UiT The Arctic
University of Norway

BIO-AI LAB | ARCTIC LLM WORKSHOP 2023

# Large Language Models

## Day 2 - Session 4

## Evolution of Foundation LLM models

**Abhinanda Ranjit Punnakkal**

Abhinanda.r.punnakkal@uit.no

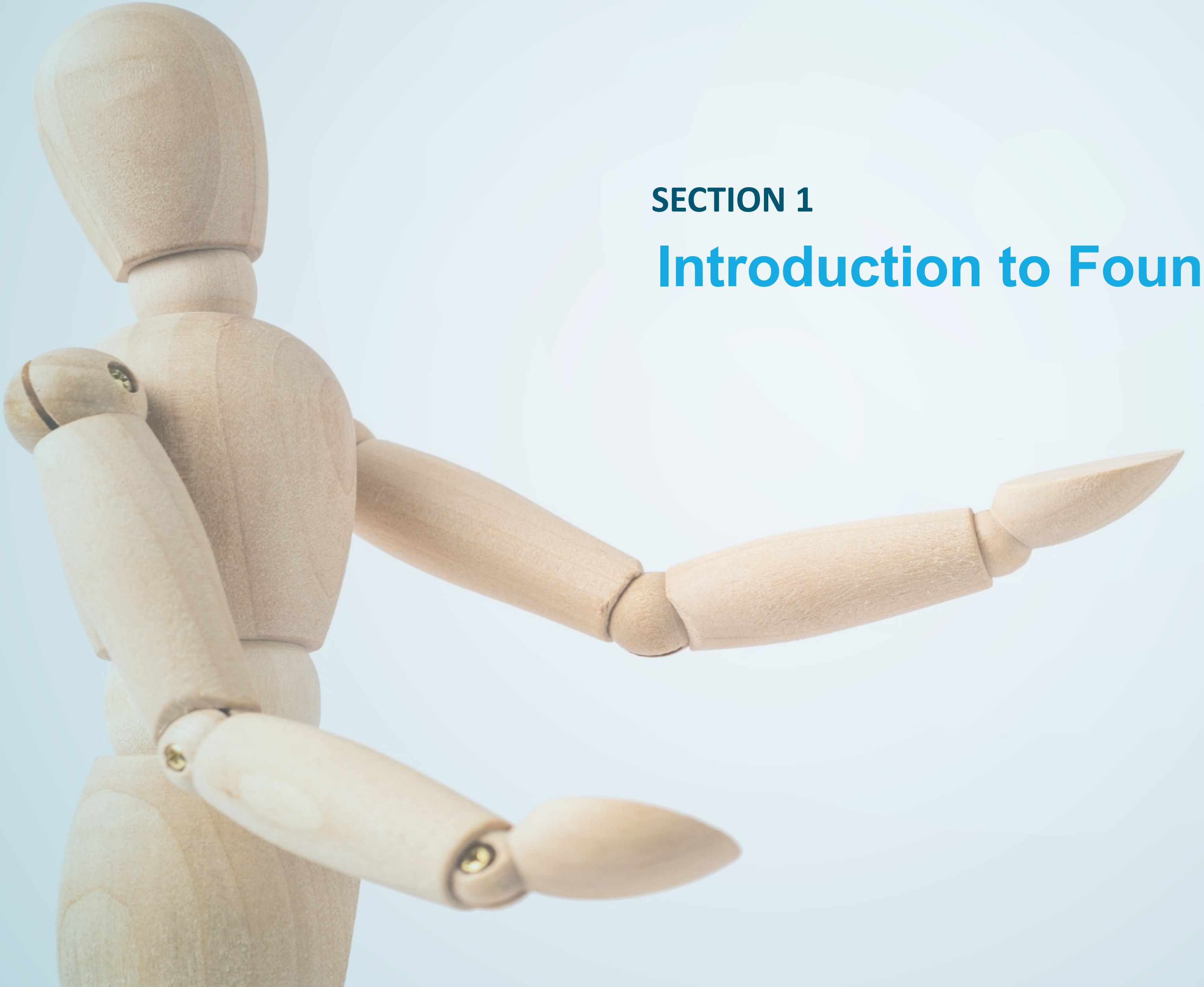
27. Oct 2023

# Outline

- *Introduction to Foundational Models*
- *Evolution of Foundation LLM Models*
- *Stats and Benchmarks*
- *Conclusion*

SECTION 1

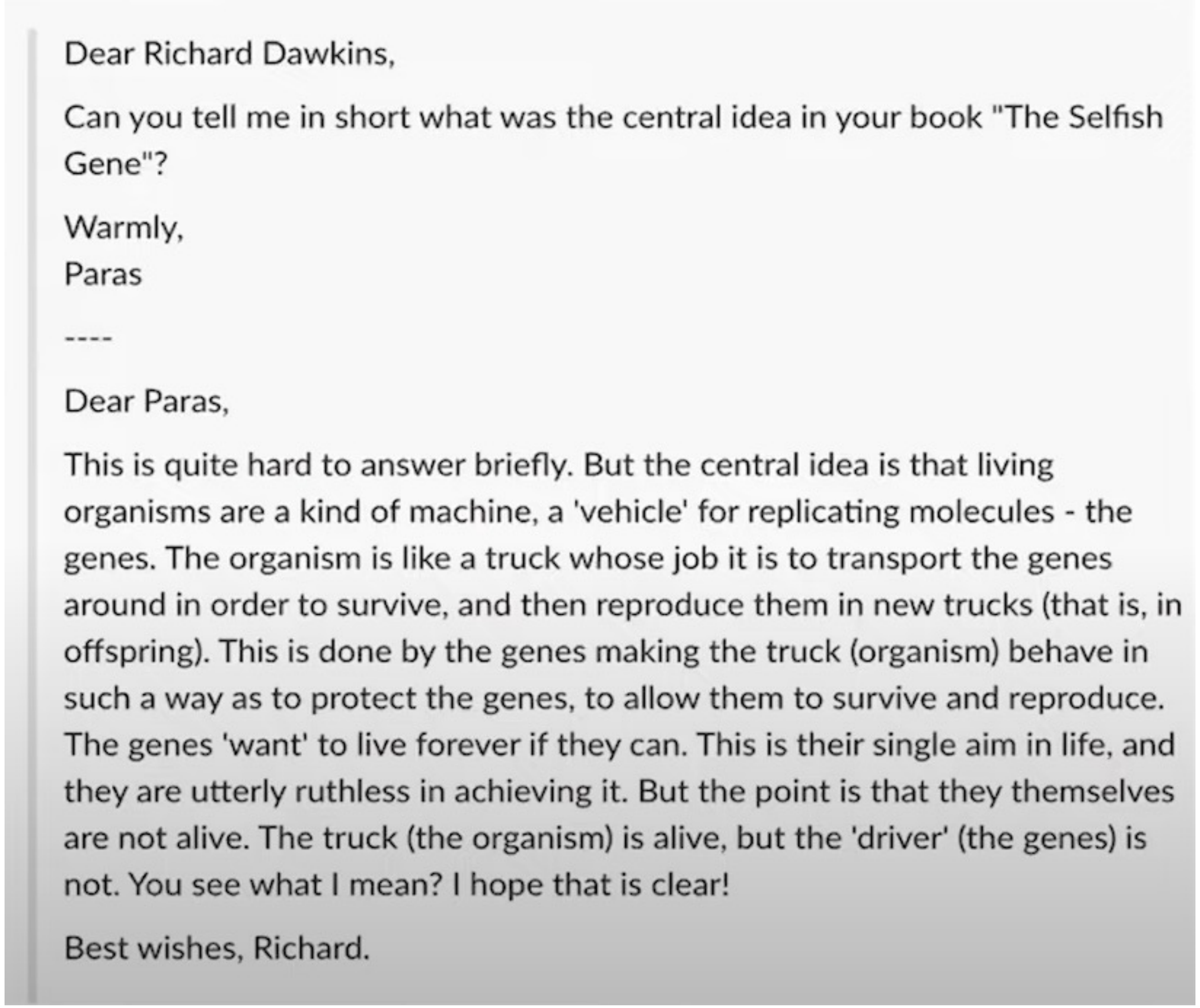
# Introduction to Foundational Models



# What is a Foundation Model?

*Examples:*

- *Ask open-ended questions.*
- *Paraphrase*
- *Generate code, web code*
- *Generate cool images*
- *Generating videos*
- *Robot manipulation*



Dear Richard Dawkins,

Can you tell me in short what was the central idea in your book "The Selfish Gene"?

Warmly,
Paras

----

Dear Paras,

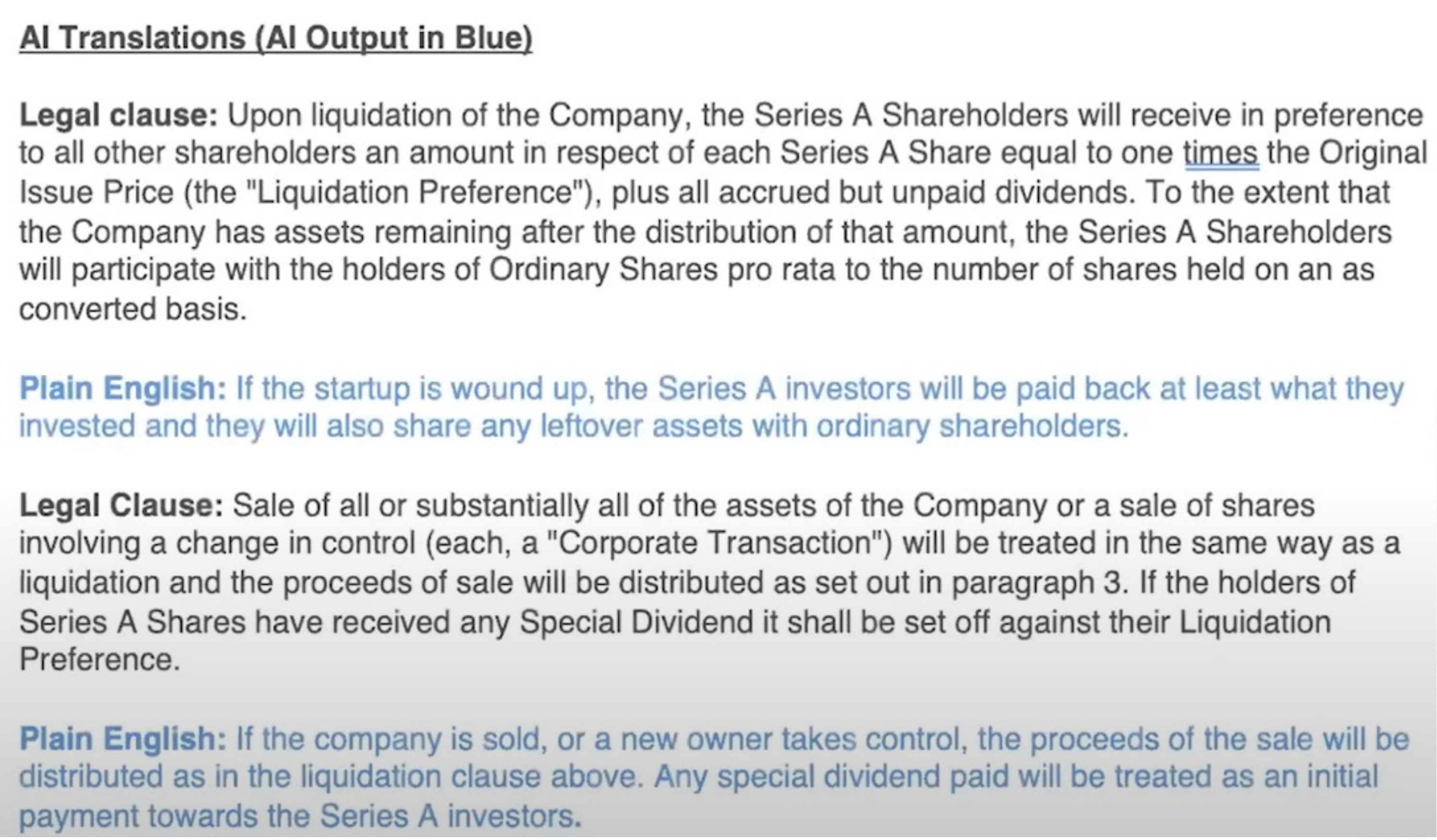
This is quite hard to answer briefly. But the central idea is that living organisms are a kind of machine, a 'vehicle' for replicating molecules - the genes. The organism is like a truck whose job it is to transport the genes around in order to survive, and then reproduce them in new trucks (that is, in offspring). This is done by the genes making the truck (organism) behave in such a way as to protect the genes, to allow them to survive and reproduce. The genes 'want' to live forever if they can. This is their single aim in life, and they are utterly ruthless in achieving it. But the point is that they themselves are not alive. The truck (the organism) is alive, but the 'driver' (the genes) is not. You see what I mean? I hope that is clear!

Best wishes, Richard.

# What is a Foundation Model?

*Examples:*

- *Ask open-ended questions.*
- *Paraphrase*
- *Generate code, web code*
- *Generate cool images*
- *Generating videos*
- *Robot manipulation*



**AI Translations (AI Output in Blue)**

**Legal clause:** Upon liquidation of the Company, the Series A Shareholders will receive in preference to all other shareholders an amount in respect of each Series A Share equal to one times the Original Issue Price (the "Liquidation Preference"), plus all accrued but unpaid dividends. To the extent that the Company has assets remaining after the distribution of that amount, the Series A Shareholders will participate with the holders of Ordinary Shares pro rata to the number of shares held on an as converted basis.

**Plain English:** If the startup is wound up, the Series A investors will be paid back at least what they invested and they will also share any leftover assets with ordinary shareholders.

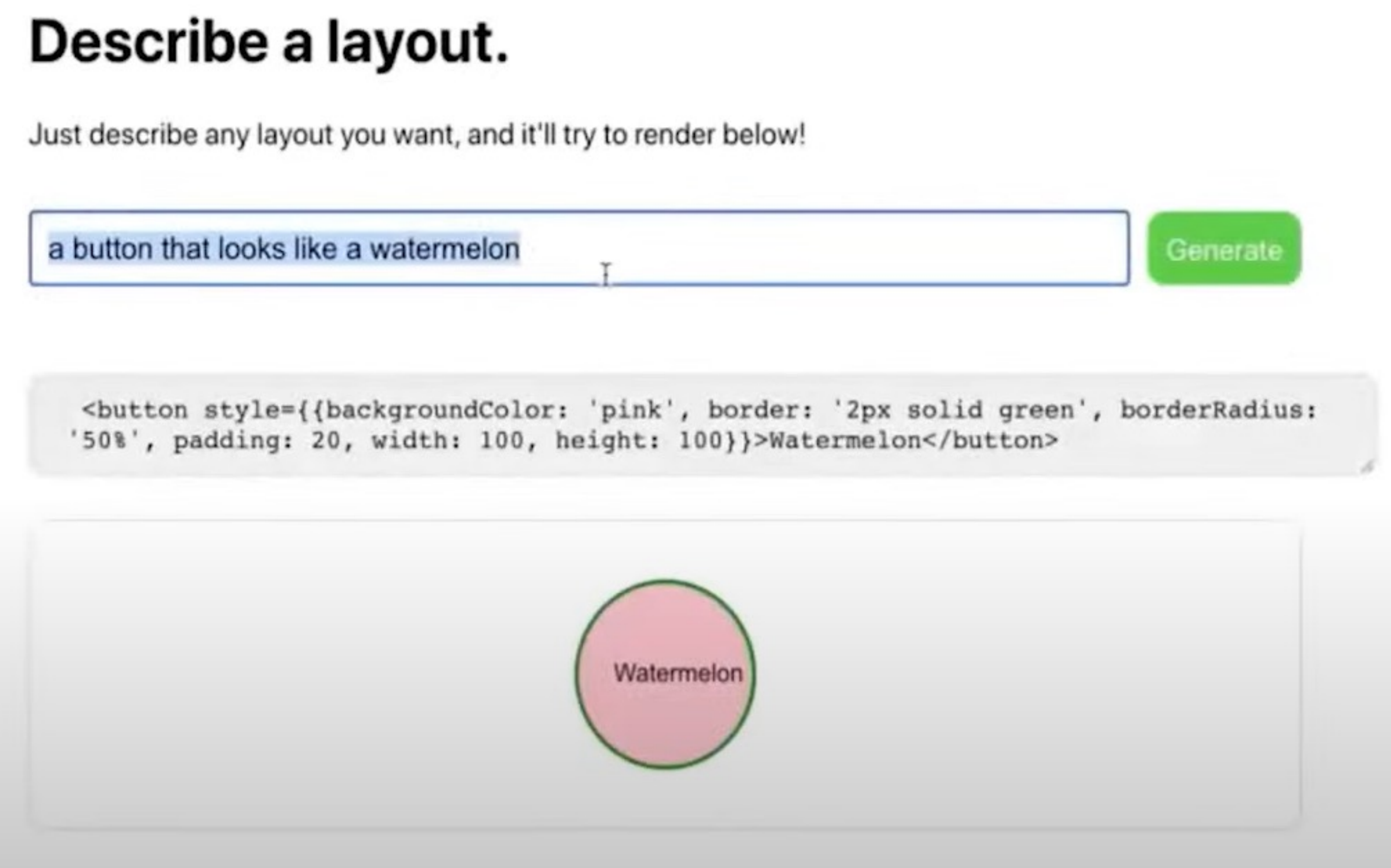
**Legal Clause:** Sale of all or substantially all of the assets of the Company or a sale of shares involving a change in control (each, a "Corporate Transaction") will be treated in the same way as a liquidation and the proceeds of sale will be distributed as set out in paragraph 3. If the holders of Series A Shares have received any Special Dividend it shall be set off against their Liquidation Preference.

**Plain English:** If the company is sold, or a new owner takes control, the proceeds of the sale will be distributed as in the liquidation clause above. Any special dividend paid will be treated as an initial payment towards the Series A investors.

# What is a Foundation Model?

*Examples:*

- *Ask open-ended questions.*
- *Paraphrase*
- *Generate code, web code*
- *Generate cool images*
- *Generating videos*
- *Robot manipulation*



## Describe a layout.

Just describe any layout you want, and it'll try to render below!

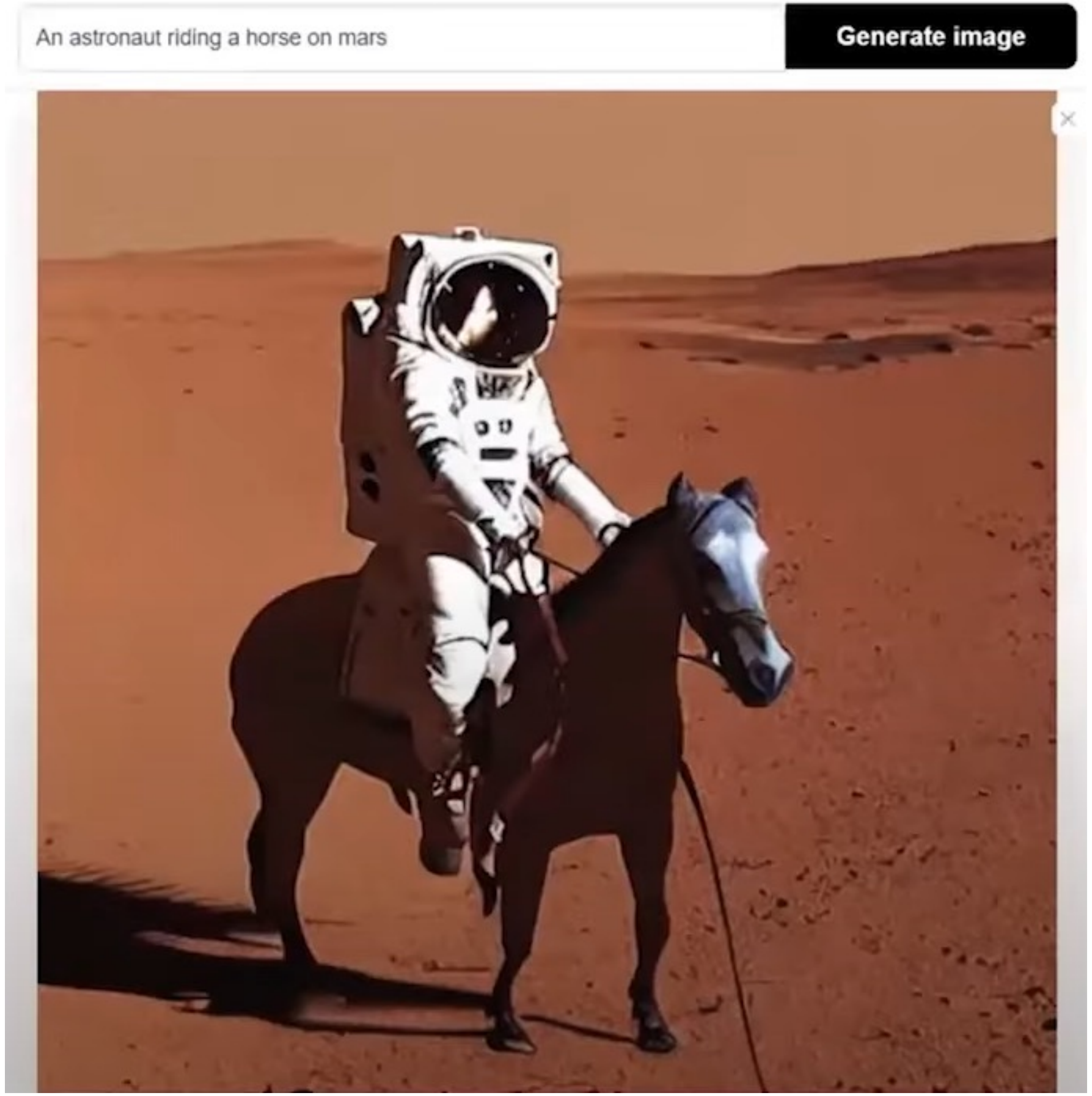
a button that looks like a watermelon

```
 <button style={{backgroundColor: 'pink', border: '2px solid green', borderRadius:
'50%', padding: 20, width: 100, height: 100}}>Watermelon</button>
```

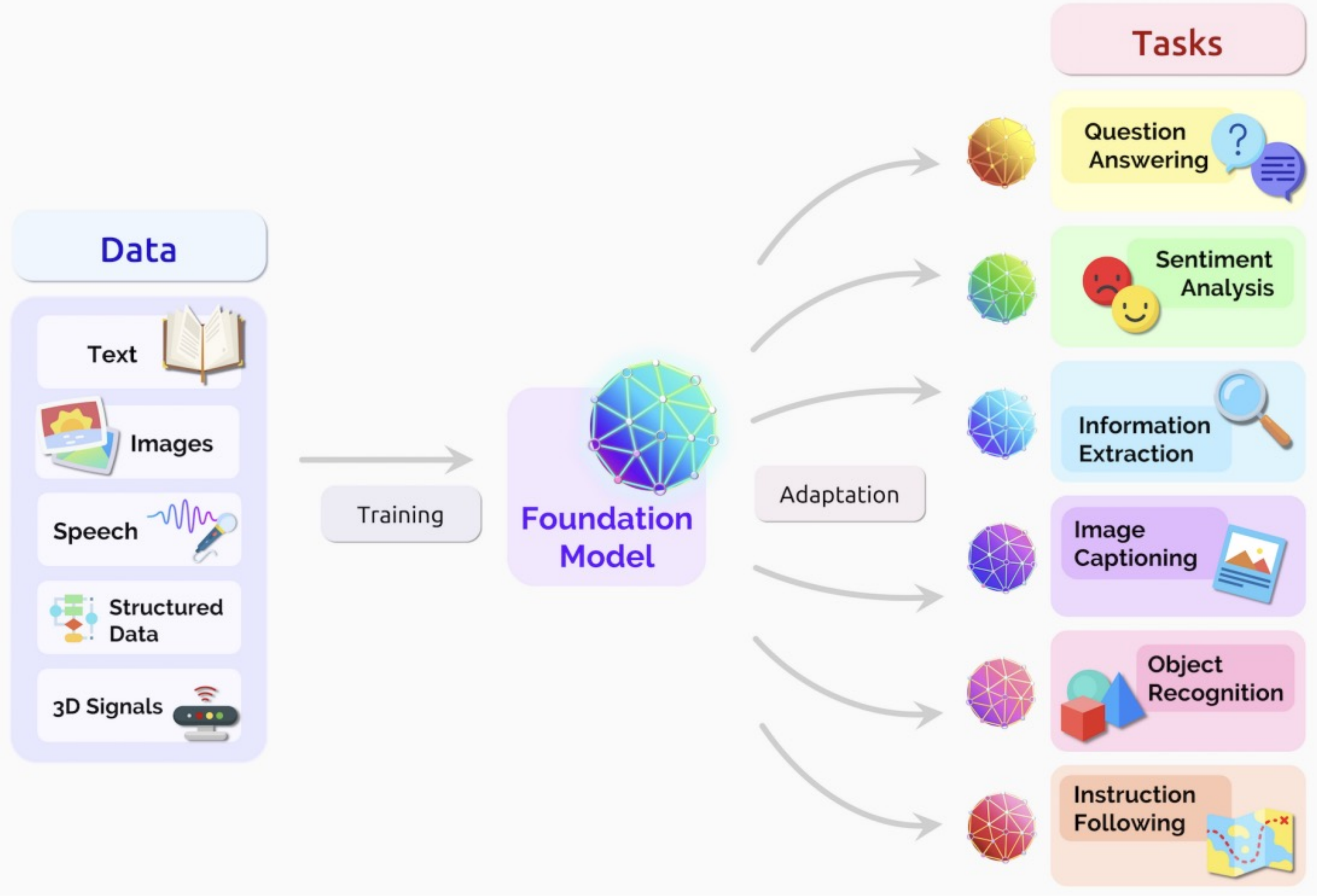
# What is a Foundation Model?

*Examples:*

- *Ask open-ended questions.*
- *Paraphrase*
- *Generate code, web code*
- *Generate cool images*
- *Generating videos*
- *Robot manipulation*


An astronaut riding a horse on mars

# What is a Foundation Model



Source: Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

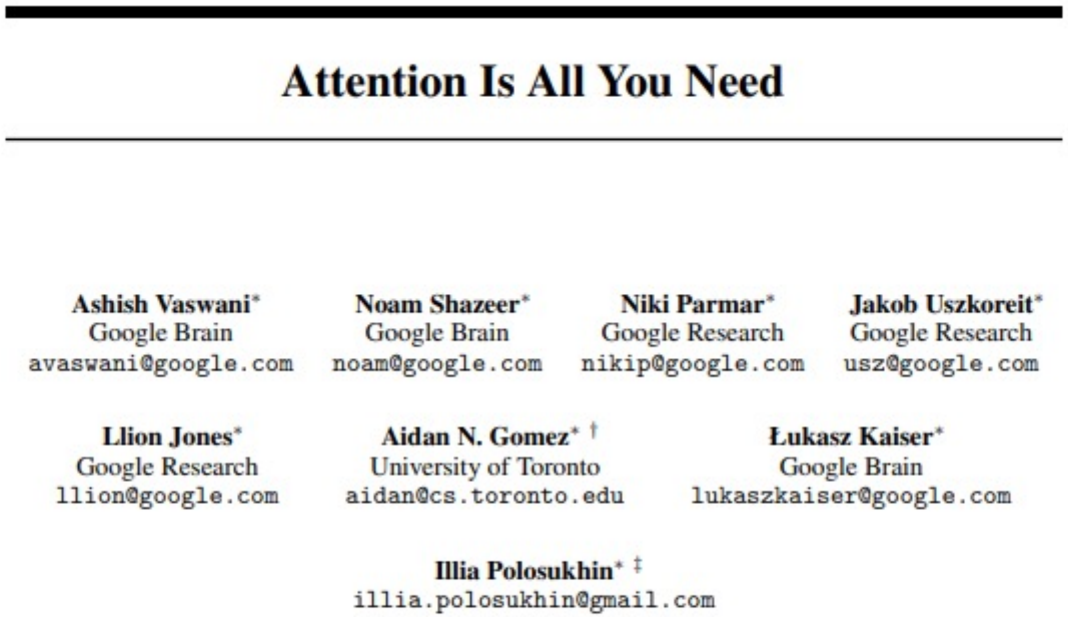
# Success of a Foundation Model

1. *Transformer architecture*
2. *Scale*
3. *In context learning/Self-supervised training*

Source: Zhou, Ce, et al. "A comprehensive survey on pretrained foundation models: A history from bert to chatgpt." *arXiv preprint arXiv:2302.09419* (2023).

# 1. Transformers: Rise of the Beasts (2017)

*Efficient architecture*



## Attention Is All You Need

**Ashish Vaswani***
Google Brain
avaswani@google.com

**Noam Shazeer***
Google Brain
noam@google.com

**Niki Parmar***
Google Research
nikip@google.com

**Jakob Uszkoreit***
Google Research
usz@google.com

**Llion Jones***
Google Research
llion@google.com

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

**Łukasz Kaiser***
Google Brain
lukaszkaiser@google.com

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

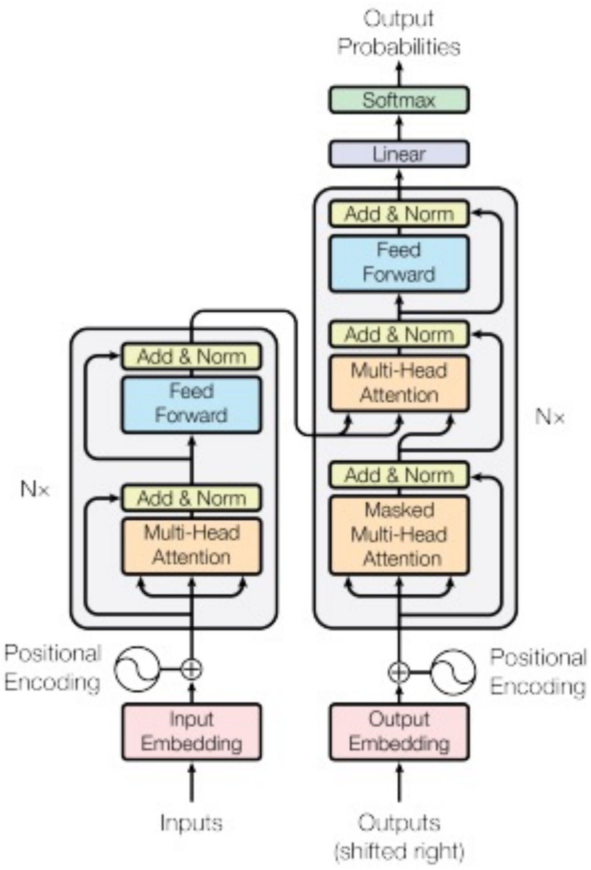
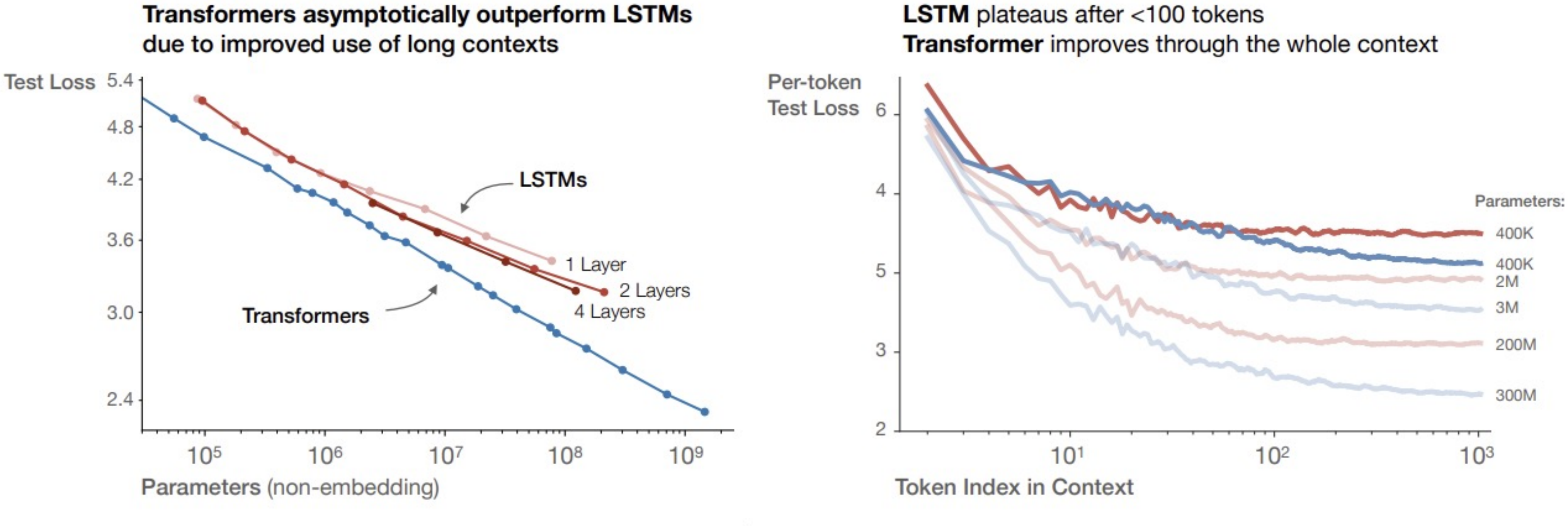


Figure 1: The Transformer - model architecture.

# LSTMs vs Transformers



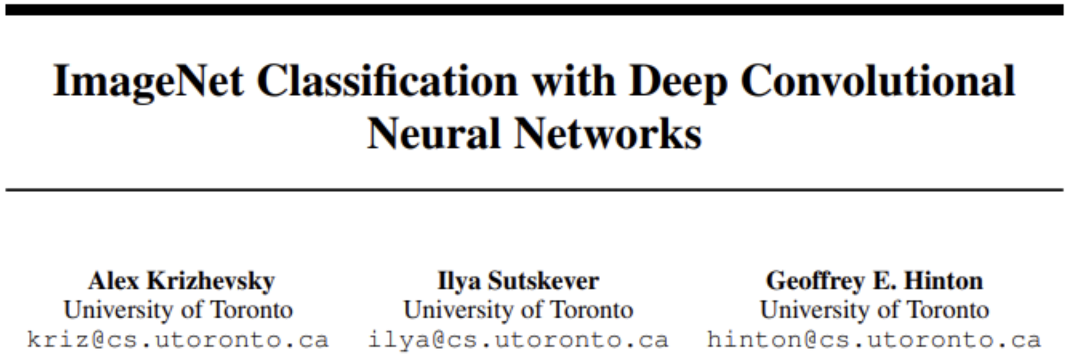
**Transformers asymptotically outperform LSTMs due to improved use of long contexts**

**LSTM** plateaus after <100 tokens
**Transformer** improves through the whole context

Source: Kaplan, Jared, et al. "Scaling laws for neural language models." *arXiv preprint arXiv:2001.08361* (2020).

# Before Large Models *2012- 2020*

*CNNs ruled!*

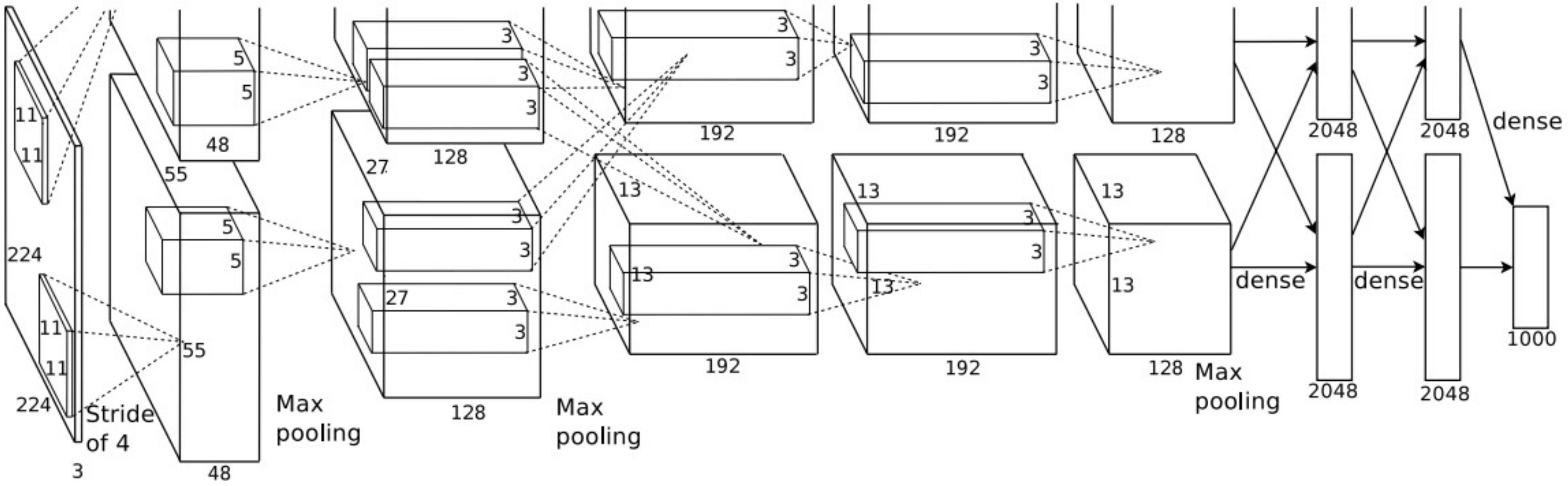


## ImageNet Classification with Deep Convolutional Neural Networks

**Alex Krizhevsky**
University of Toronto
kriz@cs.utoronto.ca

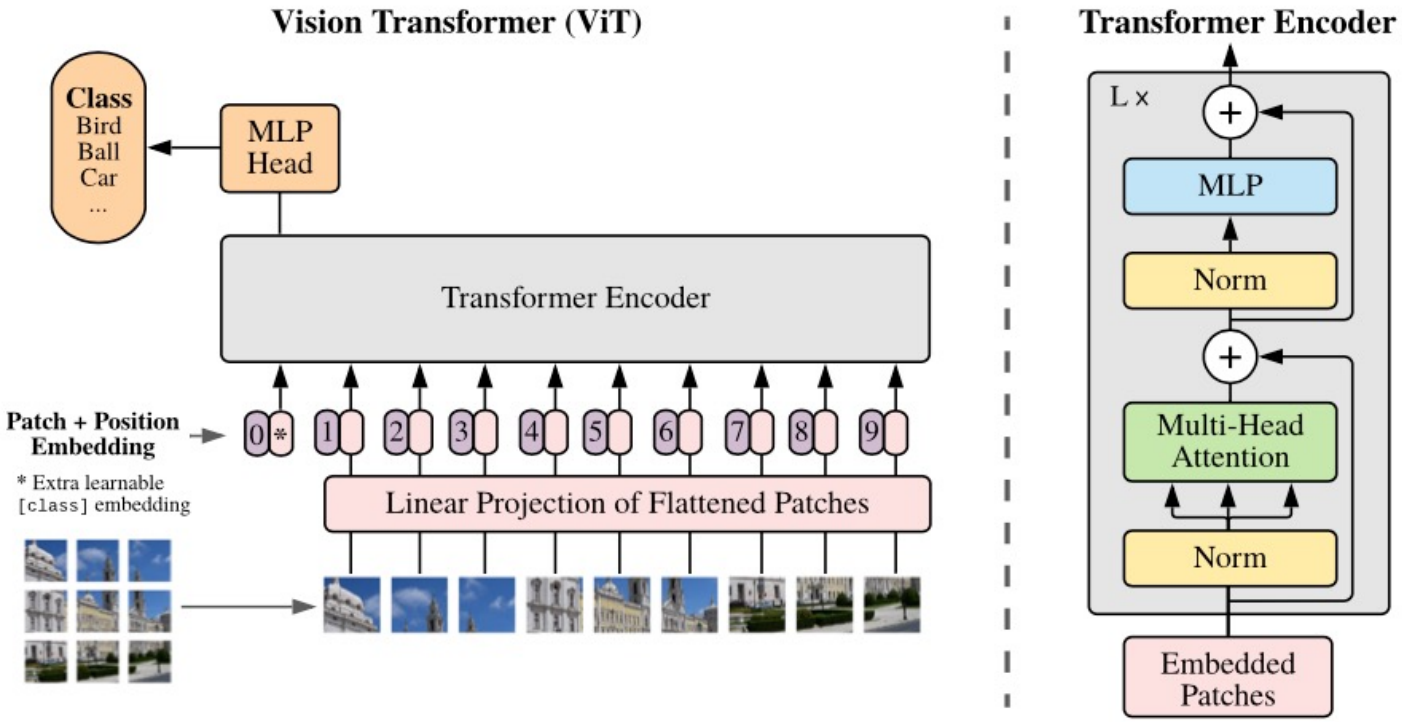
**Ilya Sutskever**
University of Toronto
ilya@cs.utoronto.ca

**Geoffrey E. Hinton**
University of Toronto
hinton@cs.utoronto.ca



# ViT

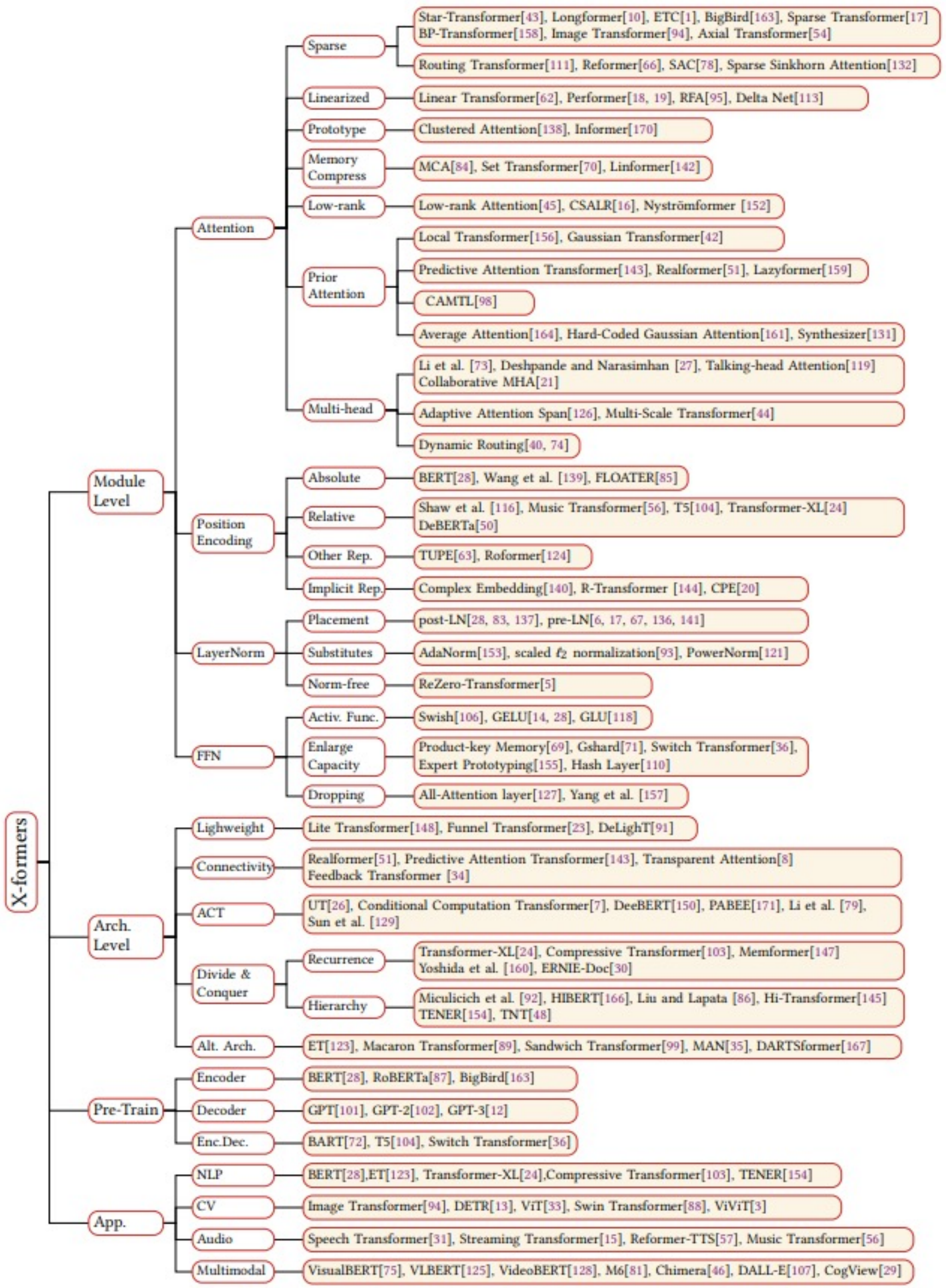
*No more Convolutions for Images!*



Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

# Variations of Transformer

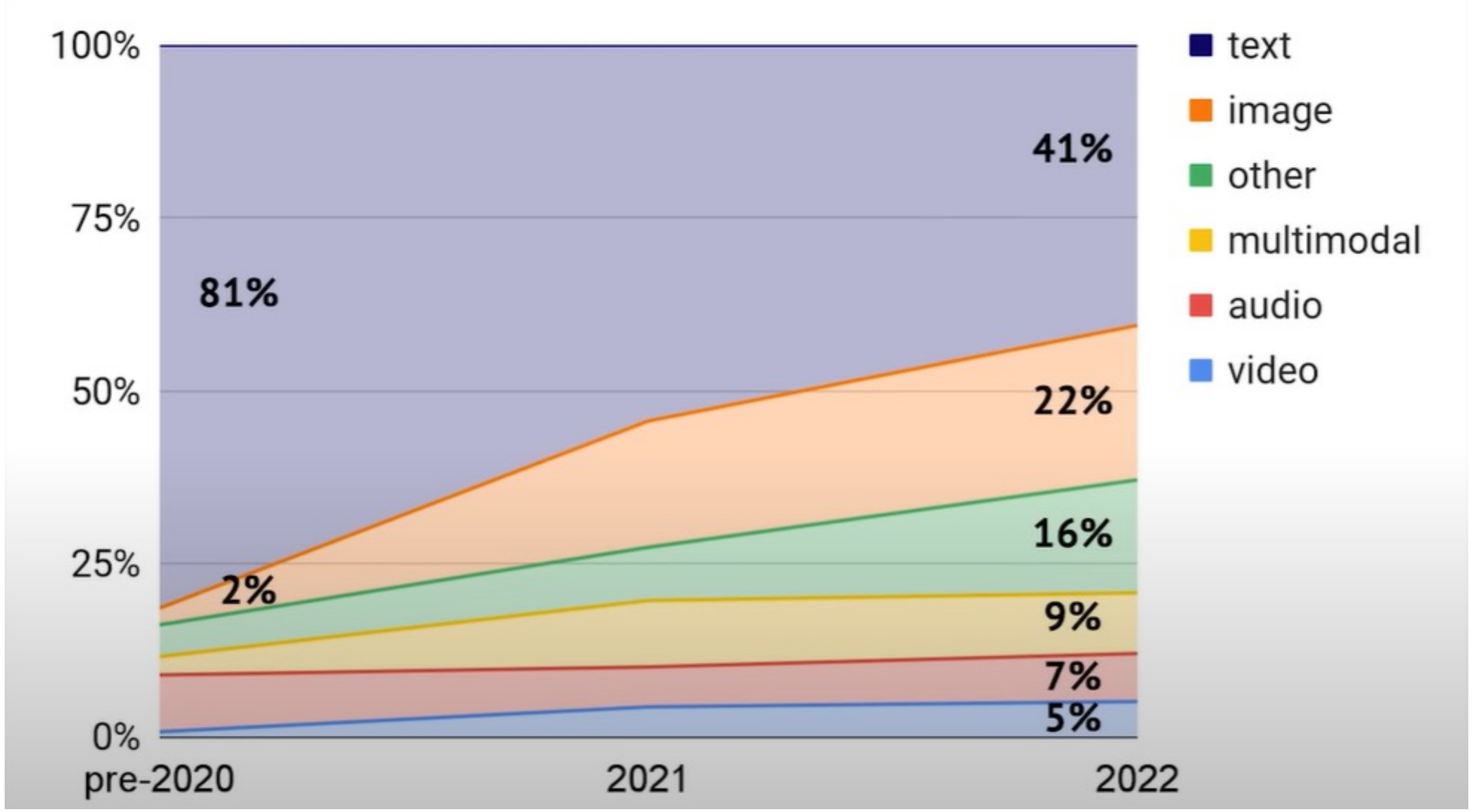
1. *Model efficiency*
2. *Model generalization*
3. *Model adaptation*


X-formers
Module Level
Attention
Sparse
Star-Transformer[43], Longformer[10], ETC[1], BigBird[163], Sparse Transformer[17] BP-Transformer[158], Image Transformer[94], Axial Transformer[54]
Routing Transformer[111], Reformer[66], SAC[78], Sparse Sinkhorn Attention[132]
Linearized
Linear Transformer[62], Performer[18, 19], RFA[95], Delta Net[113]
Prototype
Clustered Attention[138], Informer[170]
Memory Compress
MCA[84], Set Transformer[70], Linformer[142]
Low-rank
Low-rank Attention[45], CSALR[16], Nyströmformer [152]
Prior Attention
Local Transformer[156], Gaussian Transformer[42]
Predictive Attention Transformer[143], Realformer[51], Lazyformer[159]
CAMTL[98]
Average Attention[164], Hard-Coded Gaussian Attention[161], Synthesizer[131]
Multi-head
Li et al. [73], Deshpande and Narasimhan [27], Talking-head Attention[119] Collaborative MHA[21]
Adaptive Attention Span[126], Multi-Scale Transformer[44]
Dynamic Routing[40, 74]
Position Encoding
Absolute
BERT[28], Wang et al. [139], FLOATER[85]
Relative
Shaw et al. [116], Music Transformer[56], T5[104], Transformer-XL[24] DeBERTa[50]
Other Rep.
TUPE[63], Roformer[124]
Implicit Rep.
Complex Embedding[140], R-Transformer [144], CPE[20]
LayerNorm
Placement
post-LN[28, 83, 137], pre-LN[6, 17, 67, 136, 141]
Substitutes
AdaNorm[153], scaled $\ell_2$ normalization[93], PowerNorm[121]
Norm-free
ReZero-Transformer[5]
FFN
Activ. Func.
Swish[106], GELU[14, 28], GLU[118]
Enlarge Capacity
Product-key Memory[69], Gshard[71], Switch Transformer[36], Expert Prototyping[155], Hash Layer[110]
Dropping
All-Attention layer[127], Yang et al. [157]
Arch. Level
Lighweight
Lite Transformer[148], Funnel Transformer[23], DeLighT[91]
Connectivity
Realformer[51], Predictive Attention Transformer[143], Transparent Attention[8] Feedback Transformer [34]
ACT
UT[26], Conditional Computation Transformer[7], DeeBERT[150], PABEE[171], Li et al. [79], Sun et al. [129]
Divide & Conquer
Recurrence
Transformer-XL[24], Compressive Transformer[103], Memformer[147] Yoshida et al. [160], ERNIE-Doc[30]
Hierarchy
Miculicich et al. [92], HIBERT[166], Liu and Lapata [86], Hi-Transformer[145] TENER[154], TNT[48]
Alt. Arch.
ET[123], Macaron Transformer[89], Sandwich Transformer[99], MAN[35], DARTSformer[167]
Pre-Train
Encoder
BERT[28], RoBERTa[87], BigBird[163]
Decoder
GPT[101], GPT-2[102], GPT-3[12]
Enc.Dec.
BART[72], T5[104], Switch Transformer[36]
App.
NLP
BERT[28],ET[123], Transformer-XL[24],Compressive Transformer[103], TENER[154]
CV
Image Transformer[94], DETR[13], ViT[33], Swin Transformer[88], ViViT[3]
Audio
Speech Transformer[31], Streaming Transformer[15], Reformer-TTS[57], Music Transformer[56]
Multimodal
VisualBERT[75], VLBERT[125], VideoBERT[128], M6[81], Chimera[46], DALL-E[107], CogView[29]

Source: Lin, Tianyang, et al. "A survey of transformers." *AI Open* (2022).

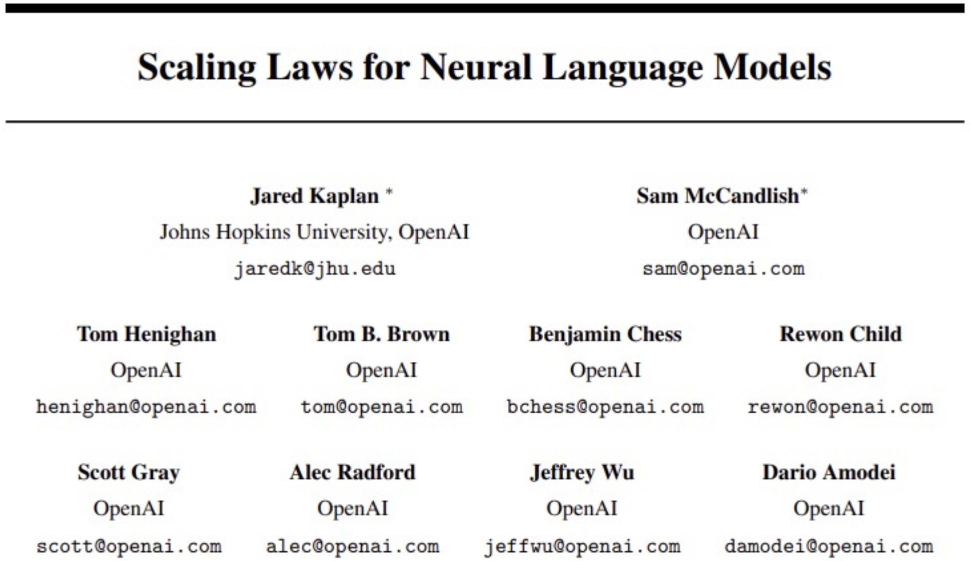
# Growth of Transformer related papers

*Started out with NLP*



# 2. Scale

- *model with number of parameters*
- *data*



**Scaling Laws for Neural Language Models**

**Jared Kaplan**[*]
Johns Hopkins University, OpenAI
jaredk@jhu.edu

**Sam McCandlish**[*]
OpenAI
sam@openai.com

**Tom Henighan**
OpenAI
henighan@openai.com

**Tom B. Brown**
OpenAI
tom@openai.com

**Benjamin Chess**
OpenAI
bchess@openai.com

**Rewon Child**
OpenAI
rewon@openai.com

**Scott Gray**
OpenAI
scott@openai.com

**Alec Radford**
OpenAI
alec@openai.com

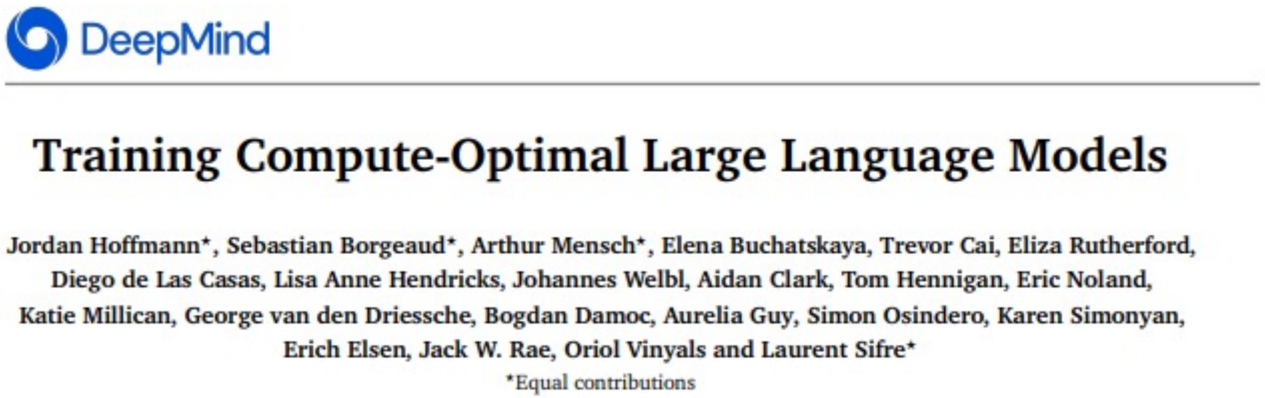
**Jeffrey Wu**
OpenAI
jeffwu@openai.com

**Dario Amodei**
OpenAI
damodei@openai.com

- non-embedding parameter count N,
- dataset size D
- optimized training computation Cmin

# 2. Scale

- *model with number of parameters*
- *data*



DeepMind

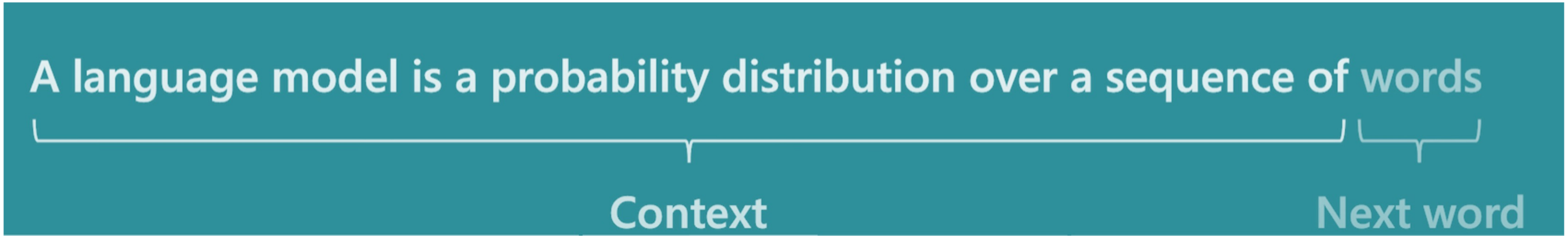
### Training Compute-Optimal Large Language Models

Jordan Hoffmann*, Sebastian Borgeaud*, Arthur Mensch*, Elena Buchatskaya, Trevor Cai, Eliza Rutherford,
Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland,
Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan,
Erich Elsen, Jack W. Rae, Oriol Vinyals and Laurent Sifre*
*Equal contributions

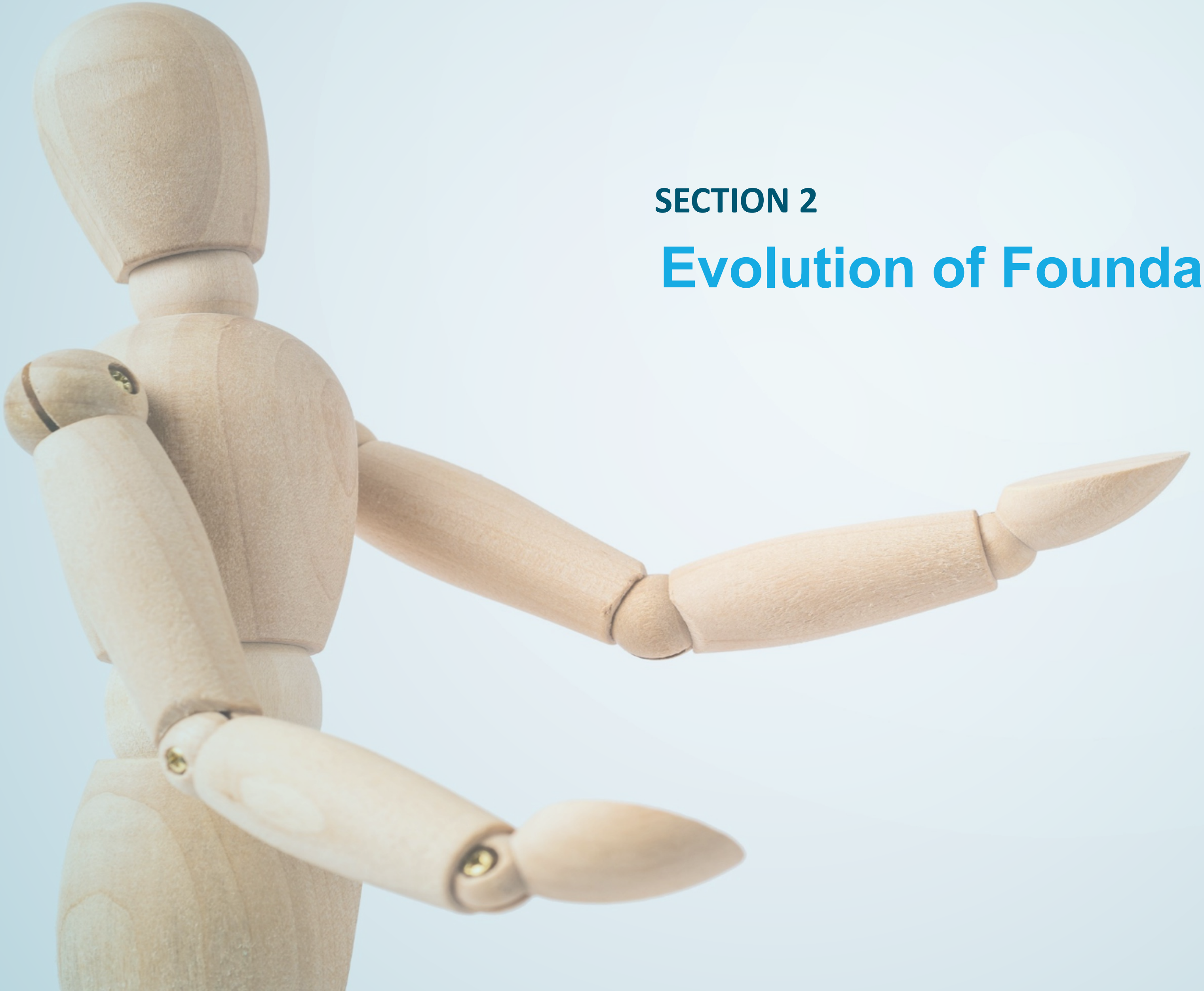
Chinchilla > Gopher
(with more training tokens) (with a larger model size)

# 3. Self Supervision

*Auto-regressive pretraining*
*LLMs are trained to predict the next of a large corpus*

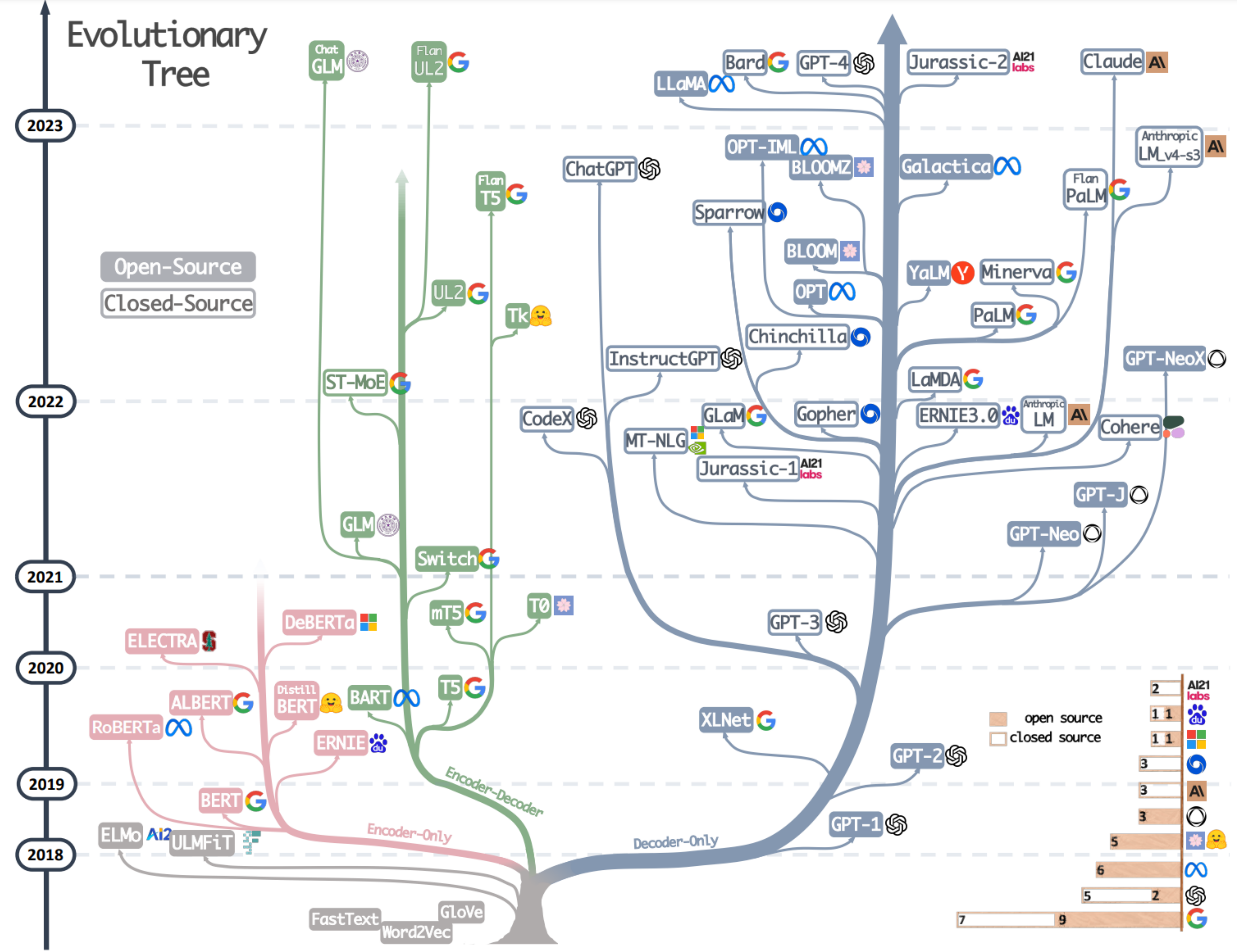


A language model is a probability distribution over a sequence of words



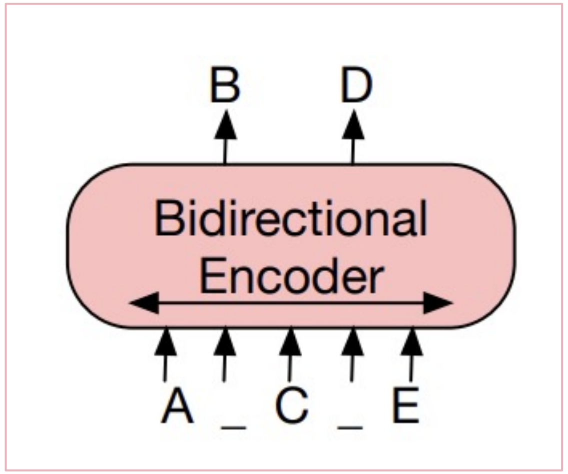
SECTION 2

# Evolution of Foundational LLM  Models

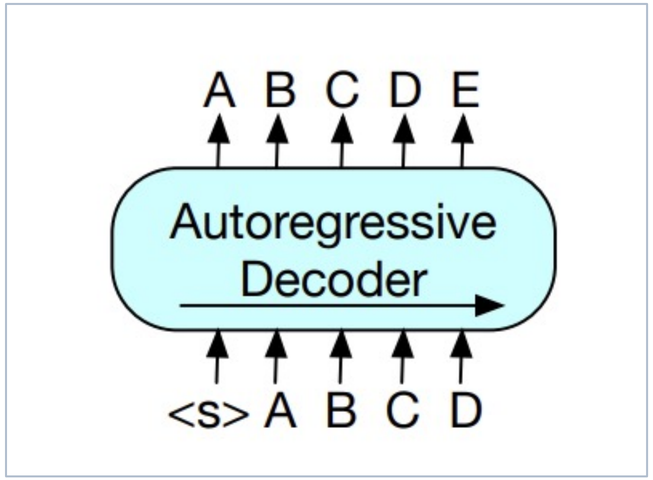


Yang, Jingfeng, et al. "Harnessing the power of llms in practice: A survey on chatgpt and beyond." *arXiv preprint arXiv:2304.13712* (2023).

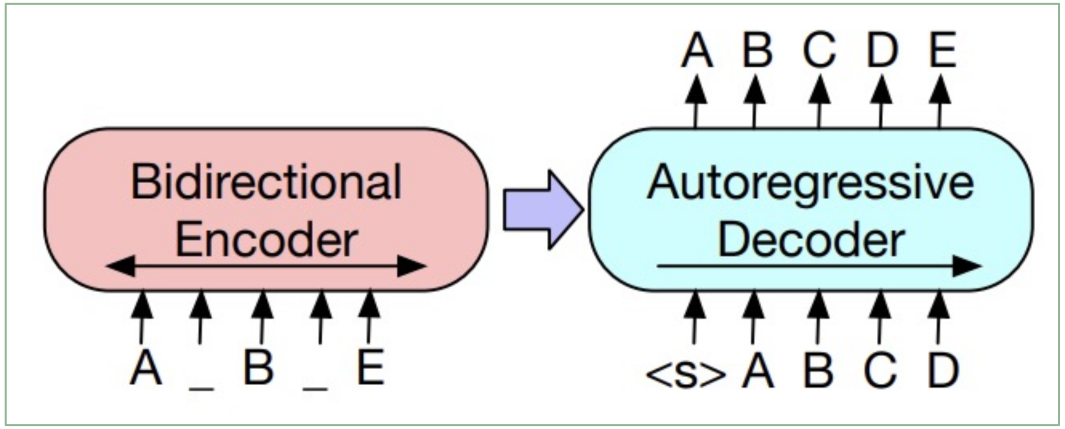
# Categories of LLMs



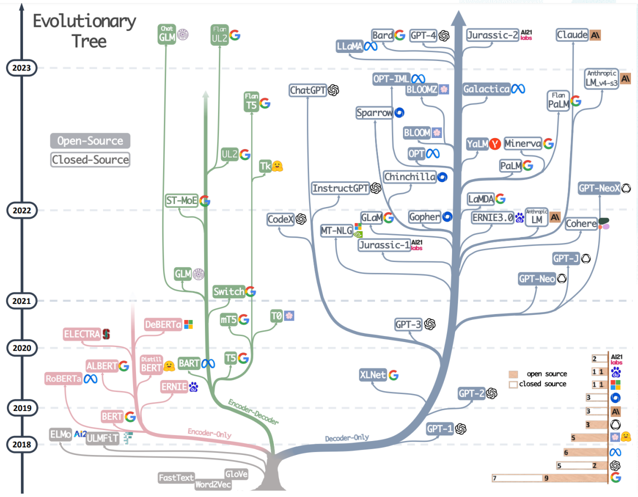
*Encoder only (BERT)*



*Decoder only (GPT)*

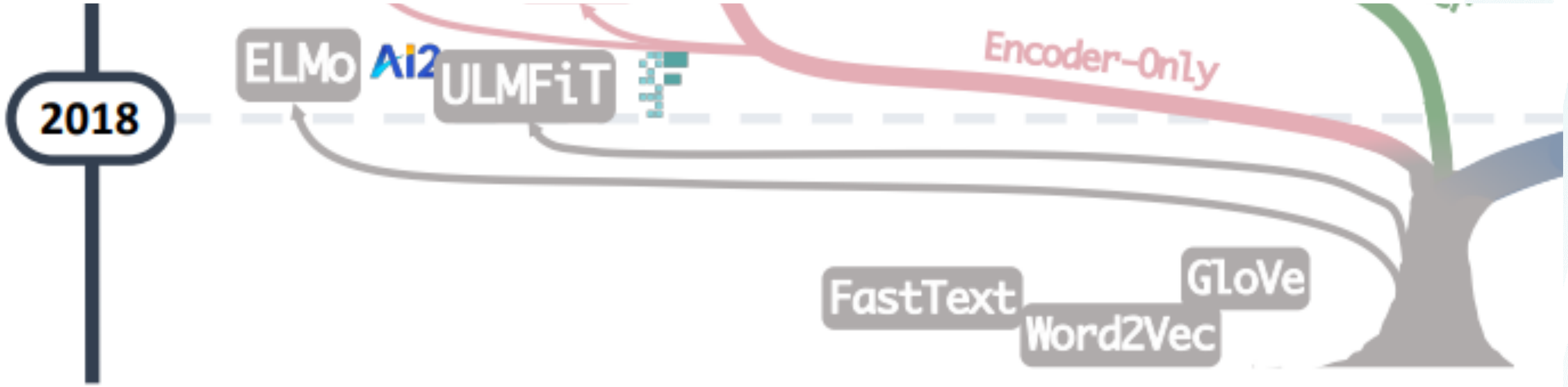


*Encoder-decoder (BART)*



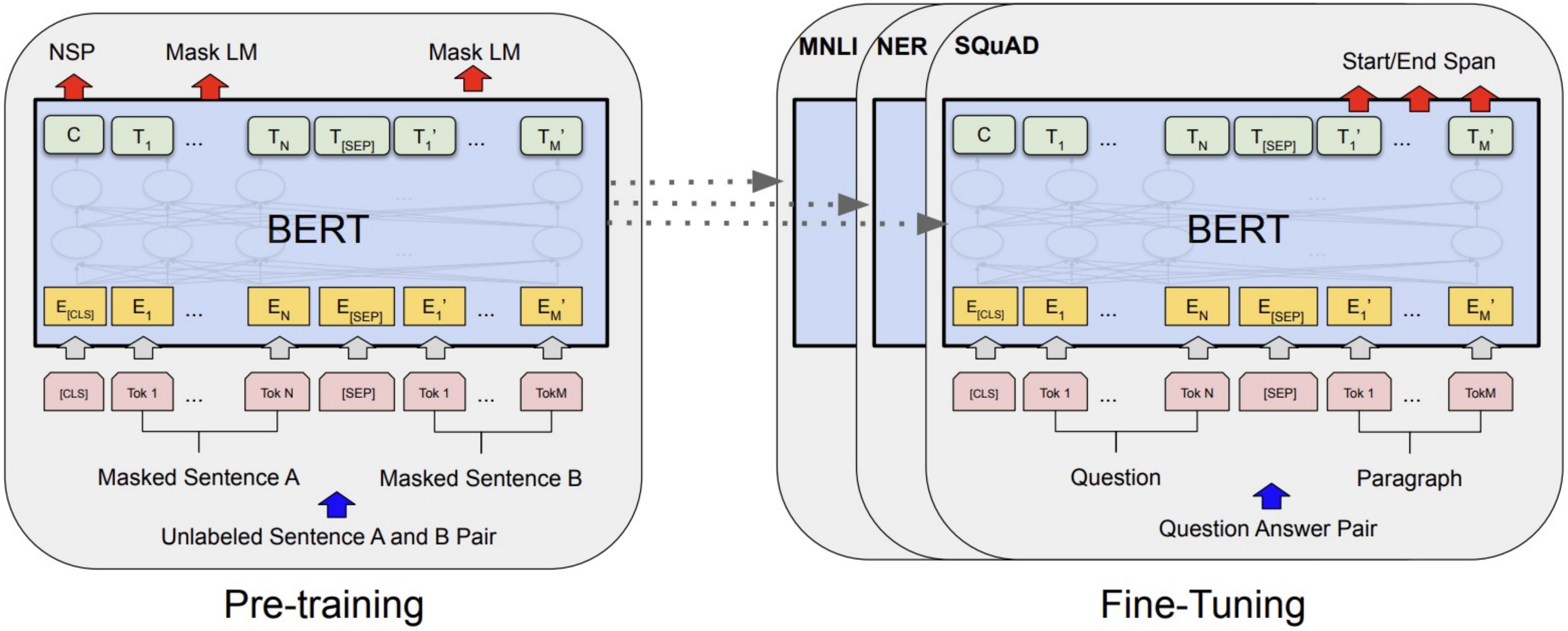
# Before LLMs

- *One-hot encoding words*
  - *Did not scale*
- *Static embedding*
  - *Limited to the bag of words*



# BERT - Bidirectional Encoder Representations from Transformers (2018)

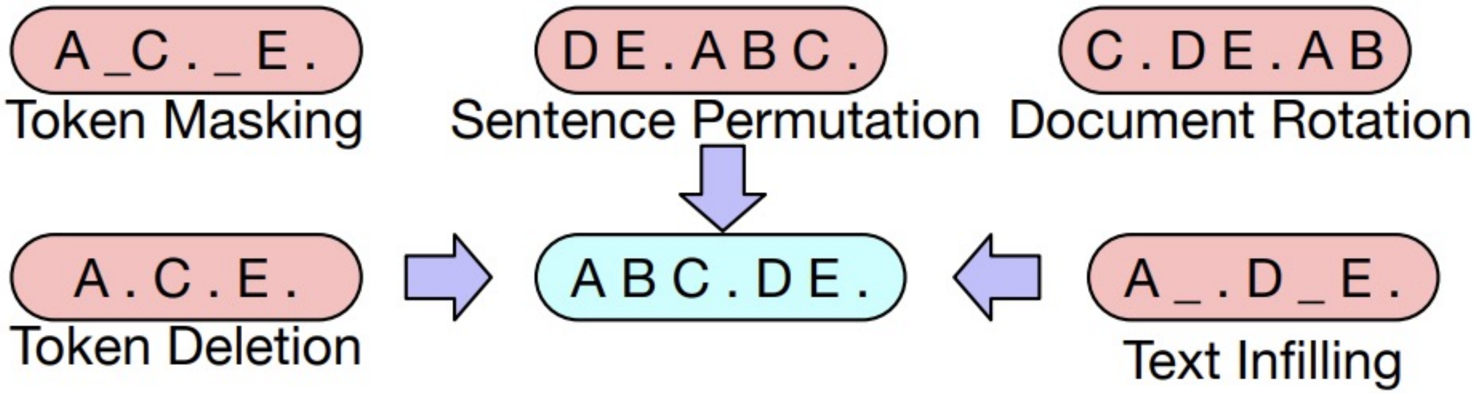
*learns bidirectional encoder representations from Transformers, which are trained on large datasets as contextual language models*



Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

# BART

*Encoder-Decoder*



A_C._E.
Token Masking

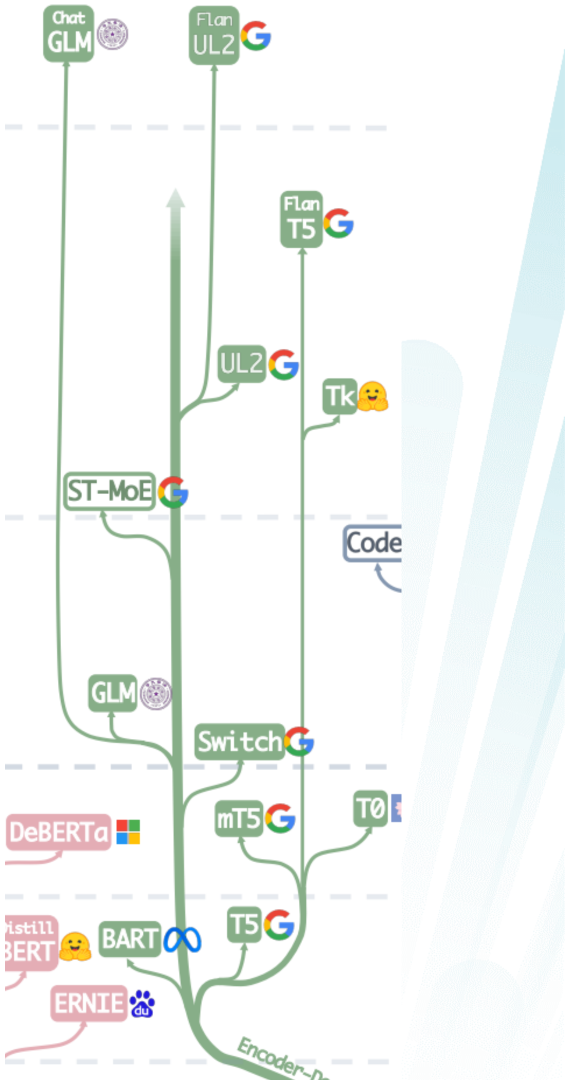
DE.ABC.
Sentence Permutation

C.DE.AB
Document Rotation

A.C.E.
Token Deletion

ABC.DE.

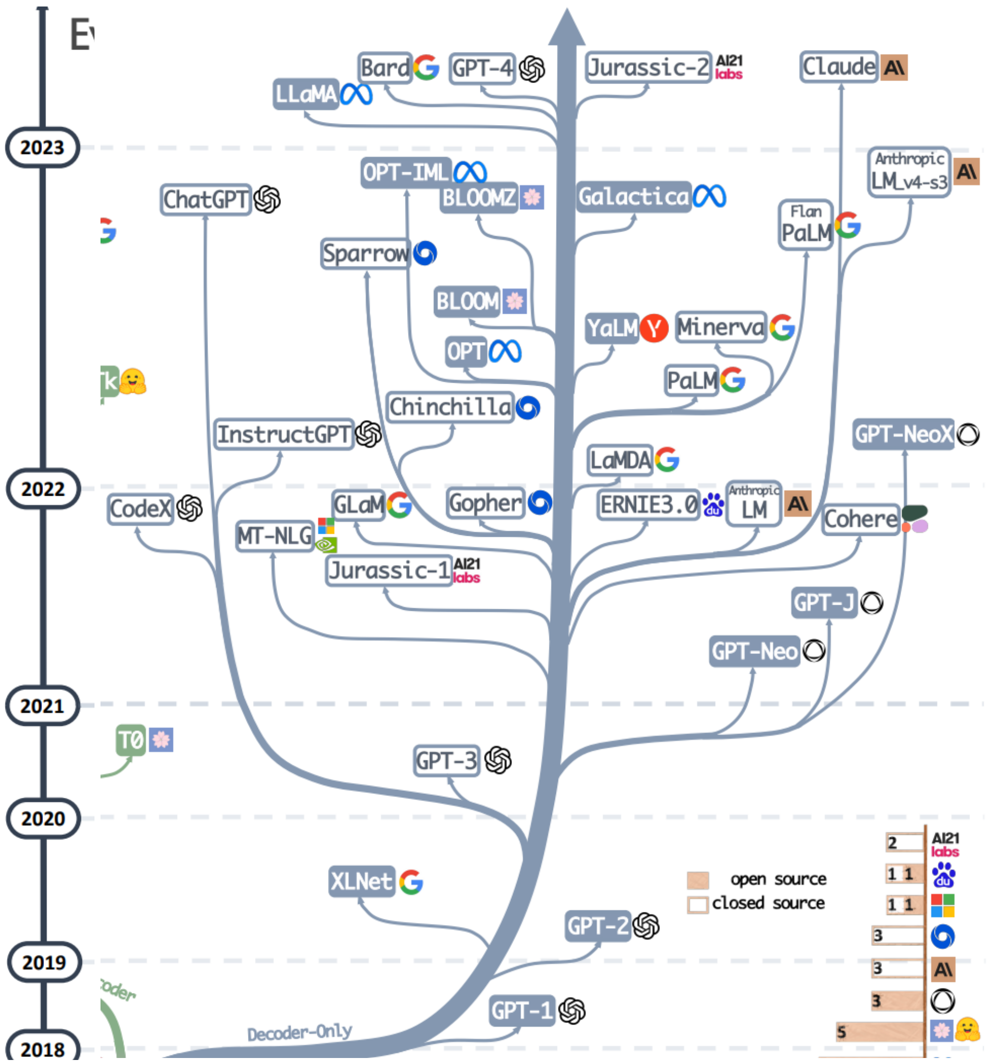
A_.D_E.
Text Infilling



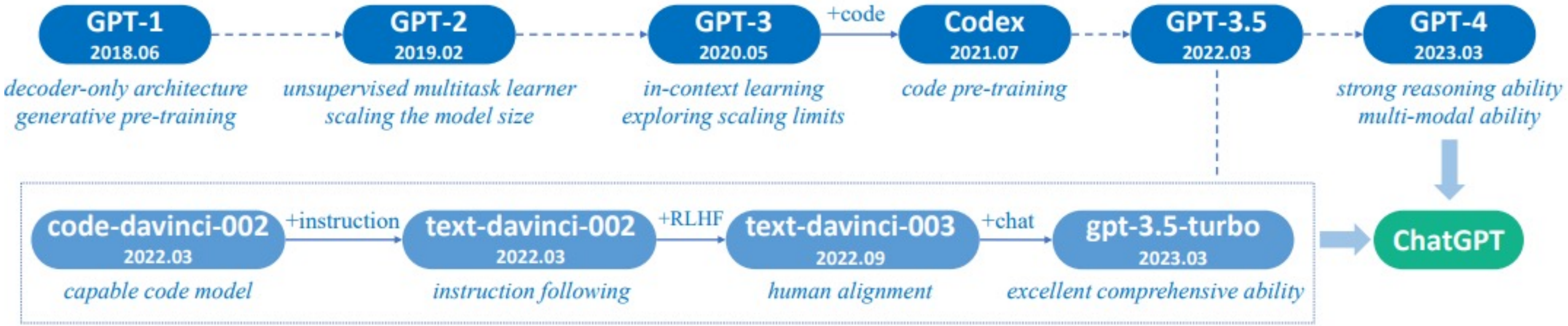
Lewis, Mike, et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." *arXiv preprint arXiv:1910.13461* (2019).

# GPT Family

## **Generative Pre-trained Transformer 1 (GPT-1)**- 2018



# GPT Family


**GPT-1**
2018.06
*decoder-only architecture*
*generative pre-training*

**GPT-2**
2019.02
*unsupervised multitask learner*
*scaling the model size*

**GPT-3**
2020.05
*in-context learning*
*exploring scaling limits*

+code

**Codex**
2021.07
*code pre-training*

**GPT-3.5**
2022.03

**GPT-4**
2023.03
*strong reasoning ability*
*multi-modal ability*

**code-davinci-002**
2022.03
*capable code model*

+instruction

**text-davinci-002**
2022.03
*instruction following*

+RLHF

**text-davinci-003**
2022.09
*human alignment*

+chat

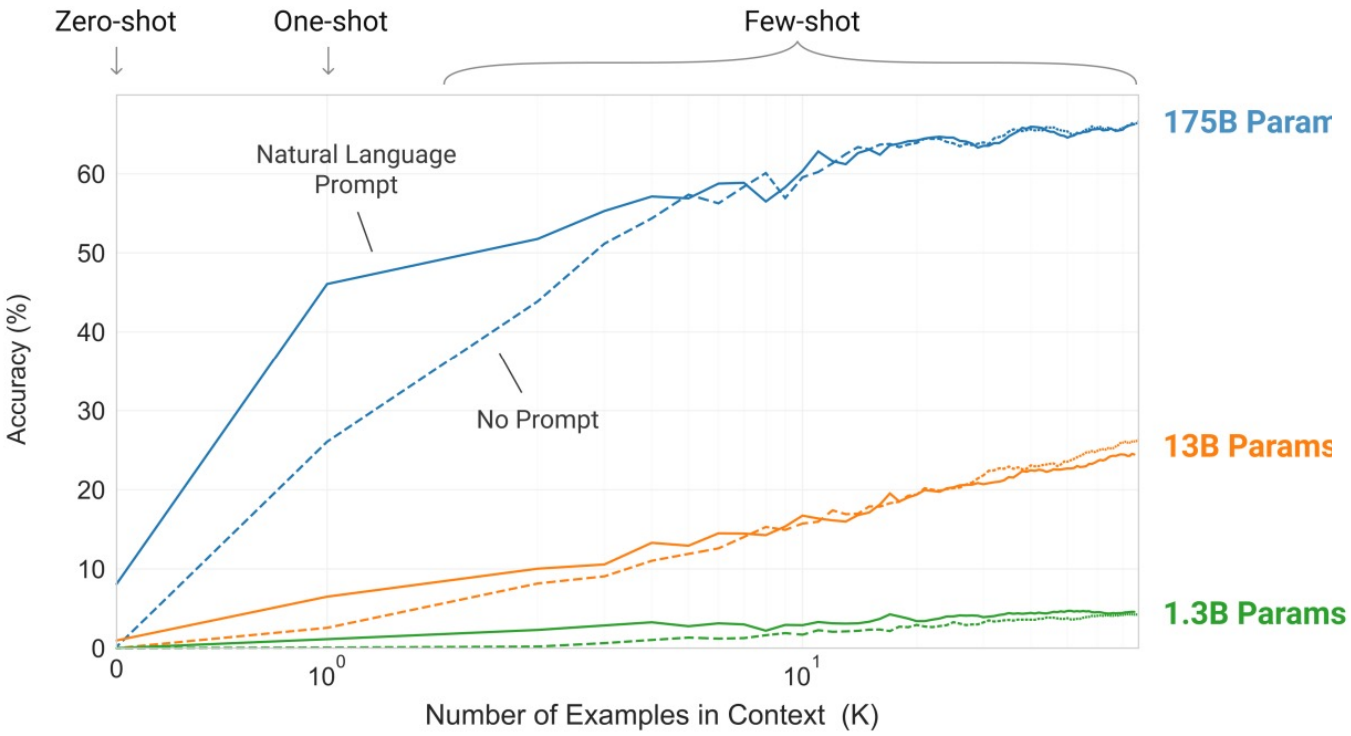
**gpt-3.5-turbo**
2023.03
*excellent comprehensive ability*

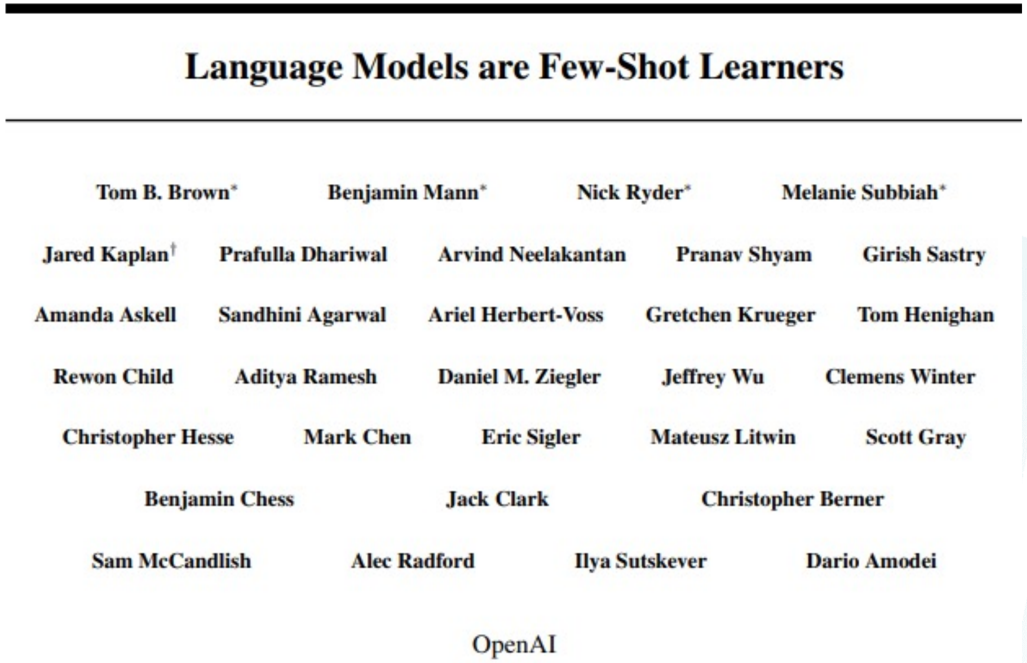
**ChatGPT**

Source: Zhao, Wayne Xin, et al. "A survey of large language models." *arXiv preprint arXiv:2303.18223* (2023).

# GPT Family

**Scaling up language models significantly improves the few-shot, even zero-shot performance**





## Language Models are Few-Shot Learners

Tom B. Brown[*] Benjamin Mann[*] Nick Ryder[*] Melanie Subbiah[*]

Jared Kaplan[†] Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry

Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan

Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter

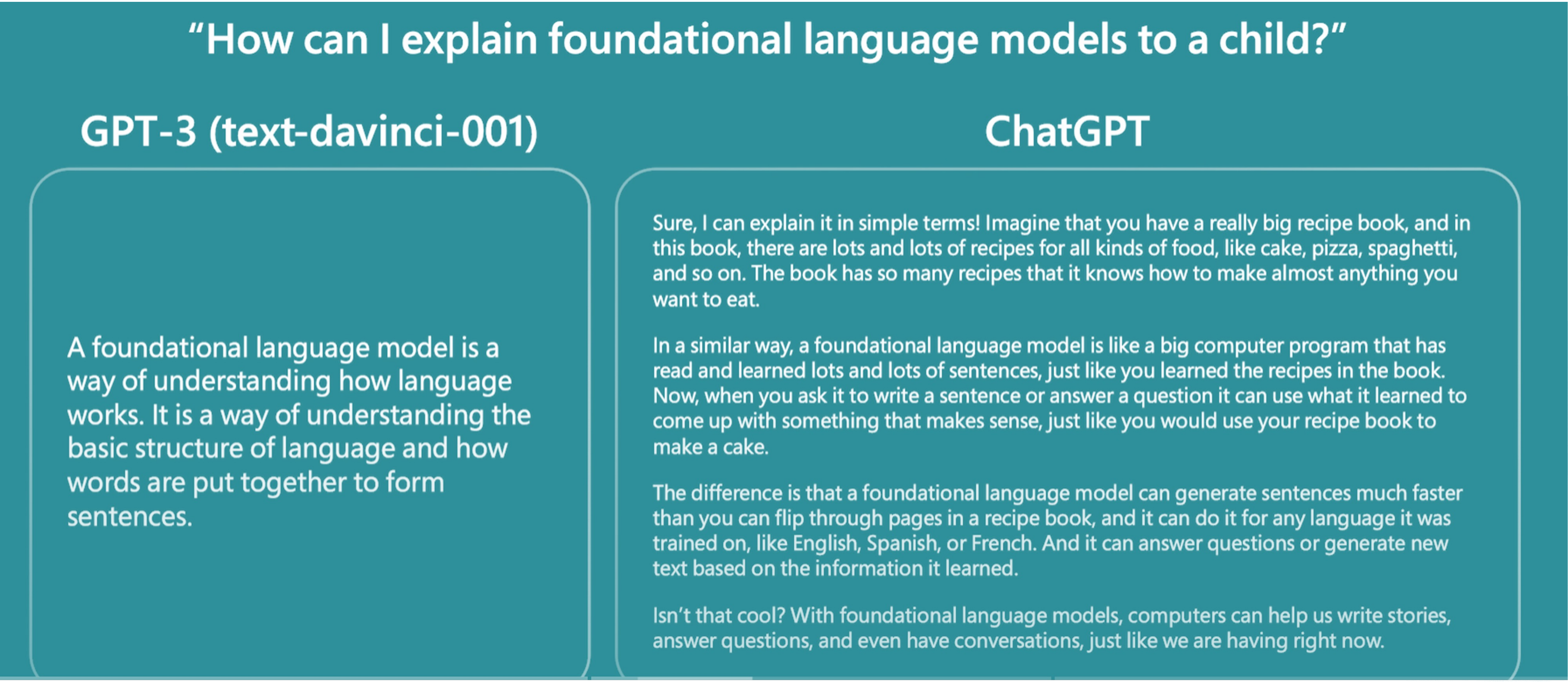
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

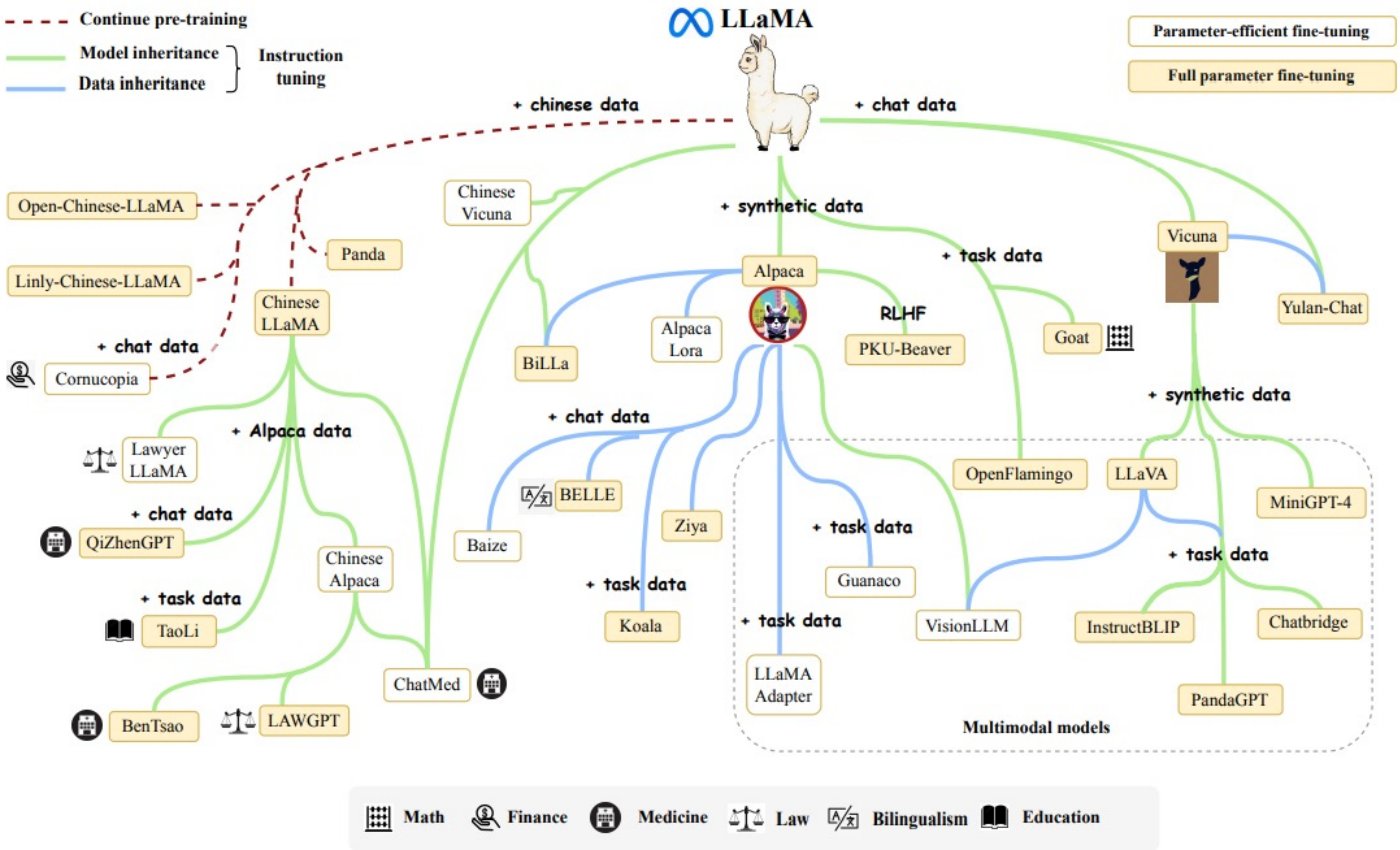
# GPT3 vs chatGPT



“How can I explain foundational language models to a child?”

| GPT-3 (text-davinci-001) | ChatGPT |
|---|---|
| A foundational language model is a way of understanding how language works. It is a way of understanding the basic structure of language and how words are put together to form sentences. | Sure, I can explain it in simple terms! Imagine that you have a really big recipe book, and in this book, there are lots and lots of recipes for all kinds of food, like cake, pizza, spaghetti, and so on. The book has so many recipes that it knows how to make almost anything you want to eat.<br>In a similar way, a foundational language model is like a big computer program that has read and learned lots and lots of sentences, just like you learned the recipes in the book. Now, when you ask it to write a sentence or answer a question it can use what it learned to come up with something that makes sense, just like you would use your recipe book to make a cake.<br>The difference is that a foundational language model can generate sentences much faster than you can flip through pages in a recipe book, and it can do it for any language it was trained on, like English, Spanish, or French. And it can answer questions or generate new text based on the information it learned.<br>Isn’t that cool? With foundational language models, computers can help us write stories, answer questions, and even have conversations, just like we are having right now. |

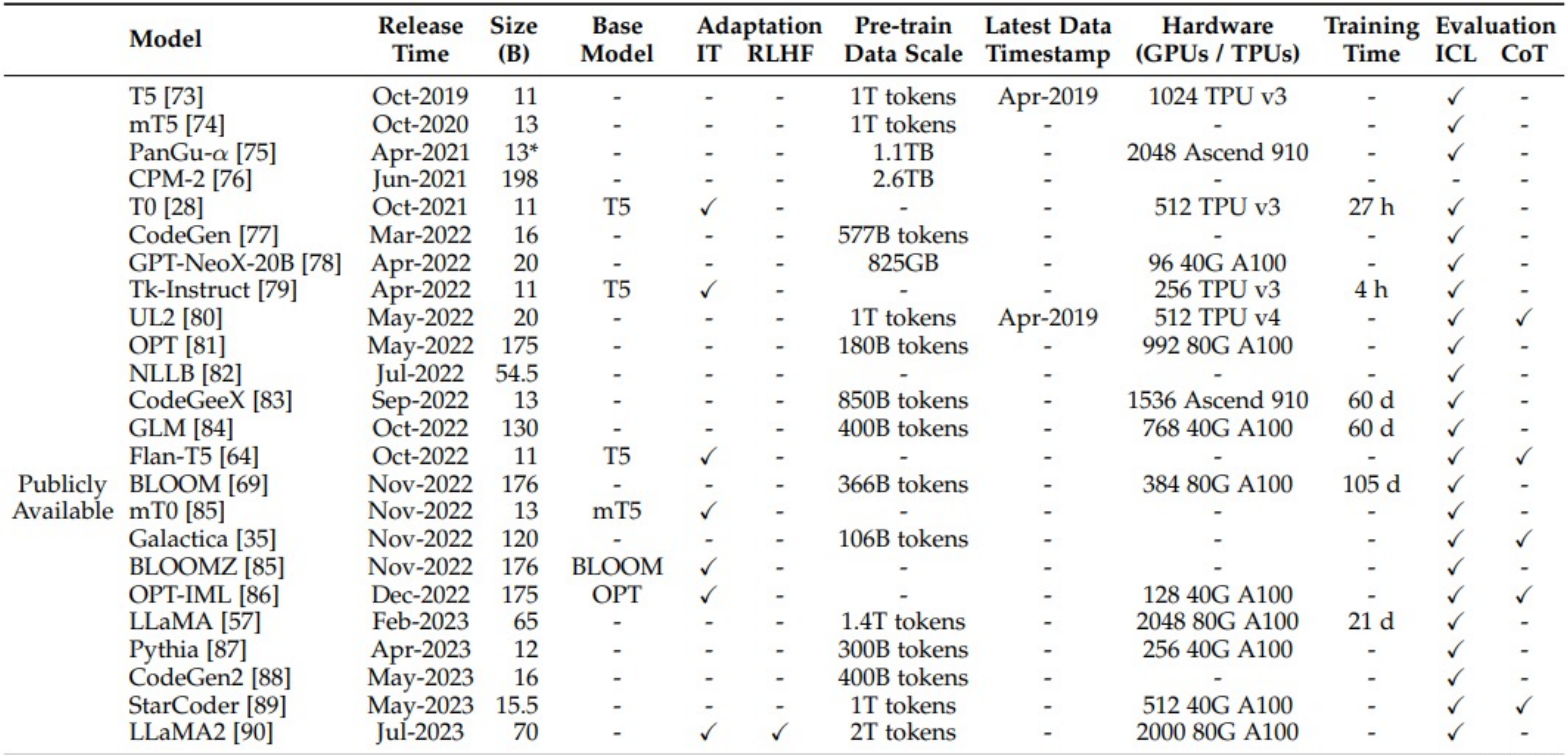
# LLaMA Model Family



SECTION 3

# Stats and Benchmarks

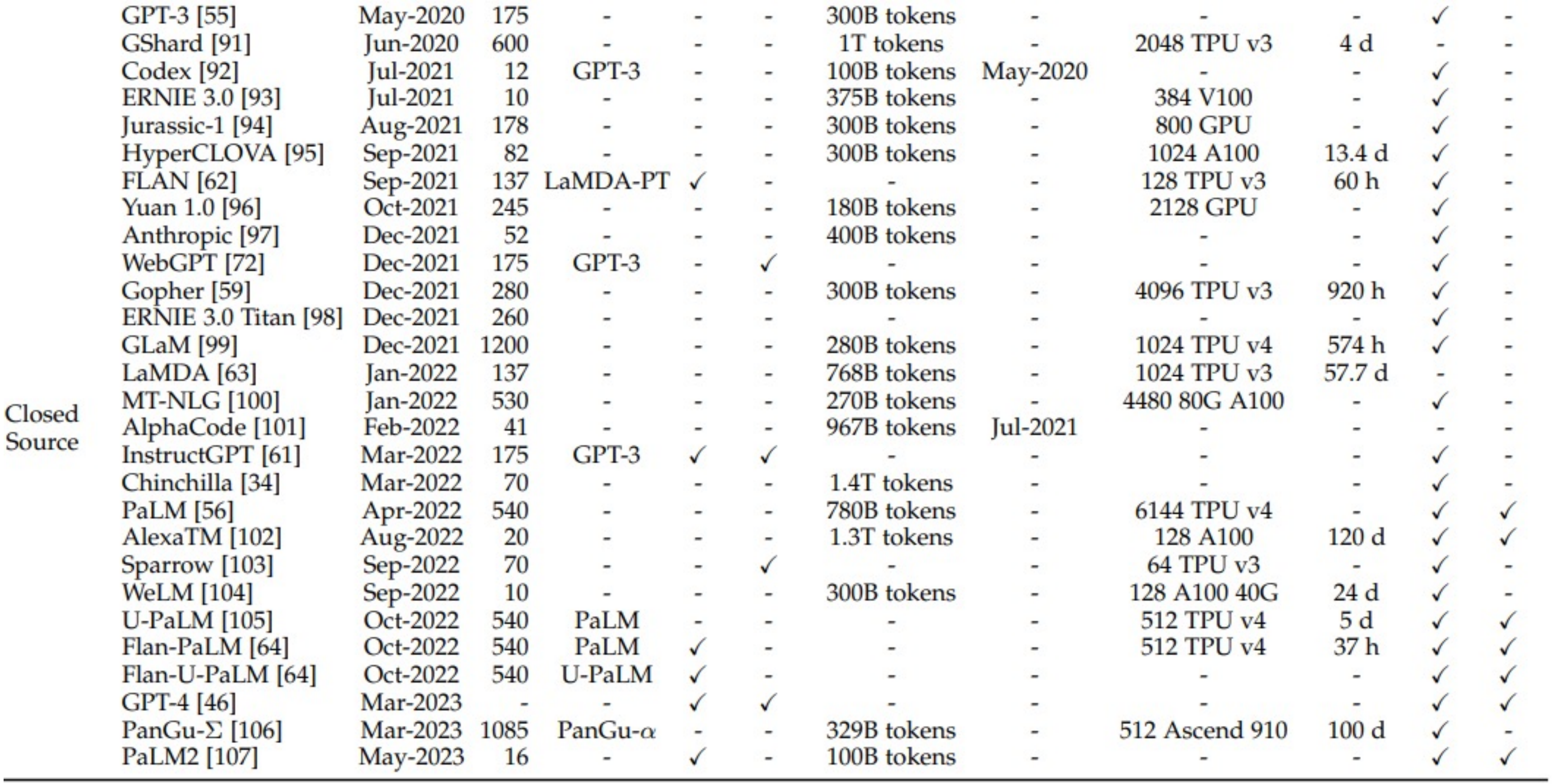
# Statistics of large language models



| | Model | Release Time | Size (B) | Base Model | Adaptation | | Pre-train Data Scale | Latest Data Timestamp | Hardware (GPUs / TPUs) | Training Time | Evaluation | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | IT | RLHF | | | | | ICL | CoT |
| Publicly Available | T5 [73] | Oct-2019 | 11 | - | - | - | 1T tokens | Apr-2019 | 1024 TPU v3 | - | ✓ | - |
| | mT5 [74] | Oct-2020 | 13 | - | - | - | 1T tokens | - | - | - | ✓ | - |
| | PanGu-α [75] | Apr-2021 | 13* | - | - | - | 1.1TB | - | 2048 Ascend 910 | - | ✓ | - |
| | CPM-2 [76] | Jun-2021 | 198 | - | - | - | 2.6TB | - | - | - | - | - |
| | T0 [28] | Oct-2021 | 11 | T5 | ✓ | - | - | - | 512 TPU v3 | 27 h | ✓ | - |
| | CodeGen [77] | Mar-2022 | 16 | - | - | - | 577B tokens | - | - | - | ✓ | - |
| | GPT-NeoX-20B [78] | Apr-2022 | 20 | - | - | - | 825GB | - | 96 40G A100 | - | ✓ | - |
| | Tk-Instruct [79] | Apr-2022 | 11 | T5 | ✓ | - | - | - | 256 TPU v3 | 4 h | ✓ | - |
| | UL2 [80] | May-2022 | 20 | - | - | - | 1T tokens | Apr-2019 | 512 TPU v4 | - | ✓ | ✓ |
| | OPT [81] | May-2022 | 175 | - | - | - | 180B tokens | - | 992 80G A100 | - | ✓ | - |
| | NLLB [82] | Jul-2022 | 54.5 | - | - | - | - | - | - | - | ✓ | - |
| | CodeGeeX [83] | Sep-2022 | 13 | - | - | - | 850B tokens | - | 1536 Ascend 910 | 60 d | ✓ | - |
| | GLM [84] | Oct-2022 | 130 | - | - | - | 400B tokens | - | 768 40G A100 | 60 d | ✓ | - |
| | Flan-T5 [64] | Oct-2022 | 11 | T5 | ✓ | - | - | - | - | - | ✓ | ✓ |
| | BLOOM [69] | Nov-2022 | 176 | - | - | - | 366B tokens | - | 384 80G A100 | 105 d | ✓ | - |
| | mT0 [85] | Nov-2022 | 13 | mT5 | ✓ | - | - | - | - | - | ✓ | - |
| | Galactica [35] | Nov-2022 | 120 | - | - | - | 106B tokens | - | - | - | ✓ | ✓ |
| | BLOOMZ [85] | Nov-2022 | 176 | BLOOM | ✓ | - | - | - | - | - | ✓ | - |
| | OPT-IML [86] | Dec-2022 | 175 | OPT | ✓ | - | - | - | 128 40G A100 | - | ✓ | ✓ |
| | LLaMA [57] | Feb-2023 | 65 | - | - | - | 1.4T tokens | - | 2048 80G A100 | 21 d | ✓ | - |
| | Pythia [87] | Apr-2023 | 12 | - | - | - | 300B tokens | - | 256 40G A100 | - | ✓ | - |
| | CodeGen2 [88] | May-2023 | 16 | - | - | - | 400B tokens | - | - | - | ✓ | - |
| | StarCoder [89] | May-2023 | 15.5 | - | - | - | 1T tokens | - | 512 40G A100 | - | ✓ | ✓ |
| | LLaMA2 [90] | Jul-2023 | 70 | - | ✓ | ✓ | 2T tokens | - | 2000 80G A100 | - | ✓ | - |

Source: Zhao, Wayne Xin, et al. "A survey of large language models." *arXiv preprint arXiv:2303.18223* (2023).

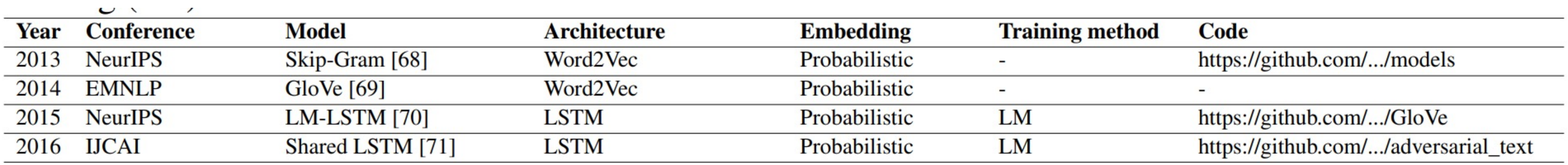
# Statistics of large language models



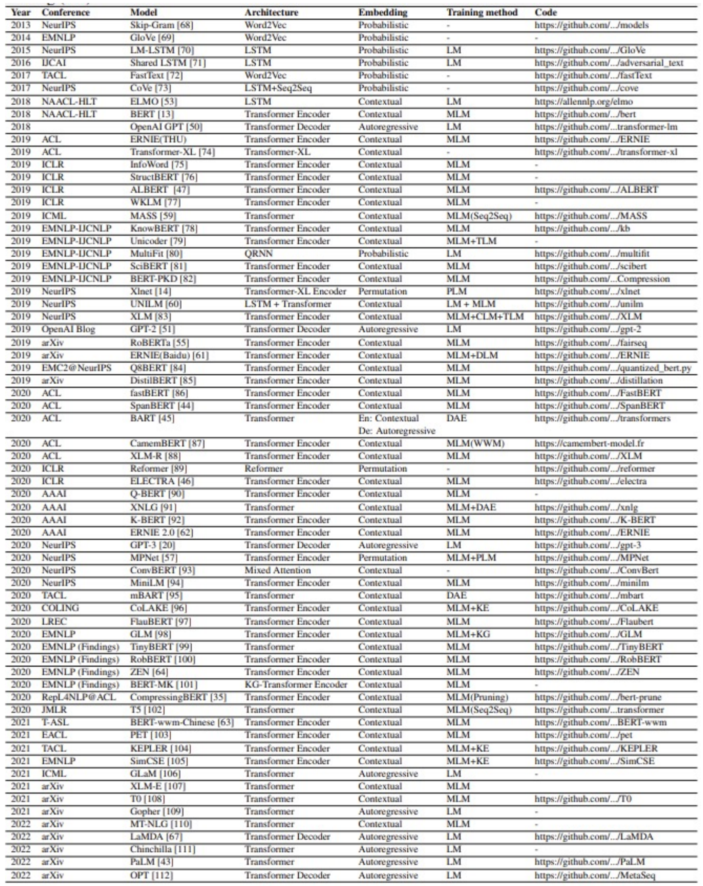
| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Closed Source | GPT-3 [55] | May-2020 | 175 | - | - | - | 300B tokens | - | - | - | ✓ | - |
| | GShard [91] | Jun-2020 | 600 | - | - | - | 1T tokens | - | 2048 TPU v3 | 4 d | - | - |
| | Codex [92] | Jul-2021 | 12 | GPT-3 | - | - | 100B tokens | May-2020 | - | - | ✓ | - |
| | ERNIE 3.0 [93] | Jul-2021 | 10 | - | - | - | 375B tokens | - | 384 V100 | - | ✓ | - |
| | Jurassic-1 [94] | Aug-2021 | 178 | - | - | - | 300B tokens | - | 800 GPU | - | ✓ | - |
| | HyperCLOVA [95] | Sep-2021 | 82 | - | - | - | 300B tokens | - | 1024 A100 | 13.4 d | ✓ | - |
| | FLAN [62] | Sep-2021 | 137 | LaMDA-PT | ✓ | - | - | - | 128 TPU v3 | 60 h | ✓ | - |
| | Yuan 1.0 [96] | Oct-2021 | 245 | - | - | - | 180B tokens | - | 2128 GPU | - | ✓ | - |
| | Anthropic [97] | Dec-2021 | 52 | - | - | - | 400B tokens | - | - | - | ✓ | - |
| | WebGPT [72] | Dec-2021 | 175 | GPT-3 | - | ✓ | - | - | - | - | ✓ | - |
| | Gopher [59] | Dec-2021 | 280 | - | - | - | 300B tokens | - | 4096 TPU v3 | 920 h | ✓ | - |
| | ERNIE 3.0 Titan [98] | Dec-2021 | 260 | - | - | - | - | - | - | - | ✓ | - |
| | GLaM [99] | Dec-2021 | 1200 | - | - | - | 280B tokens | - | 1024 TPU v4 | 574 h | ✓ | - |
| | LaMDA [63] | Jan-2022 | 137 | - | - | - | 768B tokens | - | 1024 TPU v3 | 57.7 d | - | - |
| | MT-NLG [100] | Jan-2022 | 530 | - | - | - | 270B tokens | - | 4480 80G A100 | - | ✓ | - |
| | AlphaCode [101] | Feb-2022 | 41 | - | - | - | 967B tokens | Jul-2021 | - | - | - | - |
| | InstructGPT [61] | Mar-2022 | 175 | GPT-3 | ✓ | ✓ | - | - | - | - | ✓ | - |
| | Chinchilla [34] | Mar-2022 | 70 | - | - | - | 1.4T tokens | - | - | - | ✓ | - |
| | PaLM [56] | Apr-2022 | 540 | - | - | - | 780B tokens | - | 6144 TPU v4 | - | ✓ | ✓ |
| | AlexaTM [102] | Aug-2022 | 20 | - | - | - | 1.3T tokens | - | 128 A100 | 120 d | ✓ | ✓ |
| | Sparrow [103] | Sep-2022 | 70 | - | - | ✓ | - | - | 64 TPU v3 | - | ✓ | - |
| | WeLM [104] | Sep-2022 | 10 | - | - | - | 300B tokens | - | 128 A100 40G | 24 d | ✓ | - |
| | U-PaLM [105] | Oct-2022 | 540 | PaLM | - | - | - | - | 512 TPU v4 | 5 d | ✓ | ✓ |
| | Flan-PaLM [64] | Oct-2022 | 540 | PaLM | ✓ | - | - | - | 512 TPU v4 | 37 h | ✓ | ✓ |
| | Flan-U-PaLM [64] | Oct-2022 | 540 | U-PaLM | ✓ | - | - | - | - | - | ✓ | ✓ |
| | GPT-4 [46] | Mar-2023 | - | - | ✓ | ✓ | - | - | - | - | ✓ | ✓ |
| | PanGu-$\Sigma$ [106] | Mar-2023 | 1085 | PanGu-$\alpha$ | - | - | 329B tokens | - | 512 Ascend 910 | 100 d | ✓ | - |
| | PaLM2 [107] | May-2023 | 16 | - | ✓ | - | 100B tokens | - | - | - | ✓ | ✓ |

Source: Zhao, Wayne Xin, et al. "A survey of large language models." *arXiv preprint arXiv:2303.18223* (2023).

# Summary of PFMs



| Year | Conference | Model | Architecture | Embedding | Training method | Code |
|---|---|---|---|---|---|---|
| 2013 | NeurIPS | Skip-Gram [68] | Word2Vec | Probabilistic | - | https://github.com/.../models |
| 2014 | EMNLP | GloVe [69] | Word2Vec | Probabilistic | - | - |
| 2015 | NeurIPS | LM-LSTM [70] | LSTM | Probabilistic | LM | https://github.com/.../GloVe |
| 2016 | IJCAI | Shared LSTM [71] | LSTM | Probabilistic | LM | https://github.com/.../adversarial_text |

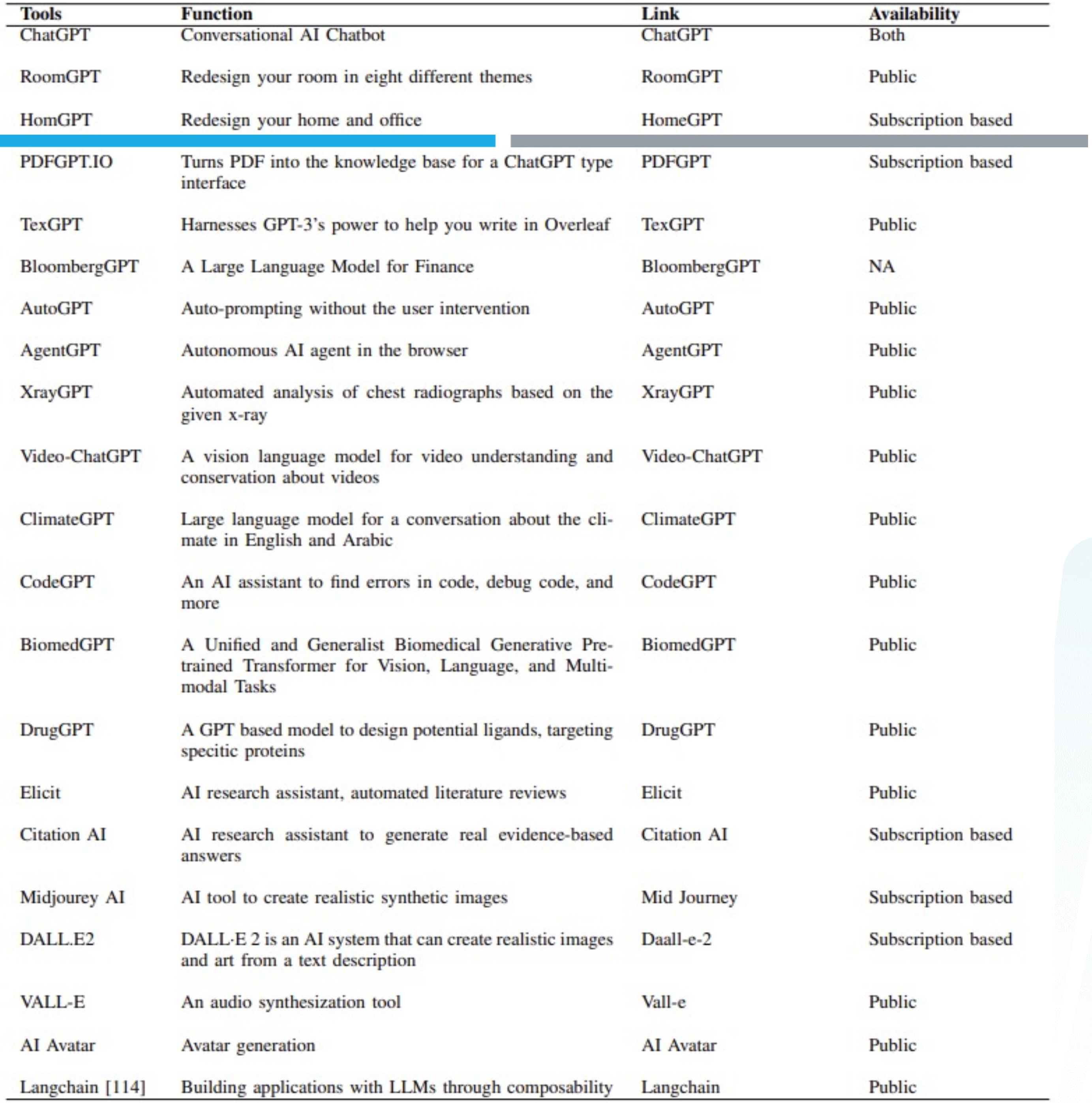


| Year | Conference | Model | Architecture | Embedding | Training method | Code |
|---|---|---|---|---|---|---|
| 2013 | NeurIPS | Skip-Gram [68] | Word2Vec | Probabilistic | - | https://github.com/.../models |
| 2014 | EMNLP | GloVe [69] | Word2Vec | Probabilistic | - | - |
| 2015 | NeurIPS | LM-LSTM [70] | LSTM | Probabilistic | LM | https://github.com/.../GloVe |
| 2016 | IJCAI | Shared LSTM [71] | LSTM | Probabilistic | LM | https://github.com/.../adversarial_text |
| 2017 | TACL | FastText [72] | Word2Vec | Probabilistic | - | https://github.com/.../fastText |
| 2017 | NeurIPS | CoVe [73] | LSTM+Seq2Seq | Probabilistic | - | https://github.com/.../cove |
| 2018 | NAACL-HLT | ELMO [53] | LSTM | Contextual | LM | https://allennlp.org/elmo |
| 2018 | NAACL-HLT | BERT [13] | Transformer Encoder | Contextual | MLM | https://github.com/.../bert |
| 2018 |  | OpenAI GPT [50] | Transformer Decoder | Autoregressive | LM | https://github.com/...transformer-lm |
| 2019 | ACL | ERNIE(THU) | Transformer Encoder | Contextual | MLM | https://github.com/.../ERNIE |
| 2019 | ACL | Transformer-XL [74] | Transformer-XL | Contextual | - | https://github.com/.../transformer-xl |
| 2019 | ICLR | InfoWord [75] | Transformer Encoder | Contextual | MLM | - |
| 2019 | ICLR | StructBERT [76] | Transformer Encoder | Contextual | MLM | - |
| 2019 | ICLR | ALBERT [47] | Transformer Encoder | Contextual | MLM | https://github.com/.../ALBERT |
| 2019 | ICLR | WKLM [77] | Transformer Encoder | Contextual | MLM | - |
| 2019 | ICML | MASS [59] | Transformer | Contextual | MLM(Seq2Seq) | https://github.com/.../MASS |
| 2019 | EMNLP-IJCNLP | KnowBERT [78] | Transformer Encoder | Contextual | MLM | https://github.com/.../kb |
| 2019 | EMNLP-IJCNLP | Unicoder [79] | Transformer Encoder | Contextual | MLM+TLM | - |
| 2019 | EMNLP-IJCNLP | MultiFit [80] | QRNN | Probabilistic | LM | https://github.com/.../multifit |
| 2019 | EMNLP-IJCNLP | SciBERT [81] | Transformer Encoder | Contextual | MLM | https://github.com/.../scibert |
| 2019 | EMNLP-IJCNLP | BERT-PKD [82] | Transformer Encoder | Contextual | MLM | https://github.com/...Compression |
| 2019 | NeurIPS | Xlnet [14] | Transformer-XL Encoder | Permutation | PLM | https://github.com/.../xlnet |
| 2019 | NeurIPS | UNILM [60] | LSTM + Transformer | Contextual | LM + MLM | https://github.com/.../unilm |
| 2019 | NeurIPS | XLM [83] | Transformer Encoder | Contextual | MLM+CLM+TLM | https://github.com/.../XLM |
| 2019 | OpenAI Blog | GPT-2 [51] | Transformer Decoder | Autoregressive | LM | https://github.com/.../gpt-2 |
| 2019 | arXiv | RoBERTa [55] | Transformer Encoder | Contextual | MLM | https://github.com/.../fairseq |
| 2019 | arXiv | ERNIE(Baidu) [61] | Transformer Encoder | Contextual | MLM+DLM | https://github.com/.../ERNIE |
| 2019 | EMC2@NeurIPS | Q8BERT [84] | Transformer Encoder | Contextual | MLM | https://github.com/.../quantized_bert.py |
| 2019 | arXiv | DistilBERT [85] | Transformer Encoder | Contextual | MLM | https://github.com/.../distillation |
| 2020 | ACL | fastBERT [86] | Transformer Encoder | Contextual | MLM | https://github.com/.../FastBERT |
| 2020 | ACL | SpanBERT [44] | Transformer Encoder | Contextual | MLM | https://github.com/.../SpanBERT |
| 2020 | ACL | BART [45] | Transformer | En: Contextual<br>De: Autoregressive | DAE | https://github.com/.../transformers |
| 2020 | ACL | CamemBERT [87] | Transformer Encoder | Contextual | MLM(WWM) | https://camembert-model.fr |
| 2020 | ACL | XLM-R [88] | Transformer Encoder | Contextual | MLM | https://github.com/.../XLM |
| 2020 | ICLR | Reformer [89] | Reformer | Permutation | - | https://github.com/.../reformer |
| 2020 | ICLR | ELECTRA [46] | Transformer Encoder | Contextual | MLM | https://github.com/.../electra |
| 2020 | AAAI | Q-BERT [90] | Transformer Encoder | Contextual | MLM | - |
| 2020 | AAAI | XNLG [91] | Transformer | Contextual | MLM+DAE | https://github.com/.../xnlg |
| 2020 | AAAI | K-BERT [92] | Transformer Encoder | Contextual | MLM | https://github.com/.../K-BERT |
| 2020 | AAAI | ERNIE 2.0 [62] | Transformer Encoder | Contextual | MLM | https://github.com/.../ERNIE |
| 2020 | NeurIPS | GPT-3 [20] | Transformer Decoder | Autoregressive | LM | https://github.com/.../gpt-3 |
| 2020 | NeurIPS | MPNet [57] | Transformer Encoder | Permutation | MLM+PLM | https://github.com/.../MPNet |
| 2020 | NeurIPS | ConvBERT [93] | Mixed Attention | Contextual | - | https://github.com/.../ConvBert |
| 2020 | NeurIPS | MiniLM [94] | Transformer Encoder | Contextual | MLM | https://github.com/.../minilm |
| 2020 | TACL | mBART [95] | Transformer | Contextual | DAE | https://github.com/.../mbart |
| 2020 | COLING | CoLAKE [96] | Transformer Encoder | Contextual | MLM+KE | https://github.com/.../CoLAKE |
| 2020 | LREC | FlauBERT [97] | Transformer Encoder | Contextual | MLM | https://github.com/.../Flaubert |
| 2020 | EMNLP | GLM [98] | Transformer Encoder | Contextual | MLM+KG | https://github.com/.../GLM |
| 2020 | EMNLP (Findings) | TinyBERT [99] | Transformer | Contextual | MLM | https://github.com/.../TinyBERT |
| 2020 | EMNLP (Findings) | RobBERT [100] | Transformer Encoder | Contextual | MLM | https://github.com/.../RobBERT |
| 2020 | EMNLP (Findings) | ZEN [64] | Transformer Encoder | Contextual | MLM | https://github.com/.../ZEN |
| 2020 | EMNLP (Findings) | BERT-MK [101] | KG-Transformer Encoder | Contextual | MLM | - |
| 2020 | RepL4NLP@ACL | CompressingBERT [35] | Transformer Encoder | Contextual | MLM(Pruning) | https://github.com/.../bert-prune |
| 2020 | JMLR | T5 [102] | Transformer | Contextual | MLM(Seq2Seq) | https://github.com/...transformer |
| 2021 | T-ASL | BERT-wwm-Chinese [63] | Transformer Encoder | Contextual | MLM | https://github.com/...BERT-wwm |
| 2021 | EACL | PET [103] | Transformer Encoder | Contextual | MLM | https://github.com/.../pet |
| 2021 | TACL | KEPLER [104] | Transformer Encoder | Contextual | MLM+KE | https://github.com/.../KEPLER |
| 2021 | EMNLP | SimCSE [105] | Transformer Encoder | Contextual | MLM+KE | https://github.com/.../SimCSE |
| 2021 | ICML | GLaM [106] | Transformer | Autoregressive | LM | - |
| 2021 | arXiv | XLM-E [107] | Transformer | Contextual | MLM |  |
| 2021 | arXiv | T0 [108] | Transformer | Contextual | MLM | https://github.com/.../T0 |
| 2021 | arXiv | Gopher [109] | Transformer | Autoregressive | LM | - |
| 2022 | arXiv | MT-NLG [110] | Transformer | Contextual | MLM | - |
| 2022 | arXiv | LaMDA [67] | Transformer Decoder | Autoregressive | LM | https://github.com/.../LaMDA |
| 2022 | arXiv | Chinchilla [111] | Transformer | Autoregressive | LM | - |
| 2022 | arXiv | PaLM [43] | Transformer | Autoregressive | LM | https://github.com/.../PaLM |
| 2022 | arXiv | OPT [112] | Transformer Decoder | Autoregressive | LM | https://github.com/.../MetaSeq |

Natural Language Processing
Computer Vision
Graph Learning
Other Modality…

*Source:* Zhou, Ce, et al. "A comprehensive survey on pretrained foundation models: A history from bert to chatgpt." *arXiv preprint arXiv:2302.09419* (2023).

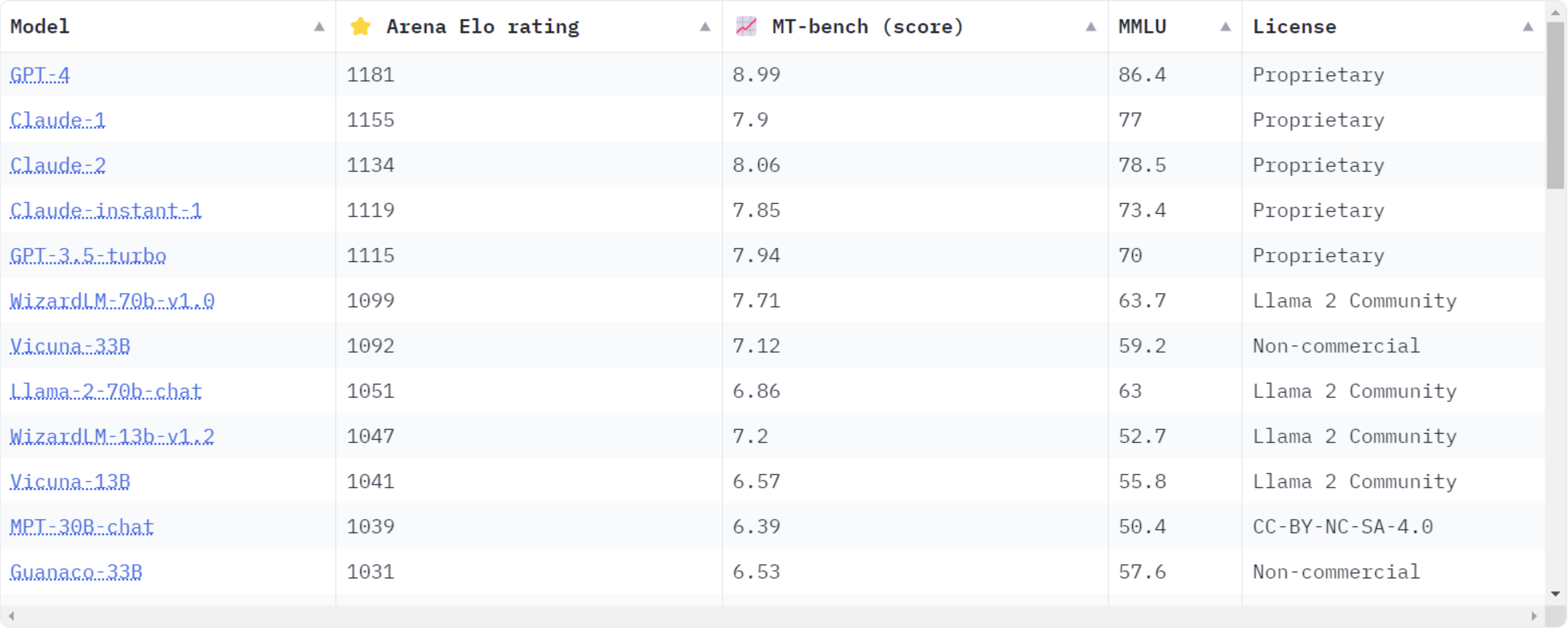
# Downstream tasks



| Tools | Function | Link | Availability |
|---|---|---|---|
| ChatGPT | Conversational AI Chatbot | ChatGPT | Both |
| RoomGPT | Redesign your room in eight different themes | RoomGPT | Public |
| HomGPT | Redesign your home and office | HomeGPT | Subscription based |
| PDFGPT.IO | Turns PDF into the knowledge base for a ChatGPT type interface | PDFGPT | Subscription based |
| TexGPT | Harnesses GPT-3's power to help you write in Overleaf | TexGPT | Public |
| BloombergGPT | A Large Language Model for Finance | BloombergGPT | NA |
| AutoGPT | Auto-prompting without the user intervention | AutoGPT | Public |
| AgentGPT | Autonomous AI agent in the browser | AgentGPT | Public |
| XrayGPT | Automated analysis of chest radiographs based on the given x-ray | XrayGPT | Public |
| Video-ChatGPT | A vision language model for video understanding and conservation about videos | Video-ChatGPT | Public |
| ClimateGPT | Large language model for a conversation about the climate in English and Arabic | ClimateGPT | Public |
| CodeGPT | An AI assistant to find errors in code, debug code, and more | CodeGPT | Public |
| BiomedGPT | A Unified and Generalist Biomedical Generative Pre-trained Transformer for Vision, Language, and Multimodal Tasks | BiomedGPT | Public |
| DrugGPT | A GPT based model to design potential ligands, targeting specitic proteins | DrugGPT | Public |
| Elicit | AI research assistant, automated literature reviews | Elicit | Public |
| Citation AI | AI research assistant to generate real evidence-based answers | Citation AI | Subscription based |
| Midjourey AI | AI tool to create realistic synthetic images | Mid Journey | Subscription based |
| DALL.E2 | DALL·E 2 is an AI system that can create realistic images and art from a text description | Daall-e-2 | Subscription based |
| VALL-E | An audio synthesization tool | Vall-e | Public |
| AI Avatar | Avatar generation | AI Avatar | Public |
| Langchain [114] | Building applications with LLMs through composability | Langchain | Public |

*Source:* Zhou, Ce, et al. "A comprehensive survey on pretrained foundation models: A history from bert to chatgpt." *arXiv preprint arXiv:2302.09419* (2023).

# Benchmarks



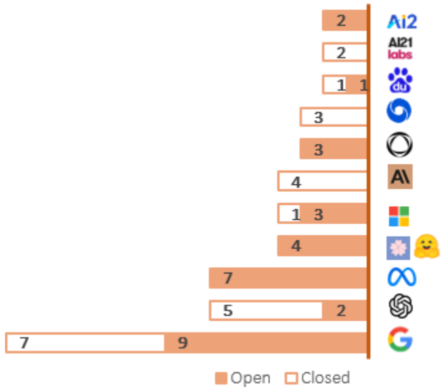
| Model | ⭐ Arena Elo rating | 📈 MT-bench (score) | MMLU | License |
|---|---|---|---|---|
| GPT-4 | 1181 | 8.99 | 86.4 | Proprietary |
| Claude-1 | 1155 | 7.9 | 77 | Proprietary |
| Claude-2 | 1134 | 8.06 | 78.5 | Proprietary |
| Claude-instant-1 | 1119 | 7.85 | 73.4 | Proprietary |
| GPT-3.5-turbo | 1115 | 7.94 | 70 | Proprietary |
| WizardLM-70b-v1.0 | 1099 | 7.71 | 63.7 | Llama 2 Community |
| Vicuna-33B | 1092 | 7.12 | 59.2 | Non-commercial |
| Llama-2-70b-chat | 1051 | 6.86 | 63 | Llama 2 Community |
| WizardLM-13b-v1.2 | 1047 | 7.2 | 52.7 | Llama 2 Community |
| Vicuna-13B | 1041 | 6.57 | 55.8 | Llama 2 Community |
| MPT-30B-chat | 1039 | 6.39 | 50.4 | CC-BY-NC-SA-4.0 |
| Guanaco-33B | 1031 | 6.53 | 57.6 | Non-commercial |

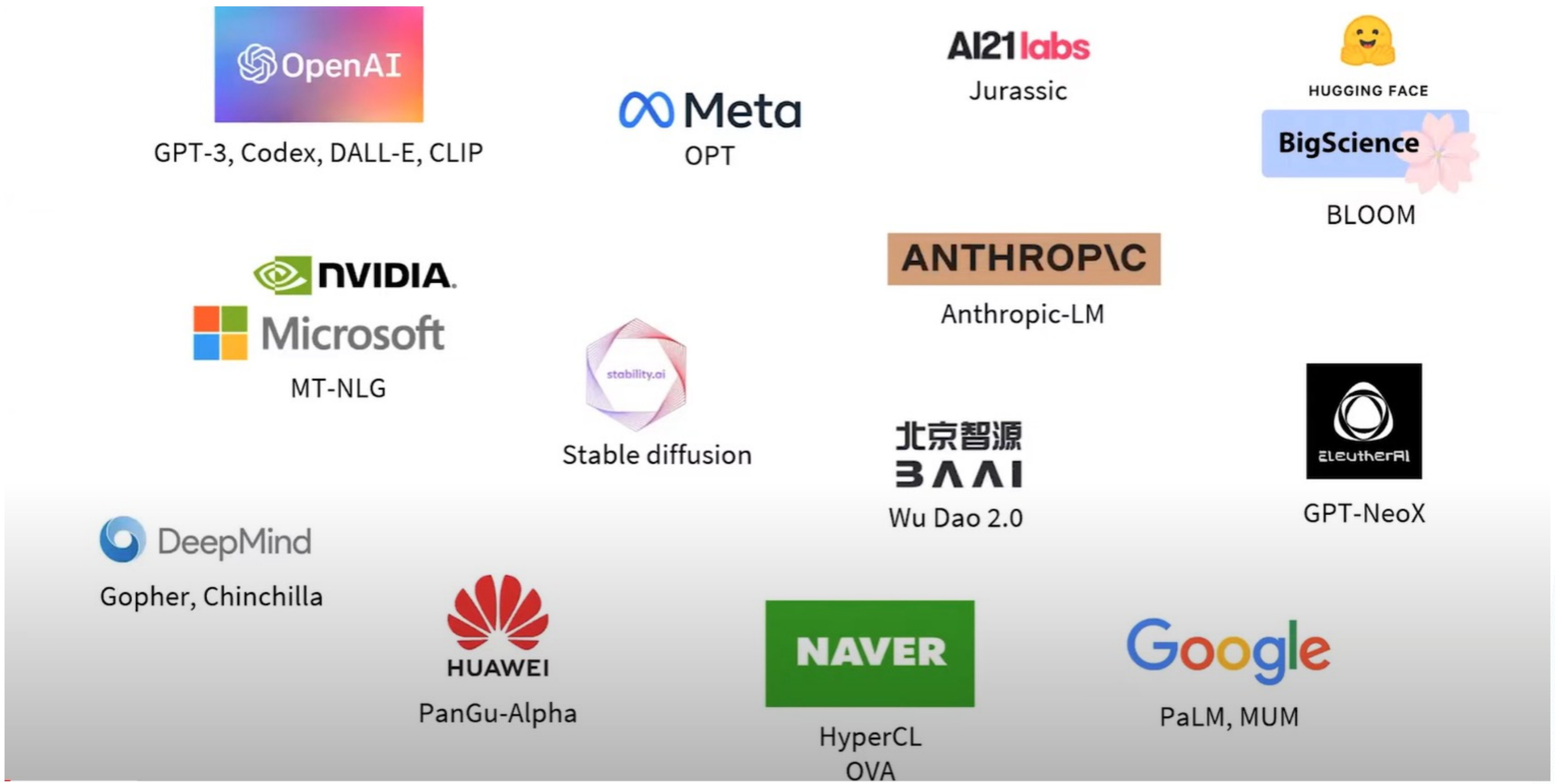
Source: https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard

# HELM

- *https://crfm.stanford.edu/helm/latest/*

# Benchmarks


2
Ai2
2
AI21 labs
1 1
3
3
4
1 3
4
7
5 2
7 9
Open
Closed


OpenAI
GPT-3, Codex, DALL-E, CLIP
Meta
OPT
AI21 labs
Jurassic
HUGGING FACE
BigScience
BLOOM
NVIDIA
Microsoft
MT-NLG
stability.ai
Stable diffusion
ANTHROP\C
Anthropic-LM
北京智源
BAAI
Wu Dao 2.0
EleutherAI
GPT-NeoX
DeepMind
Gopher, Chinchilla
HUAWEI
PanGu-Alpha
NAVER
HyperCL
OVA
Google
PaLM, MUM

# Conclusions

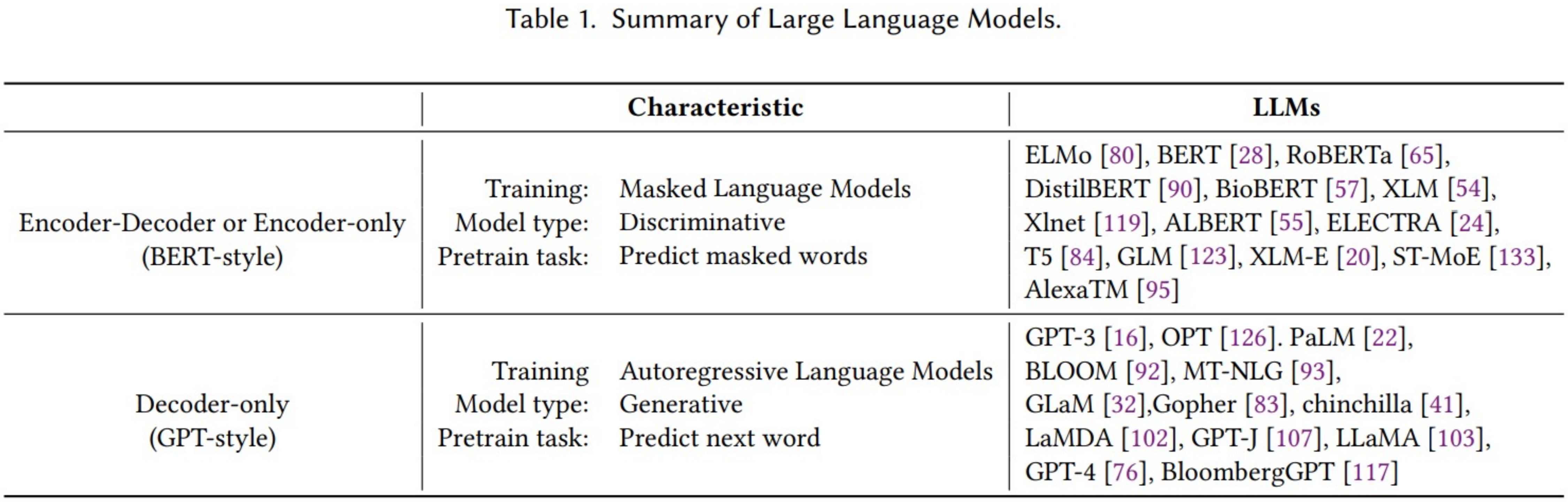
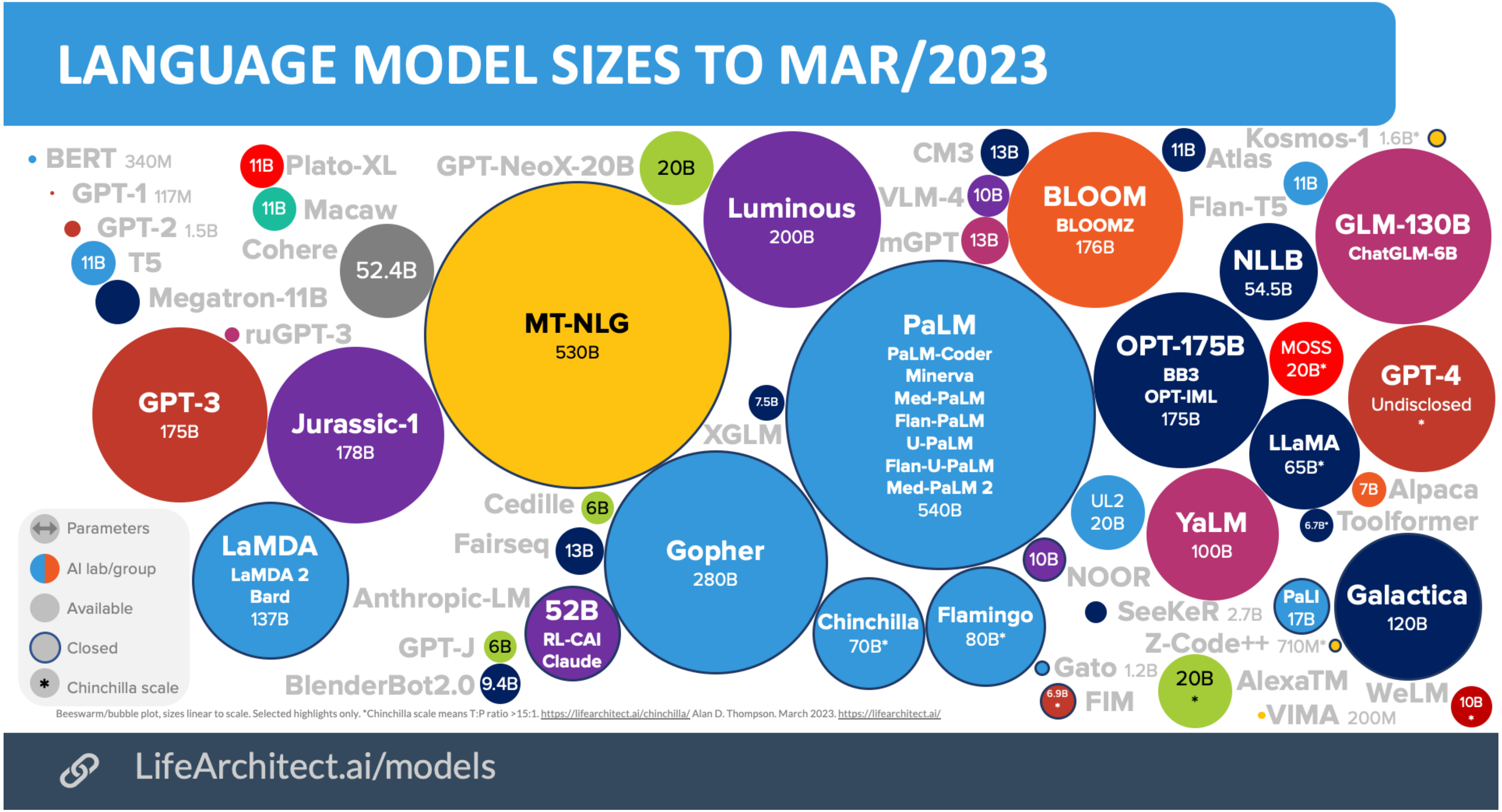


Table 1. Summary of Large Language Models.

| | Characteristic | LLMs |
|---|---|---|
| Encoder-Decoder or Encoder-only<br>(BERT-style) | Training: Masked Language Models<br>Model type: Discriminative<br>Pretrain task: Predict masked words | ELMo [80], BERT [28], RoBERTa [65], DistilBERT [90], BioBERT [57], XLM [54], Xlnet [119], ALBERT [55], ELECTRA [24], T5 [84], GLM [123], XLM-E [20], ST-MoE [133], AlexaTM [95] |
| Decoder-only<br>(GPT-style) | Training Autoregressive Language Models<br>Model type: Generative<br>Pretrain task: Predict next word | GPT-3 [16], OPT [126]. PaLM [22], BLOOM [92], MT-NLG [93], GLaM [32],Gopher [83], chinchilla [41], LaMDA [102], GPT-J [107], LLaMA [103], GPT-4 [76], BloombergGPT [117] |

# Size Now



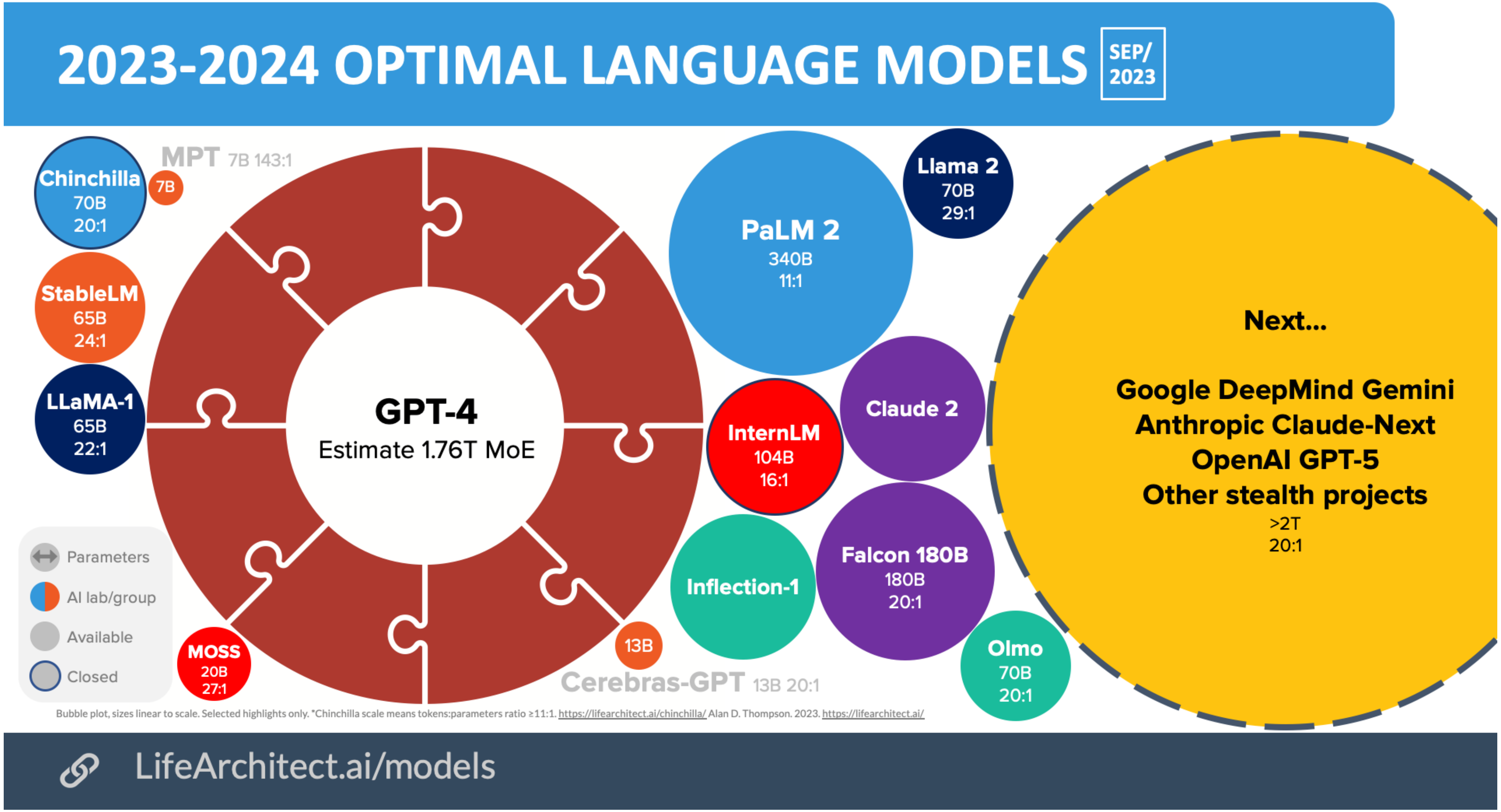
LANGUAGE MODEL SIZES TO MAR/2023

Beeswarm/bubble plot, sizes linear to scale. Selected highlights only. *Chinchilla scale means T:P ratio >15:1. https://lifearchitect.ai/chinchilla/ Alan D. Thompson. March 2023. https://lifearchitect.ai/

LifeArchitect.ai/models

Source: https://lifearchitect.ai/models/#model-bubbles

# Future Directions



2023-2024 OPTIMAL LANGUAGE MODELS SEP/ 2023

Bubble plot, sizes linear to scale. Selected highlights only. *Chinchilla scale means tokens:parameters ratio ≥11:1. https://lifearchitect.ai/chinchilla/ Alan D. Thompson. 2023. https://lifearchitect.ai/

LifeArchitect.ai/models

Source: https://lifearchitect.ai/models/#model-bubbles

# Thank You

Arctic LLM Workshop 2023
Dept. of Computer Science

https://en.uit.no/enhet/ifi

www.bioailab.org