大模型时代的危与机

张奇

复旦大学

ChatGPT仿佛电影中的AI就在眼前

ChatGPT给公众展现出来的是无所不能

GM

选择题，请运用你的中国哲学知识选择正确的答案，并说出你推理的过程。汉晋时期有多种文本记载，帝尧之时，“天下太和，百姓无事”。有老者“击壤”而戏，围观者称颂帝尧。老者歌云：“吾日出而作，日入而息，凿井而饮，耕地而食，帝何德于我哉！”上述记载所体现的政治理念最接近（ ）

A．孔子　B．老子　C．韩非　D．墨子

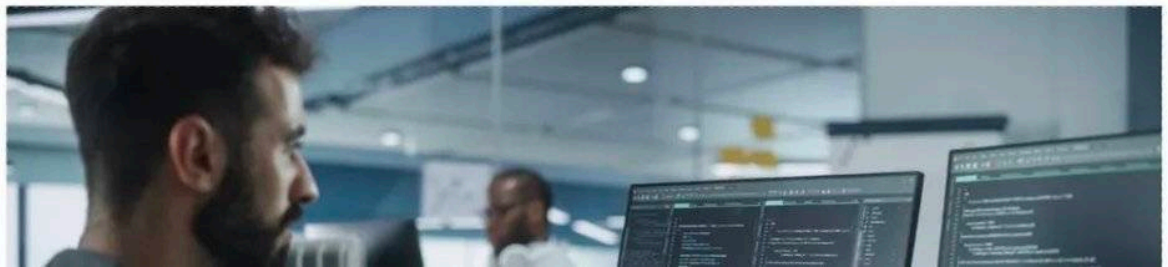
根据题干中给出的信息，上述记载所体现的
太平时人们不必干预政治，只需专注于自己
老子。

Home > News > Artificial Intelligence

ChatGPT Passes Google Coding Interview for Level 3 Engineer With $183K Salary

'Amazingly, ChatGPT gets hired at L3 when interviewed for a coding position,' reads a Google document, but ChatGPT itself says it can't replicate human creativity and problem-solving skills.

By Emily Dreibelbis　February 1, 2023



大模型认知

语言模型预训练完成后，仍然需要大量**标注数据**

How to Label 1M Data Points/Week

by Akshat Bubna on November 12th, 2019

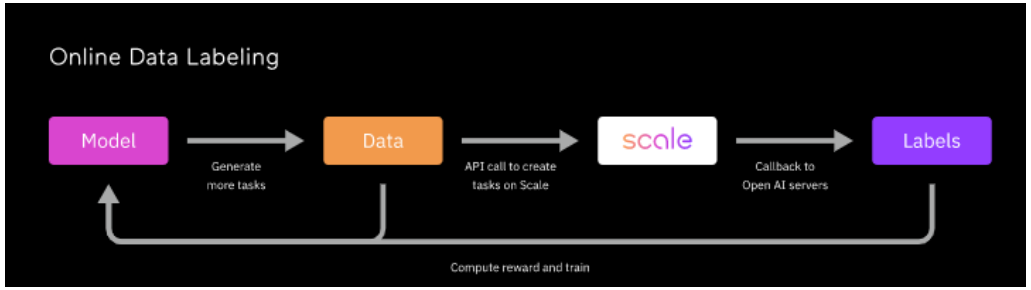
TEXT

After decades of research, scientist finally have an answer to the age old mystery: what does the fox say?

QUESTION

Which is the most descriptive continuation of the above paragraph?

- A) It turns out, it meows.
- B) It turns out that every full moon, it raises its skinny red snout to the heavens and lets out deep, soulful meows.



very low latency (<30 mins)
high throughput (~5,000 labels/hr)

https://scale.com/blog/how-to-label-1m-data-points-week

大模型认知

语言模型预训练完成后，仍然需要大量**标注数据**

OpenAI LLC has reportedly hired about **1,000 contractors** over the past six months.

About 60% of the 1,000 hires reportedly onboarded in the past six months work on data labeling projects.

The remaining 40% of the contractors are programmers.

> A software developer in South America who completed a five-hour unpaid coding test for OpenAI told Semafor he was asked to tackle a series of two-part assignments. **First**, he was given a coding problem and asked to explain in written English how he would approach it. **Then**, the developer was asked to provide a solution. **If he found a bug**, OpenAI told him to detail what the problem was and how it should be corrected, instead of simply fixing it.

https://siliconangle.com/2023/01/27/report-openai-hired-1000-contractors-improve-ai-models-coding-capabilities/

大模型认知

语言模型预训练完成后，仍然需要大量**标注数据**

GPT-NeoXT-Chat-Base-20B-v0.16

*TLDR: As part of OpenChatKit (codebase available here), GPT-NeoXT-Chat-Base-20B-v0.16 is a 20B parameter language model, fine-tuned from EleutherAI's GPT-NeoX with over **40 million instructions** on 100% carbon negative compute.*

**Summarize a long document into a single sentence and conduct question ar

<human>: Last year, the travel industry saw a big rebound in demand — and

To be honest up front, I am great at strategically losing things — or rath

I've been a big fan of Apple's $29 item tracker, dubbed AirTag. The long-r

The Federal Aviation Administration has even ruled that having an AirTag i

"Luggage tracking devices powered by lithium metal cells that have 0.3 gra

But, you may be wondering what makes them so special and why you should ca

Summarize the above into a single paragraph.

<bot>: If you're traveling this spring break, you might want to consider i

<human>: But isn't it unsafe to put AirTag in a luggage?

<bot>: The Federal Aviation Administration has ruled that having an AirTag

<human>: Ah, why is that?

<bot>: The devices only contain 0.1 grams of lithium, which falls below th

**Extract structured information from unstructured documents, as illustrat

<human>: Generate a table summarizing the options outlined in this email.

Team, we have to make a decision on the product plan for Goocrux v2.

The first option will take us 3 weeks of development, and will achieve per

The second option will take us 4 weeks of development, and will achieve pe

The third option will take the longest -- 6 weeks of development, will ach

A markdown table with four columns: (1) Option ID, (2) Development time, (

<bot>: #.

ID	Option ID	Development time	Performance	UI Change
1	1	3 weeks	80ms	Yes
2	2	4 weeks	60ms	No
3	3	6 weeks	30ms	No

Classifying a sentence or paragraph into different categories.

<human>: Classify the sentiment of the following sentence into Positive, N
<bot>: Positive

<human>: Classify the sentiment of the following sentence into Positive, N
<bot>: Negative

Itamar Golan 🤓
@ItakGol

Introducing OpenChatKit -
The first open-source alternative to ChatGPT!

A team of ex-OpenAI fellows at Together have released a 20B chat-GPT model, fine-tuned for chat using EleutherAI's GPT-NeoX-20B, with over 43 million instructions under the Apache-2.0 license.

https://huggingface.co/togethercomputer/GPT-NeoXT-Chat-Base-20B

能力是**涌现**出来的吗？

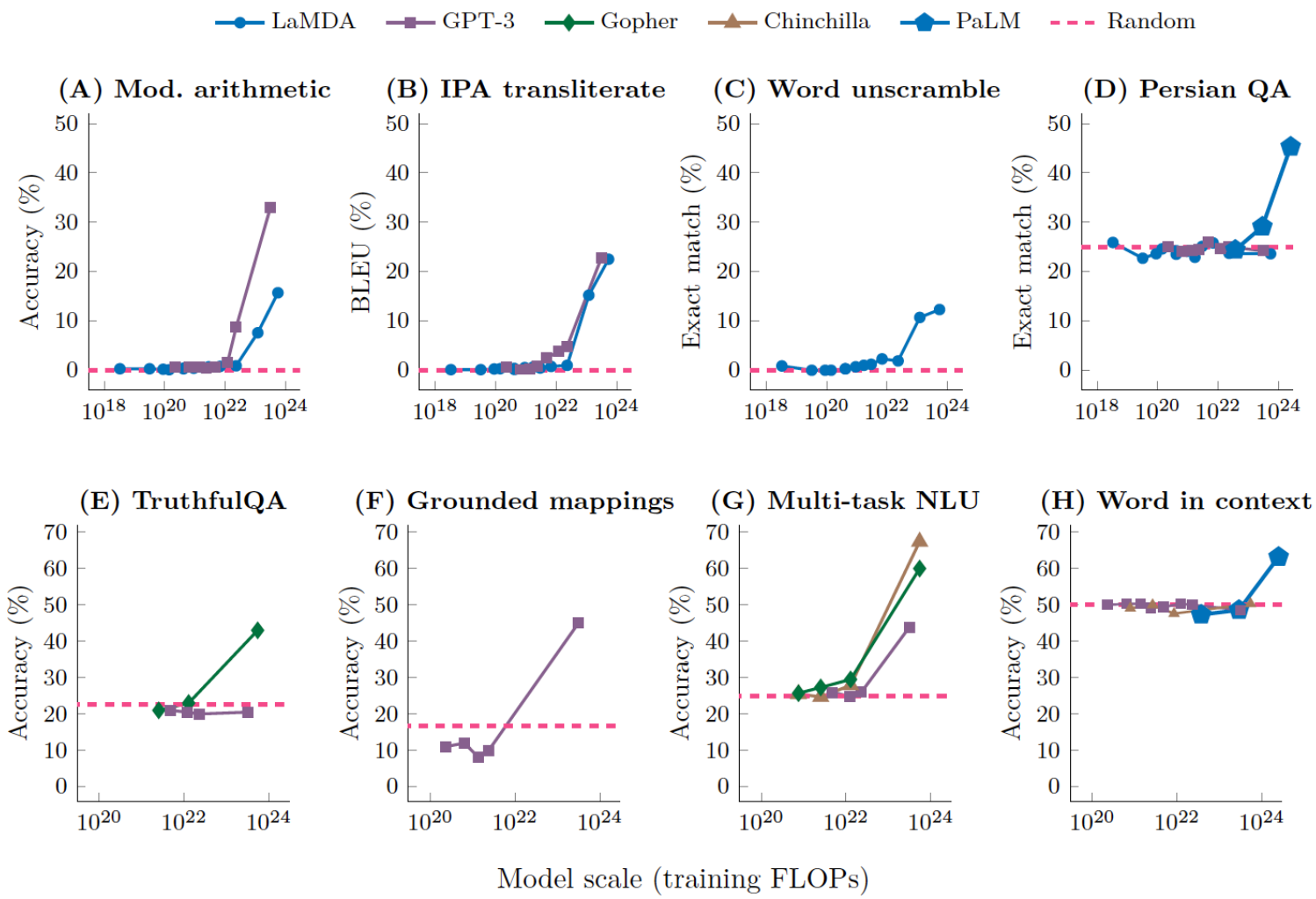


Figure 2: Eight examples of emergence in the few-shot prompting setting. Each point is a separate model. The ability to perform a task via few-shot prompting is emergent when a language model achieves random performance until a certain scale, after which performance significantly increases to well-above random. Note

大模型认知

所谓的能力**涌现**也许仅是评价指标的问题

LLM 涌现能力的**两个决定性属性**：

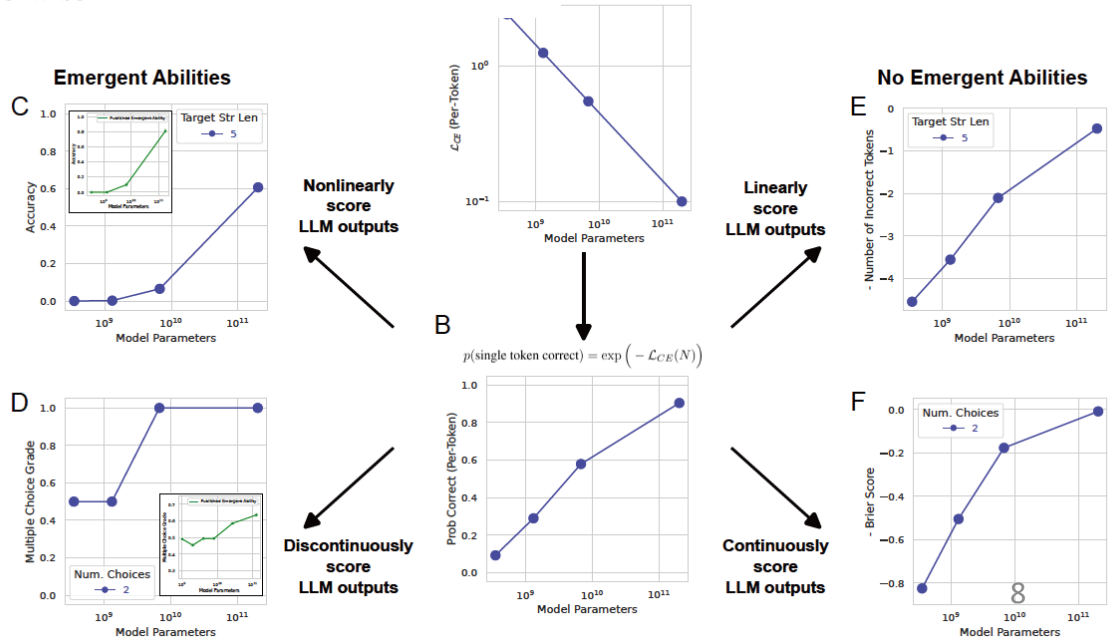
1. 敏锐性，从「不存在」到「存在」似乎只是瞬间的过渡；**2. 不可预测性**，在看似不可预见的模型规模内过渡。

在 BIG-Bench 任务中，>92% 的涌现能力是这两个度量下出现的：

$$\text{Multiple Choice Grade} \overset{\text{def}}{=} \begin{cases} 1 & \text{if highest probability mass on correct option} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Exact String Match} \overset{\text{def}}{=} \begin{cases} 1 & \text{if output string exactly matches target string} \\ 0 & \text{otherwise} \end{cases}$$

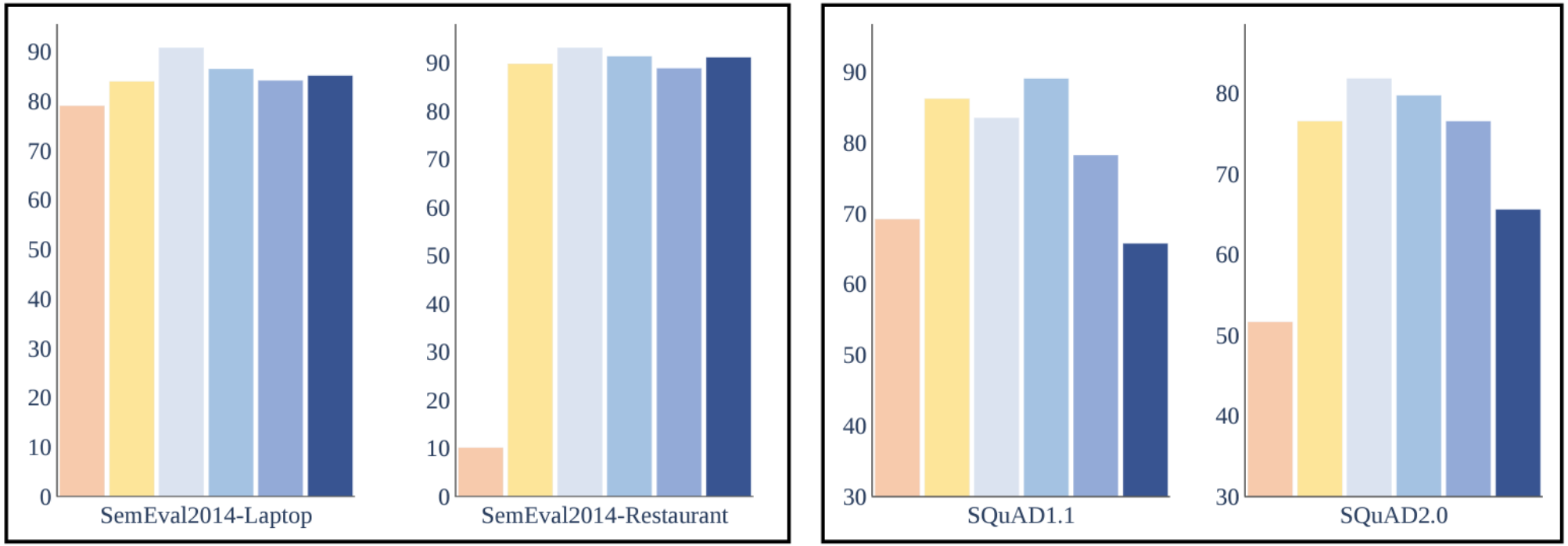
随着模型规模提升，如果将度量**从非线性 / 不连续的度量**（图 2CD）换**成线性 / 连续的度量**（图 2EF），具有有平滑的、连续的、可预测的性能提升。



Are Emergent Abilities of Large Language Models a Mirage? Arxiv, 2023

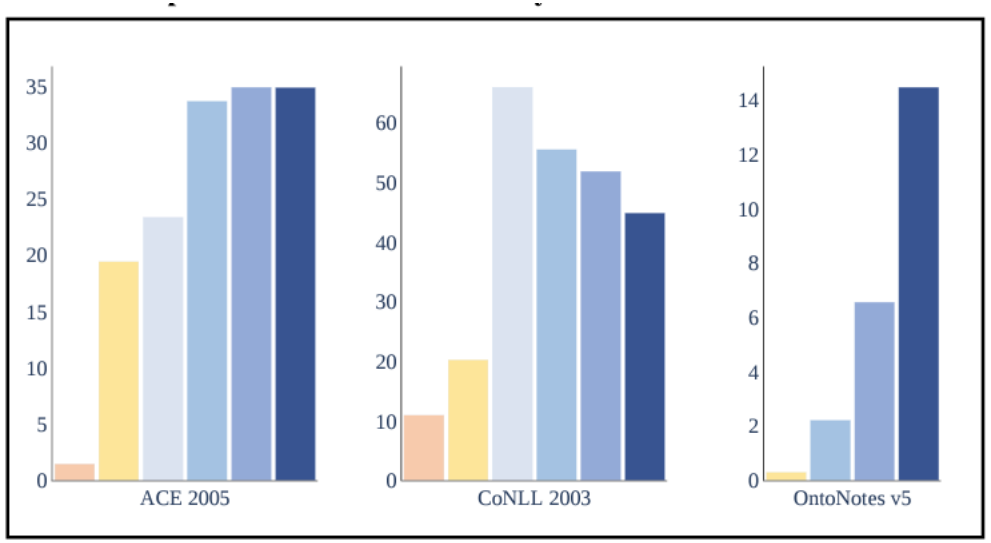
大模型认知

不能迷信和依赖任务能力**涌现**

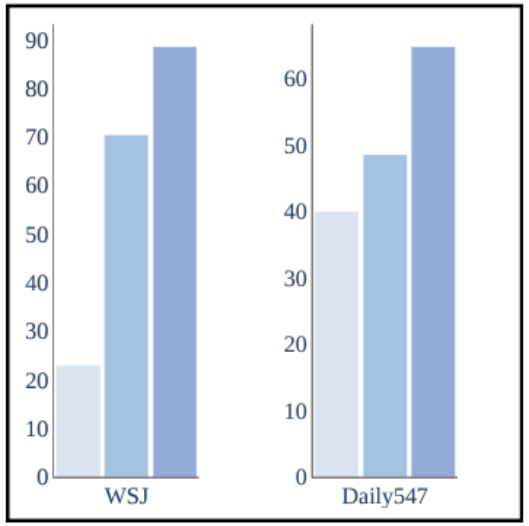


大模型认知

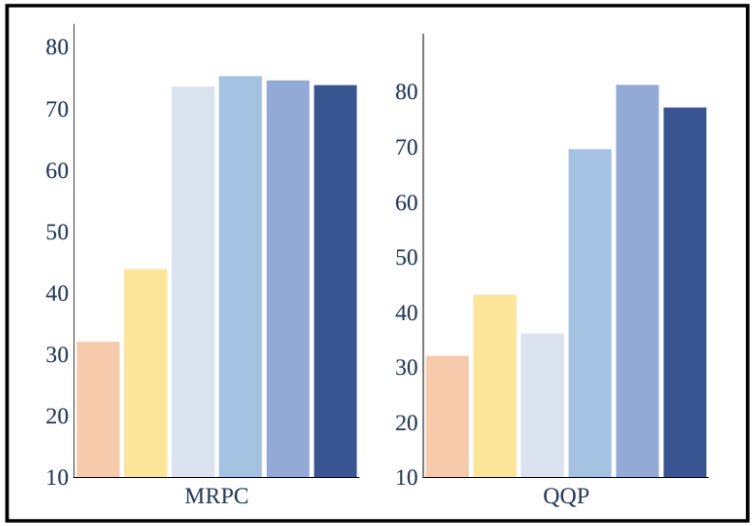
不能迷信和依赖任务能力**涌现**



Named Entity Recognition (ENG)



Part-of-speech Tagging (ENG)



Semantic Matching

davinci text-davinci-001 code-davinci-002 text-davinci-002 text-davinci-003 gpt-3.5-turbo

大模型认知

大模型是否真的具备**推理能力**仍然有待讨论

ChatGPT和GPT-4都擅长解决**著名的逻辑推理阅读理解基准**，但在处理**分布外数据集**方面很困难。

它们在需要逻辑推理的自然语言推理任务上的性能仍有待提高。

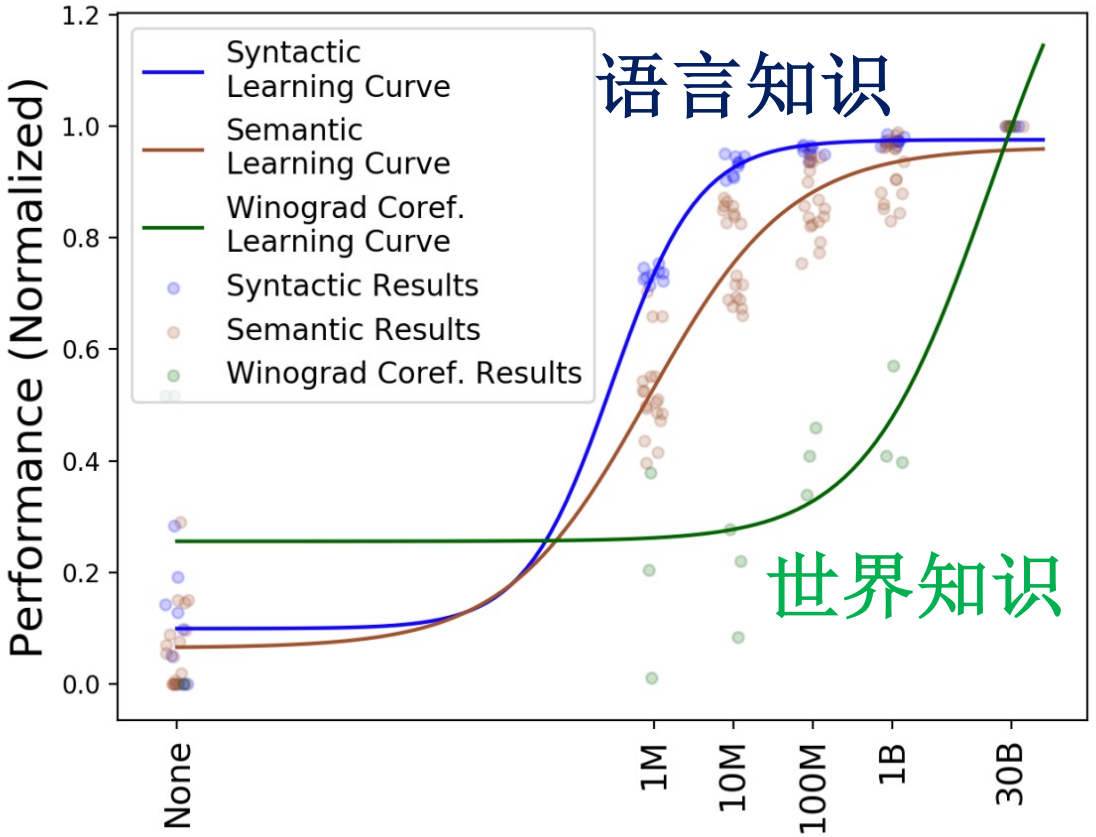
Dataset	LogiQA 2.0 test	LogiQA 2.0 zh test	ReClor dev	AR-LSAT test	LogiQA 2.0 ood
Size	1572	1594	500	230	1354
Human avg.	86.00	88.00	63.00	56.00	83.00
human ceiling	95.00	96.00	100.00	91.00	99.00
RoBERTa	48.76	35.64	55.01	23.14	33.22
ChatGPT	52.37	53.18	57.38	20.42	38.44
GPT-4	72.25	70.56	87.20	33.48	58.49

2022年新创建数据集合　　人工创建的分布外数据

Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4, Arxiv 2023

大模型认知

不是所有场景下都必须要使用**千亿**大模型



语言学任务：词性、句法、组块

语义任务：语义匹配、关系抽取等

世界知识任务：指代消解

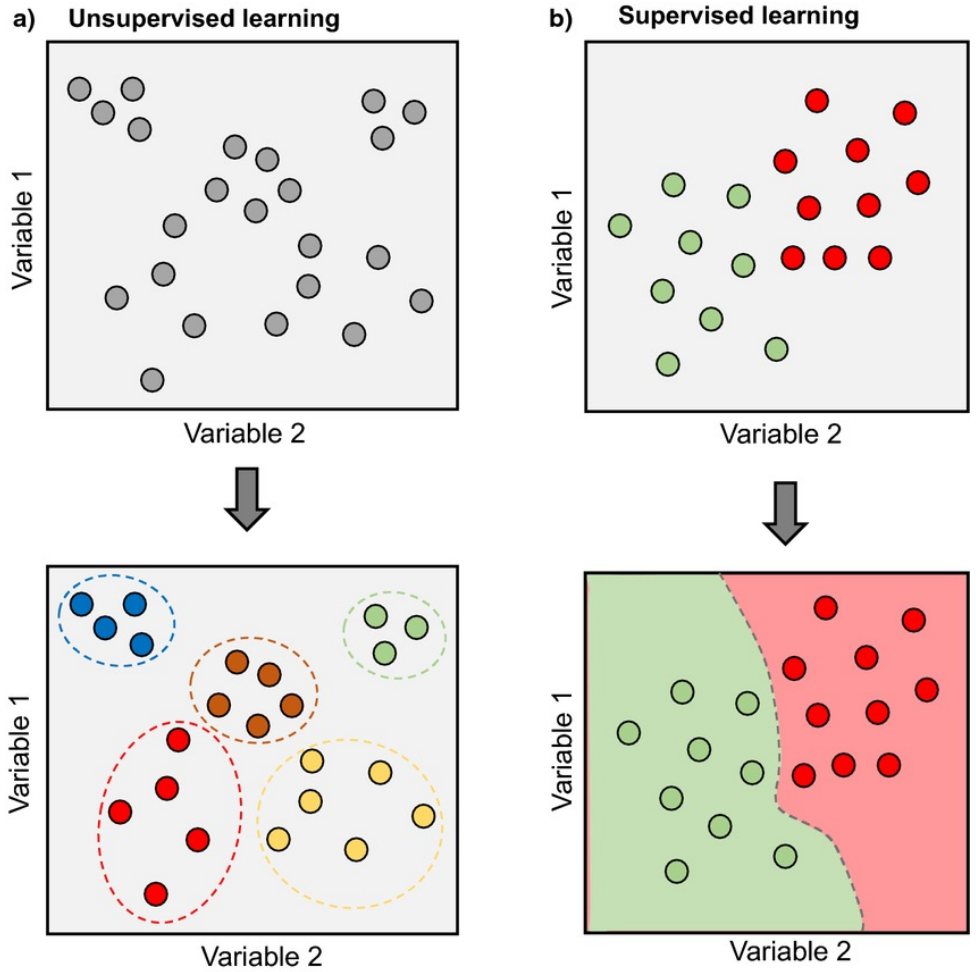
猴子吃了那个香蕉，因为它饿了

猴子吃了那个香蕉，因为它熟透了

https://arxiv.org/abs/2011.04946

同一个时代没有什么黑科技

当前时代机器学习基础理论是什么？



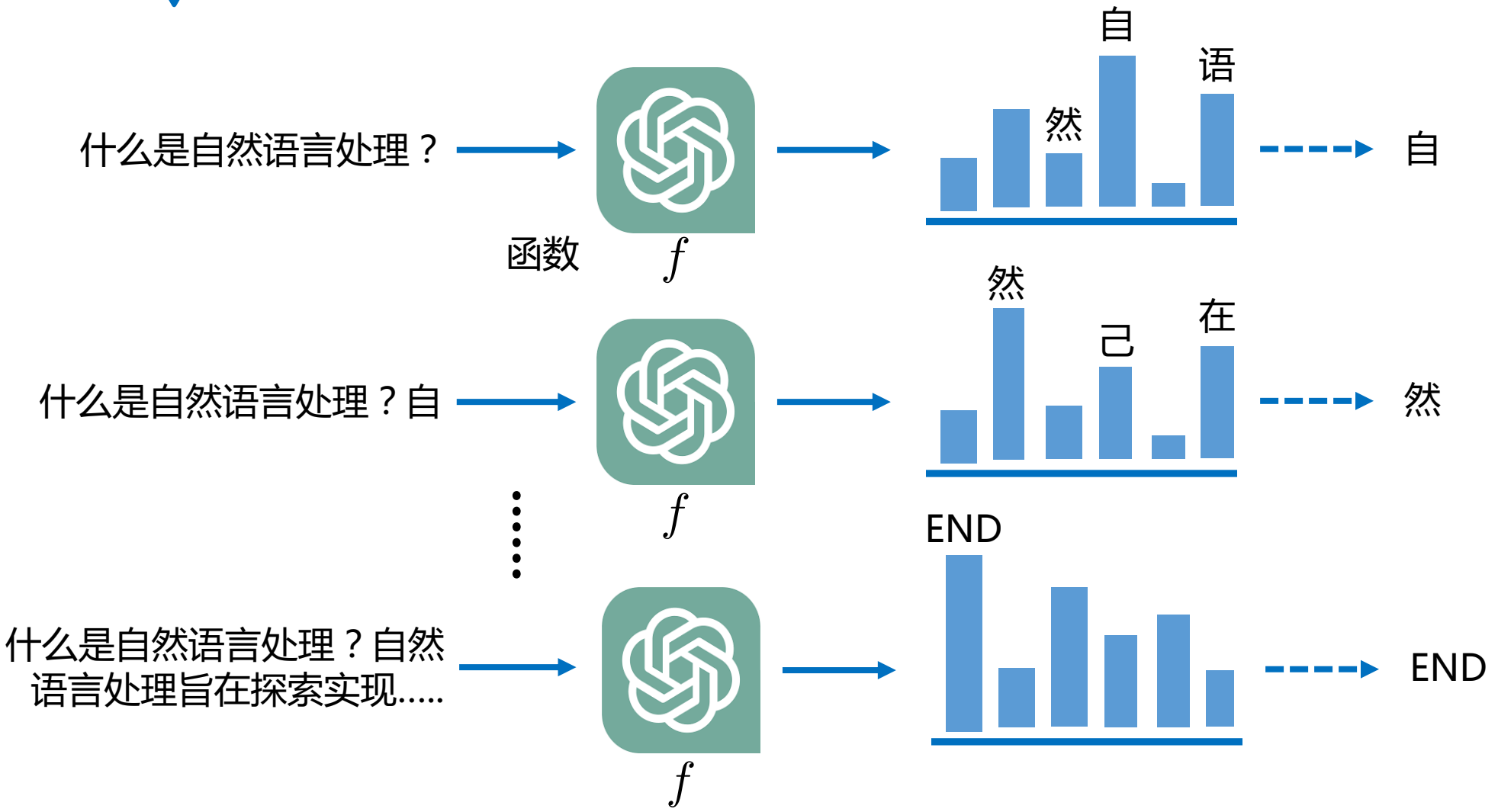
模型：学习什么样的模型

策略：按照什么样的准则学习或选择最优的模型

算法：学习模型的具体计算方法

数据：高质量、大规模训练数据

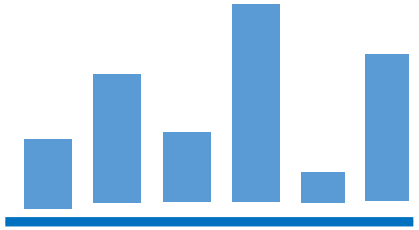
ChatGPT真正做的事情—文字接龙



概率如何得到的？

测试

用户输入

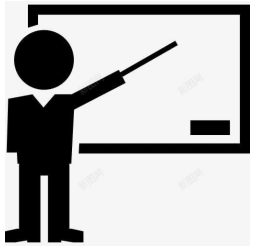


训练

预训练语言模型

大量高质量
互联网无监督数据

f



大量对话文本
以及真实回答

人类反馈

数千种针对特
定任务的有监
督训练数据

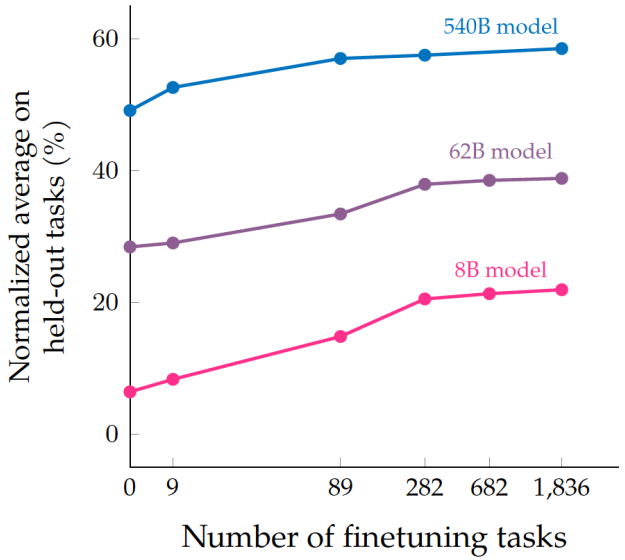
指令微调

大模型真正展现了什么能力？

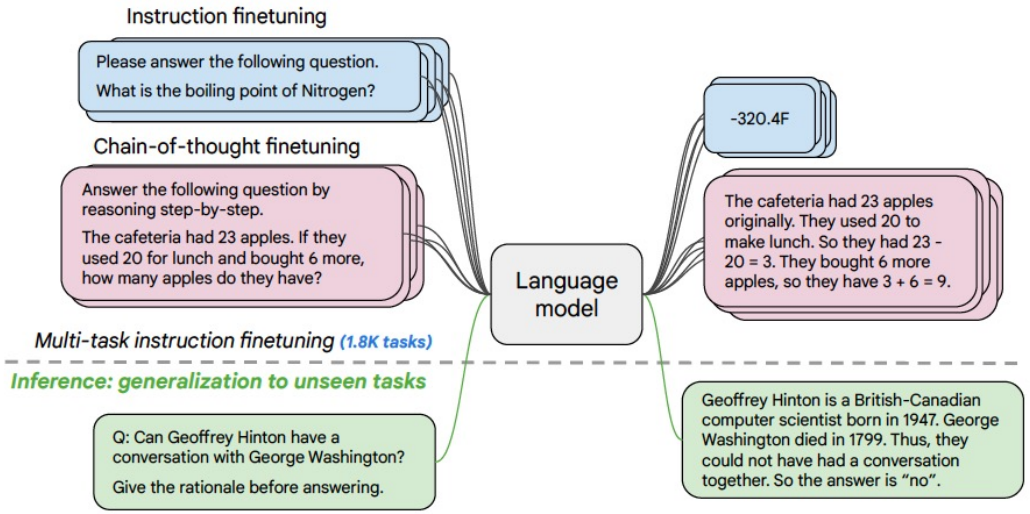
超强的文本建模能力：25K英文单词的理解和生成能力

多任务统一学习范式：将所有NLP任务统一到**生成式多任务学习框架**

任务泛化能力：数千种任务统一框架学习后，部分无标注数据任务上效果很好



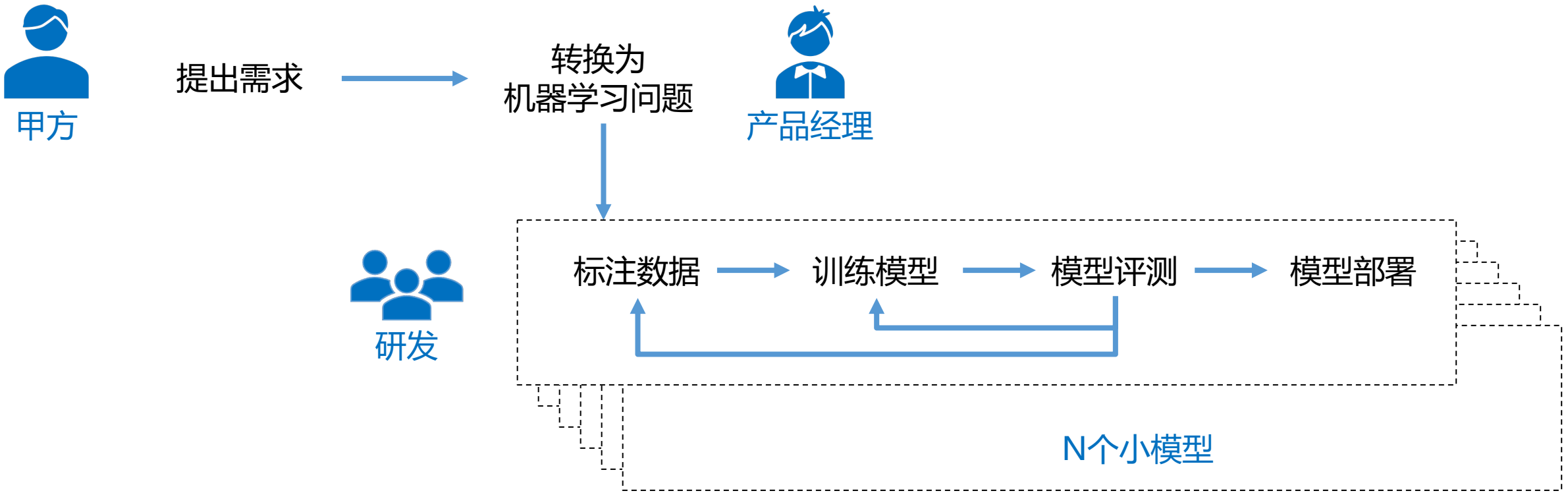
Chung et al. *Scaling Instruction-Finetuned Language Models*, arxiv



Chung et al. *Scaling Instruction-Finetuned Language Models, Arxiv Oct. 2022*

机遇在哪里？

小模型时代的工作流程



小模型时代的难点

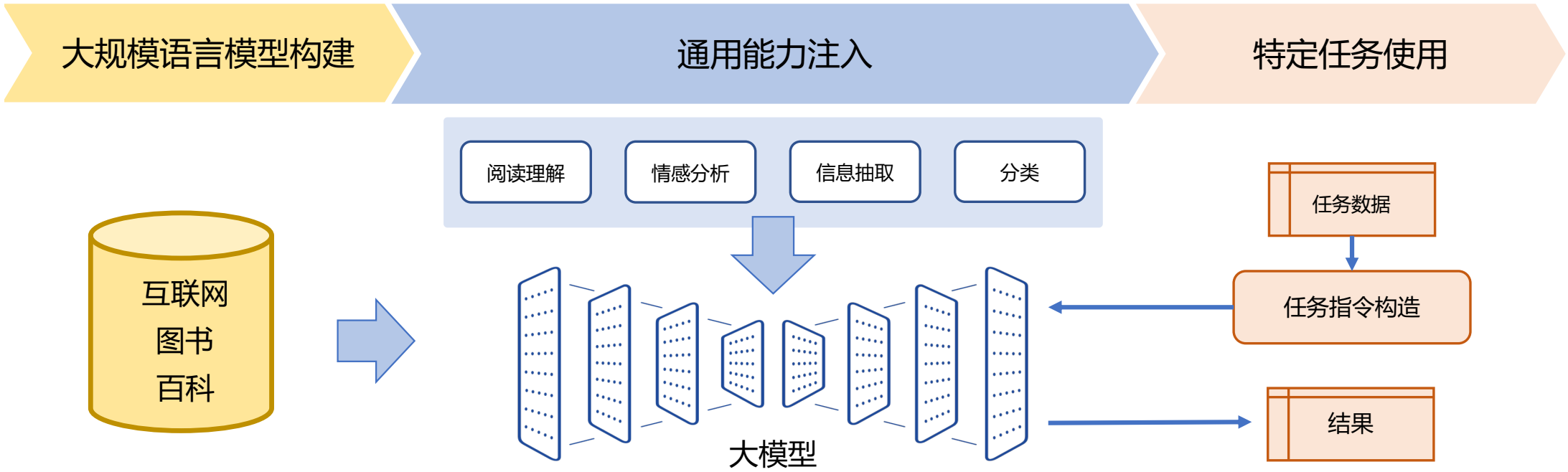
任务开发成本高、时间周期长

相同任务的微小需求变化，需要30%-70%的重新开发成本，**无法产品化**

模型开发和维护成本高

高水平算法研究员稀缺

将来会演变成什么样子？


大规模语言模型构建
通用能力注入
特定任务使用
互联网
图书
百科
阅读理解
情感分析
信息抽取
分类
大模型
任务数据
任务指令构造
结果

任务范式发生巨大变化

使用自然语言对模型进行训练和使用

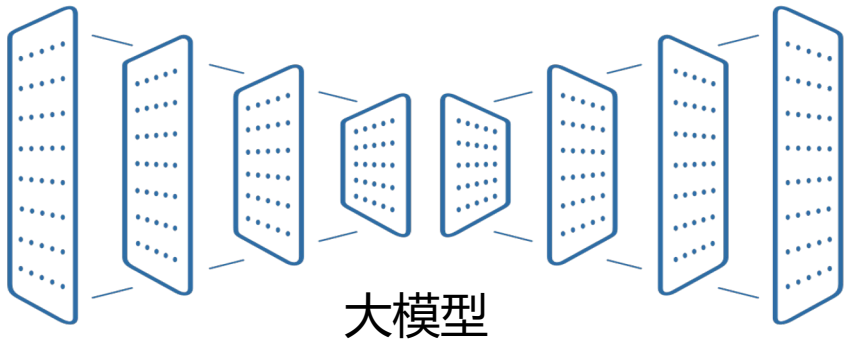
【句子输入】，上述文字表达了褒义和贬义倾向？

【文章输入】，上述文章中包含哪些公司主体？

【文章输入】，从上述文章提取2022年公司主营业务收入

【财报输入】，从上述财报，输出核心提示

【数据】，根据上述数据，书写2023年第一季度财报摘要



大模型优势

非常少量算法研究员

新任务可以**快速训练**，并且**不需要模型重新部署**

低成本**产品化**

可以应用于哪些场景—银行

场景一: 信贷辅助决策 业务部门: 风险/公司/信等	财务报告自动化解析	NLP-抽取/OCR/表格解折
	公司舆情数据智能挖据	NLP-分类/情感/实体识别/事件标签
	信贷流水材料智能处理	NLP-表格解折/OCR
	宏观行业景气度分析	NLP-抽取/事件标签/专家知识
	风险事件传导分析	NLP-事件/专家知识
场景二: 公司债业务审核业务部门: 投行	募集债券说明书智能审核	NLP-抽取/审核/比对
	审计报告自动化比对	NLP-抽取/比对
场景三，智能客服业务部门: 远程银行/信用卡等	智能客户服务	NLP-多轮对话
	工单数据处理	NLP-分类/聚类/0CR/估感
	多元营销辅助	NLP-分类/聚类/实体/情感
场景四:企业知识库建设业务部门，科技部门	企业内部数据	NLP·搜索/问答/文本解析

银行应用方式

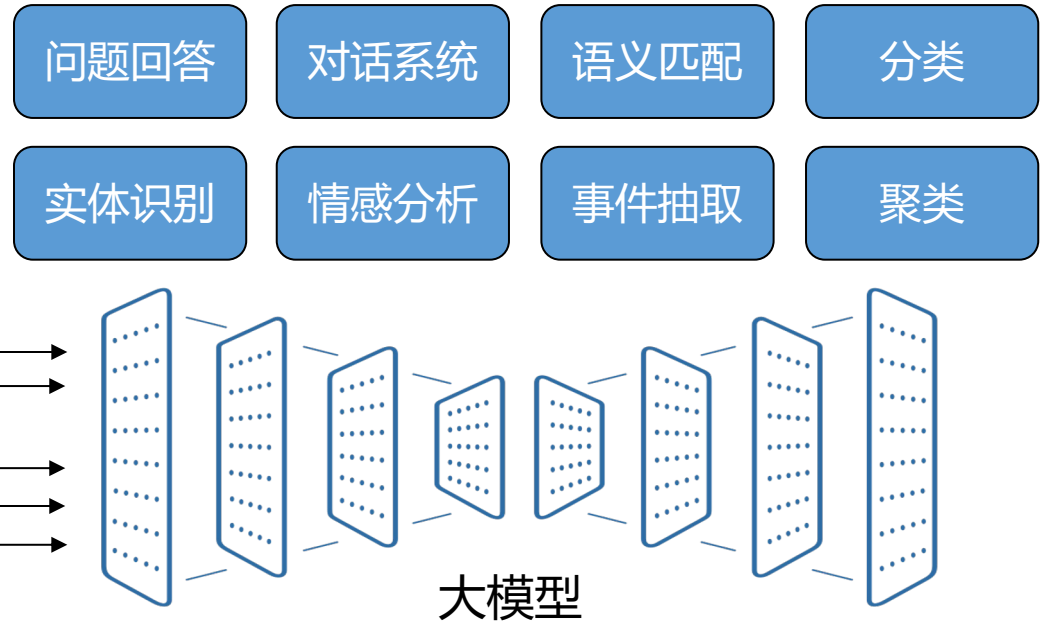
【句子输入】，上述文字表达了褒义和贬义倾向？

【文章输入】，上述文章中包含哪些公司主体？

【文章输入】，从上述文章提取2022年公司主营业务收入

【财报输入】，从上述财报，输出核心提示

【数据】，根据上述数据，书写2023年第一季度财报摘要



增加一个新任务时怎么办？

甲方

提出需求

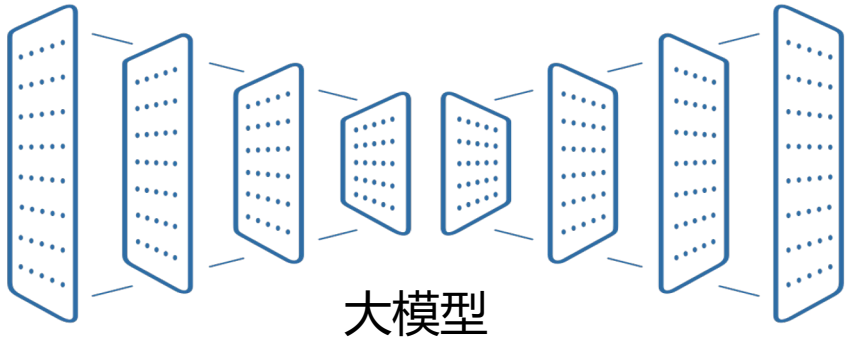
转换为
机器学习问题

产品经理

产生训练数据

研发

【句子输入】，上述文本是关于信用卡还是贷款业务？



大模型

大模型时代的自然语言处理研究变化

大模型时代从**单个小任务**的研究，转向**一类问题统一框架**研究

挑战：赢者通吃，不紧跟步伐就会造成代差，并且不容易弥补

对话交互方式，使得先发优势更加明显，用户数据无法获取的情况下，追赶代价更大

机遇：NLP从**手工作坊**转向**蒸汽机时代**，人工智能的Iphone时代来临

面向B端私有化部署的"**小**"模型

大模型时代的NLP研究重点

领域大模型----对话、摘要、改写、 NLP任务集成等　1B-50B模型族

大模型可控文本生成

大模型隐私

大模型鲁棒性评测与提升

超小规模任务模型构建　1M-50M　--> 是否可以脱离深度神经网络？

大模型推理能力构建

大模型可解释性

因果机制与大模型融合

中文开源大模型族　1-20B

怎么搞？

不同的业务类型差别很大

模型需求稳定，模型数量少，目前效果较好 ✗

模型需求稳定，但效果亟待提升 ✓ 大模型的长距离建模能力，解决歧义问题

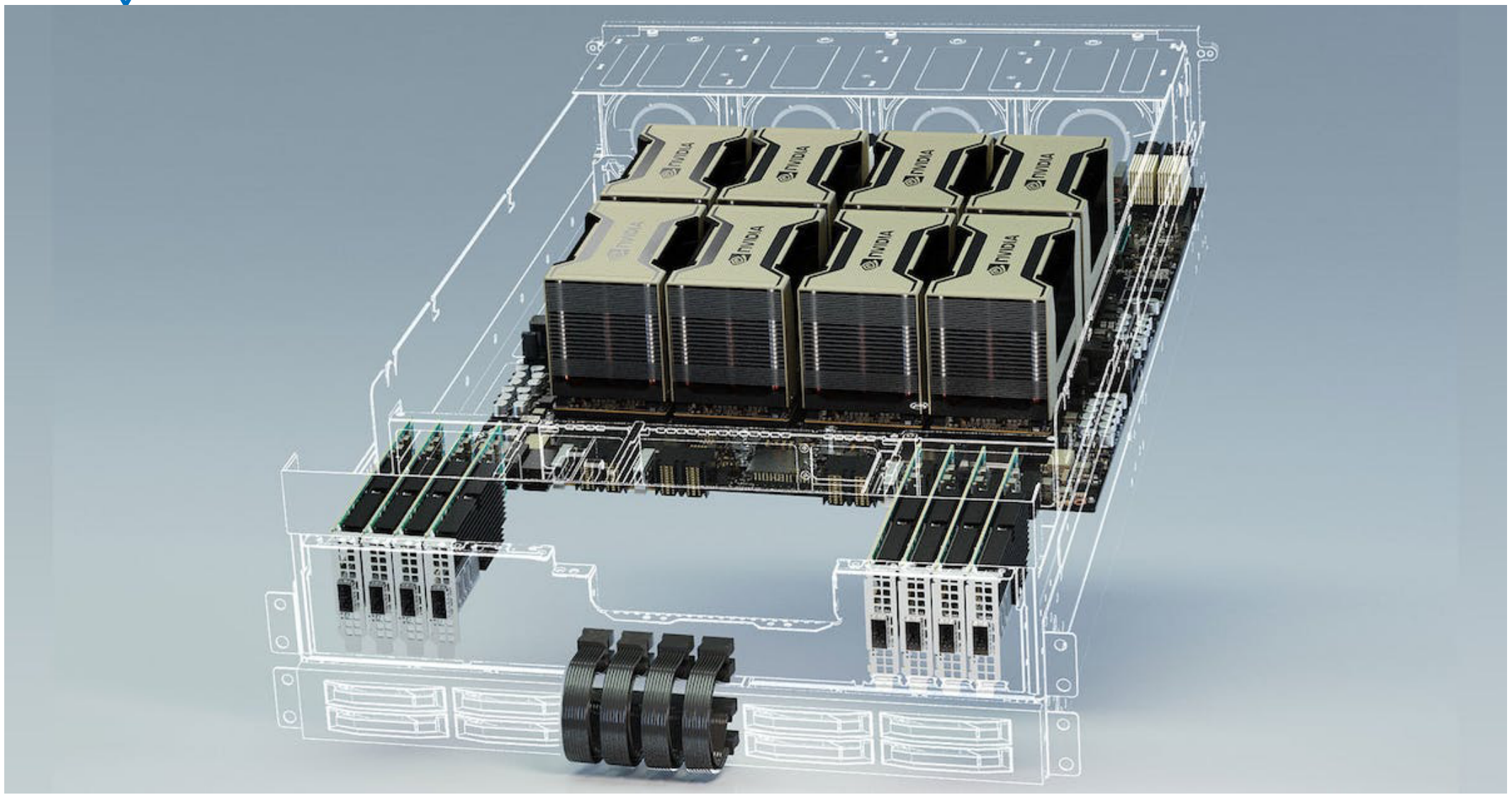
模型需求稳定、模型数量大 ✓ 大模型的统一任务建模能力，解决模型模型数量多、管理难问题

小模型不能解决的问题 ✓ AIGC：对话、问题、生成等

模型需求不稳定 ✓ 大模型的领域和任务泛化能力，解决领域迁移问题

特别提醒：大多数情况下仍然依赖一定数量的训练数据

如何搞



如何搞

DeepSpeed

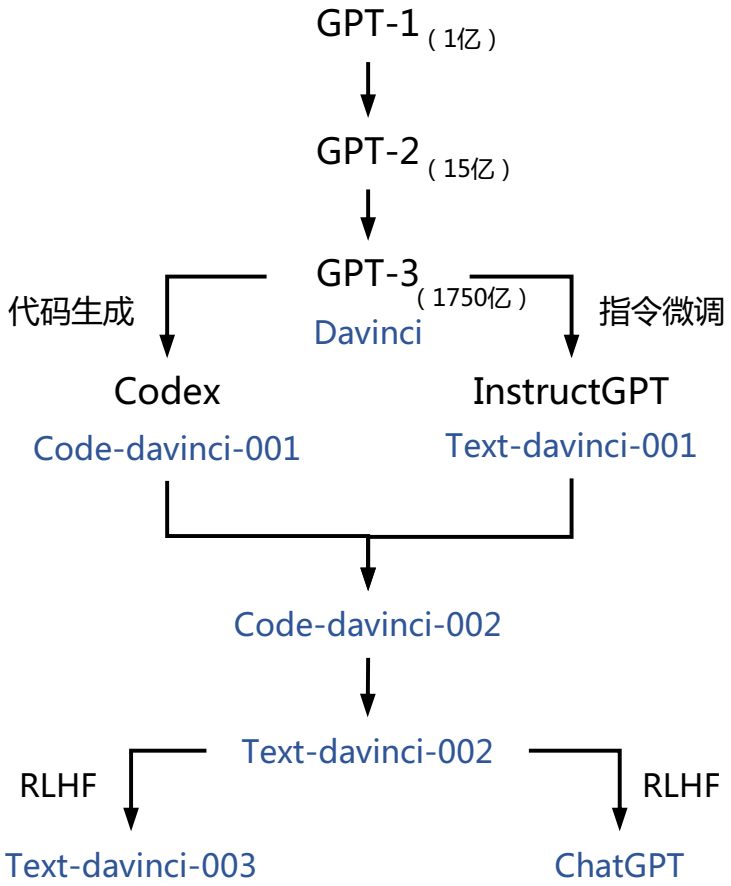
Megatron-DeepSpeed

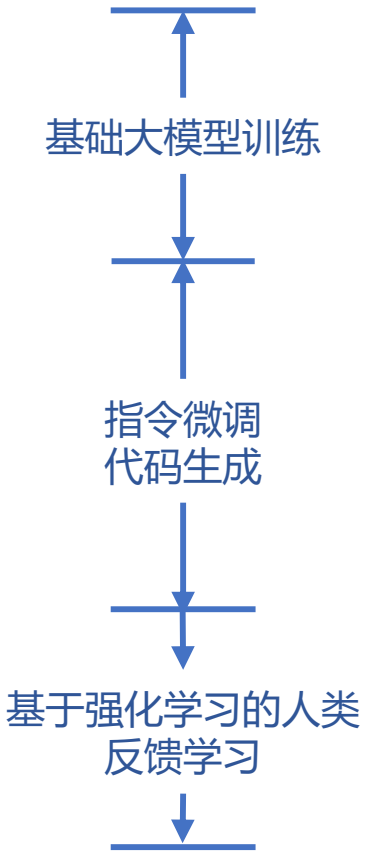
NVIDIA/Megatron-LM

Colossal-AI

LMFLOW

如何搞





长距离语言建模能力

理解人类指令能力

回答符合人类习惯的能力

如何搞

大模型需要：

1. 好的基础语言模型
2. 任务训练数据要足够
3. 任务训练方法要合适

如何搞

1. **基础语言模型选择**
   - 中英文能力、Code能力、模型大小
   - 任务相关文本增强
2. **任务数据构造**
   - 任务种类、指令构造 、数据构造
3. **类人回答**
   - Award函数选择、Award函数训练

如何搞

InstructUIE: Multi-task Instruction Tuning for Unified Information Extraction

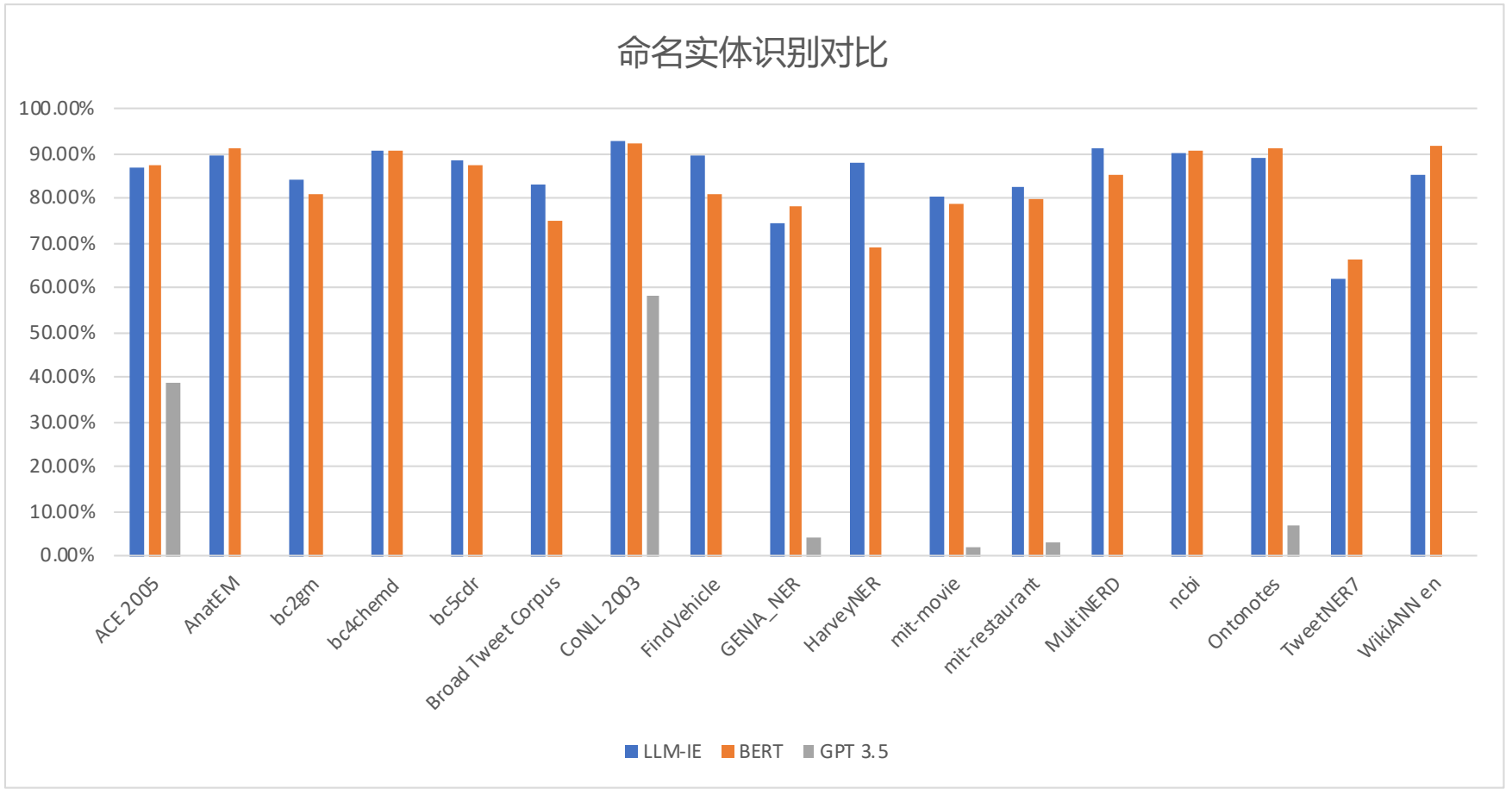
Task:NER

Please list all entity words in the text that fit the category.Output format is (word1, type1), (word2, type2).

Option: location, else, organization, person

Text: Germany 's representative to the European Union 's veterinary committee Werner Zwingmann said on Wednesday consumers should buy sheepmeat from countries other than Britain until the scientific advice was clearer .

Answer: (Germany, location), (European Union, organization), (Werner Zwingmann, person), (Britain, location)



不要神话和拟人化大模型

依然是统计机器学习范畴

C端 — 大模型

B端 — "小"模型


FUDAN UNIVERSITY
1905


FUDAN UNIVERSITY
NLP
NATURAL LANGUAGE PROCESSING

谢谢！