



大模型—人工智能 的大力神和弄潮儿

清华大学人工智能研究院常务副院长
孙茂松

第四届OpenI/O启智开发者大会NLP大模型论坛
2023年2月24日

提纲

一、大模型“画像”

二、从跟跑到领跑:下一步怎么办?

提纲

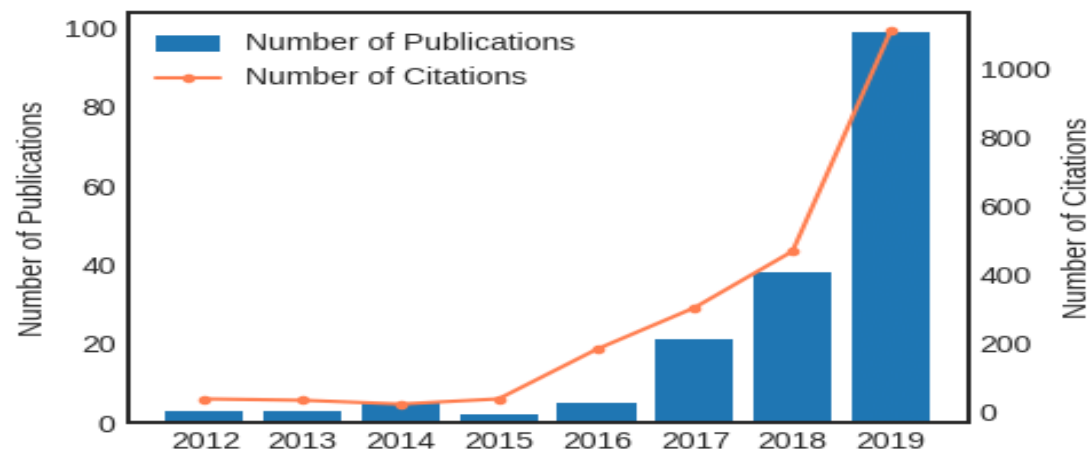
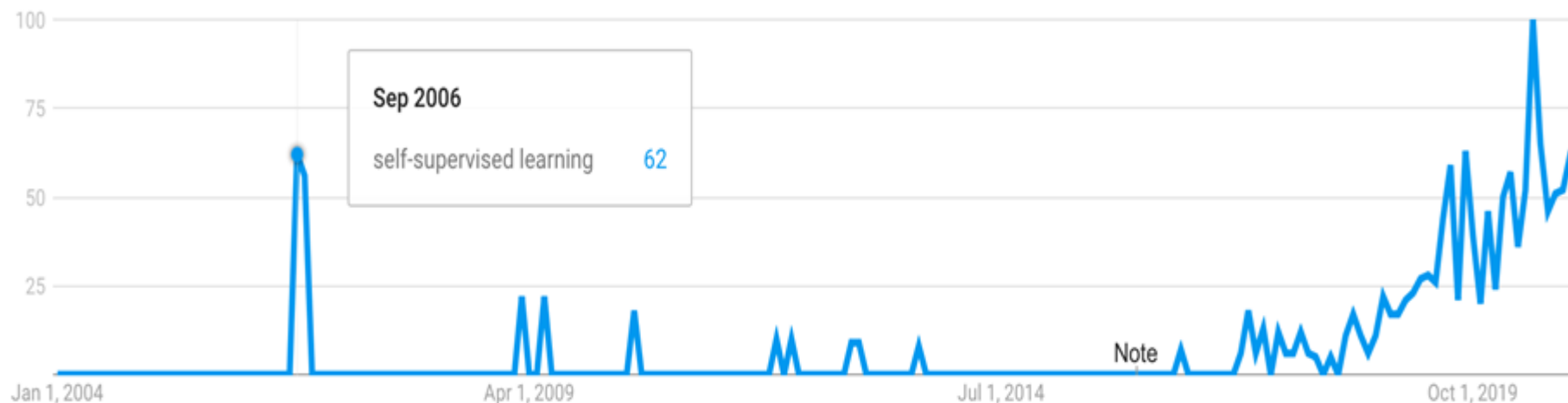
一、大模型“画像”

二、从跟跑到领跑:下一步怎么办?

大道至简：语言的生成模型(GPT-3)

- 我昨天晚上吃了？
- 我昨天晚上吃了北京？
- 经典问题：香农信息论
- N阶Markov模型：
 - 离散化符号，上下文长度受到严重限制
- 向量化表示：超越上下文长度限制
- 词向量：自动发现词与词之间的相关性
- 从词向量(有限集)到句子、篇章向量（无限集）

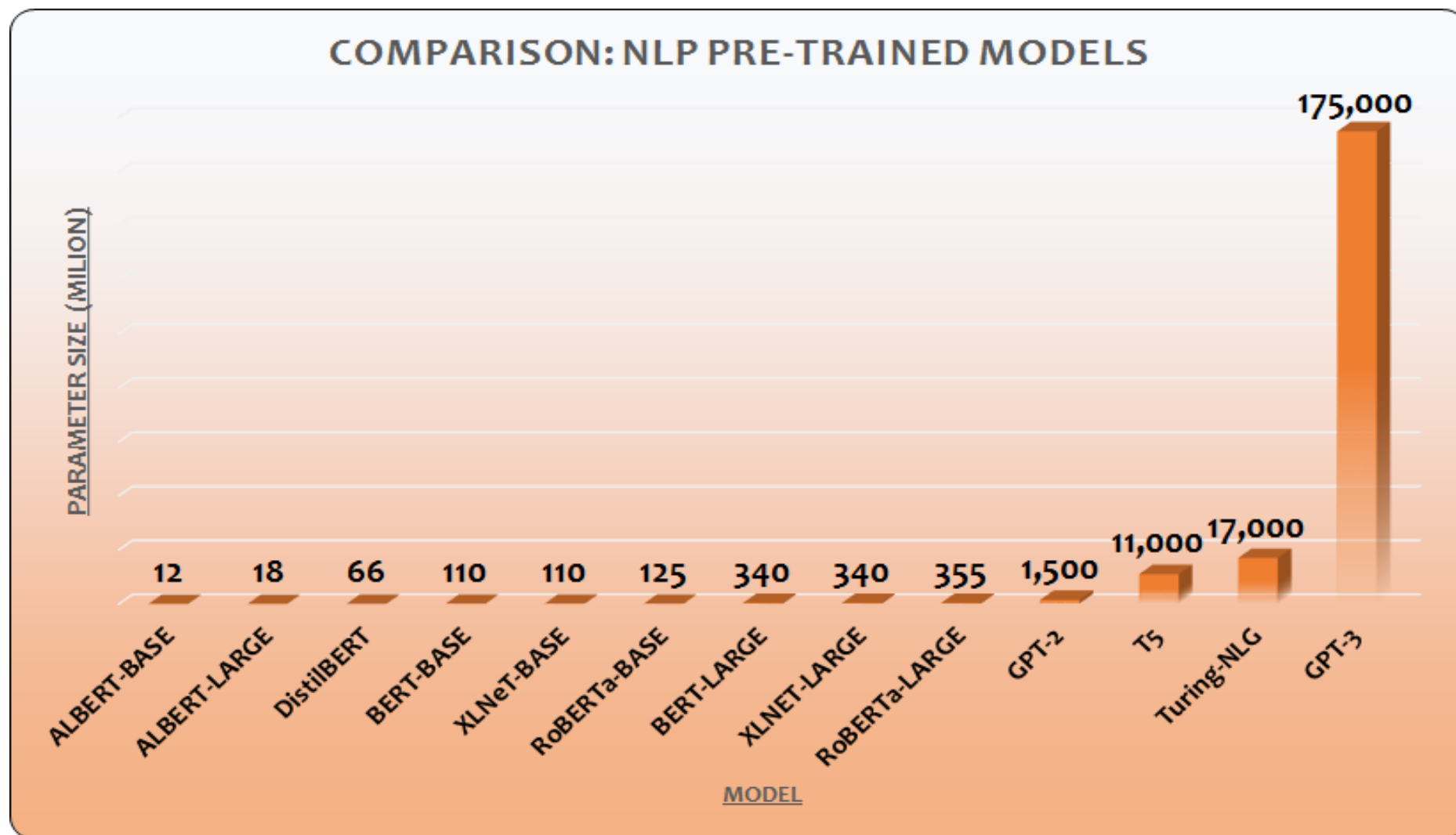
大道的的方法内涵是什么？自监督学习



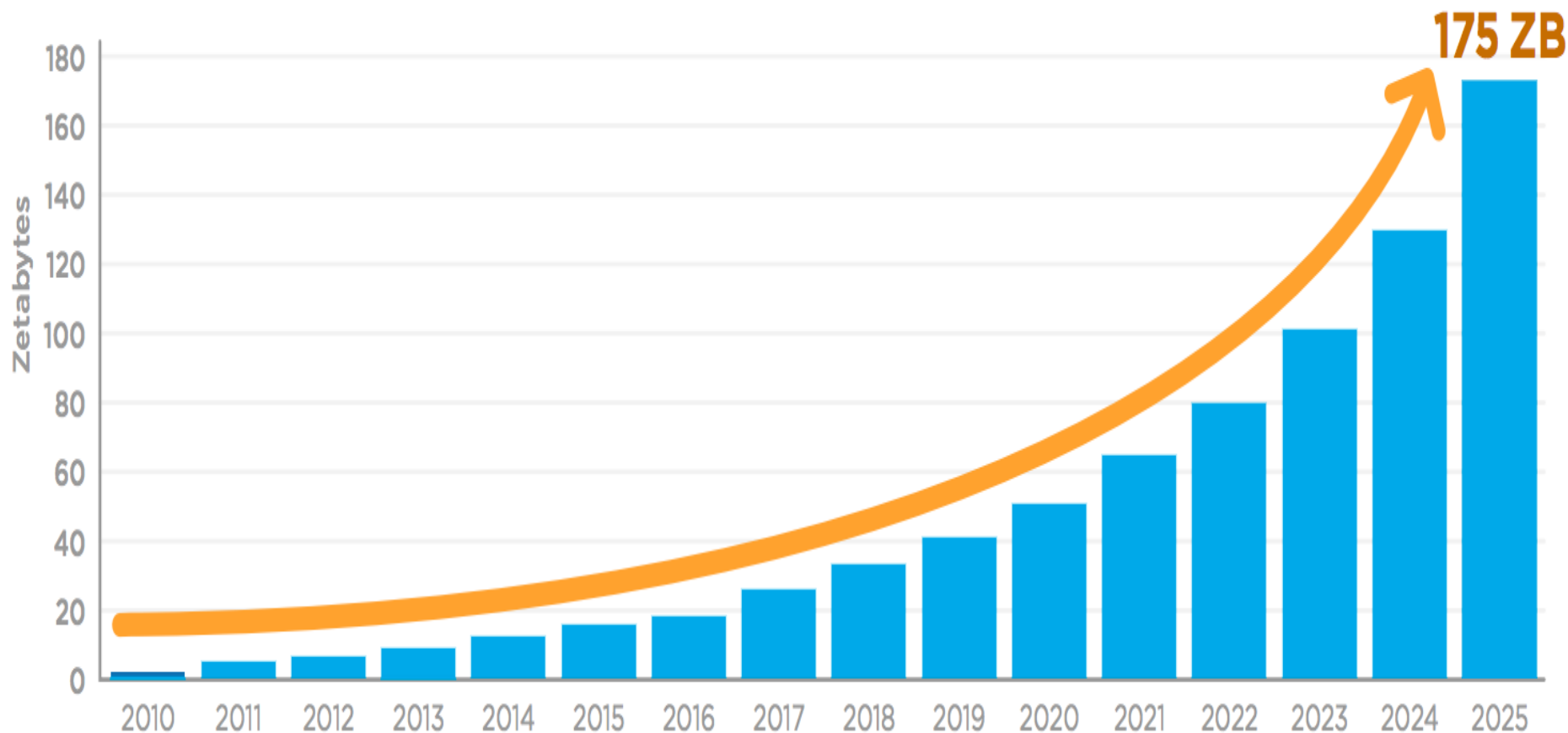
Number of papers containing the term “self-supervised learning”

<https://jonathanbgn.com/2020/12/31/self-supervised-learning.html>

语言大模型：力拔山兮气盖世



语言大模型：与大数据相得益彰



Annual Size of the Global Datasphere (Source: IDC Data Age 2025)

<https://jonathanbgn.com/2020/12/31/self-supervised-learning.html>

SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems



ML²



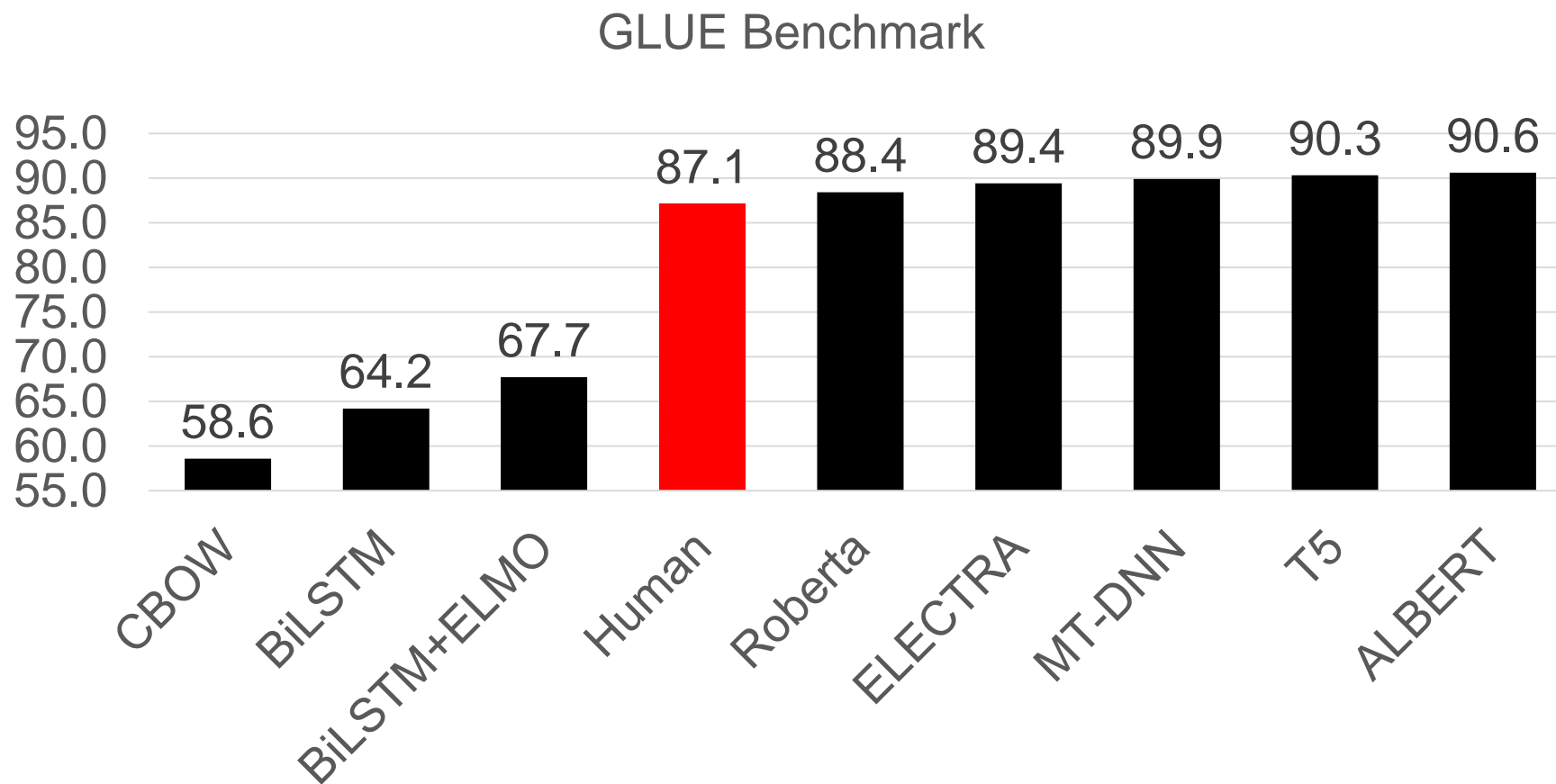
facebook Artificial Intelligence



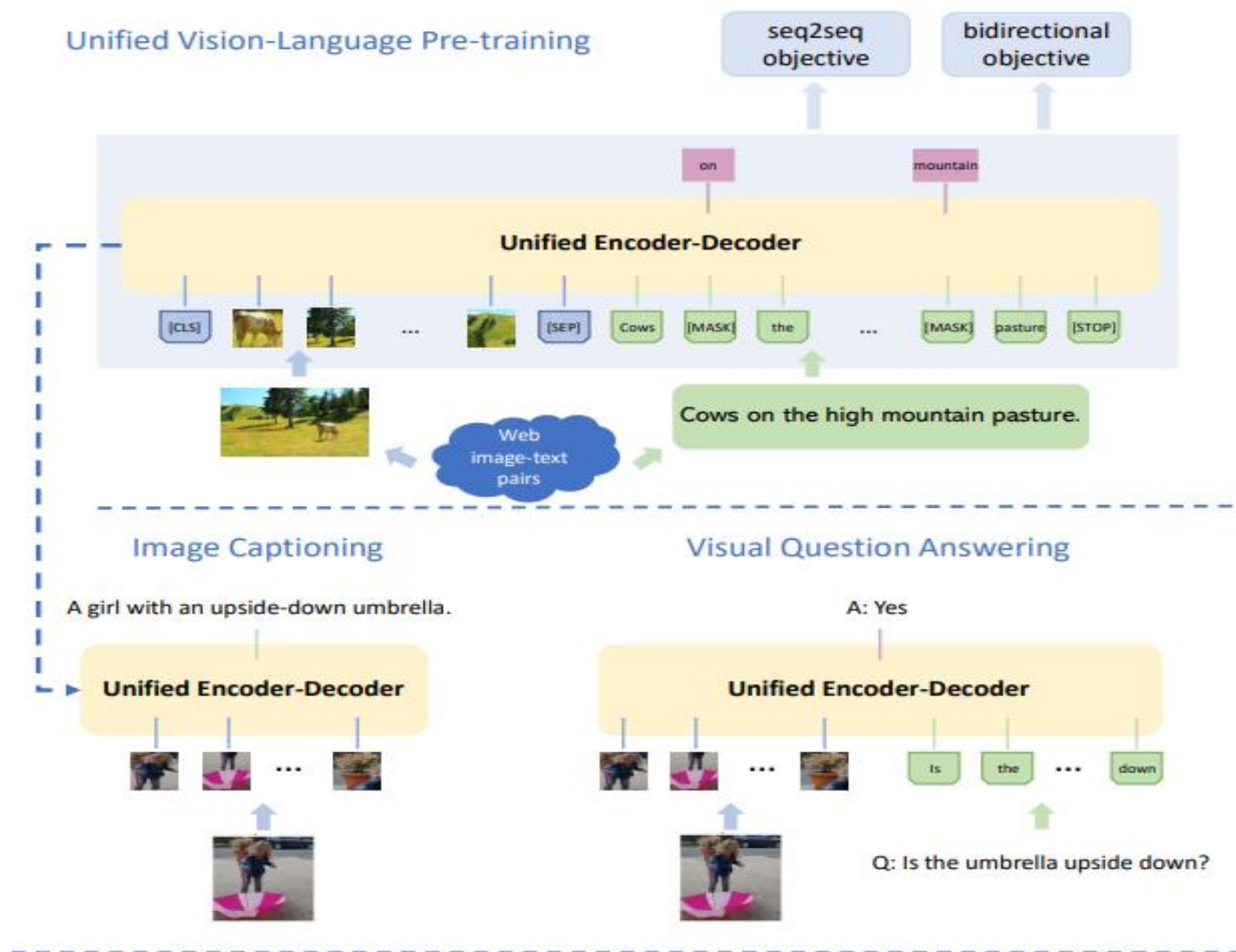
SAMSUNG Research

<https://arxiv.org/pdf/1905.00537.pdf>

Superior Performance on Language Understanding



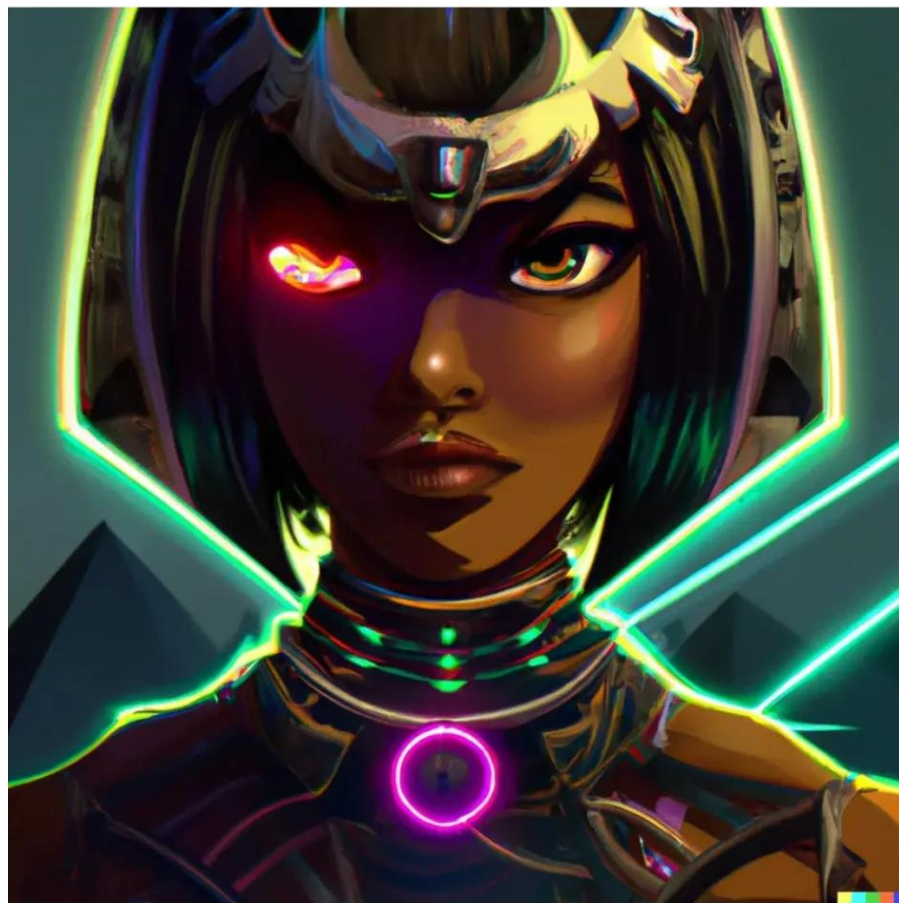
大模型之“一招鲜吃遍天”：多模态融合



大模型： 效果很奇幻

OpenAI: DALL-E2

赛博朋克埃及艳后，动漫，数字艺术

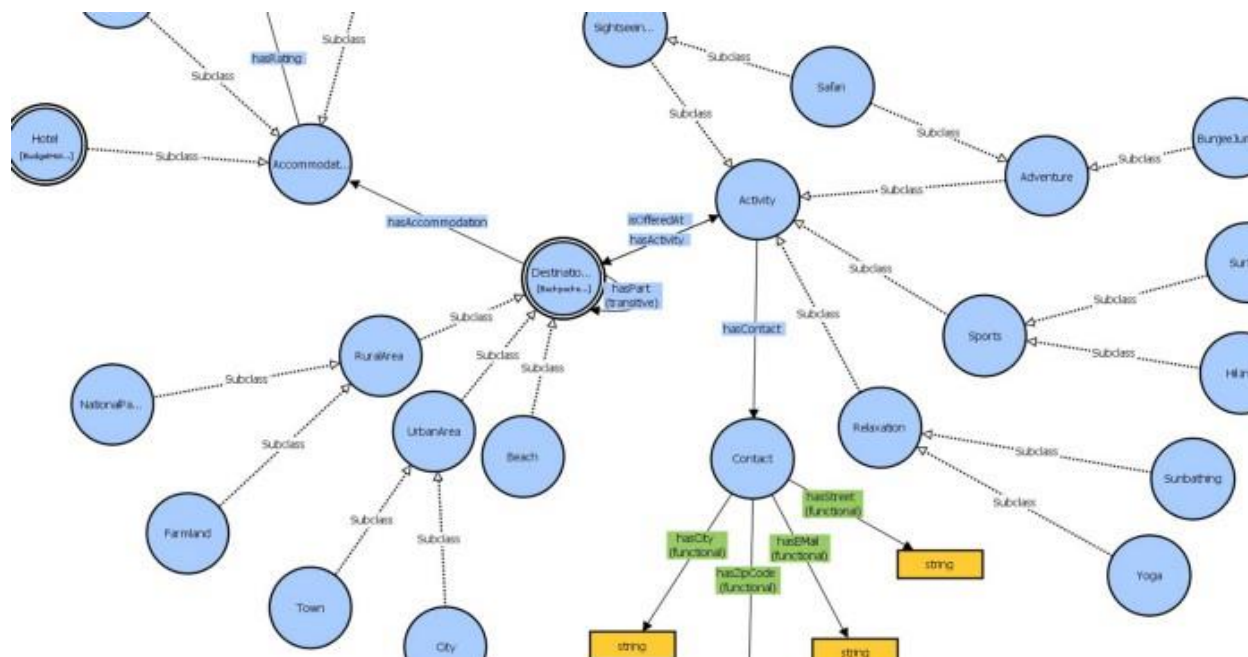
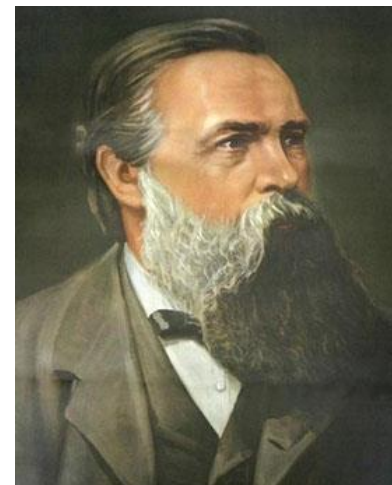


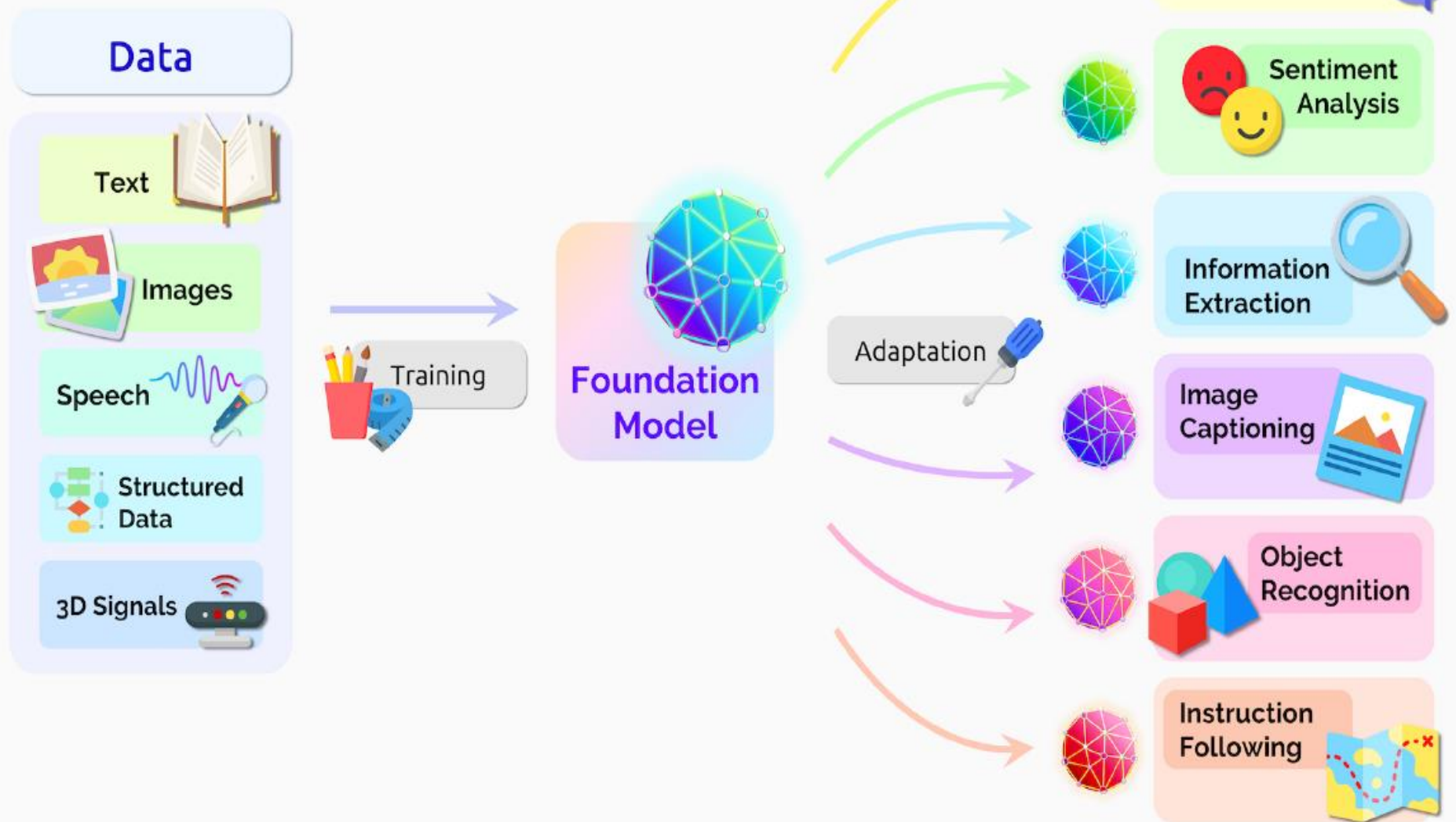
提示：美丽华丽的罗马尼亚女神，白色莲花花冠，栗色发髻，华丽的莲花。16K分辨率的概念艺术，萨金特、Alena Aenami、虚幻引擎5、动态照明、立体照明、丰富的暖色、日落



进一步的哲学思考

恩格斯谈事物普遍联系的“辩证图景”：
“当我们深思熟虑地考察自然界或人类历史或我们自己的精神活动的时候，首先呈现在我们眼前的，是一幅由种种联系和相互作用无穷无尽地交织起来的画面。”





Yann LeCun: AI Doesn't Need Our Supervision

> Meta's AI chief says self-supervised learning can build the metaverse and maybe even human-level AI

BY [ELIZA STRICKLAND](#) | 22 FEB 2022 | 6 MIN READ | [🔖](#)

Judea Pearl tweeted the caustic question: “What is the scientific principle by which ‘Foundation models’ can circumvent the theoretical limitations of data-centric methods as we know them...?”

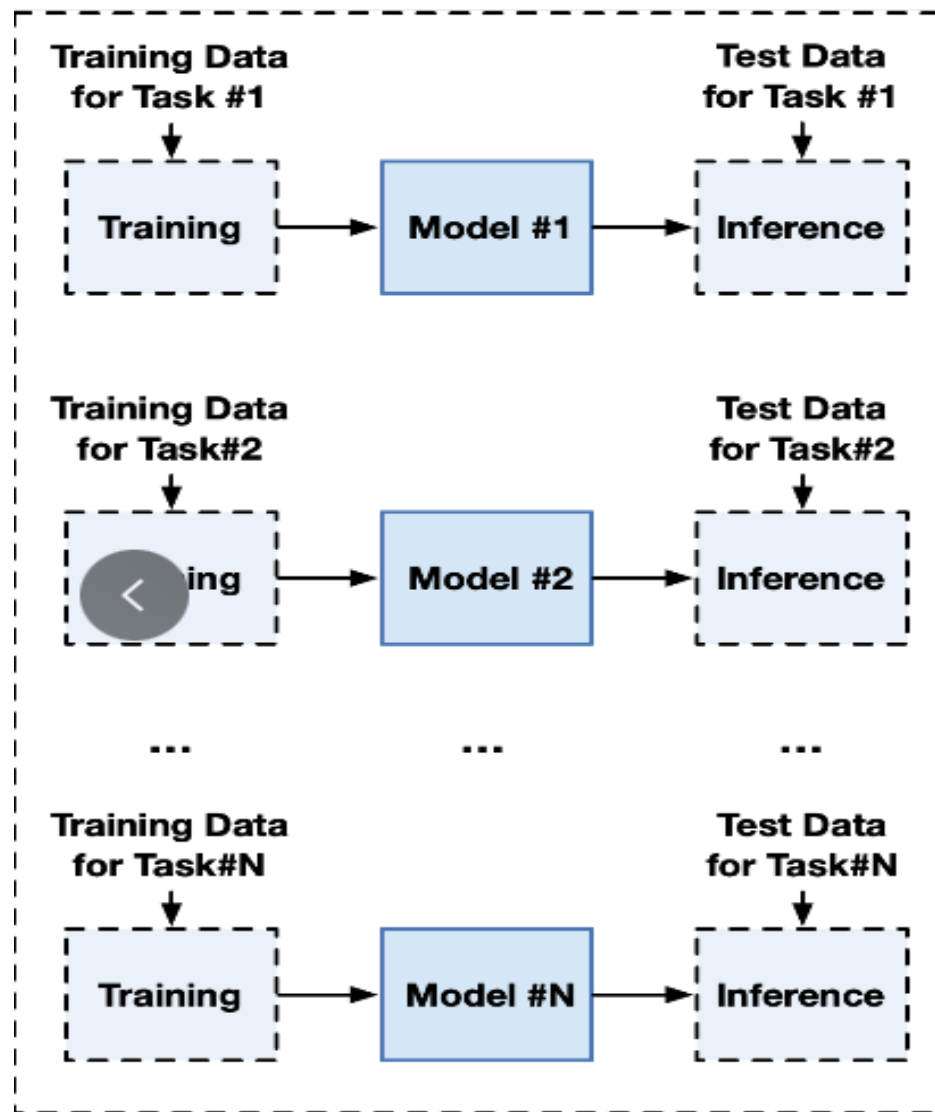


提纲

一、大模型“画像”

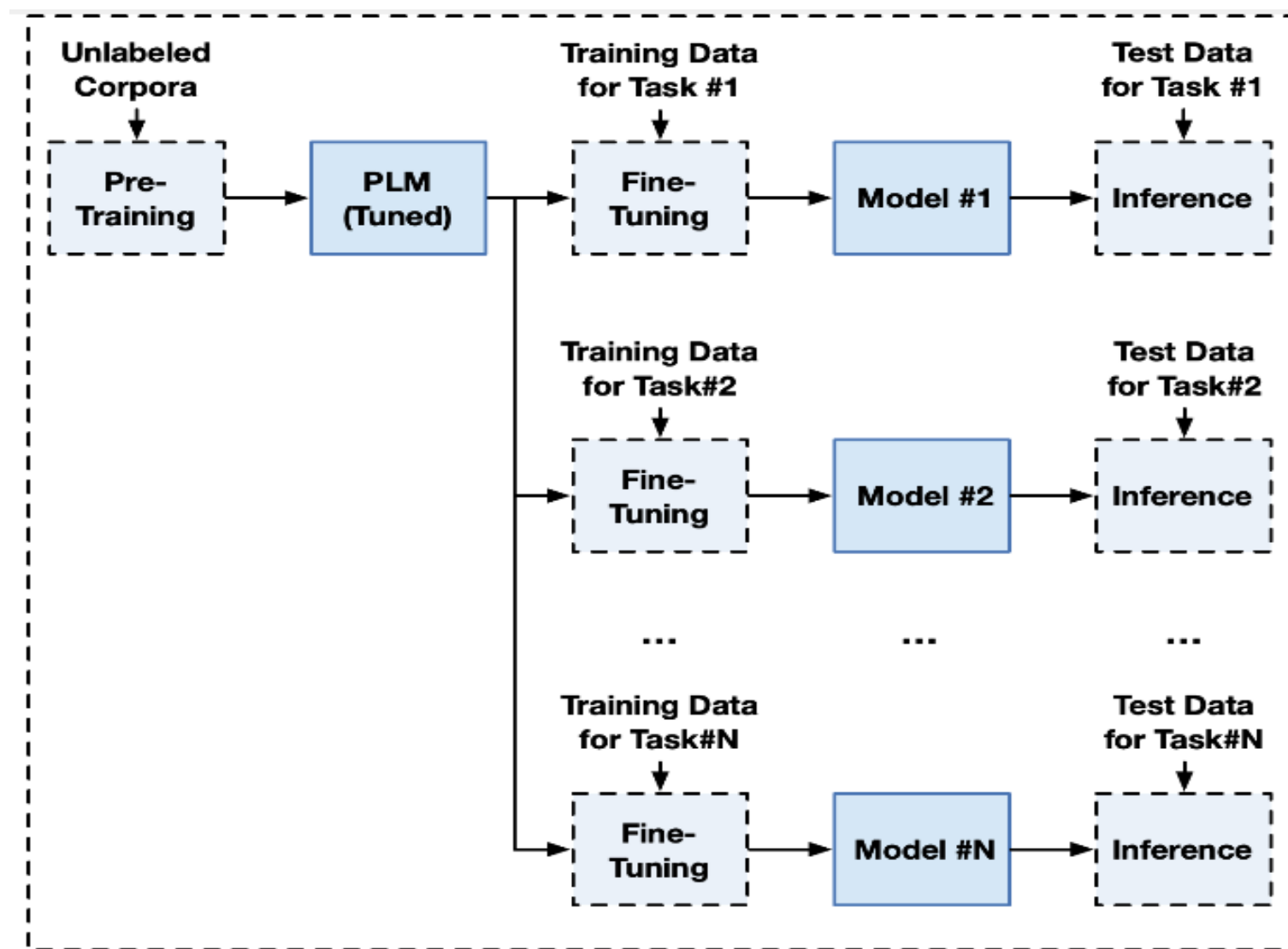
二、从跟跑到领跑：下一步怎么办？

深度学习第一重境界：白手起家 + 各家自扫门前雪



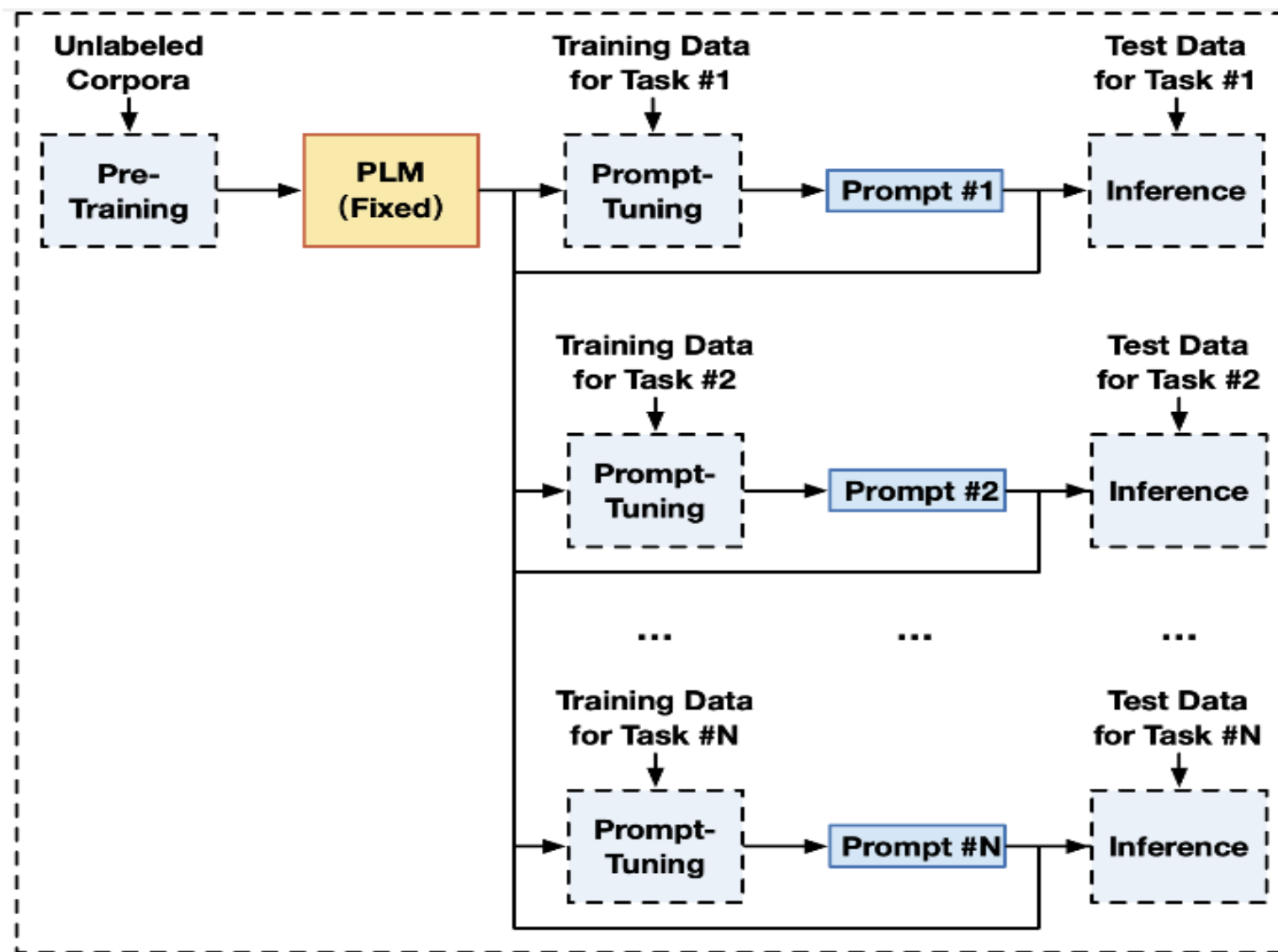
(a) Learning from Scratch

深度学习第二重境界：预训练模型+大小联调



(b) Pre-Training and Fine-Tuning

深度学习第三重境界：预训练大模型 + 一巨托众小



(c) Pre-Training and Prompt-Tuning

大模型：性能很奇特

Language Models are Few-Shot Learners

The three settings we explore for in-context learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1 Translate English to French:  ← task description
2 cheese => .....           ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1 Translate English to French:  ← task description
2 sea otter => loutre de mer    ← example
3 cheese => .....              ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1 Translate English to French:  ← task description
2 sea otter => loutre de mer    ← examples
3 peppermint => menthe poivrée ←
4 plush girafe => girafe peluche ←
5 cheese => .....              ← prompt
```

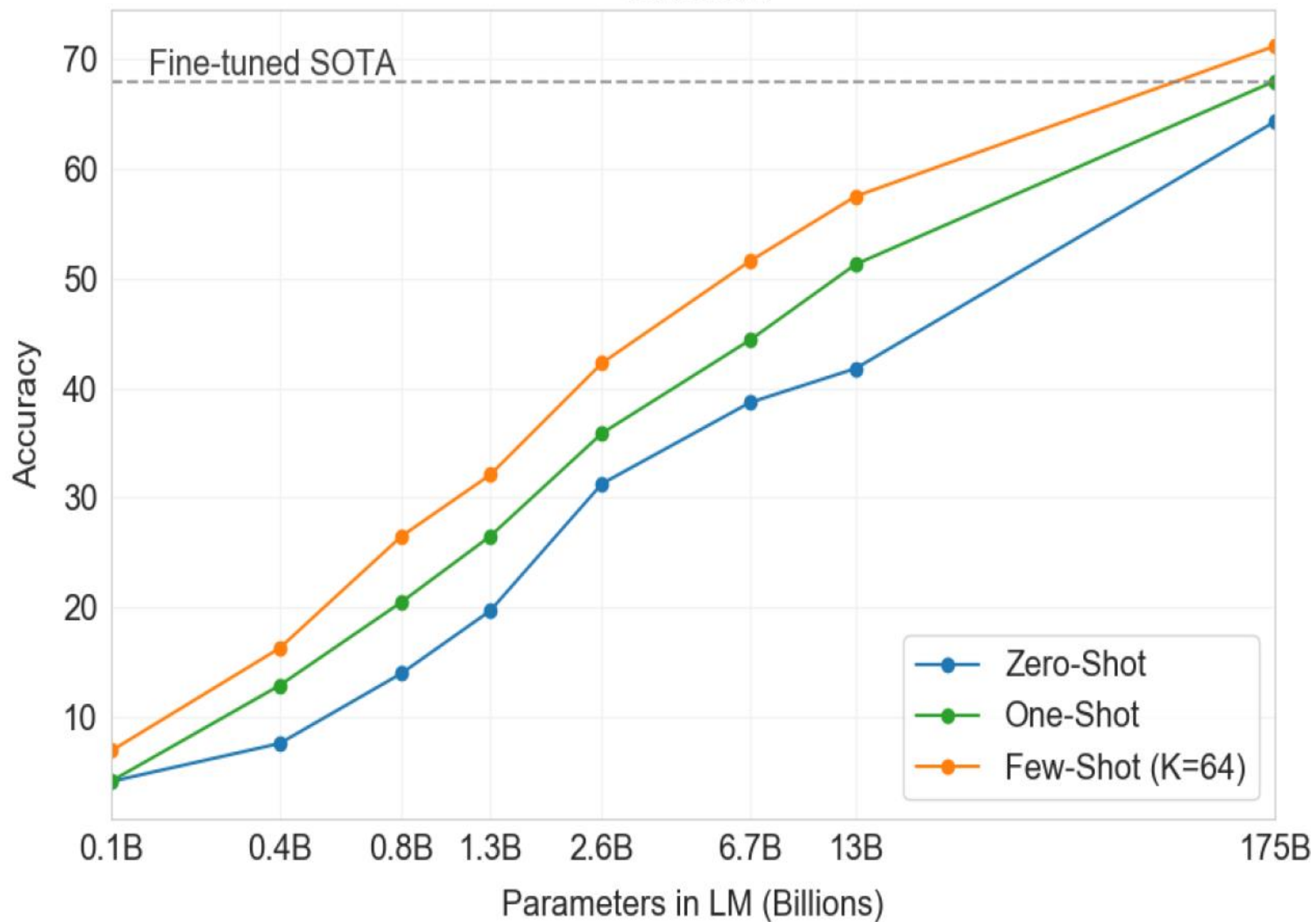
Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.



TriviaQA

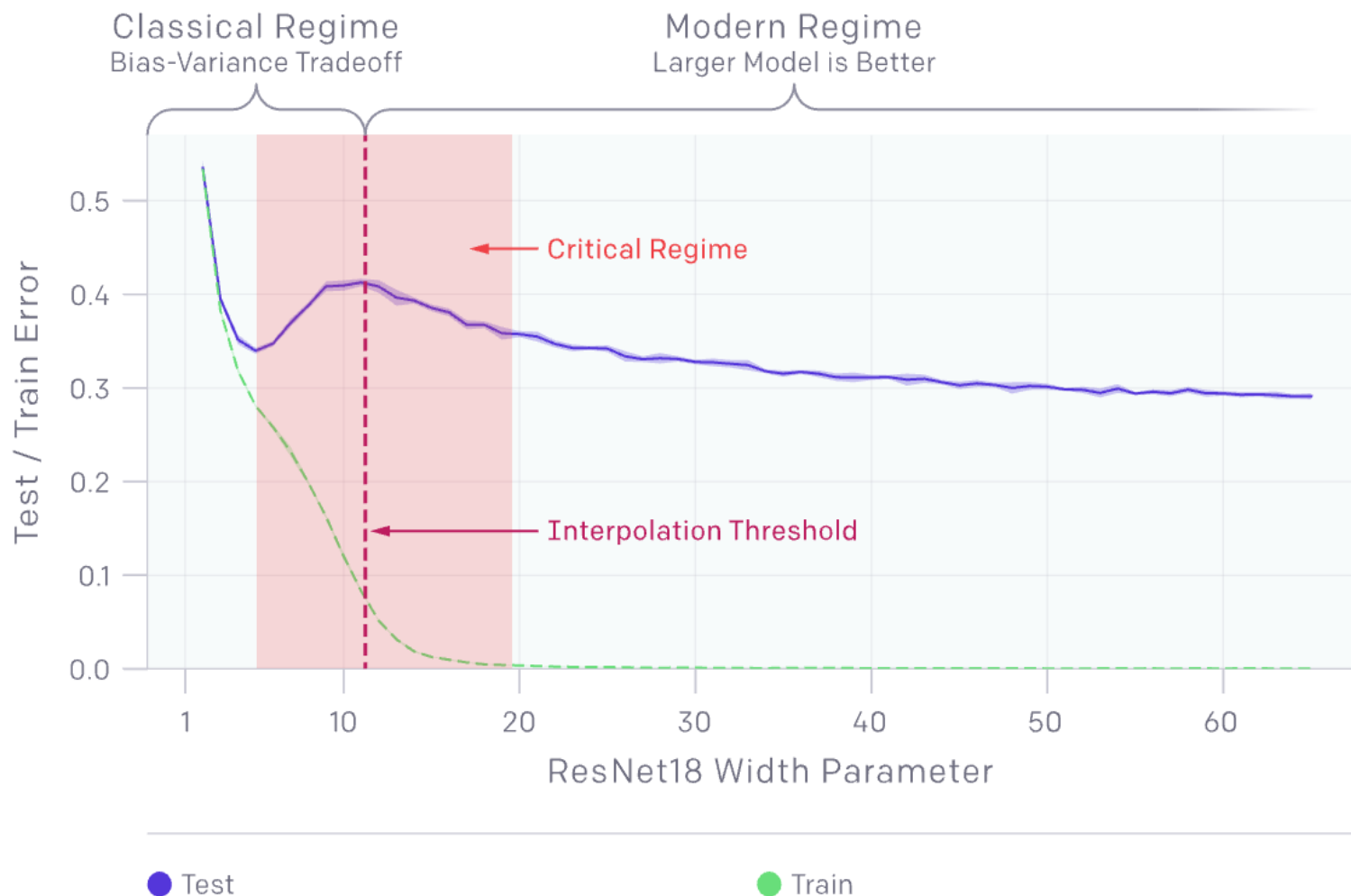


Setting	LAMBADA (acc)	LAMBADA (ppl)	StoryCloze (acc)	HellaSwag (acc)
SOTA	68.0 ^a	8.63 ^b	91.8^c	85.6^d
GPT-3 Zero-Shot	76.2	3.00	83.2	78.9
GPT-3 One-Shot	72.5	3.35	84.7	78.1
GPT-3 Few-Shot	86.4	1.92	87.7	79.3

Setting	PIQA	ARC (Easy)	ARC (Challenge)	OpenBookQA
Fine-tuned SOTA	79.4	92.0 [KKS ⁺ 20]	78.5 [KKS ⁺ 20]	87.2 [KKS ⁺ 20]
GPT-3 Zero-Shot	80.5*	68.8	51.4	57.6
GPT-3 One-Shot	80.5*	71.2	53.2	58.8
GPT-3 Few-Shot	82.8*	70.1	51.5	65.4

大模型：性质很奇妙

Deep Double Descent（哈佛大学、OpenAI, 2019）



行到水穷处： 自然语言处理下一步该怎么走？

孙茂松

清华大学计算机科学与技术系教授
清华大学人工智能研究院常务副院长

第十六届全国人机语音通讯学术会议
2021年10月16日，江苏徐州

关于大模型的思考

- 非常重要的语言信息公共基础资源
 - 工程应用上：大体ok
 - 科学研究上：量变能引起质变吗？
 - 表现很神奇，也比较费解：值得深入研究其性质
 - 我的一个猜测：巨大神经网络或实现了功能分区？
-
- 规模（数据、模型、算力）极致化能走多远？
 - 应该还是不能解决深层次语义问题
 - 包罗万象一定就好吗？
 - 需要杀手应用证明自己的能力（语音识别\输入法）

大模型的两重性：开发与探索

孙茂松

清华大学计算机科学与技术系教授

智源自然语言处理重大研究方向首席科学家

2021年11月14日，智源

大模型的探索性：不成熟思考

- 对大模型科学意义的探索也许大于其工程意义
- 对脑科学研究也许有启发性

是否出现了涌现现象？

一个典型的复杂系统是由无数个个体成分或因子构成的，它们聚集在一起会呈现出集体特性，这种集体特性通常不会体现在个体的特性中，又无法轻易地从个体的特性中预测。



ChatGPT

Step 1

Collect demonstration data and train a supervised policy.

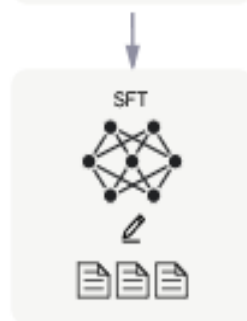
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



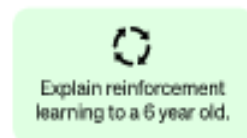
This data is used to fine-tune GPT-3.5 with supervised learning.



Step 2

Collect comparison data and train a reward model.

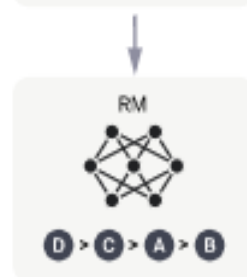
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



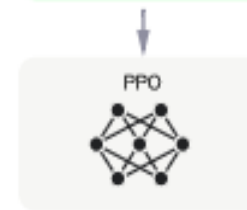
Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.



The PPO model is initialized from the supervised policy.



The policy generates an output.



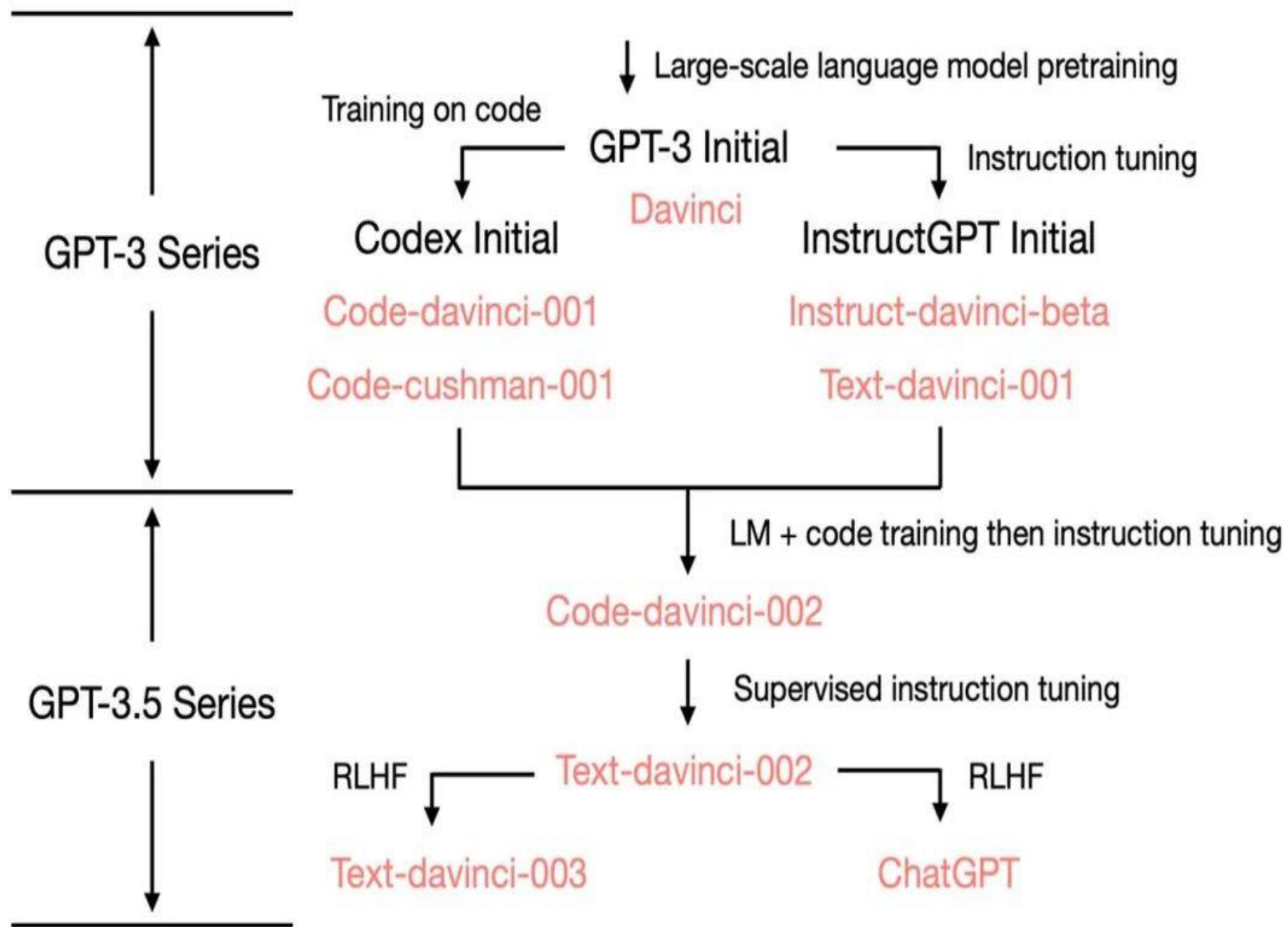
The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



Reinforcement Learning from Human Feedback, RLHF



From “Zero-Shot” To “Chain Of Thought” : Prompt Engineering

I need short, bullet-point answers to my questions.

Question: How does the sun work?

Answer:

- fusing of hydrogen nuclei to form helium
- releasing an enormous amount of energy in the form of light and heat
- estimated that the sun has enough "fuel" for around 5 billion more years

Question: How does a bird work?

Answer:

- generate lift using wings
- use feathers to provide insulation and streamlining for flight
- use a muscular system for flight control
- use a digestive and respiratory system to provide energy and intake oxygen for powering flight muscles

<https://medium.com/mlearning-ai/from-zero-shot-to-chain-of-thought-prompt-engineering-choosing-the-right-prompt-types-88800f242137>

Take the last letters in the words "Zeitgeist" and "AI" and concatenate them.

ei

Input: Take the last letters in the words "Zeitgeist" and "AI" and concatenate them.

Answer: The last letter in the word "Zeitgeist" is "t". The last letter in the word "AI" is "I". So if concatenated they become: "tI". The answer must be "tI".

Take the last two letters of "One" and "Two" and concatenate them:

The last two letters of "One" are "ne". The last two letters of "Two" are "wo". Concatenating them gives "ne wo". The answer must be "ne wo".

Do this again but drop any spacing:

The last two letters of "One" are "ne". The last two letters of "Two" are "wo". Concatenating them gives "newo". The answer must be "newo".

思维链好于直接用提示词的模型参数： 63B

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei

Xuezhi Wang

Dale Schuurmans

Maarten Bosma

Brian Ichter

Fei Xia

Ed H. Chi

Quoc V. Le

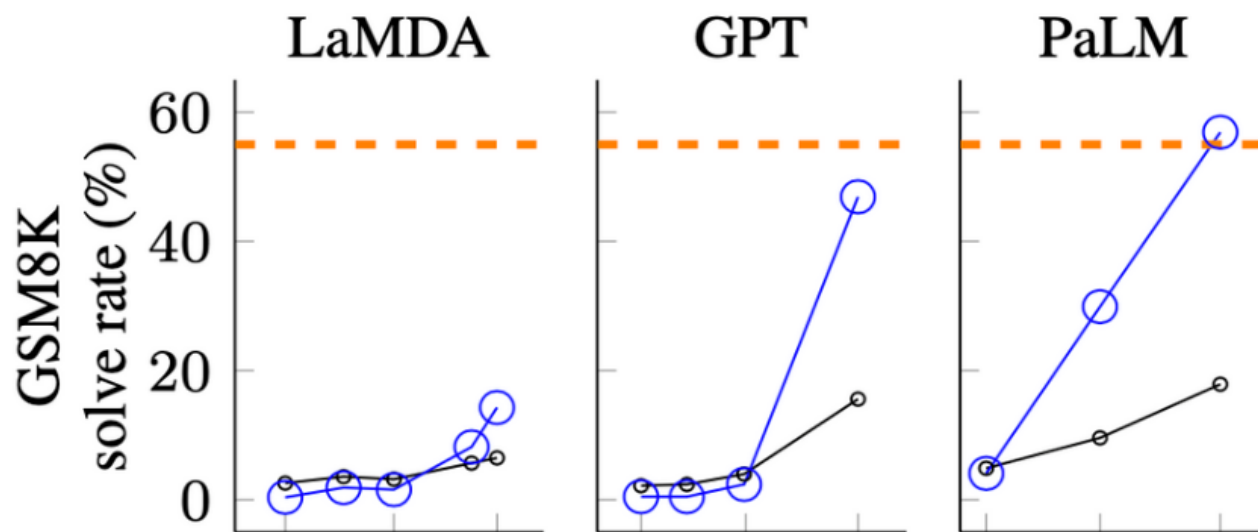
Denny Zhou

Google Research, Brain Team
{jasonwei, dennyzhou}@google.com

—○— Standard prompting

—○— Chain-of-thought prompting




- - - Prior supervised best



下一步对策

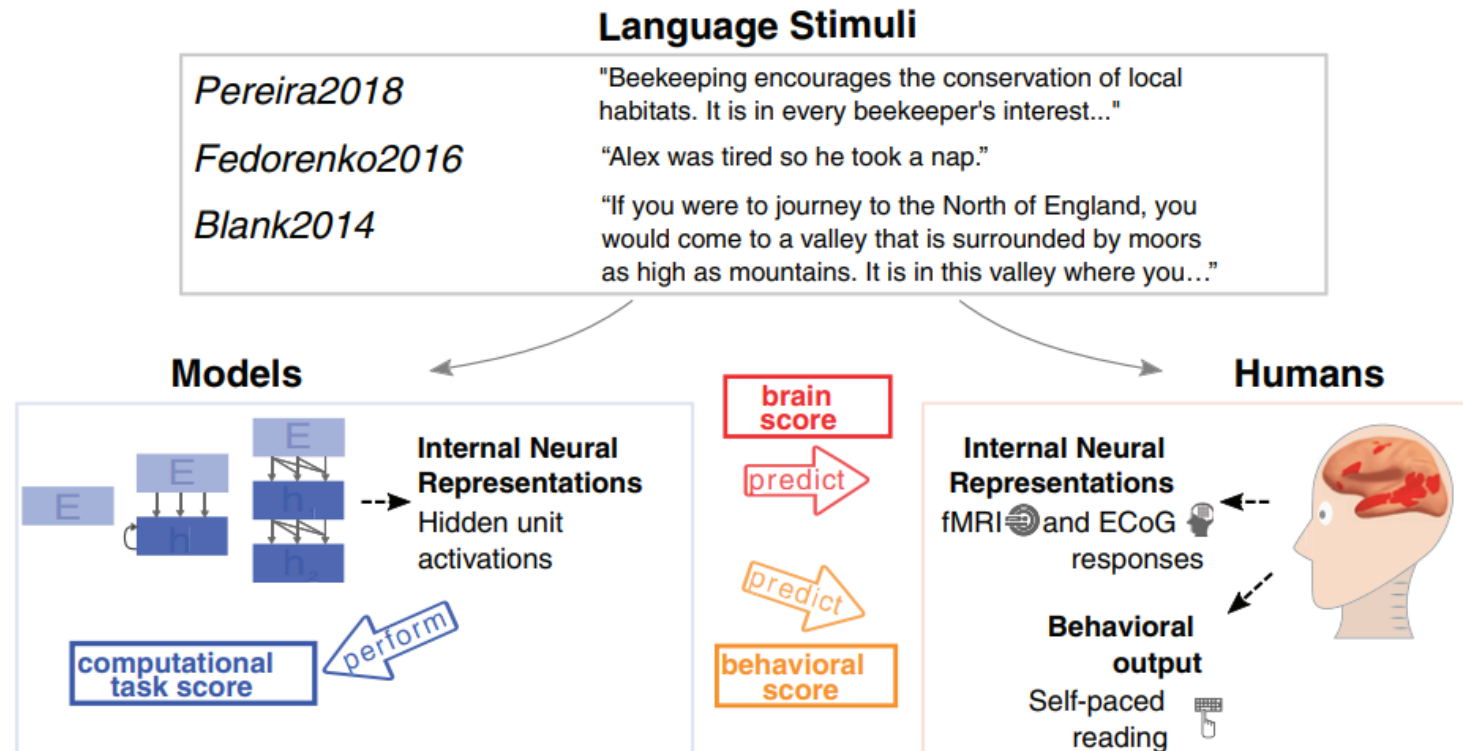
- 尽快做出性能具可比性的中国版ChatGPT
不是为了比而比，而是补差距
- 建立以中文为核心的大模型
- 加强对大模型机理的科学探索研究
对脑科学研究也许有直接启发意义
- 加强中文测试平台建设

The neural architecture of language: Integrative modeling converges on predictive processing

Martin Schrimpf^{a,b,c,1} , Idan Asher Blank^{a,d,2}, Greta Tuckute^{a,b,2}, Carina Kauf^{a,b,2}, Eghbal A. Hosseini^{a,b}, Nancy Kanwisher^{a,b,c,1} , Joshua B. Tenenbaum^{a,c,3}, and Evelina Fedorenko^{a,b,1,3} 

^aDepartment of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139; ^bMcGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139; ^cCenter for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139; and ^dDepartment of Psychology, University of California, Los Angeles, CA 90095

Contributed by Nancy Kanwisher, September 3, 2021 (sent for review April 27, 2021; reviewed by Matthew M. Botvinick and Adele E. Goldberg)





Blog post

Paper

GitHub

A language model takes in text and produces text:

A helm is a → Language Model → *wheel for steering a ship...*

Despite their simplicity, language models are increasingly functioning as the foundation for almost all language technologies from question answering to summarization. But their immense capabilities and risks are not well understood. Holistic Evaluation of Language Models (HELM) is a living benchmark that aims to improve the transparency of language models.

1. **Broad coverage and recognition of incompleteness.** We define a taxonomy over the scenarios we would ideally like to evaluate, select scenarios and metrics to cover the space and make explicit what is missing.



[Browse State-of-the-Art](#)[Datasets](#)[Methods](#)[More ▾](#)[Sign In](#)[Multi-task Language Understanding](#)

Multi-task Language Understanding on MMLU

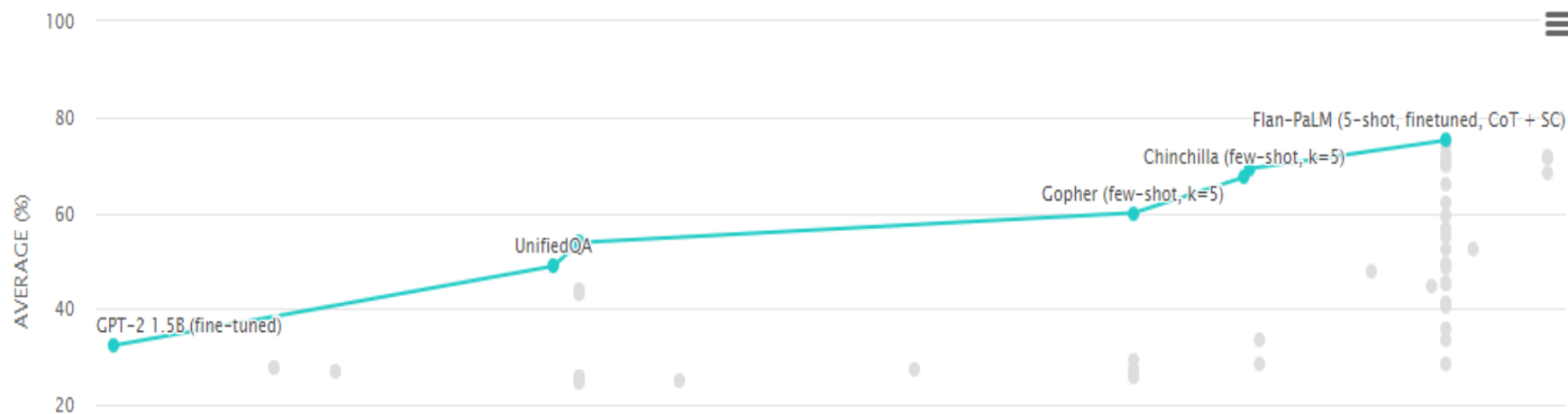
[Leaderboard](#)[Dataset](#)

View

Average (%) ▾

by

Date ▾



[Browse State-of-the-Art](#)[Datasets](#)[Methods](#)[More ▾](#)[Sign In](#)[Aa Texts](#)

BIG-bench (Beyond the Imitation Game Benchmark)

[✎ Edit](#)

Introduced by Srivastava et al. in [Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models](#)

The **Beyond the Imitation Game Benchmark** (BIG-bench) is a collaborative benchmark intended to probe large language models and extrapolate their future capabilities. Big-bench include more than 200 tasks.

Image shows a word cloud of task keywords.

[Homepage](#)

Source: <https://arxiv.org/pdf/2206.04615.pdf>

Usage 📊

Benchmarks

[✎ Edit](#)

sets
40

CUGE

中文语言理解和生成评测基准

CUGE包含高质量中文自然语言处理数据集、排行榜与在线评测平台，旨在构建全面系统的中文机器语言能力评测体系，形成多层次维度的评测方案，力求更加科学、规范、高质量地推进中文自然语言处理技术的标准评测。

委员单位 | Committee Organizations



清华大学
Tsinghua University



北京大学
Peking University



中国科学院
Chinese Academy of Sciences



中国人民大学
Renmin University of China



北京语言大学
Beijing Language and Culture University



山西大学
Shanxi University



大连理工大学
Dalian University of Technology



哈尔滨工业大学
Harbin Institute of Technology



复旦大学
Fudan University



苏州大学
Soochow University



京东科技
JD Tech

支持单位 | Supporting Organization



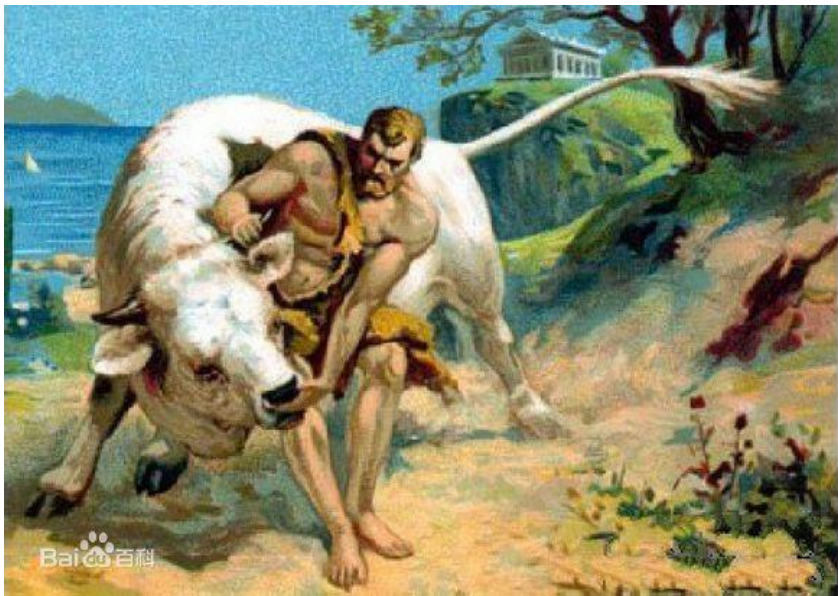
清华大学人工智能研究院
Institute of Artificial Intelligence, Tsinghua University



中国中文信息学会计算语言学专委会
Technical Committee on Computational Linguistics, Chinese Information Processing Society of China



北京智源人工智能研究院
Beijing Academy of Artificial Intelligence, BAAI



古希腊神话：赫拉克勒斯

《竞渡诗》（唐代卢肇）

石溪久住思端午，馆驿楼前看发机。
鼙鼓动时雷隐隐，兽头凌处雪微微。
冲波突出人齐讫，跃浪争先鸟退飞。
向道是龙刚不信，果然夺得锦标归。

冲波突出大模型，跃浪争先GPT

谢谢！