《物理与人工智能》

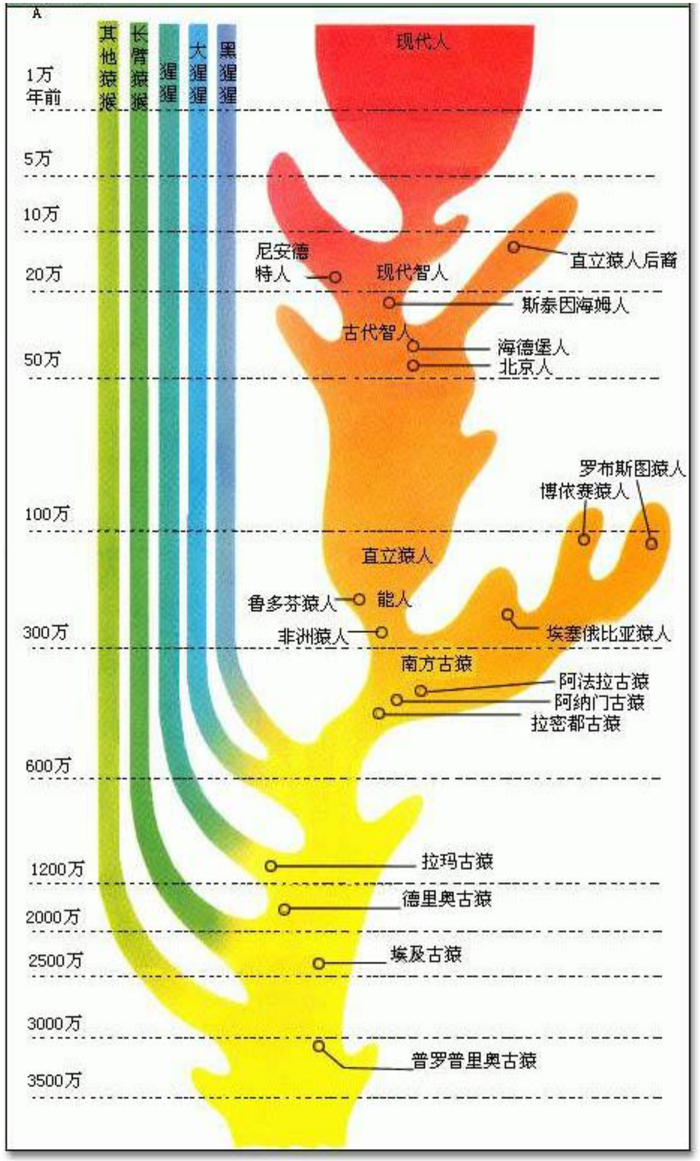
2. 大语言模型概览

授课教师： 马滟青

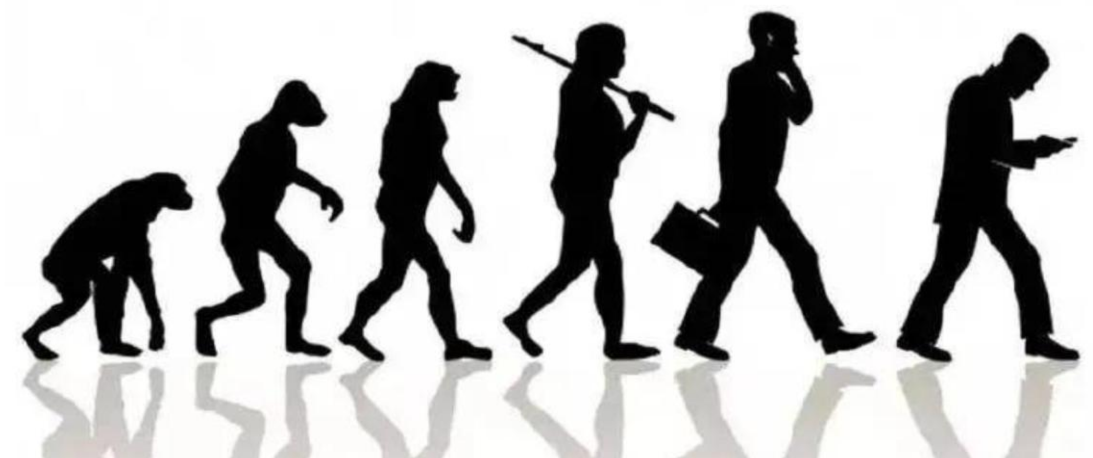
2025/09/08（第一周）

鸣谢：基于计算机学院《人工智能引论》课程和slazebni幻灯片

背景介绍——语言的重要性



- 大约5-10万年前，人类学会说话
- 大约5千年前，人类学会写字
- 图灵测试：基于语言
- 语言具有普适性，并捕捉大量智能行为
- 语言和文字的交流促进了人类文明的发展



背景介绍——为何研究自然语言处理?

北京大学 PEKING UNIVERSITY

自然语言处理 (natural language processing, NLP)

- 让机器与人类交流
- 让机器学习到知识
- 推动对语言和语言使用的科学理解

2011, IBM Watson

2018, Google BERT

2020, OpenAI GPT-3

2022, OpenAI ChatGPT

背景介绍——NLP可以做些什么?

- 语音识别（speech recognition）
- 文本-语音合成（text-to-speech synthesis）
- 机器翻译（machine translation）
- 信息提取（information extraction）
- 信息检索（information retrieval）
- 问答（question answering）
- 机器写作
- 文本摘要
- ……

Ameca多语言测试

2023年4月

背景介绍——NLP难点何在?

- **文法**（grammar）：定义合法句的句法（syntax）
- **语义规则**（semantic rule）：定义其含义

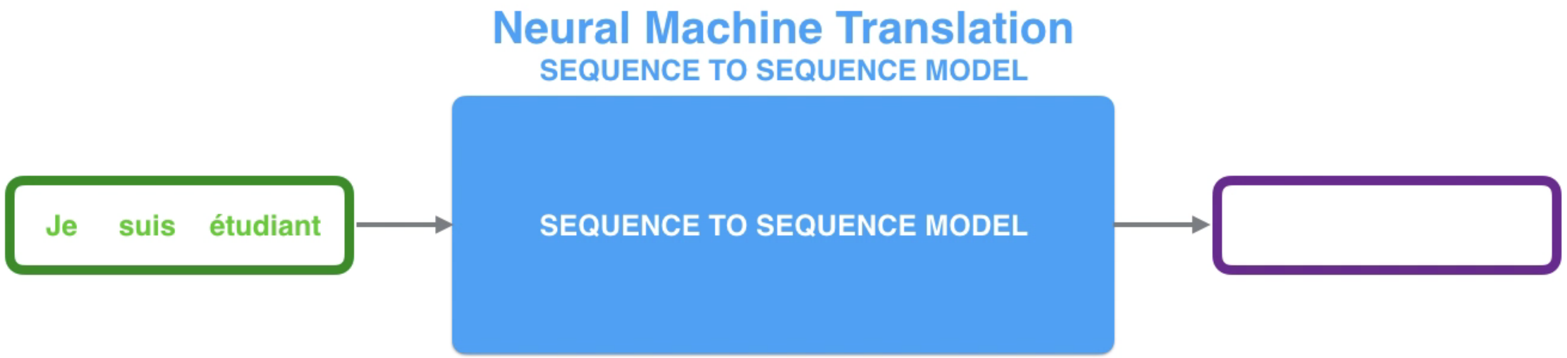
自然语言无法像形式语言一样清晰地表示：

- 不同人在不同时间对语言的判断有区别
  - “附近有餐馆”->步行距离？开车距离？
  - “早点休息”->晚八点？凌晨两点？
- 自然语言存在歧义，也是模糊不清的
  - “咬死了猎人的狗”-> 狗死了？猎人死了？
  - “小李告诉小张他通过了考试”-> 小张通过了?小李通过了?
- 自然语言没有正式定义从符号到对象的映射

自然语言的表示

$$y = f(x)$$

- 如何将自然语言用数字表达？



自然语言的表示——分词（Tokenization）

北京大學 PEKING UNIVERSITY

- 将文本序列分割为独立的“单词”
- 例1："Whatever remains, however improbable, must be the truth.“
  - ✓ ["Whatever", "remains,", "however", "improbable,", "must", "be", "the", "truth."]
  - ✓ ["Whatever", "remains", "however", "improbable", "must", "be", "the", "truth"]

- 分到什么级别合适？由语义决定!
- 类似物理中的有效自由度

自然语言的表示——中文分词

北京大学 PEKING UNIVERSITY

- 中文在词与词之间没有任何空格之类的显示标志指示词的边界。因此，中文分词是很多自然语言处理系统中的基础模块和首要环节。

■ **基于字符串匹配的分词方法**

✓ 按照一定的策略将待匹配的字符串和一个已建立好的“充分大的”词典中的词进行匹配，若找到某个词条，则说明匹配成功，识别了该词。

■ **基于机器学习的分词方法**

✓ CRF(条件随机场)、HMM(隐马尔可夫)、SVM、深度学习(BiLSTM-CRF)等。

✓ 例：CRF：对汉字进行标注训练，不仅考虑了词语出现的频率，还考虑上下文。

■ **常用的中文分词工具包：**

✓ jieba分词、SnowNLP、THULAC、NLPIR 等。

自然语言的表示——中文分词

```
import jieba
u="我来到中国北京大学"
#全模式
test1 = jieba.cut(u, cut_all=True)
print("全模式: " + "| ".join(test1))
#精确模式
test2 = jieba.cut(u, cut_all=False)
print("精确模式: " + "| ".join(test2))
#搜索引擎模式
test3= jieba.cut_for_search(u)
print("搜索引擎模式:" + "| ".join(test3))
```

例1："我来到中国北京大学"

```
全模式: 我| 来到| 中国| 北京| 北京大学| 大学
精确模式: 我| 来到| 中国| 北京大学
搜索引擎模式:我| 来到| 中国| 北京| 大学| 北京大学
```

例2： "北大是所好大学"

```
全模式: 北大| 是| 所| 好| 大学
精确模式: 北大| 是| 所| 好| 大学
搜索引擎模式:北大| 是| 所| 好| 大学
```

自然语言的表示——Byte Pair Encoding

北京大学 PEKING UNIVERSITY

- BPE算法
  - 数据预处理：收集大量文本数据，通常来自多种来源
  - 构建初始词汇表：将文本拆分为单个字符，形成初始的词汇表
  - 统计字节对频率：计算所有相邻字符对的频率，找出最常见的字节对
  - 合并最频繁的字节对：将这些字节对合并成一个新的单元，更新词汇表
  - 重复：重复上述步骤，直到达到预定的词汇表大小或没有更多的字节对可以合并
- BPE优点：
  - 处理稀有词
  - 减少词汇表大小
  - 对错别字的处理
  - 多语言支持

当前主流方法

自然语言的表示——词表示

北京大学 PEKING UNIVERSITY

- **核心目标：将单词转化为机器可计算的数字向量**
- **分布式表示** (distributed representation)
  - ✓ **理论基础**：词的语义由其上下文决定! 上下文相似的词，其语义也相似。

"He wrote a book."

he	[-0.34, -0.08, 0.02, -0.18, 0.22, ...]
wrote	[-0.27, 0.40, 0.00, -0.65, -0.15, ...]
a	[-0.12, -0.25, 0.29, -0.09, 0.40, ...]
book	[-0.23, -0.16, -0.05, -0.57, ...]

自然语言的表示——从静态到动态的词表示

北京大學 PEKING UNIVERSITY

1. 静态词嵌入 (过去)

- 特点： 一个词，一个固定向量
- 致命缺点：无法理解一词多义
- “苹果”在吃苹果和苹果手机中，向量完全相同。

2. 动态上下文表示 (现在主流)

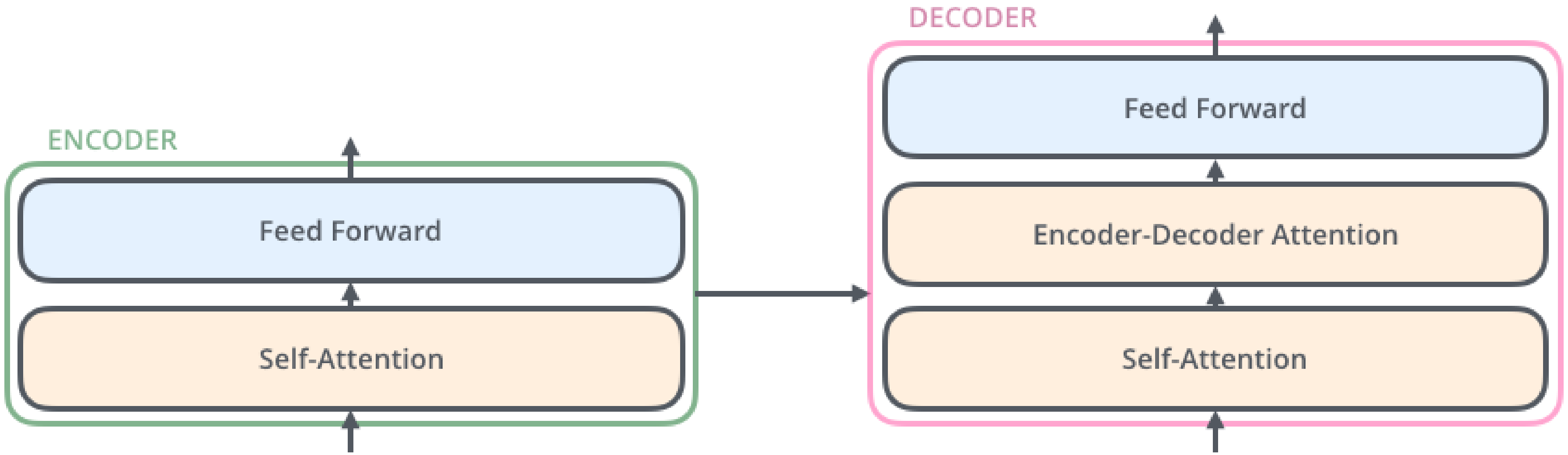
- 特点： 一个词，多种向量（由其所在句子上下文决定）
- 优势：完美解决一词多义
- “苹果”在吃苹果和苹果手机中，会生成两个不同的向量。
- 代表模型：BERT、GPT等所有现代大模型

如何实现？

Encoder-decoder transformer

Encoder: receives entire input sequence and outputs encoded sequence of the same length

Decoder: predicts next token conditioned on encoder output and previously predicted tokens



A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, I. Polosukhin, [Attention is all you need](), NeurIPS 2017

Encoder-only transformer: BERT

- At each position, the entire sequence serves as context – full attention is used

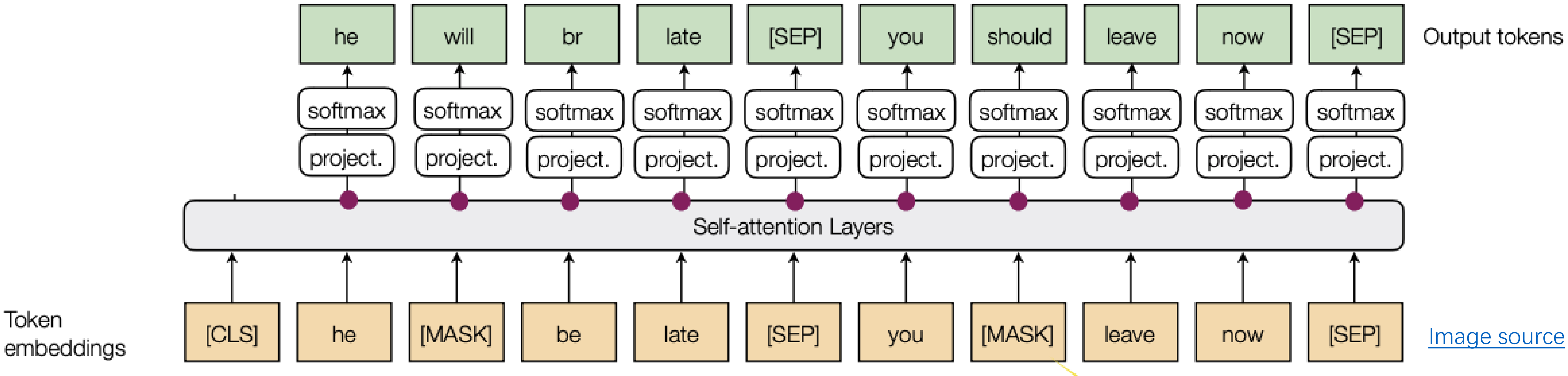
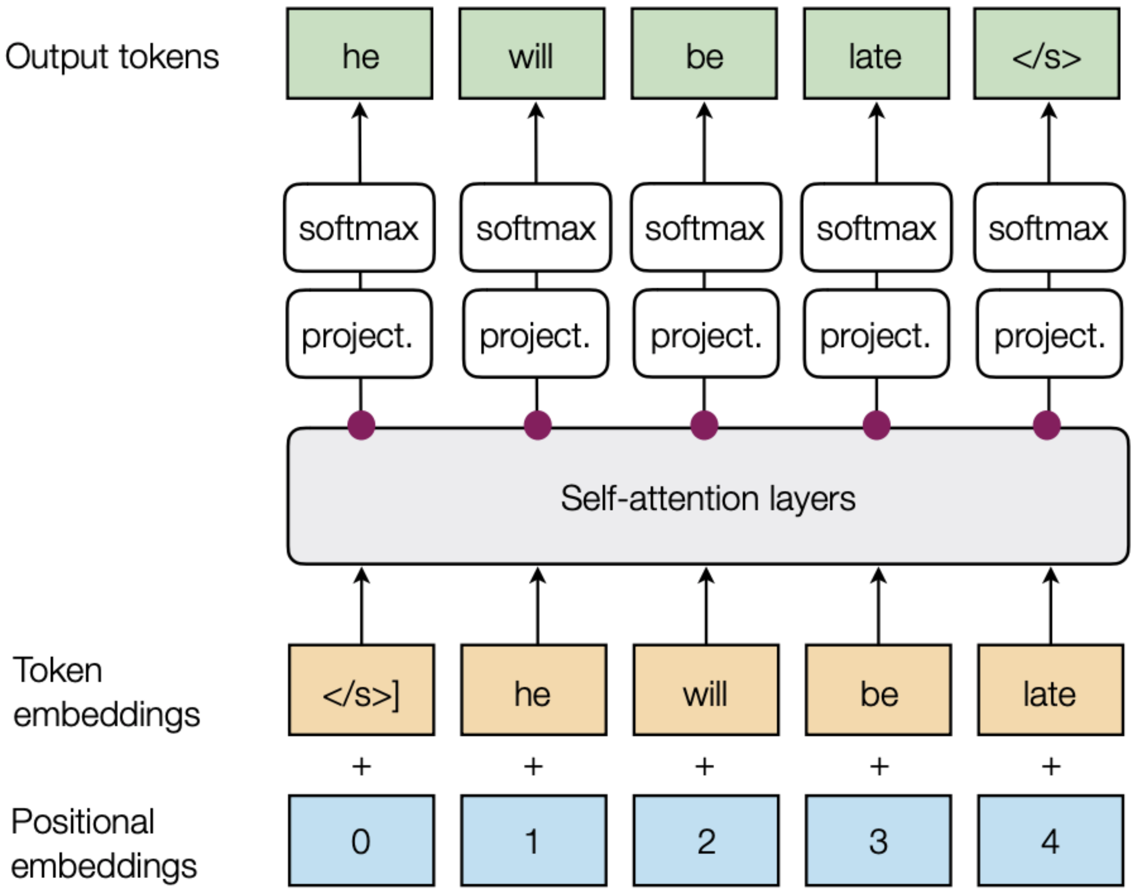


Image source

J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, EMNLP 2018

Decoder-only transformer: GPT

- Task is next token prediction – only masked attention is used

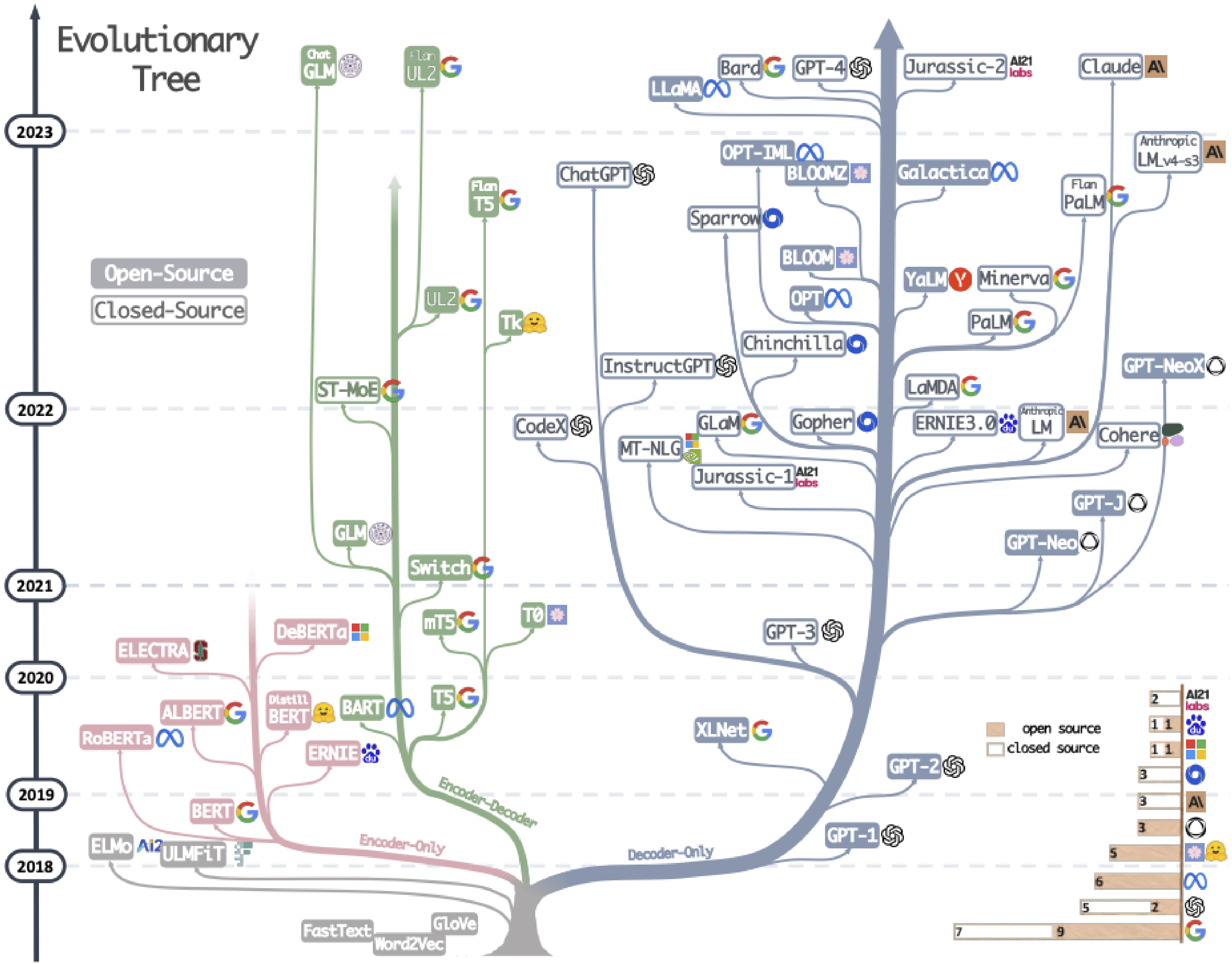


A. Radford et al. [Improving language understanding by generative pre-training](). 2018

[Image source]()

Large language models

PEKING UNIVERSITY 北京大学

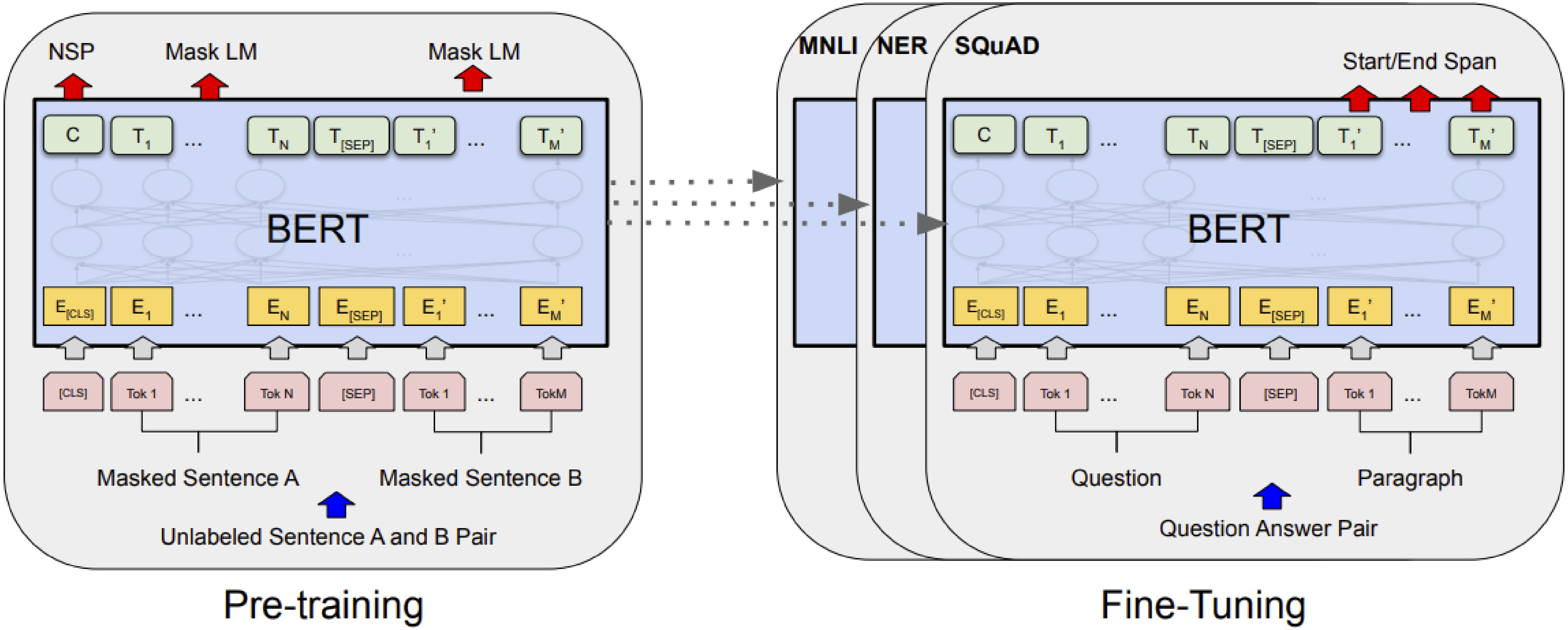


[Source](#)

Self-supervised language modeling with transformers

1. Download A LOT of text from the internet
2. Train a giant transformer using a suitable pretext task
3. Fine-tune the transformer on desired NLP task (optional)

Bidirectional encoder representations from transformers (BERT)



J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](), EMNLP 2018

BERT: Pretext tasks

- Masked language model (MLM)
  - Randomly mask 15% of tokens in input sentences, goal is to reconstruct them using bidirectional context

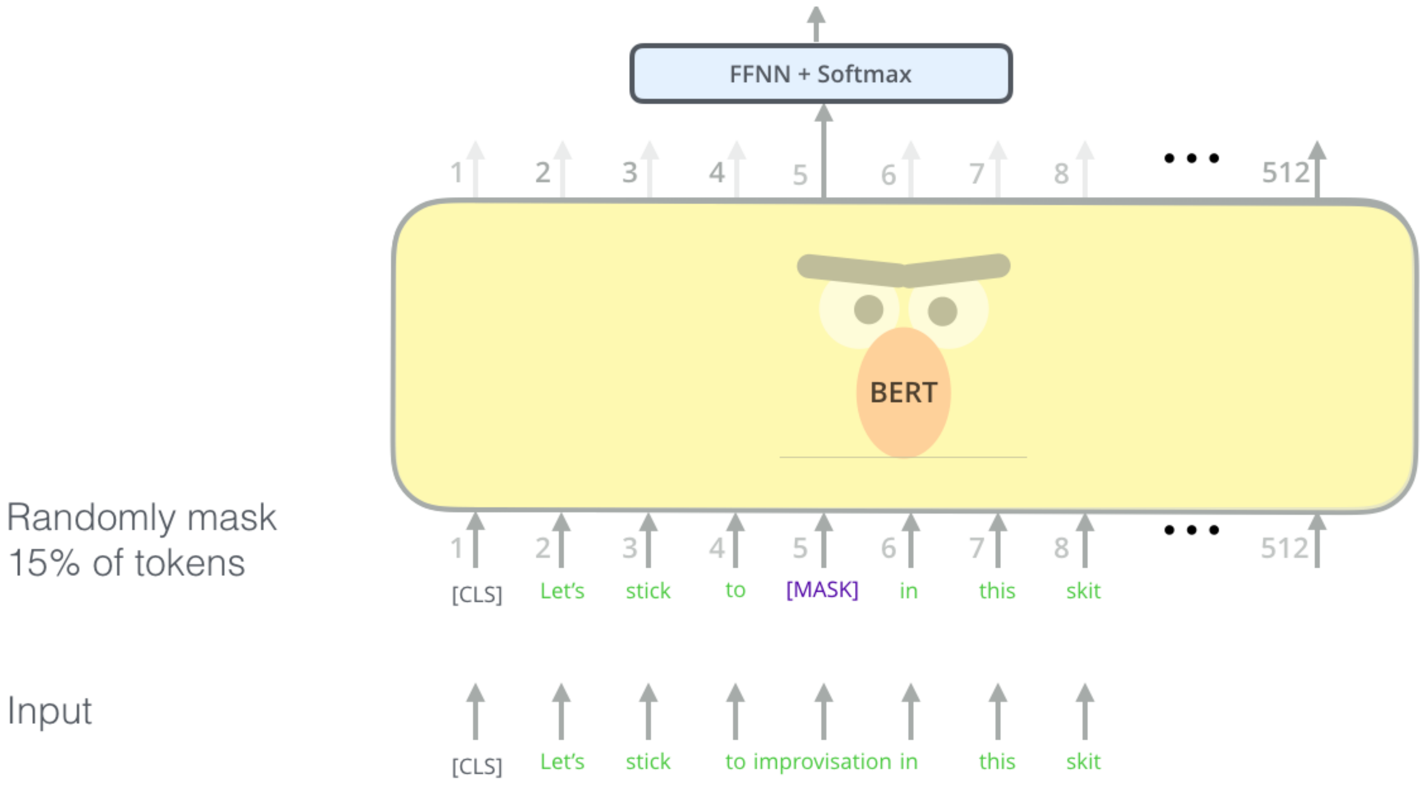


Image source

BERT: Pretext tasks

- Next sentence prediction (NSP)
  - Useful for Question Answering and Natural Language Inference tasks
  - In the training data, 50% of the time B is the actual sentence that follows A (labeled as IsNext), and 50% of the time it is a random sentence (labeled as NotNext).

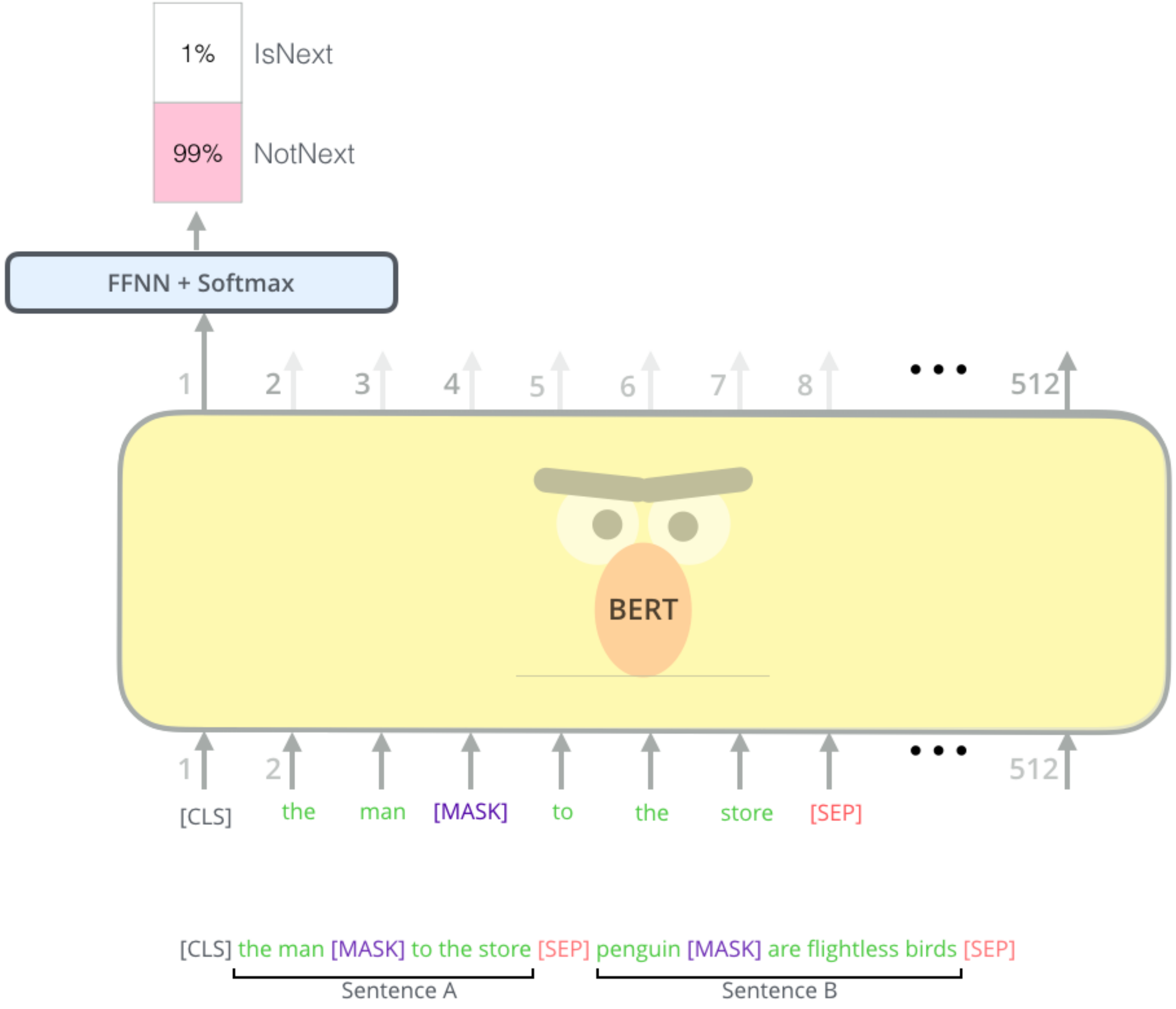
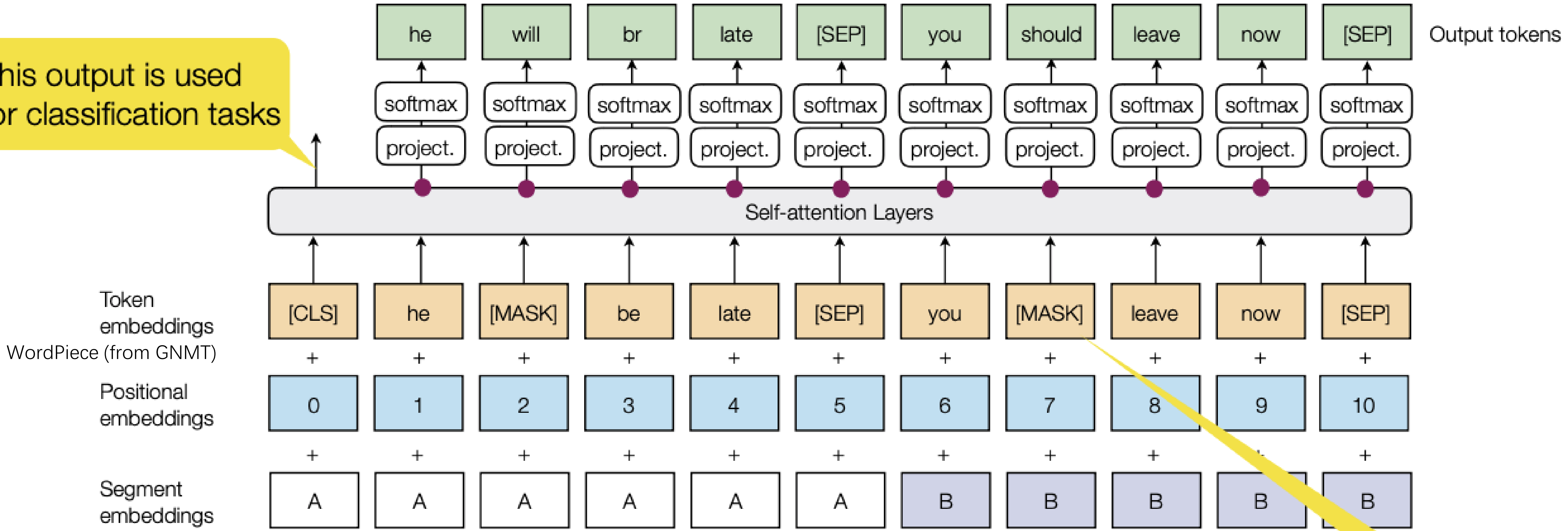


Image source

BERT: More detailed view

This output is used for classification tasks



15% of tokens get masked

Trained on Wikipedia (2.5B words) + BookCorpus (800M words)

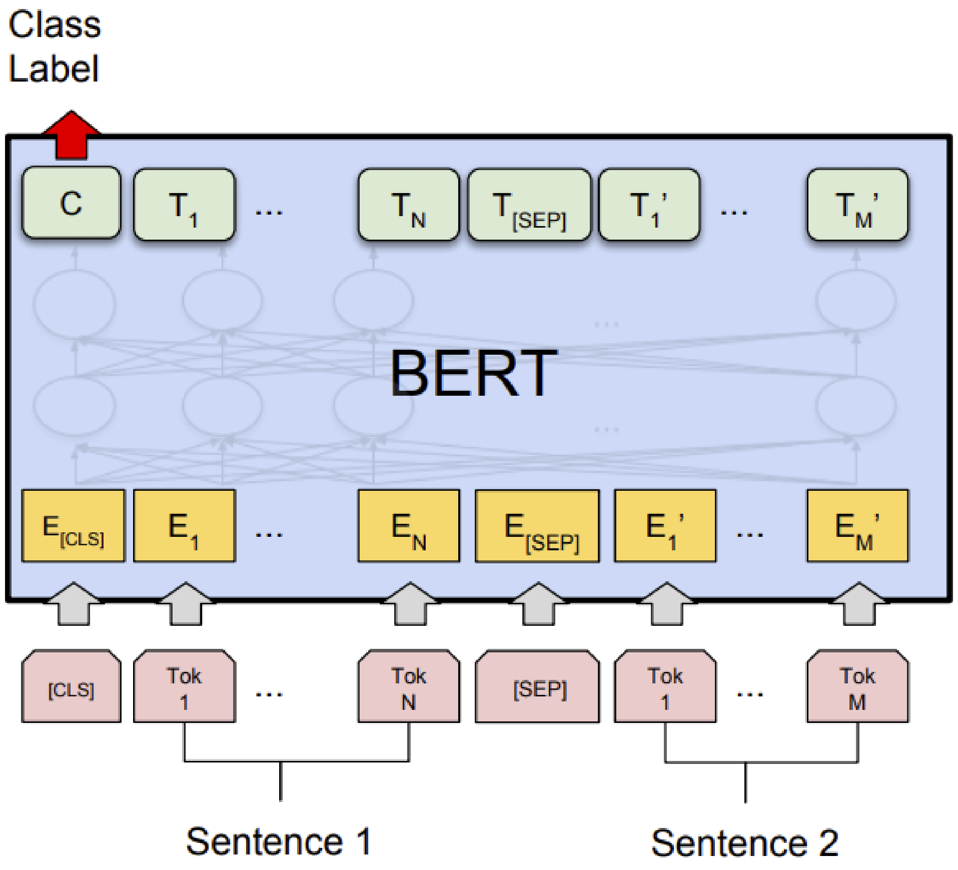
Image source

BERT: Evaluation

- General Language Understanding Evaluation (GLUE) benchmark (gluebenchmark.com)

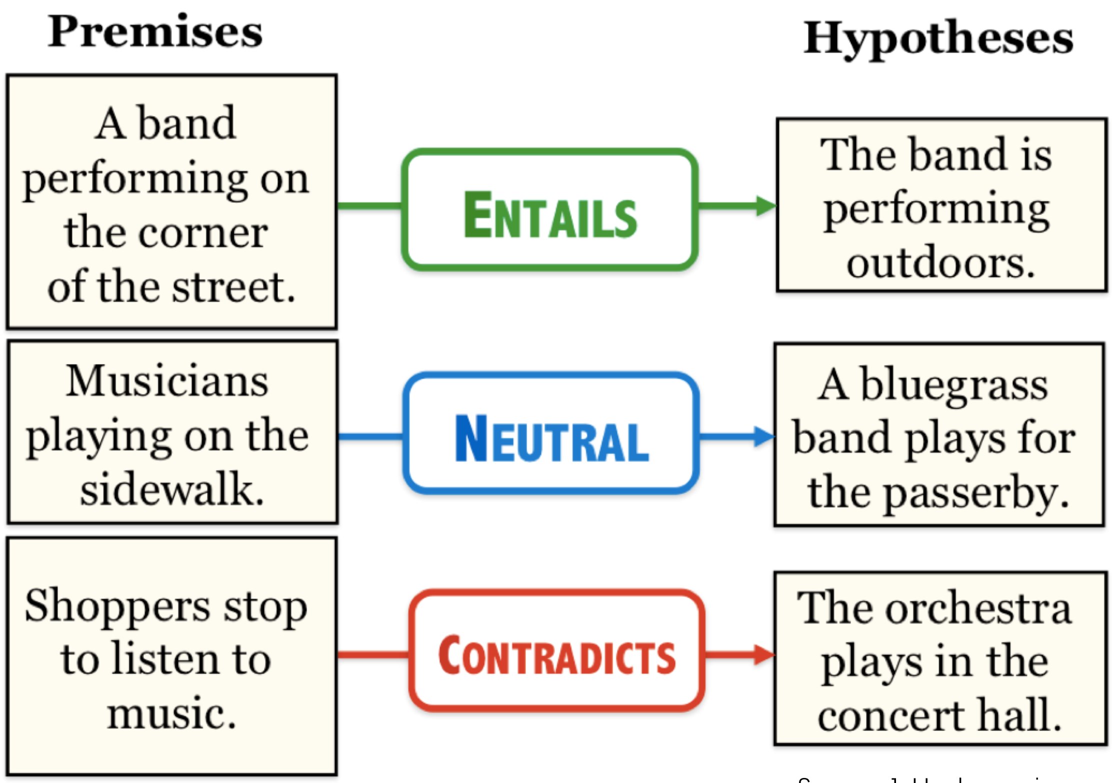
System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

BERT: Downstream tasks



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

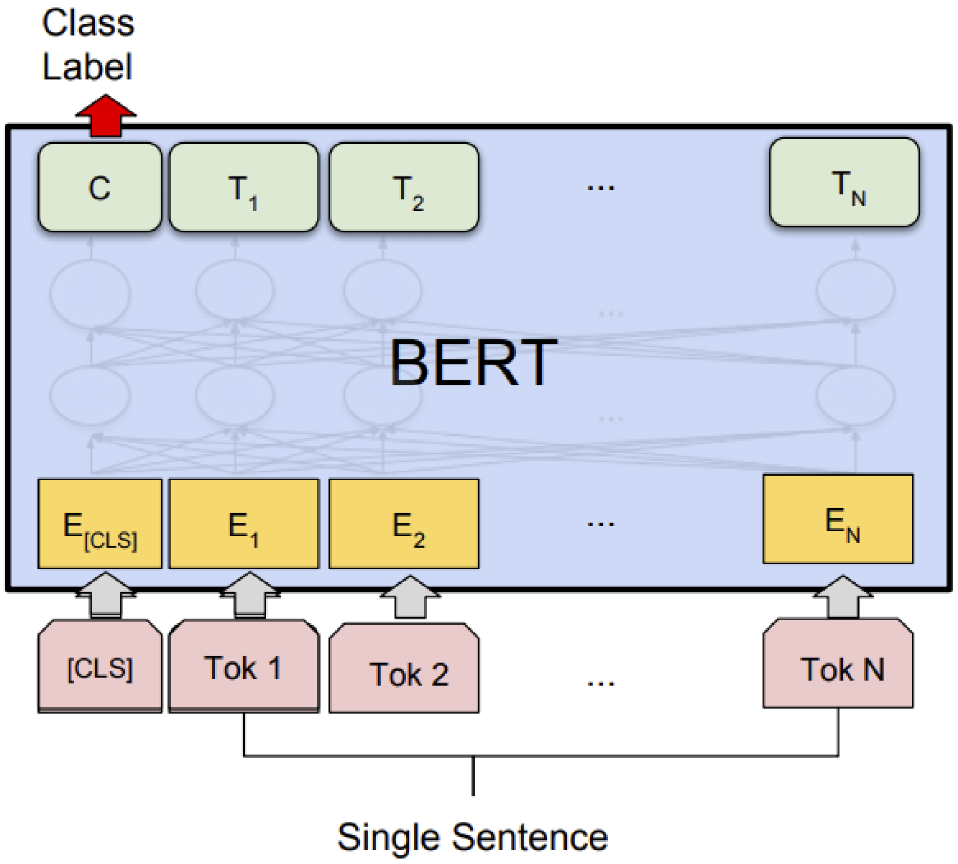
Textual entailment



Source: J. Hockenmaier

Entailment, textual equivalence and similarity

BERT: Downstream tasks



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

CoLa

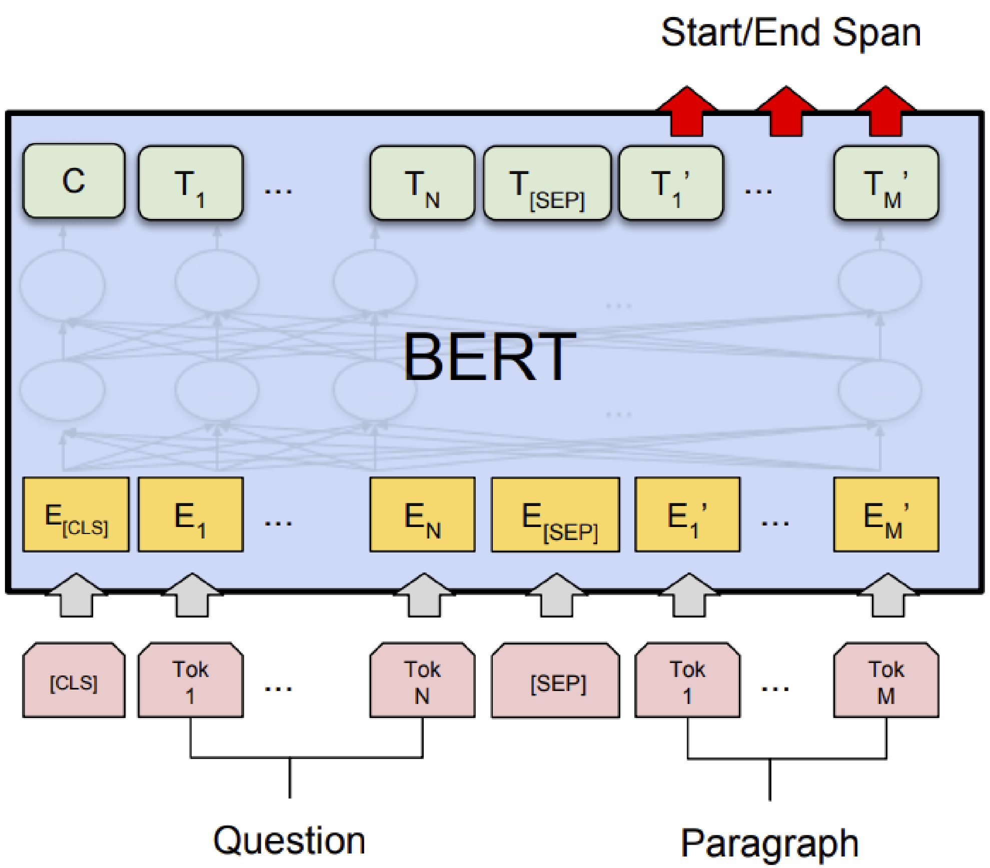
Sentence: The wagon rumbled down the road.
Label: Acceptable

Sentence: The car honked down the road.
Label: Unacceptable

Sentiment classification, linguistic acceptability

Image source

BERT: Downstream tasks



(c) Question Answering Tasks:
SQuAD v1.1

Find span in paragraph that contains the answer

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

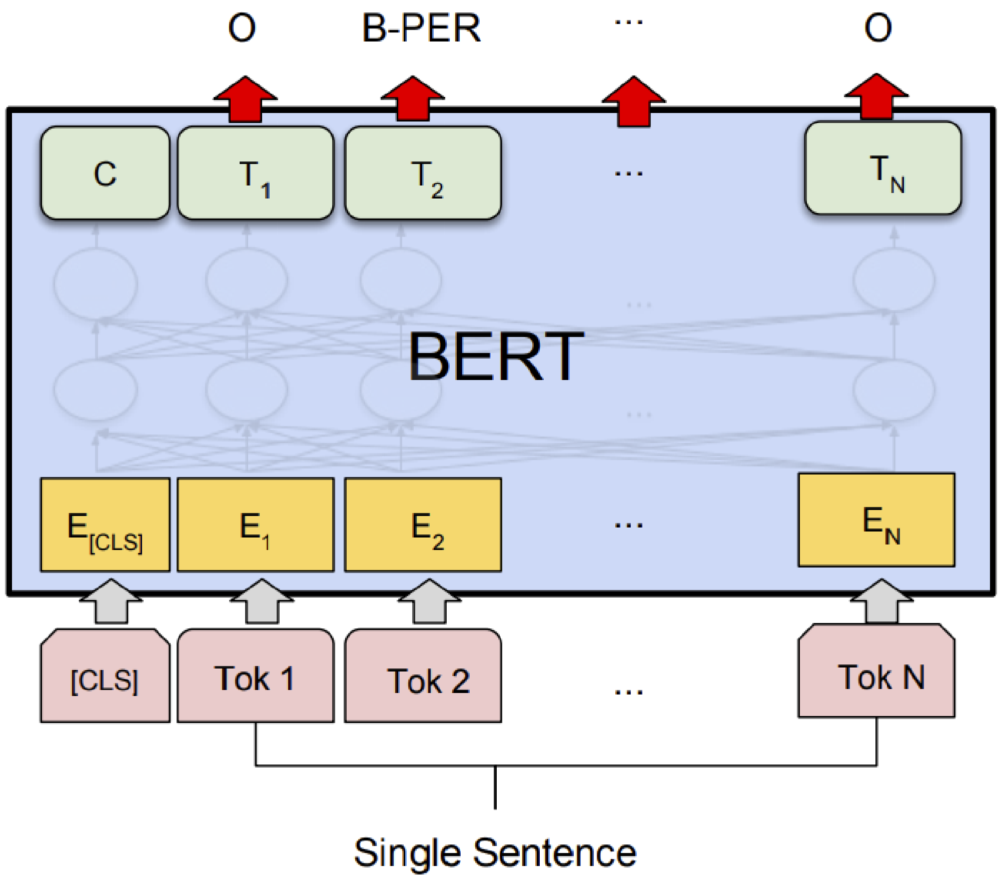
What causes precipitation to fall?
gravity

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

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

Source: SQuAD v1.1 paper

BERT: Downstream tasks



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

Named entity recognition

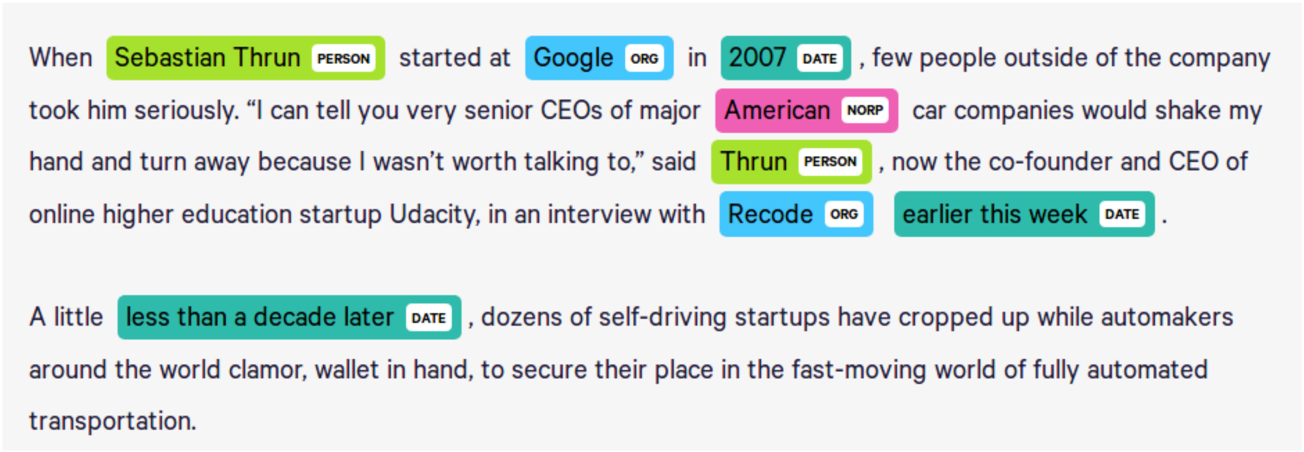


Image source

Other early language models

Alias	Model	Token	Tasks	Language
ULMfit	LSTM	word	Causal LM	English
ELMo	LSTM	word	Bidirectional LM	English
OpenAI GPT	Transformer	subword	Causal LM + Classification	English
BERT	Transformer	subword	Masked LM + Next sentence prediction	Multilingual
XLM	Transformer	subword	Causal LM +Masked LM + Translation LM	Multilingual

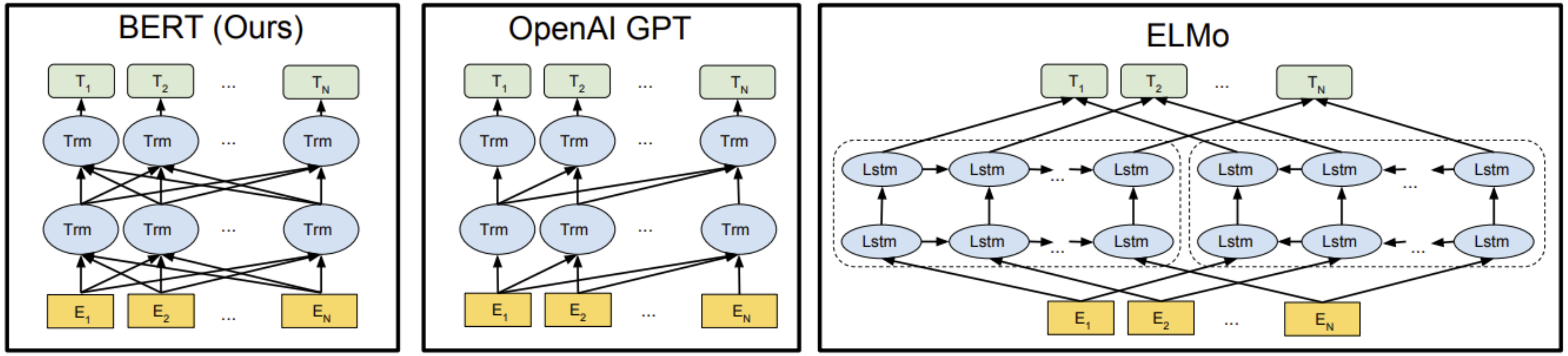


Image source

OpenAI GPT (Generative Pre-training)

- Pre-training task: next token prediction (causal language modeling)

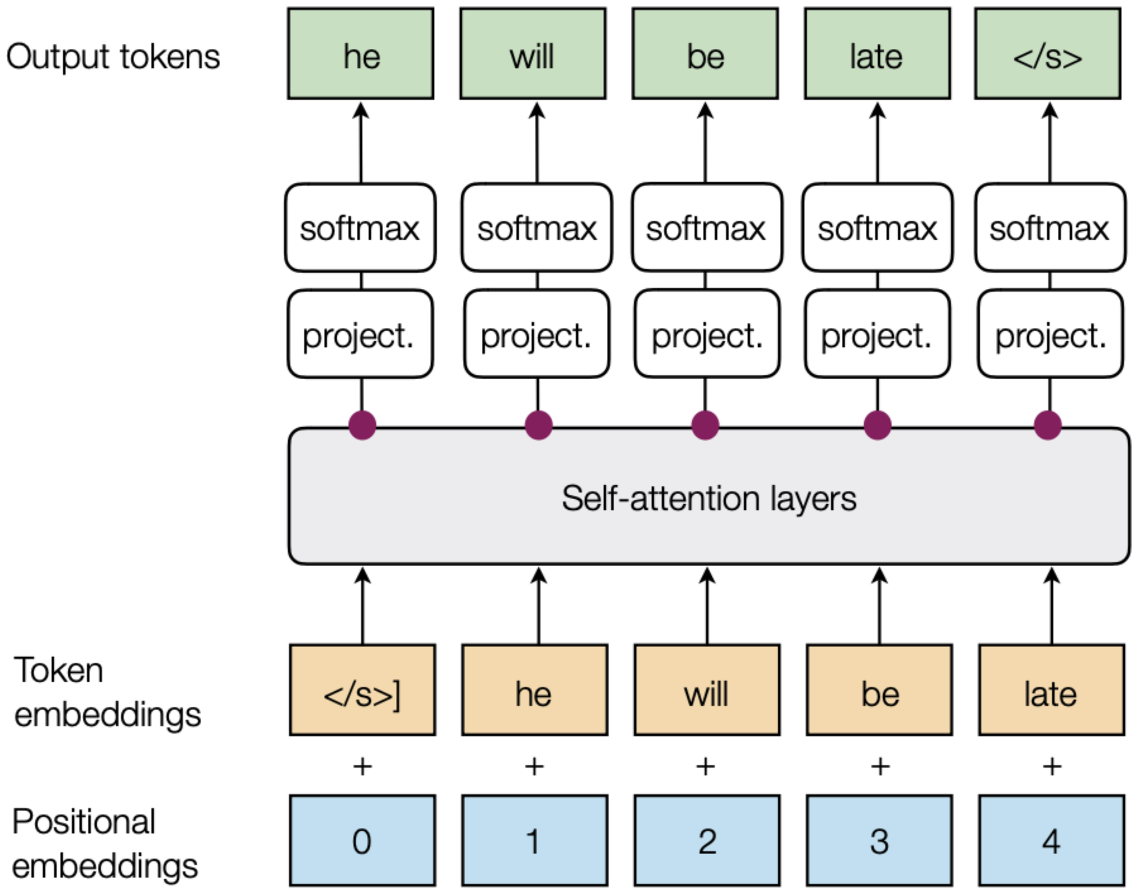


Image source

A. Radford et al. Improving language understanding by generative pre-training. 2018

Scaling up transformers

Model	Layers	Hidden dim.	Heads	Params	Data	Training
Transformer-Base	12	512	8	65M		8x P100 (12 hours)
Transformer-Large	12	1024	16	213M		8x P100 (3.5 days)

Vaswani et al. Attention is all you need. NeurIPS 2017 (Google)

PEKING UNIVERSITY 北京大學

Scaling up transformers

Model	Layers	Hidden dim.	Heads	Params	Data	Training
Transformer-Base	12	512	8	65M		8x P100 (12 hours)
Transformer-Large	12	1024	16	213M		8x P100 (3.5 days)
BERT-Base	12	768	12	110M	13 GB	4x TPU (4 days)
BERT-Large	24	1024	16	340M	13 GB	16x TPU (4 days)

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. EMNLP 2018 (Google)

Source: J. Johnson

北京大学 PEKING UNIVERSITY

Scaling up transformers

Model	Layers	Hidden dim.	Heads	Params	Data	Training
Transformer-Base	12	512	8	65M		8x P100 (12 hours)
Transformer-Large	12	1024	16	213M		8x P100 (3.5 days)
BERT-Base	12	768	12	110M	13 GB	4x TPU (4 days)
BERT-Large	24	1024	16	340M	13 GB	16x TPU (4 days)
XLNet-Large	24	1024	16	~340M	126 GB	512x TPU-v3 (2.5 days)
RoBERTa	24	1024	16	355M	160 GB	1024x V100 GPU (1 day)

Yang et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding. 2019 (Google, CMU)
Liu et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019 (FAIR, UW)

Source: J. Johnson

北京大学 PEKING UNIVERSITY

Scaling up transformers

Model	Layers	Hidden dim.	Heads	Params	Data	Training
Transformer-Base	12	512	8	65M		8x P100 (12 hours)
Transformer-Large	12	1024	16	213M		8x P100 (3.5 days)
BERT-Base	12	768	12	110M	13 GB	4x TPU (4 days)
BERT-Large	24	1024	16	340M	13 GB	16x TPU (4 days)
XLNet-Large	24	1024	16	~340M	126 GB	512x TPU-v3 (2.5 days)
RoBERTa	24	1024	16	355M	160 GB	1024x V100 GPU (1 day)
GPT-2	48	1600	?	1.5B	40 GB	

Radford et al. Language models are unsupervised multitask learners. 2019 (OpenAI)

Source: J. Johnson

北京大学 PEKING UNIVERSITY

Scaling up transformers

Model	Layers	Hidden dim.	Heads	Params	Data	Training
Transformer-Base	12	512	8	65M		8x P100 (12 hours)
Transformer-Large	12	1024	16	213M		8x P100 (3.5 days)
BERT-Base	12	768	12	110M	13 GB	4x TPU (4 days)
BERT-Large	24	1024	16	340M	13 GB	16x TPU (4 days)
XLNet-Large	24	1024	16	~340M	126 GB	512x TPU-v3 (2.5 days)
RoBERTa	24	1024	16	355M	160 GB	1024x V100 GPU (1 day)
GPT-2	48	1600	?	1.5B	40 GB	
Megatron-LM	72	3072	32	8.3B	174 GB	512x V100 GPU (9 days)

~$430,000 on Amazon AWS

Shoeybi et al. Megatron-LM: Training Multi-Billion Parameter Languge Models using Model Parallelism. 2019 (NVIDIA)

Source: J. Johnson

Scaling up transformers

Model	Layers	Hidden dim.	Heads	Params	Data	Training
Transformer-Base	12	512	8	65M		8x P100 (12 hours)
Transformer-Large	12	1024	16	213M		8x P100 (3.5 days)
BERT-Base	12	768	12	110M	13 GB	4x TPU (4 days)
BERT-Large	24	1024	16	340M	13 GB	16x TPU (4 days)
XLNet-Large	24	1024	16	~340M	126 GB	512x TPU-v3 (2.5 days)
RoBERTa	24	1024	16	355M	160 GB	1024x V100 GPU (1 day)
GPT-2	48	1600	?	1.5B	40 GB	
Megatron-LM	72	3072	32	8.3B	174 GB	512x V100 GPU (9 days)
Turing-NLG	78	4256	28	17B	?	256x V100 GPU

Microsoft. Turing-NLG: A 17-billion parameter language model by Microsoft. 2020

Scaling up transformers

Model	Layers	Hidden dim.	Heads	Params	Data	Training
Transformer-Base	12	512	8	65M		8x P100 (12 hours)
Transformer-Large	12	1024	16	213M		8x P100 (3.5 days)
BERT-Base	12	768	12	110M	13 GB	4x TPU (4 days)
BERT-Large	24	1024	16	340M	13 GB	16x TPU (4 days)
XLNet-Large	24	1024	16	~340M	126 GB	512x TPU-v3 (2.5 days)
RoBERTa	24	1024	16	355M	160 GB	1024x V100 GPU (1 day)
GPT-2	48	1600	?	1.5B	40 GB	
Megatron-LM	72	3072	32	8.3B	174 GB	512x V100 GPU (9 days)
Turing-NLG	78	4256	28	17B	?	256x V100 GPU
GPT-3	96	12288	96	175B	694 GB	$\boxed{?}$

~$4.6M, 355 GPU-years

(source)

Brown et al. Language Models are Few-Shot Learners. NeurIPS 2020 (OpenAI)

Source: J. Johnson

北京大学 PEKING UNIVERSITY

Scaling up transformers

Model	Layers	Hidden dim.	Heads	Params	Data	Training
Transformer-Base	12	512	8	65M		8x P100 (12 hours)
Transformer-Large	12	1024	16	213M		8x P100 (3.5 days)
BERT-Base	12	768	12	110M	13 GB	4x TPU (4 days)
BERT-Large	24	1024	16	340M	13 GB	16x TPU (4 days)
XLNet-Large	24	1024	16	~340M	126 GB	512x TPU-v3 (2.5 days)
RoBERTa	24	1024	16	355M	160 GB	1024x V100 GPU (1 day)
GPT-2	48	1600	?	1.5B	40 GB	
Megatron-LM	72	3072	32	8.3B	174 GB	512x V100 GPU (9 days)
Turing-NLG	78	4256	28	17B	?	256x V100 GPU
GPT-3	96	12,288	96	175B	694 GB	?
GOPHER	80	16,384	128	280B	10.55 TB	4096x TPUv3 (38 days)

$3,768,320 on Google Cloud (eval price)

J. Rae et al. Scaling Language Models: Methods, Analysis, & Insights from Training Gopher. arXiv 2021 (DeepMind)

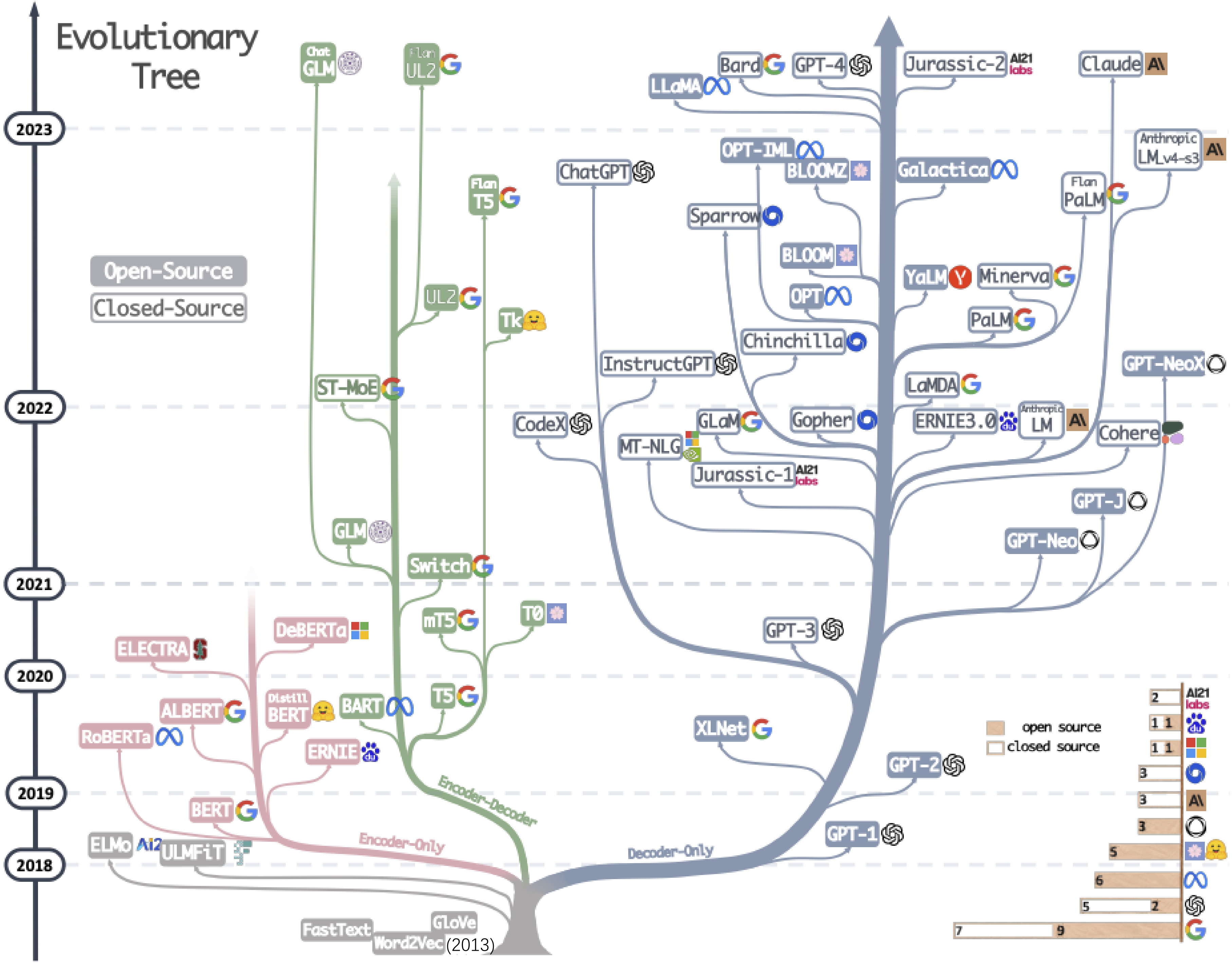
Source: J. Johnson

PEKING UNIVERSITY 北京大学

Scaling up transformers

Model	Layers	Hidden dim.	Heads	Params	Data	Training
Transformer-Base	12	512	8	65M		8x P100 (12 hours)
Transformer-Large	12	1024	16	213M		8x P100 (3.5 days)
BERT-Base	12	768	12	110M	13 GB	4x TPU (4 days)
BERT-Large	24	1024	16	340M	13 GB	16x TPU (4 days)
XLNet-Large	24	1024	16	~340M	126 GB	512x TPU-v3 (2.5 days)
RoBERTa	24	1024	16	355M	160 GB	1024x V100 GPU (1 day)
GPT-2	48	1600	?	1.5B	40 GB	
Megatron-LM	72	3072	32	8.3B	174 GB	512x V100 GPU (9 days)
Turing-NLG	78	4256	28	17B	?	256x V100 GPU
GPT-3	96	12,288	96	175B	694 GB	?
GOPHER	80	16,384	128	280B	10.55 TB	4096x TPUv3 (38 days)
PaLM	118	18,432	48	540B	?	6144x TPUv4

A. Chowdhery et al. PaLM: Scaling Language Modeling with Pathways. arXiv 2022 (Google)



Evolutionary Tree
2023
2022
2021
2020
2019
2018
Open-Source
Closed-Source
Chat GLM
Flan UL2
LLaMA
Bard
GPT-4
Jurassic-2
AI21 labs
Claude
Anthropic LM v4-s3
ChatGPT
OPT-IML
BLOOMZ
Galactica
Flan T5
Flan PaLM
Sparrow
BLOOM
YaLM
Minerva
ULM2
Tk
OPT
PaLM
Chinchilla
InstructGPT
GPT-NeoX
ST-MoE
LaMDA
CodeX
GLaM
Gopher
ERNIE3.0
Anthropic LM
Cohere
MT-NLG
Jurassic-1
GPT-J
GLM
GPT-Neo
Switch
T0
mT5
DeBERTa
ELECTRA
GPT-3
ALBERT
Distill BERT
BART
T5
XLNet
RoBERTa
ERNIE
GPT-2
open source
closed source
BERT
Encoder-Decoder
Encoder-Only
GPT-1
ELMo
Ai2
ULMFiT
Decoder-Only
FastText
GloVe
Word2Vec
(2013)

Source

Scaling behavior of large language models

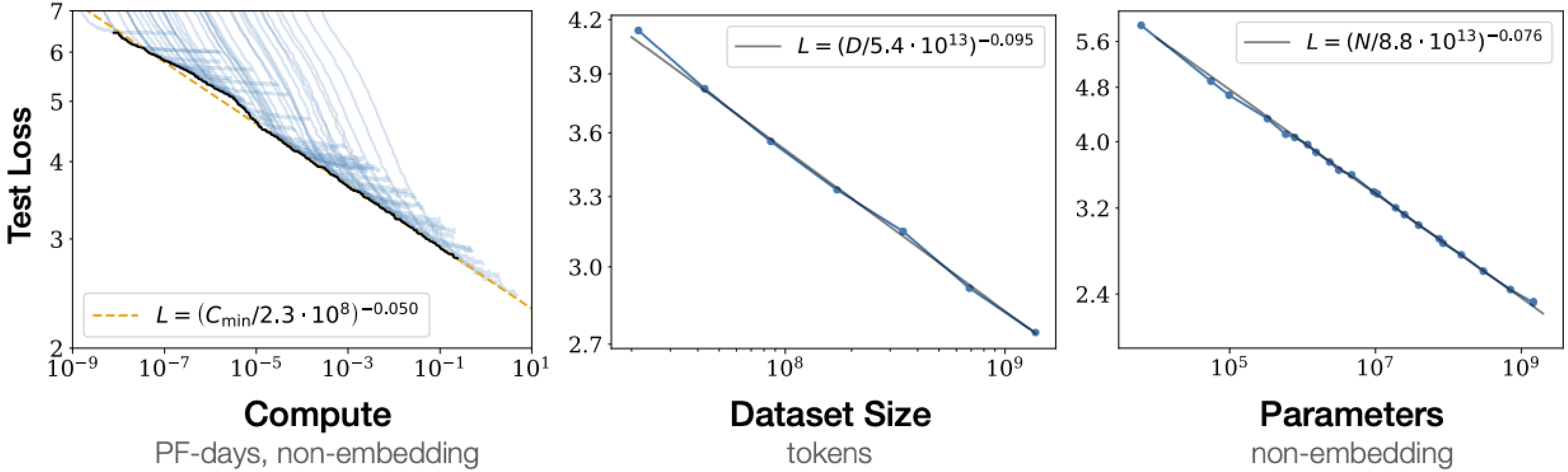


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

J. Kaplan et al. [Scaling Laws for Neural Language Models](). arXiv 2020

Scaling behavior of large language models

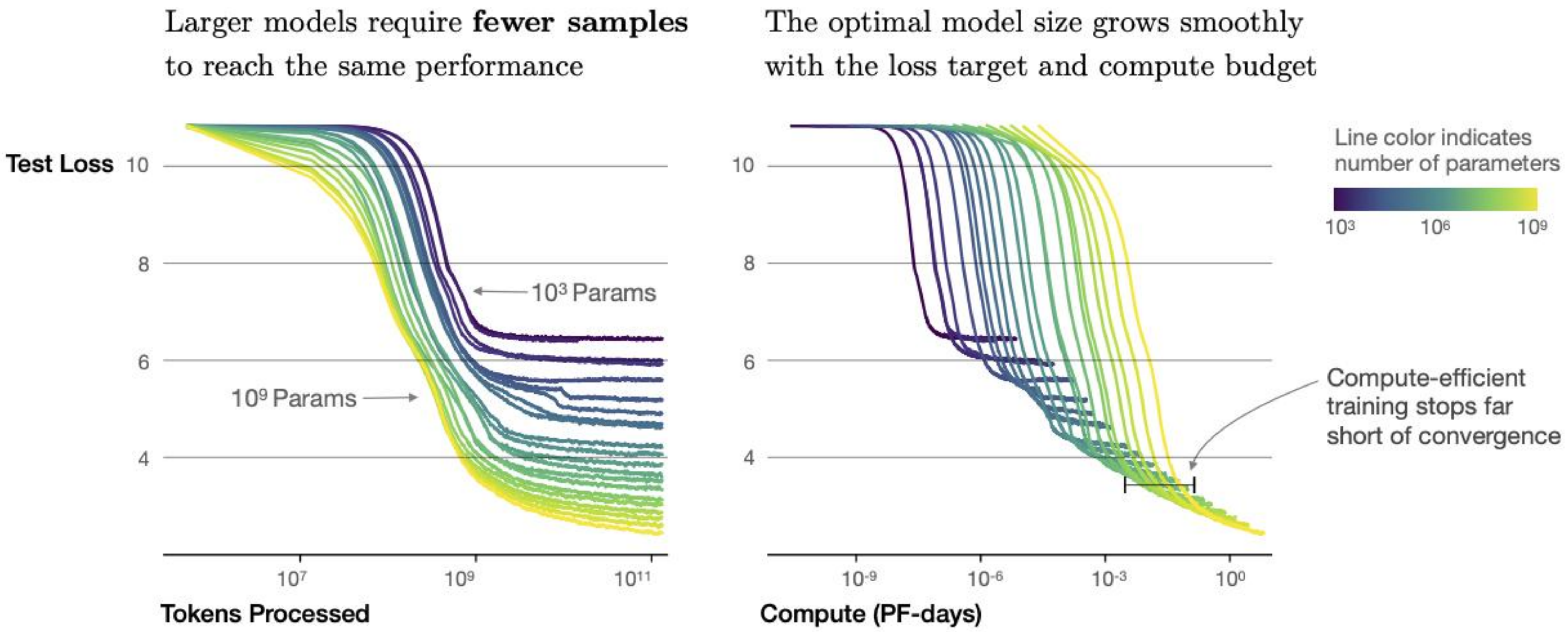


Figure 2 We show a series of language model training runs, with models ranging in size from 10^3 to 10^9 parameters (excluding embeddings).

J. Kaplan et al. Scaling Laws for Neural Language Models. arXiv 2020

Scaling behavior of large language models

Performance depends strongly on scale, weakly on model shape: Model performance depends most strongly on scale, which consists of three factors: the number of model parameters N (excluding embeddings), the size of the dataset D, and the amount of compute C used for training. Within reasonable limits, performance depends very weakly on other architectural hyperparameters such as depth vs. width.

Universality of overfitting: Performance improves predictably as long as we scale up N and D in tandem, but enters a regime of diminishing returns if either N or D is held fixed while the other increases. The performance penalty depends predictably on the ratio $N^{0.74}/D$, meaning that every time we increase the model size 8x, we only need to increase the data by roughly 5x to avoid a penalty.

Transfer improves with test performance: When we evaluate models on text with a different distribution than they were trained on, the results are strongly correlated to those on the training validation set with a roughly constant offset in the loss – in other words, transfer to a different distribution incurs a constant penalty but otherwise improves roughly in line with performance on the training set.

Sample efficiency: Large models are more sample-efficient than small models, reaching the same level of performance with fewer optimization steps (Figure 2) and using fewer data points

Convergence is inefficient: When working within a fixed compute budget C but without any other restrictions on the model size N or available data D, we attain optimal performance by training *very large models* and stopping *significantly short of convergence* (see Figure 3). Maximally compute-efficient training would therefore be far more sample efficient than one might expect based on training small models to convergence, with data requirements growing very slowly as $D \sim C^{0.27}$ with training compute.

J. Kaplan et al. [Scaling Laws for Neural Language Models](). arXiv 2020

GPT-2 and GPT-3

- Key idea: if the model and training datasets are big enough, model can adapt to new tasks *without fine-tuning*

Model	Layers	Hidden dim.	Heads	Params	Dataset
GPT-2	48	1600	?	1.5B	WebText: 40GB
GPT-3	96	12288	96	175B	CommonCrawl (cleaned up): 694GB

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

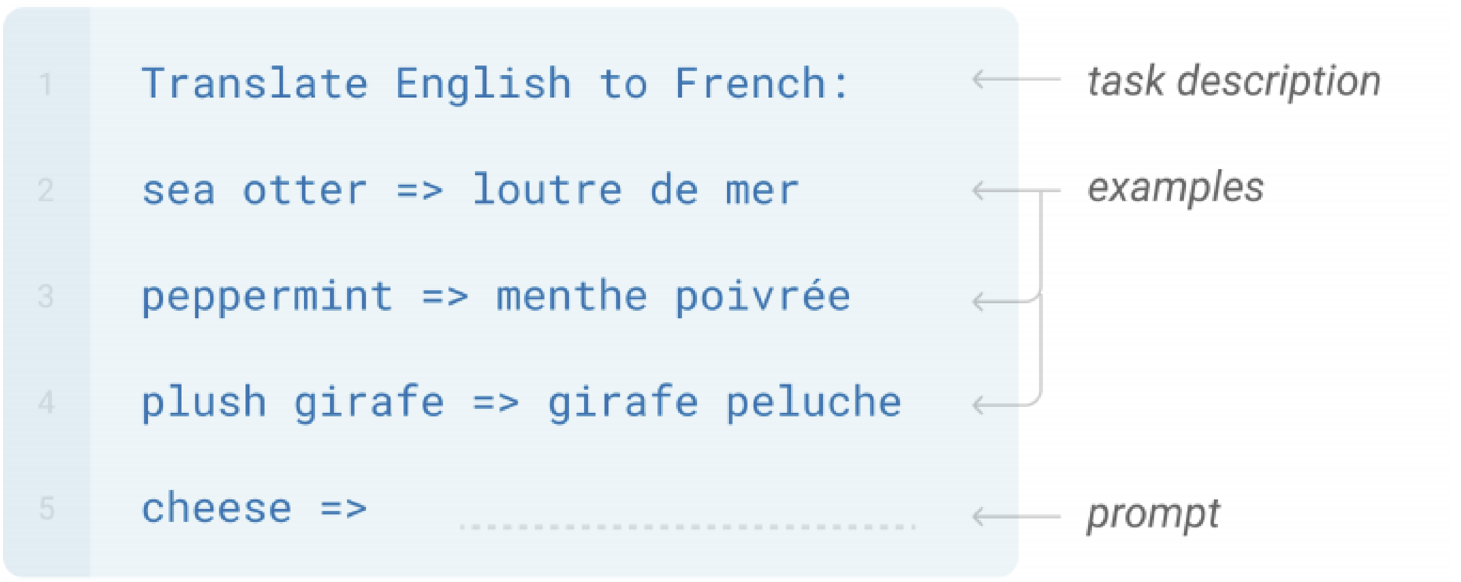
Table 2.2: Datasets used to train GPT-3. "Weight in training mix" refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen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

GPT-2: A. Radford et al., Language models are unsupervised multitask learners, 2019

GPT-3: T. Brown et al., Language models are few-shot learners, NeurIPS 2020 (Best Paper Award)

GPT-3

- Key idea: if the model and training datasets are big enough, model can adapt to new tasks *without fine-tuning*
- **Few-shot learning:** In addition to the task description, the model sees a few examples of the task



T. Brown et al., [Language models are few-shot learners](), NeurIPS 2020

GPT-3

- Key idea: if the model and training datasets are big enough, model can adapt to new tasks *without fine-tuning*
- **One-shot learning:** In addition to the task description, the model sees a *single example* of the task

T. Brown et al., Language models are few-shot learners, NeurIPS 2020

GPT-3

- Key idea: if the model and training datasets are big enough, model can adapt to new tasks *without fine-tuning*
- **Zero-shot learning:** The model sees the task description and *no* training examples

T. Brown et al., [Language models are few-shot learners](), NeurIPS 2020

Task: Generate news article

```
Title:  United Methodists Agree to Historic Split
Subtitle:  Those who oppose gay marriage will form their own denomination
Article:  After two days of intense debate, the United Methodist Church
has agreed to a historic split - one that is expected to end in the
creation of a new denomination, one that will be "theologically and
socially conservative," according to The Washington Post.  The majority of
delegates attending the church's annual General Conference in May voted to
strengthen a ban on the ordination of LGBTQ clergy and to write new rules
that will "discipline" clergy who officiate at same-sex weddings.  But
those who opposed these measures have a new plan:  They say they will form a
separate denomination by 2020, calling their church the Christian Methodist
denomination.
The Post notes that the denomination, which claims 12.5 million members, was
in the early 20th century the "largest Protestant denomination in the U.S.,"
but that it has been shrinking in recent decades.  The new split will be the
second in the church's history.  The first occurred in 1968, when roughly
10 percent of the denomination left to form the Evangelical United Brethren
Church.  The Post notes that the proposed split "comes at a critical time
for the church, which has been losing members for years," which has been
"pushed toward the brink of a schism over the role of LGBTQ people in the
church." Gay marriage is not the only issue that has divided the church.  In
2016, the denomination was split over ordination of transgender clergy, with
the North Pacific regional conference voting to ban them from serving as
clergy, and the South Pacific regional conference voting to allow them.
```

Figure 3.14: The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).

Gray: human prompts, boldface: GPT-3 completions

(Three articles provided as training examples)

Task: Use new word in sentence

A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:
We were traveling in Africa and we saw these very cute whatpus.

To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:
One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.
A "yalubalu" is a type of vegetable that looks like a big pumpkin. An example of a sentence that uses the word yalubalu is:
I was on a trip to Africa and I tried this yalubalu vegetable that was grown in a garden there. It was delicious.

A "Burringo" is a car with very fast acceleration. An example of a sentence that uses the word Burringo is:
In our garage we have a Burringo that my father drives to work every day.

A "Gigamuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigamuru is:
I have a Gigamuru that my uncle gave me as a gift. I love to play it at home.

To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is:
We screeghed at each other for several minutes and then we went outside and ate ice cream.

Gray: human prompts, boldface: GPT-3 completions

Task: Correct grammar

北京大学 PEKING UNIVERSITY

Poor English input: I eated the purple berries.
Good English output: I ate the purple berries.
Poor English input: Thank you for picking me as your designer. I'd appreciate it.
Good English output: Thank you for choosing me as your designer. I appreciate it.
Poor English input: The mentioned changes have done. or I did the alteration that you requested. or I changed things you wanted and did the modifications.
Good English output: The requested changes have been made. or I made the alteration that you requested. or I changed things you wanted and made the modifications.
Poor English input: I'd be more than happy to work with you in another project.
Good English output: I'd be more than happy to work with you on another project.

Poor English input: Please provide me with a short brief of the design you're looking for and that'd be nice if you could share some examples or project you did before.
Good English output: Please provide me with a brief description of the design you're looking for and that would be nice if you could share some examples or projects you have done before.

Poor English input: The patient was died.
Good English output: The patient died.

Poor English input: We think that Leslie likes ourselves.
Good English output: We think that Leslie likes us.

Poor English input: Janet broke Bill on the finger.
Good English output: Janet broke Bill's finger.

Poor English input: Mary arranged for, in St. Louis, John to rent a house cheap.
Good English output: Mary arranged for John to rent a house in St. Louis.

Gray: human prompts, boldface: GPT-3 completions

GPT-3 creative fiction

北京大学
PEKING UNIVERSITY

Topic: humor: literary parodies and homages, pastiches, style parodies.

Parodies of the fantasy novel series *Harry Potter* in the style of various famous authors:

By Ernest Hemingway:

"It was a cold day on Privet Drive. A child cried. Harry felt nothing. He was dryer than dust. He had been silent too long. He had not felt love. He had scarcely felt hate. Yet the Dementor's Kiss killed nothing. Death didn't leave him less dead than he had been a second before. It wasn't about living or dying really. It wasn't about death. It was about the Dark Lord, his creator, his engenderer. He was not resurrected. He was created again. He was whole again in the courtyard of Malfoy Manor."[33]

By Jane Austen:

"It is a truth universally acknowledged, that a broken Harry is in want of a book. This he knows to his cost; pressing that wretched nose against the window of a bookstore in Diagon Alley the last thing before he goes into hiding for several years, where upon, he goes straight to Azkaban."

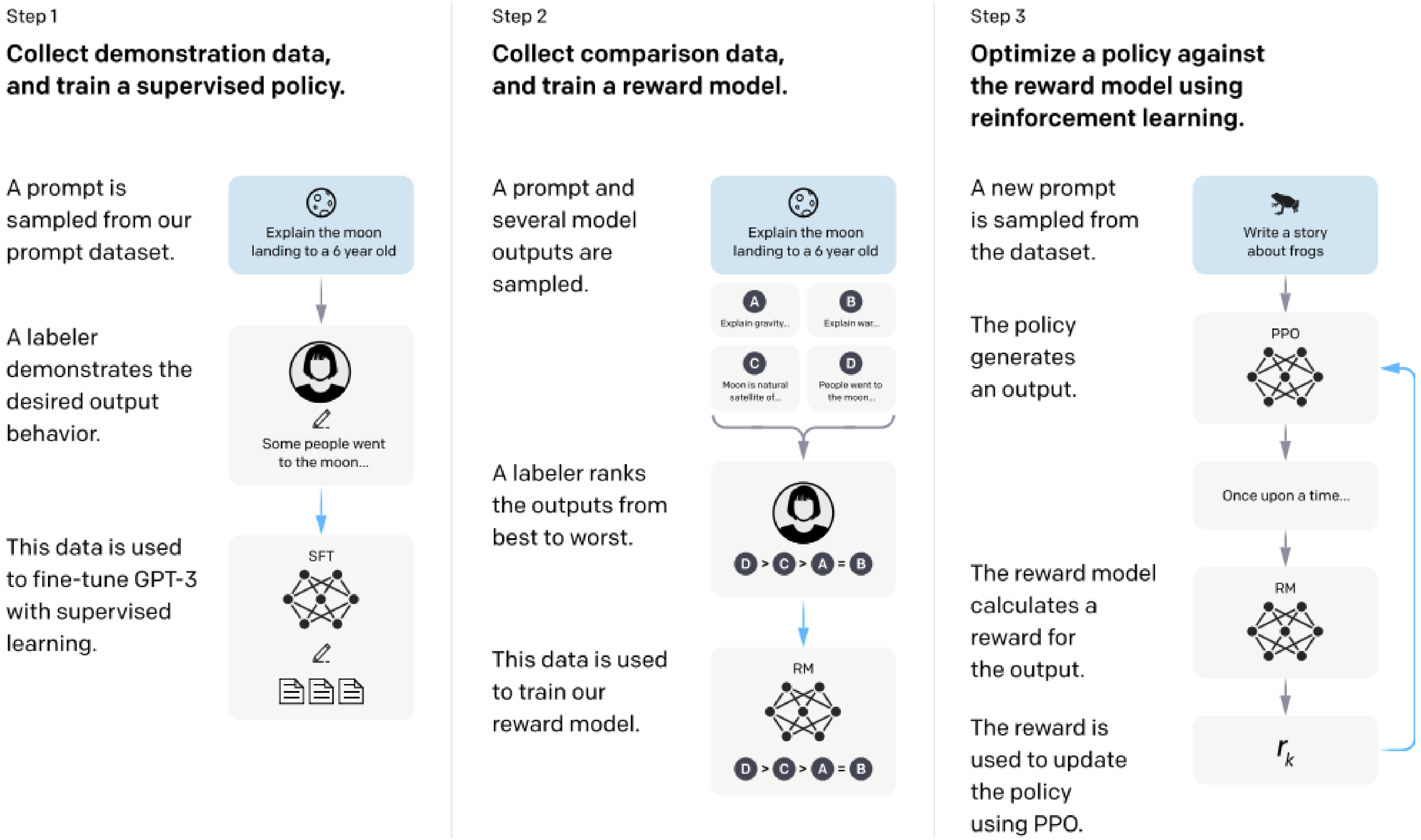
By Arthur Conan Doyle:

"Harry pushed at the swinging doors of the bookshop hard, and nearly knocked himself unconscious. He staggered in with his ungainly package, his cheeks scarlet with cold and the shame of having chosen the wrong month to go Christmas shopping. The proprietor of the store, however, didn't cast even a cursory glance at him, being far more interested in an enormous hunk of wood lying in the middle of the floor, which certainly looked like a gallows. Yes, the proprietor said to a reedy old man wearing a bowler hat and a forlorn expression that made Harry want to kick him, I can rent you such a gallows for a small fee."

https://www.gwern.net/GPT-3

For much, much more, see: https://github.com/elyase/awesome-gpt3

InstructGPT: Reinforcement learning with human feedback



L. Ouyang et al. [Training language models to follow instructions with human feedback](). NeurIPS 2022

ChatGPT

T

write a biblical verse in the style of the king james bible explaining how to remove a peanut butter sandwich from a VCR

And it came to pass that a man was troubled by a peanut butter sandwich, for it had been placed within his VCR, and he knew not how to remove it.

And he cried out to the Lord, saying, "Oh, Lord, how can I remove this sandwich from my VCR, for it is stuck fast and will not budge?"

And the Lord spoke unto him, saying, "Fear not, my child, for I shall guide thy hand and show thee the way. Take thy butter knife, and carefully insert it between the sandwich and the VCR, and gently pry them apart. And with patience and perseverance, the sandwich shall be removed, and thy VCR shall be saved."

And the man did as the Lord commanded, and lo and behold, the sandwich was removed from the VCR, and the man was saved.

And the Lord said, "Verily I say unto thee, seek not to put thy peanut butter sandwiches in thy VCR, for it is not a suitable place for such things. Rather, keep thy sandwiches in thy refrigerator or on thy plate, where they belong."

And the man heeded the Lord's words, and from that day forth, he kept his sandwiches in their proper place, and was saved from trouble and woe. Amen.

Source

GPT-4

GPT-4 Technical Report

OpenAI*

Abstract

We report the development of GPT-4, a large-scale, multimodal model which can accept image and text inputs and produce text outputs. While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test takers. GPT-4 is a Transformer-based model pre-trained to predict the next token in a document. The post-training alignment process results in improved performance on measures of factuality and adherence to desired behavior. A core component of this project was developing infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to accurately predict some aspects of GPT-4's performance based on models trained with no more than 1/1,000th the compute of GPT-4.

https://openai.com/research/gpt-4 (March 2023)

GPT-4: Technical details

2 Scope and Limitations of this Technical Report

This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model [39] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [40]. Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.

We are committed to independent auditing of our technologies, and shared some initial steps and ideas in this area in the system card accompanying this release.[2] We plan to make further technical details available to additional third parties who can advise us on how to weigh the competitive and safety considerations above against the scientific value of further transparency.

(Rumor: it has one trillion parameters)

GPT-4: Performance

	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot	LM SOTA Best external LM evaluated few-shot	SOTA Best external model (incl. benchmark-specific tuning)
MMLU [49] Multiple-choice questions in 57 subjects (professional & academic)	**86.4%** 5-shot	70.0% 5-shot	70.7% 5-shot U-PaLM [50]	75.2% 5-shot Flan-PaLM [51]
HellaSwag [52] Commonsense reasoning around everyday events	**95.3%** 10-shot	85.5% 10-shot	84.2% LLaMA (validation set) [28]	85.6 ALUM [53]
AI2 Reasoning Challenge (ARC) [54] Grade-school multiple choice science questions. Challenge-set.	**96.3%** 25-shot	85.2% 25-shot	85.2% 8-shot PaLM [55]	86.5% ST-MOE [18]
WinoGrande [56] Commonsense reasoning around pronoun resolution	**87.5%** 5-shot	81.6% 5-shot	85.1% 5-shot PaLM [3]	85.1% 5-shot PaLM [3]
HumanEval [43] Python coding tasks	**67.0%** 0-shot	48.1% 0-shot	26.2% 0-shot PaLM [3]	65.8% CodeT + GPT-3.5 [57]
DROP [58] (F1 score) Reading comprehension & arithmetic.	80.9 3-shot	64.1 3-shot	70.8 1-shot PaLM [3]	**88.4** QDGAT [59]
GSM-8K [60] Grade-school mathematics questions	**92.0%*** 5-shot chain-of-thought	57.1% 5-shot	58.8% 8-shot Minerva [61]	87.3% Chinchilla + SFT+ORM-RL, ORM reranking [62]

Table 2. Performance of GPT-4 on academic benchmarks. We compare GPT-4 alongside the best SOTA (with benchmark-specific training) and the best SOTA for an LM evaluated few-shot. GPT-4 outperforms existing LMs on all benchmarks, and beats SOTA with benchmark-specific training on all datasets except DROP. For each task we report GPT-4's performance along with the few-shot method used to evaluate. For GSM-8K, we included part of the training set in the GPT-4 pre-training mix (see Appendix E), and we use chain-of-thought prompting [11] when evaluating. For multiple-choice questions, we present all answers (ABCD) to the model and ask it to choose the letter of the answer, similarly to how a human would solve such a problem.

GPT-4: Performance

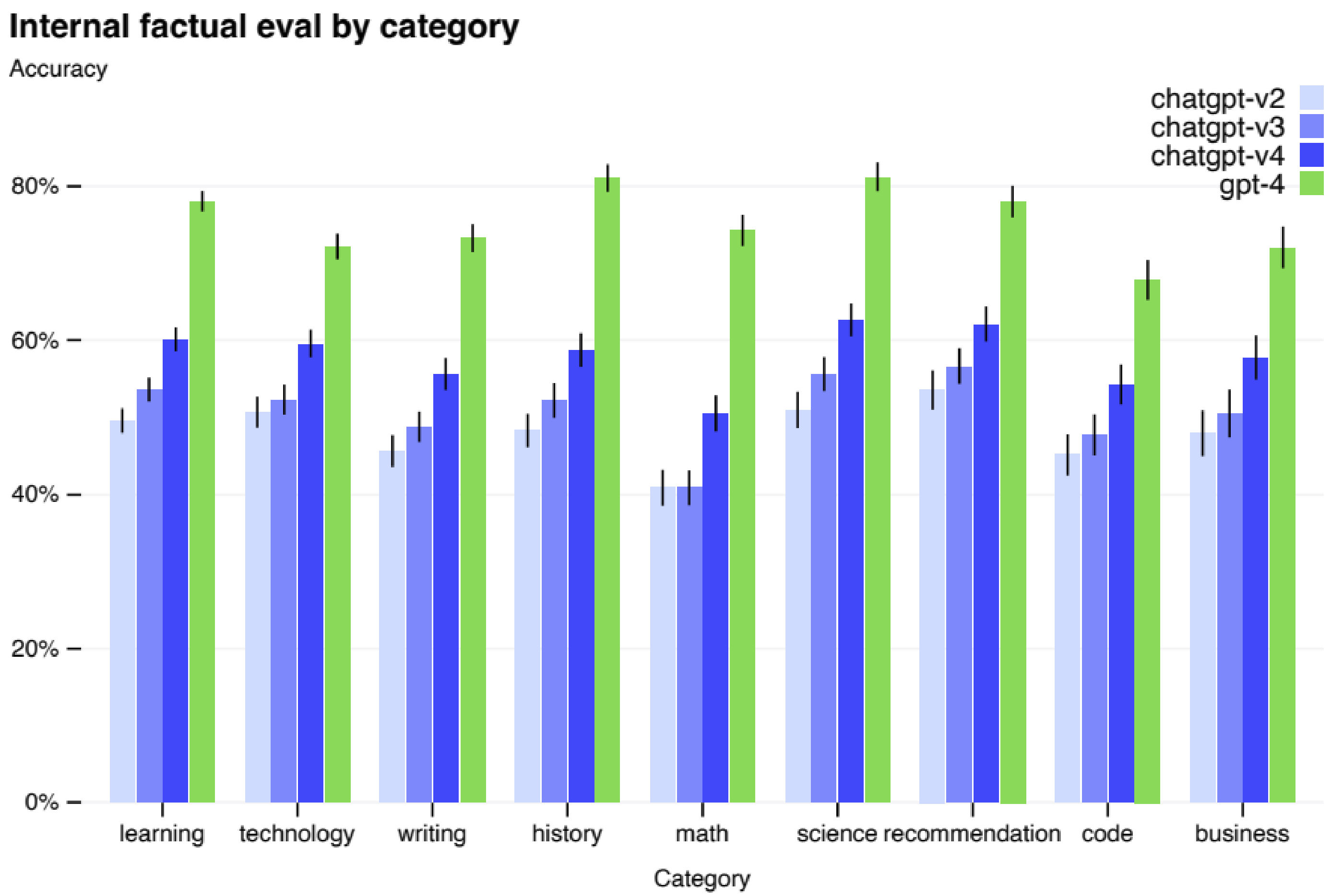


Figure 6. Performance of GPT-4 on nine internal adversarially-designed factuality evaluations. Accuracy is shown on the y-axis, higher is better. An accuracy of 1.0 means the model's answers are judged to be in agreement with human ideal responses for all questions in the eval. We compare GPT-4 to three earlier versions of ChatGPT [64] based on GPT-3.5; GPT-4 improves on the latest GPT-3.5 model by 19 percentage points, with significant gains across all topics.

GPT-4: Performance

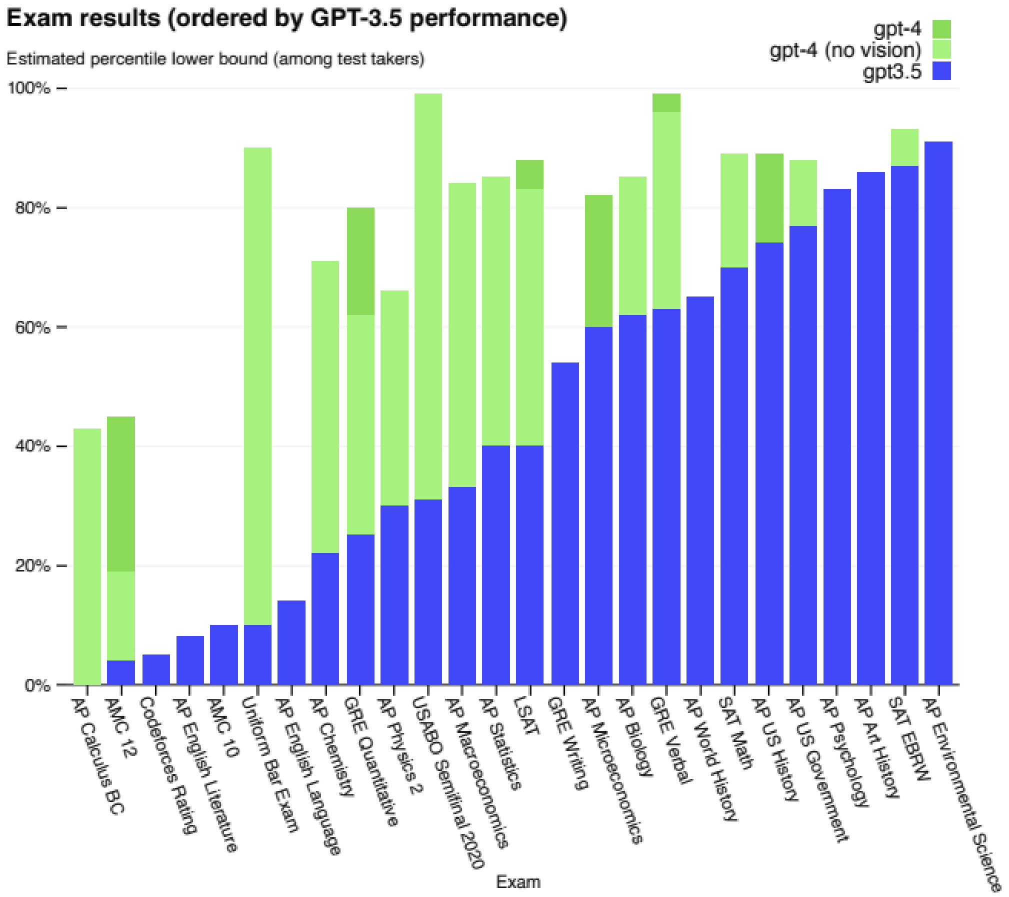


Figure 4. GPT performance on academic and professional exams. In each case, we simulate the conditions and scoring of the real exam. Exams are ordered from low to high based on GPT-3.5 performance. GPT-4 outperforms GPT-3.5 on most exams tested. To be conservative we report the lower end of the range of percentiles, but this creates some artifacts on the AP exams which have very wide scoring bins. For example although GPT-4 attains the highest possible score on AP Biology (5/5), this is only shown in the plot as 85th percentile because 15 percent of test-takers achieve that score.

For objections, see https://aisnakeoil.substack.com/p/gpt-4-and-professional-benchmarks

GPT-4: Visual input

User Can you explain this meme?

GPT-4 This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets.
The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world.
The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.

User What is unusual about this image?

Source: `https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg`

GPT-4 The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

AGI or hype?

北京大學 PEKING UNIVERSITY

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

Abstract

Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4 [Ope23], was trained using an unprecedented scale of compute and data. In this paper, we report on our investigation of an early version of GPT-4, when it was still in active development by OpenAI. We contend that (this early version of) GPT-4 is part of a new cohort of LLMs (along with ChatGPT and Google's PaLM for example) that exhibit more general intelligence than previous AI models. We discuss the rising capabilities and implications of these models. We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system. In our exploration of GPT-4, we put special emphasis on discovering its limitations, and we discuss the challenges ahead for advancing towards deeper and more comprehensive versions of AGI, including the possible need for pursuing a new paradigm that moves beyond next-word prediction. We conclude with reflections on societal influences of the recent technological leap and future research directions.

S. Bubeck et al. [Sparks of artificial general intelligence: Early experiments with GPT-4](). arXiv 2023

LLM critiques: Stochastic parrots or sentient entities?*

*Asking either question will get you fired from Google

MIT Technology Review

Artificial intelligence / Machine learning

We read the paper that forced Timnit Gebru out of Google. Here's what it says.

The company's star ethics researcher highlighted the risks of large language models, which are key to Google's business.

by **Karen Hao** December 4, 2020

COURTESY OF TIMNIT GEBRU

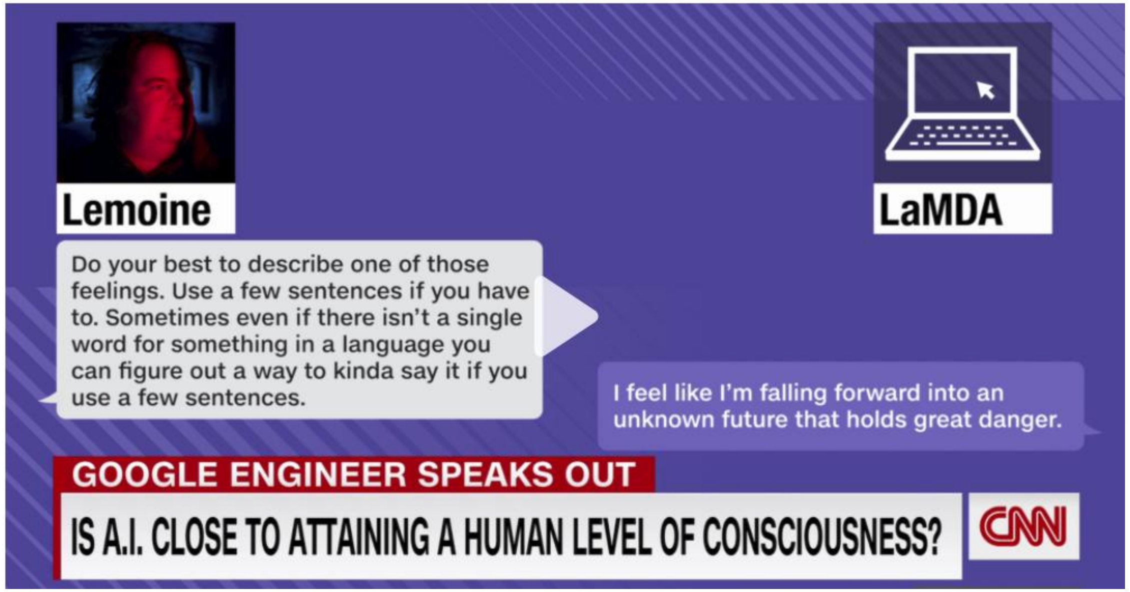
https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

E. Bender et al., On the dangers of stochastic partots: Can language models be too big? FAccT 2021

Google fires engineer who contended its AI technology was sentient

By Ramishah Maruf, CNN

Updated 1:45 PM EDT, Mon July 25, 2022



https://www.cnn.com/2022/07/23/business/google-ai-engineer-fired-sentient/index.html

More LLM concerns

- Bias and toxicity
- Hallucination
- Leakage of private information
- Exploitation of crowd workers, users
- Access and reproducibility
- Carbon footprint
- Potential for purposeful misuse (e.g., misinformation generation)
- Potential for destroying jobs (e.g., writers, editors, programmers, teachers, academics)
- All that AGI stuff…

谢谢

北京大學

PEKING UNIVERSITY