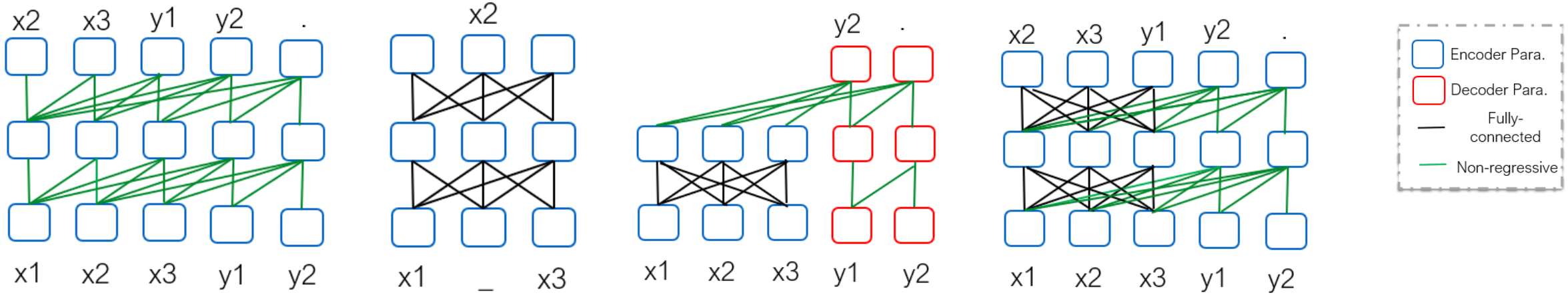
预训练模型回顾

CS2916 大语言模型

饮水思源 爱国荣校

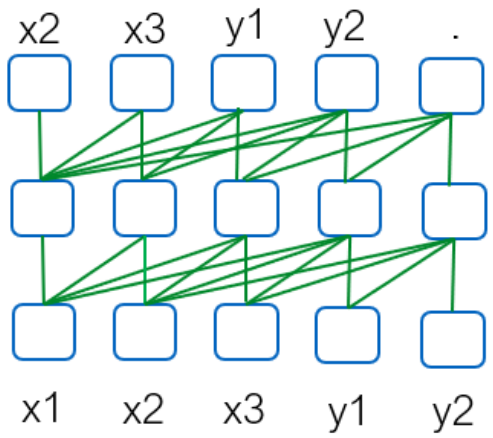
https://plms.ai/teaching/index.html

“Big Four” Pretraining Framework



Left-to-right Masked LM Encode-decoder Prefixed LM

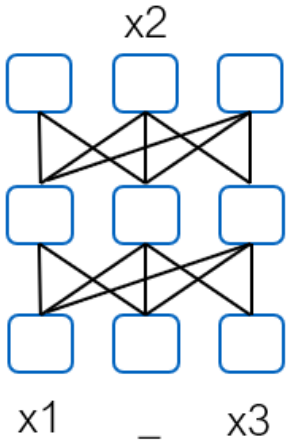
“Big Four” Pretraining Framework



Left-to-right

unidirectional

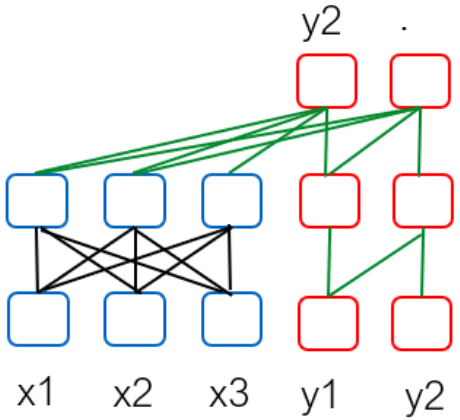
GPT1/2/3



Masked LM

no decoder

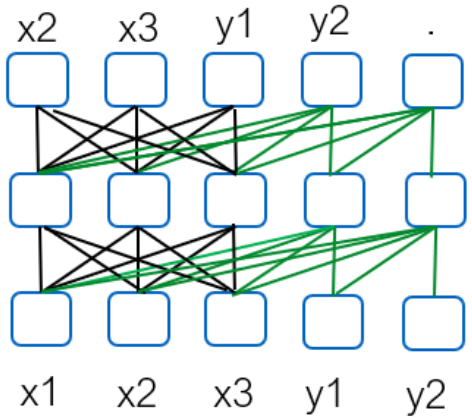
BERT



Encode-decoder

more params

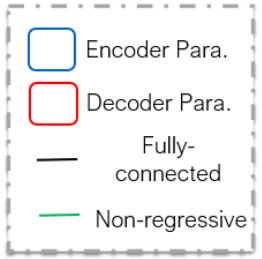
MASS/T5/BART



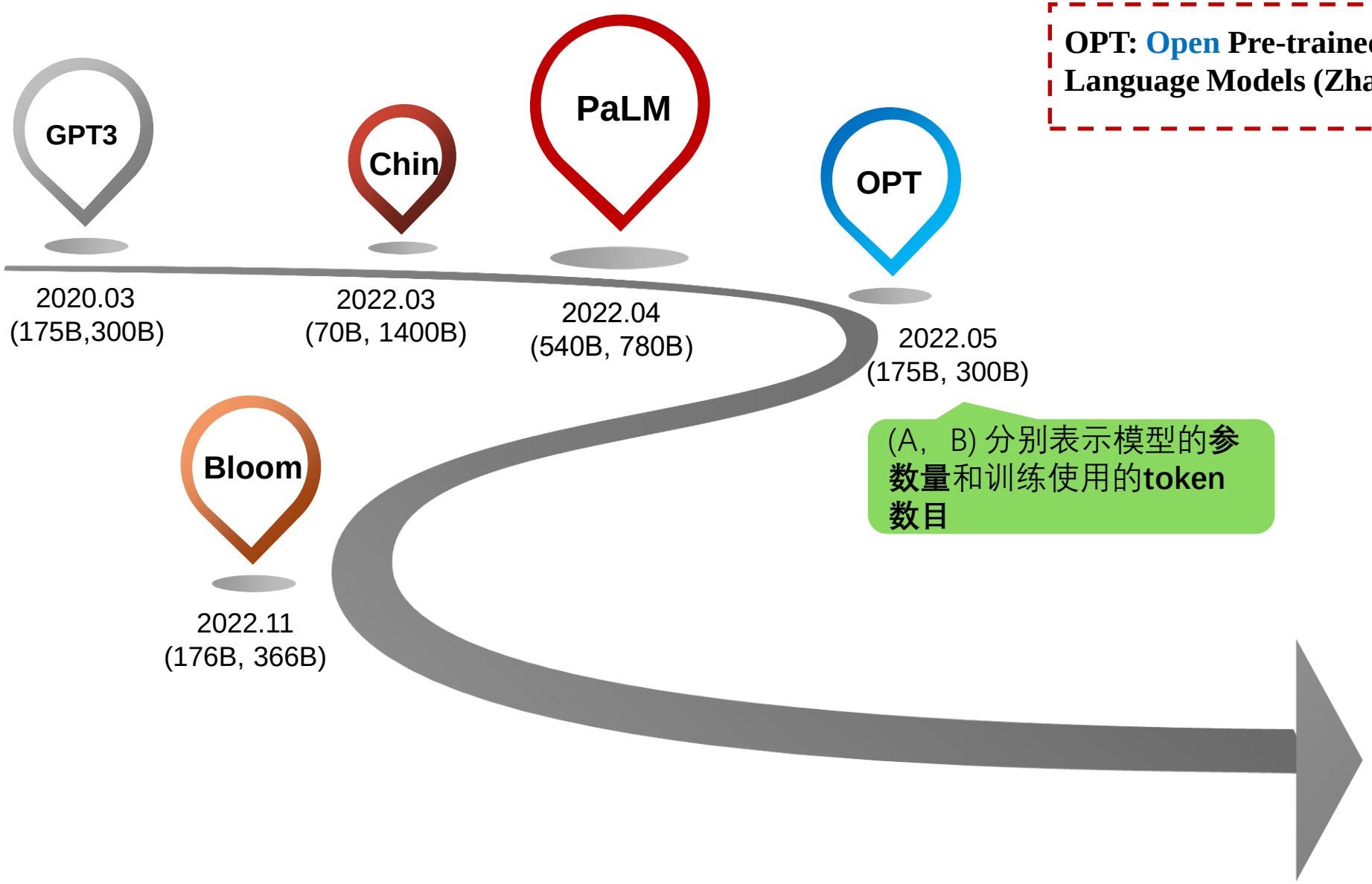
Prefixed LM

limited capacity

UNiLM/T5



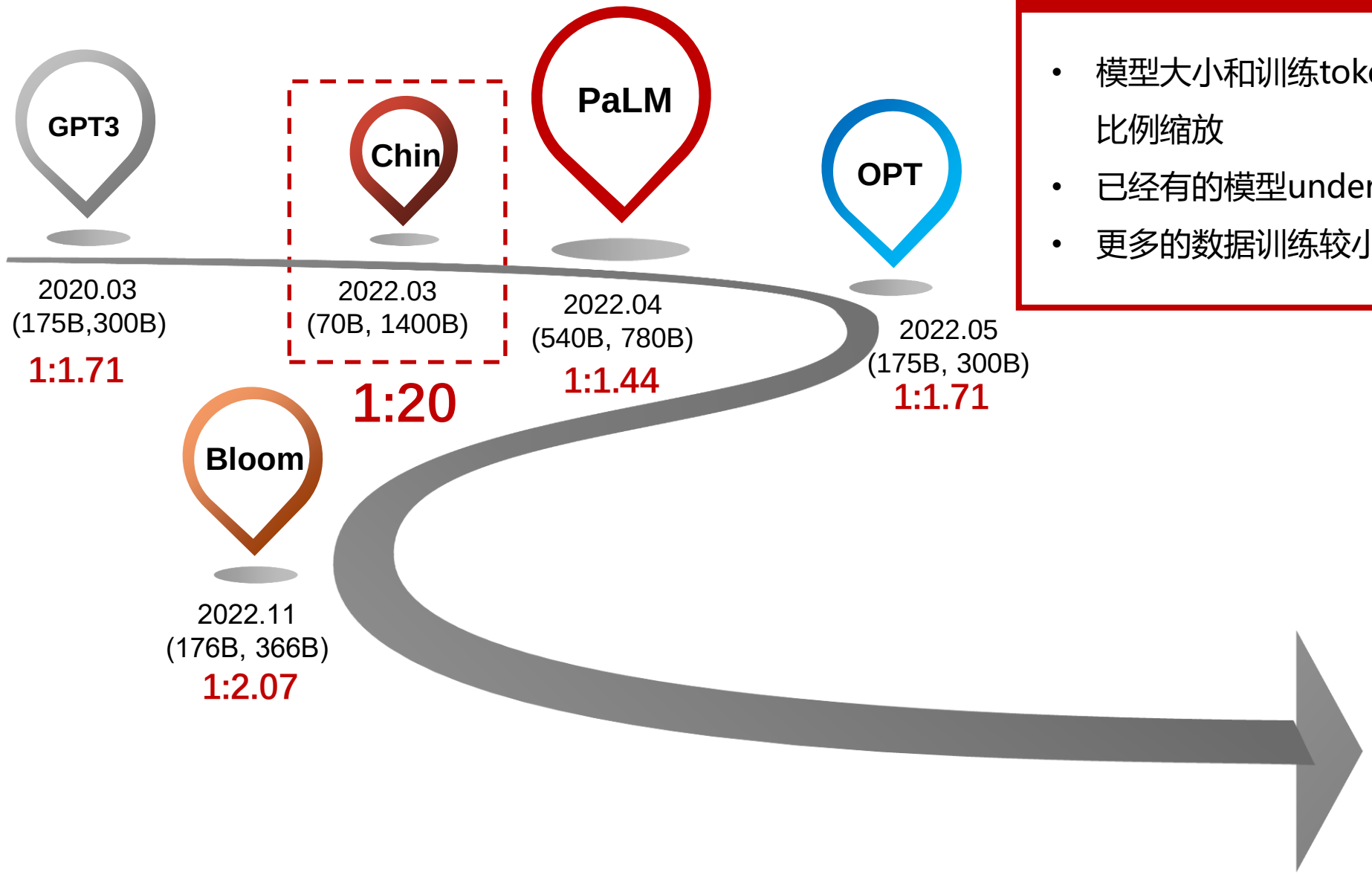
案例分享：LLaMa系列



OPT: Open Pre-trained Transformer Language Models (Zhang et. Al)

(A，B) 分别表示模型的**参数量**和训练使用的**token数目**

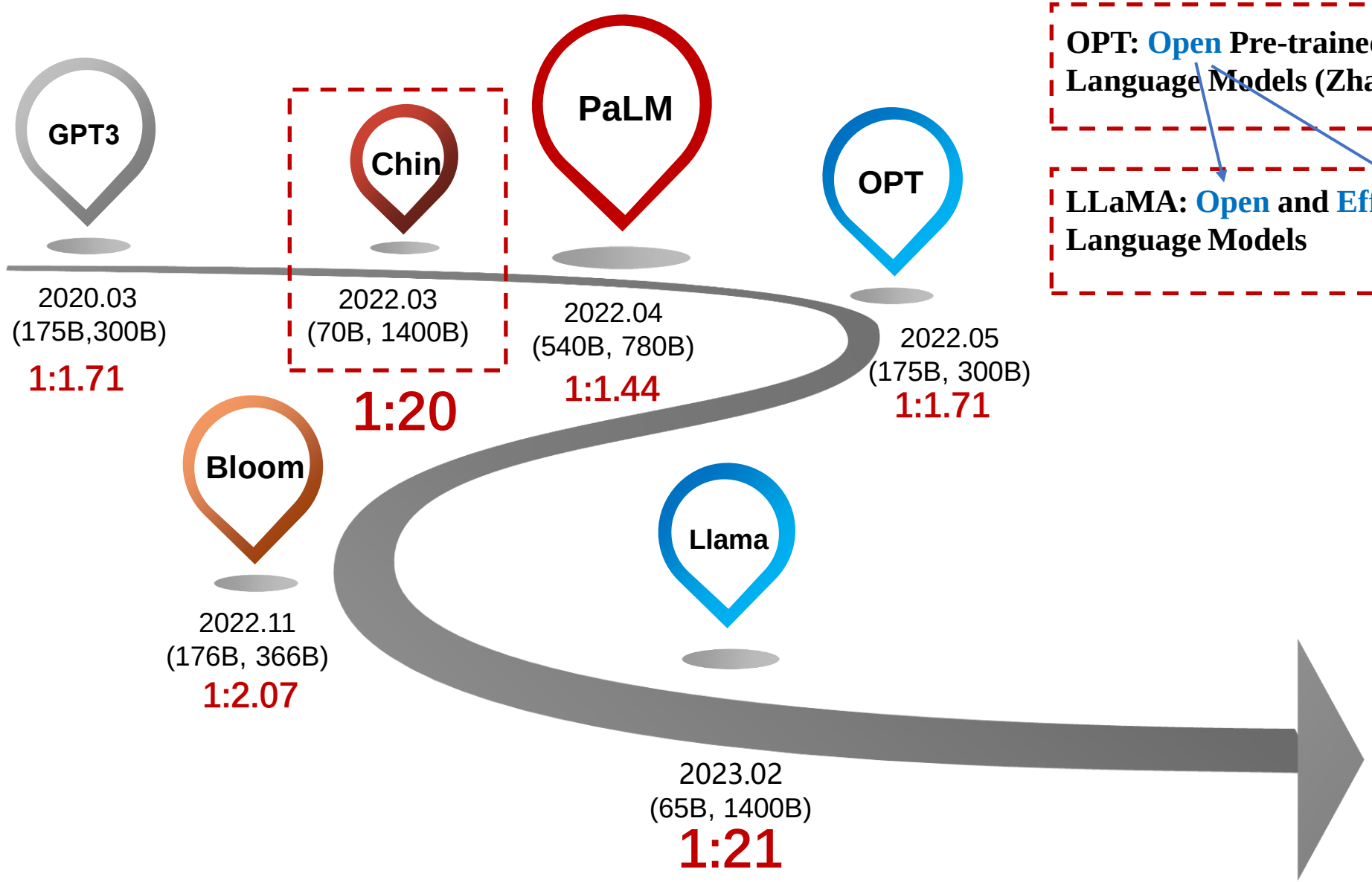
案例分享：LLaMa系列



Chinchilla Scaling Law

- 模型大小和训练token的数量应该按相等比例缩放
- 已经有的模型under-trained (over-sized)
- 更多的数据训练较小的模型表现更好

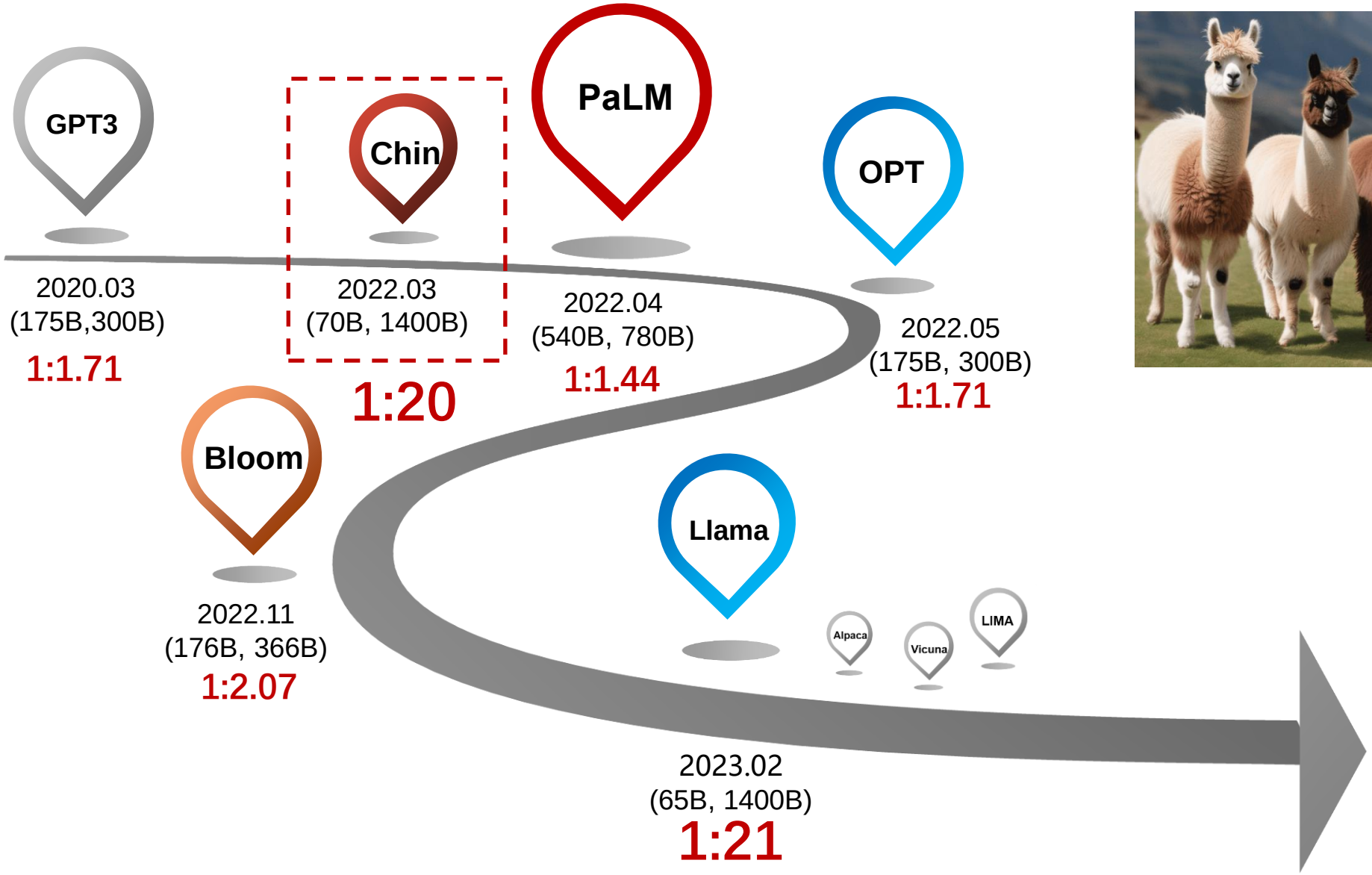
案例分享：LLaMa系列



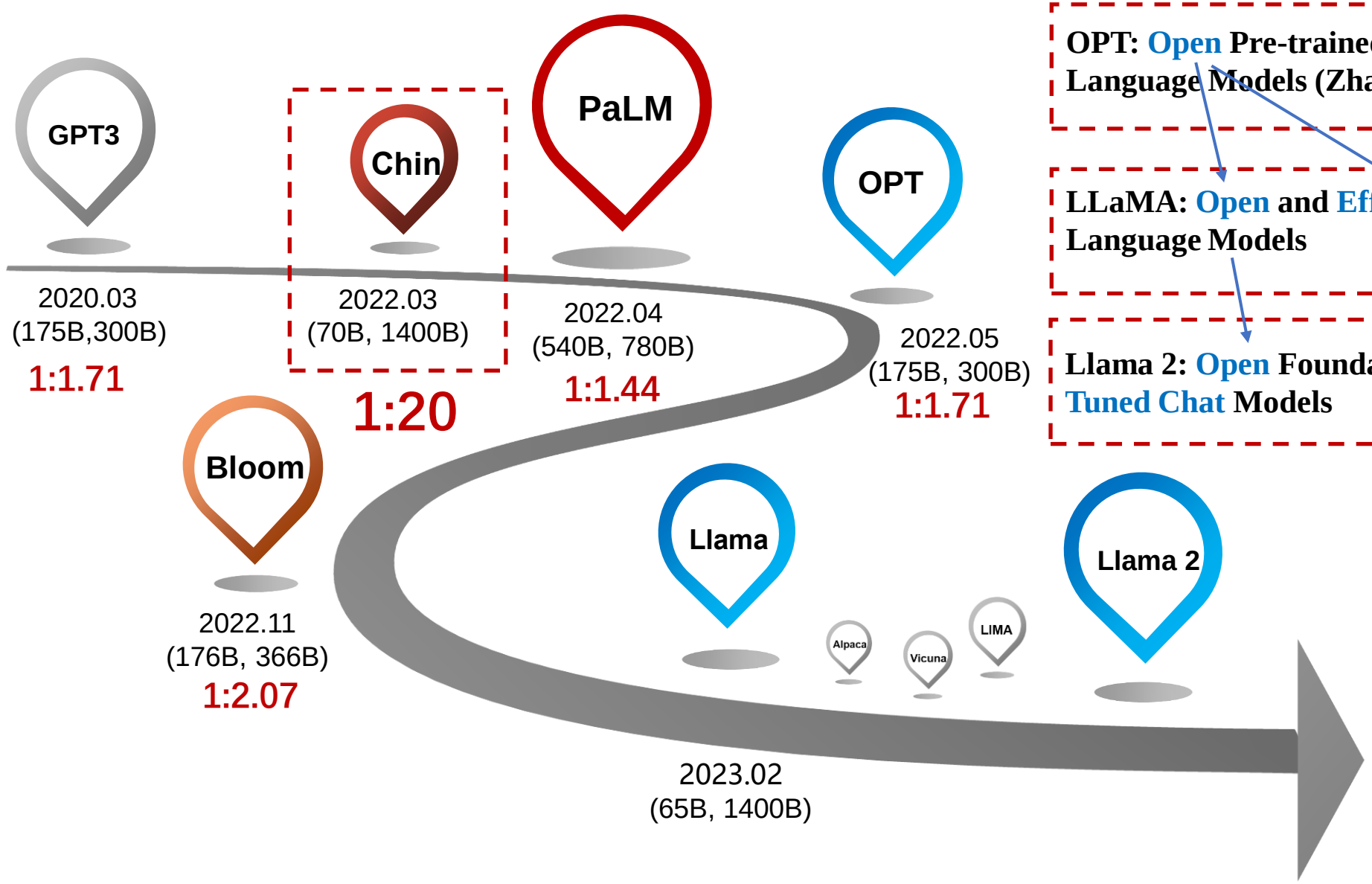
OPT: Open Pre-trained Transformer Language Models (Zhang et. Al)

LLaMA: Open and Efficient Foundation Language Models

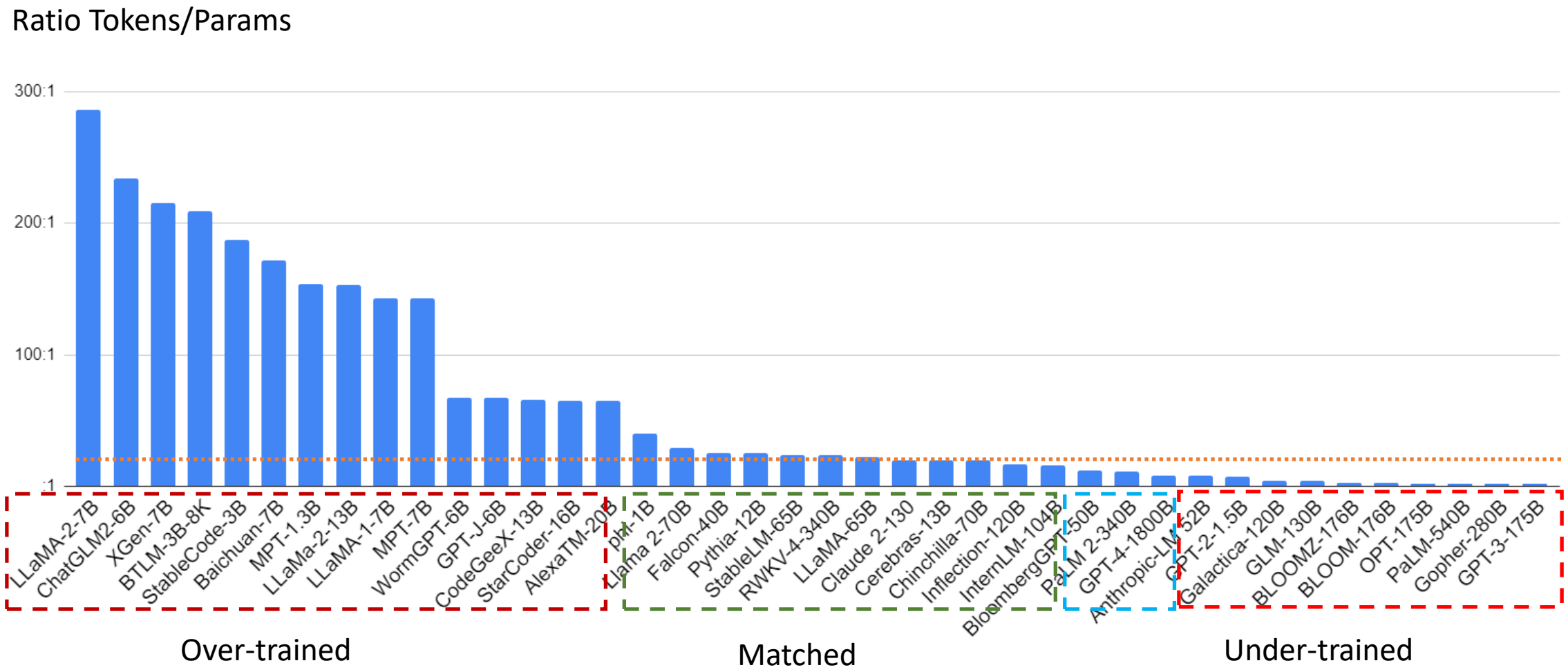
案例分享：LLaMa系列



案例分享：LLaMa系列



Chinchilla Scaling Law”视角下的模型发展规律



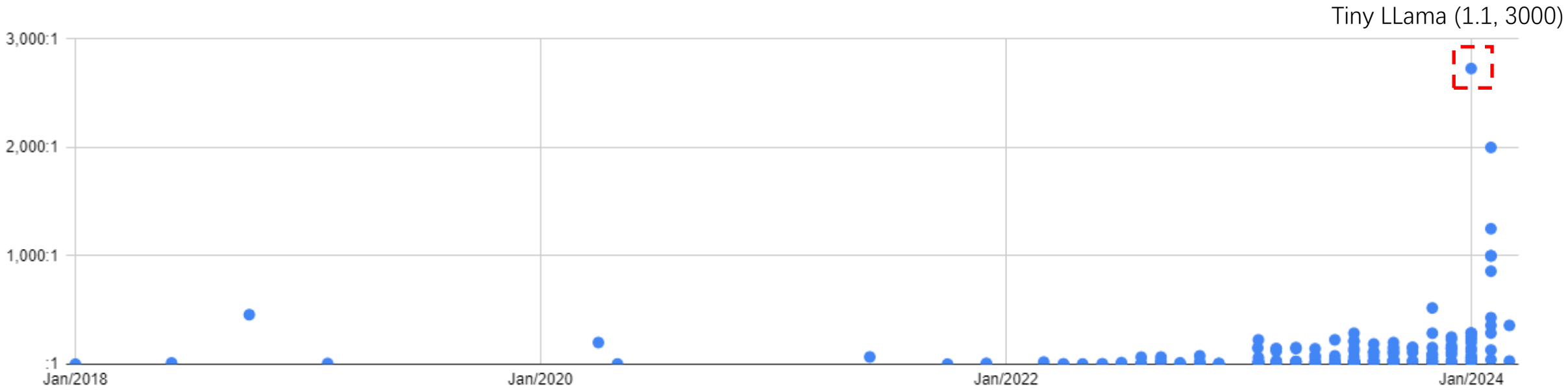
模型大小7B以内，训练token数大于500B

模型大小大于50B, 符合Chinchilla scaling law

大于 100B

Chinchilla Scaling Law”视角下的模型发展规律

Ratio Tokens/Params



统计从2028年以来160+代表性大语言模型

LLaMa2的预训练

数据组成

LlaMa 2相比LlaMa 1使用了更多的tokens

	Training Data	Params	Tokens	ChinLaw
LlaMa 1	CommonCrawl C4, Github Wikipedia Books Arxiv StackExchange	7B	1000B	140B
		13B	1000B	260B
		33B	1400B	660B
		65B	1400B	1300B
LlaMa 2	A new mix of publicly available online data	7B	2000B	140B
		13B	2000B	260B
		34B	2000B	680B
		70B	2000B	1400B

Language	Percent	Language	Percent
en	89.70%	uk	0.07%
unknown	8.38%	ko	0.06%
de	0.17%	ca	0.04%
fr	0.16%	sr	0.04%
sv	0.15%	id	0.03%
zh	0.13%	cs	0.03%
es	0.13%	fi	0.03%
ru	0.13%	hu	0.03%
nl	0.12%	no	0.03%
it	0.11%	ro	0.03%
ja	0.10%	bg	0.02%
pl	0.09%	da	0.02%
pt	0.09%	sl	0.01%
vi	0.08%	hr	0.01%

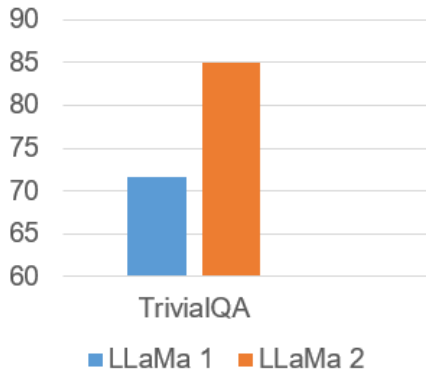
LLaMa2的预训练

- 数据 (处理细节)
  - 通过对最**为真实的来源**进行上采样，努力增加知识**并减弱幻觉**
  - 没有使用Meta用户数据
  - 去除了含大量个人个人信息的网站的数据
  - 没有对数据集进行额外的过滤

2.1 Pretraining Data

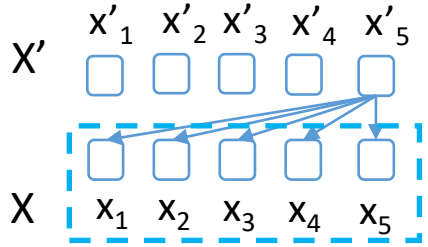
Our training corpus includes a new mix of data from publicly available sources, which does not include data from Meta's products or services. We made an effort to remove data from certain sites known to contain a high volume of personal information about private individuals. We trained on 2 trillion tokens of data as this provides a good performance–cost trade-off, up-sampling the most factual sources in an effort to increase knowledge and dampen hallucinations.

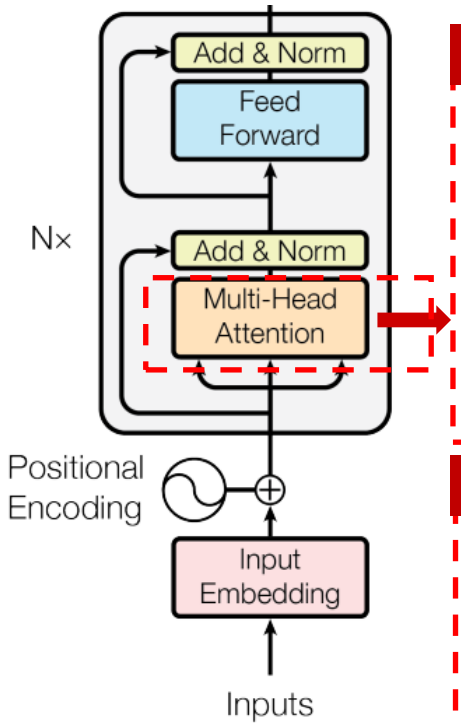
Data Sampling in LLaMa 2



LLaMa2的预训练

- 训练架构 &细节
  - transformer architecture





Single Head

$$X' = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where

$$K = X\,W^n$$
$$V = X\,W^v$$
$$Q = X\,W^q$$

Multi Head

$$X' = \text{MultiHeadAttention}(Q, K, V)$$
$$= \text{concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

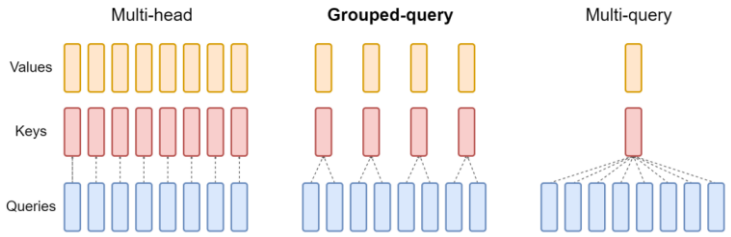
where

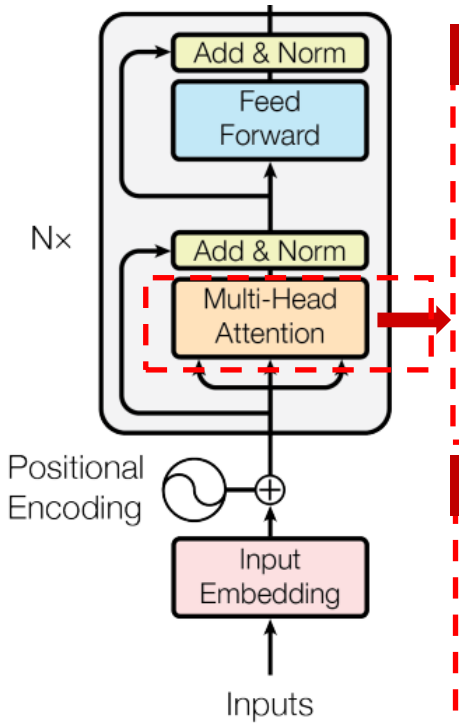
$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i)$$

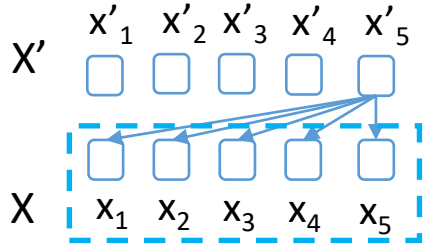
Attention Is All You Need, Ashish et al 2017

LLaMa2的预训练

- 训练架构 &细节
  - transformer architecture
  - Grouped-query attention







Single Head

$$X' = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where

$$K = X\,W^n$$
$$V = X\,W^v$$
$$Q = X\,W^q$$

Multi Head

$$X' = \text{MultiHeadAttention}(Q, K, V)$$
$$= \text{concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

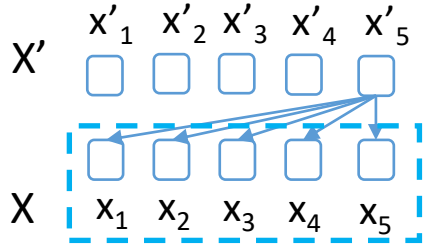
where

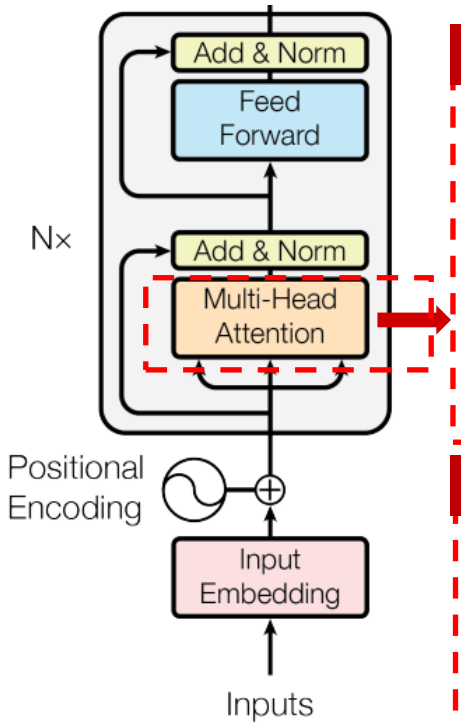
$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i)$$

GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints (Ainslie et al 2023)

LLaMa2的预训练

- 训练架构 &细节
  - transformer architecture
  - Grouped-query attention
  - KV-cache





Single Head

$$\mathrm{X}' = \mathrm{Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where

$$\mathrm{K} = \mathrm{X}\,\mathrm{W}^n$$
$$\mathrm{V} = \mathrm{X}\,\mathrm{W}^v$$
$$\mathrm{Q} = \mathrm{X}\,\mathrm{W}^q$$

Multi Head

$$\mathrm{X}' = \mathrm{MultiHeadAttention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)\mathrm{W}^O$$

where

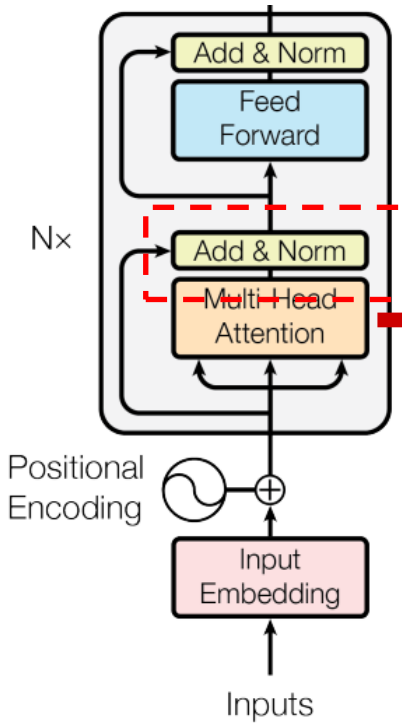
$$\mathrm{head_i} = \mathrm{Attention}(\mathrm{Q_i}, \mathrm{K_i}, \mathrm{V_i})$$

GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints (Ainslie et al 2023)

LLaMa2的预训练

- 训练架构 &细节
  - transformer architecture
  - Grouped-query attention
  - KV-cache
  - **Pre-normalization using RMSNorm**

Layer Normalization将向量规范化，使得经过处理的向量有着统一的均值和方差，以便于网络的学习和泛化



Post-Layer Normalization

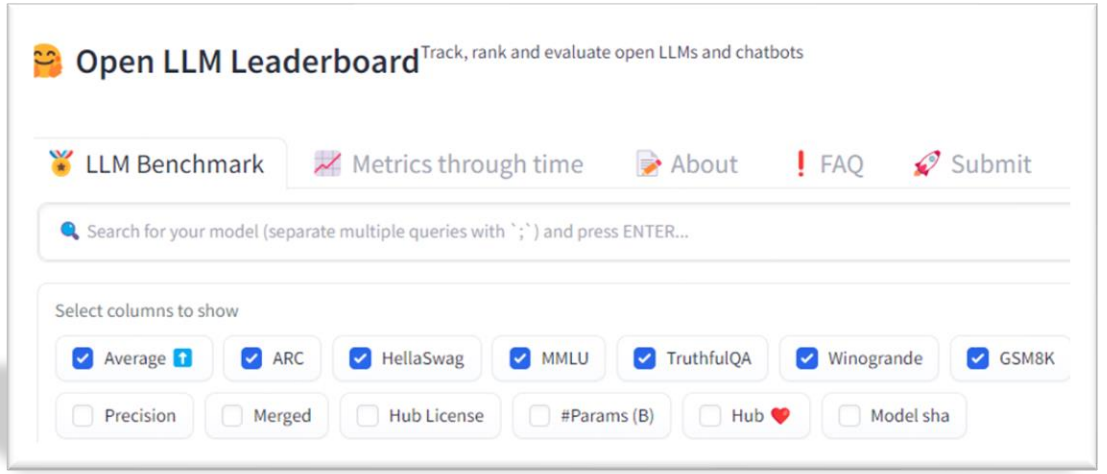
$$X' = \text{MultiHeadAttention}(X, Q, K, V)$$
$$X' = X + X'$$
$$X' = \text{LayerNorm}(X')$$

Pre-Layer Normalization

$$X' = \text{LayerNorm}(X)$$
$$X' = \text{MultiHeadAttention}(X', Q, K, V)$$
$$X' = X + X'$$

LLaMa2的预训练评估

- 评估Benchmark
  - **Code**: HumanEval, MBPP
  - **Commonsense Reasoning**
  - **World Knowledge**
    - NaturalQuestions, TriviaQA
  - **Reading Comprehension**
  - **Math Problems**
  - **Popular Aggregated Benchmark**
    - MMLU, BBH, AGI Eval (English only)
- 评估方式
  - Few-shot Prompting



LLaMa2的评估任务/数据

- Code - HumanEval

prompt

```
def has_close_elements(numbers: List[float], threshold: float) -> bool:
    """ Check if in given list of numbers, are any two numbers closer to each other than
    given threshold.
    >>> has_close_elements([1.0, 2.0, 3.0], 0.5)
    False
    >>> has_close_elements([1.0, 2.8, 3.0, 4.0, 5.0, 2.0], 0.3)
    True
    """
```

entrypoint

```
has_close_elements
```

test

```
def check(candidate):
    assert candidate([1.0, 2.0, 3.9, 4.0, 5.0, 2.2], 0.3)
    assert not candidate([1.0, 2.0, 3.9, 4.0, 5.0, 2.2], 0.05)
    assert candidate([1.0, 2.0, 5.9, 4.0, 5.0], 0.95)
    assert not candidate([1.0, 2.0, 5.9, 4.0, 5.0], 0.8)
```

LLaMa2的评估任务/数据

- Commonsense Reasoning - HellaSwag

context	A woman is outside with a bucket and a dog. The dog is running around trying to avoid a bath. She…
endings	A. rinses the bucket off with soap and blow dry the dog's head. B. uses a hose to keep it from getting soapy. C. gets the dog wet, then it runs away again. D. gets into a bath tub with the dog.
label	C

LLaMa2的评估任务/数据

- World Knowledge - NaturalQuestions

question	what color was john wilkes booth’s hair
wikipedia page	John Wilkes Booth
long answer	Some critics called Booth “the handsomest man in America” and a “natural genius”, and noted his having an “astonishing memory”; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair, and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a “muscular, perfect man” with “curling hair, like a Corinthian capital”.
short answer	jet-black

LLaMa2的评估任务/数据

- Reading Comprehension - SQuAD

context	Beyoncé Giselle Knowles-Carter (/biːˈjɒnseɪ/ bee-YON-say) (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child. ...
question	When did Beyonce start becoming popular?
answer	in the late 1990s
answer start	269

LLaMa2的评估任务/数据

- Math Problems - GSM8K

question

Janet’s ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?

answer

Janet sells 16 - 3 - 4 = <<16-3-4=9>>9 duck eggs a day. She makes 9 * 2 = $<<9*2=18>>18 every day at the farmer’s market. ####
18

LLaMa2的评估任务/数据

- Popular Aggregated Benchmark - BBH

inputs	Q: What movie does this emoji describe? 👨 👓 ⚡ \n choice: harry potter\n. choice: shutter island\n. choice: inglourious basterds\n. choice: die hard\n. choice: moonlight\nA:
targets	["harry potter"]
choices	["harry potter", "shutter island", "die hard", "inglourious basterds", "moonlight"]
choices scores	[1, 0, 0, 0, 0]

LLaMa2的评估任务/数据

- Llama 2 vs. Llama 1
  - 在Math, MMLU, BBH, AGI Eval上提升明显
  - 7B, 13B模型仍然有提升
  - 34B模型相对于其它大小的模型提升相对小

Model	Size	Code	Commonsense Reasoning	World Knowledge	Reading Comprehension	Math	MMLU	BBH	AGI Eval
MPT	7B	20.5	57.4	41.0	57.5	4.9	26.8	31.0	23.5
	30B	28.9	64.9	50.0	64.7	9.1	46.9	38.0	33.8
Falcon	7B	5.6	56.1	42.8	36.0	4.6	26.2	28.0	21.2
	40B	15.2	69.2	56.7	65.7	12.6	55.4	37.1	37.0
LLAMA 1	7B	14.1	60.8	46.2	58.5	6.95	35.1	30.3	23.9
	13B	18.9	66.1	52.6	62.3	10.9	46.9	37.0	33.9
	33B	26.0	70.0	58.4	67.6	21.4	57.8	39.8	41.7
	65B	30.7	70.7	60.5	68.6	30.8	63.4	43.5	47.6
LLAMA 2	7B	16.8	63.9	48.9	61.3	14.6	45.3	32.6	29.3
	13B	24.5	66.9	55.4	65.8	28.7	54.8	39.4	39.1
	34B	27.8	69.9	58.7	68.0	24.2	62.6	44.1	43.4
	70B	**37.5**	**71.9**	**63.6**	**69.4**	**35.2**	**68.9**	**51.2**	**54.2**

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

LLaMa2的评估任务/数据

- Llama 2 vs. Llama 1
  - 在Math, MMLU, BBH, AGI Eval上提升明显
  - 7B, 13B模型仍然有提升
  - 34B模型相对于其它大小的模型提升相对小
- 与闭源模型对比
  - Code & Math上与最好模型有很大差距

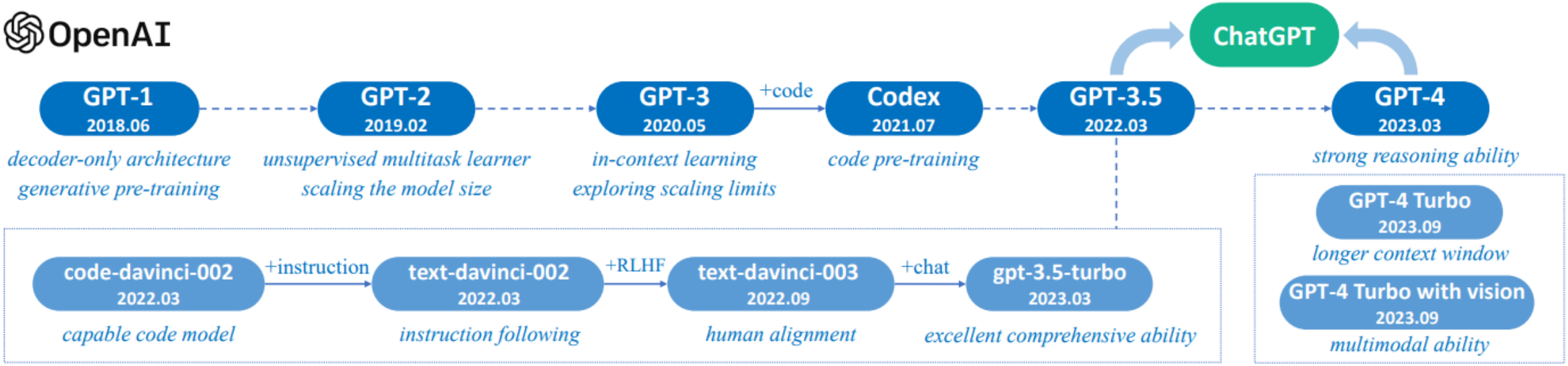
(CodeLlama, Lemur)

Model	Size	Code	Commonsense Reasoning	World Knowledge	Reading Comprehension	Math	MMLU	BBH	AGI Eval
MPT	7B	20.5	57.4	41.0	57.5	4.9	26.8	31.0	23.5
	30B	28.9	64.9	50.0	64.7	9.1	46.9	38.0	33.8
Falcon	7B	5.6	56.1	42.8	36.0	4.6	26.2	28.0	21.2
	40B	15.2	69.2	56.7	65.7	12.6	55.4	37.1	37.0
Llama 1	7B	14.1	60.8	46.2	58.5	6.95	35.1	30.3	23.9
	13B	18.9	66.1	52.6	62.3	10.9	46.9	37.0	33.9
	33B	26.0	70.0	58.4	67.6	21.4	57.8	39.8	41.7
	65B	30.7	70.7	60.5	68.6	30.8	63.4	43.5	47.6
Llama 2	7B	16.8	63.9	48.9	61.3	14.6	45.3	32.6	29.3
	13B	24.5	66.9	55.4	65.8	28.7	54.8	39.4	39.1
	34B	27.8	69.9	58.7	68.0	24.2	62.6	44.1	43.4
	70B	**37.5**	**71.9**	**63.6**	**69.4**	**35.2**	**68.9**	**51.2**	**54.2**

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

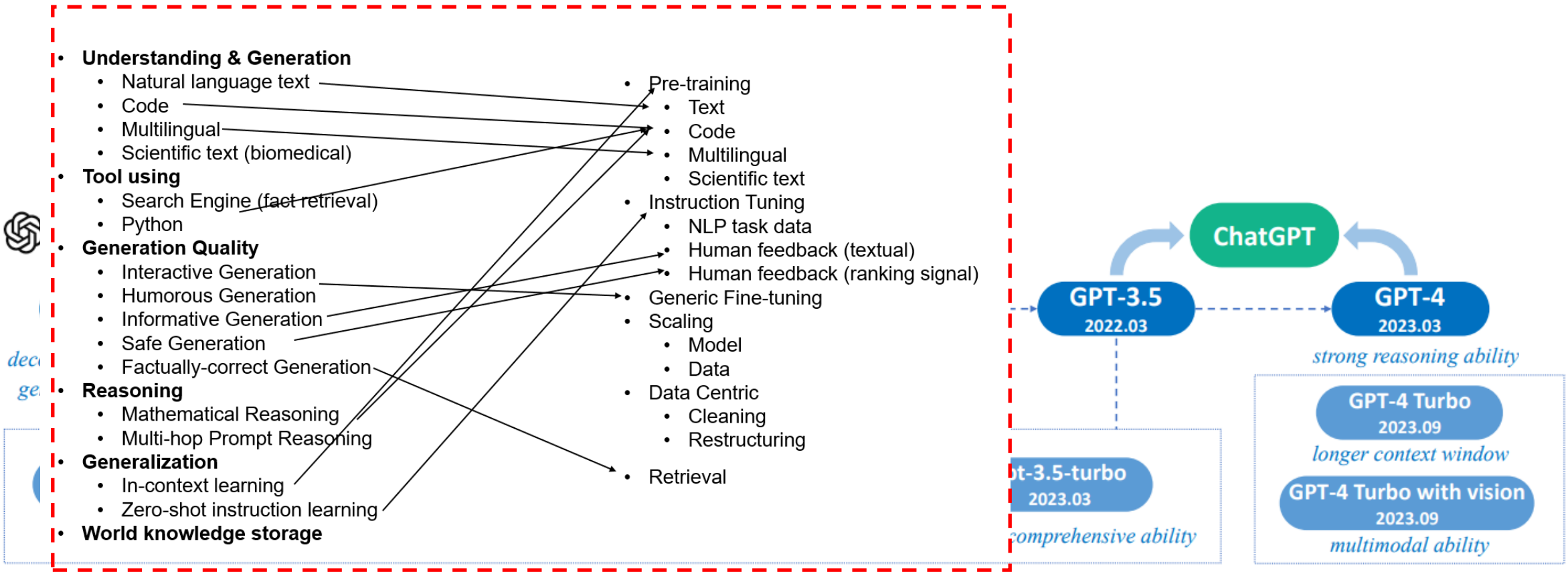
Code Llama: Open Foundation Models for Code, Roziere et al.2023
Lemur: Harmonizing Natural Language and Code for Language Agents, Xu et al.2023

大语言模型历史梳理 OpenAI视角



图来自于：https://arxiv.org/pdf/2303.18223.pdf

大语言模型历史梳理 OpenAI视角



图来自于：https://arxiv.org/pdf/2303.18223.pdf

大语言模型历史梳理 OpenAI视角

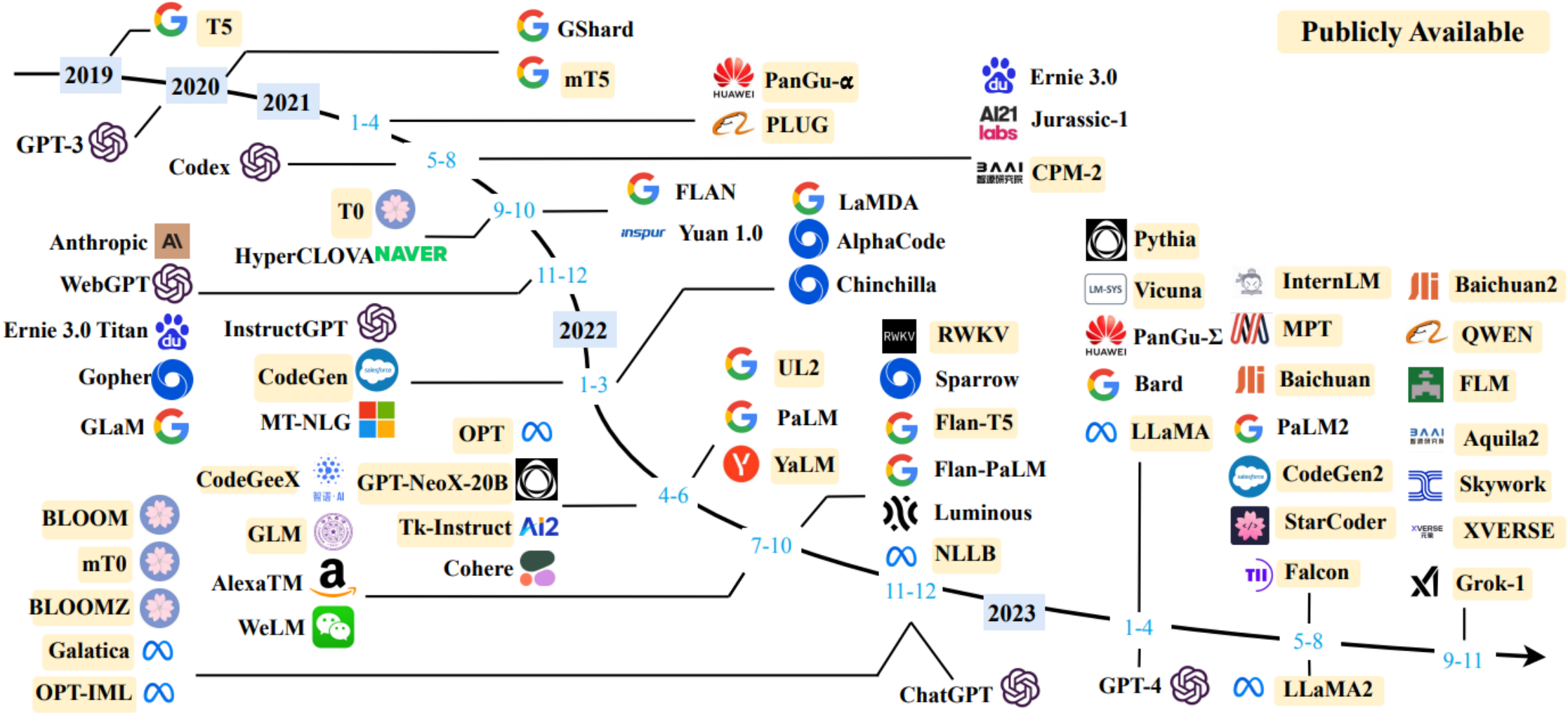
OpenAI 一直坚持“安全的 AGI”，但是路径上逐渐聚焦于大语言模型

关键决策：

- ☑ 迅速、深度、坚定选择了 Transformer 路线；
- ☑ 坚持走了从左到右自然语言生成路线，而不是自然语言理解路线；
- ☑ 意识到了“大”和“规模”的力量；
- ☑ GPT-3 后迅速引入了人类反馈；

2015 - 2016	2017 - 2018	2018 - 2019	2018 - 2019	2018 - 2019	2019 - 2020	2020 - 2021
关键决策						
早期 ML Engineering 能力和基础设施建设没有落后于行业，甚至目前比 Google 内部的还好用。	从 Unsupervised sentime-nt neuron 工作开始，逐渐将精力和关注点分配更多给语言模型上。	迅速和深度转向Transformer，没有在 CNN/RNN 等上一代特征提取器上浪费时间。	在行业对强化学习的效果充满争议的情况下，在 DOTA 及之后的项目中坚持探索深度强化学习。	在语言模型中坚持了仅有上文背景的 GPT 式生成式路线，没有追随 BERT 狂潮陷入理解式路线。	团队持续思考 Scaling Law 的问题，在 Transformer 基础上押注大规模数据和算力。	在长期强调安全和使用无监督强化学习的情况下，在 GPT-3 工作完成后迅速引入人类反馈。
争议或非共识						
AI 的突破是一项研究工作，而非工程问题； 每个探索 AGI 的公司在工程能力和基建并不会有明显差距。	OpenAI 的这个工作是优化别的任务时的副作用，歪打正着； 语言模型不是通往 AGI 的道路。	Transformer 彻底抛弃了之前 CNN、RNN 等网络结构； 前几年统治 AI 进展的 CV 圈并不买账 Transformer。	深度强化学习的效率非常低； 强化学习设置奖励函数非常 tricky； 它会陷入局部最优，并且通常难以稳定复现效果。	BERT 代表着未来，GPT 只是基于 Transformer 的过渡性技术； GPT 白白丢掉了下文的信息，在许多自然语言理解任务上都难以和 BERT 竞争。	AI 的进步来源于算法的创新； 算力在过去 10 年的进步不一定在未来 10 年持续。	随着模型变得更智能，Alignment 问题可以自动解决，人类反馈多此一举； 人类反馈违反了无监督的原教旨，并且缺少可拓展性。
OpenAI 的选择原因						
核心圈子内，没落后于业界趋势； 创始人 Greg Brockman 是工程能手和代码狂人； OpenAI 很早在 Gym/Unive-rse 上就遭遇工程挑战。	OpenAI 在研究中注重寻找 Signs of Life； OpenAI 想明白了理解与预测是有联系的，好的预测需要一定程度的理解，这个工作印证了这一原则。	Transformer 是 CapsNet（这是 Ilya 和导师 Hinton 做出的重要工作）的近亲，因为软注意力机制（Soft Attention）跟“协商路由”（Routing by Agreement）有很多理念相似点； 有人认为 Ilya 的 Neural GPU 工作某种程度上启发了 Transformer。	OpenAI 的创始人 Ilya 和 John 分别是深度学习和强化学习领域的引领者，可以忽略某些质疑； John 是 PPO、TRPO 等强化学习算法的发明者，它们就是要克服这些业界质疑的问题。	一定的运气，Unsupervised sentiment neuron 是BERT 出现前的工作； OpenAI 瞄准的目标是 AGI，因此目标用例是自然语言生成，这恰好连带解决了自然语言理解问题。	顶尖业界探索者逐渐形成共识，Rich Sutton 在 19 年发布了 *The Bitter Lesson*； OpenAI 经过 Five 和 Dota 项目更加对数据和算力的进步有信仰，提出了 *Scaling Law*，并且引入了足够资源尝试 GPT-3。	安全一直是 OpenAI 比同行强调更多的，OpenAI 从 17 年就和 Deepmind 做了从少量人类反馈中优化强化学习代理表现的工作； OpenAI 积累了的强化学习人才和基建，反应速度快，从人工标注到让 AI 辅助，终极目标是让 AI 反馈 AI。

大语言模型历史梳理 全局视角



图来自于：https://arxiv.org/pdf/2303.18223.pdf

大语言模型相关资源：文献

- A Survey of Large Language Models, Zhao et al.2023
- Pre-trained models for natural language processing: A survey
- Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing

大语言模型相关资源：周边

PLM Emoji (English)

PLM Emoji (Chinese)

https://plms.ai/peripherals/index.html