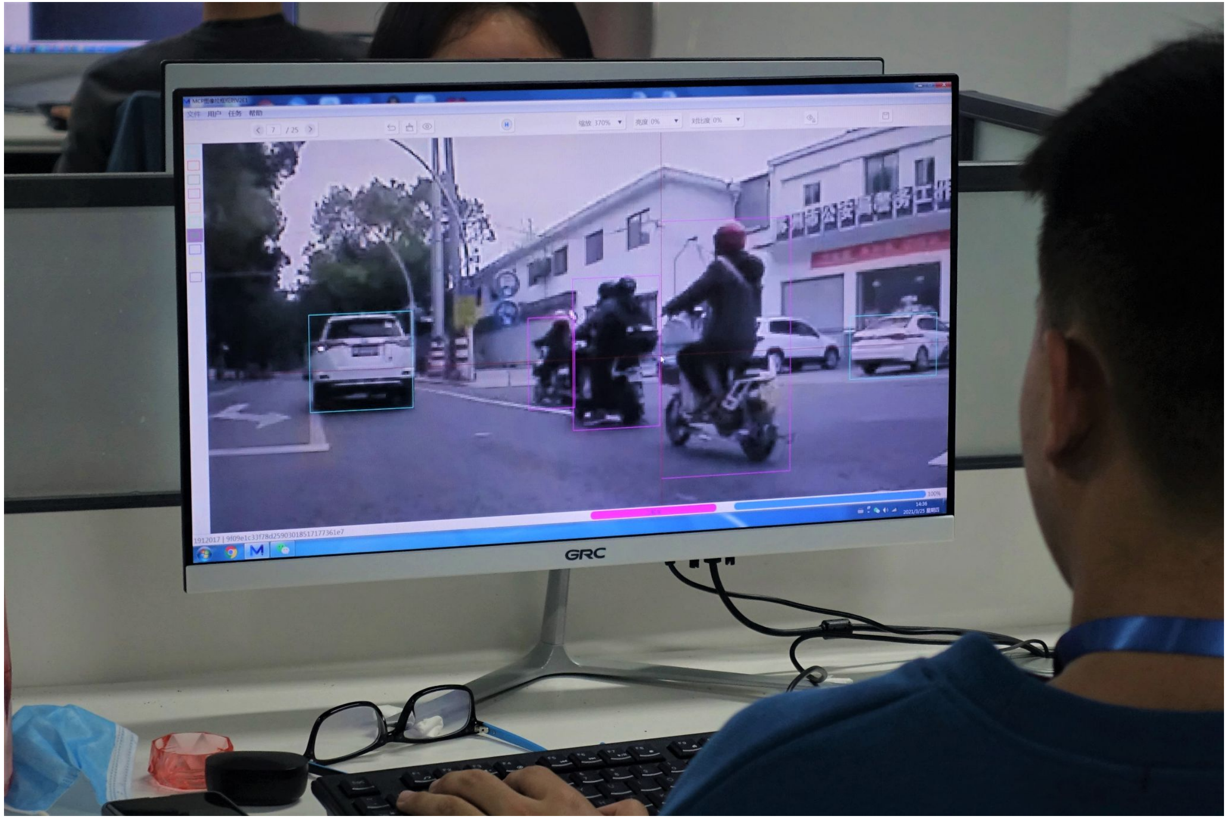
Changpinyo et al. 2021



Chen et al. 2023

Post-Training

Appen, Sama, iMerit, Scale AI, Labelbox, and Amazon SageMaker Ground Truth...



https://www.marketplace.org/2021/05/04/the-human-labor-behind-artificial-intelligence/

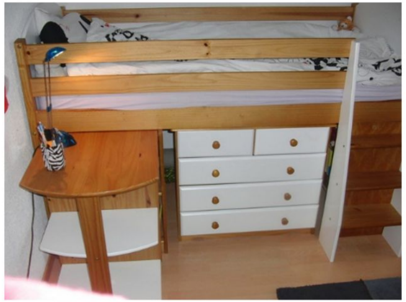
Evaluation

- **Early days**: Academic benchmarks focusing on captioning and question answering, with automatic metrics
- **These days**: Also real-world prompts, using humans and LLMs to evaluate long(er) responses. Other considerations: safety, fairness, localization, multi-turn, etc.

Image Captioning

MSCOCO Captions (2015)

A large bus sitting next to a very tall building.



Bunk bed with a narrow shelf sitting underneath it.

nocaps (2019)

1. A **man** sitting in the saddle on a **camel**.
2. A **person** is sitting on a **camel** with another **camel** behind him.
3. A **man** with long hair and blue jeans sitting on a **camel**.
4. **Man** sitting on a **camel** with a standing **camel** behind them.
5. Long haired **man** wearing sitting on blanket draped **camel**
6. A **camel** stands behind a sitting **camel** with a **man** on its back.
7. The standing **camel** is near a sitting one with a **man** on its back.
8. Someone is sitting on a **camel** and is in front of another **camel**.
9. Two **camels** in the dessert and a **man** sitting on the sitting one.
10. Two **camels** are featured in the sand with a **man** sitting on one of the seated **camels**.

Localized Narratives (2020)

In the front portion of the picture we can see a dried grass area with dried twigs. There is a woman standing wearing a light blue jeans and ash colour long sleeve length shirt. This woman is holding a black jacket in her hand. On the other hand she is holding a balloon which is peach in colour. on the top of the picture we can see a clear blue sky with clouds. The hair colour of the woman is brownish.

VQA

VizWiz

[redacted]

VQAv2 (2017)

Who is wearing glasses?

man | woman

Where is the child sitting?

fridge | arms

Is the umbrella upside down?

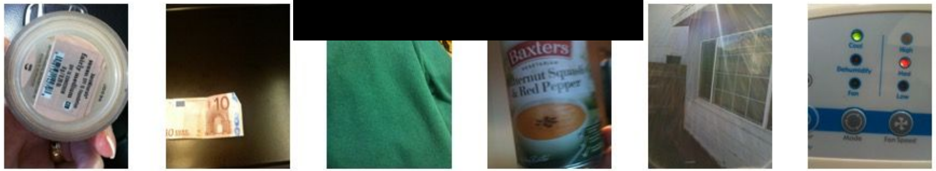
yes | no

How many children are in the bed?

2 | 1

GQA (2019)

A1. Is the tray on top of the table black or light brown? light brown
A2. Are the napkin and the cup the same color? yes
A3. Is the small table both oval and wooden? yes
A4. Is there any fruit to the left of the tray the cup is on top of? yes
A5. Are there any cups to the left of the tray on top of the table? no
B1. What is the brown animal sitting inside of? box
B2. What is the large container made of? cardboard
B3. What animal is in the box? bear
B4. Is there a bag to the right of the green door? no
B5. Is there a box inside the plastic bag? no



Q: Does this foundation have any sunscreen?
A: yes

Q: What is this?
A: 10 euros

Q: What color is this?
A: green

Q: Please can you tell me what this item is?
A: butternut squash red pepper soup

Q: Is it sunny outside?
A: yes

Q: Is this air conditioner on fan, dehumidifier, or air conditioning?
A: air conditioning

Q: What type of pills are these?
A: unsuitable image

Q: What type of soup is this?
A: unsuitable image

Q: Who is this mail for?
A: unanswerable

Q: When is the expiration date?
A: unanswerable

Q: What is this?
A: unanswerable

Q: Can you please tell me what the oven temperature is set to?
A: unanswerable

OK-VQA (2019)

Vehicles and Transportation | Brands, Companies and Products | Objects, Material and Clothing | Sports and Recreation | Cooking and Food

Q: What sort of vehicle uses this item?
A: firetruck

Q: When was the soft drink company shown first created?
A: 1898

Q: What is the material used to make the vessels in this picture?
A: copper

Q: What is the sports position of the man in the orange shirt?
A: goalie

Q: What is the name of the object used to eat this food?
A: chopsticks

Geography, History, Language and Culture | People and Everyday Life | Plants and Animals | Science and Technology | Weather and Climate

Q: What days might I most commonly go to this building?
A: Sunday

Q: Is this photo from the 50's or the 90's?
A: 50's

Q: What phylum does this animal belong to?
A: chordate, chordata

Q: How many chromosomes do these creatures have?
A: 23

Q: What is the warmest outdoor temperature at which this kind of weather can happen?
A: 32 degrees

MaXM

Q: Combien de nems sont servis dans une assiette rectangulaire blanche avec de la sauce soja?
A: 5, 5 nems, cinq, cinq nems
*(Q: How many spring rolls are served in a white rectangular plate with soy sauce?
A: 5, 5 spring rolls, five, five spring rolls)*

Q: Ce fel de cafea este în ceașca albă?
A: cu caimac, cafea cu lapte, cafea cu spumă, cafea cu spumă de lapte, cafea neagră cu spumă
*(Q: What kind of coffee is in the white cup?
A: with cream, coffee with milk, coffee with foam, coffee with milk foam, black coffee with foam)*

Q: काली कार किस सतह पर दिखाई दे रही है?
A: सड़क पर, काली सड़क पर, सड़क पे, काली सड़क पे
*(Q: On which surface is the black car visible?
A: on the road, on the black road, on the road, on the black road)*

Q: กระโปรงบนตุ๊กตาสีอะไร?
A: สีขาว, ขาว
*(Q: What color is the skirt on the doll?
A: white, white)*

Q: מה השם של העוף הזה?
A: קזואר
*(Q: What is the name of this bird?
A: Cassowary)*

Q: 除了可回收物，垃圾桶上还印有什么?
A: 其他垃圾
*(Q: Besides recyclables, what else is printed on the trash can?
A: other waste)*

Changpinyo et al. 2023

Model	Text Caps	VizWiz Cap	Text VQA	VizWiz VQA	ST VQA	OCR VQA	Info VQA	Doc VQA	AI2D	Chart QA	Screen2 Words	Widget Cap	OVEN	Info Seek
***with** OCR pipeline input*														
SoTA	160.4 [5]	124.7 [5]	73.67 [52]	73.3 [5]	79.9 [5]	67.5 [53]	47.4 [54]	84.7 [54]	38.5 [45]	45.5 [46]	- -	- -	- -	- -
PaLI-X	**163.7**	**125.7**	**80.78**	**74.6**	**84.5**	**77.3**	**54.8**	**86.8**	**81.4**	**72.3**	-	-	-	-
***without** OCR pipeline input*														
SoTA	145.0 [9]	120.8 [9]	67.27 [9]	70.7 [5]	75.8 [9]	71.3 [27]	40.0 [27]	76.6 [27]	42.1 [27]	70.5 [8]	109.4 [27]	141.8 [20]	20.0 [47]	17.7 [48]
PaLI-X	**147.0**	**122.7**	**71.44**	**70.9**	**79.9**	**75.0**	**49.2**	**80.0**	**81.2**	**70.9**	**127.9**	**153.0**	**23.1**	**21.8**

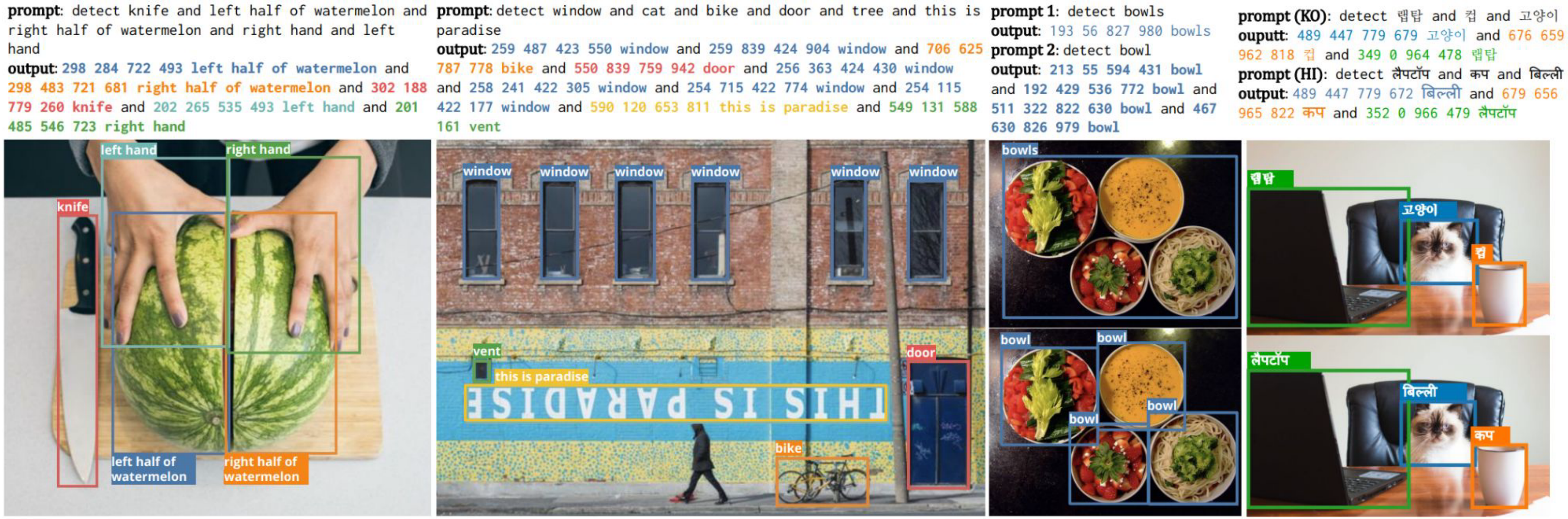
Table 2: Results on benchmarks more focused on text understanding capabilities. For OVEN [47] & InfoSeek [48], we follow the proposed 224×224 resolution settings for fair comparison.

Chen et al. 2024 https://arxiv.org/abs/2305.18565

Method	MSR-VTT Cap. [55]	MSR-VTT QA [60]	Activity-Net Cap. [57]	Activity-Net QA [61]	VATEX Cap. [56]	SMIT Cap. [58]	NExT-QA QA [59]
Prior SOTA	75.9 GIT2 [9]	**47.4** Flamingo [10]	52.5 PDVC [62]	44.7 VINDLU [63]	94.0† GIT2 [9]	28.1‡ MV-GPT [64]	33.5§ Flamingo 32shot [10]
PaLI-X (8fr)	74.6	46.9	49.0	48.4	66.0	42.5	37.0
PaLI-X (16fr)	**76.8**	**47.1**	**54.9**	**49.4**	69.3	**43.5**	**38.3**

Table 5: Results for Video Captioning and Video-QA using 8 frames (8fr) or 16 frames (16fr). †GIT2 uses Self-Critical Sequence Training to directly optimize the CIDEr metric for VATEX. ‡SMIT has not been used for video captioning before, we apply MV-GPT [64] and report results on the test set. §Numbers were obtained using 32-shot; since Flamingo 32-shot outperforms fine-tuning SOTA on this open-ended QA task, they did not conduct further fine-tuning experiments for this task.

Chen et al. 2024 https://arxiv.org/abs/2305.18565



Credits: Watermelon/Cat; Sarah Pflug (burst), Bowls; ariesandrea (flickr), Wall; Matthew Henry (burst)

Figure 2: Examples demonstrating multilingual, OCR and other capabilities transferred to detection.

Chen et al. 2024 https://arxiv.org/abs/2305.18565

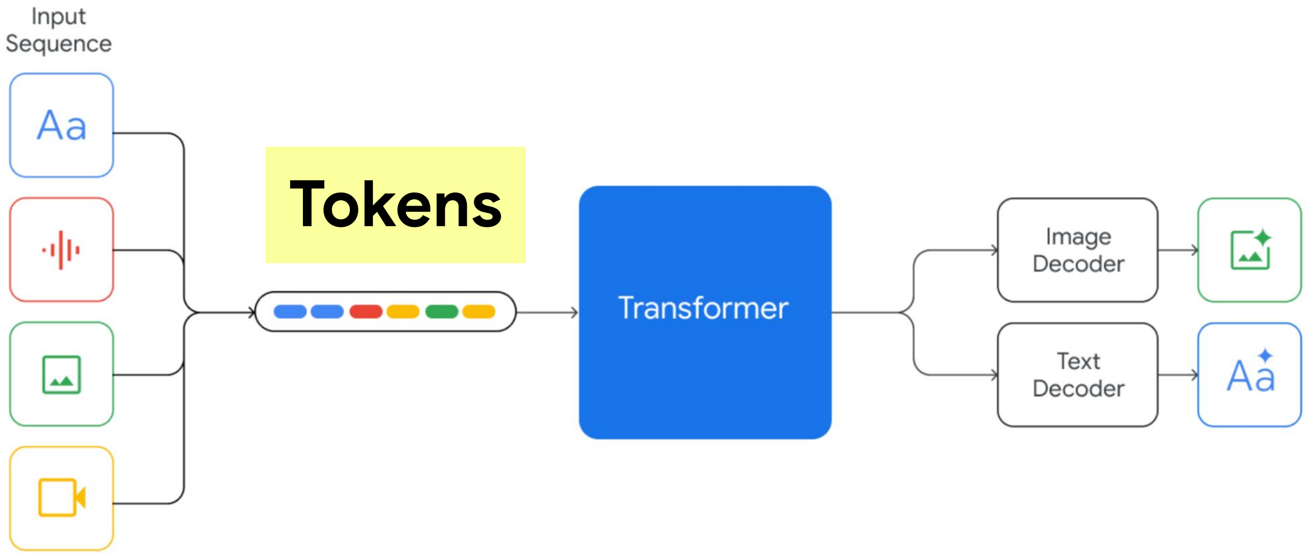
Summary

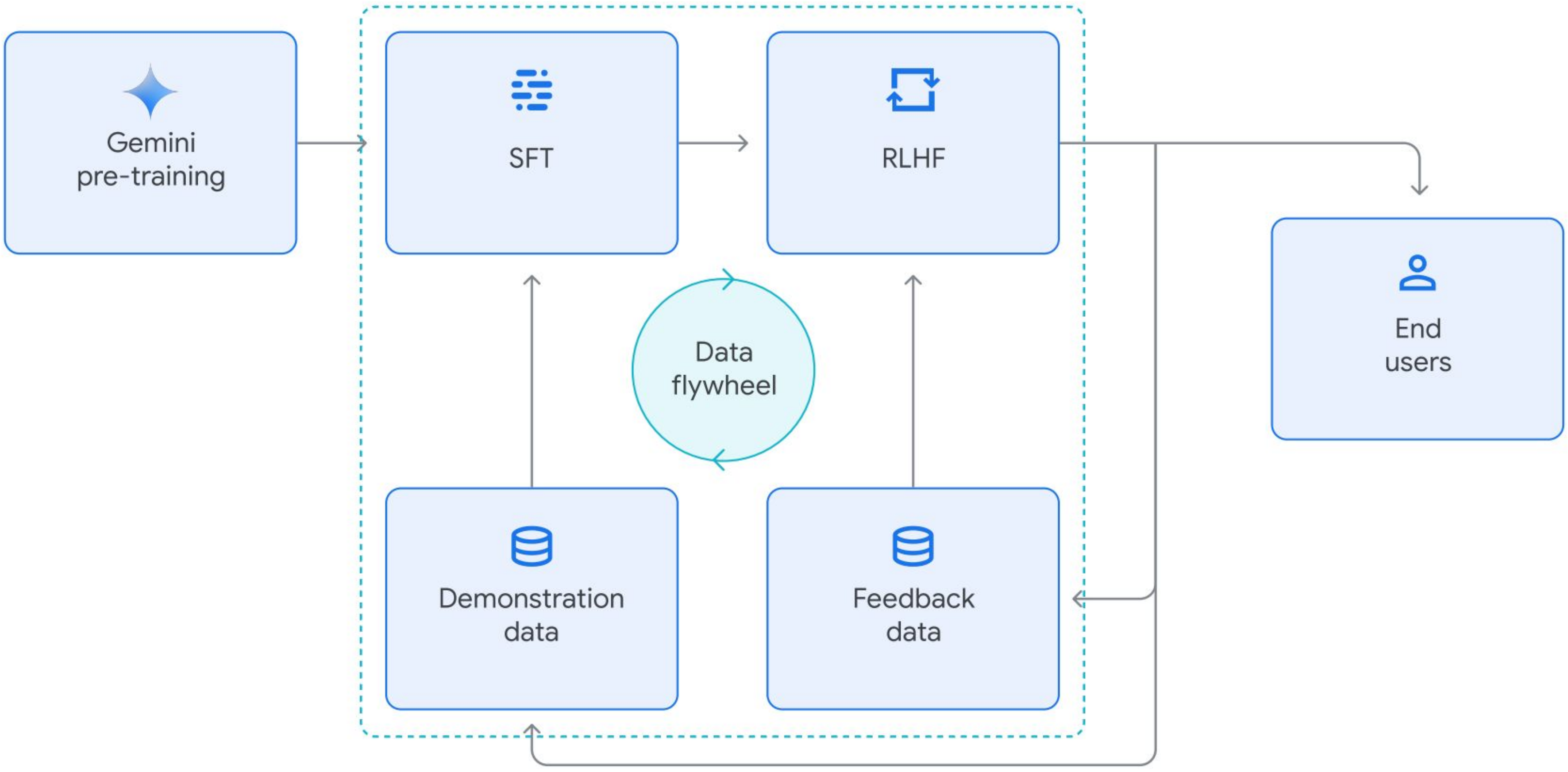
- **Vision**+**Language** took whatever worked best **at the moment**
- **Extensive engineering** goes into modeling (bridging the gaps between modalities), data, and evaluation.
- **Data scale and diversity** are important
- **Metrics for “hill-climbing”** are important

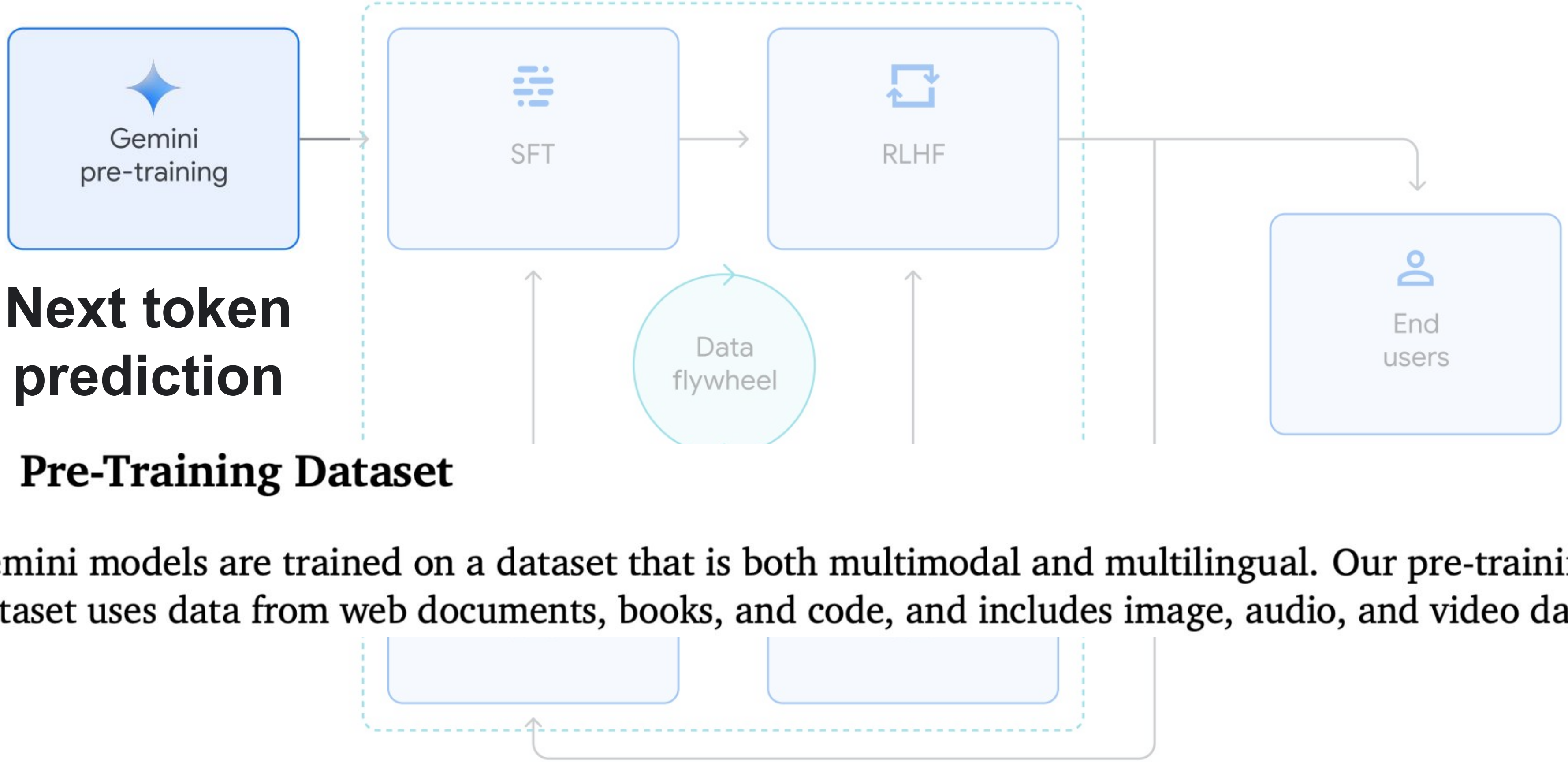
Gemini

Gemini: A Family of Highly Capable Multimodal Models

Gemini Team, Google[1]



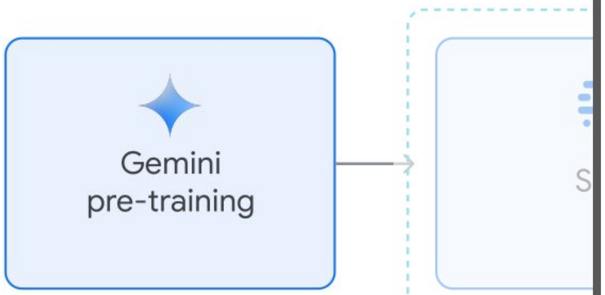

Gemini
pre-training
SFT
RLHF
End
users
Data
flywheel
Demonstration
data
Feedback
data

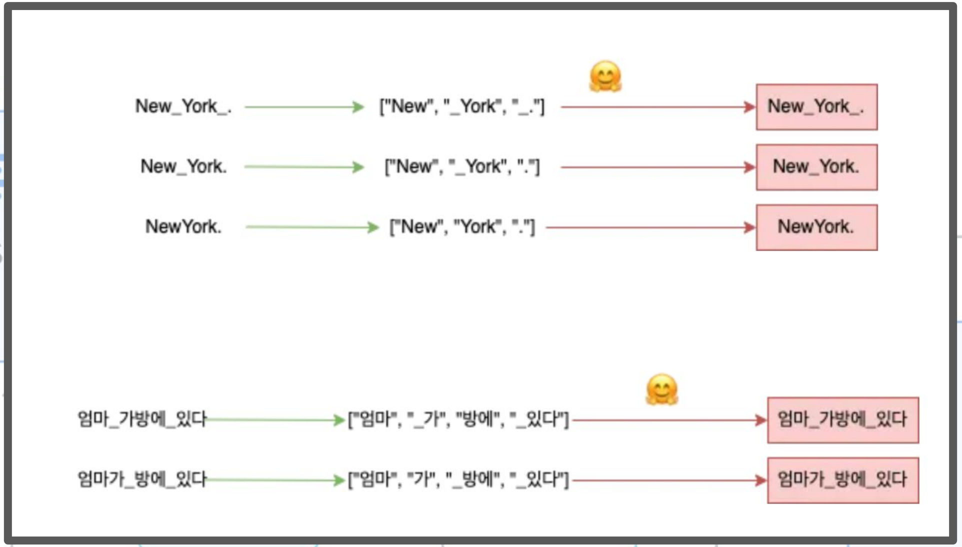


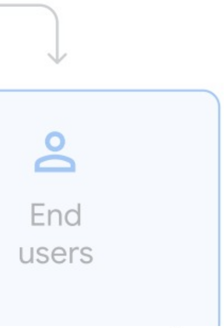
Next token prediction

4. Pre-Training Dataset

Gemini models are trained on a dataset that is both multimodal and multilingual. Our pre-training dataset uses data from web documents, books, and code, and includes image, audio, and video data.

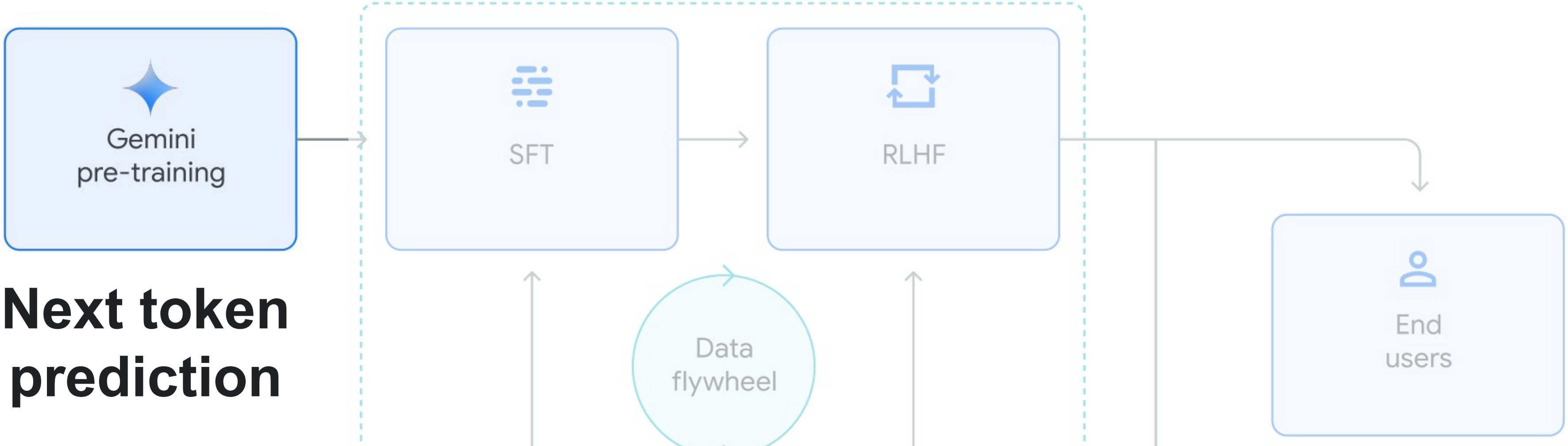






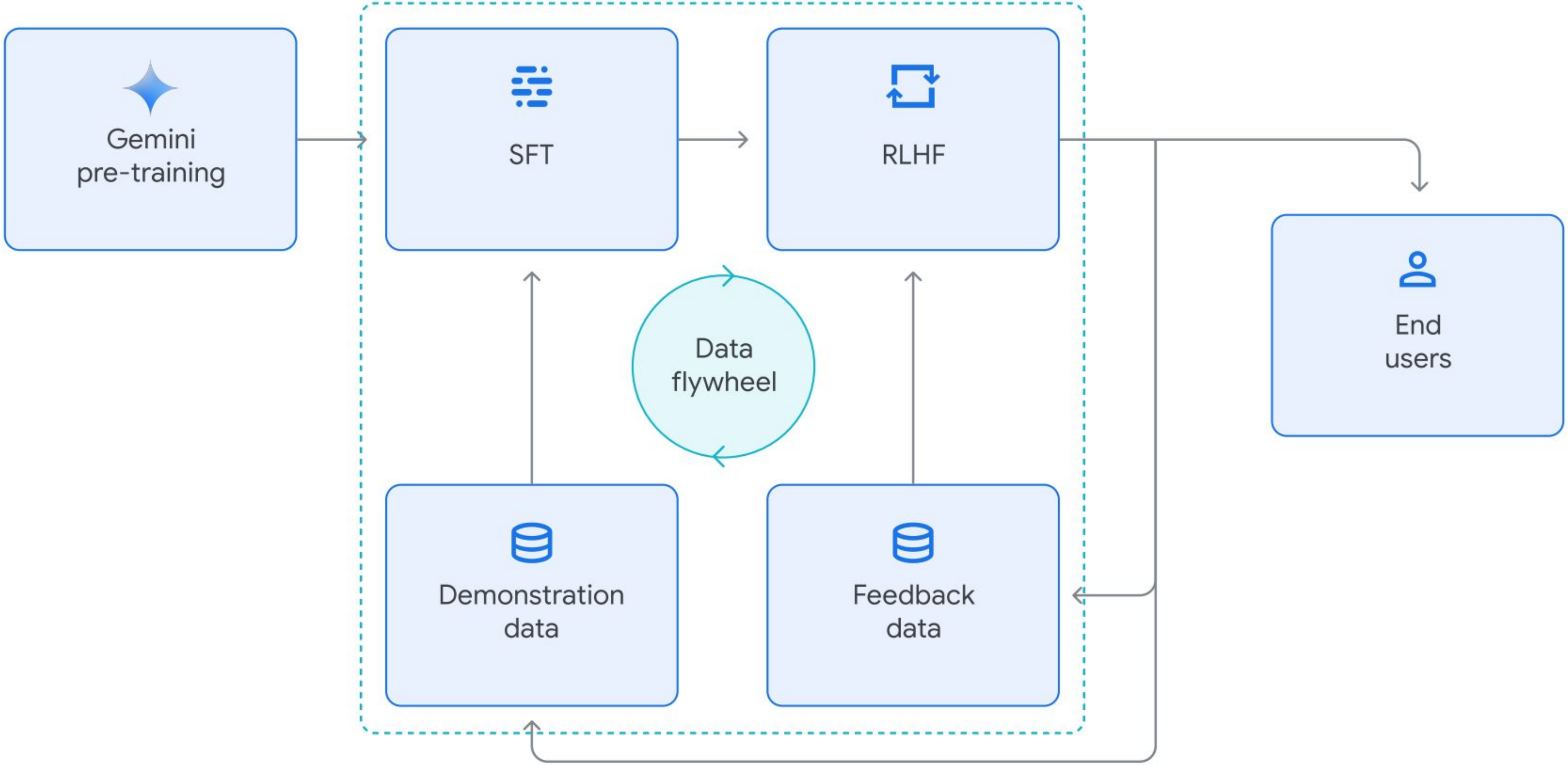
Next token prediction

We use the SentencePiece tokenizer (Kudo and Richardson, 2018) and find that training the tokenizer on a large sample of the entire training corpus improves the inferred vocabulary and subsequently improves model performance. For example, we find Gemini models can efficiently tokenize non-Latin scripts which can, in turn, benefit model quality as well as training and inference speed.

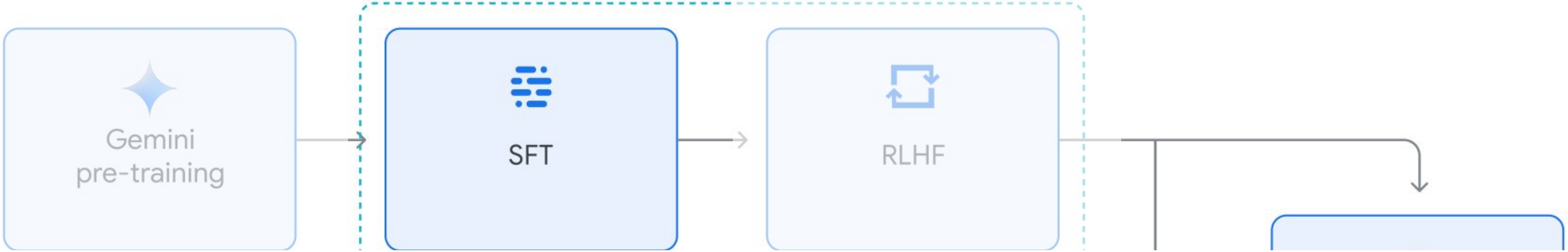


Next token prediction

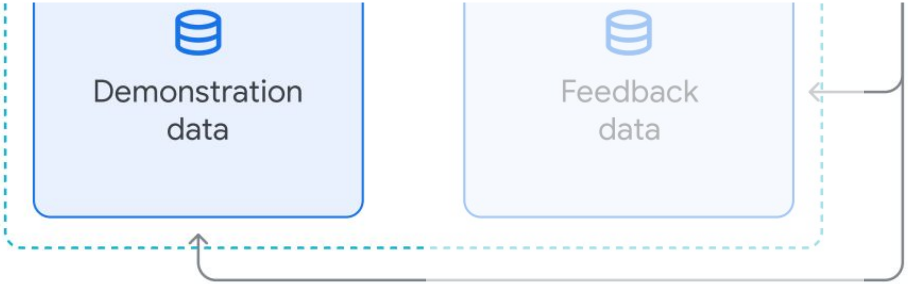
The number of tokens used to train the largest models were determined following the approach in Hoffmann et al. (2022). The smaller models are trained for significantly more tokens to improve performance for a given inference budget, similar to the approach advocated in Touvron et al. (2023a).



Gemini pre-training
SFT
RLHF
End users
Data flywheel
Demonstration data
Feedback data



(2) SFT on Demonstration Data: SFT trains the model to output a desired target response given a prompt. Our Demonstration Data target responses can be directly written by a human expert, or generated by a model and in some cases revised or reviewed by a human. Additionally, we use data analysis tools and heuristics to ensure high data diversity across capabilities, use cases, and semantic clusters.



Instruction Tuning

Input: text

Output: text

Is the following sentence acceptable?
"The course is jumping well."

On the scale of 1 to 5, how similar are the following two sentences?

1. The rhino grazed on the grass.
2. A rhino is grazing in a field.

Instruction finetuning*

"It is not acceptable"

"3.8"

Wei et al., 2021
Sanh et al. 2021
Slide Credit: Hyung Won Chung

Finetuning tasks

T0-SF

Commonsense reasoning
Question generation
Closed-book QA
Adversarial QA
Extractive QA
Title/context generation
Topic classification
Struct-to-text
...

55 Datasets, 14 Categories, 193 Tasks

Muffin

Natural language inference
Code instruction gen.
Program synthesis
Dialog context generation
Closed-book QA
Conversational QA
Code repair
...

69 Datasets, 27 Categories, 80 Tasks

CoT (Reasoning)

Arithmetic reasoning
Commonsense Reasoning
Implicit reasoning
Explanation generation
Sentence composition
...

9 Datasets, 1 Category, 9 Tasks

Natural Instructions v2

Cause effect classification
Commonsense reasoning
Named entity recognition
Toxic language detection
Question answering
Question generation
Program execution
Text categorization
...

372 Datasets, 108 Categories, 1554 Tasks

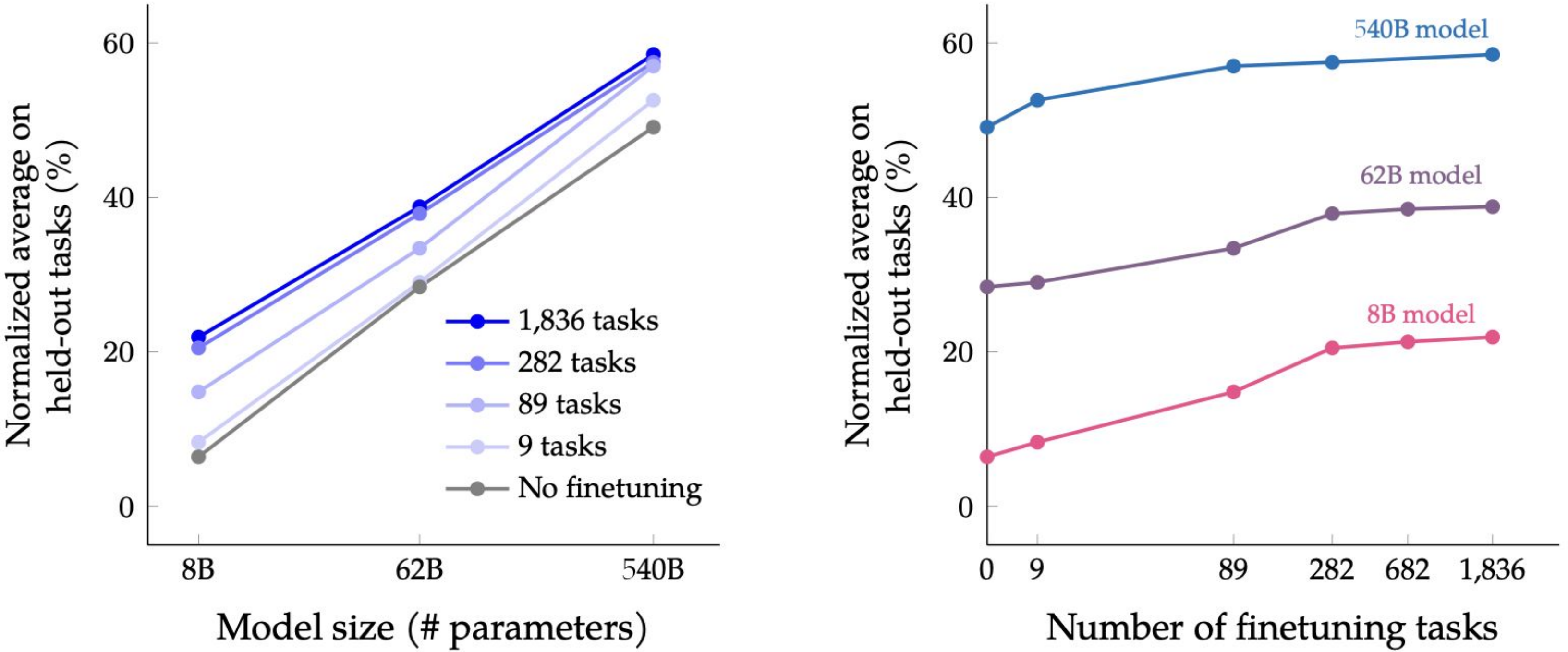
- A **Dataset** is an original data source (e.g. SQuAD).
- A **Task Category** is unique task setup (e.g. the SQuAD dataset is configurable for multiple task categories such as extractive question answering, query generation, and context generation).
- A **Task** is a unique <dataset, task category> pair, with any number of templates which preserve the task category (e.g. query generation on the SQuAD dataset.)

Instruction finetuning on 1836 (!!) academic tasks

Chung et al., 2022

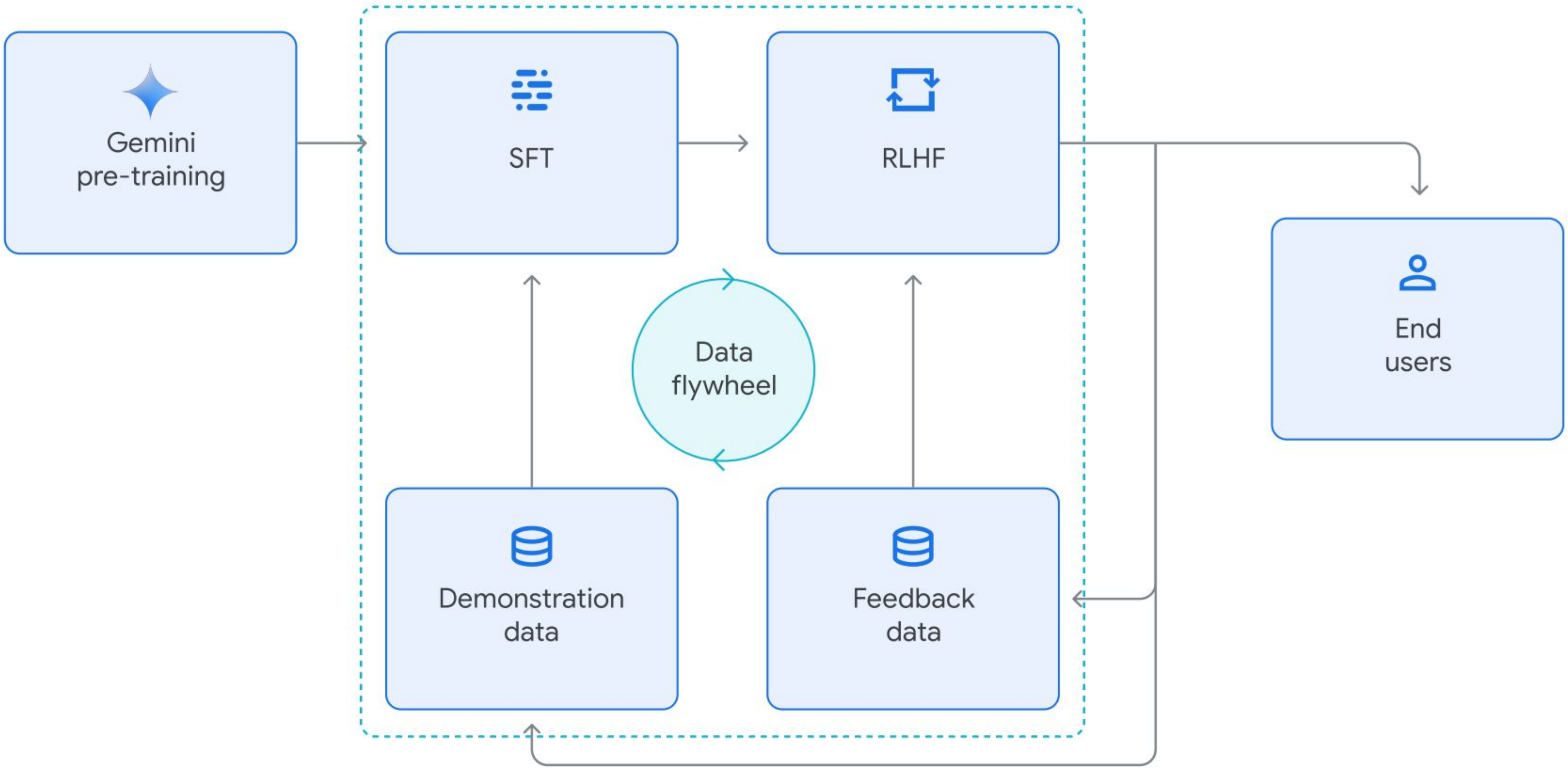
Slide Credit: Hyung Won Chung

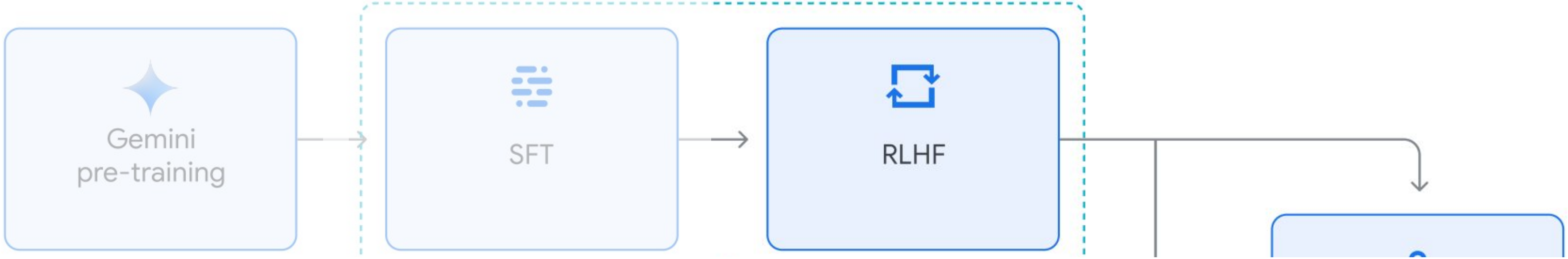
Scaling the number of tasks and model size improves the



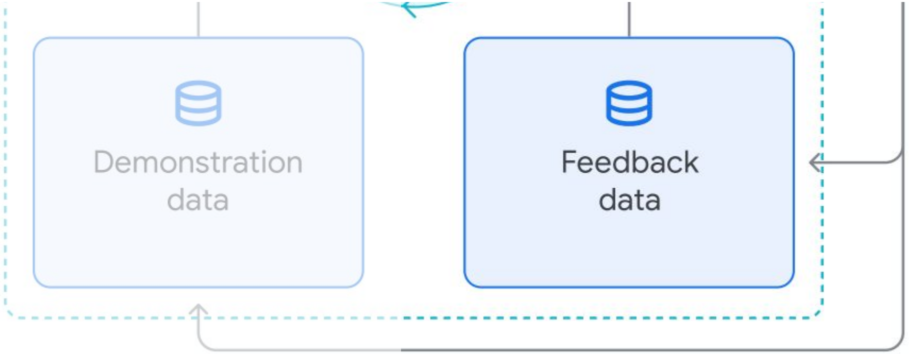
Chung et al., 2022

Slide Credit: [Hyung Won Chung]()


Gemini
pre-training
SFT
RLHF
End
users
Data
flywheel
Demonstration
data
Feedback
data



(4) RLHF: Applying reinforcement learning from human feedback (RLHF) to our models provides further gains over SFT alone. Our approach creates an iterative process in which RL continually pushes the boundaries of the RM, while the RM is continuously improved through evaluation and data collection, leading to progressive improvements in both.



Reward Model (RM)
(Relative Preference)

Input

Explain the moon landing to a 6 year old in a few sentences

Completion 1

The Moon is a natural satellite of the Earth. It is the fifth largest moon in the Solar System and the largest relative to the size of its host planet.

<

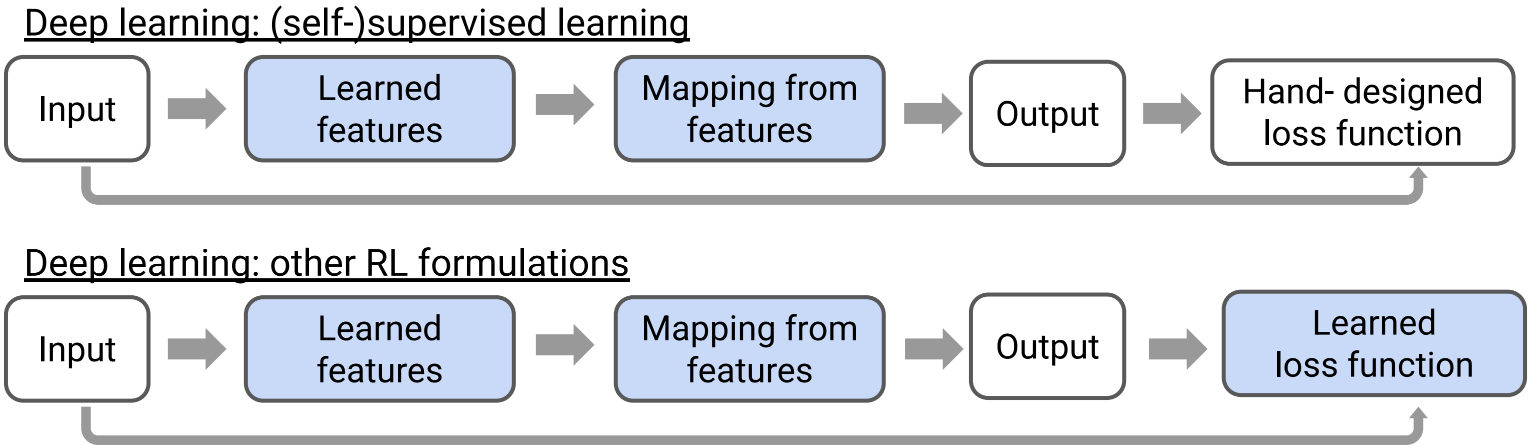
Completion 2

People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.

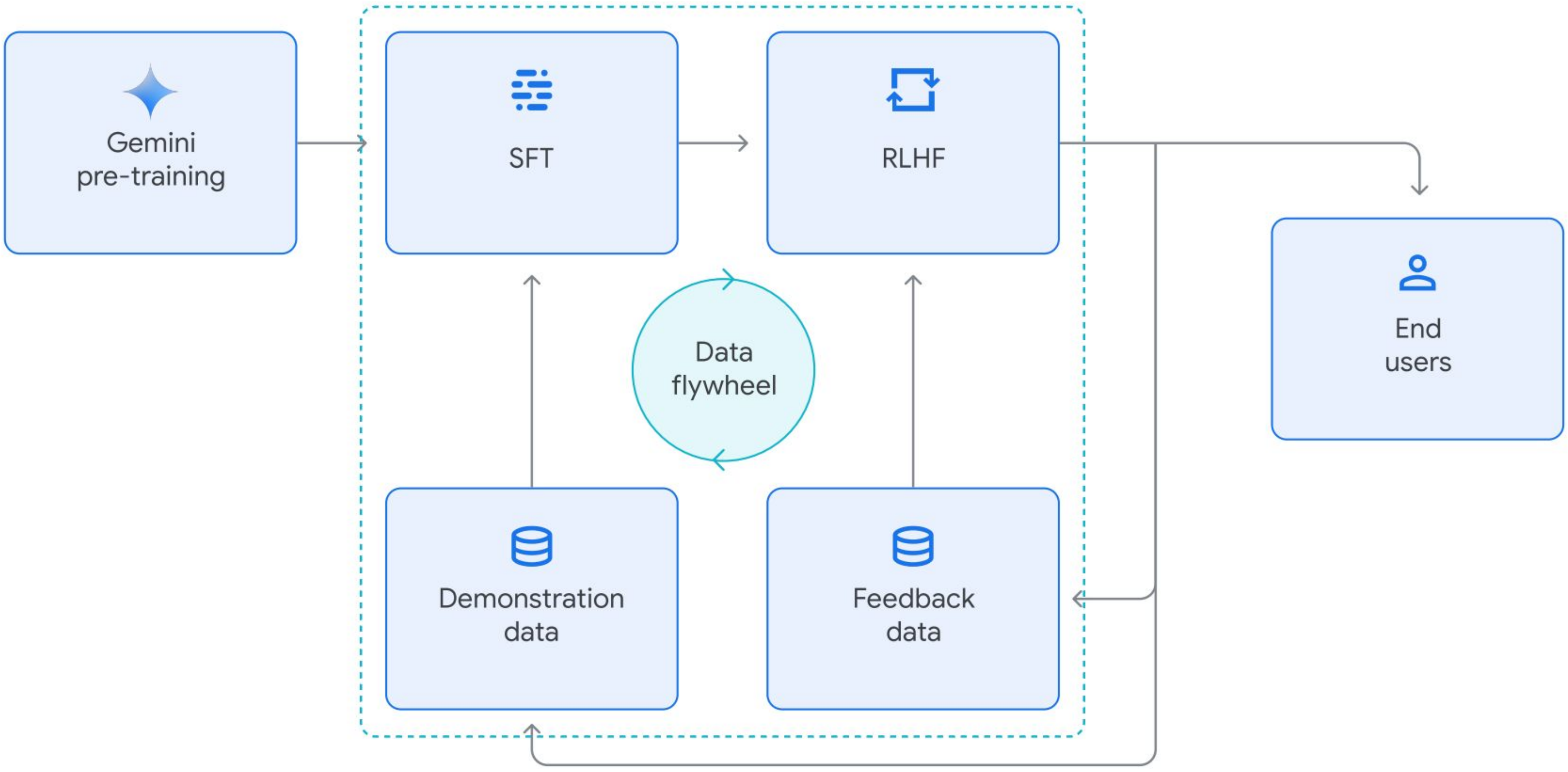
Slide Credit: Hyung Won Chung

Why RLHF?

Learnable part of the system



Slide Credit: [Hyung Won Chung]



Gemini
pre-training
SFT
RLHF
End
users
Data
flywheel
Demonstration
data
Feedback
data

Gemini Team Organization

Gemini Leads
Rohan Anil, *Co-Lead, Text*
Sebastian Borgeaud, *Co-Lead, Text*
Jean-Baptiste Alayrac, *Co-Lead, MM Vision*
Jiahui Yu, *Co-Lead, MM Vision*
Radu Soricut, *Co-Lead, MM Vision*
Johan Schalkwyk, *Lead, MM Audio*
Andrew M. Dai, *Co-Lead, Data*
Anja Hauth, *Co-Lead, Data*
Katie Millican, *Co-Lead, Data*
David Silver, *Co-Lead, Fine-Tuning*
Melvin Johnson, *Lead, Instruction Tuning*
Ioannis Antonoglou, *Co-Lead, RL Techniques*
Julian Schrittwieser, *Co-Lead, RL Techniques*
Amelia Glaese, *Lead, Human Data*
Jilin Chen, *Lead, Safety*
Emily Pitler, *Co-Lead, Tool Use*
Timothy Lillicrap, *Co-Lead, Tool Use*
Angeliki Lazaridou, *Co-Lead, Eval*
Orhan Firat, *Co-Lead, Eval*
James Molloy, *Co-Lead, Infra*
Michael Isard, *Co-Lead, Infra*
Paul R. Barham, *Co-Lead, Infra*
Tom Hennigan, *Co-Lead, Infra*
Benjamin Lee, *Co-Lead, Codebase & Parallelism*
Fabio Viola, *Co-Lead, Codebase & Parallelism*
Malcolm Reynolds, *Co-Lead, Codebase & Parallelism*
Yuanzhong Xu, *Co-Lead, Codebase & Parallelism*
Ryan Doherty, *Lead, Ecosystem*
Eli Collins, *Lead, Product*
Clemens Meyer, *Co-Lead, Operations*
Eliza Rutherford, *Co-Lead, Operations*
Erica Moreira, *Co-Lead, Operations*
Kareem Ayoub, *Co-Lead, Operations*
Megha Goel, *Co-Lead, Operations*

- Text
- **Multimodal Vision**
- **Multimodal Audio**
- Data
- Fine-Tuning
- Instruction-Tuning
- RL Techniques
- Human Data
- Safety
- Tool Use
- Eval
- Infra
- Codebase & Parallelism
- Ecosystem
- Product
- Operations

Introducing Gemini

By Demis Hassabis, CEO and Co-Founder of Google DeepMind, on behalf of the Gemini team

AI has been the focus of my life's work, as for many of my research colleagues. Ever since programming AI for computer games as a teenager, and throughout my years as a neuroscience researcher trying to understand the workings of the brain, I've always believed that if we could build smarter machines, we could harness them to benefit humanity in incredible ways.

This promise of a world responsibly empowered by AI continues to drive our work at Google DeepMind. For a long time, we've wanted to build a new generation of AI models, inspired by the way people understand and interact with the world. AI that feels less like a smart piece of software and more like something useful and intuitive — an expert helper or assistant.

Today, we're a step closer to this vision as we introduce Gemini, the most capable and general model we've ever built.

Gemini is the result of large-scale collaborative efforts by teams across Google, including our colleagues at Google Research. It was built from the ground up to be multimodal, which means it can generalize and seamlessly understand, operate across and combine different types of information including text, code, audio, image and video.

https://blog.google/technology/ai/google-gemini-ai/#introducing-gemini

Introducing Gemini

By Demis Hassabis, CEO and Co-Founder of Google DeepMind, on behalf of the Gemini team

AI has been the focus of my life's work, as for many of my research colleagues. Ever since programming AI for computer games as a teenager, and throughout my years as a neuroscience researcher trying to understand the workings of the brain, I've always believed that if we could build smarter machines, we could harness them to benefit humanity in incredible ways.

This promise of a world responsibly empowered by AI continues to drive our work at Google DeepMind. For a long time, we've wanted to build a new generation of AI models, inspired by the way people understand and interact with the world. AI that feels less like a smart piece of software and more like something useful and intuitive — an expert helper or assistant.

Today, we're a step closer to this vision as we introduce Gemini, the most capable and general model

Gemini is the result of large-scale collaborative efforts by teams across Google, including our colleagues at Google Research. It was built from the ground up to be multimodal, which means it can generalize and seamlessly understand, operate across and combine different types of information including text, code, audio, image and video.

https://blog.google/technology/ai/google-gemini-ai/#introducing-gemini

Multimodal

- *Data at all stages*

Product Solutions Resources Open Source Enterprise Pricing Sign in Sign up

BradyFU / **Awesome-Multimodal-Large-Language-Models** Public

Notifications Fork 779 Star 12.2k

- Awesome Datasets
  - Datasets of Pre-Training for Alignment
  - Datasets of Multimodal Instruction Tuning
  - Datasets of In-Context Learning
  - Datasets of Multimodal Chain-of-Thought
  - Datasets of Multimodal RLHF
  - Benchmarks for Evaluation
  - Others

https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models

Multimodal

- *Data at all stages*
  Example: Llama 3

7.1.1 Image Data

Our image encoder and adapter are trained on image-text pairs. We construct this dataset via a complex data processing pipeline that consists of four main stages: **(1)** quality filtering, **(2)** perceptual de-duplication, **(3)** resampling, and **(4)** optical character recognition. We also apply a series of safety mitigations.

Transcribing documents. To improve the performance of our models on document understanding tasks, we render pages from documents as images and paired the images with their respective text. The document text is obtained either directly from the source or via a document parsing pipeline.

Grattafiori et al. 2024 https://arxiv.org/abs/2407.21783

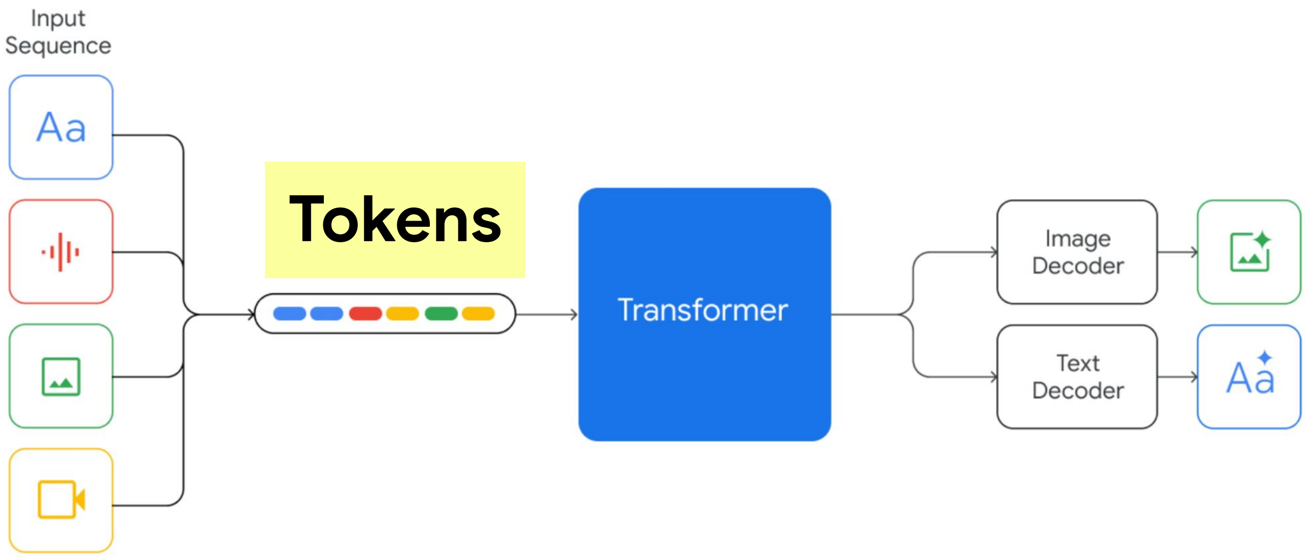
Annealing data. We create an annealing dataset by resampling the image-caption pairs to a smaller volume of ~350M examples using n-grams. Since the n-grams resampling favor richer text descriptions, this selects a higher-quality data subset. We augment the resulting data with ~150M examples from five additional sources:

- **Visual grounding.** We link noun phrases in the text to bounding boxes or masks in the image. The grounding information (bounding boxes and masks) are specified in the image-text pair in two ways. (1) We overlay boxes or masks with marks on the image and use marks in the text as reference, akin to set-of-marks (Yang et al., 2023a). (2) We insert normalized $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$ coordinates directly into the text, demarcated by special tokens.
- **Screenshot parsing.** We render screenshots from HTML code and task the model with predicting the code that produced a specific element in the screenshot, akin to Lee et al. (2023). The element of interest is indicated in the screenshot via a bounding box.
- **Question-answer pairs.** We include question-answer pairs, enabling us to use volumes of question-answering data that are too large to be used in model finetuning.
- **Synthetic captions.** We include images with synthetic captions that were generated by an early version of the model. Compared to original captions, we find that synthetic captions provide a more comprehensive description of images than the original captions.
- **Synthetically-generated structured images.** We also include synthetically generated images for a variety of domains such as charts, tables, flowcharts, math equations and textual data. These images are accompanied by a structured representation such as the corresponding markdown or LaTeX notation. Besides improving recognition capabilities of the model for these domains, we find this data useful to generate question-answer pairs via the text model for finetuning.

Grattafiori et al. 2024 https://arxiv.org/abs/2407.21783

Multimodal

- *How to process multimodal input?*



Multimodal

- *How to process multimodal input?*

Here's how tokens are calculated for images:

- **Gemini 1.0 Pro Vision**: Each image accounts for 258 tokens.
- **Gemini 1.5 Flash and Gemini 1.5 Pro**:
  - If both dimensions of an image are less than or equal to 384 pixels, then 258 tokens are used.
  - If one dimension of an image is greater than 384 pixels, then the image is cropped into tiles. Each tile size defaults to the smallest dimension (width or height) divided by 1.5. If necessary, each tile is adjusted so that it's not smaller than 256 pixels and not greater than 768 pixels. Each tile is then resized to 768x768 and uses 258 tokens.

https://cloud.google.com/vertex-ai/generative-ai/docs/multimodal/image-understanding#image-requirements

Multimodal

- *Image generation?*
- *Video? Audio? Code?*
- *Interleaved multimodal inputs?*
- *Interleaved multimodal outputs?*

Efficiency

- *Model architecture*
- *Parallelization and Hardware*
- *Precision*

Efficiency

- *Model architecture*

Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context

Gemini Team, Google[1]

3. Model Architecture

3.1. Gemini 1.5 Pro

Gemini 1.5 Pro is a sparse mixture-of-expert (MoE) Transformer-based model that builds on Gemini 1.0's (Gemini-Team et al., 2023) research advances and multimodal capabilities. Gemini 1.5 Pro also builds on a much longer history of MoE research at Google (Clark et al., 2022; Du et al., 2022; Fedus et al., 2021; Lepikhin et al., 2020; Riquelme et al., 2021; Shazeer et al., 2017; Zoph et al., 2022) and language model research in the broader literature (Anil et al., 2023b; Anthropic, 2023a; Brown et al., 2020; Chowdhery et al., 2023b; Hoffmann et al., 2022; Jiang et al., 2024; Kim et al., 2021; OpenAI, 2023a; Rae et al., 2021; Raffel et al., 2020; Roller et al., 2021; Thoppilan et al., 2022; Touvron et al., 2023a,b; Vaswani et al., 2017). MoE models use a learned routing function to direct inputs to a subset of the model's parameters for processing. This form of conditional computation (Bengio et al., 2013; Davis and Arel, 2014; Jacobs et al., 1991) allows models to grow their total parameter count while keeping the number of parameters that are activated for any given input constant.

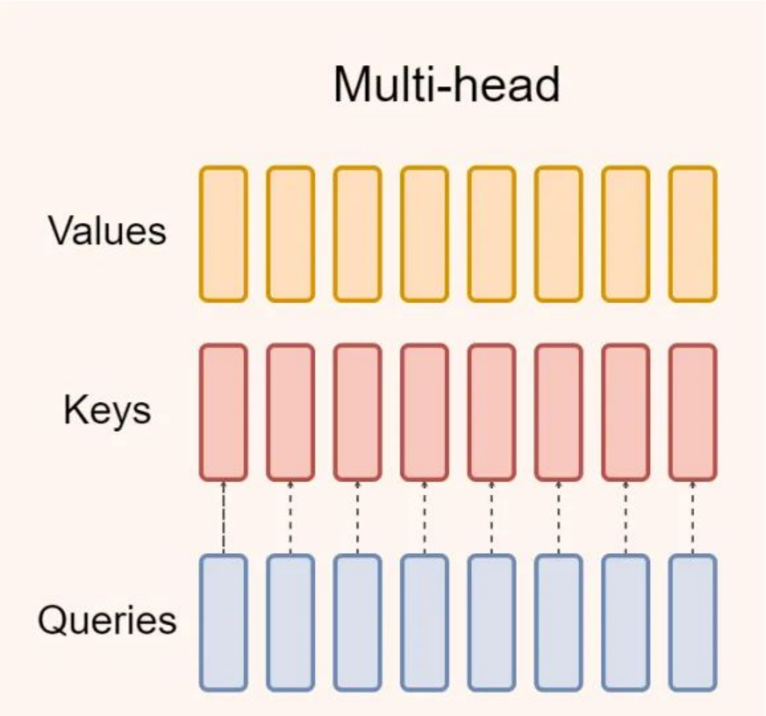
Efficiency

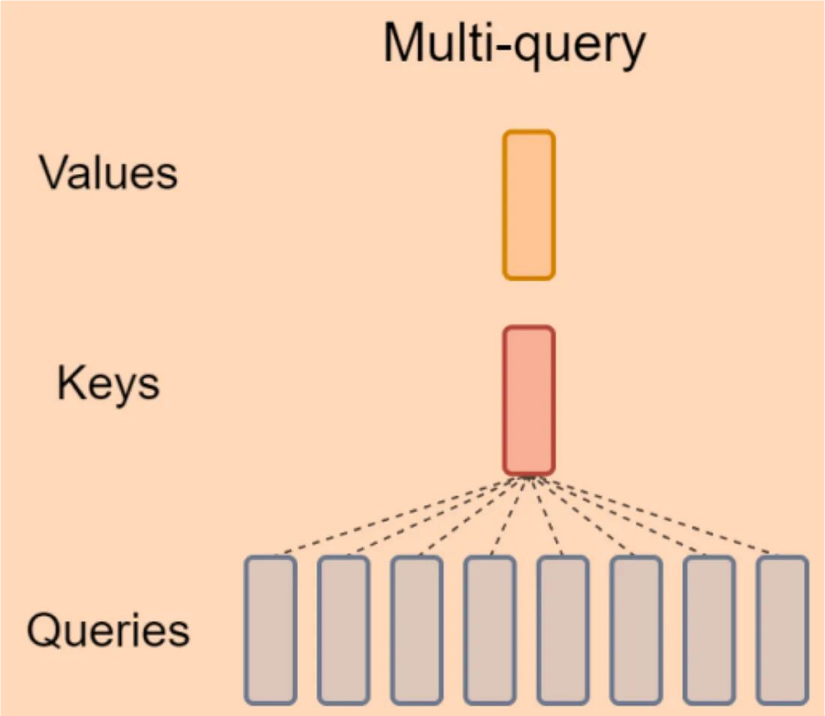
- *Attention mechanisms*

2. Model Architecture

Gemini models build on top of Transformer decoders (Vaswani et al., 2017b) that are enhanced with improvements in architecture and model optimization to enable stable training at scale and optimized inference on Google's Tensor Processing Units. They are trained to support 32k context length, employing efficient attention mechanisms (for e.g. multi-query attention (Shazeer, 2019a)).

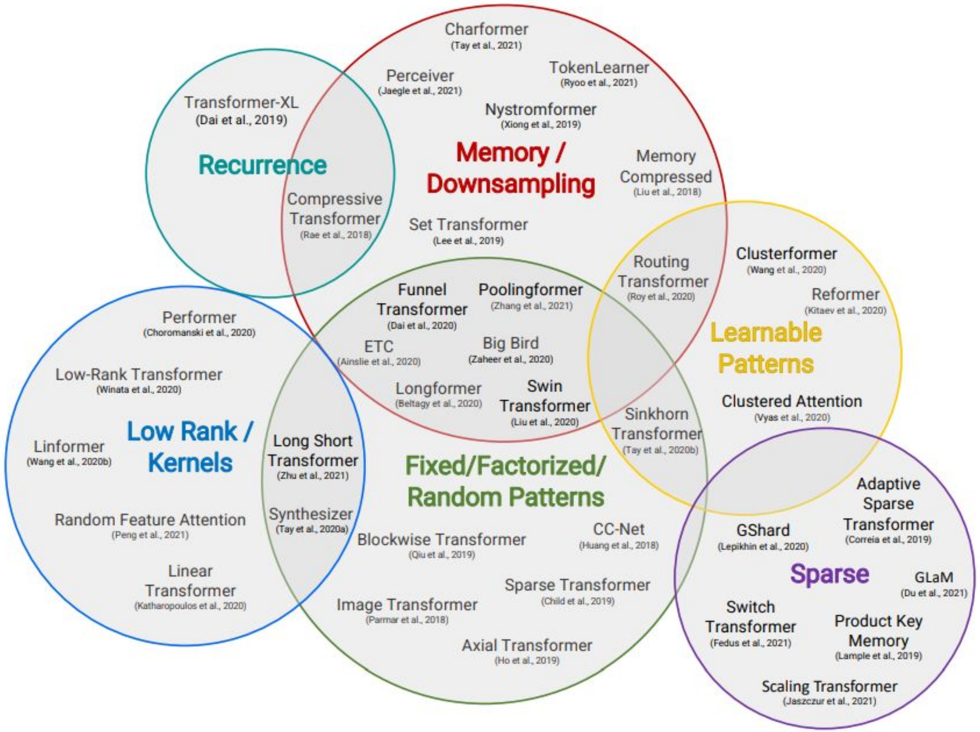
Multi-Query Attention





https://pub.towardsai.net/multi-query-attention-explained-844dfc4935bf

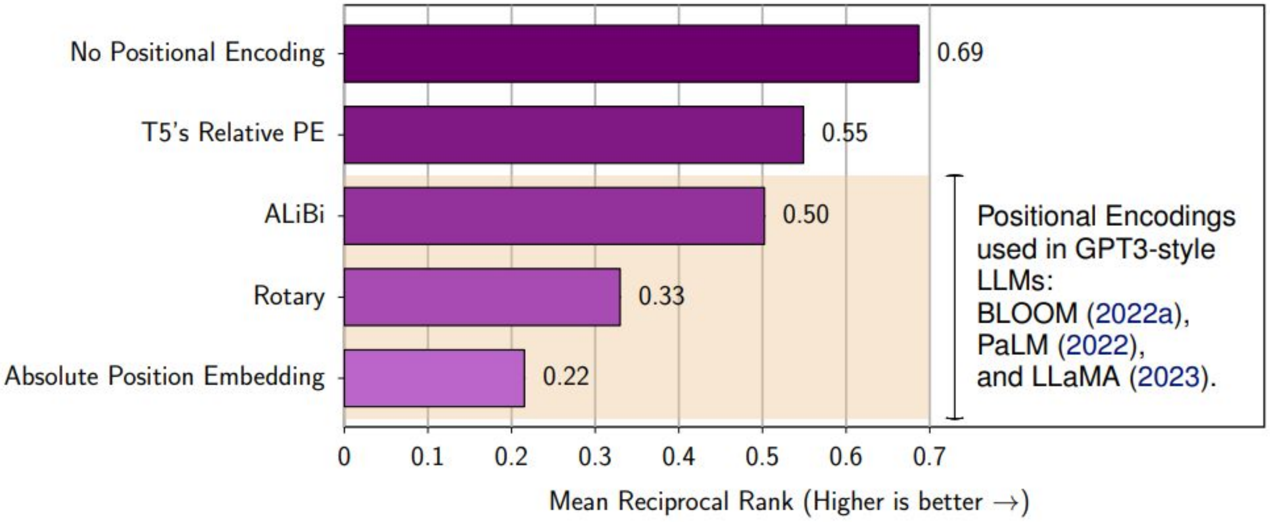
Efficiency



Tay et al. 2022

Efficiency

- *Positional Encoding & Length Generalization*



Kazemnejad et al. 2023

Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context

Gemini Team, Google[1]

In practice, 1 million tokens would look like:

- 50,000 lines of code (with the standard 80 characters per line)
- All the text messages you have sent in the last 5 years
- 8 average length English novels
- Transcripts of over 200 average length podcast episodes

Summary

- **Devil is in the detail**: infrastructure, architecture, data, evaluation. **Massive engineering!**
- **Multimodal brings further challenges**: data sources, how to bridge the gap between different modalities (e.g. tokenization, encoders, adaptors), how to process long inputs, etc.
- **R&D** are on-going on multimodal + X, where X = multilinguality, reasoning, coding, agent, etc.
- Will **enrich/automate existing pipelines** and **enable new applications**

Nando de Freitas
@NandoDF

RL is not all you need, nor attention nor Bayesianism nor free energy minimisation, nor an age of first person experience. Such statements are propaganda.

You need thousands of people working hard on data pipelines, scaling infrastructure, HPC, apps with feedback to drive benchmarks and data, tons of research and engineering on generative models, data mixtures, ablations, RL/selftraining, etc etc and we will probably need lots of people working hard to figure out safety, causal world models, awareness, models that create abstractions comparable to infinity and zero and use these to predict the existence of things like black holes and suggest experiments to verify such hypothesis, or come up with novel engineering designs to generate energy more efficiently, robotics, etc etc.

Applications
(Brief)

Generative AI is a hammer and no one knows what is and isn’t a nail

Colin Fraser · Follow

35 min read · Feb 22, 2024

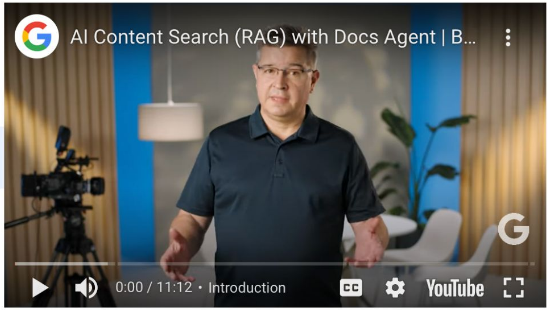
1.3K 21

https://medium.com/@colin.fraser/generative-ai-is-a-hammer-and-no-one-knows-what-is-and-isnt-a-nail-4c7f3f0911aa

Build an AI content search with Docs Agent

Send feedback

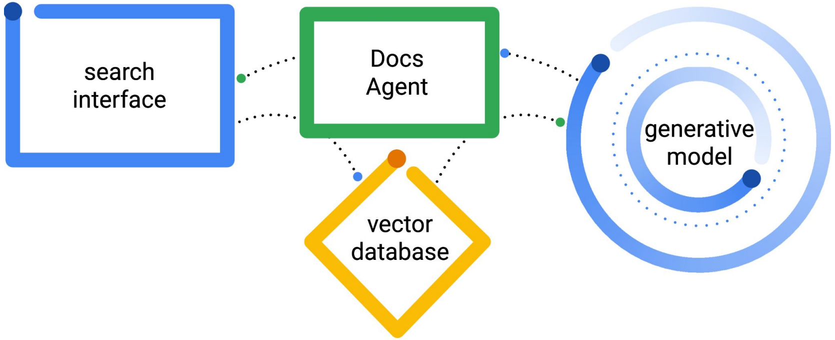
Searching for information is one of the most common uses of artificial intelligence (AI) generative models. Building a conversational search interface for your content using AI allows your users to ask specific questions and get direct answers.



Note: This example is updated for use with Gemini API.

This tutorial shows you how to build an AI-powered, conversational search interface for your content. It's based on Docs Agent, an open source project that uses Google Gemini API to create a conversational search interface, *without training a new AI model or doing model tuning with Gemini models*. That means you can get this search capability built quickly and use it for small and large content sets.

For a video overview of the project and how to extend it, including insights from the folks who build it, check out: AI Content Search | Build with Google AI. Otherwise you can get started extending the project following the instructions below.



https://ai.google.dev/gemini-api/tutorials/docs-agent https://www.youtube.com/watch?v=LTJb76UHuJg Based on this video, what affects the performance the most?

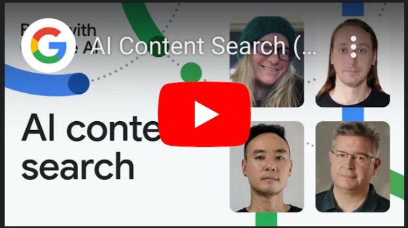
Regenerate draft

YouTube

According to the video, splitting the text into meaningful chunks, such as headings and related paragraphs, is the most important thing for good performance.

YouTube

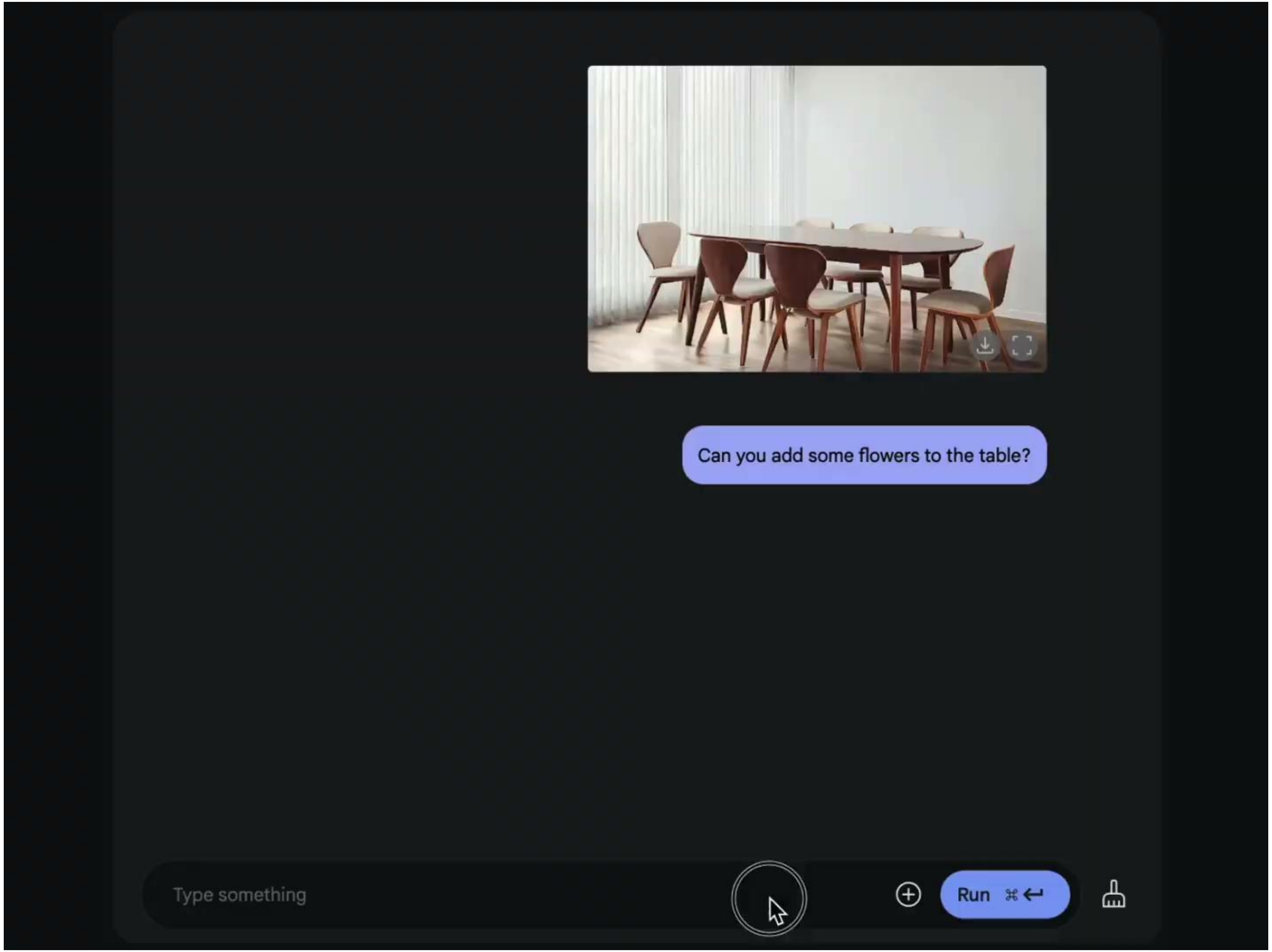
what affects the performance the most



AI Content Search (RAG) with Docs ...

Google for Developers

Learn how to build an AI-powered conversational search interface for your content using the Googl...


Can you add some flowers to the table?
Type something
Run ⌘↵


What will you build?
Push Gemini to the limits of what AI can do using the Gemini API
Image Editing
Edit an image of croissants.
Visual Story
Generate a story with images.
Birthday Card
Design a custom birthday card.
Give me a recipe for a chocolate chip cookie.
Run ⌘↵

Agents

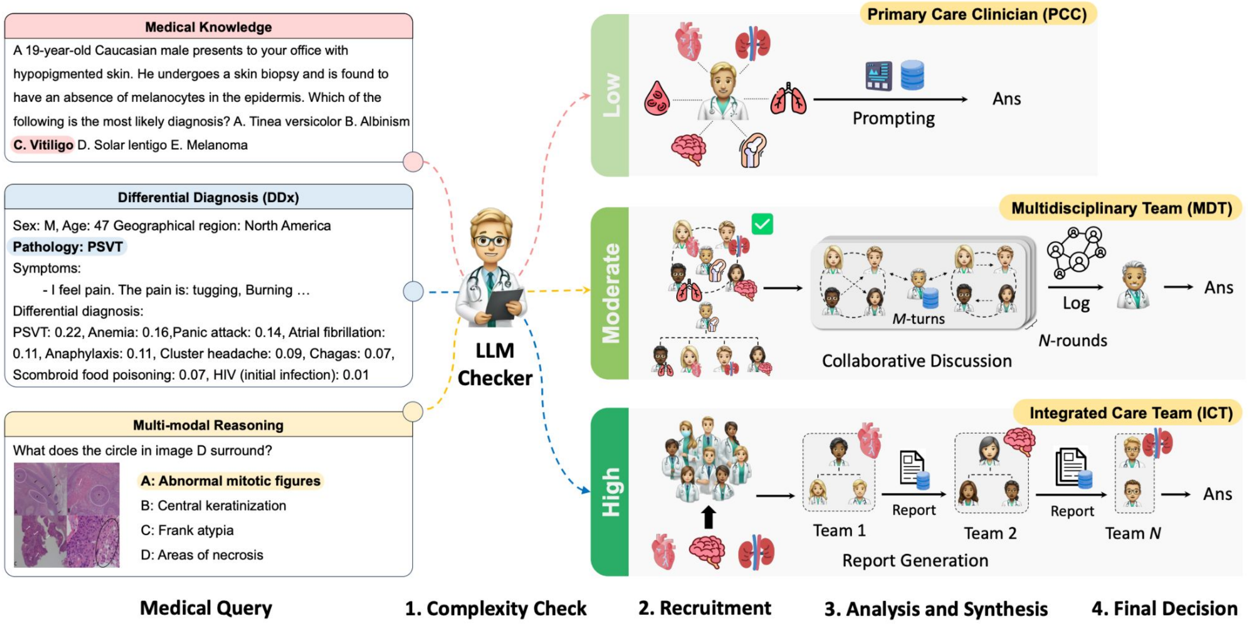
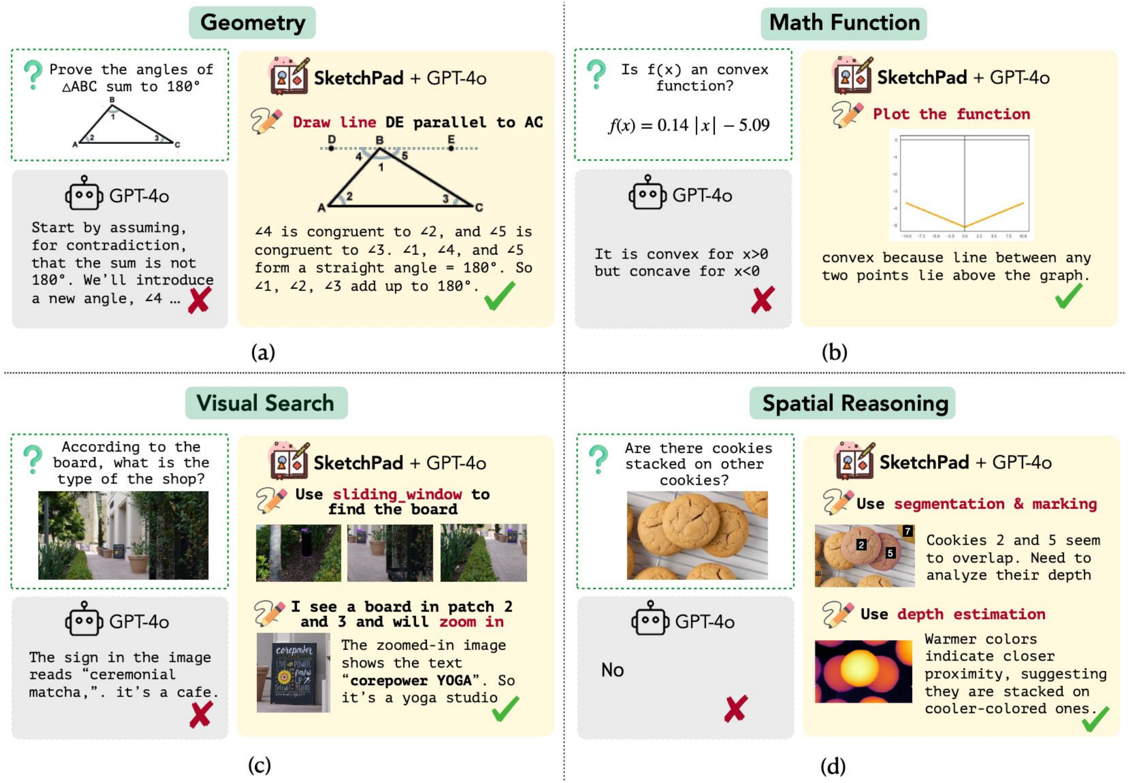


Figure 1: **Medical Decision-Making Agents (MDAgents) framework**. Given a medical query from different medical datasets, the framework performs 1) medical complexity check, 2) recruitment, 3) analysis and synthesis, and 4) decision-making steps.

Kim et al. 2024

Agents



Hu et al., 2024

Conclusion

- Fundamentals & Evolution
- Multimodal LLMs: Modeling, Data, Evaluation
- Example: Gemini

Conclusion

- Fundamentals & Evolution
- Multimodal LLMs: Modeling, Data, Evaluation
- Example: Gemini

But this is just the beginning!

- Learn, explore, research, experiment (perhaps with LLMs) to advance the sciences and engineering
- Tailor (multimodal) LLMs to your specific tasks or domains