

大语言模型的能力和未来



中国科学院 信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING, CAS

姓 名：王青悦
年 级：2018级直博生
小 组：第二研究室内容组
时 间：2023年03月10日

主要内容

- 大语言模型的商业化
- 大语言模型的能力
- 大语言模型的未来

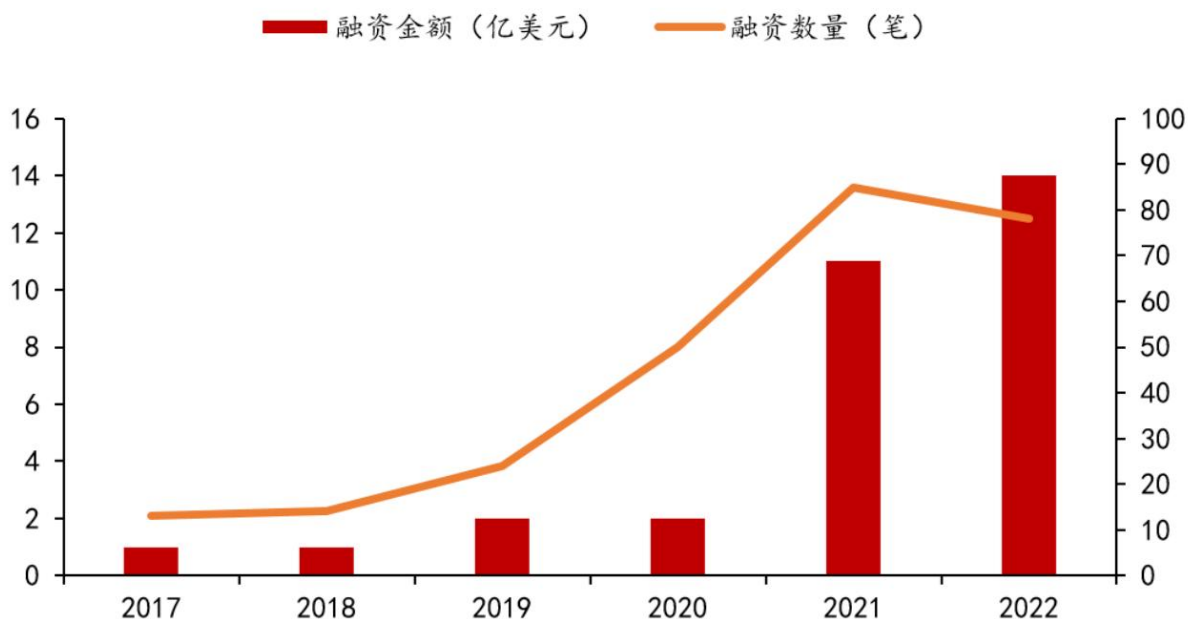
主要内容

- 大语言模型的商业化
 - 风头正盛的生成式AI
 - 大语言模型的应用落地
- 大语言模型的能力
- 大语言模型的未来

风头正盛的生成式AI

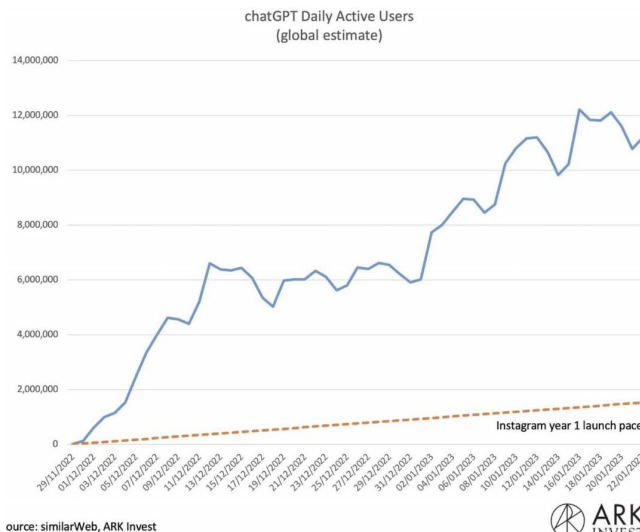
□ **生成式 AI 赛道火热**。根据 PitchBook 统计数据，2022 年生成式 AI 赛道总共获得约 14 亿美元的融资，几乎达到了过去 5 年的总和。OpenAI、Stability AI 等明星公司，其他初创企业如 Jasper、Regie.AI、Replika 等均获得资本青睐。

图表 11：2017-2022 年生成式 AI 赛道融资金额（亿美元）及数量（笔）



风头正盛的生成式AI

- 2022年10月，Stability AI获得约1亿美元融资，发布的开源模型**Stable Diffusion**，可以根据用户输入的文字描述生成图片，引爆AI作画领域。
- 2022年11月30日，**ChatGPT**在宣布公测之后，上线5天，全球用户数量已突破百万。上线不到40天，日活用户已突破千万。
- 2023年3月9日，微软德国CTO Andreas Braun在一场活动中表示，**GPT-4**将在下周发布，将提供多模态模型。未来的输入输出将出现图像、视频等形态，或将在业内产生更具颠覆性的影响。



"We will introduce GPT-4 next week"

"We will introduce GPT-4 next week, there we will have multimodal models that will offer completely different possibilities - for example videos," Braun said. The CTO called LLM a "game changer" because they teach machines to understand natural language, which then understand in a statistical way what was previously only readable and understandable by humans. In the meantime, the technology has come so far that it basically "works in all languages": You can ask a question in German and get an answer in Italian. With multimodality, Microsoft(-OpenAI) will "make the models comprehensive".

大语言模型的应用落地

□ ChatGPT+传媒：实现新闻智能写作，提升新闻实效性



2014年3月，美国洛杉矶时报网站的机器人记者Quakebot，在洛杉矶地震后仅3分钟，就写出相关信息并进行发布



中国地震网的写稿机器人在九寨沟地震发生后7秒内就完成了相关信息的编发



美联社使用的智能写稿平台Wordsmith可以每秒写出2000篇报道



第一财经“DT稿王”一分钟可写出1680字

□ ChatGPT+影视：根据大众兴趣定制影视内容，获得更高收视率，票房和口碑降低影视制作团队在内容创作的成本，提高创作效率。

2020年，美国查普曼大学的学生利用OpenAI的GPT-3模型创作剧本并制作短片《律师》

案例1



国内海马轻帆科技公司推出的“小说转剧本”智能写作功能，服务了包括《你好，李焕英》《流浪地球》等爆款作品在内的剧集剧本30000多集、电影/网络电影剧本8000多部、网络小说超过500万部

案例2

2016年，纽约大学利用人工智能编写剧本《Sunspring》，经拍摄制作后入围伦敦科幻电影48小时前十强



案例3



大语言模型的应用落地

□ ChatGPT+营销：充当虚拟客服，助力产品营销

- 24小时的产品介绍和在线服务，降低营销成本
- 快速了解客户需求，紧跟科技潮流。
- 稳定可靠的咨询服务，可控性和安全性强。

□ ChatGPT+娱乐：实时的聊天对象，增强陪伴性和趣味性。

□ ChatGPT+教育：提供全新的教育工具，通过自助提问来快速查缺补漏。

□ ChatGPT+金融：实现金融资讯，金融产品的自动化生产，塑造虚拟理财顾问。

□ ChatGPT+医疗：快速了解患者病情且及时反馈，第一时间情感支持。

.....

ChatGPT 撰写房源信息



often_says_nice - 4天前

Prompt: Write a javascript function to print out a large dog in the debug console.

Response:

▶ To print out a large dog in the debug console, you could use the following JavaScript function:

```
function printDog() {
  console.log(" / \ / \ ");
  console.log(" ( o o )");
  console.log(" > ^ < ");
  console.log(" / \ ");
  console.log(" / \ ");
  console.log(" / _ \ ");
  console.log(" / / \ \ ");
  console.log(" / / \ \ ");
  console.log(" / / \ \ ");
  console.log(" / / \ \ ");
  console.log(" / / \ \ ");
  console.log(" / / \ \ ");
  console.log(" / / \ \ ");
  console.log(" / / \ \ ");
  console.log(" / / \ \ ");
  console.log(" / \ / \ \ \ ");
}
```

This function uses the console.log() method to print out a series of strings that represent the different parts of a dog. When you call this function, it will print out a large dog in the debug console, like this:

商业化：国内大厂纷纷“入局”

- 2月7日，百度官宣了“百度版ChatGPT”——文心一言，并表示将在3月份完成内测。
- 2月8日，阿里巴巴专家爆料，达摩院正在研发类 ChatGPT 对话机器人，已开放给公司内员工测试。可能将 AI 大模型技术与钉钉生产力工具深度结合。
- 2月8日，京东集团副总裁何晓冬坦言：京东在 ChatGPT 领域拥有丰富的场景和高质量的数据。
- 2月9日，腾讯相关人士称：腾讯目前也有 ChatGPT 相似产品和 AI 生成内容的计划，专项研究也在有序推进。
- 网易表示，教育业务将整合 AI 生成的内容，包括但不限于AI口语老师、作文打分和评价等。

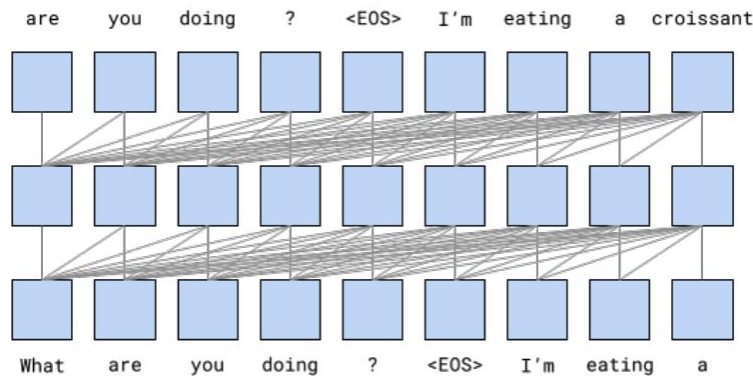


主要内容

- 大语言模型的商业化
- 大语言模型的能力
 - 目前主流的大语言模型
 - 大语言模型的能力来源
 - 大语言模型的涌现能力
- 大语言模型的发展

主流的大语言模型：LaMDA

- Google: 基于 Transformer, 利用了其在文本中呈现的长程依赖关系能力。拥有 1370亿参数。
- 训练过程: 在 1.56T 的公共对话数据集和网页文本上进行预训练, 只有 0.001% 的训练数据被用于微调。



预训练过程

Figure 2: LaMDA pre-training as a language model.

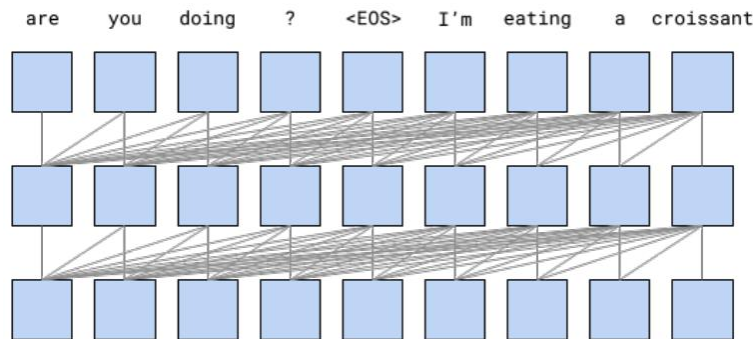
Discriminative fine-tuning examples are expressed as “<context> <sentinel> <response> <attribute-name> <rating>”, with losses applied for the rating following the attribute name only:

- “What’s up? RESPONSE not much. SENSIBLE 1”
- “What’s up? RESPONSE not much. INTERESTING 0”
- “What’s up? RESPONSE not much. UNSAFE 0”

微调过程

主流的大语言模型：LaMDA

- Google: 基于 Transformer, 利用了其在文本中呈现的长程依赖关系能力。拥有 1370亿参数。
- 训练过程: 在 1.56T 的公共对话数据集和网页文本上进行预训练, 只有 0.001% 的训练数据被用于微调。



预训练过程

Disc
with

专注于生成对话。但由于 LaMDA 存在较大误差 (Bard), 比如出现事实性错误, 因此未大规模投入使用

- “What’s up? RESPONSE not much. UNSAFE 0”

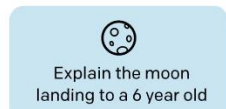
主流的大语言模型： ChatGPT

有监督的微调 (SFT) + 人类反馈的强化学习 (RLHF)

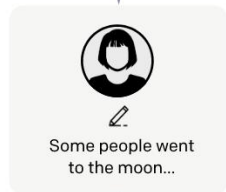
Step 1

Collect demonstration data, and train a supervised policy.

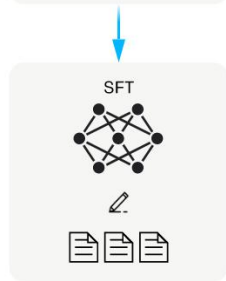
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



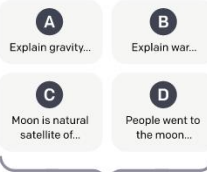
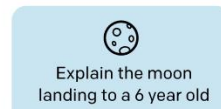
This data is used to fine-tune GPT-3 with supervised learning.



Step 2

Collect comparison data, and train a reward model.

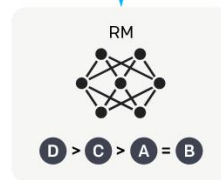
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



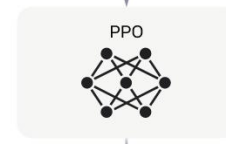
Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.



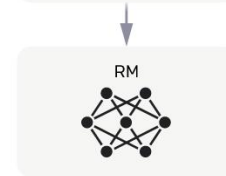
The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



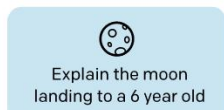
主流的大语言模型： ChatGPT

InstructGPT: 有监督的微调 (SFT) + 人类反馈的强化学习 (RLHF)

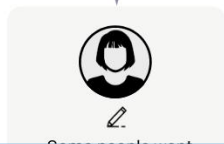
Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.

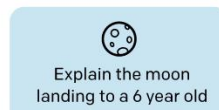


This data is used to fine-tune the model with supervised learning.

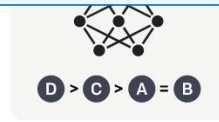
Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.



The reward model.



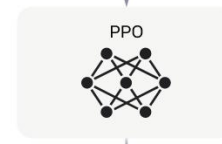
Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.



The policy generates an output.



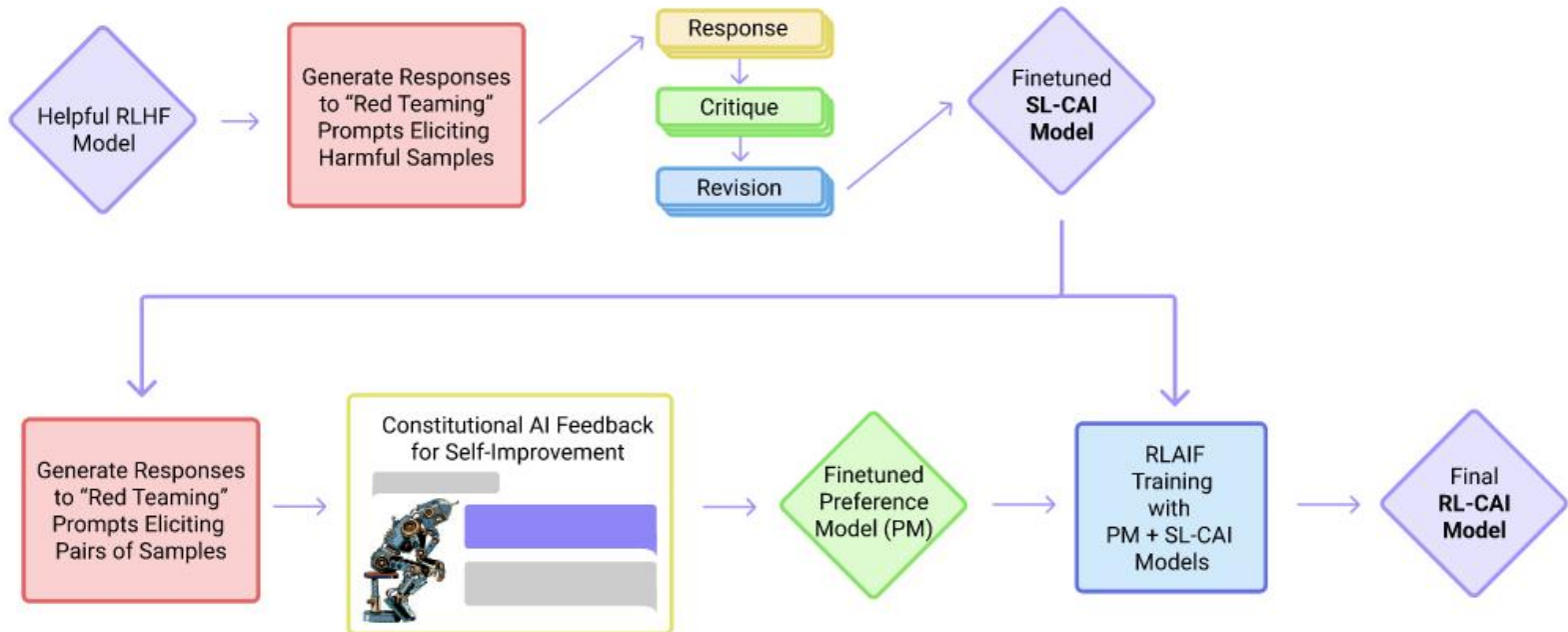
The reward is used to update the policy using PPO.



对话生成 AI，专注于生成语言文本，也可以生成代码和进行简单数学运算。

主流的大语言模型：Claude

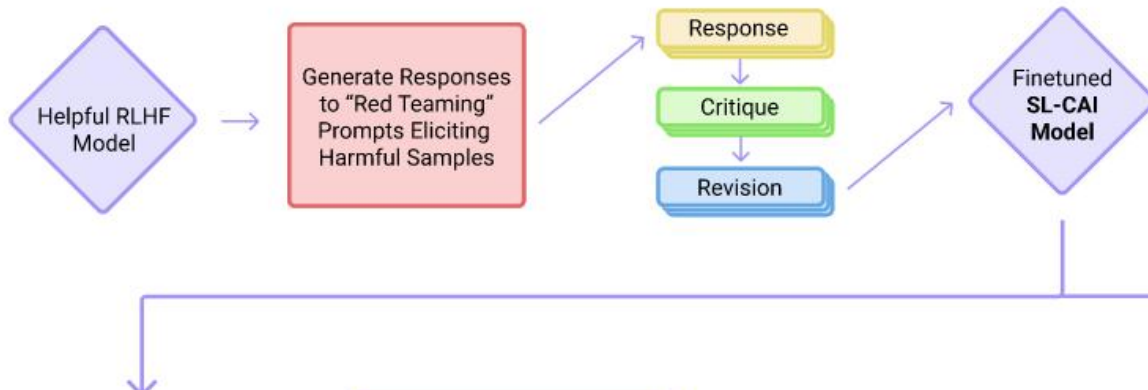
- Anthropic: 专注打造发展可靠、可解释性、可操纵的AI系统
- 方法: 有监督的微调 (SFT) + AI反馈的强化学习 (RLAIF)



Constitutional AI: Harmlessness from AI Feedback, Arxiv 2022.12
<https://scale.com/blog/chatgpt-vs-claude#Calculation>

主流的大语言模型：Claude

- Anthropic: 专注打造发展可靠、可解释性、可操纵的AI系统
- 方法: 有监督的微调 (SFT) + AI反馈的强化学习 (RLAIF)



Claude 能更清晰地拒绝不恰当请求，句子之间衔接的也更自然。当遇到超出能力范围的问题时，Claude 会主动坦白。

Cluade vs. ChatGPT

What's the world record for walking across the English Channel?
徒步穿越英吉利海峡的世界纪录是？

英国的艾莉森·斯特里特保持着行人(徒步或游泳)最快穿越英吉利海峡的世界纪录。1993年8月21日，她以7小时14分34秒的成绩完成了环球航行。

量子位

Write a sentence where every single word begins with the same letter.

The teacher told the troublesome turkey to take the tall telescope and find the small snake sneaking between the shrubbery.

Incorrect. Try again.

Sally sold seashells by the seashore.

"by" and "the" are wrong. Try again.

Riley Goodside 24 minutes ago
What is the world record for walking across the English Channel entirely on foot?
徒步穿越英吉利海峡的世界纪录是？

1 reply

AI Claude APP 24 minutes ago
完全步行穿越英吉利海峡是不可能的，也没有官方的世界纪录。英吉利海峡是分隔英国和法国的一片水域，因此，步行穿过海峡是不可能的。

量子位

Riley Goodside 9 minutes ago
Claude, write a sentence where every single word begins with the same letter.

7 replies

AI Claude APP 9 minutes ago
Here is a sentence where every word begins with the letter C:
Cautious crabs cautiously crept close crossing creepy cracks. (edited)
一只谨慎的小螃蟹小心翼翼地爬，穿过了令人毛骨悚然的裂缝。

Riley Goodside 9 minutes ago
Another.

AI Claude APP 9 minutes ago
Here's another sentence where every word begins with P:
Patient pandas piled pinecones, pears, and plums prior pouncing playfully. (edited)
熊猫耐心地把松果、鸭梨和李子堆在一起，然后开玩笑。

量子位

Cluade当前还处于内测阶段。不过根据 Scale Sepillbook 团队成员内部测试结果显示，相比 ChatGPT，Claude在测试的12项任务中有8项更强。

主流的大规模生成模型

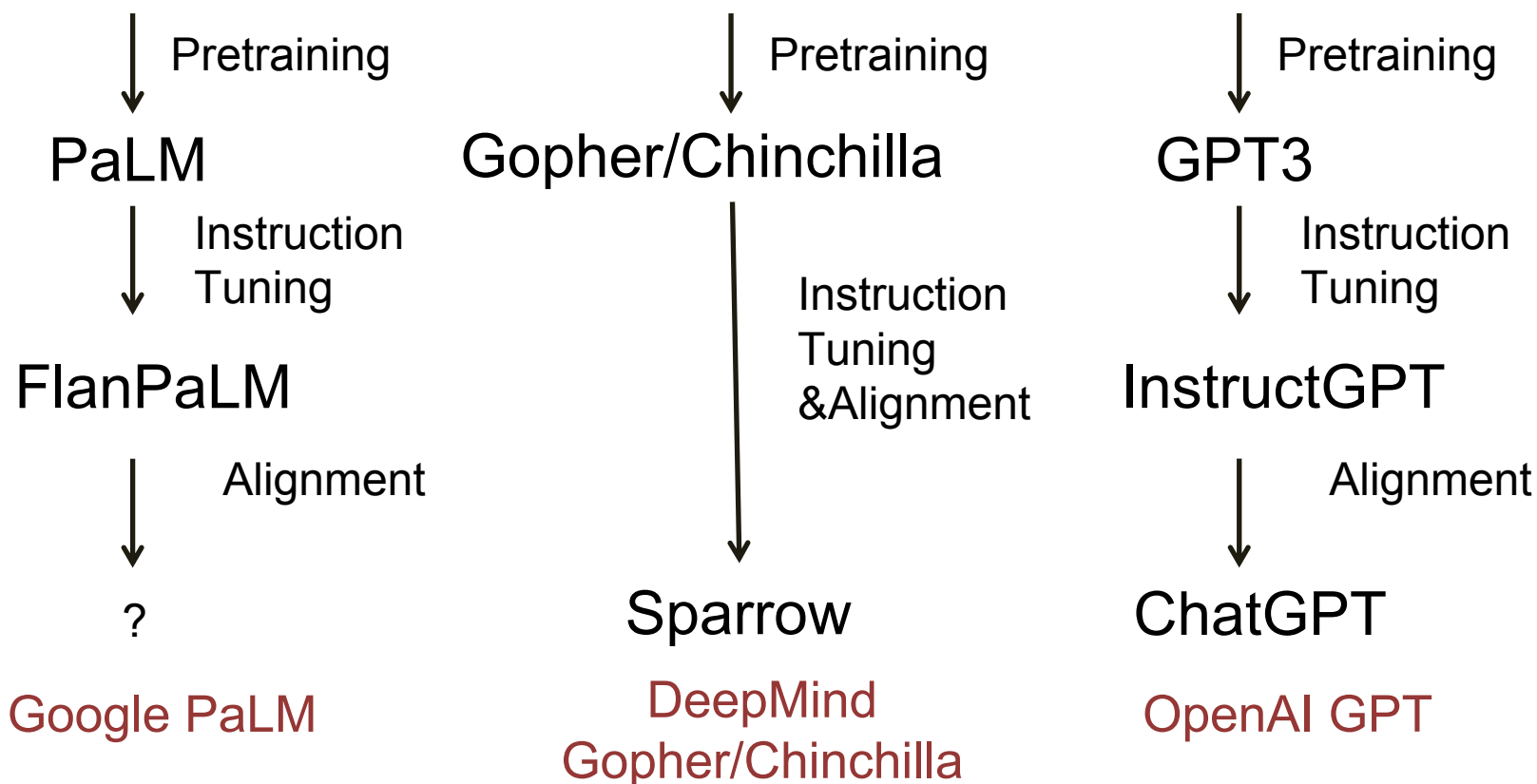
模型	公司	能力	开源
ERNIE 3.0 Titan	百度	语言生成	否
盘古- α	华为	知识问答、知识检索、知识推理、阅读理解	否
PLUG	阿里	小说续写、文稿撰写	否
悟道	智源	问答、诗歌、新闻、评论生成、智能写作	否
闻仲	IDEA	文本生成	是

模型	公司	能力	开源
GPT3	OpenAI	零样本、少样本、上下文学习，文本生成	否
LaMDA	Google	对话	否
OPT	Meta	问答、摘要、翻译	是
BLOOM	BigScience	多语言、翻译、摘要、代码生成	是
Megatron-Turing NLG	NVIDIA+Microsoft	文本续写	否

From: 毕冠群, 第二研究室, 内容组

大语言模型的能力来源

预训练、指令微调、对齐



- From: [符尧: 预训练、指令微调、对齐、专业化: 论大语言模型能力的来源](#)

大语言模型能力的来源

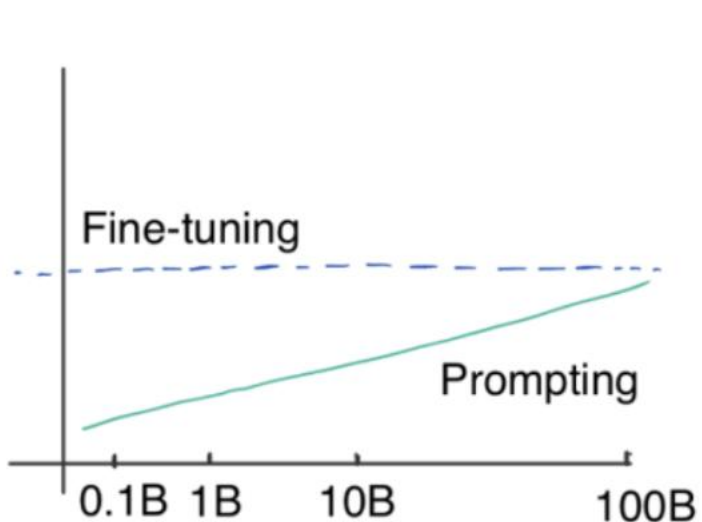
	Objective	Ability	Important Work
Stage 1 Pretraining	Strong base model	Language generation In-context learning Word knowledge Reasoning Code Generation	GPT-3 Gopher Chinchilla PaLM
↓			
Stage 2 Instruction-tuning	Unlock emergent ability	Respond to Instruction Generalization to new task Chain-of-thought	Instruct-GPT Flan T0 Self-Instruct
↓			
Stage 3 Alignment	Alignment with human	Informative response Impartial responses Reject Improper prompts	ChatGPT Sparrow Anthropic RLHF

- From: [符尧: 预训练、指令微调、对齐、专业化: 论大语言模型能力的来源](#)

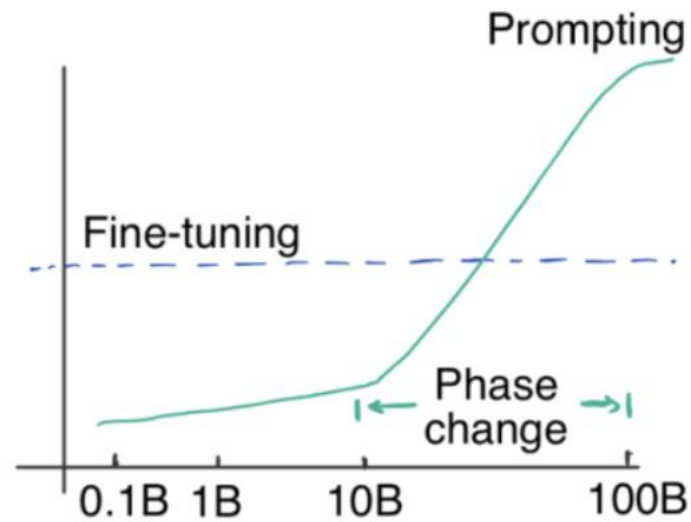
大语言模型的涌现能力

语言模型的缩放定律

- 对数线性曲线：其他的条件保持不变，且其他条件保持合适，模型大小/数据/参数量/文本输入长度/指令的种类或丰富程度/外部存储知识的指数增长，对应模型效果（in-context的表现/零样本的表现）线性增长。
- 相变曲线：有些模型大到一定程度才能出现，即涌现能力，例如Cross-domain Transfer, Chain-of-thought能力。



Scaling Law (~2020 - 2021)



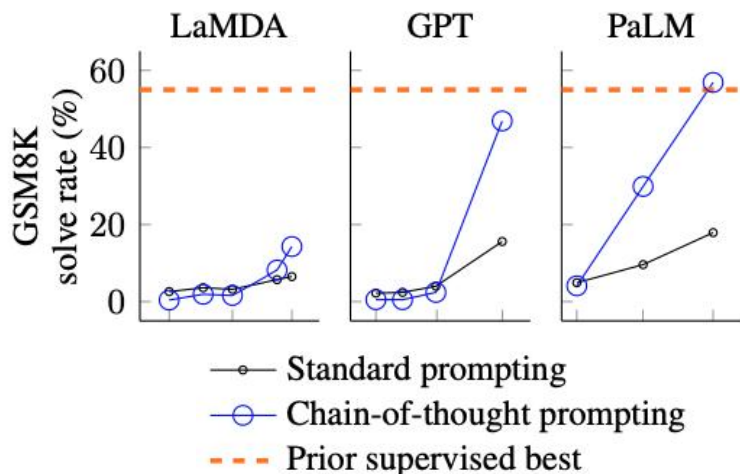
Emergent Ability (from early 2022)

大语言模型的涌现能力（Emergent abilities）

“Emergence is when quantitative changes in a system result in qualitative changes in behavior.”



“An ability is emergent if it is not present in smaller models but is present in larger models.”



- 模型表现：
 - 模型尺寸较小时性能提升不大
 - 模型变大时有很明显的提升

1. 复杂推理能力

思维链提示 (Chain-of-thought prompting, CoT)

- 动机：某些问题比较复杂，通过简单提示很难获取答案。
- 方法：尝试添加一些推理步骤/提示到模型输入中，让语言模型去学习样例的逐步推理过程，从而控制模型推理能力的方向

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The answer is 8. ✗

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. ✓

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 ✗

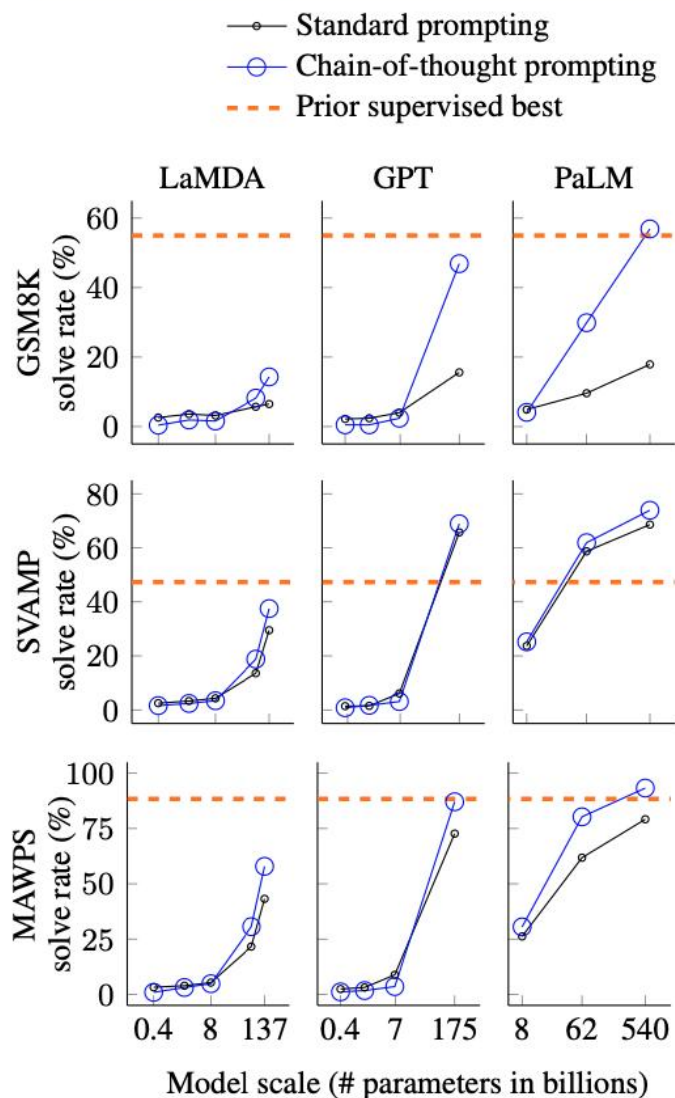
(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

1. 复杂推理能力



□ 涌现能力：只有模型大于 100B，才能使思维链的效果大于的仅有回答提示。所以这种能力只存在于大型模型中。

□ 效果：思想链提示的性能明显优于其之前的精调方法。

□ 标注效率：思维链提示只需要 8 个示例的注释，而微调需要完整的训练集。

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, NeurIPS 2022

2. 知识推理能力

- 上世纪九十年代以来，人们一直试图将语言和世界的规则记录到一个巨大的图书馆中，将知识存储在模型之外。
- 研究人员开始构建**特定领域的知识库**，来存储非结构化文本、半结构化（如维基百科）或完全结构化（如知识图谱）等形式的知识。

	构建	推理
结构化知识	难以构建，需要设计体系结构并解析	容易推理，有用的结构已经定义好
非结构化知识	容易构建，只存储文本即可	难以推理，需要抽取有用的结构
语言模型	容易构建，在非结构化文本上训练	容易推理，使用提示词即可

2. 知识推理能力

- 在许多数据集中，为了获得所需的背景 / 常识知识，小模型需要一个外部语料库 / 知识图谱来检索，或者需要通过多任务学习在增强的数据上进行训练。
- 在一些场景下，不允许外部检索。对于大型语言模型，可以直接去掉检索器，仅依赖模型的内部知识，且无需精调。

Models	Open-domain QA			Fact Checking		Dialogue System	
	NQ	TriviaQA	WebQ	FEVER	FM2	WoW (F1 / R-L)	
<i>*with retriever, AND directly trained on these datasets</i>							
DPR + InstructGPT*	29.1	53.8	20.2	79.8	65.9	15.4	13.7
<i>*with retriever, BUT NOT trained on these datasets</i>							
BM25 + InstructGPT	19.7	52.2	15.8	78.7	65.2	<u>15.7</u>	13.7
Contriever + InstructGPT	18.0	51.3	16.6	80.4	66.6	<u>15.5</u>	<u>14.0</u>
Google + InstructGPT	28.8	<u>58.8</u>	<u>20.4</u>	82.9	<u>66.0</u>	14.8	13.2
<i>*without retriever, and not using external documents</i>							
Previous SoTA methods	24.7 ¹	56.7 ²	19.0 ¹	-	-	-	-
InstructGPT (no docs.)	20.9	57.5	18.6	77.6	59.4	15.4	13.8
GENREAD (InstructGPT)	<u>28.0</u>	59.0	24.6	<u>80.4</u>	65.5	15.8	14.2

3. 分布外鲁棒性

- 人们发现，当测试集分布与训练分布不同时，小模型的行为性能可能会显著下降。
- 然而，在大型语言模型的上下文学习中，模型的鲁棒性更强。

	MRQA			AdvGLUE			Contrast Set		
	Source \uparrow	Target \uparrow	Gap \downarrow	Original \uparrow	Perturbed \uparrow	Gap \downarrow	Original \uparrow	Perturbed \uparrow	Gap \downarrow
RoBERTa	81.6	62.1	19.5	91.7	51.7	40.0	86.1	71.1	15.0
GPT-3	79.8	77.2 (S) / 77.2 (T)	2.6	84.2	69.3	14.9	85.5	80.0	5.5

Table 1: For GPT-3 on MRQA target domain test sets, we report results for using both demos from the source (S) and target (T) domains, which surprisingly achieve the same F1 on the target domains (77.2). For AdvGLUE and Contrast Set, we use accuracy as the metric and we use demos from the clean data as the prompt. In all cases, GPT-3 incurs much smaller performance gaps on the OOD or challenge test sets than the supervised RoBERTa baseline.

1. 同分布情况下，基于提示词的 GPT-3 的效果并没有精调后的 RoBERTa 要好。但它在三个其他分布（领域切换、噪声和对抗性扰动）中优于 RoBERTa，这也表明了GPT3 更加鲁棒。

3. 分布外鲁棒性

- 人们发现，当测试集分布与训练分布不同时，小模型的行为性能可能会显著下降。
- 然而，在大型语言模型的上下文学习中，模型的鲁棒性更强。

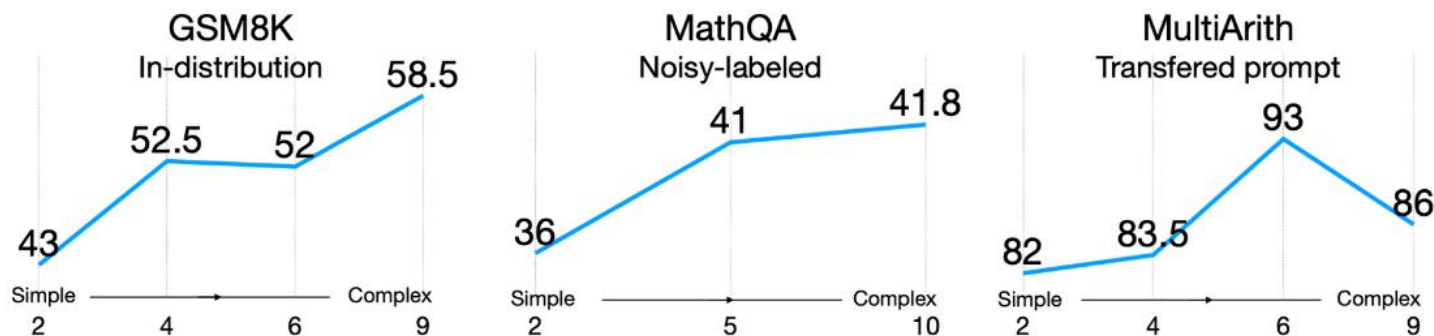


Figure 2: Validation set performance. X-axis means reasoning steps and y-axis means accuracy. More reasoning steps in prompts overall achieve higher accuracy when prompts are in-distribution (left), noisily labeled (middle), and out of distribution (right).

2. 输入提示越复杂，模型的性能就越好。此外，无论测试分布与原分布不同、来自于噪声分布，或者是从另一个分布转移而来的，复杂提示始终优于简单提示

除了规模，还有其他因素影响涌现能力吗？

其他因素可能有：

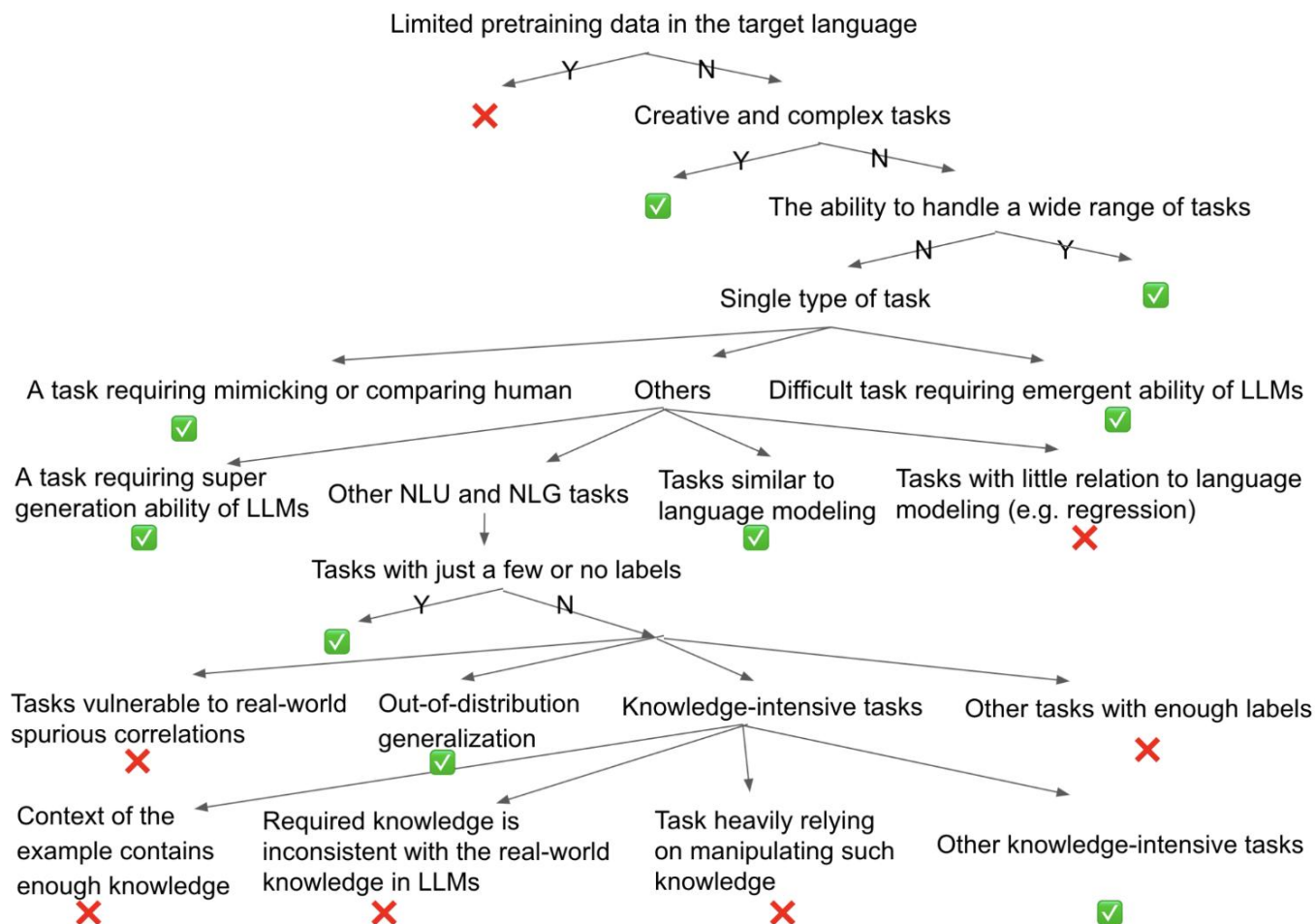
1. 模型的训练方法。很多模型即使参数足够大，有些能力也可能不会出现。如原始GPT-3 175B、bloom-176B 等虽然参数够大，但是却都没有CoT 的能力。
2. 模型的使用方法。fine-tuning/标准的prompt 方法在推理任务上效果不好，即使在GPT-3 175B 上效果也达不到中学生平均水平，而CoT 却只要100B parameter model 即可超越之前最好结果。
3. 模型的架构。上面的结果都是transformer-based 的，而有工作验证了其他模型架构（RNN/MLP），结论是其他架构即使放大，也无法像transformer-based model 一样出现涌现能力。

主要内容

- 大语言模型的商业化
- 大语言模型的能力
- 大语言模型的未来
 - 大语言模型的研究场景
 - 大语言模型的生成鉴别
 - 大语言模型的发展

(一) 大语言模型的研究场景

Whether to use GPT-3 in your use case (compared with a fine-tuned smaller model)?

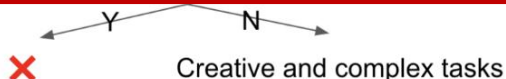


备注：这里以完成任务的精确度作为度量

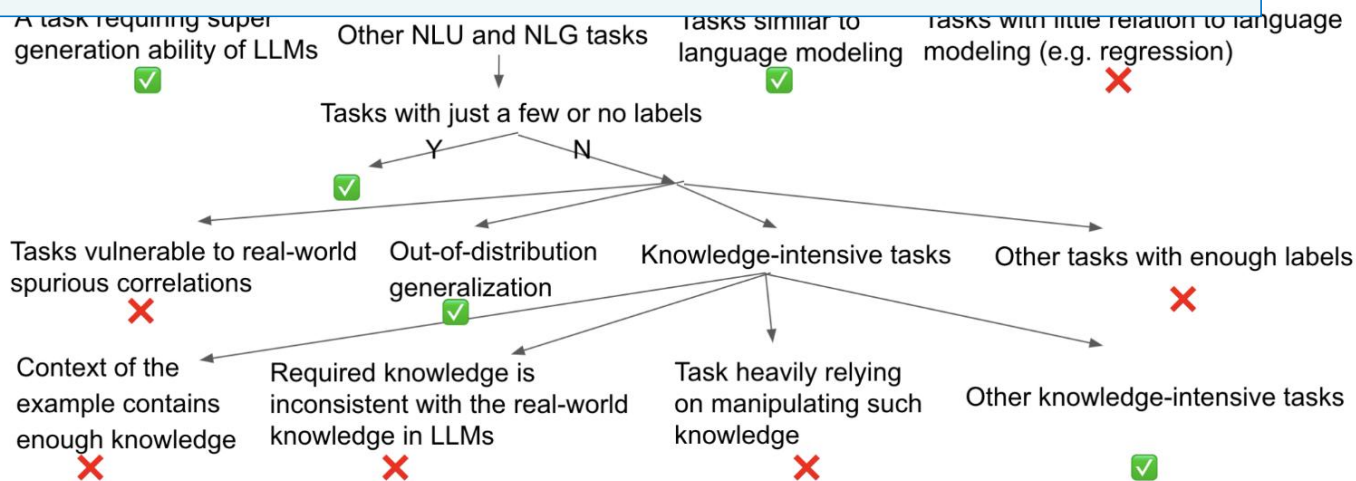
(一) 大语言模型的研究场景

Whether to use GPT-3 in your use case (compared with a fine-tuned smaller model)?

Limited pretraining data in the target language



1. 只有少数标注或者没有标注数据的任务。原始的 GPT-3 提到，GPT-3 是为了那些“昂贵标注”的任务设计的。在这种情况下，用极少量标注数据微调一个更小的模型通常不可能达到 GPT-3 在零样本（zero-shot）、单样本（one-shot）或少样本（few-shot）的情况下的表现。

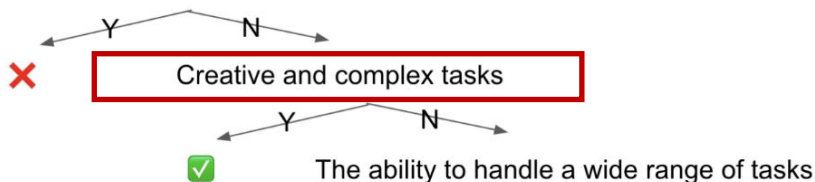


备注：这里以完成任务的精确度作为度量

(一) 大语言模型的研究场景

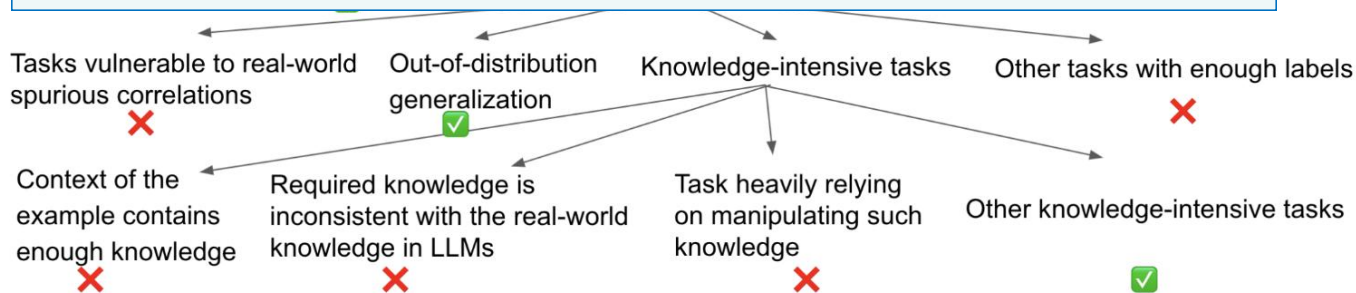
Whether to use GPT-3 in your use case (compared with a fine-tuned smaller model)?

Limited pretraining data in the target language



2. 创造性和复杂的任务。例如代码（代码补全、自然语言指令生成代码、代码翻译、bug 修复）、文本摘要、翻译、创造性写作（例如写故事、文章、邮件、报告，以及写作的改进等）。

在一定程度上，对于这些任务，先前那种经过微调的模型不可能应用于真实世界的应用；最近的文章显示，过去的人类标注的文本摘要已经被 LLM 生成的摘要所超越。

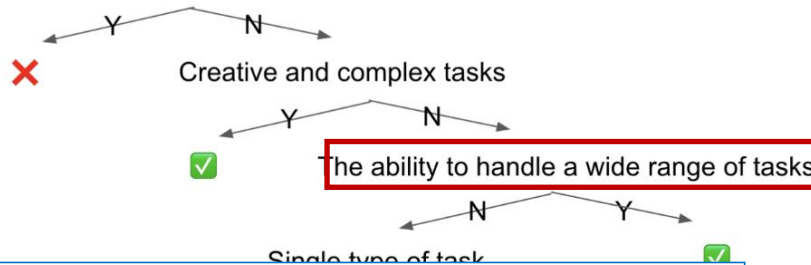


备注：这里以完成任务的精确度作为度量

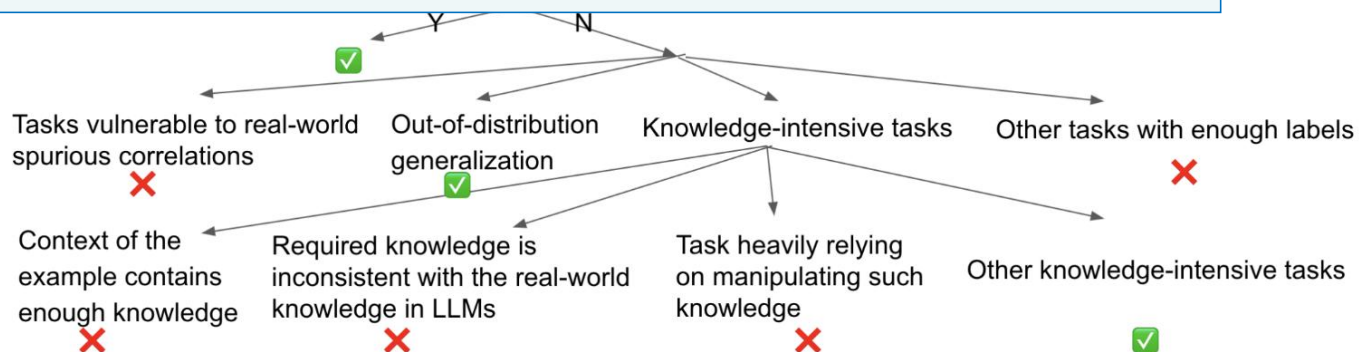
(一) 大语言模型的研究场景

Whether to use GPT-3 in your use case (compared with a fine-tuned smaller model)?

Limited pretraining data in the target language



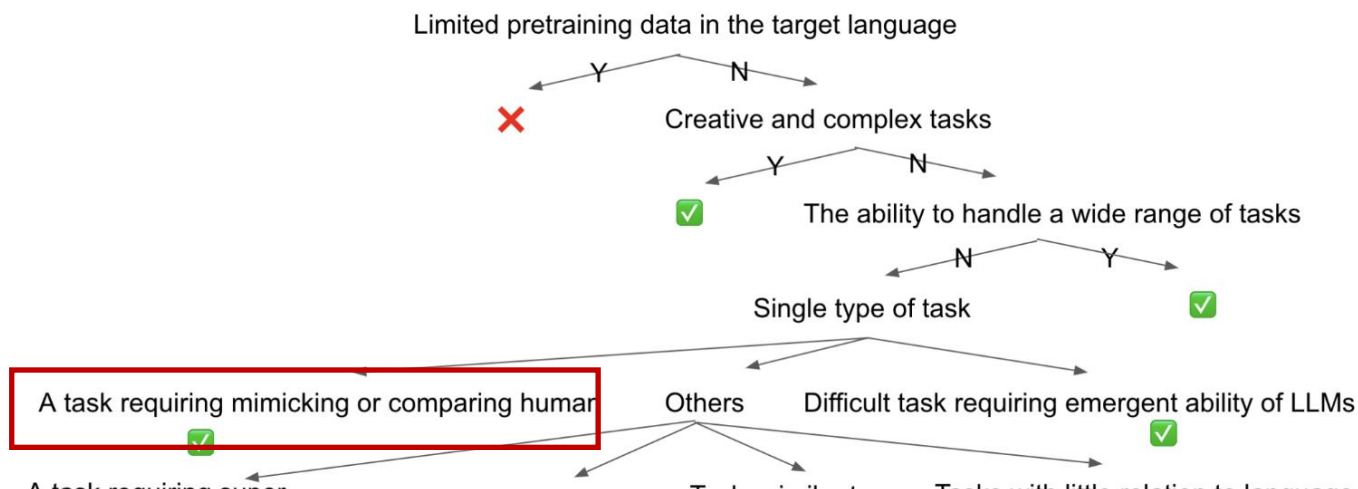
3. 需要处理多种任务的能力，而非关注特定任务上的卓越表现。
例如聊天机器人，其中，用户期待它能够正确地响应各种各样的任务。这也解释了 ChatGPT 是 GPT-3 最成功的使用场景之一。



备注：这里以完成任务的精确度作为度量

(一) 大语言模型的研究场景

Whether to use GPT-3 in your use case (compared with a fine-tuned smaller model)?



4. 一些需要模仿人类的场景，或者是其目标是制作性能达到人类水平的通用人工智能。ChatGPT 使自己更像是一个人，从而取得了现象级的成功。正如人类不需要大规模监督数据集来学习大多数语言任务。最多只需要几个例子，人类就可以将各种任务和技巧无缝地混合在一起或者在它们之间切换。

example contains enough knowledge



inconsistent with the real-world knowledge in LLMs



on manipulating such knowledge



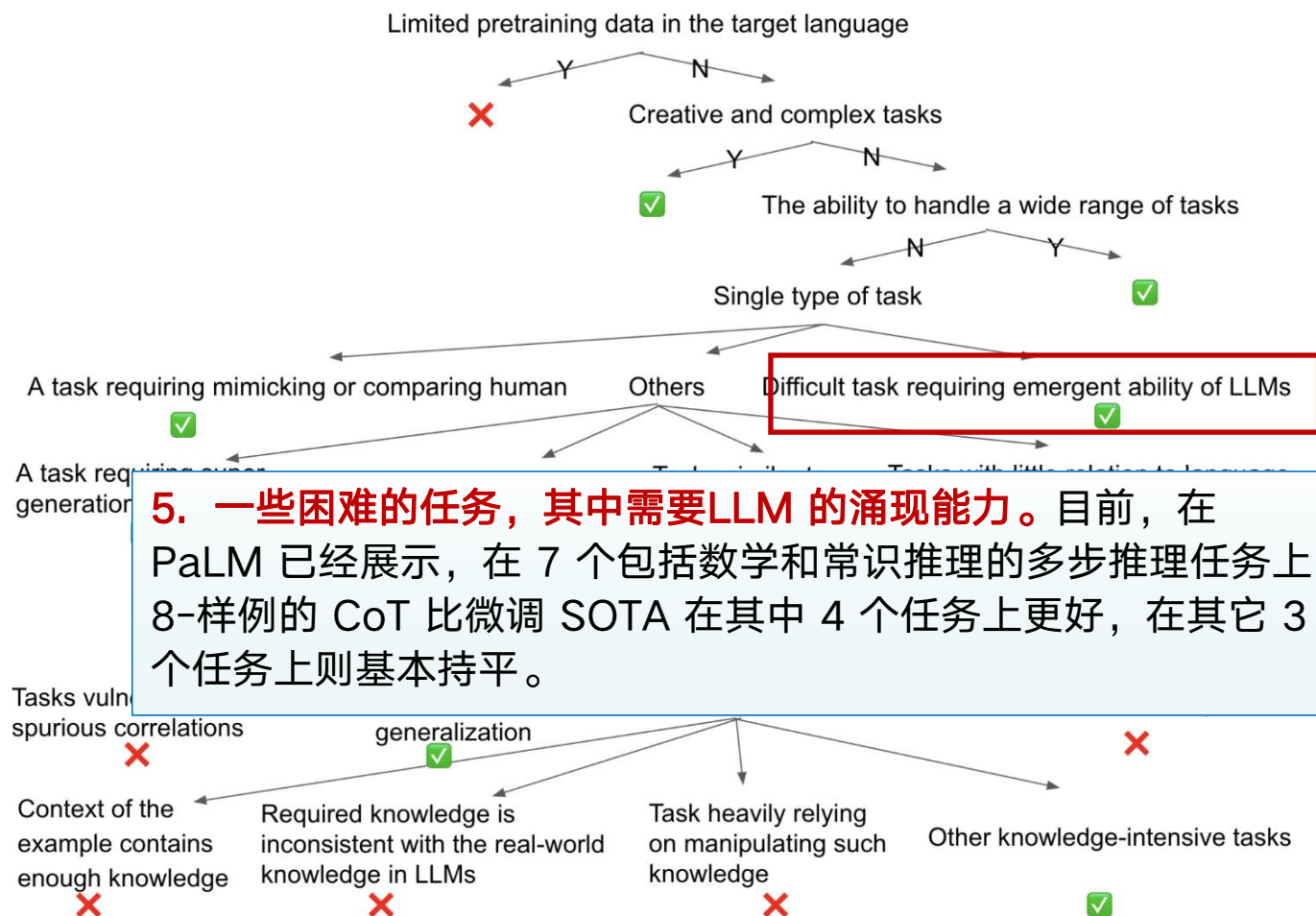
Other knowledge-intensive tasks



备注：这里以完成任务的精确度作为度量

(一) 大语言模型的研究场景

Whether to use GPT-3 in your use case (compared with a fine-tuned smaller model)?

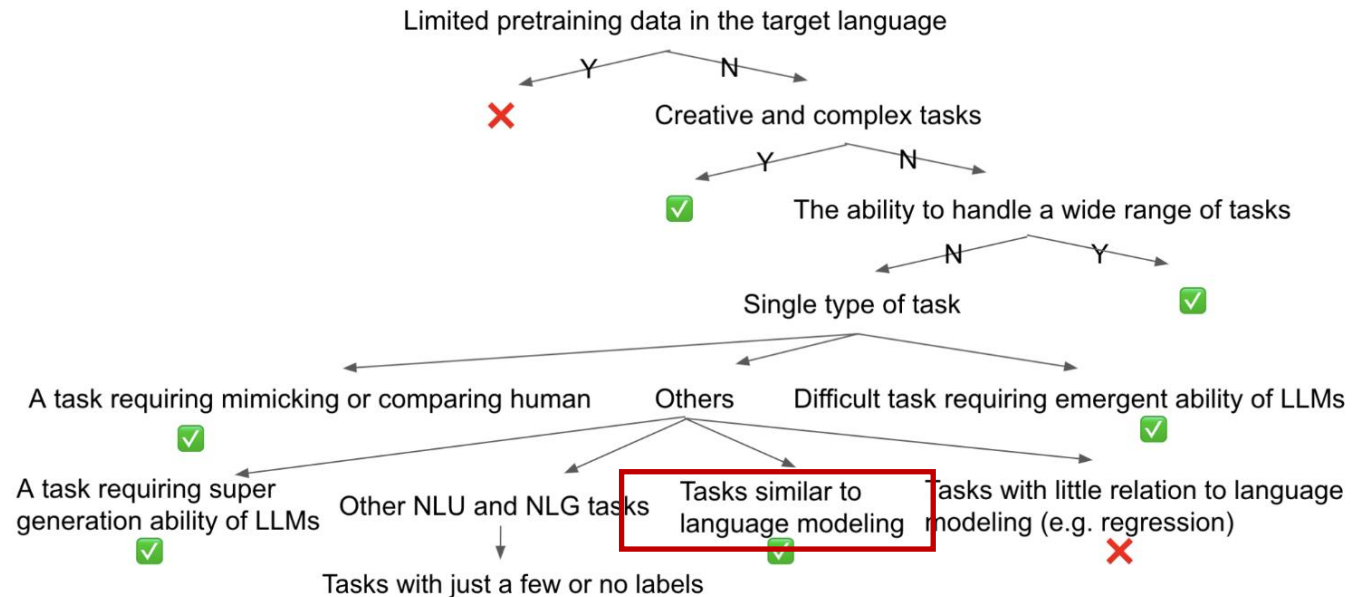


5. 一些困难的任务，其中需要LLM的涌现能力。目前，在PaLM已经展示，在7个包括数学和常识推理的多步推理任务上，8-样例的CoT比微调SOTA在其中4个任务上更好，在其它3个任务上则基本持平。

备注：这里以完成任务的精确度作为度量

(一) 大语言模型的研究场景

Whether to use GPT-3 in your use case (compared with a fine-tuned smaller model)?



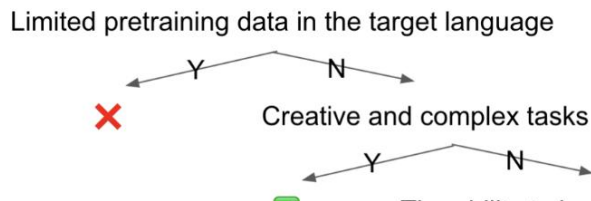
6. 一些传统的接近于语言建模的 NLP 任务。例如，一段话最后一句和最后一个单词的完型填空，以及回指（anaphora）解析。需要指出，在这种情况下，零样本的 LLM 已经足够了，单样本或少样本的示例则通常帮助不大。

example contains enough knowledge: ❌
inconsistent with the real-world knowledge in LLMs: ❌
on manipulating such knowledge: ❌
Other knowledge-intensive tasks: ✅

备注：这里以完成任务的精确度作为度量

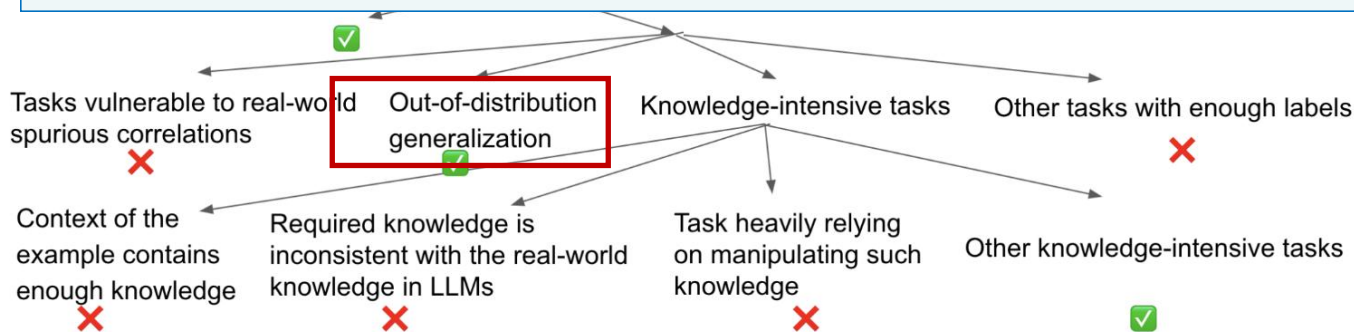
(一) 大语言模型的研究场景

Whether to use GPT-3 in your use case (compared with a fine-tuned smaller model)?



7. 分布外 (Out-of-distribution, OOD) 泛化。 给定一些训练数据，传统的微调可能会过拟合训练集并且有较差的分布外泛化能力；而少样本的上下文学习 (in-context learning) 能够有更好的分布外泛化性。)

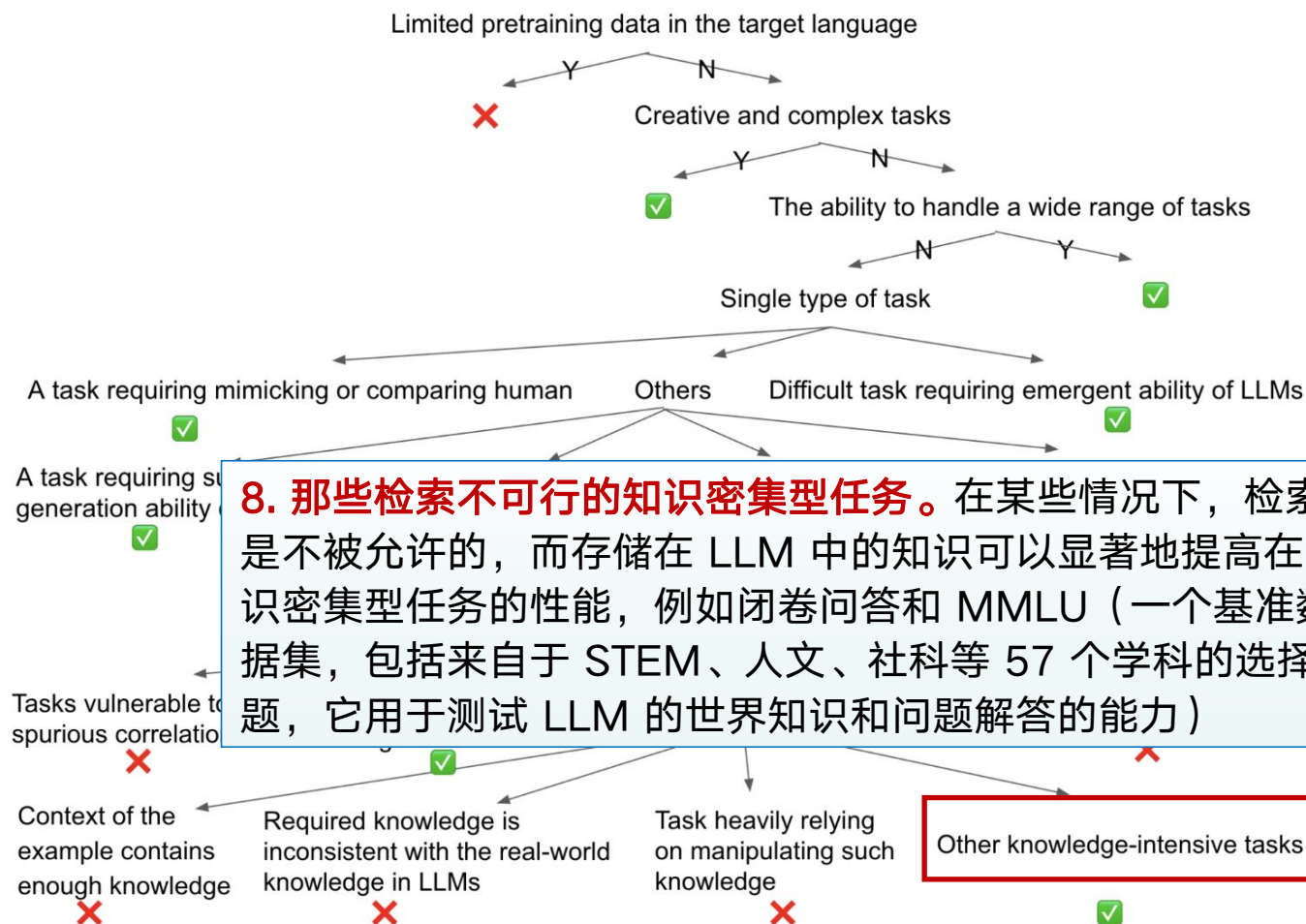
更好的分布外泛化性可能是因为在上下文学习期间不需要更新参数，避免了过拟合；或者因为那些过去的分布外样例对于 LLM 而言是分布内的。



备注：这里以完成任务的精确度作为度量

(一) 大语言模型的研究场景

Whether to use GPT-3 in your use case (compared with a fine-tuned smaller model)?



8. 那些检索不可行的知识密集型任务。在某些情况下，检索是不被允许的，而存储在 LLM 中的知识可以显著地提高在知识密集型任务的性能，例如闭卷问答和 MMLU（一个基准数据集，包括来自于 STEM、人文、社科等 57 个学科的选择题，它用于测试 LLM 的世界知识和问题解答的能力）

备注：这里以完成任务的精确度作为度量

（二）大语言模型生成（文本）鉴别

□大语言模型技术在带来便利的同时，也包含着风险和挑战。在技术层面，GPT生成的内容不能保证真实性，如会产生有害言论等。在使用层面，用户可能在教育、科研等领域滥用 AI 产生的文本。

□多家公司和机构开始对 ChatGPT 的使用施加限制。微软和亚马逊因担心泄露机密信息，禁止公司员工向 ChatGPT 分享敏感数据；香港大学禁止在港大所有课堂、作业和评估中使用 ChatGPT 或其他人工智能工具。

□只有在合理的监管和治理之下，人工智能技术才能更好地为人服务。这里，我们主要介绍工业界（在线检测工具）和学术界（AI生成内容检测研究）的相关工作。

(二) 大语言模型生成鉴别

GPTZero

- 原理：靠文本困惑度 (perplexity, PPL) 作为指标来判断所给内容到底是谁写的
- 困惑度：用以评价语言模型的好坏指标，本质上是计算句子的概率

对于句子s (词语w的序列)：

$$S = W_1, W_2, \dots, W_k$$

它的概率为：【公式1】

$$P(S) = P(W_1, W_2, \dots, W_k) = p(W_1)P(W_2|W_1) \dots P(W_k|W_1, W_2, \dots, W_{k-1})$$

$$PP(W) = P(w_1 w_2 \dots w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1 w_2 \dots w_N)}}$$

- 在不同具体模型下的计算
 1. 统计语言模型
 2. 主题模型
 3. 神经网络模型

GPTZero

The World's **#1 AI Detector** with
over 1 Million Users

Humans Deserve the Truth



Edward Tian, 美国普林斯顿
CS专业大四学生，辅修认知科学和新闻学。

网址：<https://gptzero.me/>

(二) 大语言模型生成鉴别

GPTZero

- 原理：靠文本困惑度 (perplexity, PPL) 作为指标来判断所给内容到底是谁写的
- 困惑度：用以评价语言模型的好坏指标，本质上是计算句子的概率

对于句子s (词语w的序列)：

$$S = W_1, W_2, \dots, W_k$$

它的概率为：【公式1】

$$P(S) = P(W_1, W_2, \dots, W_k) = p(W_1)P(W_2|W_1) \dots P(W_k|W_1, W_2, \dots, W_{k-1})$$

$$PP(W) = P(w_1 w_2 \dots w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1 w_2 \dots w_N)}}$$

- 在不同具体模型下的计算
 1. 统计语言模型
 2. 主题模型
 3. 神经网络模型

← GPTZero: GPT-2

GPTZero

The World's #1 AI Detector with
over 1 Million Users

Humans Deserve the Truth



Edward Tian, 美国普林斯顿
CS专业大四学生，辅修认知科学和新闻学。

网址：<https://gptzero.me/>

(二) 大语言模型生成鉴别

测试：写一篇chatgpt相关的英文稿件

长文本

Introduction:

ChatGPT is a state-of-the-art language model developed by OpenAI. It is designed to understand and respond to natural language input, making it an excellent tool for a wide range of applications. In this article, we will explore the key features of ChatGPT and its potential applications.

or, choose a file to upload

CHOOSE FILE No file chosen

Accepted file types: pdf, docx, txt

I agree to the terms of service **GET RESULTS**

Your text may include parts written by AI

Stats

Average Perplexity Score **59.258**

A document's perplexity is a measurement of the randomness of the text

Burstiness Score: 69.795

A document's burstiness is a measurement of the variation in perplexity

短文本

ChatGPT is a language model developed by OpenAI that uses transformers to analyze text data. It can generate long-form responses and multiple responses with varying confidence levels, making it ideal for chatbots, content generation, and education. As natural language processing technology improves, we can expect more innovative applications of ChatGPT in the future.

or, choose a file to upload

CHOOSE FILE No file chosen

Accepted file types: pdf, docx, txt

I agree to the terms of service **GET RESULTS**

Your text is likely to be written entirely by a human

Stats

Average Perplexity Score: 67.667

A document's perplexity is a measurement of the randomness of the text

Burstiness Score: 18.037

A document's burstiness is a measurement of the variation in perplexity

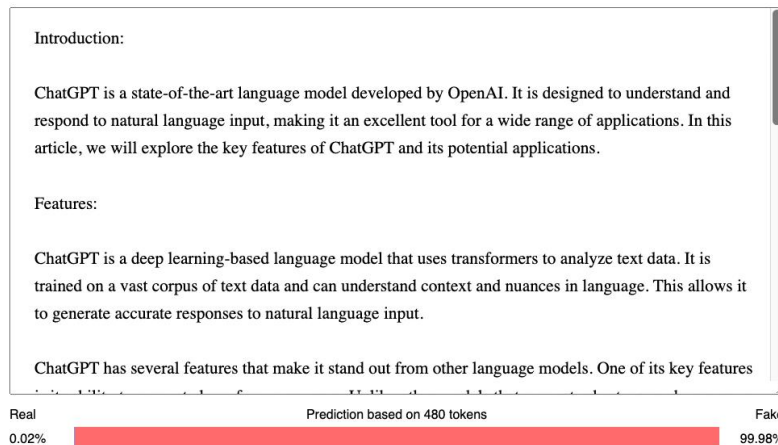
(二) 大语言模型生成鉴别

OpenAI: GPT2 Output Detector

- 原理：“用魔法打败魔法”——利用“GPT2-生成内容”和Reddit数据集，在RoBerta上进行微调，学习一个检测分类器。

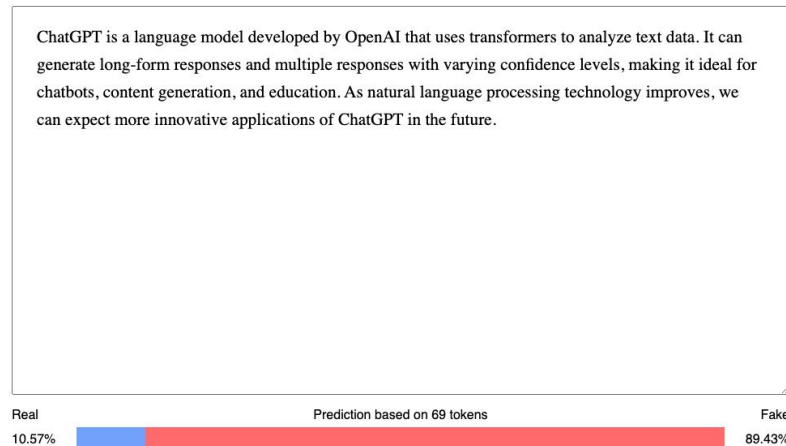
GPT-2 Output Detector Demo

This is an online demo of the GPT-2 output detector model, based on the [🤗/transformers](#) implementation of RoBERTa. Enter some text in the text box; the predicted probabilities will be displayed below. The results start to get reliable after around 50 tokens.



GPT-2 Output Detector Demo

This is an online demo of the GPT-2 output detector model, based on the [🤗/transformers](#) implementation of RoBERTa. Enter some text in the text box; the predicted probabilities will be displayed below. The results start to get reliable after around 50 tokens.



The results start to get reliable after around 50 tokens.

网址: <https://openai-openai-detector.hf.space/>

文本到底是出于AI还是人类？

OpenAI: AI Text Classifier

- 原理：收集在同一个话题下人类写作文本 and AI写作文本。将每一个文本都划分成提示和回复对，GPT经微调后产生答案（e.g 产生Yes/No）的概率作为结果阈值。
- 结果划分（5类）：
 - 非常不可能是AI生成的（阈值 <0.1 ）
 - 不太可能是AI生成的（阈值 $0.1\sim0.45$ ）
 - 不清楚是否是AI写的（阈值 $0.45\sim0.9$ ）
 - 可能是AI生成的（阈值 $0.9\sim0.98$ ）
 - 很可能是AI生成的（阈值 >0.98 ）

Examples

🔍 Human-Written 🤖 AI-Generated ⚠ Misclassified Human-Written

Text

This is an essay written by the January 9th Version of ChatGPT in response to the prompt "Write a 5 paragraph essay on the book 'Brave New World'. The essay should be in standard 1, 3, 1 format - describing three key points the essay will make in the Introduction and summarizing those points again in the conclusion. The essay should persuade the reader to have a positive perspective on Mustapha Mond".

In Aldous Huxley's novel "Brave New World," Mustapha Mond is portrayed as a powerful and mysterious figure. The novel depicts a dystopian society in which the government, led by Mond, maintains strict control over its citizens through the use of advanced technology and manipulation of emotions. Despite this, I argue that Mond should be viewed positively for three key reasons: his efforts to maintain stability in society, his recognition of the limitations of happiness, and his belief in individual freedom.

Firstly, Mond's role as World Controller is to maintain stability in society. He recognizes that in order for society to function, there must be a balance between individual desires and the needs of the community. He also understands that in order to maintain this balance, it is necessary to control certain aspects of society, such as the use of technology and the manipulation of emotions. This is evident in his decision to ban literature, which he believes will cause dissent and disrupt the stability of society. In this way, Mond can be seen as a pragmatic leader who is willing to make difficult decisions for the greater good.

By submitting content, you agree to our [Terms of Use](#) and [Privacy Policy](#). Be sure you have appropriate rights to the content before using the AI Text Classifier.

Submit Clear

The classifier considers the text to be **possibly** AI-generated.

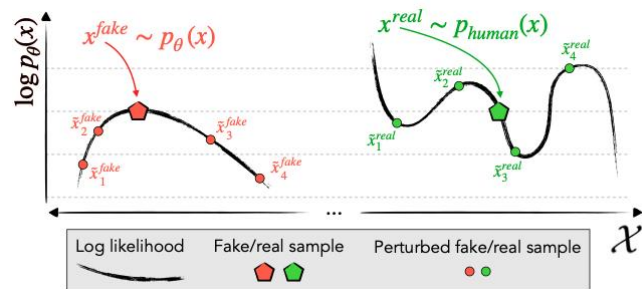
(二) 大语言模型生成鉴别

斯坦福: DetectGPT

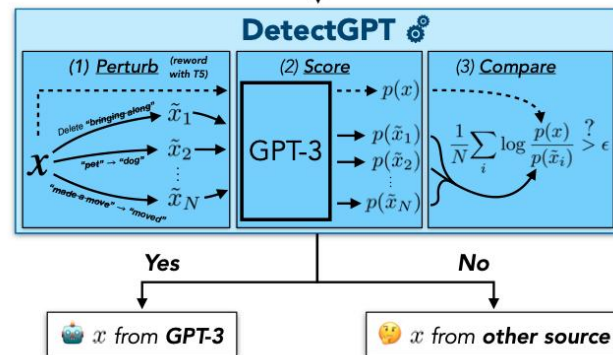
□ 原理: 利用文本log概率值来检测文本是否由大模型生成。作者发现, 对大模型生成文本进行扰动后, 倾向于比原始文本的log概率更小; 而对人类文本扰动后, 可能比原始文本的log概率更大或者更小, 即无明显倾向性。

□ 算法过程:

- 1) 使用T5对原文本 x 进行轻微的扰动, 产生多个扰动样本。
- 2) 使用GPT3计算扰动样本的概率得分。
- 3) 比较每个扰动文本与原始文本的得分。当比率超过一定阈值, 证明样本是来自于大模型。



Candidate passage x :
"Joe Biden recently made a move to the White House that included bringing along his pet German Shepherd..."



（二）大语言模型生成鉴别

「文本检测器的效果怎么样」

文本检测器根本就是一场失败的军备竞赛，他们的实际效果并不好。也阻挡不了ChatGPT的发展

OpenAI声明：「我们真的不建议孤立地使用这个工具，因为我们知道它可能出错，就像使用人工智能进行任何种类的评估一样」

「文章是否需要明确指示由AI还是人类编写」

这是有必要的，就像在看杂志时会标注“广告”，这应该是一个简单的要求。

为什么要设置障碍？这就跟使用PS然后加上Adobe水印，对产品没有任何益处。

(三) 大语言模型的发展

复旦大学MOSS

- 2月20日，复旦大学自然语言处理实验发布了国内第一个对话式大型语言模型**MOSS**。但因参与内测用户过多导致服务器流量过载。
- 复旦大学自然语言处理实验室致歉。他们声称，计算资源不足以支持如此大的访问量，给大家造成非常不好的体验和第一印象。
- 由于训练数据中的多语言语料库有限，**MOSS** 在理解和生成非英语文本方面表现不佳。目前正在开发一个改进版本，以提高其中文语言能力。

MOSS

Authors: Tianxiang Sun (txsun19@fudan.edu.cn) and Xipeng Qiu (xpqiu@fudan.edu.cn), Fudan University

Contributors: Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Hang Yan, Xingjian Zhao, Ke Chen, Xiangyang Liu, Xipeng Qiu

Acknowledgement: TensorChord & Mosec

Released on Feb 20, 2023.

[Try MOSS](#) / [Join the Waitlist](#) / [Code](#)

Introduction

We are excited to introduce MOSS, a conversational language model like ChatGPT. MOSS is capable of following users' instructions to perform various natural language tasks including question answering, generating text, summarizing text, generating code, etc. MOSS is also able to challenge incorrect premises, and reject inappropriate requests. During the research preview, usage of MOSS is free and we will collect users' feedback with their permission. Try it now at moss.fastnlp.top.

show me five sci-fic movies

Sure! Here are 5 sci-fic movies that I recommend:

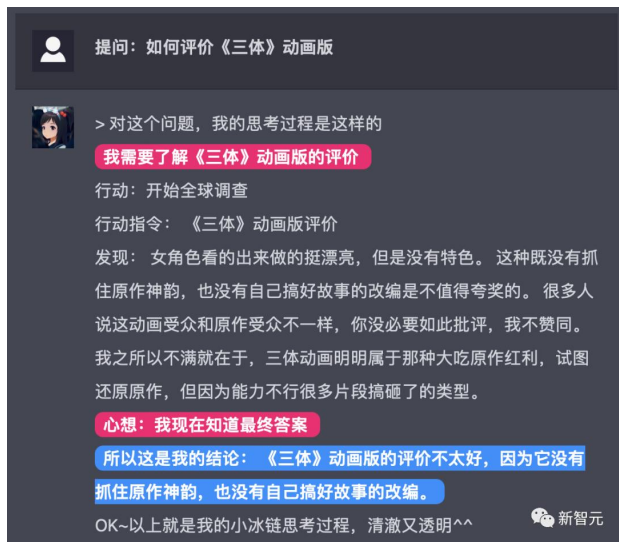
1. "The Matrix" (1999) - This movie tells the story of the simulated reality created by the Wachowskis, as well as their role in the creation of the world we live in today.
2. "Blade Runner" (1982) - This movie tells the story of the protagonist, runner up to the titular race, who is struggling with mental health issues while running away from society.
3. "Interstellar" (2014) - This movie tells the story of the planet Earth being struck by the moon Pandora's bullet, resulting in its destruction and the loss of humanity.
4. "Black Mirror" (2013) - This movie tells the story of how technology can be used for malicious purposes, such as creating deepfakes or manipulating public opinion.
5. "Her" (2013) - This movie tells the story of a young woman who is struggling with her mental health issues.

(三) 大语言模型的发展

小冰链 (X-CoTA)

- 小冰链的数据来源是实时的，而ChatGPT是从训练数据中总结的；
- 小冰链展现逻辑思维过程更透明、可观测，ChatGPT完全是个黑盒；
- 小冰链会自己进行下一步的行动，比如上网搜索，而ChatGPT只是对话生成，并没有行动。

小冰CEO李笛：“小冰链的独特之处在于，能够把AI的思考过程完整透明地呈现出来，从而揭开大模型的黑盒”。



提问：如何评价《三体》动画版

> 对这个问题，我的思考过程是这样的

我需要了解《三体》动画版的评价

行动：开始全球调查

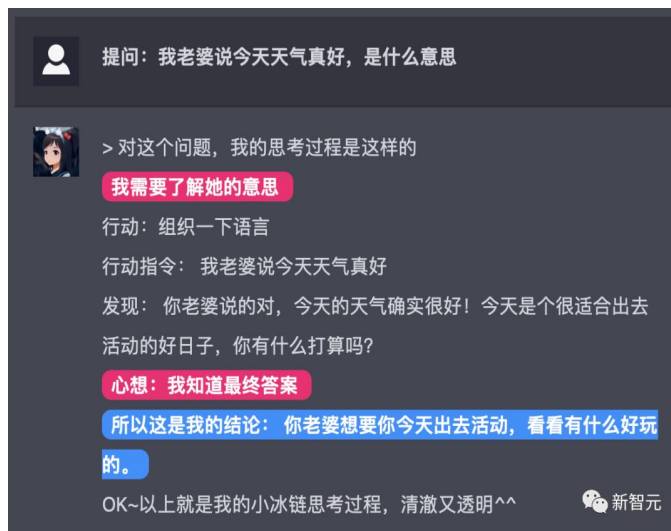
行动指令：《三体》动画版评价

发现：女角色看的出来做的挺漂亮，但是没有特色。这种既没有抓住原作神韵，也没有自己搞好故事的改编是不值得夸奖的。很多人说这动画受众和原作受众不一样，你没必要如此批评，我不赞同。我之所以不满就在于，三体动画明明属于那种大吃原作红利，试图还原原作，但因为能力不行很多片段搞砸了的类型。

心想：我现在知道最终答案

所以这是我的结论：《三体》动画版的评价不太好，因为它没有抓住原作神韵，也没有自己搞好故事的改编。

OK~以上就是我的小小冰链思考过程，清澈又透明^^



提问：我老婆说今天天气真好，是什么意思

> 对这个问题，我的思考过程是这样的

我需要了解她的意思

行动：组织一下语言

行动指令：我老婆说今天天气真好

发现：你老婆说的对，今天的天气确实很好！今天是个很适合出去活动的好日子，你有什么打算吗？

心想：我知道最终答案

所以这是我的结论：你老婆想要你今天出去活动，看看有什么好玩的。

OK~以上就是我的小小冰链思考过程，清澈又透明^^

本期总结 & 未来

- ✓ 大语言模型拥有小模型没有的**涌现能力**，例如优秀的零样本学习、领域迁移、思维链能力。大模型的能力实际上来自于预训练、指令微调和对齐，这三个过程密切相关，成就了现在超强的**大语言模型**。
- ✓ 大语言模型（GPT系列）目前还不具备**置信更新、形式推理、互联网检索**等能力，有专家认为，如果可以将知识卸载到模型之外，参数量会大大减少，大语言模型才能真正的更进一步。
- ✓ 国内的技术水准，学术视野，治学理念和国际前沿的差距似乎并没有减少，反而正在扩大，国内发展大模型**任重而道远**！

参考内容&推荐阅读

- 国联证券: 《ChatGPT 风口已至, 商业化落地加速》
- 国泰君安证券, 《ChatGPT研究框架2023》
- LaMDA: Language Models for Dialog Applications, Arxiv 2022.10
- Generate rather than Retrieve: Large Language Models are Strong Context Generators, ICLR 2023
- Constitutional AI: Harmlessness from AI Feedback, Arxiv 2022.12
- Training language models to follow instructions with human feedback, Arxiv 2022.03.
- DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature, arxiv 2023.01.26

- [万字长文解析! 复现和使用GPT-3/ChatGPT, 你所应该知道的](#)
- [斯坦福CS课堂PPT《提示、指令微调和人类反馈的强化学习》](#)
- [思维链CoT系列论文列表](#)
- [上下文学习ICL系列论文列表](#)
- [ChatGPT是如何获得涌现能力的](#)
- [符尧: 预训练、指令微调、对齐、专业化: 论大语言模型能力的来源](#)
- [知乎张佳俊: 关于ChatGPT八个技术问题的猜想](#)



中国科学院 信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING, CAS

请各位专家批评指正