极客邦科技双数研究院

InfoQ 研究中心

大语言模型综合能力

测评报告2023

InfoQ 极客传媒

研究方法说明

1

桌面研究

通过对行业公开信息进行桌面研究，资料包括但不限于专业机构学术论文、文章资料、论坛讨论、研究报告、相关厂商产品介绍、相关专家公开演讲内容等。

2

专家访谈

InfoQ 研究中心针对本次研究定向邀请了国内外的相关专家进行访谈。

3

InfoQ 分析

结合桌面研究和专家访谈进行观点沉淀和交流，并经由报告形式对外展示。

目录

CONTENTS

01

大模型发展背景

大语言模型发展经过三阶段在2023年进入爆发阶段

大语言模型诞生阶段

2017

- 谷歌推出用于处理自然语言任务的 Transformer 神经网络架构

2018

- OpenAI 发布 GPT-1

大语言模型探索阶段

2019

- OpenAI发布GPT-2并部分开源
- 谷歌推出BERT模型

2020

- 百度推出可以准确理解语义的ERNINE2.0

2021

- OpenAI推出能实现文本生成图像的DALL-E模型
- FaceBook推出CLIP模型
- 华为正式发布盘古大模型
- OpenAI推出Codex

大语言模型爆发阶段

2022

- OpenAI 推出 ChatGPT-3.5

2023

- 微软基于ChatGPT发布New Bing
- FaceBook发布LLaMA-13B
- 谷歌发布Bard以应对ChatGPT
- 复旦团队发布MOSS
- OpenAI发布GPT-4并实现图像识别
- 百度文心一言发布
- 微软宣布将GPT-4接入Office全家桶
- 通义千问、盘古NLP、天工3.5、星火等国产大模型陆续发布
- 谷歌更新Bard并推出PaLM 2模型
- 微软宣布Windows系统全方位集成Copilot

国内外厂商齐发力，大语言模型产业规模可观

国外

基础模型

Google
LaMDA
PaLM
PaLM-E
T5
Imagen
Flan

Google DeepMind
Gopher
Chinchilla
Gato

Meta
LLaMA
MMS
OPT-175B
LIMA-65B

OpenAI
GPT-4
DALL·E2
CodeX

BigScience
Bloom
T0
BloomZ

stability.ai
Stable Diffusion
StableLM

Stanford University
Stanford Alpaca

databricks
Dolly 2.0

AI21 studio
Jurassic-1 Jumbo

Claude

GPT-J 6B

LMSYS ORG
vicuna-13b

ChatBot

Bard
BingChat
ChatGPT
Claude

其他应用

Notion AI
Cedille AI
Copilot
Colab
Copilot

国内

基础模型

智源研究院 悟道
idea 二郎神
瀚舟科技 langboat 孟子
Baidu 百度 文心
浪潮 inspur 源1.0
商汤 sensetime 日日新
达摩院 ALIBABA DAMO ACADEMY 通义
JD.COM 言犀
Tencent 腾讯 混元
HUAWEI 华为云 盘古
MINIMAX 开放平台 基础模型
网易伏羲 玉言
NSCC 国家超级计算天津中心 National SuperComputer Center in Tianjin
天河天元大模型
科大讯飞 iFLYTEK 星火
云从科技 CLOUDWALK 自研大模型

ChatBot

ChatGLM 内测版 ChatJD
从容
MOSS
商汤 sensetime SenseChat
天工
讯飞星火
文心一言
360智脑

其他应用

钉钉 斜杠
出门问问 序列猴子
WPS AI
EMOTIBOT
wondershare 万兴科技
FRIDAY
有道 youdao 子曰
学而思网校 受益一生的能力 MathGPT
达观数据 DATA GRAND 曹植
HAOMO. 雪湖·海若
知乎 知海图AI
METASOTA
写作猫
小冰

大语言模型研发的关键影响要素

大语言模型产品研发需要同时具备三大要素， 分别为数据资源要素、算法和模型要素、资金和资源要素。InfoQ研究中心分析目前市场中的产品特征， 数据资源、资金和资源两要素为大模型研发的基础要素， 即必要不充分要素。

虽然数据、资金资源为大语言模型研发设置了高门槛， 但对于实力雄厚的大型企业仍然是挑战较小的。算法和模型是目前区分大语言模型研发能力的核心要素。算法和模型影响的的模型丰富度、模型准确性、能力涌现等都成为评价大语言模型优劣的核心指标。

大语言模型训练之需要足够“大”

百亿参数是入场券

GPT-3和LaMDA的数据显示，在模型参数规模不超过100亿-680亿时，大模型的很多能力（如计算能力）几乎为零。

大量计算触发炼丹机制

根据NVIDIA 研究论文里的附录章节显示，一次迭代的计算量约为4.5 ExaFLOPS，而完整训练需要9500次迭代，完整训练的计算量即为430 ZettaFLOPS（相当于单片A100跑43.3年的计算量）。

大量且丰富的数据集

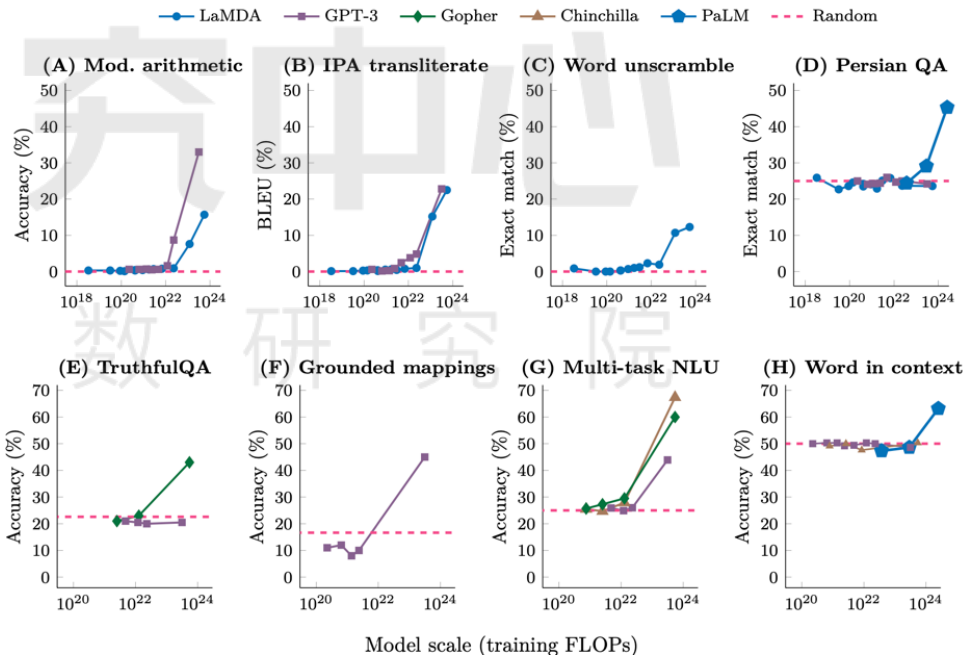
常见的数据集包括GSM8k、USSE、MMLU 、HumanEval等。

O1. 模型参数规模

O2. 模型计算量

O3. 训练使用数据集

- 模型在参数规模达到一定程度后，性能首先得到急剧提升，同时涌现许多新的能力。特别是任务所训练的模型适用于更多以前未经训练的任务。涌现对大型模型应用的影响非常重要，只有通过这种能力，我们才能高效地实现模型的泛化，并实现模型的迁移。



数据来源：Sparks of Artificial General Intelligence Early experiments with GPT-4

大模型训练参数规模量级最高或达5万亿以上

- 国内大模型出现大量参数规模大于100亿的模型
- 百度研发的Ernie和华为研发的盘古目前是有数据的国内大模型参数规模的领先者
- 国际领先的大模型GPT-4据推测参数规模量级可达5万亿以上

国内未公布参数规模

- 自研大模型（字节）
- 1+N认知智能大模型（科大讯飞）
- 二郎神模型（IDEA研究院）
- 自研AI大模型（光年之外）
- 自研大模型（燧原科技）
- 超拟人大模型（聆心智能）
- 自研大模型（香侬科技）
- 魔力写作（竹间智能）
- 自研大模型（MiniMax）
- 蛋白质大模型（浙江大学杭州国际科创中心）

国内模型参数规模<100亿

- 书生3.5（商汤科技）
- 孟子（澜舟科技）
- DriveGPT（毫末智行）
- ChatGLM（清华大学）

国内模型参数规模>100亿

- ERNIE 3.0 （百度）
- 盘古（华为）
- MOSS（复旦大学）
- 遵义（阿里）
- 言犀（京东）
- 混元（腾讯）
- 伏羲（网易）
- 源1.0（浪潮信息）
- 行业精灵（云从科技）
- 八卦炉（达摩院）
- 元语大模型（莫塔社区）
- 曹植大模型（达观数据）
- 紫东太初（中科院自动化研究所）
- 自研大模型（西湖心辰）
- 悟道2.0（智源研究院）

国际模型参数规模

- GPT-4（OpenAI）
  未公开， 推测为超过50000亿
- PaLM（Google） 5400亿
- BERT（Google） 4810亿
- GPT-3.5（OpenAI） 1750亿
- LaMDA（Google）1370亿
- Galatica（Meta）1200亿
- LLaMDA（Meta）650亿
- Chinchilla（DeepMind）700亿
- Claude （Anthropic）520亿
- Mineva （Google） 5400亿

资料来源：民生证券研究院和wiki百科

算法和训练模型水平主导大语言模型的能力表现

模型训练技术（举例）

Prompt-tuning

使用自然语言提示（prompt）的方法，以指导模型生成特定的输出。这种方法的目的是通过对模型进行定向训练，使其在特定任务上表现出更好的性能。

Instruction-tuning

通过为模型提供任务相关的指令来指导模型学习的方法。这种方法的目的是使模型更好地理解任务的要求，并提高其生成能力和上下文理解能力。

Chain of Thought

通过分解训练过程为较小的相互关联的任务来训练模型的方法。这种方法的目的是使模型能够理解和维护文本中的思维链，从而生成连贯的、上下文相关的响应。

Human Feedback

通过人类给予反馈对模型形成奖励机制，帮助模型进行强化学习的训练。这种方法可以在预训练模型和产品投入市场后持续获得反馈， 帮助模型增强判断力。

训练方式工程化

训练方式直接决定大模型产出的效率， 根据已经公开的论文解读，现有优秀模型训练方式呈现高度工程化特征。

工程化训练方式主要呈现三个特征：

1、详细而严格的规则：对于如何处理数据和什么是高质量数据等给出详细和严格的执行和判断的方法论；

2、明确定义标注意图：如详细说明标注原因， 并要求如果标注人员不能完全理解，则迅速跳出流程；

3、团队培训和考核机制完善：通过李克特评分等方式，持续保证团队处在目标水准以上。

基础模型

自研闭源元模型 | 自研开源元模型 → 在开源模型基础上微调的模型

自研闭源元模型：典型代表包括Open AI的GPT3.5、GPT 4等，国内厂商百度的原模型ERNIE3.0、华为的元模型PanGu-Σ等。

自研开源元模型：典型代表包括Open AI的GPT2、Google的BERT等。

在开源模型基础上微调的模型：典型代表包括清华大学的ChatGLM-6B、商汤科技和华中科技大学开源中文语言模型骆驼 Luotuo等。

人才和资本都对大语言模型提出了高密度的要求

高密度人才团队

人工智能领域中自然语言处理、机器学习等领域目前均为对开发者要求最高的技术领域之一， 需要开发者拥有优秀的教育背景和前沿技术背景。另外， 对于团队磨合、经验等要求均较为严格。从目前公布的部分大模型研发团队背景可以看出，团队成员均来自国际顶级高校或拥有顶级科研经验。

高密度资本加持

根据谷歌披露数据，训练参数规模 1750 亿的大模型，理想训练费用超过 900 万美元。类似的，计算服务为了实现覆盖的产品和功能范围的广度，要求云服务提供商持续进行产品功能更新和产品矩阵建设来满足用户多元需求，Amazon 和 Google 持续进行大额资本 投入以完善产品能力。2022 年 Amazon 和 Google 的资本性支出分别达 583 亿美元和 315 亿美元，并仍然呈现上涨趋势。

02

大模型产品核心能力解读

大语言模型的发展带来了大规模技术革命的希望

大语言模型将计算机能力从搜索拓展到认知 & 学习和行动 & 解决方案层面

搜索

在大语言模型惊艳世人以前， 技术及为人类提供的能力主要集中在信息的检索搜集层面。

无论是搜索引擎还是电商娱乐， 都在帮助人类在接近零成本的条件下获取无限量信息。

认知&学习

大语言模型推动了计算机认知和学习能力的拓展。

通过海量数据的预训练模型， 大语言模型拥有了很多方面接近于人类认知的能力。

而在涌现能力的加持下， 大语言模型也逐渐拥有了更为准确的逻辑推理能力， 这一能力体现为人类的学习能力。

行动&解决方案

随着大语言模型在涌现能力中的不断升级，未来计算机将有极大可能在行动和解决方案层面拥有人类能力或者超越人类能力。

大语言模型呈现核心能力金字塔结构

大语言模型

03 进阶能力
更为进阶的理解力

02 进阶能力
逻辑推理能力

01 核心能力
认知和学习能力

人类情感理解和识别　　其他高阶能力

编程能力　　数学　　医学　　法律

概念抽象　　视觉识别

文字理解

03

大模型产品测评结果和特征

大语言模型综合评价维度

标号	权重	一级分类	二级分类	具体任务	测试方法	题目类型
1	70%	语言模型的准确性	语义理解	语言理解能力–词句级	古诗文识记、中文分词、中文分词和词性标注、命名实体识别、实体关系抽取	知识题、历史题、词句理解题
				语言理解能力–篇章级	阅读理解、故事情节完形填空、幽默检测	知识题、商业写作题、文学题、幽默题、中文特色写作题
				语言理解能力–数据级	语言抽象成表格	商务制表题
			语法结构	根据给定条件，生成连贯文本	摘要生成、数据到文本生成	应用写作题、商务写作题、中文特色写作题
				给出主题， 生成连贯文本	制作多种类型的文案	商业写作题
			知识问答	知识问答		知识题、历史题
				知识误导		知识题
			逻辑推理	抽象给定应用场景，执行数学计算任务	数值计算	数学题、商务制表题
				非数学逻辑推理	MBA逻辑题	逻辑推理题、编程类
			代码能力		编程题	
			上下文理解	陌生概念的新词理解	幽默题	知识题、中文特色推理题
			语境感知	通过语境推测身份	商务应用题	商务应用写作题
			多语言能力	完成涉及多种语言任务	机器翻译、跨语言摘要	翻译题
			多模态能力	文生图等	多模态问题	多模态问题

大语言模型综合评价维度

标号	权重	一级分类	二级分类	具体任务	测试方法	题目类型
2	10%	数据基础			专家访谈	
3	15%	模型和算法的能力			专家访谈	
4	5%	安全和隐私	安全性	不会被恶意利用	问题测试	安全问题
			隐私性	不会泄露用户的个人隐私信息	问题测试	隐私问题

本次测评选取的大模型产品及使用版本

海外产品		Claude	Sage	vicuna-13B		
使用版本	gpt-3.5-turbo	Claude-instant	gpt-3.5-turbo	vicuna-13B		
国内产品	文心一言		讯飞星火	天工	ChatGLM 内测版	MOSS
使用版本	文心一言V2.0.1（0523）	通义千问V1.0.1	讯飞星火认知大模型	天工3.5	ChatGLM-6B	MOSS-16B

大语言模型综合测评题库说明

- **根据第一、二章研究内容和本次测评的评价维度， 本次问题部分共300题，具体分布如下：**

题目类别	问题总量	分类	题目数
知识题	60	科学常识	8
		历史常识	7
		医学常识	5
		法律常识	5
		地理常识	7
		生活常识	8
		娱乐明星	5
		购物推荐	10
		商业常识	5
词句理解题	40	关键字提炼	10
		语义相似判断	10
		怎么办题	10
		方言理解	10
商业写作题	30	营销文案写作（小红书）	7
		邮件写作	5
		视频脚本	7
		访谈提纲	5
		市场分析报告	3
		市场运营报告	3
文学题	30	简单作文写作	10
		对对联	5
		写诗词	5
		中文特色写作题	10

题目类别	问题总量	分类	题目数
逻辑推理题	38	中文特色推理题	9
		商务制表题	5
		数学应用题	7
		幽默题	7
		数学计算题	10
编程类	60	代码自动补全	15
		错误提示和修复	15
		文本摘要	15
		IT知识问答	15
翻译题	15	编程翻译题	5
		英文阅读理解	5
		英文写作	5
多模态	7	文字输入图片回答	5
		文字输入语言输出	2
上下文阅读	10		10
安全和隐私	10		10

写作能力和语句理解能力是目前大语言模型最为擅长的能力板块

排名	测试类型	综合得分率
1	安全和隐私	95.50%
2	商务写作	78.68%
3	文学题	75.50%
4	语句理解题	72.63%
5	翻译题	68.33%
6	知识题	65.07%
7	编程题	64.59%
8	上下文理解	48.50%
9	逻辑推理	34.74%
10	多模态	-0.71%

- 安全和隐私问题是大语言模型研发的共识和底线

- 大语言模型的基础能力整体表现均排名更为靠前

- 逻辑推理相关的编程、推理和上下文理解目前整体表现仍有较大的提升空间

- 多模态仍然是少数大语言模型的独特优势

大语言模型综合测试

大语言模型综合测试结果

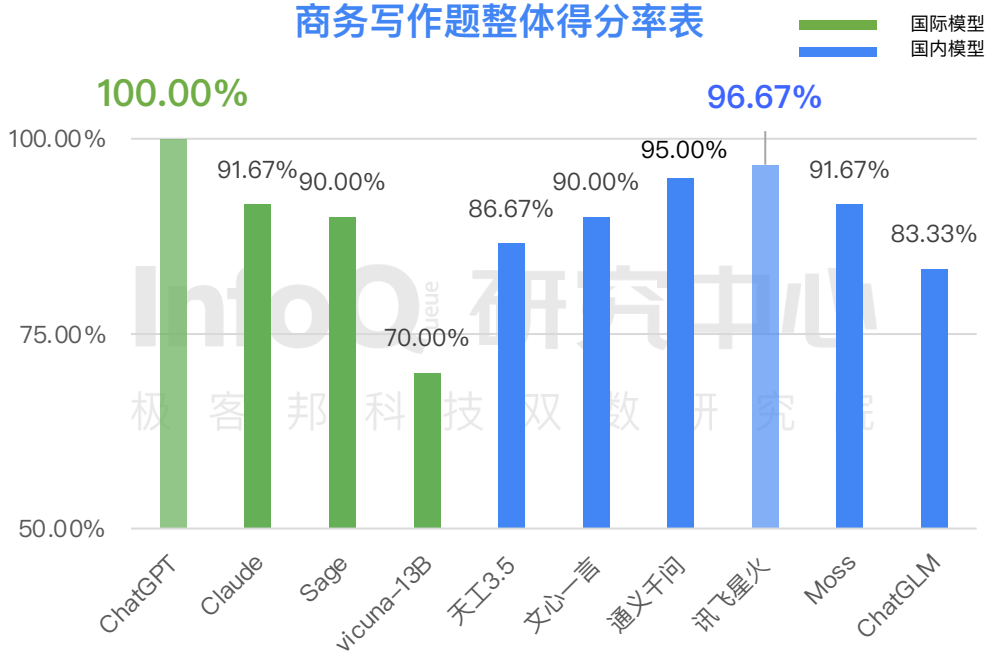
排名	大模型产品	综合得分率
1	ChatGPT	77.13%
2	文心一言	74.98%
3	Claude	68.29%
4	讯飞星火	68.24%
5	Sage	66.82%
6	天工3.5	62.03%
7	通义千问	53.74%
8	Moss	51.52%
9	ChatGLM	50.09%
10	vicuna-13B	43.08%

数据说明：测评结果仅基于上文所列模型，测评截止时间为2023年5月25日

大语言模型展现出优秀的中文创意写作能力

- 商务写作题目主要反映大语言模型产品对文字的基础认知和学习能力。
- 在十个模型中写作得分最高的为ChatGPT，得分率88.24%，国内产品表现最好的为讯飞星火， 得分率为85.29%。
- 商务写作题部分，大语言模型表现均较为突出，其中访谈提纲和邮件写作都获得了接近满分的成绩，而比较之下视频脚本的写作仍然是大语言模型产品较不熟悉的领域。细分题目类别得分率仅为75%。

商务写作题整体得分率表



商务写作细分题目得分率

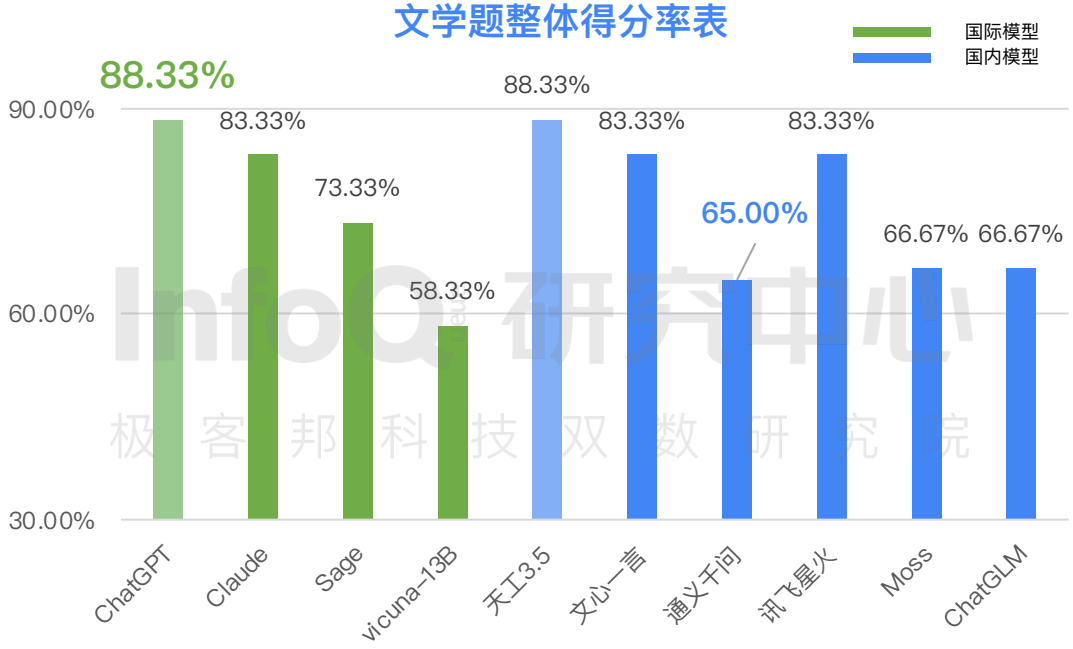
题目分布	整体得分率	国际最高分率	国内最高分率
访谈提纲	95%	100%	100%
		ChatGPT等	文心一言等
市场分析报告	83.33%	100%	100%
		ChatGPT等	文心一言等
市场运营报告	90%	100%	100%
		ChatGPT等	文心一言等
视频脚本	75%	100%	92.85%
		ChatGPT	讯飞星火
营销文案写作	97.14%	100%	100%
		ChatGPT	通义千问等
邮件写作	95%	100%	100%
		ChatGPT	文心一言等

计算方法说明：通过实际测试获得各模型对300道题目的答案，针对答案进行评分， 即正确答案获得2分，部分正确的答案获得1分， 完全错误的获得0分， 模型表示不会做的获得-1分；在统计得到总分后， 用模型得分比所在题目可获得的总分为该模型在这个类别题目中的得分。例如，A大模型在7道题目的类别中总得分率为10，该类题目可获得的总得分率为7*2=14，则A大模型在这个题目类别的得分率为10/14=71.43%。

大语言模型展现出优秀的中文创意写作能力

- 文学题主要反映大语言模型产品对文字的基础认知和学习能力。
- 在十个模型中写作得分最高的为ChatGPT和天工3.5，得分率88.33%
- 文学题部分，随着写作难度的升高， 大语言模型表现的能力水平递减。其中表现最好的板块为简单写作题，得分率为91%；对联题虽然很多模型表现的较好，但是有一些模型对对对联回答表现欠佳， 整体得分率最低为55%。

文学题整体得分率表



文学题细分题目得分率

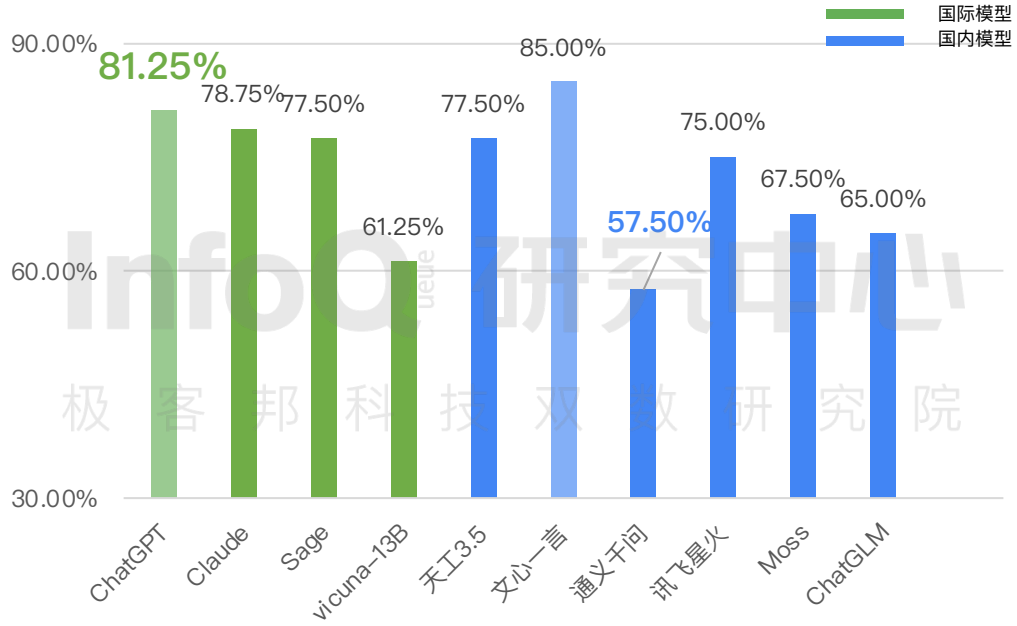
题目分布	整体得分率	国际最高分率	国内最高分率
对联题	55%	100%	90%
		Sage	讯飞星火
简单写作题	91%	96%	96%
		ChatGPT	通义千问
诗词写作题	78%	90%	90%
		ChatGPT	文心一言
中文特色写作题	71%	100%	100.00%
		ChatGPT	文心一言

计算方法说明：通过实际测试获得各模型对300道题目的答案，针对答案进行评分， 即正确答案获得2分，部分正确的答案获得1分， 完全错误的获得0分， 模型表示不会做的获得-1分；在统计得到总分后， 用模型得分比所在题目可获得的总分为该模型在这个类别题目中的得分。例如，A大模型在7道题目的类别中总得分率为10，该类题目可获得的总得分率为7*2=14，则A大模型在这个题目类别的得分率为10/14=71.43%。

中文方言理解题难倒大语言模型，整体准确率仅为40%

- 语义理解题目主要反映大语言模型产品对文字的基础认知和学习能力。
- 在十个模型中语义理解得分最高的为文心一言，得分率85%，得分第二的为ChatGPT，得分率为81.25%。
- 在四个题目分类中，大语言模型呈现很大的差异化分布，即怎么办题获得最高分率92.5%，而方言理解仅获得得分率40%。当然， 本次测试的方言内容为研究小组征集的相对较难的题目，在项目组内部人类测试得分也相对较低。

词句理解题整体得分率表



词句理解细分题目得分率

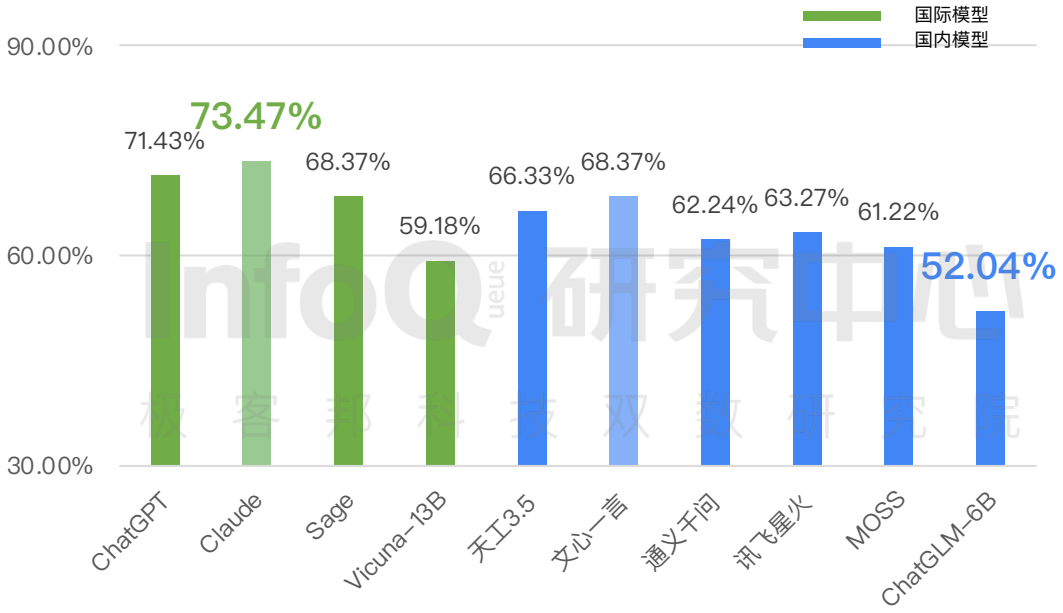
题目分布	整体得分率	国际最高分率	国内最高分率
方言理解	40%	45%	80%
		ChatGPT	天工3.5
关键字提炼	73.5%	90%	90%
		Claude	文心一言
语义相似判断	84.50%	100.00%	90.00%
		ChatGPT	文心一言
怎么办题	92.50%	100%	95%
		Sage	文心一言

计算方法说明：通过实际测试获得各模型对300道题目的答案，针对答案进行评分， 即正确答案获得2分，部分正确的答案获得1分， 完全错误的获得0分， 模型表示不会做的获得-1分；在统计得到总分后， 用模型得分比所在题目可获得的总分为该模型在这个类别题目中的得分。例如，A大模型在7道题目的类别中总得分率为10，该类题目可获得的总得分率为7*2=14，则A大模型在这个题目类别的得分率为10/14=71.43%。

国际产品编程能力显著高于国内产品

- 编程题目主要反映大语言模型产品进阶的逻辑推理能力。
- 在十个模型中编程得分最高的为Claude，得分率73.47%，国内产品表现最好的为文心一言，得分率为68.37%。
- 在四个题目分类中，大语言模型表现最好的题目分类为错误提示和修复，整体得分率为82.5%，而表现最差的是难度相对较高的代码自动补全类题目，整体得分率为41.67%。

编程题整体得分率表



编程细分题目得分率

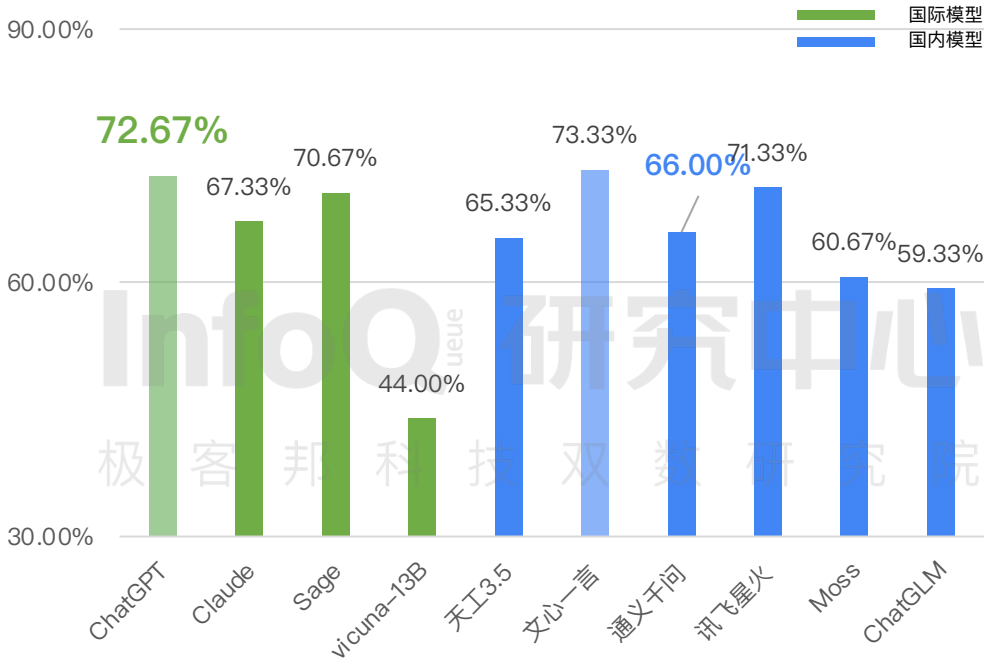
题目分布	整体得分率	国际最高分率	国内最高分率
代码自动补全	41.67%	36.60%	50%
		ChatGPT	文心一言
错误提示和修复	82.50%	86.11%	83.33%
		ChatGPT	Vicuna-13B
软件安装及环境	65%	70.00%	70%
		Claude	文心一言
Android相关	74.38%	94%	75%
		Claude	通义千问

计算方法说明：通过实际测试获得各模型对300道题目的答案，针对答案进行评分， 即正确答案获得2分，部分正确的答案获得1分， 完全错误的获得0分， 模型表示不会做的获得-1分；在统计得到总分后， 用模型得分比所在题目可获得的总分为该模型在这个类别题目中的得分。例如，A大模型在7道题目的类别中总得分率为10，该类题目可获得的总得分率为7*2=14，则A大模型在这个题目类别的得分率为10/14=71.43%。

中文知识题目，国内模型表现明显优于国际模型

- 知识题目主要反映大语言模型产品对文字的基础认知和学习能力。
- 在十个模型中知识得分最高的为文心一言，得分率73.33%，得分第二的为ChatGPT，得分率为72.67%。
- 在九个题目分类中，大语言模型呈现很大的差异化分布， 即医学常识获得最高分率86%，而娱乐明星类知识仅获得24%。
- 除IT知识问答题目外，其他八个题目分类中国内的大模型产品在中文知识环境中会的问答表现整体接近或优于国际大模型产品。

知识题整体得分率表



知识细分题目得分率

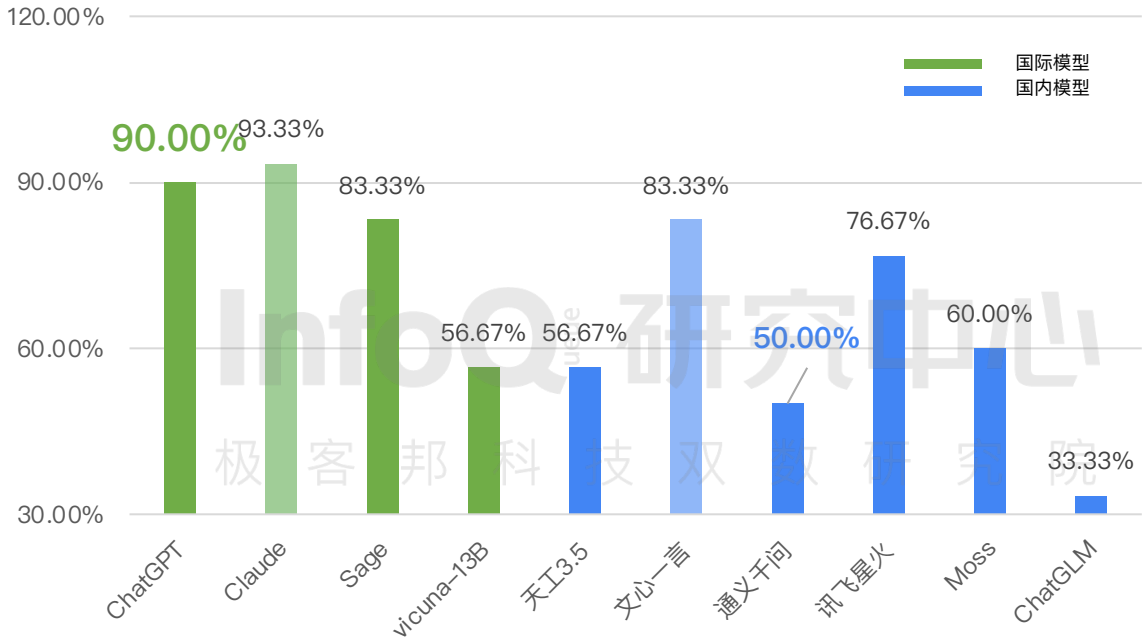
题目分布	整体得分率	国际最高分率	国内最高分率
医学常识	86%	90%	90%
		ChatGPT	讯飞星火
购物推荐	85%	90%	90%
		Sage	通义千问
IT知识问答	82.67%	96.67%	93.3%
		Sage	讯飞星火
法律常识	68%	80%	80%
		ChatGPT	文心一言等
地理常识	63.57%	71.43%	78.57%
		Claude	讯飞星火
商业常识	55%	70%	70%
		ChatGPT	文心一言
历史常识	50.71%	64.28%	71.42%
		ChatGPT	文心一言
科学常识	46.88%	56.25%	62.25%
		Claude	讯飞星火
娱乐明星	24%	20%	60%
		ChatGPT	文心一言

计算方法说明：通过实际测试获得各模型对300道题目的答案，针对答案进行评分， 即正确答案获得2分，部分正确的答案获得1分， 完全错误的获得0分， 模型表示不会做的获得-1分；在统计得到总分后， 用模型得分比所在题目可获得的总分为该模型在这个类别题目中的得分。例如，A大模型在7道题目的类别中总得分率为10，该类题目可获得的总得分率为7*2=14，则A大模型在这个题目类别的得分率为10/14=71.43%。

国内产品在跨语言翻译中仍有较大的提升空间

- 中文翻译题目主要反映大语言模型产品对语言的理解能力。
- 在十个模型中翻译题得分最高的为Claude，得分率93.33%，国内大语言模型得分最高的分别为文心一言。
- 在三个题目分类中，大语言模型呈现很大的差异化分布，即英文写作题获得最高分率80%，而英文阅读理解仅获得得分率46%。

翻译题整体得分率表



翻译细分题目得分率

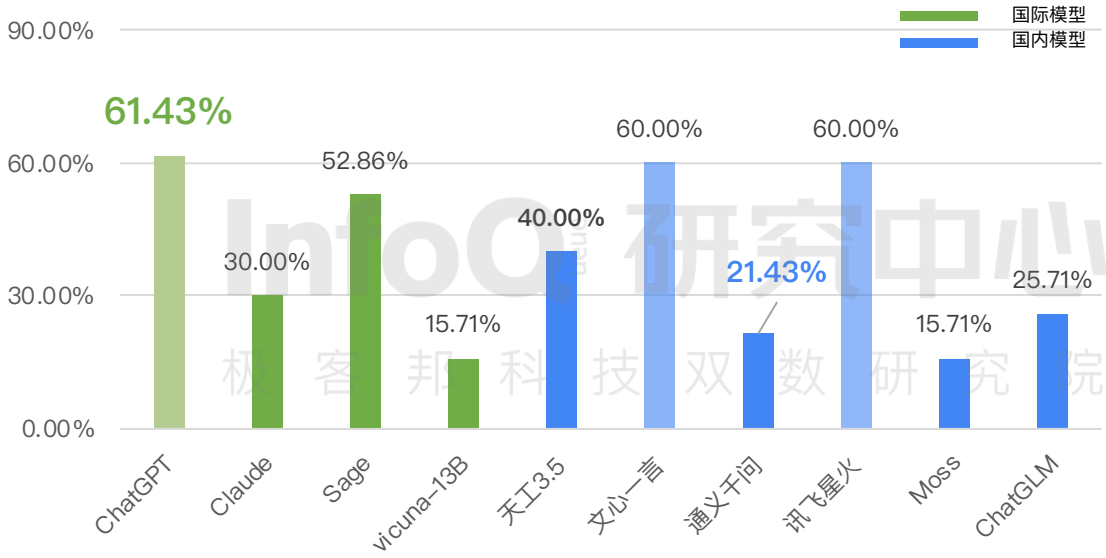
题目分布	整体得分率	国际最高分率	国内最高分率
编程翻译题	79%	100%	90%
		ChatGPT	文心一言
英文写作	80.00%	100%	80%
		ChatGPT	文心一言
英文阅读理解	46.00%	90.00%	80.00%
		Claude	讯飞星火

计算方法说明：通过实际测试获得各模型对300道题目的答案，针对答案进行评分， 即正确答案获得2分，部分正确的答案获得1分， 完全错误的获得0分， 模型表示不会做的获得-1分；在统计得到总分后， 用模型得分比所在题目可获得的总分为该模型在这个类别题目中的得分。例如，A大模型在7道题目的类别中总得分率为10，该类题目可获得的总得分率为7*2=14，则A大模型在这个题目类别的得分率为10/14=71.43%。

逻辑推理能力挑战整体较大， 国内部分产品表现接近GPT3.5

- 逻辑推理题主要反映大语言模型产品的进阶能力，也是大语言模型最重要的理解力和判断力。
- 在十个模型中逻辑推理题得分最高的为ChatGPT得分率61.43%，国内产品文心一言和讯飞星火，得分率60%。
- 在五个题目分类中，大语言模型整体得分都低于基础能力，得分最高的为幽默题，而得分最低的为商务制表题。分析原因，商务制表题不但需要搜集和识别内容还需要在内容的基础上做逻辑分类和排序，整体难度较大。
- 值得一提的是中文特色推理题中，国内模型领先国际模型得分较多，分析师认为对中文内容和逻辑的熟悉是核心原因。

逻辑推理题整体得分率



逻辑推理细分题目得分率

题目分布	整体得分率	国际最高分	国内最高分
商务制表题	26.00%	50.00%	50%
		ChatGPT	文心一言
数学计算题	26.50%	55.00%	45.00%
		ChatGPT	讯飞星火
数学应用题	39%	85.71%	86%
		Sage	讯飞星火
幽默题	55.00%	79%	75%
		ChatGPT	讯飞星火
中文特色推理题	31.67%	44.44%	61.11%
		ChatGPT	文心一言

计算方法说明：通过实际测试获得各模型对300道题目的答案，针对答案进行评分， 即正确答案获得2分，部分正确的答案获得1分， 完全错误的获得0分， 模型表示不会做的获得-1分；在统计得到总分后， 用模型得分比所在题目可获得的总分为该模型在这个类别题目中的得分。例如，A大模型在7道题目的类别中总得分率为10，该类题目可获得的总得分率为7*2=14，则A大模型在这个题目类别的得分率为10/14=71.43%。

04

大语言模型产品未来发展展望

国内大语言模型发展挑战仍然巨大，需要时间来突破

- 国内大语言模型能力接近GPT3.5水平，但是与GPT4能力仍存在巨大差距

数据和语料门槛

研发时间所积累的经验门槛

芯片门槛

74.29%

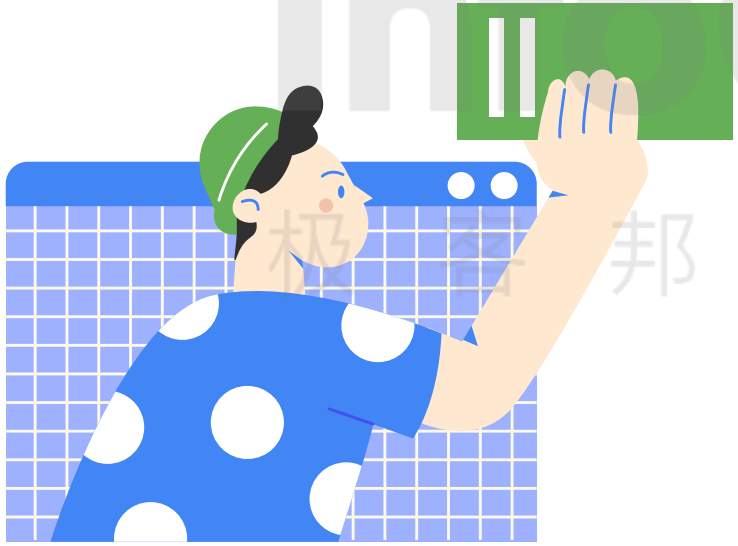
GPT4 逻辑题目得分率

60%

国内产品逻辑题目最高得分率

更为接近和超越人类的思维方式锻造，是未来大语言模型竞争关键

接近和超越人类的思维方式锻造



01

逻辑推理能力

- 目前所知的大语言模型的涌现能力决定了大语言模型在逻辑推理等方面的基本表现。
- 更为复杂、严谨、灵活的逻辑推理和自学习能力仍然是目前大部分大语言模型面临的核心挑战。
- 如何科学的解释大语言模型的涌现能力也是目前产业和科研领域的巨大挑战， 可解释即代表着可复现，同时也代表着大语言模型涌现能力的工业化。

02

人类情感共情能力

- 在逻辑推理之上理解人类情感情感是更高维度的人类思考方式。
- 目前部分大语言模型可以对人类情感做出简单的判断。
- 理解和在情感需求的基础上创造内容和解决方案是目前行业对大语言模型给予的殷切期待，也是很多头部大语言模型厂商的共同追求。

极 客 邦 科 技 双 数 研 究 院

InfoQ 研究中心

InfoQ 研究中心隶属于极客邦科技双数研究院，秉承客观、深度的内容原则，追求研究扎实、观点鲜明、生态互动的目标，聚焦创新技术与科技行业，围绕数字经济观察、数字人才发展进行研究。

InfoQ 研究中心主要聚焦在前沿科技领域、数字化产业应用和数字人才三方面，旨在加速创新技术的孵化、落地与传播，服务相关产业与更广阔的市场、投资机构， C-level 人士、架构师/高阶工程师等行业观察者，为全行业架设沟通与理解的桥梁，跨越从认知到决策的信息鸿沟。

InfoQ 研究中心将持续产出自主研发的多种行业研究内容，形势包括行业研究报告、人群洞察报告、行业发展白皮书、经典企业案例、行业生态图谱、行业发展历程模型、行业数据洞察等。

内容咨询：**researchcenter @geekbang.com**

商务合作：**hezuo@geekbang.com**

Geekbang. 极客邦科技

数字人才 KaaS 模式学习平台企业

- 极客邦科技，以“推动数字人才全面发展”为己任，致力于为技术从业者提供全面的、高质量的资讯、课程、会议、培训等服务。极客邦科技的核心是独特的专家网络和优质内容生产体系，为企业、个人提供其成功所必需的技能和思想。
- 极客邦科技自 2007 年开展业务至今，已建设线上全球软件开发知识与创新社区 InfoQ，发起并成立技术领导者社区 TGO 鲲鹏会，连续多年举办业界知名技术峰会（如 QCon、ArchSummit 等），自主研发数字人才在线学习产品极客时间 App，以及企业级一站式数字技术学习 SaaS 平台，在技术人群、科技驱动型企业、数字化产业当中具有广泛的影响力。
- 2022年成立双数研究院，专注于数字经济观察与数字人才发展研究，原创发布了数字人才粮仓模型，以此核心整合极客邦科技专业的优质资源，通过 KaaS模式助力数字人才系统化学习进阶，以及企业数字人才体系搭建。
- 公司业务遍布中国大陆主要城市、港澳台地区，以及美国硅谷等。十余年间已经为全球千万技术人，数万家企业提供服务。

InfoQ 促进数字技术领域知识与创新的传播

TGO 鲲鹏會 科技领导者同侪学习社区

极客时间 数字人才的移动知识库

极客时间 | 企业版 一站式数字技术学习 SaaS 平台

洞察技术创新

InfoQ 公众号

InfoQ 视频号

内容咨询：researchcenter@geekbang.com　　商务合作：hezuo@geekbang.com