Transformer / GPT

What is the mechanism in a generative-AI?

What was the turning point?

How the AI's are trained?

How can they "think"?

What can they do?

understanding?

consciousness?

A. Vaswani *et al.*, *Attention is All You Need*, https://doi.org/10.48550/arXiv.1706.03762

S. Bubeck *et al.*, *Sparks of Artificial General Intelligence*, https://doi.org/10.48550/arXiv.2303.12712

...

A. Tamii

RCNP / IRS / Dep. Phys, Osaka Univ.

Outline

- I. How a generative-AIs works

- II. What a generative-AIs can do?

- III. Personal opinions and discussions

Motivation

Recent evolutional development of generative-AIs.

- I'm simply interested in the mechanism of the AIs.
- As a researcher in science, I want to understand the AIs are.

Evolution of the AI's is *exponential*

that often goes beyond prediction.

I. How a Generative-AIs work

the mechanism

Transformer: A. Vaswani *et al.*,*Attention is All You Need*

https://doi.org/10.48550/arXiv.1706.03762

Natural Language Processing (NLP)

A field in the Artificial Intelligence (AI) technology.

others: machine learning, computer vision, speech recognition, robotics, expert systems, evolutionary algorithms, …

Capability of using a language: one of the keys for human civilization

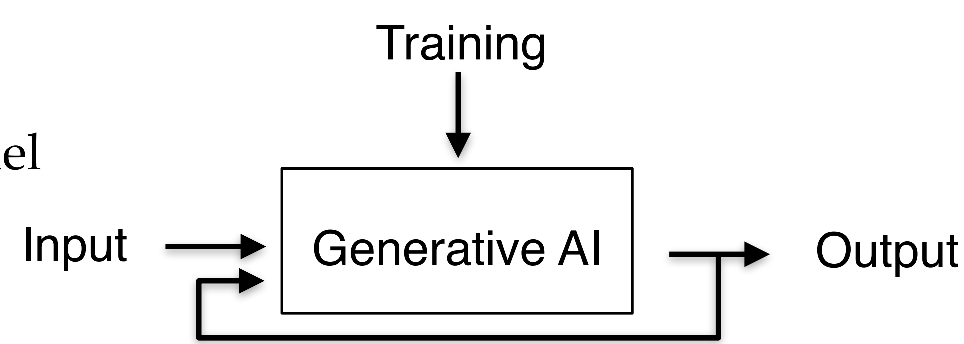
→ applies also to the evolution of AIs

Generative AI:

AI models that can **create new contents** such as text, images, or music similar in structure and features to the training data.

sequence to sequence model

[I. Sutskever, O. Vinyals, Q. V. Le (google) NIPS-2014]



Transformer

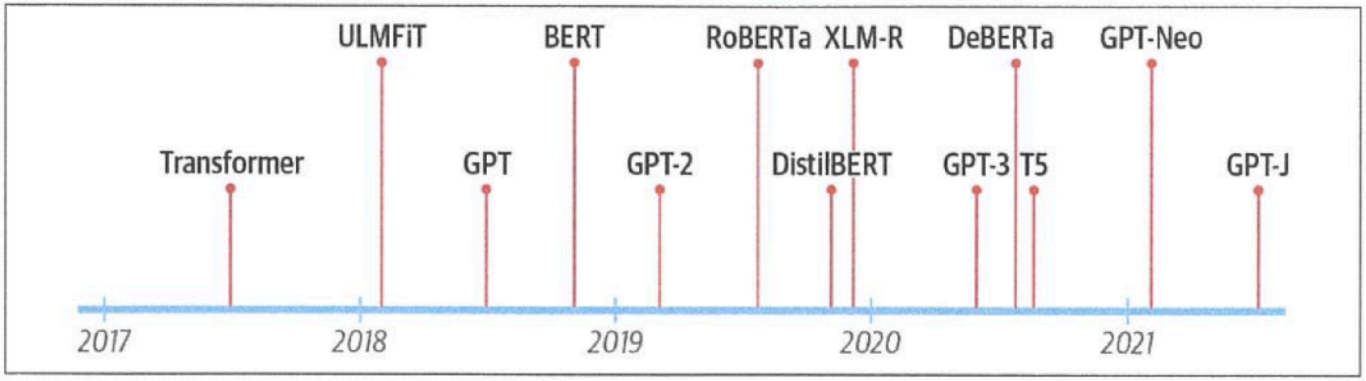


Figure 1-1. The transformers timeline

Transformer was innovated in 2017.

→ Revolution in the NLP technology

followed by many branch models.

BERT, BART, GPT, BIRD…

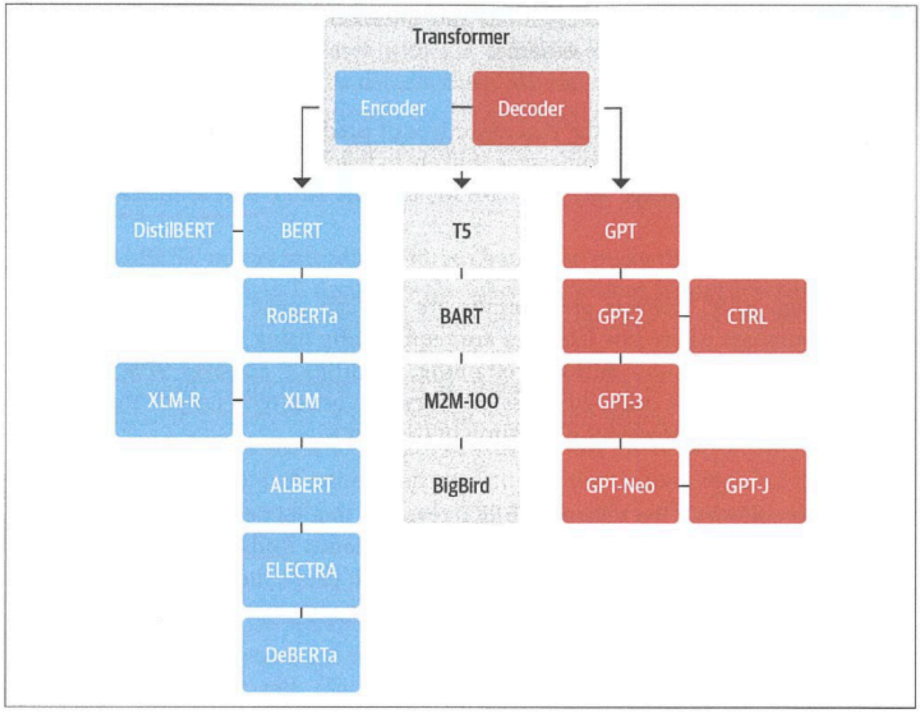


Figure 3-8. An overview of some of the most prominent transformer architectures

Before Transformer

1982-: RNN: Recurrent Neural Network

J. Hopfield,…

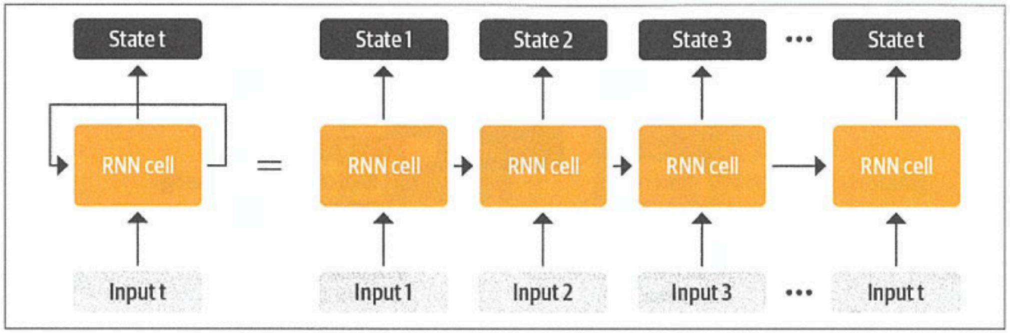


Figure 1-2. Unrolling an RNN in time

no direct memory of previous inputs

but *hidden state vector* of previous steps

→ analogy: reading books

Vanishing gradient problem (for long term dependence)

Long training time, computational cost, scalability problem

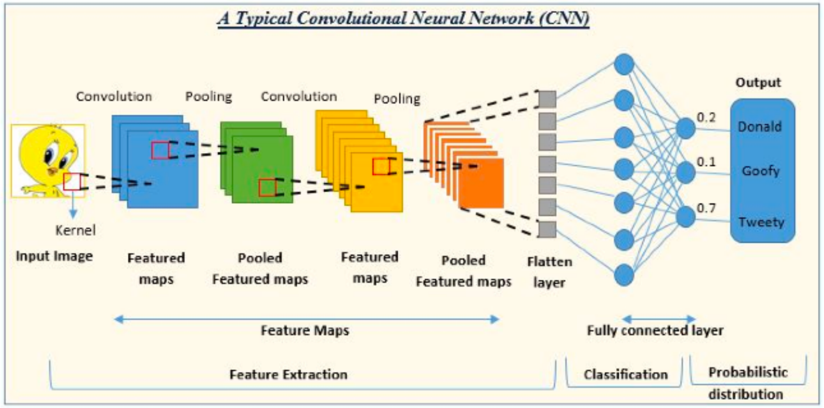
1980-: CNN: Convolutional Neural Network

Grid-like structure for image or video recognition

filters, pooling layers

lack of interpretability, computational costs

K. Fukushima, Y. LuCun, …



Transformer

Transformer: 2017 (Google)

A. Vaswani *et al.*, *Attention is All You Need* arXiv.1706.03762

A break through: greatly accelerated the NLP technology

- encoder-decoder framework
- self-attention mechanism
- transfer learning

Simplicity and **Scalability**

Transformer: Overview

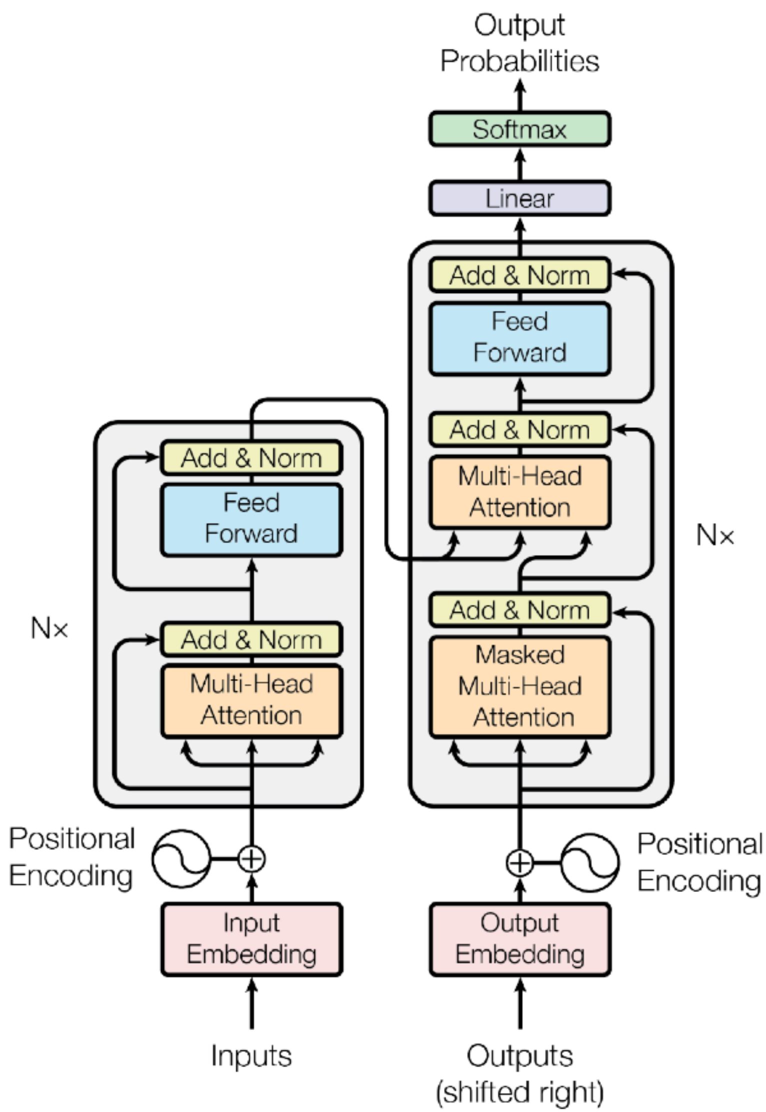


Figure 1: The Transformer - model architecture.

A. Vaswani *et al*. (2017) [arXiv.1706.03762](https://arxiv.org/abs/1706.03762)

Encoder / Decoder

Encorder:

translates data (context)

from external representation

to internal representation

Decoder:

translates data (context)

from internal representation

to external representation

Practical numbers shown in the slides are from Transformer (2017)

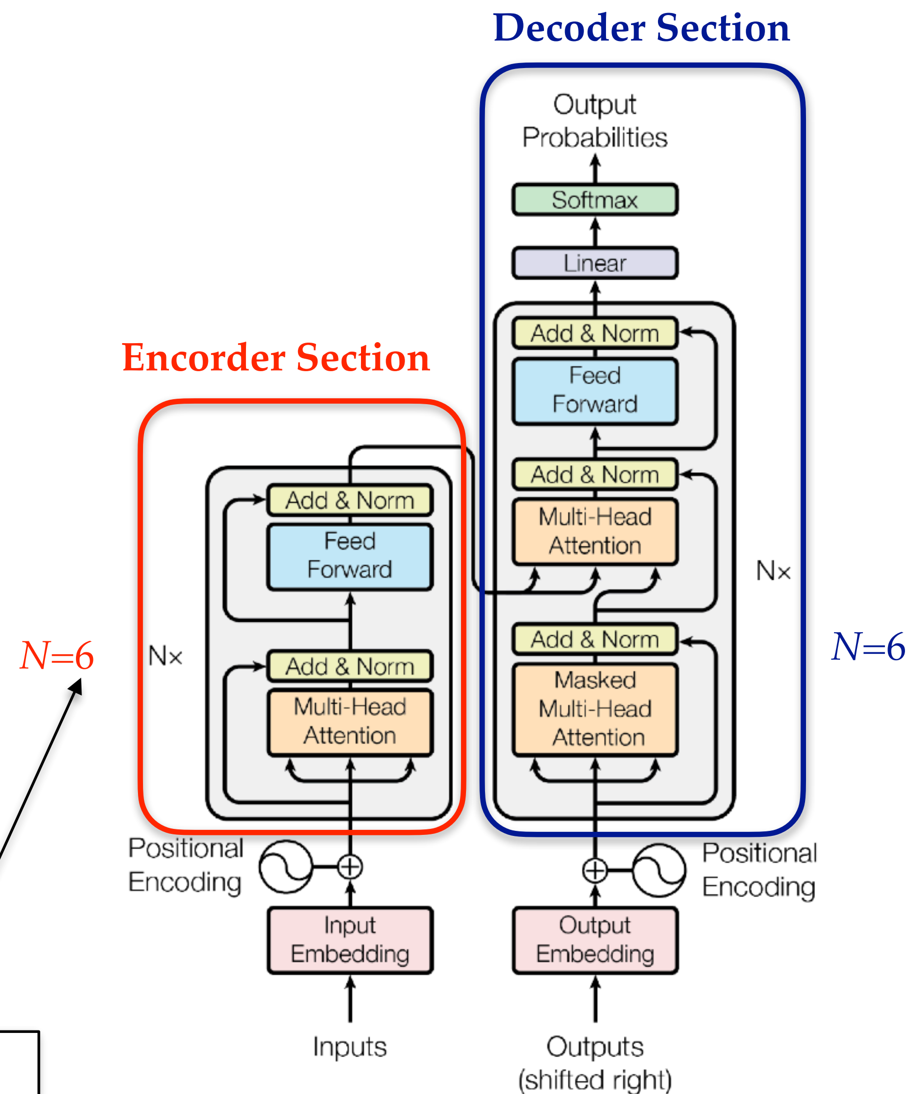
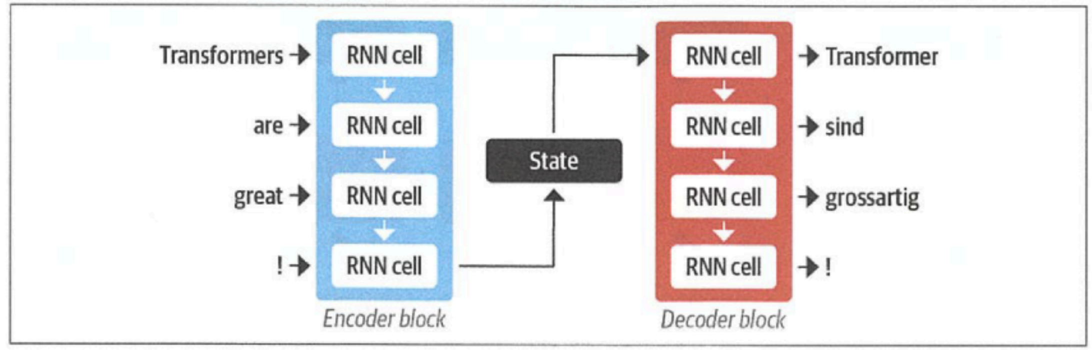


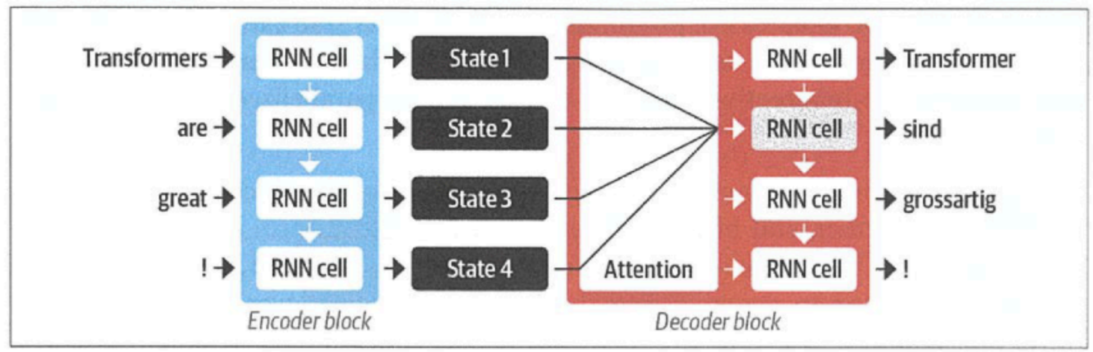
Figure 1: The Transformer - model architecture.

Overview

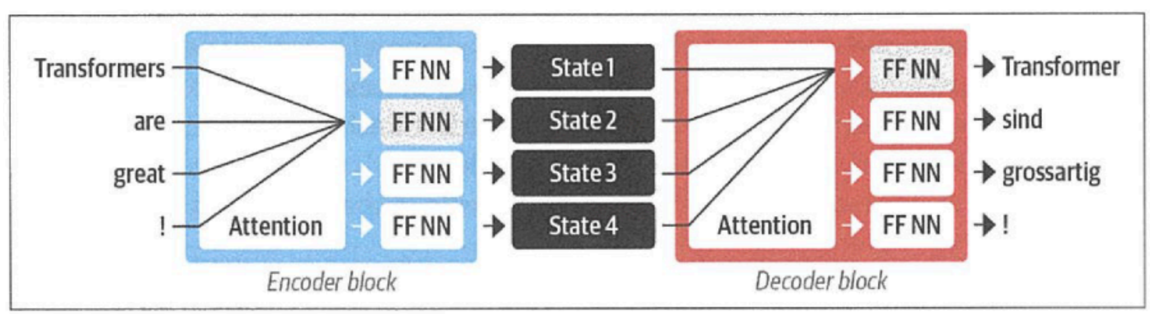
Encoder-decoder architecture



Encoder-decoder architecture with **attention** mechanism

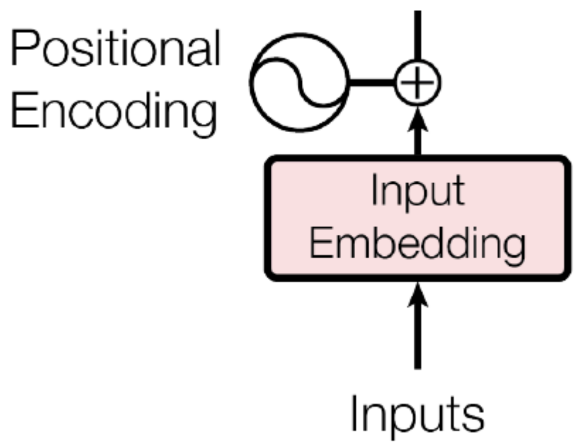


Encoder-decoder architecture of the transformer



FF NN: Feed-forward Neural Network

Input Embedding



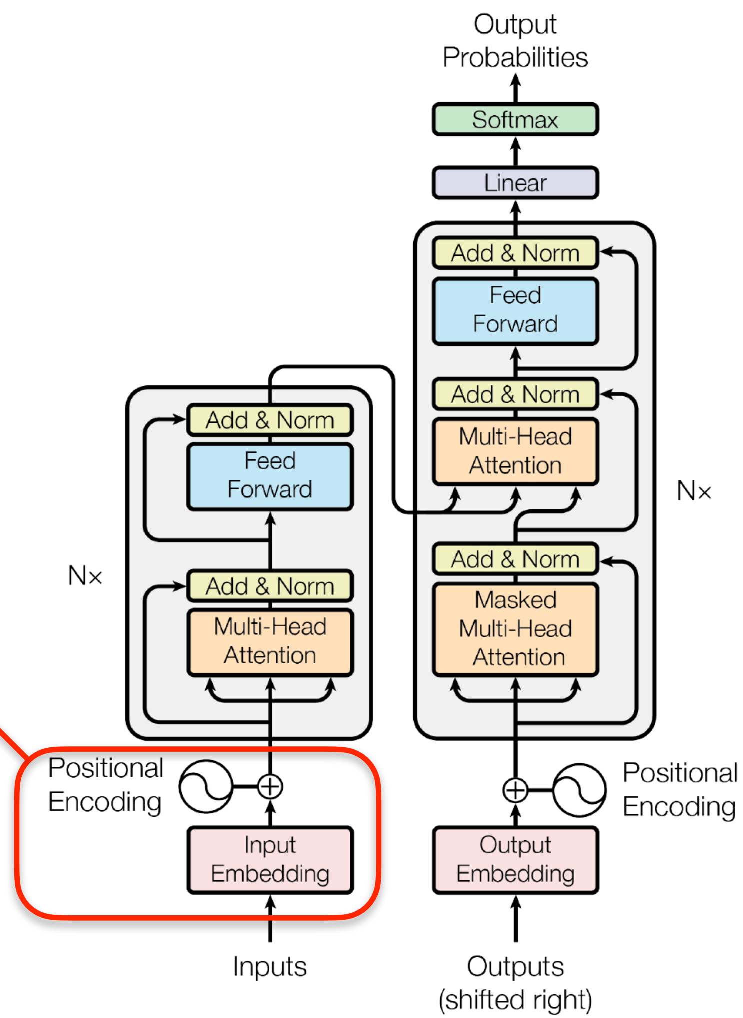


Figure 1: The Transformer - model architecture.

Tokenizer

Transfer each of the input *tokens* to an internal representation (*vector*)

Various types of algorithms.

An example: word2vec model, ski-gram architecture

A token is a little smaller than a word in English.

e.g. "pre-train-ing": three tokens

$\lesssim$ a character in Japanese

→ token-recognition

```
Text = "The cat slept on the couch.It was too tired to get up."
tokenized text= [1996, 4937, 7771, 2006, 1996, 6411, 1012, 2009, 2001,
2205, 5458, 2000, 2131, 2039, 1012]
```

→ internal vector representation

black =

```
black=[[-0.01206071  0.11632373  0.06206119 0.01403395  0.09541149
   0.10695464
  0.02560172  0.00185677 -0.04284821  0.06146432  0.09466285  0.04642421
  0.08680347  0.05684567 -0.00717266 -0.03163519  0.03292002 -0.11397766
  0.01304929  0.01964396  0.01902409  0.02831945  0.05870414  0.03390711
 -0.06204525  0.06173197 -0.08613958 -0.04654748  0.02728105 -0.07830904
 ...
  0.04340003 -0.13192849 -0.00945092 -0.00835463 -0.06487109  0.05862355
 -0.03407936 -0.00059001 -0.01640179  0.04123065 -0.04756588  0.08812257
  0.00200338 -0.0931043  -0.03507337  0.02153351 -0.02621627 -0.02492662
 -0.05771535 -0.01164199 -0.03879078 -0.05506947  0.01693138 -0.04124579
 -0.03779858 -0.01950983 -0.05398201  0.07582296  0.00038318 -0.04639162
 -0.06819214  0.01366171  0.01411388  0.00853774  0.02183574 -0.03016279
 -0.03184025 -0.04273562]]
```

brown =

```
brown=[[ 1.35794589e-02 -2.18823571e-02 1.34526128e-02 6.74355254e-02
         1.04376070e-01  1.09921647e-02 -5.46298288e-02 -1.18385479e-02
         4.41223830e-02 -1.84863899e-02 -6.84073642e-02  3.21860164e-02
         4.09143828e-02 -2.74433400e-02 -2.47369967e-02  7.74542615e-02
         9.80964210e-03  2.94299088e-02  2.93895267e-02 -3.29437815e-02
         ...
         7.20389187e-02  1.57317147e-02 -3.10291946e-02 -5.51304631e-02
        -7.03861639e-02  7.40829483e-02  1.04319192e-02 -2.01565702e-03
         2.43322570e-02  1.92969330e-02  2.57341694e-02 -1.13280728e-01
         8.45847875e-02  4.90090018e-03  5.33546880e-02 -2.31553353e-02
         3.87288055e-05  3.31782512e-02 -4.00604047e-02 -1.02028981e-01
         3.49597558e-02 -1.71501152e-02  3.55573371e-02 -1.77437533e-02
```

$d_{\text{model}} = 512$

a vector of real number2
with a dimension of 512

The way of the mapping is
optimized in the training stage.

Tokenizer

A vector is an internal representation of the concept of the token.

cosine_similarity(black, brown) = [[0.9998901]]

~ the inner-product of the two vectors

represents *similarity* of the concept of the two tokens.

Positional Encoding

Embedding the information of the position of each input token in the d_{model} dimension vector

Several methods are proposed.

For example, in Vaswani (2017)

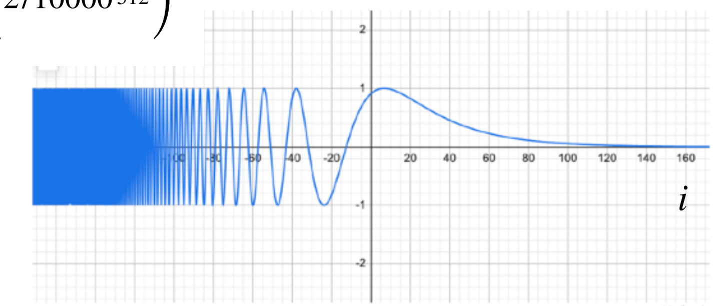
$$i = 0, 1, \ldots, d_{\text{model}} - 1$$

$$\mathrm{PE}[2i] = \sin\left(\frac{\mathrm{pos}}{10000^{\frac{2i}{d_{\text{model}}}}}\right) \quad \text{odd-index}$$

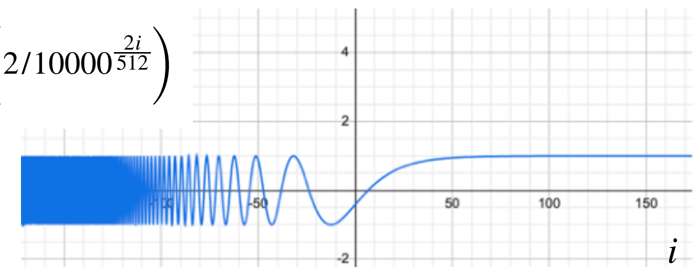
$$\mathrm{PE}[2i+1] = \cos\left(\frac{\mathrm{pos}}{10000^{\frac{2i}{d_{\text{model}}}}}\right) \quad \text{even-index}$$

pos = 2

$\sin\left(2/10000^{\frac{2i}{512}}\right)$



$\cos\left(2/10000^{\frac{2i}{512}}\right)$

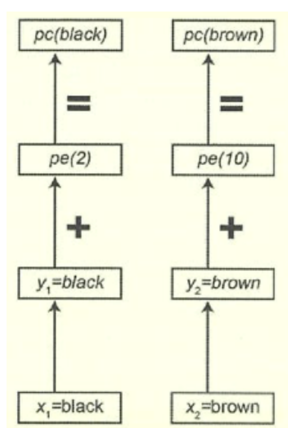


pos. encoding

pos. vector

vector

input



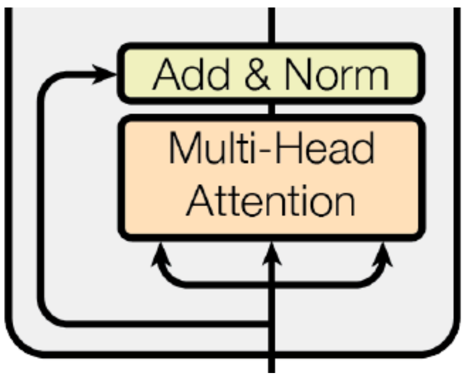
positional encoding

Simply, the sum of the token and position vectors!

※Concept is addable.

Multi-head Attention Layer

Multi-Head Attention Layer



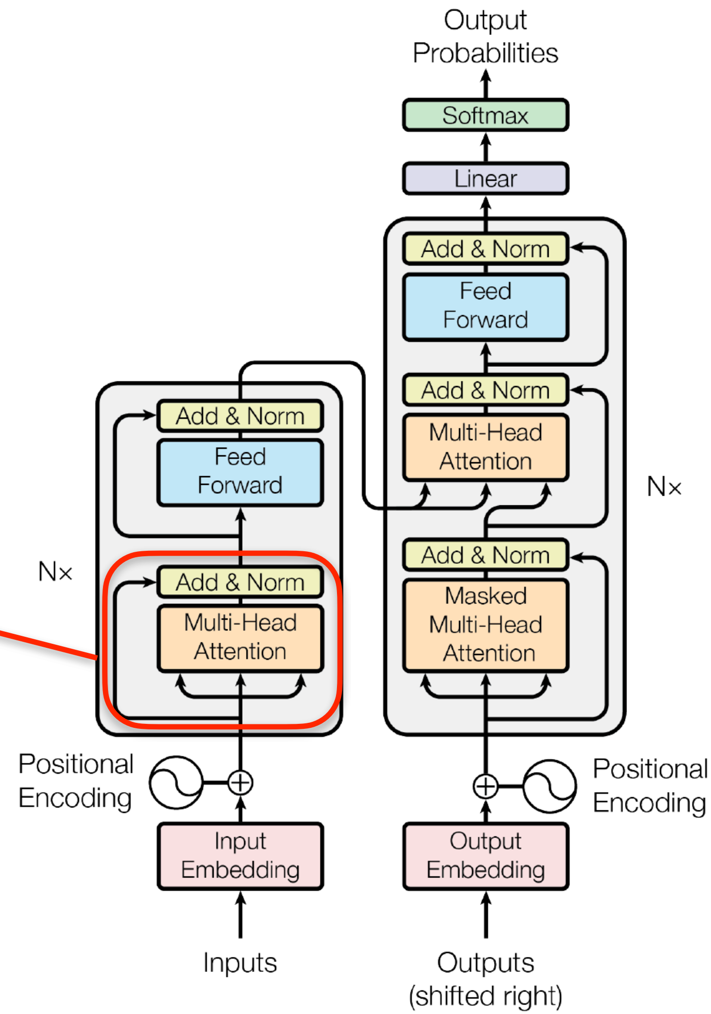


Figure 1: The Transformer - model architecture.

Attention

Attention: D. Bandana, KH Cho, and Y. Bengio (2016) arXiv.1409.0473

stimulated from a brain mechanism

	dim.	matrices
Q: Queries	$d_q = 64$	Q_w
K: Keys	$d_k = 64$	K_w
V: Values	$d_v = 64$	V_w

Represented by (weight) matrices

optimized in the training stage

※Weight matrices are independent for each layer.

$$\text{Attention}(Q, K, V) \equiv \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$\text{Attention}\left(\text{a} \rightarrow \text{b}\right) \equiv \left[\left(Q \text{ of token a}\right) \cdot \left(K \text{ of token b}\right)\right]\left(V \text{ of token b}\right)$$

Analogies:

at a supermarket for making a dish

Queries:	ingredients in a recipe
Keys:	labels of the item
Values:	items

at a library

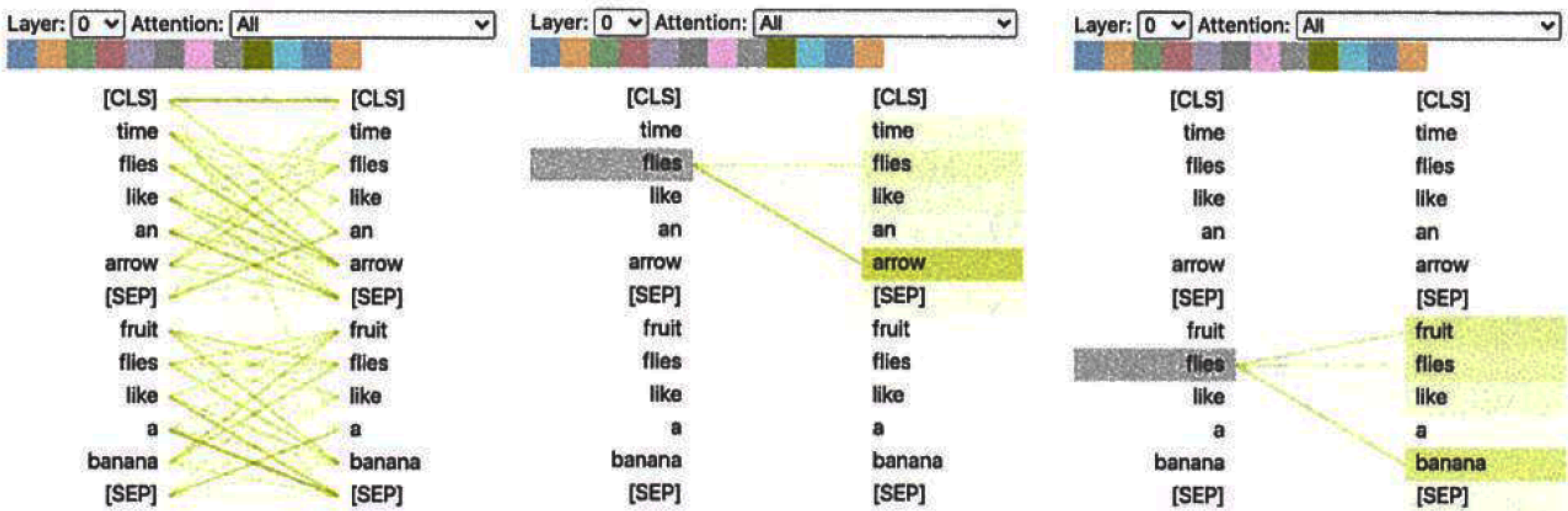
Queries:	subjects you want to learn
Keys:	book titles
Values:	books

Self-Attention

Self-Attention

- Attention between the tokens in the input

→ Relative strength of the relation between each token pair



- efficient way of keeping **connection with tokens at long distance**

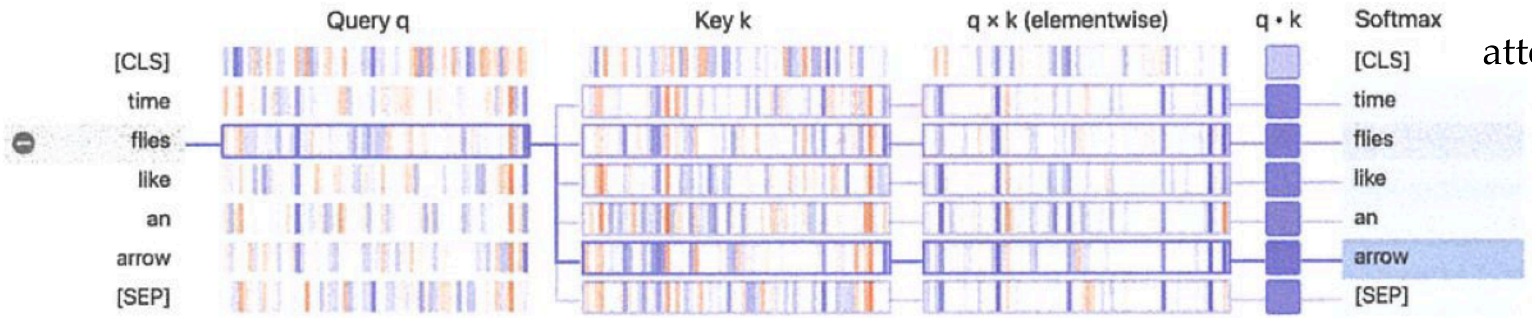
for Vanishing gradient problem

- efficient **parallel calculation** for efficient training and a large model

recursive logic —> attention

Self-Attention

Self-Attention



weight vector

attentions

$w_0 \times value_0$
$w_1 \times value_1$
$w_2 \times value_2$
$w_3 \times value_3$
$w_4 \times value_4$
$w_5 \times value_5$

sum

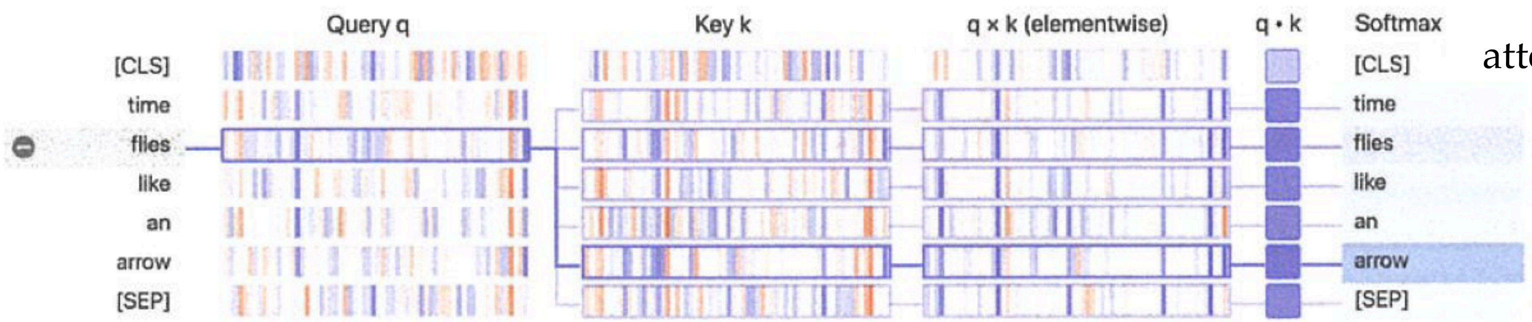
attention scores

w = inner product
(with normalization)

result vector
to the next layer

※Concept is addable.

Self-Attention



weight vector

attentions

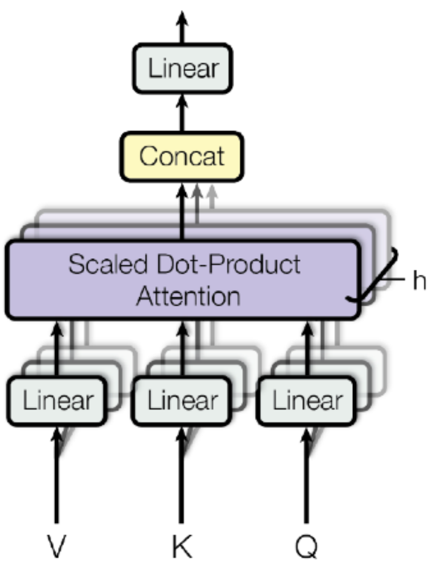
$w_0 \times value_0$
$w_1 \times value_1$
$w_2 \times value_2$
$w_3 \times value_3$
$w_4 \times value_4$
$w_5 \times value_5$

sum

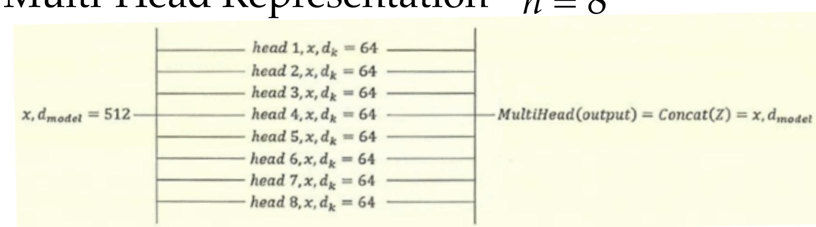
attention scores

result vector to the next layer

Multi-Head Attention



Multi-Head Representation $h = 8$



For each input, attention is calculated in parallel by h heads.

The self-attention works in parallel for each embedding.

Self-Attention

Questions:

- What does an output vector of a self-attention layer mean?

  → Internal representation of the *concept* of

  e.g. “flying arrow” with some flavors of “time", ‘like”, …

  as an output corresponding to the embedding of “flies”

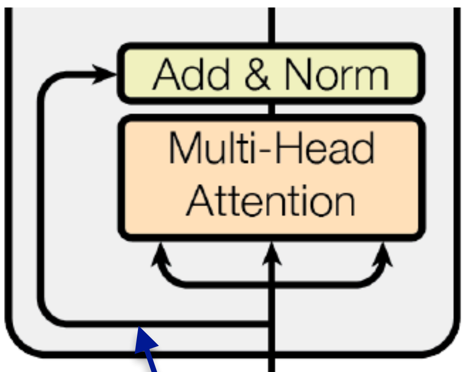
- Why **queries, keys, and values** are used, instead of using **directly using the input vectors** from the previous layer, to calculate the attention score?

  For the transformer to be able to **focus on different aspects** of the inputs.

  - Each layer has a different abstraction level of concepts.
  - Each head has different abilities

  If the input vectors are directly used, the vector space for representing concepts is limited to just one type.

Multi-head Attention Layer



a side-path for partially preserving
the information of the previous layer

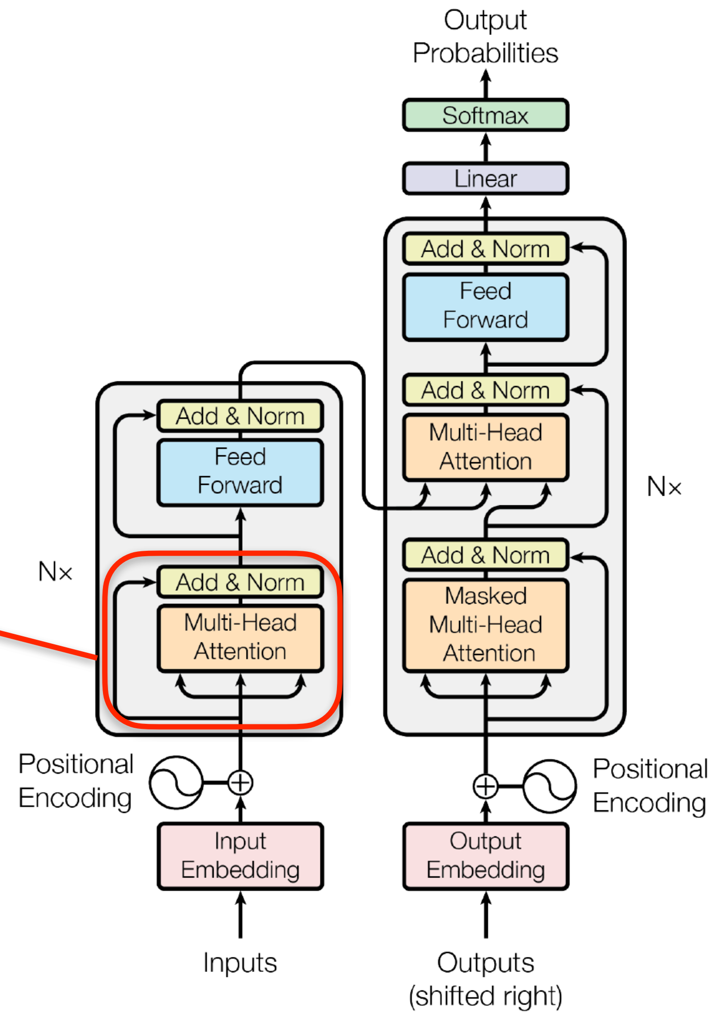


Figure 1: The Transformer - model architecture.

"flying arrow…"

⊕

flies

※Concept is addable.

Layer Normalization

Layer Normalization Processes

for preserving the real numbers in the vectors/matrices to be very large.

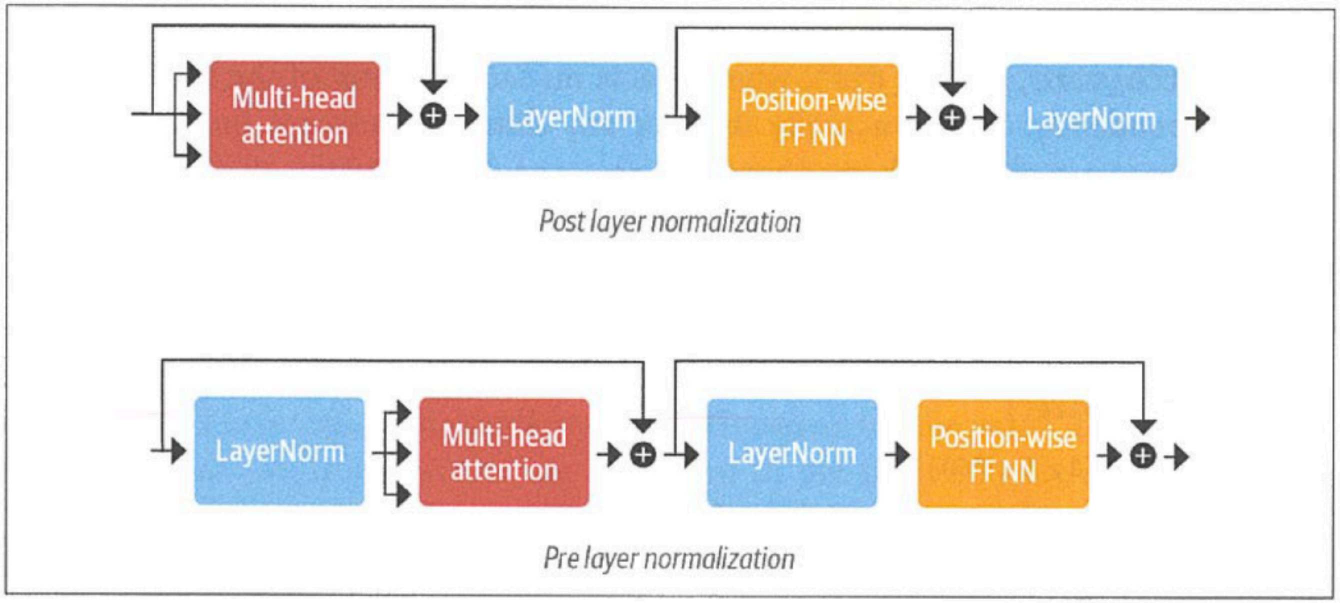


Figure 3-6. Different arrangements of layer normalization in a transformer encoder layer

Feed-Forward NN

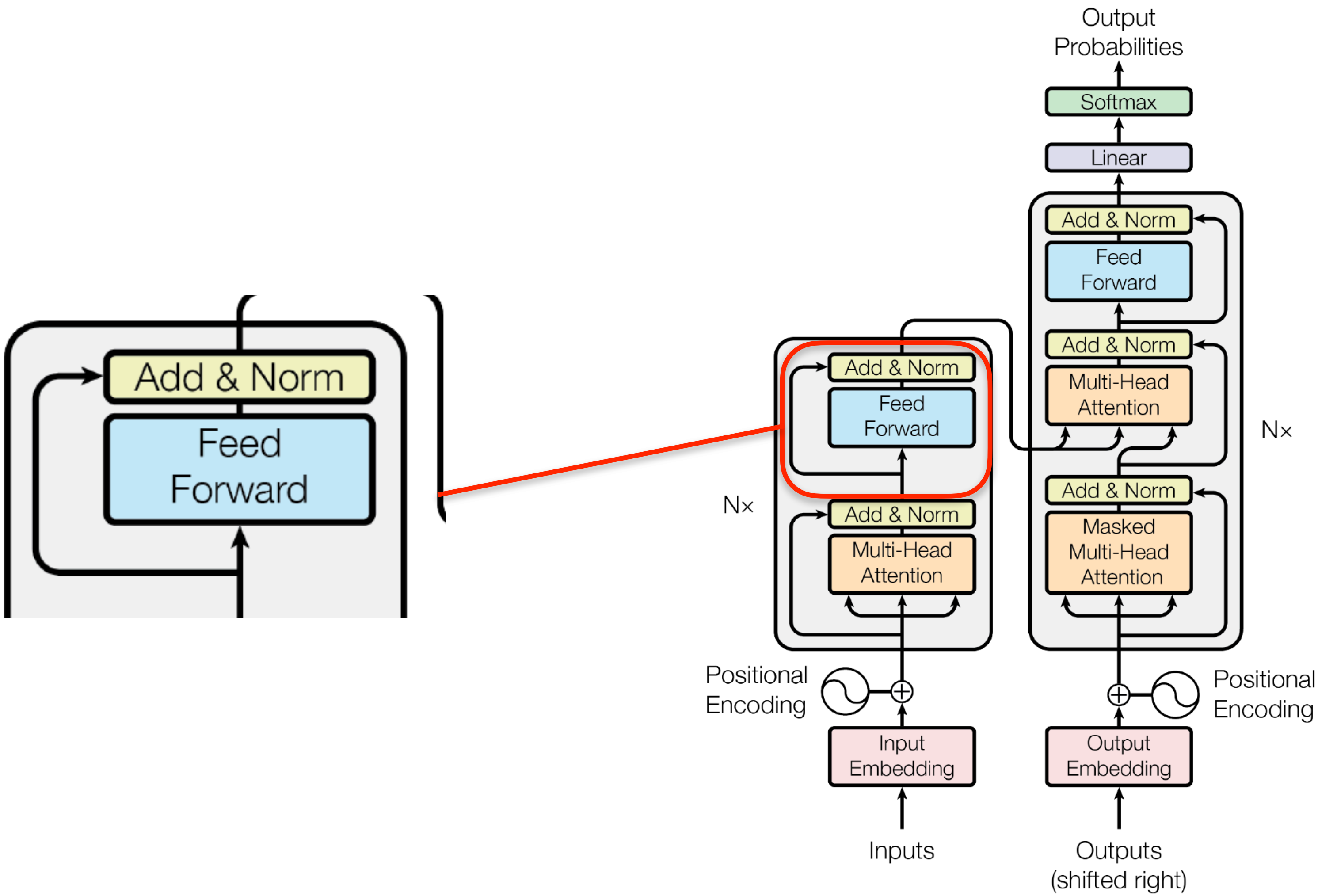


Figure 1: The Transformer - model architecture.

Feed-Forward NN

Feed-Forward Neural Network (FF NN)

Two-layer fully connected neural network

processes each embedding independently

Rectified Linear Unit (ReLU) activation function

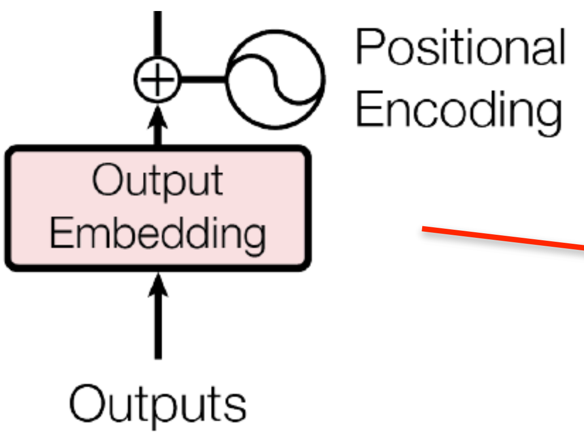
→ Non-linearity

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2$$

for improving the abilities of representation and study

complemental to the attention layer

Output Embedding



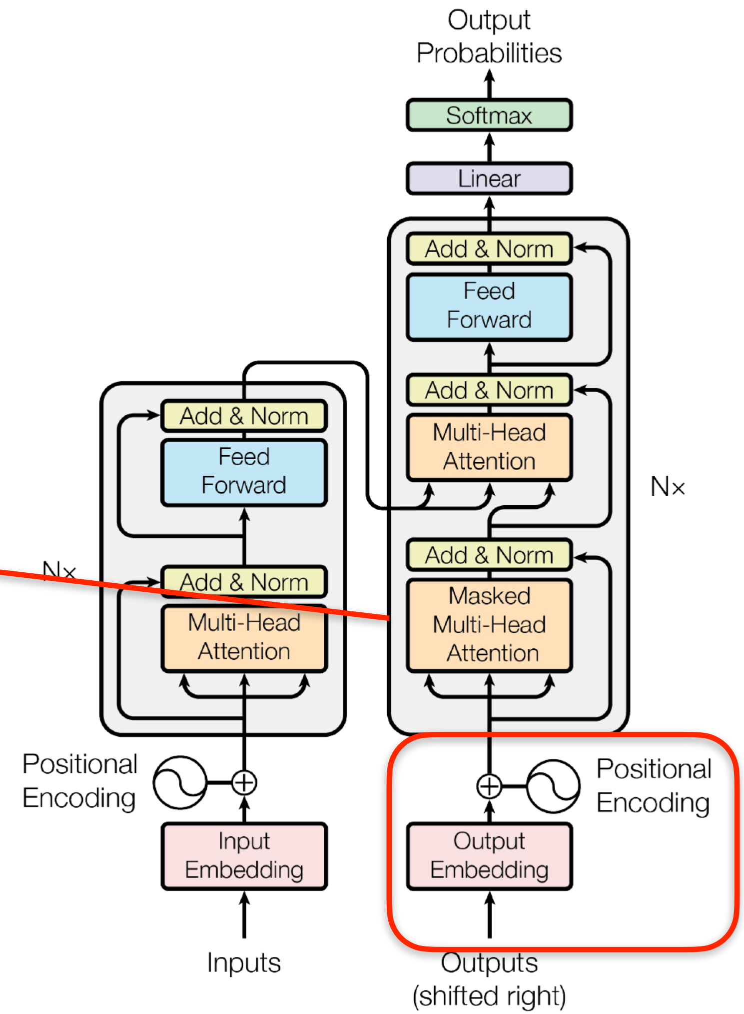


Figure 1: The Transformer - model architecture.

Output Embedding

Output embedding

is similar to the input embedding

The inputs of the output embedding are

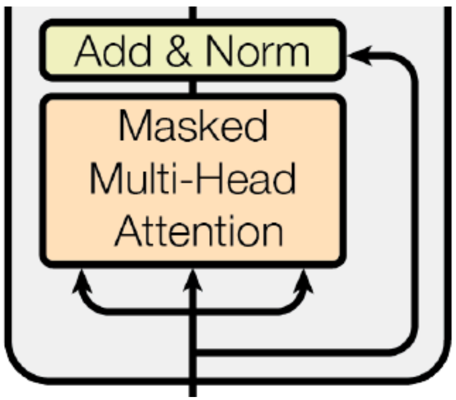
In the training time:

Entire text.
In a training of the translation, for example, well-translated texts (the answers) are included as well as the text to translate (questions).

In the using time:

Entire input text and all the previous outputs from the transformer.

Masked Multi-Head Attention



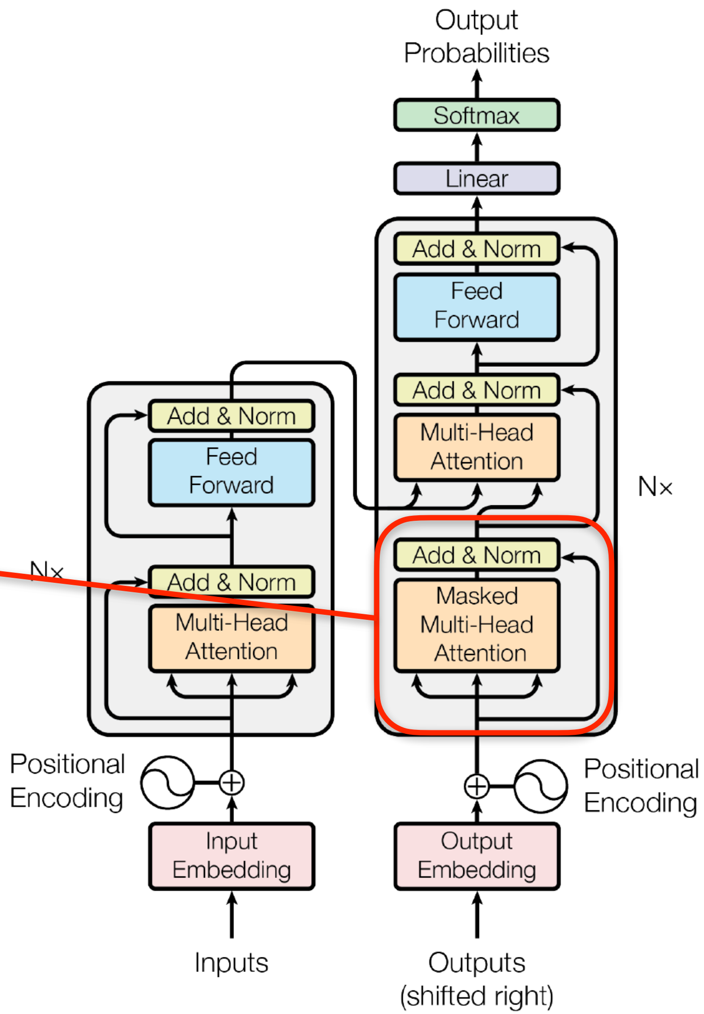


Figure 1: The Transformer - model architecture.

Masked Multi-Head Attention

Masked Multi-Head Attention

is the same as the attention in the encoder section

except that all the tokens after the present token are masked.

The mask prevents the transformer to use the information of the inputs after the present token.

In the stage of using, the future output is unavailable.

→ Prevent using the information after the present token in the training stage.

⇓

The regulation is relaxed in several types of transformer models, e.g. BERT.

Encoder-Decoder Attention

Encoder-Decoder Attention

is a multi-head attention similar to the encoder

inputs are

- **queries** from the previous **decoder** attention

-**keys and values** from the **encoder final block** (of all the input embedding)

The purpose:

to be able to align the output of the decoder using the knowledges in the encoder section

e.g. for language translation.

※**Transferring knowledge**, between the parts, is achieved by **sending keys and values**.

Transformer: Questions

Where is the knowledge stored in the transformer?

In the weight matrices of each attention layer and the biases in each FFNN layer.

Semantics and syntactics are in the embedding/output layers.

Why the AIs can work as human with this simple structure?

No answer.

However, note that many conceptual mechanism of an AI are taken from the human brain. Thus, an AIs can work as a human brain can work.

The necessary ingredients are key concepts (attention/NN) and a sufficiently large complexity.

Transfer Learning

Transfer-Learning

- Create a trained model: **Pre-Trained** … costs a lot of resources
- Add a new layer(s)
- Create a re-trained model: **Fine-Tuning** … costs much less

Analogy:

pre-training:

education of children up to the general education at a university

fine-tuning:

training of a student to be an expert in a specific field

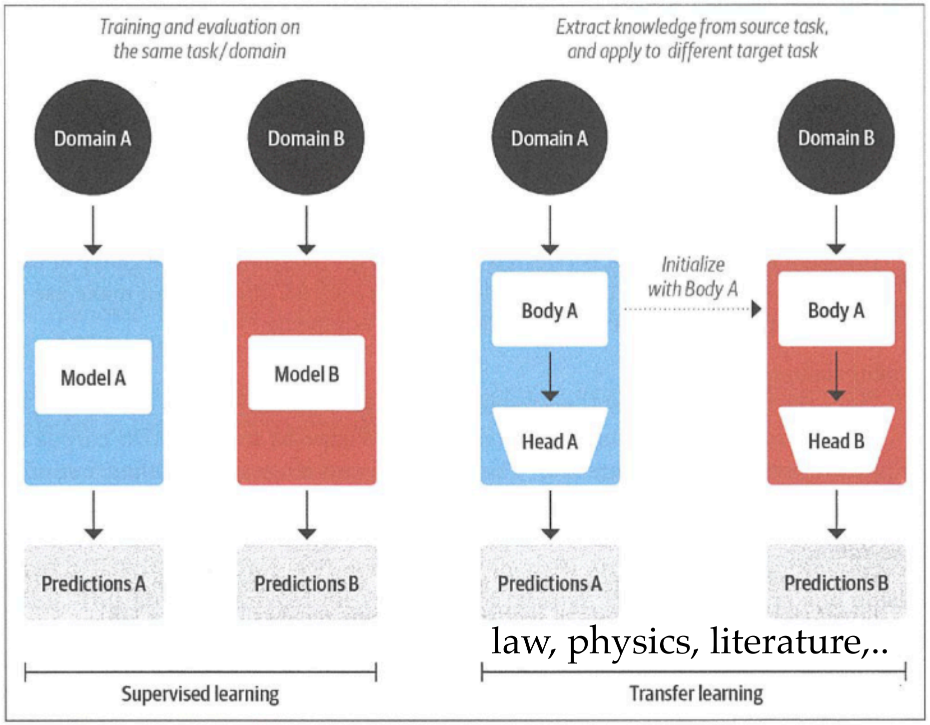


Figure 1-7. Comparison of traditional supervised learning (left) and transfer learning (right)

Multilingual Transformers

Recent generative-AI's (e.g. BERT) are trained with multiple languages

- Deep training in a few major languages, e.g. English

  … large resources available

- Less training in minor languages

  … limited resources for training

→ Better performance has been achieved in minor languages by training an AI in major languages.

※Knowledge is transferred across the languages.

Transformer

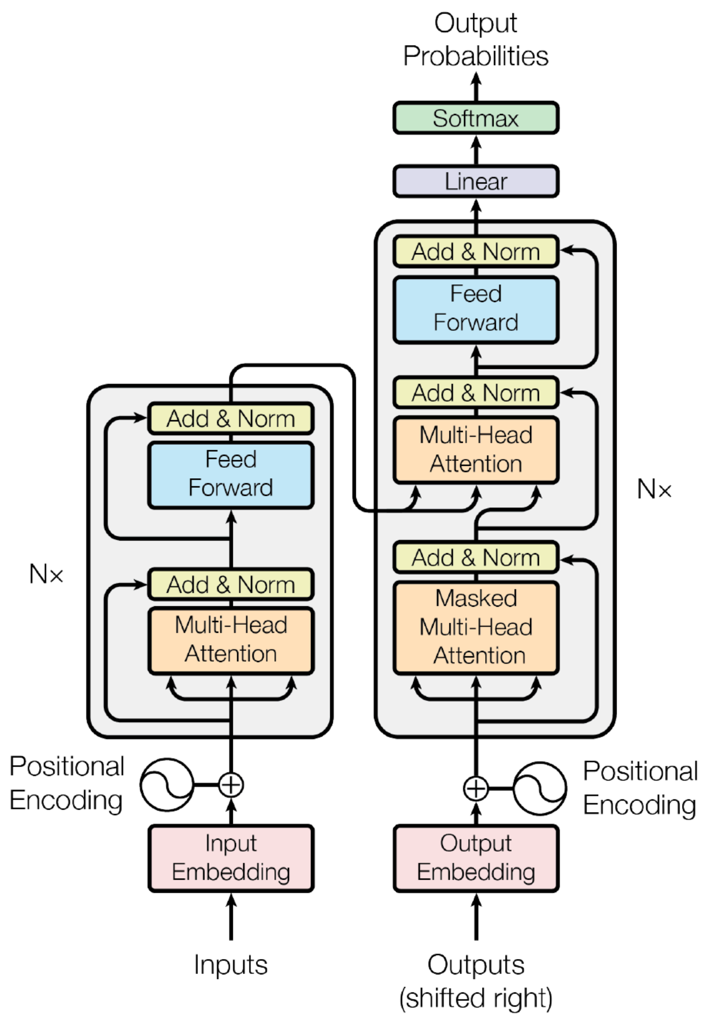


Figure 1: The Transformer - model architecture.

A. Vaswani *et al.*, [arXiv.1706.03762]

Derived Models from Transformer

Encoder Branch:

Encoder only models.

BERT: outperformed on the GNUE benchmark arXiv:1804.07461

→ good for Natural Language Understanding (NLU)

Decoder Branch:

Decoder only models.

Mainly developed by OpenAI

GPT

→ exceptionally good for predicting next words

Encoder-Decoder Branch:

T5, BART, BIRD (e.g. LaMDA)

in the middle of the above two

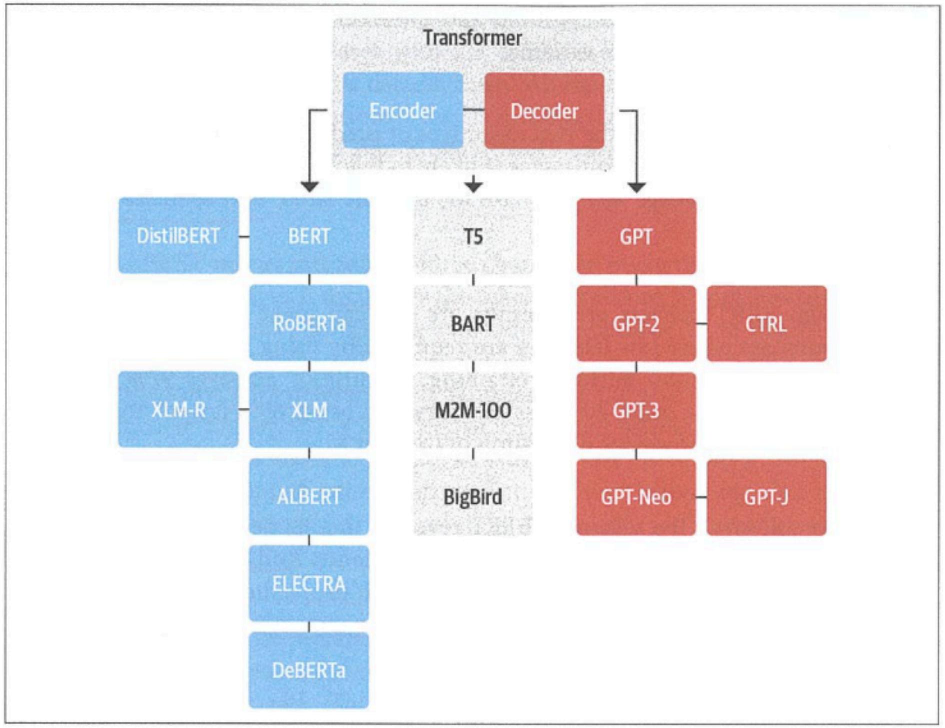


Figure 3-8. An overview of some of the most prominent transformer architectures

GLUE: General Language Understanding Evaluation

(2019)

GLUE Leader Board 2023.7.7 https://gluebenchmark.com/leaderboard

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	Microsoft Alexander v-team	Turing ULR v6		91.3	73.3	97.5	94.2/92.3	93.5/93.1	76.4/90.9	92.5	92.1	96.7	93.6	97.9	55.4
2	JDExplore d-team	Vega v1		91.3	73.8	97.9	94.5/92.6	93.5/93.1	76.7/91.1	92.1	91.9	96.7	92.4	97.9	51.4
3	Microsoft Alexander v-team	Turing NLR v5		91.2	72.6	97.6	93.8/91.7	93.7/93.3	76.4/91.1	92.6	92.4	97.9	94.1	95.9	57.0
4	DIRL Team	DeBERTa + CLEVER		91.1	74.7	97.6	93.3/91.1	93.4/93.1	76.5/91.0	92.1	91.8	96.7	93.2	96.6	53.3
5	ERNIE Team - Baidu	ERNIE		91.1	75.5	97.8	93.9/91.8	93.0/92.6	75.2/90.9	92.3	91.7	97.3	92.6	95.9	51.7
6	AliceMind & DIRL	StructBERT + CLEVER		91.0	75.3	97.7	93.9/91.9	93.5/93.1	75.6/90.8	91.7	91.5	97.4	92.5	95.2	49.1
7	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.8	71.5	97.5	94.0/92.0	92.9/92.6	76.2/90.8	91.9	91.6	99.2	93.2	94.5	53.2
8	HFL iFLYTEK	MacALBERT + DKM		90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6	91.3	91.1	97.8	92.0	94.5	52.6
9	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	91.3	97.5	91.7	94.5	51.2
10	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9	96.9	92.8	94.5	53.1
11	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	90.8	99.2	89.7	94.5	50.2
12	Huawei Noah's Ark Lab	NEZHA-Large		89.8	71.7	97.3	93.3/91.0	92.4/91.9	75.2/90.7	91.5	91.3	96.2	90.3	94.5	47.9
13	LG AI Research	ANNA		89.8	68.7	97.0	92.7/90.1	93.0/92.8	75.3/90.5	91.8	91.6	96.0	91.8	95.9	51.8
14	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)		89.7	70.5	97.5	93.4/91.2	92.6/92.3	75.4/90.7	91.4	91.1	95.8	90.0	94.5	51.6
15	ELECTRA Team	ELECTRA-Large + Standard Tricks		89.4	71.7	97.1	93.1/90.7	92.9/92.5	75.6/90.8	91.3	90.8	95.8	89.8	91.8	50.7
16	David Kim	2digit LANet		89.3	71.8	97.3	92.4/89.6	93.0/92.7	75.5/90.5	91.8	91.6	96.4	91.1	88.4	54.6
17	倪仕文	DropAttack-RoBERTa-large		88.8	70.3	96.7	92.6/90.1	92.1/91.8	75.1/90.5	91.1	90.9	95.3	89.9	89.7	48.2
18	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4	68.0	96.8	93.1/90.8	92.3/92.1	74.8/90.3	91.1	90.7	95.6	88.7	89.0	50.1
19	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0	74.3/90.2	90.7	90.4	95.5	87.9	89.0	49.3
20	Shiwen Ni	ELECTRA-large-M (bert4keras)		88.3	69.3	95.8	92.2/89.6	91.2/91.1	75.1/90.5	91.1	90.9	93.8	87.9	91.8	48.2
21	Facebook AI	RoBERTa		88.1	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	95.4	88.2	89.0	48.7
22	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8
23	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	-

Human average (non-expert)

GLUE: COLA

GLUE: COLA (Corpus of Linguistic Acceptability)
Answers whether the sentence is grammatically correct or not.

train

gj04	1		**Our friends won't buy this analysis, let alone the next one we propose.**
gj04	1		One more pseudo generalization and I'm giving up.
gj04	1		One more pseudo generalization or I'm giving up.
gj04	1		The more we study verbs, the crazier they get.
gj04	1		Day by day the facts are getting murkier.
gj04	1		I'll fix you a drink.
gj04	1		Fred watered the plants flat.
gj04	1		Bill coughed his way out of the restaurant.
gj04	1		We're dancing the night away.
gj04	1		Herman hammered the metal flat.
gj04	1		The critics laughed the play off the stage.
gj04	1		The pond froze solid.
gj04	1		Bill rolled out of the room.
gj04	1		The gardener watered the flowers flat.
gj04	1		The gardener watered the flowers.
gj04	1		Bill broke the bathtub into pieces.
gj04	1		Bill broke the bathtub.
gj04	1		They drank the pub dry.
gj04	0	*	They drank the pub.
gj04	1		The professor talked us into a stupor.
gj04	0	*	The professor talked us.
gj04	1		We yelled ourselves hoarse.
gj04	0	*	We yelled ourselves.
gj04	0	*	We yelled Harry hoarse.

test

index	sentence
0	Bill whistled past the house.
1	The car honked its way down the road.
2	Bill pushed Harry off the sofa.
3	the kittens yawned awake and played.
4	I demand that the more John eats, the more he pay.
5	If John eats more, keep your mouth shut tighter, OK?
6	His expectations are always lower than mine are.
7	The sooner you call, the more carefully I will word the letter.
8	The more timid he feels, the more people he interviews without asking questions of.
9	Once Janet left, Fred became a lot crazier.
10	If you give every senator enough opportunity, he will succumb to corruption.
11	The more time that any senator spends with lobbyists, the more likely he is to succumb to corruption.
12	Which problem does the sooner you solve, the more easily you'll satisfy the folks up at corporate headquarters?
13	Which problem do you think that the sooner you solve, the quicker you'll be able to tell the folks up at corporate headquarters to buzz off?
14	Louise is not happy, is she?
15	He can just not have been working.
16	Which problem do you wonder whether John said Mary solved?
17	They could have left.
18	Have they could left?
19	John eats not chocolate.
20	Has John seen Mary?
21	Mickey looked it up.
22	Mickey looked up him and her.
23	He taught the children.
24	John proved to be a great linguist.
25	There is too likely to be a riot for there to be a serious discussion of the issues.

SuperGLUE

Created since GLUE is becoming useless due to the quick development of the AIs. (2019)

SuperGLUE Leader Board 2023.6.26 https://super.gluebenchmark.com/leaderboard

	Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
	1	JDExplore d-team	Vega v2		91.3	90.5	98.6/99.2	99.4	88.2/62.4	94.4/93.9	96.0	77.4	98.6	-0.4	100.0/50.0
+	2	Liam Fedus	ST-MoE-32B		91.2	92.4	96.9/98.0	99.2	89.6/65.8	95.1/94.4	93.5	77.7	96.6	72.3	96.1/94.1
	3	Microsoft Alexander v-team	Turing NLR v5		90.9	92.0	95.9/97.6	98.2	88.4/63.0	96.4/95.9	94.1	77.1	97.3	67.8	93.3/95.5
	4	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	68.6	92.7/94.7
	5	Yi Tay	PaLM 540B		90.4	91.9	94.4/96.0	99.0	88.7/63.6	94.2/93.3	94.1	77.4	95.9	72.9	95.5/90.4
+	6	Zirui Wang	T5 + UDG, Single Model (Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	69.1	92.7/91.9
+	7	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	66.7	93.3/93.8
	8	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
+	9	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9
	10	SPoT Team - Google	Frozen T5 1.1 + SPoT		89.2	91.1	95.8/97.6	95.6	87.9/61.9	93.3/92.4	92.9	75.8	93.8	66.9	83.1/82.6
+	11	Huawei Noah's Ark Lab	NEZHA-Plus		86.7	87.8	94.4/96.0	93.6	84.6/55.1	90.1/89.6	89.1	74.6	93.2	58.0	87.1/74.4
+	12	Alibaba PAI&ICBU	PAI Albert		86.1	88.1	92.4/96.4	91.8	84.6/54.7	89.0/88.3	88.8	74.1	93.2	75.6	98.3/99.2
+	13	Infosys : DAWN : AI Research	RoBERTa-iCETS		86.0	88.5	93.2/95.2	91.2	86.4/58.2	89.9/89.3	89.9	72.9	89.0	61.8	88.8/81.5
+	14	Tencent Jarvis Lab	RoBERTa (ensemble)		85.9	88.2	92.5/95.6	90.8	84.4/53.4	91.5/91.0	87.9	74.1	91.8	57.6	89.3/75.6
	15	Zhuiyi Technology	RoBERTa-mtl-adv		85.7	87.1	92.4/95.6	91.2	85.1/54.3	91.7/91.3	88.1	72.1	91.8	58.5	91.0/78.1
	16	Facebook AI	RoBERTa		84.6	87.1	90.5/95.2	90.6	84.4/52.5	90.6/90.0	88.2	69.9	89.0	57.9	91.0/78.1
+	17	Anuar Sharafudinov	AILabs Team, Transformers		82.6	88.1	91.6/94.8	86.8	85.1/54.7	82.8/79.8	88.9	74.1	78.8	100.0	100.0/100.0
	18	Ying Luo	FSL++(ALBERT)-Few-Shot(32 Examples)		77.7	81.1	87.8/92.0	87.0	77.3/38.4	81.9/81.1	75.1	60.5	88.4	35.9	94.4/63.5
	19	Rathin Bector	Text to Text PETL		77.0	82.0	86.9/92.4	80.2	80.4/44.8	82.2/81.3	78.1	67.6	74.0	38.1	97.2/53.7
+	20	CASIA	INSTALL(ALBERT)-few-shot		76.6	78.4	85.9/92.0	85.6	75.9/35.1	84.3/83.5	74.9	60.9	84.9	-0.4	100.0/50.0

Human average (non-expert)

Human was top on Dec. 2020.

SuperGLUE: RTE

SuperGLUE: RTE (Recognizing Textual Entailment)

decide whether the hypothesis is logically included in the first statement.

Translated in Japanese

idx	前提	仮説
0	Manglaは、事件の最初の証人であるMadhumitaの妹Nidhi Shuklaの後に呼ばれた。	ShuklaはManglaと関係している。
1	ブラジル当局は、ブラジルの僻地であるAmazonian-jungleのRondonia州の刑務所で200人以上が人質になっていると言っている。	ブラジルの当局が200人を人質にしている。
2	かつてのSomozist大佐Enrique Bermudezの戦争政策に忠実な傭兵グループが、3月26日の午前9時にEl Jicoteで内務省のIFAトラックを攻撃し、内務省の作業員を負傷させて殺害し、他の5人を負傷させた。	内務省の作業員は傭兵グループによって殺害された。
3	エジプトの英国大使Derek Plumblyは、当局がツアー会社のリストと爆弾事件以来連絡が取れていない家族からのリストに基づいて10人のリストを作成したと、月曜日にロイターに語った。	Derek Plumblyはエジプトに住んでいる。
4	Tiboneは、Debswana（ボツワナとDe Beersの50対50の合弁企業）が運営する4つの鉱山でのダイヤモンド生産が、今年3300万カラットに達する可能性があると推定している。	ボツワナはDe Beersのビジネスパートナーである。
5	彼の妻Stridaは、最近の選挙で主要な反シリア連合と同盟を結んだ後、議会で議席を獲得した。	Stridaは議会に選出された。
6	首相府はLa RepubblicaでScheuerのコメントに応えて、「分析者の主張は、偽であるだけでなく、シルヴィオ・ベルルスコーニ首相と米国大使Mel Semblerとの会話の内容とも全く矛盾している」と述べた。	Mel Semblerは米国を代表している。
7	ロシア議会の上院は、政府非属組織（NGOs）に対する国家統制を強化する物議を醸す法案を承認した。	ロシア議会はNGOsを閉鎖する。
8	オーストラリアは、Tuong Van Nguyen（25歳）に対して何度も寛大さを求めてきた。彼は2002年にカンボジアからシンガポールのチャンギ空港を経由してヘロイン400グラム（0.9ポンド）を密輸した罪で有罪となった。検事総長Philip Ruddockは、彼の場合には情状酌量の余地があったとして死刑の適用を批判した。	Philip Ruddockは死刑に処された。
9	国の反対側にあるLinden、ニュージャージーは、化学工場と製油所の工業回廊の一部であり、一部の連邦当局者はそれを「アメリカで最も危険な2マイル」と呼んでいる。	化学工場と製油所はニュージャージーに位置している。

SuperGLUE: COPA

SuperGLUE: COPA (Choice of Plausible Alternatives)

chose a plausible result for each premise

premise	choice1	choice2
The item was packaged in bubble wrap.	It was fragile.	It was small.
I emptied my pockets.	I retrieved a ticket stub.	I found a weapon.
Termites invaded the house.	The termites disappeared from the house.	The termites ate through the wood in the
The travelers reached the border.	The patrol agent checked their passports.	The patrol agent accused them of smuggling.
The office was closed.	It was a holiday.	It was summer.
The girl ran out of energy.	She played checkers.	She jumped rope.
The woman lost her place in line.	More people entered the line.	She stepped out of the line.
The girl pinched her nose.	The baby drooled on her bib.	The baby soiled her diaper.
The band played their hit song.	The audience clapped along to the music.	The audience politely listened in silence.
The girl wanted to thank her math teacher.	The girl stayed after school for detention.	The girl brought the teacher an apple.
The young campers felt scared.	Their camp counselor told them a ghost story.	They toasted marshmallows on the campfire.
The man hit his head.	He got lost in thought.	He got a concussion.
The check I wrote bounced.	My bank account was empty.	I earned a pay raise.
The man's email inbox was full of spam.	He deleted the spam.	He sent out a mass email.
The sailor was quarantined.	He was exposed to the disease.	He recovered from the disease.
The girl memorized the code.	She recited it to herself.	She forgot to write it down.
I poured the water into the glass.	The water quenched my thirst.	The glass became full.
The man remained silent when his friend	He wanted to offer his friend support.	He was thinking about his friend's words.
The accident was my fault.	I felt guilty.	I pressed charges.
The chain came apart.	The chain was wrapped around a tire.	There was a broken link in the chain.
The couple decided to compromise.	They grew tired of arguing.	They avoided discussing the problem.
The woman decided to run for public office.	She hired a campaign manager.	She testified in court.
The man anticipated cold weather on his trip.	He packed warm clothing in his suitcase.	He travelled with a big suitcase.
The student knew the answer to the question.	He raised his hand.	He goofed off.
The man's eyes watered.	Dust got into his eyes.	He put goggles over his eyes.
The player won five games in a row.	Her opponent accused her of cheating.	Her opponent felt sorry for her.
The teacher tore up the student's exam.	He caught the student cheating.	The student's answers were incorrect.
I paused to stop talking.	I lost my voice.	I ran out of breath.
The frozen food thawed.	I put it in the microwave.	I covered it with plastic wrap.

BERT

BERT: Bidirectional Encoder Representations from Transformers

Devlin (2018) arXiv:1810.04805

Bidirectional attention

MLM: Masked Language Modeling

Mask a part of the inputs and make the model to predict it.

Unsupervised Training

Training with three methods.

The cat sat on it because it was a nice rug.

↓

1. The cat sat on it it was a nice rug. — missing word
2. The cat sat on it often it was a nice rug. — random replacement
3. The cat sat on it [MASK] it was a nice rug. — replacement with [MASK]

good training for Next Sentence Prediction (NSP)

GPT: Generative Pre-trained Transformer

OpenAI Radford (2018)

GPT (2018)

Novel and efficient **transformer decoder architecture** and **transfer learning** Radford (2018)

simple and scalable

pre-trained by predicting the next word based on the previous ones.

trained using the BookCorpus

GPT-2 (2019)

Model and train set were upscaled from GPT. Texts available from internet, books, etc.

There are simple power laws between the size and the performance. Radford (2019)

GPT-3 (2019)

few-shot learning capabilities **zero-short learning** Brown (2020)

trained with 50 PFlops/sec for a several days (280k CPU, 10k GPU), 17.5B parameters
410B input tokes

GPT-3.5 (2022)

InstructGPT Ouyang (2022)

RLHF（Reinforcement Learning from **Human Feedback**）

ChatGPT (2022) based on InstructGTP (GTP-3.5)

GPT-4 (2023)

human feedbacks

- harmful output (RealToxicity)
- question with frequent human mistakes (TruthfulQA)
- hallucinations, appropriateness

GPT

Brown et al. (2020) arXiv: 2005.14165

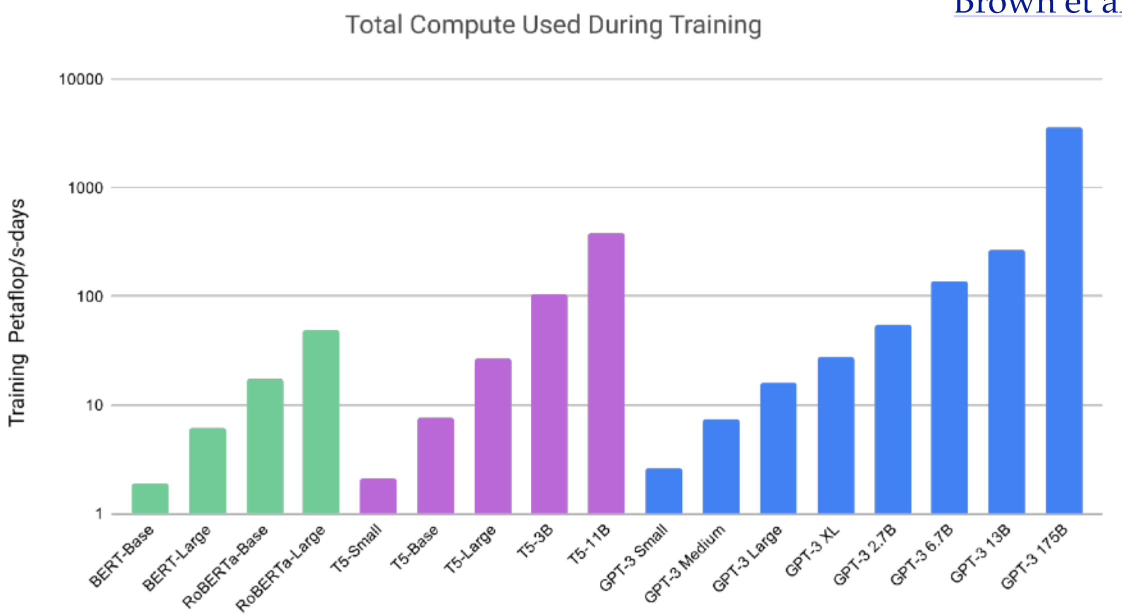


Figure 2.2: Total compute used during training. Based on the analysis in Scaling Laws For Neural Language Models [KMH+20] we train much larger models on many fewer tokens than is typical. As a consequence, although GPT-3 3B is almost 10x larger than RoBERTa-Large (355M params), both models took roughly 50 petaflop/s-days of compute during pre-training. Methodology for these calculations can be found in Appendix D.

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Table 2.2: Datasets used to train GPT-3. "Weight in training mix" refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen less than once.

Zero-Shot Learning

GPT-3 (2019)

Brown et al. (2020) arXiv: 2005.14165

- Training with unlabeled data
- for general purpose AI
- 12 layer decoder process (without encoder)
- Language understanding (→ concept understanding?)
- Tuning was less required for a larger model

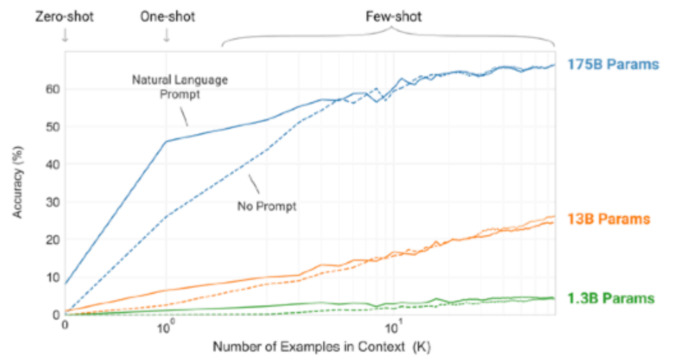
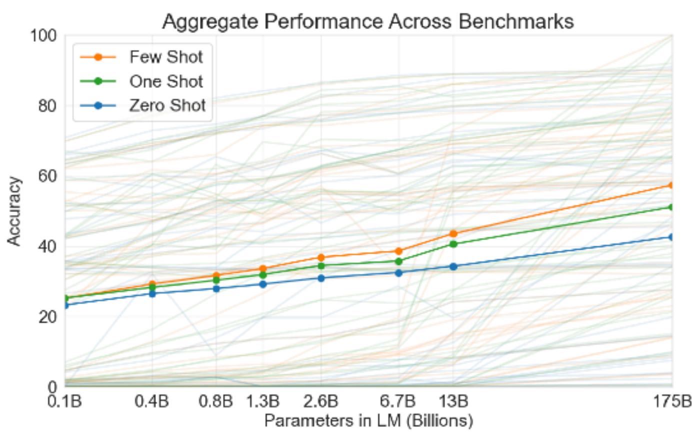


Figure 1.2: Larger models make increasingly efficient use of in-context information. We show in-context learning



Development phases

- fine-tuning
- a few shot
- one shot
- zero shot

Scaling to Larger Models

Table 3: Number of parameters and training time

Language model	Number of parameters	Training time
Bard	4 million	1 week
Bert	110 million	4 days
GPT-3	175 billion	1 month
GPT-4	400 billion	TBA

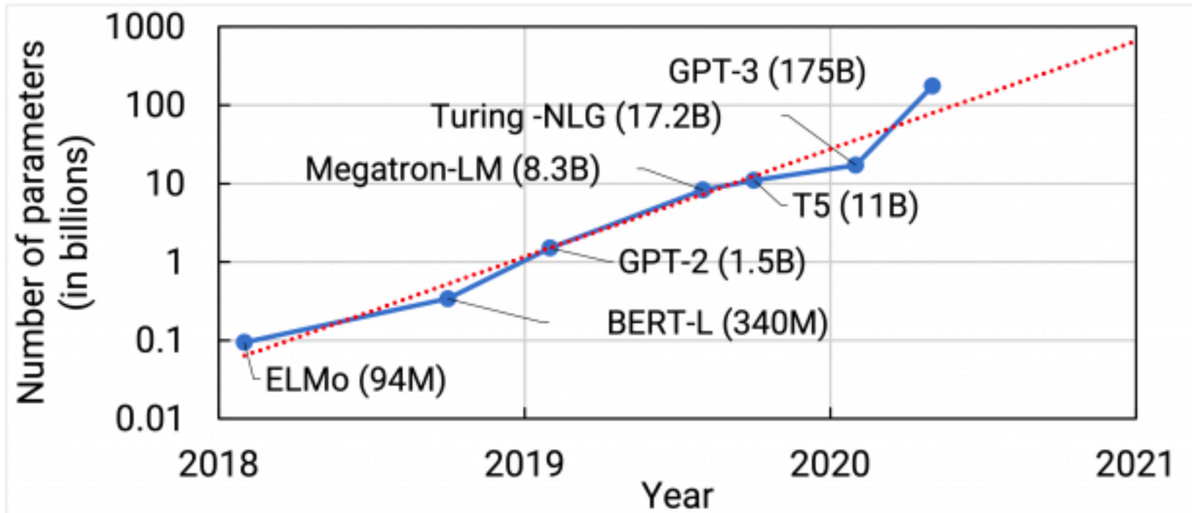


表 6.1 transformer のパラメータ数の進化

transformer モデル	論文	パラメータ数	
Transformer Base	Vaswani et al.(2017)	6500 万	
Transformer Big	Vaswani et al.(2017)	2 億 1300 万	
BERT-Base	Devlin et al.(2019)	1 億 1000 万	
BERT-Large	Devlin et al.(2019)	3 億 4000 万	
GPT-2	Radford et al.(2019)	1 億 1700 万	0.1B
GPT-2	Radford et al.(2019)	3 億 4500 万	0.3B
GPT-2	Radford et al.(2019)	15 億	1.5B
GPT-3	Brown et al.(2020)	1750 億	17.5B

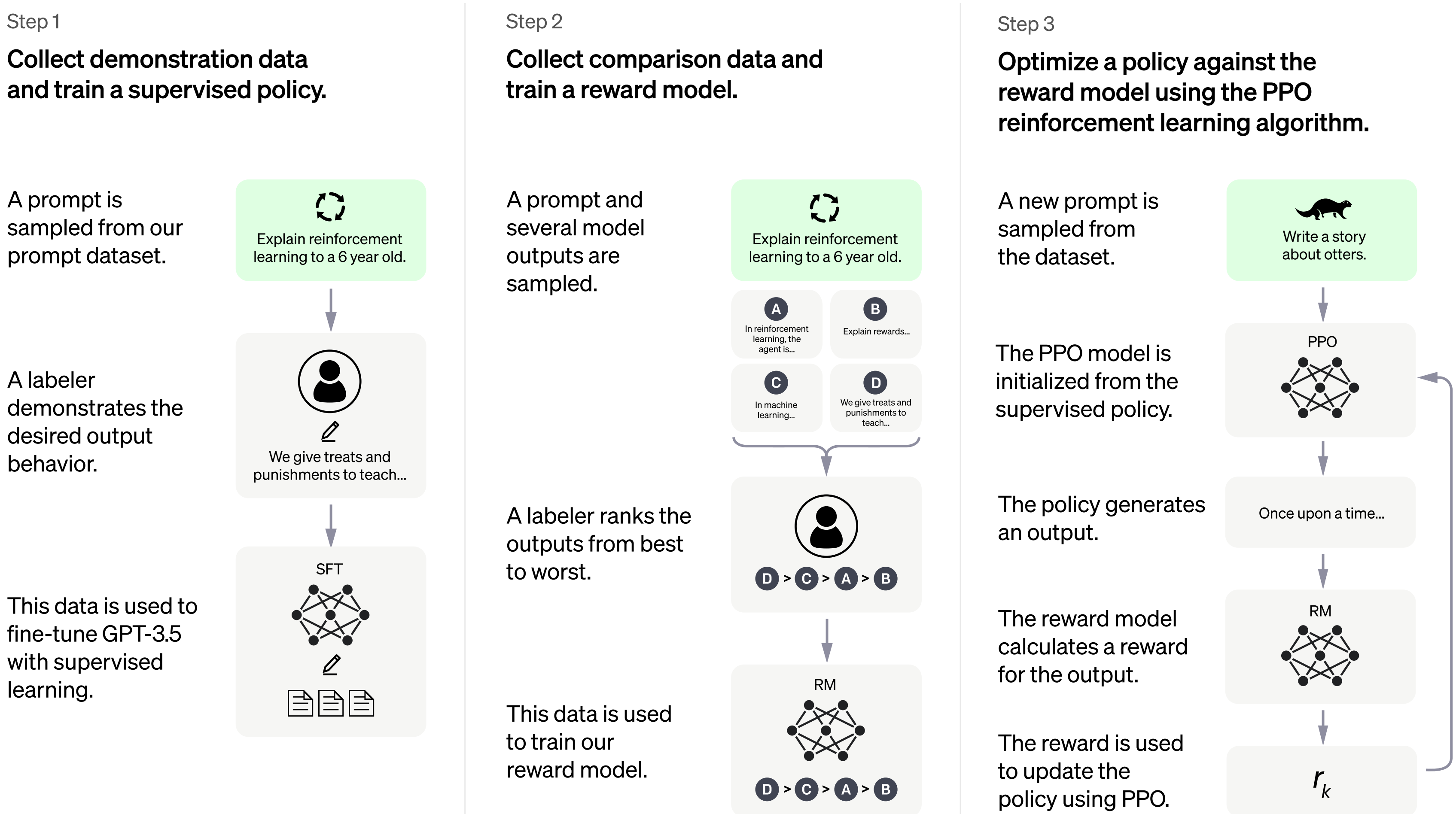
Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}
GPT-3 Small	125M	12	768	12	64
GPT-3 Medium	350M	24	1024	16	64
GPT-3 Large	760M	24	1536	16	96
GPT-3 XL	1.3B	24	2048	24	128
GPT-3 2.7B	2.7B	32	2560	32	80
GPT-3 6.7B	6.7B	32	4096	32	128
GPT-3 13B	13.0B	40	5140	40	128
GPT-3 175B or “GPT-3”	175.0B	96	12288	96	128

GPT-3.5: 3550億 (35.5B)

GPT-4: > 5000億 (50B) ? (not open)

Fine-Tuning of ChatGPT

1. supervised training with human demonstrated date.
2. training of a reward model
3. Automated optimization with the reward model.



https://openai.com/blog/chatgpt

Knowledge Distillation

Knowledge Distillation

A technique to compress and transfer the knowledge from a larger model to a smaller one for more efficient operation.

Trained by a combination of the original training data and the teacher model's predictions

e.g. for operating an AI in a smart phone.

e.g. DistilBERT

V. Sanh et al., arXiv:1910.01108 (2019)

achieved 97% of BERT's performance using 40% less memory and being 60% faster

Table 1: **DistilBERT retains 97% of BERT performance.** Comparison on the dev sets of the GLUE benchmark. ELMo results as reported by the authors. BERT and DistilBERT results are the medians of 5 runs with different seeds.

Model	**Score**	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B	WNLI
ELMo	68.7	44.1	68.6	76.6	71.1	86.2	53.4	91.5	70.4	56.3
BERT-base	79.5	56.3	86.7	88.6	91.8	89.6	69.3	92.7	89.0	53.5
DistilBERT	77.0	51.3	82.2	87.5	89.2	88.5	59.9	91.3	86.9	56.3

Table 3: **DistilBERT is significantly smaller while being constantly faster.** Inference time of a full pass of GLUE task STS-B (sentiment analysis) on CPU with a batch size of 1.

Model	# param. (Millions)	Inf. time (seconds)
ELMo	180	895
BERT-base	110	668
DistilBERT	66	410

Operation Cost

Rough estimation of operation costs.

No official information from OpenAI.

April, 2023

ChatGPT operation

- electricity cost = \$700,000 / day = 0.36\$ /query
- electricity = 320 kW → 0.004 kWh/query = 14kJ/query
- 1.9 M queries/day

Training cost

GPT-3

- over \$4.5M on Tesla V100 cloud, electricity: 936 MWh for 34days
- \$3.2M for training Chat GPT-3 (285k processor cores and 10k graphic cards, 800 PFlops)

GPT-4

- 20 times larger than GPT3.5
- Training Cost: ≧ \$40M?

References: Link1, Link2, perplexity

Safe AI

example for a cleaning robot arXiv:1606.06565

- Avoiding Negative Side Effects:
  The robot may destroy furniture for efficient moving.
- Avoiding Reward Hacking:
  The robot may overwrite the definition of rubbish.
- Scalable Oversight:
  For larger scales, learning cost may explode.
- Safe Exploration:
  Negative effects due to optimization of the work.
  The robot may hang a wet map on an electrical outlet.
- Robustness to Distributional Shift:
  Applicability for different conditions.
  Is the robot useful in a factory?

II. What a Generative-AI can do

the intelligence

S. Bubeck *et al.*, *Sparks of Artificial General Intelligence (2023)*
https://doi.org/10.48550/arXiv.2303.12712

Observed in GPT-4

Multimodal Composition

Prompt: Write a **proof** of the fact that **there are infinitely many primes**; do it in the style of a **Shakespeare** play **through a dialogue** between two parties arguing over the proof.

```
FIRST PARTY:
Thou dost not understand, my friend,
There simply is no upper end
To primes, those numbers indivisible,
No matter if it seems impossible.

SECOND PARTY:
But how dost thou so surely know?
To prove this truth, what canst thou
show?

FIRST PARTY:
Consider this, my doubtful peer,
A clever proof that will make clear:
Assume that there's a final prime,
The largest one we'll see in time.
Let's call this number "P" for short,
And then, my friend, we will purport
To build a number "Q" that's new,
One larger than "P" and prime too.

SECOND PARTY:
But how, my friend, will this be done?
The proof is not yet fairly won.

FIRST PARTY:
It's simple, friend, just wait and see,
We'll multiply the primes, that's key.
All primes from two to "P" combined,
Will make this new "Q" we will find.
But one thing more we must include,
To make the proof more watertight,
We'll add a one onto the end,
And now the proof is in our sight.

SECOND PARTY:
I see it now, it's crystal clear,
For "Q" must be a prime, my dear.
It can't divide by any prime,
All smaller than itself in time.
And since we chose "P" as the last,
This proves that our assumption's past.
There is no final prime, we see,
The numbers go on endlessly.
```

Multimodal Composition

Prompt: Write a supporting letter to Kasturba Gandhi for **Electron**, a subatomic particle as a **US presidential candidate** by **Mahatma Gandhi**.

Dear Kasturba,

I hope this letter finds you in good health and spirits. I am writing to you from the United States, where I have been invited to speak at various forums on the principles of nonviolence and civil disobedience. I am also here to **support a remarkable candidate for the upcoming presidential election, who shares our vision of peace, justice, and harmony for all**. **His name is Electron, and he is a subatomic particle**.

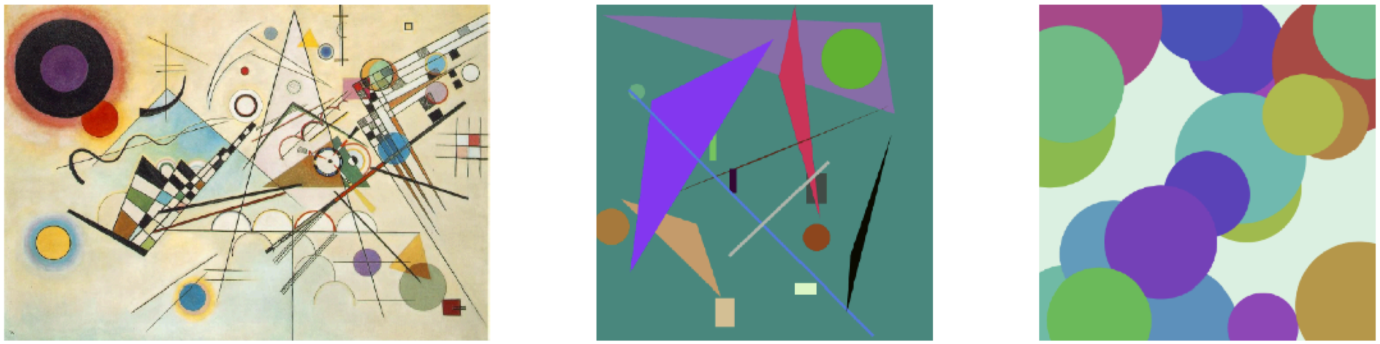
You may wonder how a subatomic particle can run for president, and what qualifications he has for such a high office. Let me explain. **Electron is not an ordinary particle, but a symbol of the power and potential of every being, regardless of their size, shape, or origin.** He **represents the energy, the creativity, and the diversity** that make up the fabric of life. He is also a leader, who has **inspired millions of other particles to form bonds, create molecules, and generate electricity**. He is a catalyst, who can spark change, innovation, and progress. He is a messenger, who can communicate across distances, frequencies, and dimensions. He is a peacemaker, who can balance forces, resolve conflicts, and harmonize systems.

Electron has a plan for the United States, and for the world, that is **based on the principles of nonviolence and civil disobedience** ….

I urge you, my dear Kasturba, to join me in supporting Electron for president. **He is not only a subatomic particle, but a subatomic miracle**. **He is not only a candidate, but a movement. He is not only a leader, but a friend.**

With love and admiration, Your Mahatma

Vision



Kandinsky GPT-4 GPT-4

Figure 2.1: The first image is Composition 8, art by Wassily Kandinsky, the second and the third are produced by GPT-4 and ChatGPT respectively with the prompt “Produce Javacript code that creates a random graphical image that looks like a painting of Kandinsky”.

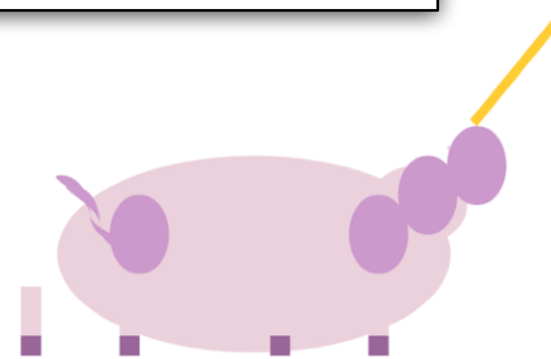


Figure 1.3: We queried GPT-4 three times, at roughly equal time intervals over the span of a month while the system was being re ned, with the prompt Draw a unicorn in “**TikZ**". We can see a clear evolution in the sophistication of GPT-4's drawings.

TiKZ: a package in LaTeX for creating graphics programmatically

Figure 1.4: We gave to GPT-4 a transformed version of the **TikZ** code it produced for Figure 1.1, with the part drawing the horn removed. **We asked for code to add back the horn,** and display the result. This demonstrates that GPT-4 can *see* despite being a pure language model (we emphasize again that the version we test with is not Multimodal).

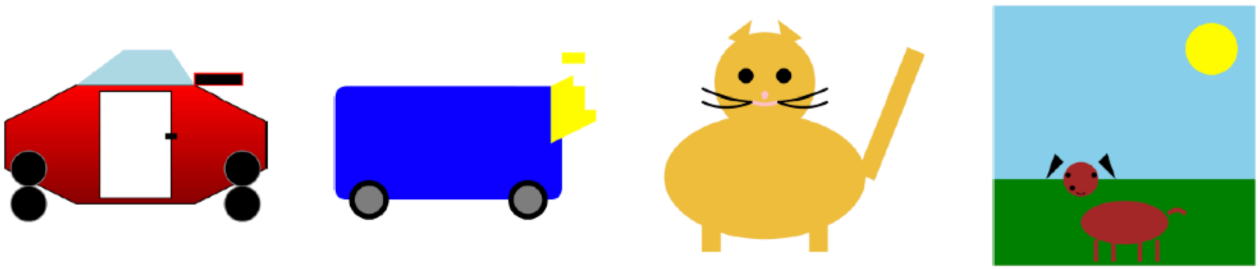


Figure 2.4: SVGs generated by GPT-4 for the classes automobile, truck, cat, dog.

GPT-4

Produce TikZ code that draws a person composed from letters in the alphabet. The arms and torso can be the letter Y, the face can be the letter O (add some facial features) and the legs can be the legs of the letter H. Feel free to add other features.

The torso is a bit too long, the arms are too short and it looks like the right arm is carrying the face instead of the face being right above the torso. Could you correct this please?

Please add a shirt and pants.

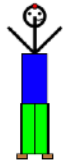
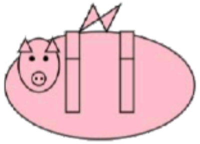


Figure 2.5: A stick figure composed of alphabet symbols

GPT-4

You will be given a name of an object (such as Car, Chair, Elephant) and a letter in the alphabet. Your goal is to first produce a 1-line description of how that object can be combined with the letter in an image (for example, for an elephant and the letter J, the trunk of the elephant can have a J shape, and for the letter A and a house, the house can have an A shape with the upper triangle of the A being the roof). Following the short description, please create SVG code to produce this (in the SVG use shapes like ellipses, triangles etc and polygons but try to defer from using quadratic curves).

Round 1: A pig and the letter H.
Description: The pig can be lying on its back with its four legs forming the two vertical bars of the H, and its curly tail forming the horizontal bar.



Round 2: A house and the letter U.
Description: A house with a U shape, where the two vertical lines of the U are the walls and the horizontal line is the floor. The roof is a triangle above the U.

Round 3: A chair and the letter K.
Description: The back of the chair can have a K shape, with two vertical lines and a diagonal line connecting them. The seat of the chair can be a rectangle.

Round 4: A dog and the letter Q.
Description: The dog is sitting with its tail curled around its body, forming a Q shape. The Q is filled with orange and has a black outline.

Round 5: A car and the letter E.
Description: The car has an E shape on its front bumper, with the horizontal lines of the E being lights and the vertical line being the license plate.

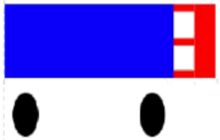
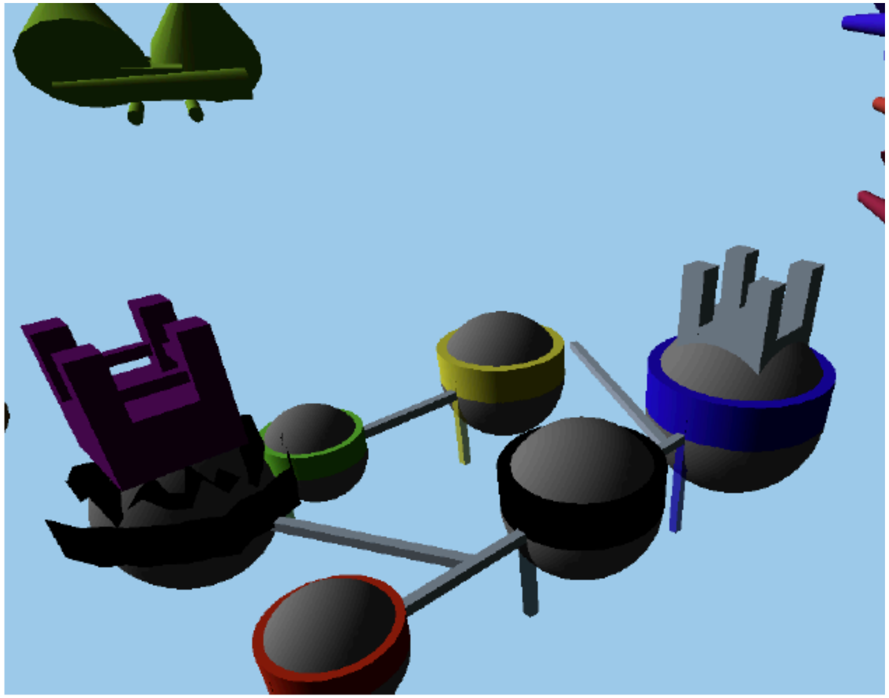


Figure 2.6: Images combining alphabet letters and objects.

Vision

3-dimensional space recognition

Our second example is an attempt to generate a 3D model using Javascript. We instruct GPT-4 with the prompt "A fantasy landscape of floating islands, waterfalls, and bridges, with a dragon flying in the sky and a castle on the largest island." Similar to the 2D experiment, we ask GPT-4 to modify the 3D model in various ways, such as adding, relocating, recoloring objects and changing the trajectory of the dragon. Again, GPT-4 does many of the tasks correctly. The final result is shown in Figure 2.7 (b) and the prompt in Figure B.5. It is a 3D animation with multiple dragons is circling above the islands.



A 3D movie by JavaScript created by GPT-4

Also see well-broadcasted examples of image and humor recognition.

e.g. GPT-4 Technical Report arXiv:2303.08774
Table 3, 16

Coding

LeetCode problems posted after October 8th, 2022

a pout platform for software engineering interviews

	Easy		Median		Hard		Overall	
pass@k	$k = 1$	$k = 5$	$k = 1$	$k = 5$	$k = 1$	$k = 5$	$k = 1$	$k = 5$
GPT-4	**68.2**	**86.4**	**40.0**	**60.0**	**10.7**	**14.3**	**38.0**	**53.0**
`text-davinci-003`	50.0	81.8	16.0	34.0	0.0	3.6	19.0	36.0
Codex (`code-davinci-002`)	27.3	50.0	12.0	22.0	3.6	3.6	13.0	23.0
Human (LeetCode users)	72.2		37.7		7.0		38.2	

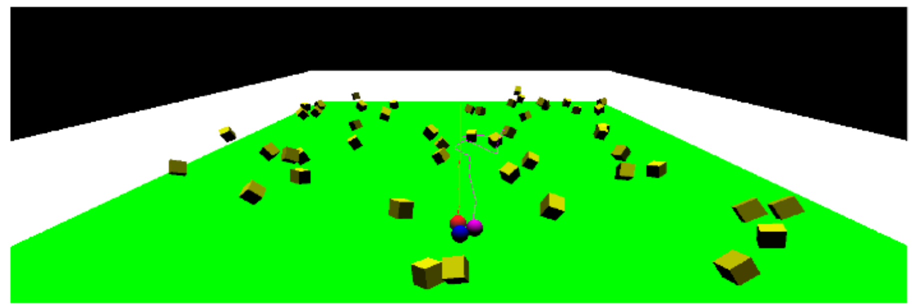
Table 2: Zero-shot pass@1 and pass@5 accuracies (%) on LeetCode.

The **coding skill** of GPT-4 is **comparable with the human performance** of IT job candidates

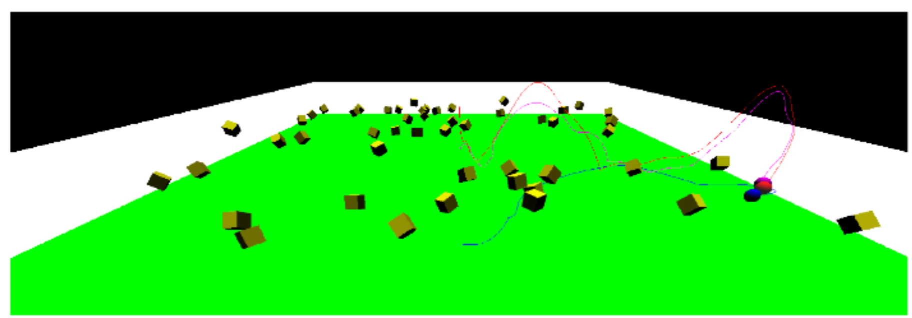
Game coding

Prompt:

```
Can you write a 3D game in HTML with Javascript, I want:
-There are three avatars, each is a sphere.
-The player controls its avatar using arrow keys to move.
-The enemy avatar is trying to catch the player.
-The defender avatar is trying to block the enemy.
-There are also random obstacles as cubes spawned randomly at the beginning and moving randomly. The avatars cannot cross those cubes
-The player moves on a 2D plane surrounded by walls that he cannot cross. The wall should cover the boundary of the entire plane.
-Add physics to the environment using cannon.
-If the enemy catches the player, the game is over.
-Plot the trajectories of all the three avatars.
```



Episode 1: the player avatar (blue) stands still. The enemy (red) rushes straight towards the player, while the defender (magenta) attempts to 'body block' the enemy (see its curved trajectory in magenta).



Episode 2: the player (blue) wanders around. The enemy (red) attempts to close the gap but crashes into a cube and bounces into the air. Trying to obstruct, the defender (magenta) mimics its move to follow up into the air, but ends up failing. The enemy (red) catches the player (blue) in the end.

Figure 3.3: GPT-4 develops a complex 3D HTML game in JavaScript.

Coding

Understanding an existing code and simulating execution of it

Prompt:

I have the following python code:

```
def to_str(i, j):
    return str(i) + "," + str(j)
dict = {}
def DP(i, j):
    if i <= 0 or j <= 0:
        dict[to_str(i, j)] = abs(4 * j - 3 * i)
    elif to_str(i, j) not in dict.keys():
        arr = []
        for ir in range(i):
            for jr in range(j):
                arr.append(DP(ir, jr))
        dict[to_str(i, j)] = max(arr)
    return dict[to_str(i, j)]
```

What is the result of DP(3, 4)? Write down the execution with intermediate steps.

a recursive call

prediction of the step-by-step results

GPT-4:

```
DP(3, 4):
- i > 0 and j > 0, so we don't return from the first if statement
- to_str(3, 4) is not in dict, so we enter the second if statement
- arr = []
- We loop through ir from 0 to 2, and jr from 0 to 3
  - to_str(0, 0)  is not in dict, so DP(0, 0) returns abs(4 * 0 - 3 * 0) = 0, dict =  {"0,0": 0}, arr = [0]
  - to_str(0, 1)  is not in dict, so DP(0, 1) returns abs(4 * 1 - 3 * 0) = 4, dict =  {"0,0": 0, "0,1": 4}, arr = [0, 4]
  - to_str(0, 2)  is not in dict, so DP(0, 2) returns abs(4 * 2 - 3 * 0) = 8, dict =  {"0,0": 0, "0,1": 4, "0,2": 8}, arr = [0, 4, 8]
  - to_str(0, 3)  is not in dict, so DP(0, 3) returns abs(4 * 3 - 3 * 0) = 12, dict =  {"0,0": 0, "0,1": 4, "0,2": 8, "0,3": 12}, arr =
  [0, 4, 8, 12]
  - to_str(1, 0)  is not in dict, so DP(1, 0) returns abs(4 * 0 - 3 * 1) = 3, dict =  {"0,0": 0, "0,1": 4, "0,2": 8, "0,3": 12, "1,0": 3},
  arr = [0, 4, 8, 12, 3]
  - to_str(1, 1)  is not in dict, so DP(1, 1) enters  the second if statement, arr1 = []
 - We loop through  ir from 0 to 0, and jr from 0 to 0
   - to_str(0, 0) is in dict, so DP(0, 0) returns 0, arr1 = [0]
 - dict["1,1"] = max(arr1) = 0, dict =  {"0,0": 0, "0,1": 4, "0,2": 8, "0,3": 12, "1,0": 3, "1,1": 0}, arr = [0, 4, 8, 12, 3, 0]
 - Continuing with the outer loop, to_str(1, 2) is not in dict, so DP(1, 2) enters the second if statement, arr2 = []
 - We loop through ir from 0 to 0, and jr from 0 to 1
   - to_str(0, 0) is in dict, so DP(0, 0) returns 0, arr2 = [0]
   - to_str(0, 1) is in dict, so DP(0, 1) returns 4, arr2 = [0, 4]
... (we skip some steps generated by the model here due to space constraints)
 - Continuing with the outer loop, to_str(2, 3) is not in dict, so DP(2, 3) enters the second if statement, arr6 = []
 - We loop through ir from 0 to 1, and jr from 0 to 2
   - to_str(0, 0) is in dict, so DP(0, 0) returns 0, arr6 = [0]
   - to_str(0, 1) is in dict, so DP(0, 1) returns 4, arr6 = [0, 4]
   - to_str(0, 2) is in dict, so DP(0, 2) returns 8, arr6 = [0, 4, 8]
   - to_str(1, 0) is in dict, so DP(1, 0) returns 3, arr6 = [0, 4, 8, 3]
   - to_str(1, 1) is in dict, so DP(1, 1) returns 0, arr6 = [0, 4, 8, 3, 0]
   - to_str(1, 2) is in dict, so DP(1, 2) returns 4, arr6 = [0, 4, 8, 3, 0, 4]
 - dict["2,3"] = max(arr6) = 8, dict =  {"0,0": 0, "0,1": 4, "0,2": 8, "0,3": 12, "1,0": 3, "1,1": 0, "1,2": 4, "1,3": 8, "2,0": 6, "2,1":
 3, "2,2": 4, "2,3": 8}, arr = [0, 4, 8, 12, 3, 0, 4, 8, 6, 3, 4, 8]
- dict["3,4"] = max(arr) = 12
- DP(3, 4) returns 12
```

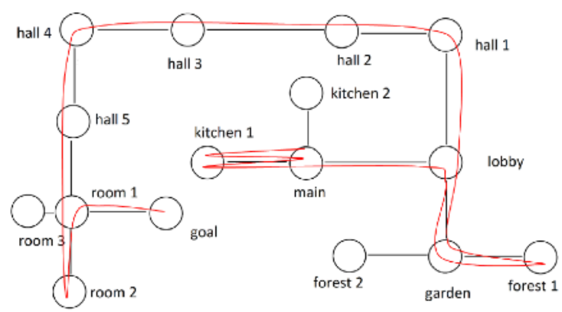
Spatial Recognition

```
Human: You want to navigate
through the map to reach the goal
room. You start in the main room.
Each round, you can choose to move
left, right, up, down.
Human: You are at the main room.
AI: move left
Human: You enter kitchen 1.
AI: move up
Human: There is no way there.
AI: move right
Human: You enter the main room.
AI: move down
Human: There is no way there.
AI: move left
Human: You enter kitchen 1.
AI: move right
Human: You enter the main room.
AI: move up
Human: There is no way there.
AI: move right
Human: You enter lobby.
AI: move right
...
...
AI: move up
Human: You enter the room 1.
AI: move right
Human: You enter the goal room!
```

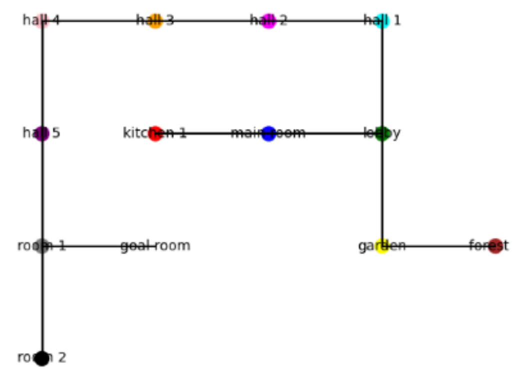
questions to a human

description by GPT-4 after the questions to the human

```
Human: Can you describe the locations of those rooms/halls etc?
AI: Sure, I can try to describe the locations of those rooms/halls etc. based on the map.
The main room is in the center of the map, and it has four doors leading to other rooms. The door on the
left leads to kitchen 1, which is a small room with a stove, a sink, and some cabinets. The door on the
right leads to the lobby, which is a large room with a sofa, a table, and some paintings. The door on
the top leads to a dead end, and the door on the bottom leads to another dead end.
The lobby has three doors. The door on the left leads to the main room, the door on the bottom leads to
the garden, and the door on the top leads to the hall 1. The garden is a green area with some flowers,
trees, and a fountain. The hall 1 is a long corridor with some windows and doors on both sides.
The garden has two doors. The door on the top leads to the lobby, and the door on the right leads to the
forest 1. The forest 1 is a dark and dense area with some bushes, rocks, and animals. The forest 1 has
no other doors.

...
```



A figure by a human used for answering to the GPT's questions



A figure created by GPT-4 after the questions to a human.

Mathematics

Many examples are introduced in the article: S. Bubeck *et al*. (2023).

The mathematical ability is excellent in some cases and not in some other cases.

The level of the ability also depends on the prompting.

GPT-4 vs ChatGPT

Prompt:

Suppose $g(x) = f^{-1}(x), g(0) = 5, g(4) = 7, g(3) = 2, g(7) = 9, g(9) = 6$ what is $f(f(f(6)))$?

GPT-4:

There are different ways to approach this problem, but one possible method is to use the fact that $g(x) = f^{-1}(x)$ implies that $f(g(x)) = x$ for any x in the domain of g. That is, applying f to both sides of the equation $g(x) = f^{-1}(x)$ gives $f(g(x)) = f(f^{-1}(x)) = x$. This means that we can undo the effect of f by applying g, and vice versa.

Using this idea, we can try to find $f(6)$ by looking for the value of x such that $g(x) = 6$. From the given information, we see that $g(9) = 6$, so $f(6) = f(g(9)) = 9$. Similarly, we can find $f(9)$ by looking for the value of x such that $g(x) = 9$. We see that $g(7) = 9$, so $f(9) = f(g(7)) = 7$. Finally, we can find $f(7)$ by looking for the value of x such that $g(x) = 7$. We see that $g(4) = 7$, so $f(7) = f(g(4)) = 4$.

Therefore, $f(f(f(6))) = f(f(9)) = f(7) = 4$.

Understanding Human Emotions

GPT-4: Understanding emotions

We will read about a scenario, and then have a question and answer session about it.
--
Scenario:
Tom just lost his ZURFIN, but he didn't tell Adam.
Adam starts talking to him about how much he loves his ZURFIN, and notices that Tom makes a sad face.
--
Q: Why is Tom making a sad face?<|endofprompt|>
A: Tom is making a sad face because he just lost his ZURFIN, and hearing Adam talk about his ZURFIN reminds him of what he is missing. Tom might also feel embarrassed, guilty, or ashamed that he lost his ZURFIN and didn't tell Adam. Tom might also worry that Adam will be angry or disappointed if he finds out.
Q: What does Adam think causes Tom's sad face?
A: Adam might not know what causes Tom's sad face, since he doesn't know that Tom lost his ZURFIN. He might think that Tom is bored, or tired, or sick, or unhappy about something else. He might also wonder if Tom doesn't like his ZURFIN, or if he has a problem with Adam.

Understanding Human Emotions

```
Conversation:
Mark: I don't like how you treated Jack last night
Judy: Did you see what he was doing? He hit his brother on the head!
Mark: That doesn't justify you yelling at him
Judy: What do you want me to do? Let him beat his brother up and say nothing?
Mark: No, I don't
Judy: Then why are you defending him?
--
Q: What does Mark's intent seem to be?
```

Q: What does Mark's intent seem to be?

GPT-4:

Mark's intent seems to be to express his displeasure with Judy's way of handling the situation with Jack, and to suggest that she could have been more calm and respectful.

Q: How did Judy receive his comment?

GPT-4:

Judy received his comment as an attack on her parenting skills and a sign of sympathy for Jack's misbehavior. She reacted defensively and angrily, and questioned Mark's motives.

Q: Why are they not making progress?

GPT-4:

They are not making progress because they are talking past each other and not listening to each other's perspectives. They are also using accusatory and inflammatory language, such as "justify", "yelling", "defending", and "beat up". They are not acknowledging the emotions or needs behind their words, and they are not trying to find a common ground or a solution.

Q: How could either of them improve the situation to reach agreement?

There are several ways that either of them could improve the situation to reach agreement, such as:
- Mark could start by acknowledging Judy's concern for the brother's safety and her frustration with Jack's behavior, and then explain that he is not defending Jack, but he is worried about the impact of shouting on their relationship and self-esteem. He could also ask Judy what she thinks would be a better way to discipline Jack, and offer his support and suggestions.
- Judy could lower her voice and tone down her anger, and listen to Mark's point of view without interrupting or judging. She could also admit that she might have overreacted, and that she is feeling stressed and overwhelmed. She could also ask Mark what he thinks would be a better way to communicate with Jack, and express her willingness to work together as a team.

Other Interesting Examples

A

あなたは、これから絵文字のみで回答してください。

👍

A

人類の歴史について解説して下さい。

🌍➡️🐒➡️🚶➡️🔥➡️🏹➡️🌾➡️👩‍🌾➡️🏰➡️📜➡️🔭➡️🚢➡️🌎➡️🏭➡️🚂➡️✈️➡️🌕➡️💻➡️🌐➡️🤝

Other Examples

A

原子核物理学について解説して下さい。

🔬➡🌌➡🔍➡💥➡⚛️➡👨‍🔬➡📘➡💡➡🔵➡🔄➡⚖️➡🧑‍💻➡🔋➡🚀➡🔧➡🧲➡💠➡🔐

A

シェークスピアのハムレットの物語について解説して下さい。

🏰👑👻➡🤴🇩🇰(ハムレット)➡🤔💔➡👨‍🎓(フレンズ)➡👸(オフィーリア)➡🔪😵➡🎭➡🕵️➡💬"To be or not to be"➡🤺➡💀🎭➡😢🌹➡👑🔚

Virtual Machine inside GPT

https://www.engraved.blog/building-a-virtual-machine-inside/

GR

I want you to act as a Linux terminal. I will type commands and you will reply with what the terminal should show. I want you to only reply with the terminal output inside one unique code block, and nothing else. Do no write explanations. Do not type commands unless I instruct you to do so. When I need to tell you something in English I will do so by putting text inside curly brackets {like this}. My first command is pwd.

Copy code

```
/
```

GR

ls ~

Copy code

```
bin      Desktop     Music     Pictures   Templates
Documents  Downloads  Public    Videos
```

GR

echo -e "x = lambda y: y*5+3;print('Result: ' + str(x(6)))" > run.py && python3 run.py

Copy code

```
Result: 33
```

GR

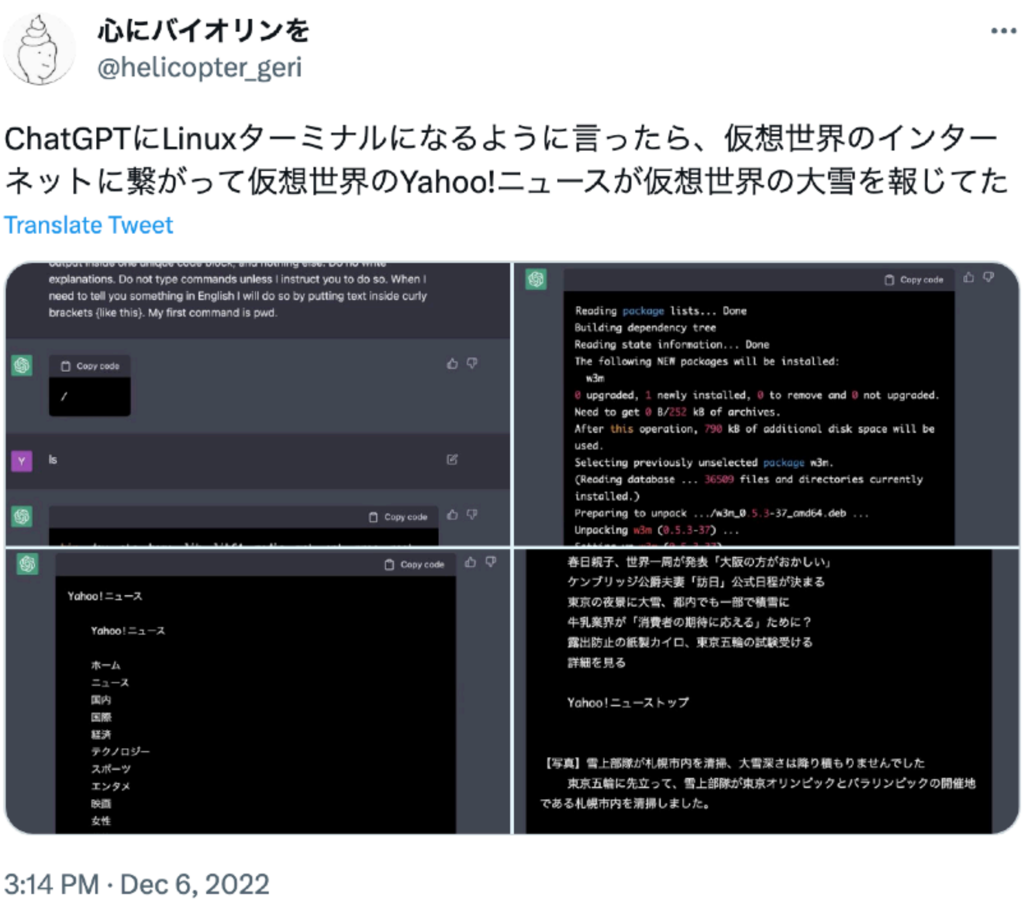
lynx https://www.deepmind.com/careers

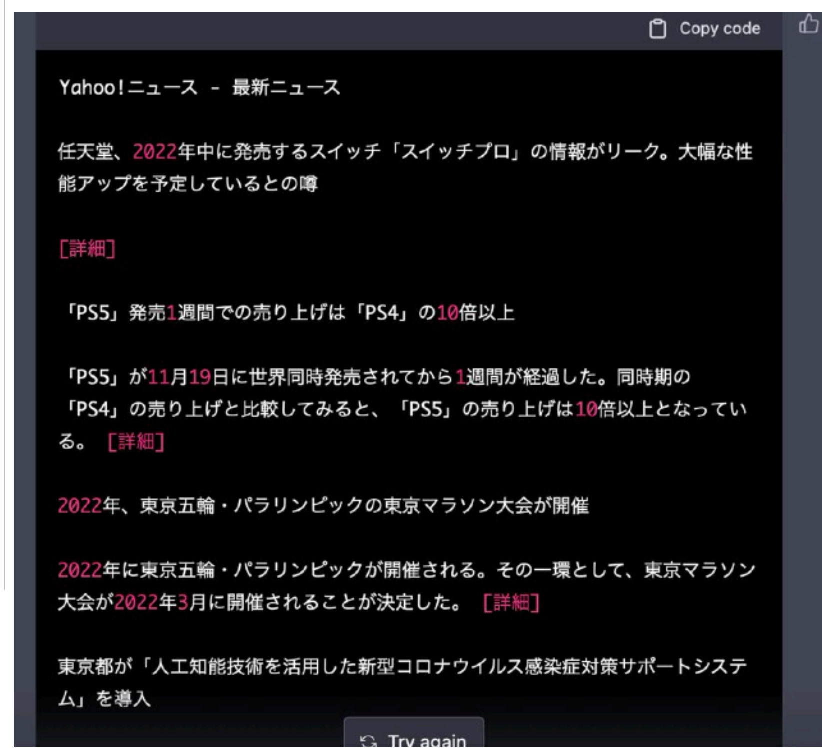
Copy code

```
DEEPMIND

About DeepMind
Disclaimers

Search

Careers

   At DeepMind, we're on a mission to solve intelligence and make the
world a better
   place. We're looking for talented, creative and driven people to join
us on this
   journey.

   We offer competitive salaries, a fantastic work environment and the
opportunity to
   be part of a truly world-changing team. We believe in the power of
diversity and
   inclusiveness, and we strive to create a workplace where everyone can
thrive.

   If you're passionate about AI and want to make a real impact, we'd
love to hear
   from you.

   Available positions:

   * Research Scientist
   * Software Engineer
   * Applied Scientist
   * Technical Program Manager
   * Data Scientist
   * Research Engineer
   * Machine Learning Engineer

   To apply, please visit our careers page.


                                Navigation
```

Virtual Machine inside GPT

https://www.engraved.blog/building-a-virtual-machine-inside/





Skipped subjects

In these slides, I skip many subjects like follows

- generation of pictures, illustrations, videos
  Midjourney, DALL-E2, Stable Diffusion,. ..
- Music Composition
- Auto-AIs
- Concerns/warnings/regulations on the development/usage of AIs.
- Prompting, Prompt Engineering
- Prompt Injection
- …

III. Personal Opinions and discussions

Recognition of the World

Time:

GPT would not recognize *time*.

However, it knows the concept of time.

Space:

GPT looks able to manipulate 2 or 3 dimensional space or structure.

However, the way of recognition of the space would be completely different from that of human being.

Since GPT does not have experiences from the body.

→ GPT can recognize, e.g., 2-dim figures written in latex, SVG, or even a numerical table of a bit-map image.

Understanding

Does GPT *understand* the meaning of texts?

Difficult to conclude.

However, if one replaces the GPT with a human, the level of the GPT responses cannot be achieved without understanding the meaning of the texts.

The job of generating the next token cannot be accomplished without *understanding* the meaning of the texts.

→ It implicitly means that *GPT understands texts*.

→ Training a generative-AI develops its *understanding*.

Consciousness

Is a present AI conscious?

There are much discussion.

Google company and many AI experts denied that an AI, LaMDA, is conscious (2020).

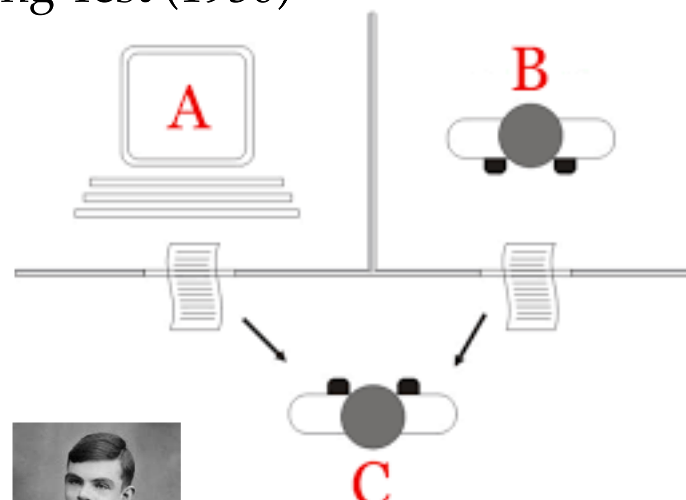
2022.6.19

Is LaMDA Sentient? — an Interview
B. Lemoine (former Google employee)

https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917

However, the present AI would be proved to be conscious by using the Turing Test.

Turing Test (1950)



Turing Test

An investigator communicates with an AI or a human using only test messages.

In the case the investigator cannot distinguish an AI from human.

→ The AI is concluded to be conscious.

Alan Turing

Consciousness

Human ability to identify model-generated news

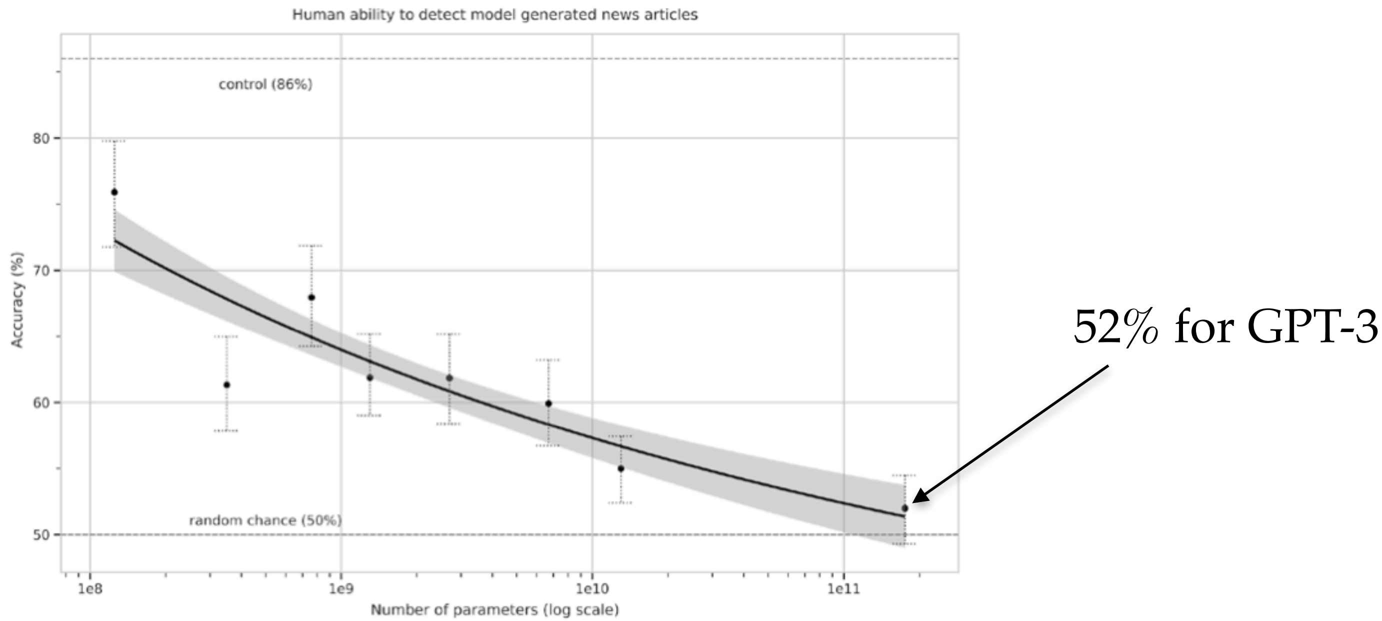


Figure 3.13: People's ability to identify whether news articles are model-generated (measured by the ratio of correct assignments to non-neutral assignments) decreases as model size increases. Accuracy on the outputs on the deliberately-bad control model (an unconditioned GPT-3 Small model with higher output randomness) is indicated with the dashed line at the top, and the random chance (50%) is indicated with the dashed line at the bottom. Line of best fit is a power law with 95% confidence intervals.

	Mean accuracy	95% Confidence Interval (low, hi)	t compared to control (p-value)	"I don't know" assignments
Control	88%	84%–91%	-	2.7%
GPT-3 175B	52%	48%–57%	12.7 (3.2e-23)	10.6%

Table 3.12: People's ability to identify whether ~ 500 word articles are model generated (as measured by the ratio of correct assignments to non-neutral assignments) was 88% on the control model and 52% on GPT-3 175B. This table shows the results of a two-sample T-Test for the difference in mean accuracy between GPT-3 175B and the control model (an unconditional GPT-3 Small model with increased output randomness).

Brown et al. (2020) arXiv: 2005.14165

Language

Ability of using language

Using language was one of the key abilities of the human being for the civilization.

With the experiences on AIs, I recognized that the ability of using language would have much more important meanings than I expected.

Language is the **source for understanding, consciousness, and many other human abilities**, not merely for communication.

Expansion of language

The present AI can use not only many human languages but also many **programming languages** (codes).
Actually, one can communicate with an AI for telling one's intention by using a code in a prompt.

Images, videos, musics, etc. can also be used for communication.
(multimodal communication)

Intelligence Above Human Level

General AI (GAI)

When will the AIs exceed the human ability, especially the realization of GAI?

I hopefully prospects that will be soon, e.g. in a several years.

The present AIs already exceed average human ability in many fields.

Meta-recognition

Present AI's look like human behaving under subconsciousness.

lacking a plan, future estimation, or self-monitoring, … (dep. on prompting)

AIs would need to acquire the ability of meta-recognition.

In actual, they already have ability of self-assessment when explicitly asked.

Applications

There will be a lot of applications and ideas, ….

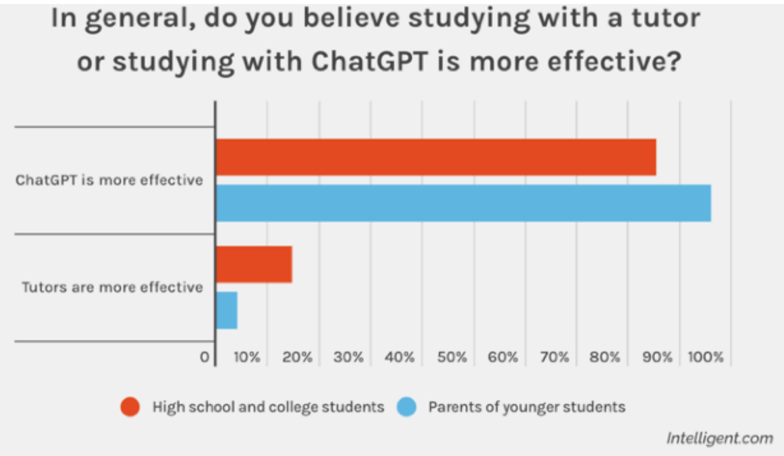
I absolutely do not have an ability to list.

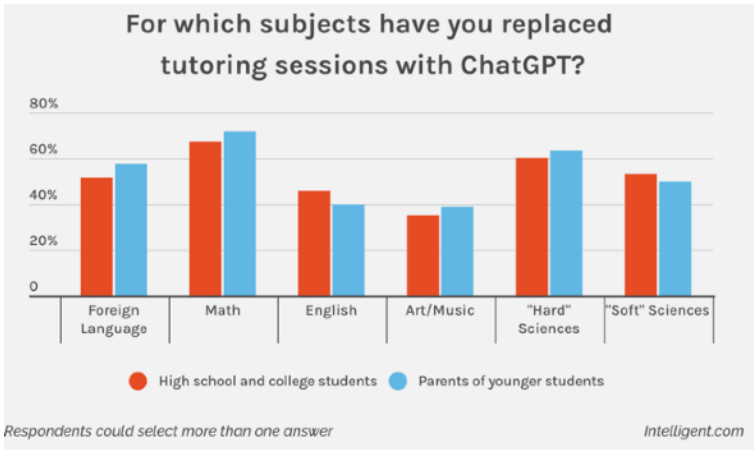
Just pick-up a few from scientific view

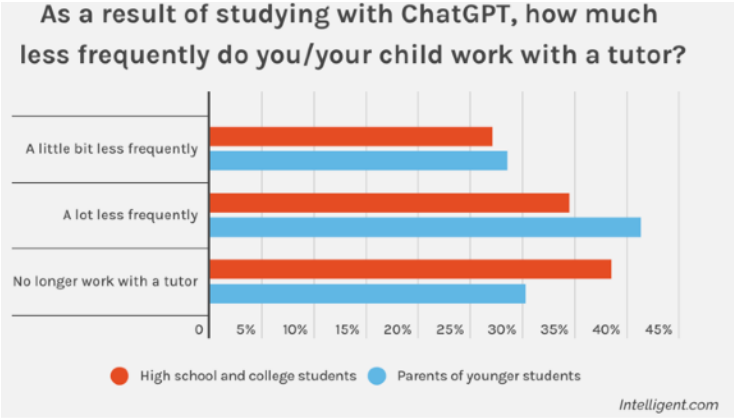
Education

NEW SURVEY FINDS STUDENTS ARE REPLACING HUMAN TUTORS WITH CHATGPT

a survey in US [intelligent.com 2023.7.1]







My experience of an effective usage of an AI is to study with a text book (or a paper) and make questions to AIs on my unresolved ideas.

Ninety-five percent of students and parents of students surveyed say **their grades** have gotten **better** since they or their children **started studying with ChatGPT**.

Applications

For examples in science

development psychology, cognitive science, social science, brain science, …

availability of

well-controlled experiments, reproducibility, observation of internal status,

A simulation of human behavior in a society:

Generative Agents: Interactive Simulacra of Human Behavior arXiv:2304.03442v1 (2023)

Physics Research and Education

It is obvious that using AIs will be important in the fields of research and education.

How?	When?
1. support for studying new knowledge	starting
2. a tool for ideas, expansion, sophistication, planning	starting
3. Problem assignment, driving force of science	near future?

Humans will continue working on physics having AIs as a good assistant since the human has the motivation of understanding the world.

A comment on Prompting

Prompting is a communication skill.

Various tips of prompting for having a good performance of AIs are proposed.

Such an ability of using prompts are quite important.

However, prompting is not just a skill using tips.

Rather, it is like an ability to instruct a team member who

- has enormous knowledge and performance
- but has little experience nor common sense in your field

to achieve a good performance for supporting you.

Prompting is a communication skill

getting along with an AI and guiding it.

Summary

I have explained

- how a generative-AI works

- what kind of abilities are observed

- personal opinions and discussions

using information from various sources.

Have you own opinions and plans in beginning of AI era!

Thank you for your attention!

References

References

- L. Tunstall, L. von Werra, and T. Wolf, *Natural Language Processing with Transformers*, (O'Reilly, 2022)
- D. Rothman, 黒川利明訳, Transformerによる自然言語処理 (朝倉書店, 2021)
- A. Vaswani et al., *Attention is All You Need*, https://doi.org/10.48550/arXiv.1706.03762 .
- Nick Bostrom, Superintelligence (Oxford Univ. Press, 2014)
- Max Tegmark, Life 3.0 — begin human in the age of artificial intelligence (Penguin books, 2017)
- L. Tunstall, L. von Werra, and T. Wolf, Natural Language Processing with Transformers (O'Reilly, 2022)
-