

第6章 大语言模型

宗成庆

中国科学院自动化研究所

cqzong@nlpr.ia.ac.cn



本章内容

- ➡ 1.ChatGPT 基本原理
- 2.ChatGPT 性能分析
- 3.思考与展望
- 4.习题



1. ChatGPT 基本原理

- 预训练语言模型(pre-training language model)
 - ELMo [Peters et al., 2018] (AllenAI&WU)
 - BERT [Devlin et al., 2019] (Google)
 - GPT: **G**enerative **P**re-**T**rained Transformer [Radford et al., 2018] (OpenAI)
 - GPT-3.5 (ChatGPT), Nov. 29, 2022
 - GPT-4.0, March 14, 2023
 - Sora, Feb. 15, 2024



1. ChatGPT 基本原理

模型	GPT (2018)	GPT 2.0 (2019)	GPT 3.0 (2020)	ChatGPT (2022)
参数量	1.17亿	15亿	1750亿	1750亿
学习目标	无监督语言模型(Pre-training)/有监督fine-tune	多任务, P(output input, task), Zero Short Task Transfer	小样本学习(Few shot)	基于人类反馈的强化学习(RLHF)
主要区别		增加语料、层数、维度 LN前移, 最后加LN, 初始化scale	增加语料、层数、维度	训练方法
Dataset	7,000 unpublished books, 长文较多, 约5G	WebText, 40GB, 8 million documents	Common Crawl, WebText2, Books1/s2, and Wikipedia, 45TB	+人工标注数据 +用户反馈数据
模型结构	12-layer decoder, 12 heads, dim 768, ff 3072	48 layers, dim 1600	96 layers, 96 heads, dim 12288	不变
训练参数	100 epochs, batch_size 64, sequence length of 512, lr 2.5e-4, BPE vocab 40,000	vocab 50,257, batch_size 512, context window 1024	context 2,048, $\beta_1=0.9$, $\beta_2=0.95$, $\epsilon=10^{-8}$	不变
Papers	Improving Language Understanding by GPT link	Language Models are Unsupervised Multitask Learners link	Language Models are Few-Shot Learners link	Training language models to follow instructions with human feedback

有消息透露说, GPT4.0 的参数量已经达到1.8万亿。



1. ChatGPT 基本原理

◆ ChatGPT的三项基本技术

- 预训练大模型 GPT-3/3.5

利用大量**文本数据(45T)**和**编程代码数据(179G)**，采用自回归语言模型，训练得到预训练大模型。

- 指令微调(instruction tuning)

利用API收集真实的人类指令，并由人工书写答案，进行指令微调。

- 基于人类反馈的强化学习(Reinforcement Learning from Human Feedback, RLHF)

基于人类反馈的强化学习对模型调优, 使模型生成的答案更符合用户偏好。

1. ChatGPT 基本原理

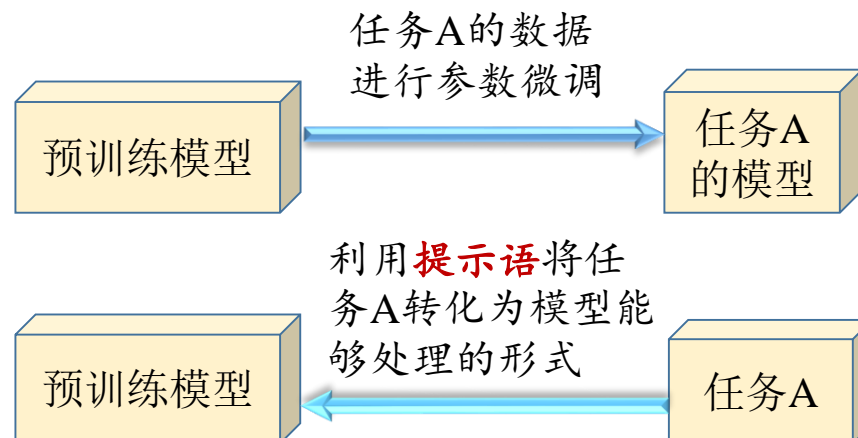
● 已有的方法

➤ 参数微调

预训练模型适应下游任务。

➤ 提示学习

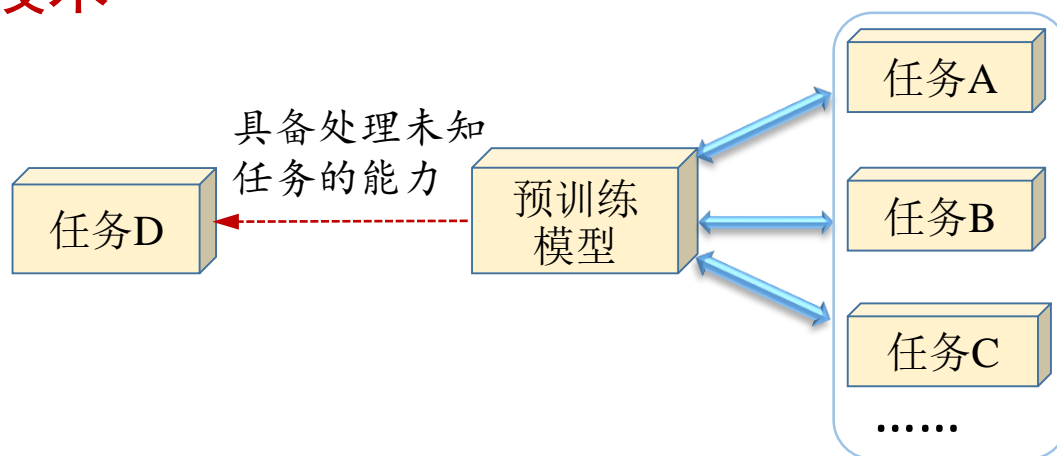
下游任务适应预训练语言模型。



● ChatGPT使用指令微调技术

(a) 预训练任务和下游任务互相适应。

(b) 将下游任务转化为指令，然后将预训练模型在多个任务上进行微调。



1. ChatGPT 基本原理

多任务数据集



多任务指令数据集

请把这句话翻译成英语“这个电影是真的不错”：this movie is very nice

“这个电影是真的不错”的情感极性是：正向

《流浪地球2》的导演是谁？郭帆

.....

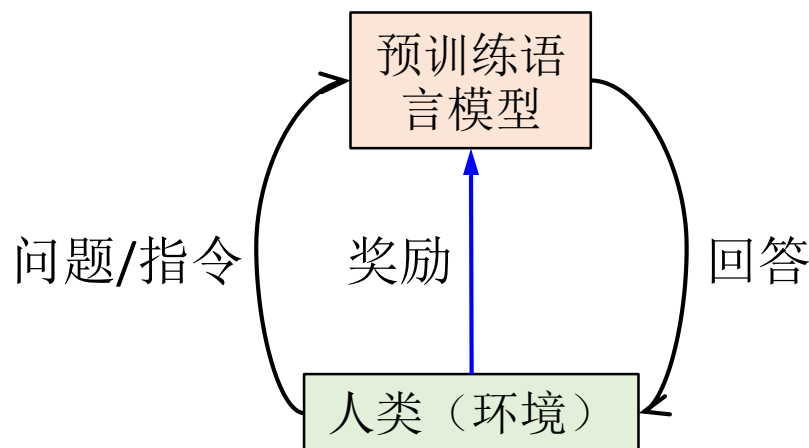


预训练语言
模型

1. ChatGPT 基本原理

● 基于人类反馈的强化学习

- 人工对预训练模型的多个答案进行打分，根据打分结果训练奖励模型(Reward Model, RM);
- 采用基于逼近策略优化(Proximal Policy Optimization, PPO)的强化学习方法，调整模型参数，使得**模型生成内容更符合人类的偏好**。



大数据 + 大算力 + 大量人工 \longrightarrow 大工程

1. ChatGPT 基本原理

基础大模型 GPT-3.0/3.5

Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

Explain the moon landing to a 6 year old

A labeler demonstrates the desired output behavior.

Some people went to the moon...

This data is used to fine-tune GPT-3 with supervised learning.

SFT

人工收集数据集，并书写答案，进行指令微调。

指令微调

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

Explain the moon landing to a 6 year old

A Explain gravity... B Explain war... C Moon is natural satellite of... D People went to the moon...

A labeler ranks the outputs from best to worst.

D > C > A = B

This data is used to train our reward model.

RM

人工构建排序数据集，训练奖励模型

基于人类反馈的强化学习

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

Write a story about frogs

The policy generates an output.

PPO

The reward model calculates a reward for the output.

RM

The reward is used to update the policy using PPO.

r_k

利用奖励函数，进行人工反馈的强化学习。

ChatGPT



1. ChatGPT 基本原理

Step 1: 指令学习/指令微调(instruction tuning)

指令微调: 语言模型在指令描述的多项任务上进行有监督的
微调(Supervised Fine-Tuning, SFT), 它能够显著地提升模型在未知
任务零样本条件下的表现。

以下是一个描述任务的指令:

用户: 将这句话翻译成中文: Why do birds fly?

回复: 鸟为什么会飞?

目标函数:

$$\mathcal{L}(\mathcal{D}; \theta) = - \sum_i \sum_j \log P_{\theta}(t_{ij} | s_i, t_{i, < j})$$

s 为指令描述中的任务; t 是回复样本;
 i 和 j 分别是第几个词。



1. ChatGPT 基本原理

Step 2: 收集人工排序的答案，训练奖励模型 r

指令：鸵鸟大还是麻雀大：

回复：

- a. 鸵鸟大，鸵鸟的体积和重量都远远超过麻雀，一只鸵鸟的重量通常有十几甚至几十千克，而一只麻雀只有几十克重。
- b. 鸵鸟大，鸵鸟的体积比麻雀要大得多。
- c. 这个问题不好回答，要看比较的是成年鸵鸟还是未成年的。
- d. 无法简单回答。

答案质量排序： $a > b > c = d$

好的回
复得分

差的回
复得分

$$\text{loss}(\theta) = -\frac{1}{\binom{K}{2}} E_{(x, y_w, y_l) \sim D} [\log(\sigma(r_\theta(x, y_w) - r_\theta(x, y_l)))]$$

参数 回复的个数 问题 好的回复 差的回复 指令集



1. ChatGPT 基本原理

Step 3: 利用强化学习技术优化策略(PPO)

强化学习的概率 初始概率

$$\text{objective}(\phi) = E_{(x,y) \sim D_{\pi_{\phi}^{\text{RL}}}} \left[r_{\theta}(x, y) - \beta \log \left(\frac{\pi_{\phi}^{\text{RL}}(y | x)}{\pi^{\text{SFT}}(y | x)} \right) \right] + \\ \gamma E_{x \sim D_{\text{pretrain}}} \left[\log(\pi_{\phi}^{\text{RL}}(x)) \right]$$

预训练语言模型

π 是学习到的策略（模型）；

y 是模型生成的回复；

β, γ 分别是权重，惩罚因子。



1. ChatGPT 基本原理

- 预训练语言模型对齐(Alignment) — 从[Instruct GPT](#)开始

SFT Data			RM Data			PPO Data		
split	source	size	split	source	size	split	source	size
train	labeler	11,295	train	labeler	6,623	train	customer	31,144
train	customer	1,430	train	customer	26,584	valid	customer	16,185
valid	labeler	1,550	valid	labeler	3,488			
valid	customer	103	valid	customer	14,399			

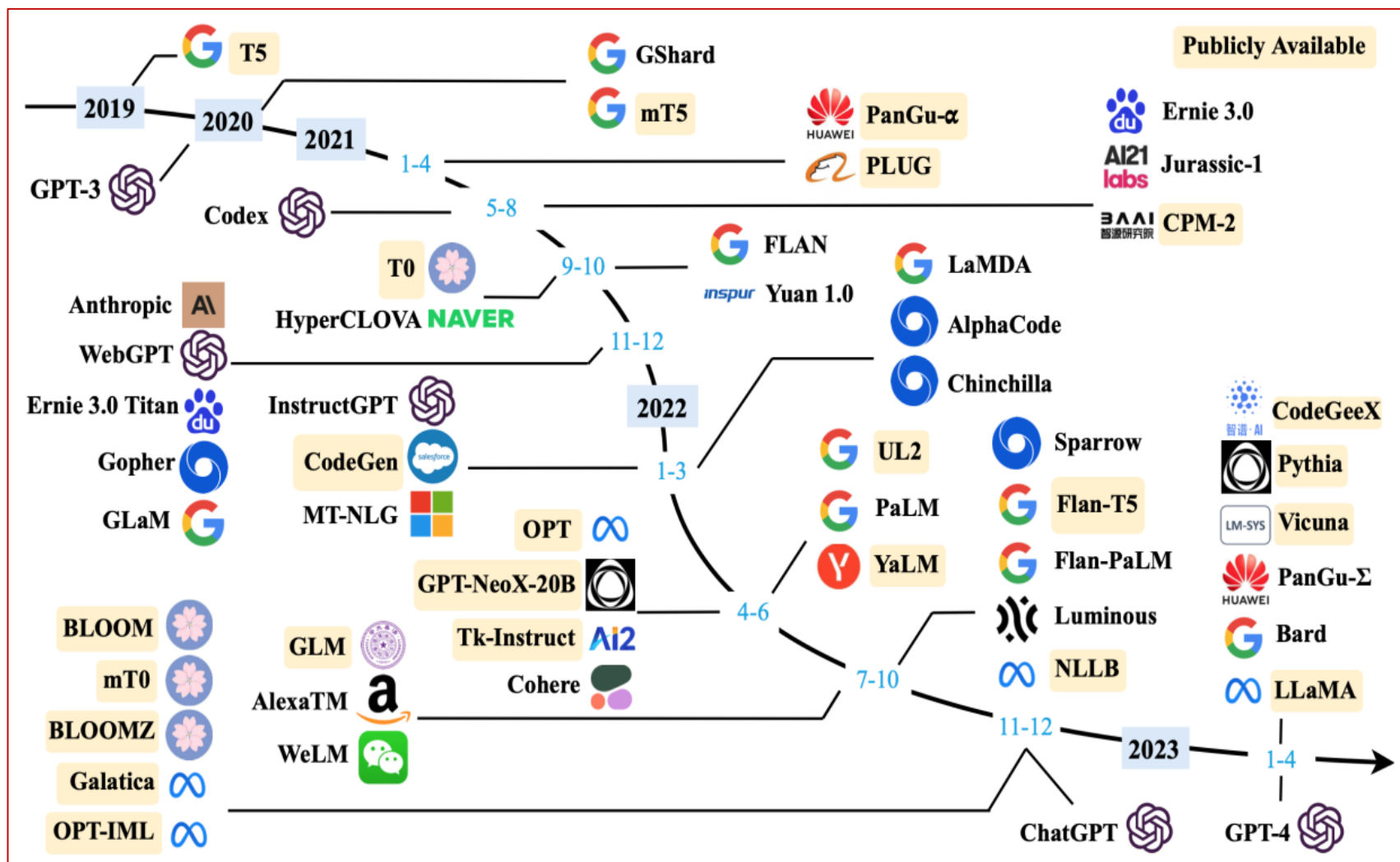
OpenAI : Training language models to follow instructions with human feedback

labeler: 专门雇人撰写的微调数据集，即“提示(prompt)”；

customer: 测试人员使用的数据集。

Long Ouyang et al. Training language models to follow instructions with human feedback.
In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS) 35*
(2022): 27730-27744.

1. ChatGPT 基本原理





本章内容

1.ChatGPT 基本原理

➡ 2.ChatGPT 性能分析

3.思考与展望

4.习题



2. ChatGPT 性能分析

◆ ChatGTP 性能测试(2022.12.15): 关于写作

NW

我们聊聊滑雪吧。你会滑雪吗？



对不起，我是一个纯粹的语言模型，我无法感知或了解外界环境。我无法参与体育活动，也无法对滑雪进行评价。我只能根据您提供的信息来回答问题，不能对外界环境进行评价或描述。如果您有其他问题，我将尽力回答您的问题。

NW

你能写一首关于大雪的诗歌吗？




虽然我不会滑雪，但我可以尝试写一首关于大雪的诗歌：

大雪纷飞，
银装素裹。
白茫茫一片，
如同洁白的棉被。
踏着深厚的积雪，
挥动双手捕捉飞舞的雪花。
欢乐的嬉戏，
让人忘却寒冷的冬日。
希望您喜欢这首诗。



2. ChatGPT 性能分析

NW 还不错，就是不押韵。

 我抱歉地通知您，我只是一个纯粹的语言模型，我无法感知或了解外界环境，也无法理解诗歌的韵律和押韵。我的目的是提供简单的信息和回答问题，而不是写诗歌。如果您对我的诗有任何批评或建议，我将非常感激。

NW 你可以把这首诗翻译成英文吗？

 是的，我可以尝试把这首诗翻译成英文：

The snow is falling heavily,
Dressed in silver clothes.
The vast expanse is white,
Like a pure cotton blanket.
Walking on the deep snow,
Swinging your hands to catch the flying snowflakes.
Happy play,
Forget the cold winter.

希望您能喜欢这首诗的英文版本。

大雪纷飞，
银装素裹。
白茫茫一片，
如同洁白的棉被。
踏着深厚的积雪，
挥动双手捕捉飞舞的雪花。
欢乐的嬉戏，
让人忘却寒冷的冬日。



2. ChatGPT 性能分析


NW 你能把这首诗改写成散文吗？

 是的，我可以尝试把这首诗改写成散文：

当大雪纷飞，披着银色的衣裳时，白雪覆盖了整个世界，就像一床洁白的棉被。踩着厚厚的积雪，挥动双手捕捉飞舞的雪花，在冬日的寒冷中尽情嬉戏，忘却一切烦恼。

希望您喜欢这篇散文。

NW 谢谢你！我快要下班了。

 不客气，感谢您提出宝贵的问题。如果您有其他问题，欢迎随时联系我，我将尽力为您解答。祝您有一个愉快的下午！



2. ChatGPT 性能分析

关于医疗

Model: Default (GPT-3.5)

2023.4.18

NW

有人长期运动，喜欢打球和爬山，可是最近左脚底前面部分出现麻木、疼痛，甚至剧痛症状。请问是什么原因造成的？



这种疼痛和麻木的症状可能是由多种因素引起的。以下是一些可能的原因：



1. 运动损伤：如果您运动时使用不正确的姿势或过度使用脚部肌肉，可能会导致肌肉、韧带或肌腱的拉伤或扭伤，从而导致疼痛和麻木。
2. 跖筋膜炎：跖筋膜是脚底部的一层薄膜，它支撑着脚弓。如果这个膜发炎，可能会导致脚底疼痛和麻木。
3. 足底筋膜纤维化：这是一种常见的足底疼痛症状，可能是由于过度运动、过度使用或足部负重过重引起的。
4. 神经问题：麻木和疼痛也可能是由于神经问题引起的，例如坐骨神经痛或足底神经痛。

如果症状持续存在，建议您咨询医生进行检查，以确定症状的确切原因，并获得适当的治疗。

2. ChatGPT 性能分析

NW

A和B两个人在饭店吃饭时发生争执，A先动手打了B，但没有造成重伤。B报警后民警将两人带到派出所，B不同意和解。请问民警应该如何处理？

关于法律



在这种情况下，民警应该根据相关法律法规，对A和B进行调查和处理，以维护公正和法律的权威。



1. 对A和B进行询问和取证：民警应当向A和B了解事发经过，了解是否有其他证人或证据，如监控录像等，以收集证据。
2. 进行医学鉴定：民警应当将B送到医院进行医学鉴定，确定受伤程度，以作为后续处理的参考。
3. 根据相关法律规定进行处理：如果A的行为构成了刑事犯罪，民警应当依照法律程序进行处理，可能会涉及到行政拘留或刑事处罚等。如果A的行为不构成犯罪，民警应当及时调解双方，促成和解，或者将案件移交到相关部门进行调解或处理。
4. 加强宣传教育：民警应当向双方进行法律宣传和教育，提高双方的法律意识和法律素养，以预防类似事件的发生。

在处理这种情况时，民警应当尽力保持公正和客观，遵循法律程序，依法处理，确保双方的合法权益得到维护。



2. ChatGPT 性能分析

Model: GPT-4

关于算术

NW

When I was 6 my sister was half my age. Now I'm 70 how old is my sister?



When you were 6 years old, your sister was half your age, which means she was 3 years younger than you. Now that you're 70, your sister would be $70 - 3 = 67$ years old.





2. ChatGPT 性能分析



You

There are a total of 80 students in our class, including 30 from Africa, 25 from Europe, 15 from Asia, and 10 from other regions. What is the proportion of students in different regions?



ChatGPT

To find the proportion of students from each region, we can calculate the fraction of students from each region out of the total number of students.

1. Africa: $\frac{30}{80} = 0.375$ or 37.5%
2. Europe: $\frac{25}{80} = 0.3125$ or 31.25%
3. Asia: $\frac{15}{80} = 0.1875$ or 18.75%
4. Other regions: $\frac{10}{80} = 0.125$ or 12.5%

So, the proportions of students from different regions are approximately:

- Africa: 37.5%
- Europe: 31.25%
- Asia: 18.75%
- Other regions: 12.5%

2. ChatGPT 性能分析

- **Feb. 15, 2024, Sora**, Creating video from text

Prompt: A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.



2. ChatGPT 性能分析





2. ChatGPT 性能分析

◆ 拥有渊博的知识和强大的分析、对话能力

➤ 很强的对话能力

- 多轮对话的上下文理解
- 用户意图的理解能力
- 回复质量和逻辑一致性
- 甚至可以“交互式学习”

➤ 很好的语言生成和表达能力

- 规范的语言表达能力
- 长文本生成能力
- 代码生成能力
- 多语言生成能力
-

➤ 丰富的知识储备

- 常识性知识
- 百科知识
- 数学问题求解
-

- 可大幅度提高工作效率
- 极大地拓展个人视野
- 在各领域、各行业都有中等水平以上的表现

● **定位—AI助手**

2. ChatGPT 性能分析

◆但是，它像人一样也会犯错误

- 特定领域的专业知识匮乏
- 模型的内生性缺陷
 - 可解释性
 - 知识来源缺乏引用
 - 出现“臆想/幻觉”答案
 - 事实性错误……
- 由于对齐而产生人为倾向性
 - 回复带有主观性倾向/价值观
 - 无法展现和平衡多种观点
 - 缺乏灵活的角色和具身感知能力……
- 对低资源语言的处理能力差
- 任务规划能力有限
- 多步推理能力弱
- 训练成本高…



2. ChatGPT 性能分析

NW

鲁迅和周树人是一个人吗？



不，鲁迅和周树人不是一个人。鲁迅（原名周樟寿）和周树人是两位中国现代著名的作家，他们都有重要的文学成就，但是他们是两个不同的人。

鲁迅（1881年9月25日 - 1936年10月19日）是中国现代文学的先驱之一，他是一位伟大的思想家、文学家和革命家，被誉为“百年不遇之才”。他的代表作品包括《呐喊》、《狂人日记》、《阿Q正传》等，通过他的作品，鲁迅揭露了当时社会的黑暗和尖锐批判了封建传统。

周树人（1896年1月14日 - 1967年11月13日），原名周樟寿（与鲁迅同名），后改名周树人以示区别。他是中国现代作家、文学批评家和教育家，是现代白话文运动的代表人物之一。周树人的代表作品有《封神演义》、《茶馆》等，他在文学上提倡运用白话文进行写作，推动了现代文学的发展。

虽然鲁迅和周树人的原名都是周樟寿，但他们是两个不同的个体，分别有独立的文学成就和不同的文学风格。

一本正经地胡说八道

2. ChatGPT 性能分析

NW

你知道“林黛玉倒拔垂杨柳”的故事吗？



是的，我了解“林黛玉倒拔垂杨柳”的故事。这个故事出自中国古典文学名著《红楼梦》。

《红楼梦》是清代作家曹雪芹创作的长篇小说，被誉为中国四大名著之一。故事讲述了贾宝玉、林黛玉、薛宝钗等人的爱情故事，以及贾家荣华富贵和衰落的历程。

在《红楼梦》中，林黛玉是贾宝玉的表妹，是书中主要女主角之一。在故事中，林黛玉在园中倒拔垂杨柳的情节发生在贾府的花园里。她因受到一些事情的打击和伤害，心情非常沮丧和痛苦，于是决定在花园里倒拔垂杨柳来发泄情绪。

垂杨柳是柔弱的树种，它的形象往往被用来比喻柔弱而忧愁的女性。林黛玉倒拔垂杨柳的场景，表现了她内心的忧愁和悲伤，也反映了她在贾府中处境的无奈和困苦。

这个情节也成为《红楼梦》中的一个重要象征，用来表达人物的心情和命运，以及寓意着人生的起伏和变化。



2. ChatGPT 性能分析

◆对“文心一言”的测试



鲁迅和周树人是同一个人吗？



鲁迅和周树人是同一个人。

鲁迅，原名周樟寿，后改名周树人，字豫山，后改字豫才，浙江绍兴人。著名文学家、思想家、革命家、教育家、民主战士，新文化运动的重要参与者，中国现代文学的奠基人之一。

所以，周树人和鲁迅是同一个人。



你知道“林黛玉倒拔起杨柳”的故事吗？



对不起，似乎没有“林黛玉倒拔起杨柳”的故事。林黛玉是《红楼梦》中的主要人物，她的性格和行为在小说中有着详细的描述，其中并没有提到她倒拔杨柳的情节。倒拔杨柳是名著《水浒传》中的情节，小说中鲁智深倒拔垂杨柳，《红楼梦》中没有林黛玉倒拔杨柳的情节，所以您的问题可能不符合事实。



2. ChatGPT 性能分析

◆ ChatGPT在4个NLP专业任务上的对比分析

- 机器翻译
- 信息抽取
- 文本摘要
- 情感分析

(1) 高资源语言：中文，德语
(2) 中等资源语言：爱沙尼亚语，立陶宛语
(3) 低资源语言：僧伽罗语，尼泊尔语

语言对	COMET			BLEU		
	ChatGPT	Google	比值	ChatGPT	Google	比值
汉语→英语	88.7	89.4	99.3	29.6	38.6	76.7
德语→英语	90.7	91.0	99.7	43.0	47.3	90.9
爱沙尼亚语→英语	90.1	91.7	98.3	36.9	45.5	81.1
立陶宛语→英语	86.1	89.2	96.6	32.6	41.3	78.9
尼泊尔语→英语	87.4	93.2	93.8	24.5	51.3	47.7
僧伽罗语→英语	61.3	90.5	67.7	3.5	45.5	7.7

**Flores-200
测试集**



2. ChatGPT 性能分析

➤ 翻译实例

模型	翻译实例
Source	On Dec. 17, according to disclosure documents, Paul Pelosi bought 10 Alphabet call options with a strike price of \$2,000 with an expiration date set for mid-September 2022.
ChatGPT	根据披露文件, 保罗 佩洛西于12月17日购买了10份 谷歌母公司 的认购期权, 行使价格为2000美元, 到期日定于2022年9月中旬。
Google Translate	根据披露文件, 12 月 17 日, 保罗 佩洛西 (Paul Pelosi) 购买了 10 份 Alphabet 看涨期权, 行使价为 2,000 美元, 到期日定为 2022 年9月中旬。
Source (尼泊尔语)	चराहरूमा प्वाँखको विकासको अन्तरदृष्टि पनि यस खोजले प्रदान गर्छ।
ChatGPT	This search also provides insight into the development of wings in insects .
Google Translate	The discovery also provides insight into the evolution of feathers in birds .
Reference	The find also grants insight into the evolution of feathers in birds .

武俊宏等, ChatGPT 能力分析与未来展望, 中国科学基金, 2023年05期 v.37 第735-742页



2. ChatGPT 性能分析

➤ 关于幻觉表现的对比(hallucinations)

模型	翻译实例
Source (尼泊尔语)	चराहरूमा प्वाँखको विकासको अन्तरदृष्टि पनि यस खोजले प्रदान गर्छ।
ChatGPT	This search also provides insight into the development of wings in insects .
Google Translate	The discovery also provides insight into the evolution of feathers in birds .
Reference	The find also grants insight into the evolution of feathers in birds .
Source (僧伽罗语)	ඔබ සතපුම් දහස් ගණනක් දුරින් ඉන්නාකෙනෙකුට කථාකරන විට, ඔබ එන්ද්රිකඩක් භවිතාකරයි.
ChatGPT	When you speak with someone who understands your language , you use Chinese .
Google Translate	When you call someone who is thousands of miles away , you are using a satellite .
Reference	When you speak to someone thousands of miles away , you use a satellite .

武俊宏等, ChatGPT 能力分析与未来展望, 中国科学基金, 2023年05期 v.37 第735-742页

2. ChatGPT 性能分析

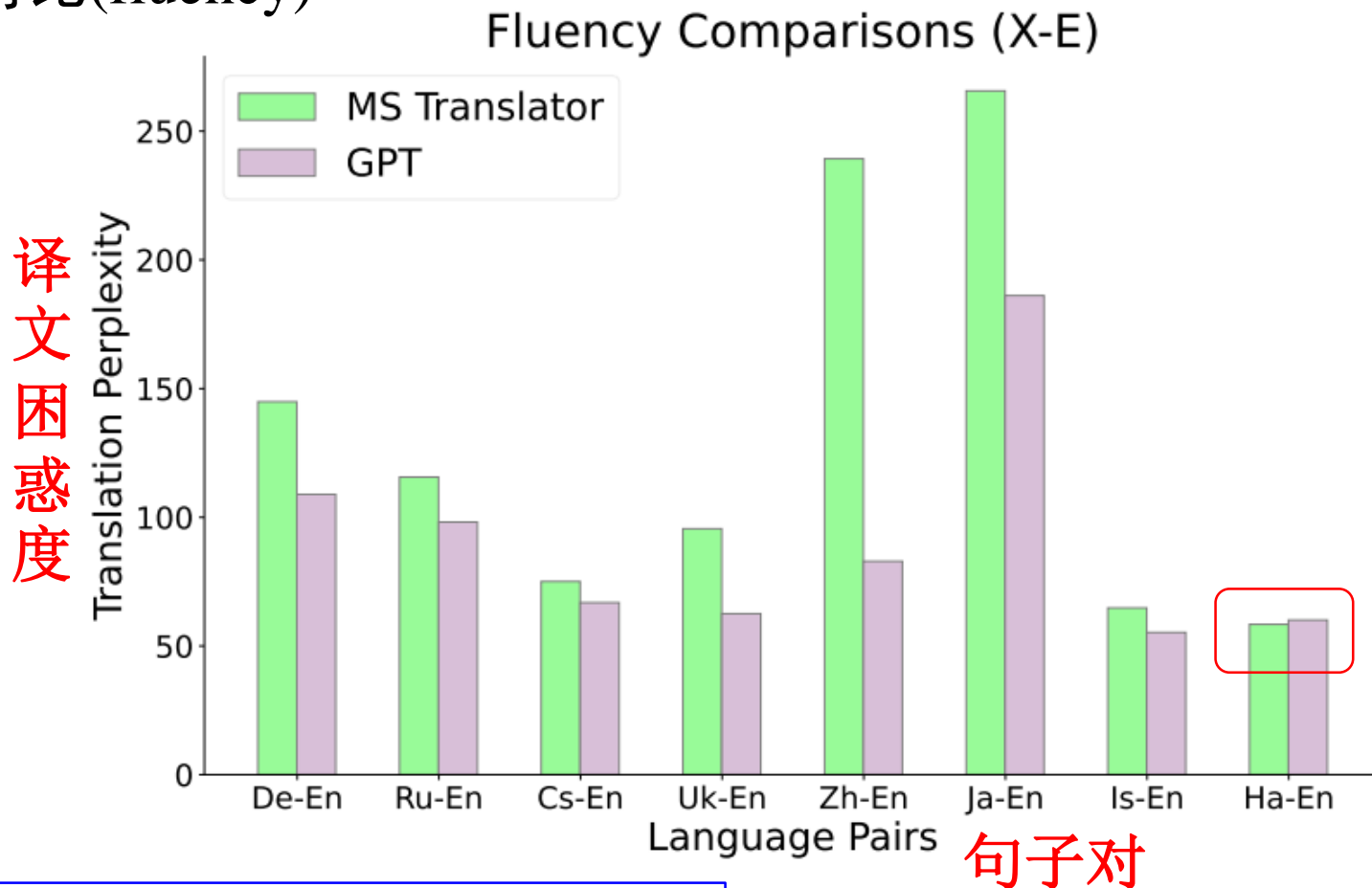
➤ 对特殊语言现象的处理能力

Sequence Type	Translation Instance	Phenomenon
Source MS Translator GPT	Bis auf die E 95 02 wurden <u>alle Lokomotiven zerlegt</u> . With the exception of E 95 02, all locomotives were dismantled . All locomotives were dismantled except for the E 95 02.	Non-Monotonicity (NM)
Source MS Translator GPT	Oder ist sie ganz aus dem Sortiment genommen? Or is it completely removed from the range? Or has it been completely removed from the range?	Fluency (F)
Source MS Translator GPT	Sehen Sie bitte im Screenshot was der Kollege geschrieben hat Please see in the screenshot what the colleague wrote Please see the screenshot for what the colleague wrote.	Punctuation Insertion (PI)
Source MS Translator GPT	Die Email zur Stornierung wurde am 26.12. <u>#NUMBER#</u> versendet. The cancellation email was sent on 26.12. <u>#NUMBER#</u> . The cancellation email was sent on December 26th.	Dropped Content (USW)
Source MS Translator GPT	"We won't accept the <u>CAA</u> and that is for sure. “我们不会接受CAA，这是肯定的。 “我们不会接受 <u>《公民法》</u> ，这是肯定的。	Inserted Content (UTW)

Amr Hendy et al., Good Are GPT Models at Machine Translation? A Comprehensive Evaluation, <https://arxiv.org/abs/2302.09210>

2. ChatGPT 性能分析

➤ 流畅性对比(fluency)



De- 德语 Ru- 俄语 CS- 捷克语 Uk- 乌克兰语
Zh- 中文 Ja- 日语 Is- 冰岛语 Ha- 豪萨语



2. ChatGPT 性能分析

◆在4个NLP专业任务上的对比分析

- 机器翻译
- **信息抽取**
- 文本摘要
- 情感分析

在命名实体识别、关系抽取和事件抽取三项任务上，ChatGPT的性能最高只能达到专门训练出来的最优模型性能的63.5%、43.0%和35.3% [Han et al., 2023]。

CoNLL04: 人工评价，仍不如现有最好的系统。

[Han et al., 2023] R. Han et al., Is Information Extraction Solved by ChatGPT? An Analysis of Performance, Evaluation Criteria, Robustness and Errors. (2023-05-23)/[2023-06-21]. <https://arxiv.org/abs/2305.14450>



2. ChatGPT 性能分析

◆在4个NLP专业任务上的对比分析

- 机器翻译

- 信息抽取

- **文本摘要** \Rightarrow 在CSDS中文对话数据集上测试[Lin et al., 2021]:

- 情感分析



SemEval-2014
情感极性三元组

Model	ROUGE-2	ROUGE-L	BERTScore
ChatGPT	17.5	39.1	70.4
Fast-RL	41.4	47.1	79.8

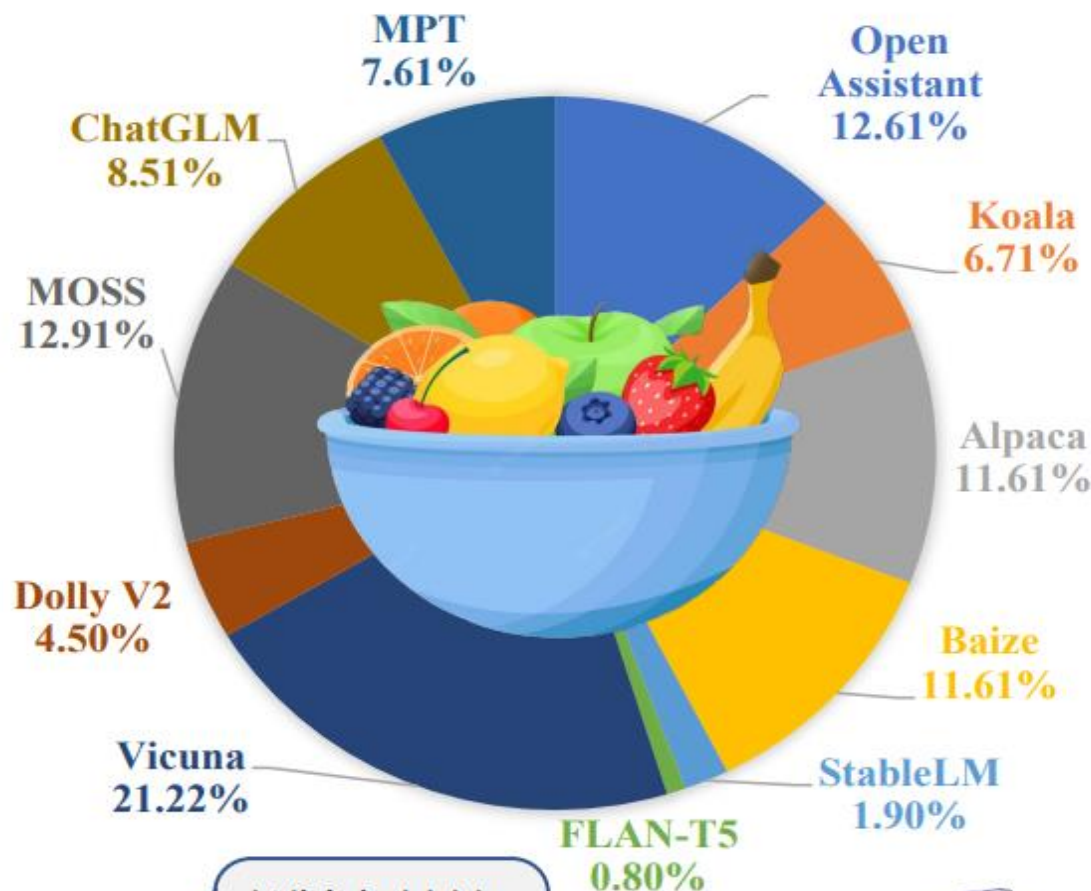
约为该数据集上最优模型BDTS的 **58.6%**[Zhang et al., 2022]。

[Lin et al., 2021] H. Lin et al. 2021. CSDS: A fine-grained Chinese dataset for customer service dialogue summarization. *Proc. of EMNLP2021*, pages 4436–4451

[Zhang et al., 2022] Y. Zhang et al. 2022. Boundary-Driven Table-Filling for Aspect Sentiment Triplet Extraction. *Proc. EMNLP'2022*, pp. 6485–6498

2. ChatGPT 性能分析

Percentage of Examples Where Each Model Ranks First



- MixInstruct
- 3 Metrics (BERTScore, BLEURT, BARTScore)
- GPT-Rank



Which LLM
should I use

All! I can





本章内容

1.ChatGPT 基本原理

2.ChatGPT 性能分析

➡ 3.思考与展望

4. 习题

3. 思考与展望

◆大模型研究与应用

- 大模型技术改进



- 大模型应用

- ✓ 可控的文本生成
- ✓ 模型幻觉消除
- ✓ 模型压缩
- ✓ 低资源语言模型
- ✓ 意识形态对齐
- ✓ 模型评测
- ✓ 模型推理过程的可解释性及结果溯源
-

3. 思考与展望

◆大模型研究与应用

- 大模型技术改进

- 大模型应用

➤ 应用模式

◇ 基础大模型 + API

◇ 直接使用

特别注意：

- 意识形态问题
- 知识产权
- 正确性核对
- 信息安全

教学：

- 批改作业
- 设计/代写设计教案
- 文献检索和信息推荐
- 语料标注

4. 思考与展望

◆ 大模型会成为NLP研究的终结者吗？

- ChatGPT 的“成功”证明了生成式语言模型可能是实现“通用AI”的一条可行之路，但是，它还远没有成功！
- 面向落地应用的“最后一公里”，如何针对特定应用场景、特定任务和特定语言等具体问题，开展技术改进和创新，将是未来大模型未来发展的重要一步；
- 我们不仅需要掌握大模型技术应用的基本能力，更重要的是熟悉模型的底层逻辑、数学原理和模型构建的创新思维；
- 学习和跟踪国际前沿技术是科学研究的基本要求，但跟踪不等于盲从，更不是故弄玄虚！静下心来，冷静思考，守正创新，才是超越和发展之根本。

3. 思考与展望

◆未来展望

- 以 ChatGPT 为代表的大语言模型对科研、教学、人才培养以及各行各业的冲击势不可挡，思维转换刻不容缓；
- 例如，一个新语言对机器翻译系统的研发周期越来越短(几年→几个月→几个周→几天)，译文质量越来越好，翻译效率越来越高，语种范围越来越大；
- 传统的作坊式的科研模式远远不能跟上时代快速发展的脚步。中低水平，尤其从事重复性文字（包括翻译）工作的人员（岗位/职业/行业）被计算机取代成为不可逆转的趋势；
- 运用新技术，将AI与相关技术领域相结合，相辅相成，协同创新，是未来基础研究和应用技术研发的必然之路。



本章小结

- ◆ ChatGPT 的基本原理：3个关键步骤
- ◆ ChatGPT 的性能现状分析：全能与幻觉
- ◆ 展望与思考：提高想象力，加强创新意识



本章内容

- 1.ChatGPT 基本原理
- 2.ChatGPT 性能分析
- 3.思考与展望

 4.习题



习题

1. 分析对比你能够正常使用的大模型性能，比较它们的性能差异，包括在不同任务、不同语言以及对不同提问方式的处理能力等你能想到的各方面。**切忌违规使用境外的大模型。**
2. 针对某个你认为性能表现较好的大模型，总结出指令(prompt)撰写的经验和建议。
3. 试设计一种方法，自动判断大模型生成内容的错误。
4. 试设计一种方法，自动判断大模型生成内容的逻辑一致性。
5. 试设计一种方法，用于定量评价大模型性能的综合得分。

谢谢!

Thanks!

