

极客邦科技双数研究院

InfoQ^{lab} 研究中心

大语言模型综合能力 测评报告 2024

InfoQ^{lab} 极客传媒



01

大模型市场发展洞察

2023年，全球大模型市场进入爆发阶段

大模型诞生阶段

2017-2018年

在诞生阶段，以Transformer为代表的全新神经网络架构，奠定了大模型的算法架构基础，使大模型技术的性能得到了显著提升。

2017

- 6.12 谷歌推出用于处理自然语言任务的Transformer神经网络架构

2018

- 6.11 OpenAI 发布 GPT-1
- 11.22 谷歌推出BERT

大模型探索阶段

2019-2021年

在探索期，基于人类反馈的强化学习（RHLF）、代码预训练、指令微调等开始出现，被用于进一步提高推理能力和任务泛化。

2019

- 3.16 百度推出可以准确理解语义的 ERNIE 1.0
- 8.21 OpenAI 发布 GPT-2 并部分开源
- 10.11 谷歌推出 BERT 模型的两个新版本 RoBERTa 和 XLNet

2020

- 11.5 OpenAI发布120亿参数图像版GPT-3

2021

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

大模型爆发阶段

2022-2023年

在爆发期，大数据、大算力和大算法完美结合，具备了多模态理解与多类型内容生成能力。

2022

- 11.30 OpenAI 推出 ChatGPT

2023

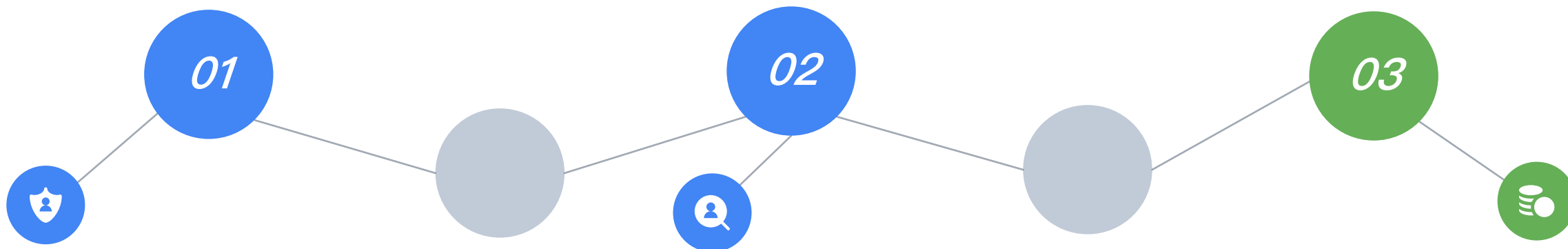
- 2.7 谷歌发布Bard以应对ChatGPT
- 2.20 复旦团队发布MOSS
- 2.23 微软基于ChatGPT发布New Bing
- 2.25 FaceBook发布LLaMA-13B
- 3.14 OpenAI发布GPT-4并实现图像识别
- 3.16 百度发布文心一言
- 3.17 微软宣布将GPT-4接入Office全家桶
- 4.10 昆仑万维发布天工3.5
- 4.11 阿里云发布通义千问
- 5月 百度发布文心大模型3.5
- 5.6 科大讯飞发布认知大模型
- 5.11 谷歌更新Bard并推出PaLM 2模型
- 5.24 微软宣布Windows系统全方位集成Copilot

2023年中央及地方政府积极制定政策鼓励和规范大模型产业发展

2023年，中央及地方围绕基础设施、算法、算力、数据等领域出台多项政策，坚持发展和安全并重、促进创新和依法治理相结合的原则，鼓励和规范作为硬科技的AI大模型产业发展。

序号	时间	文件名	主要内容
1	2023年7月	《生成式人工智能服务管理暂行办法》	明确了生成式人工智能“提供者”内容生产、数据保护、隐私安全等方面的法定责任及法律依据，确立了人工智能产品的安全评估规定及管理办法。
2	2023年7月	《上海市推动人工智能大模型创新发展的若干措施》	推动上海大模型创新发展，营造通用人工智能创新生态，加快打造世界级人工智能产业集群。
3	2023年8月	《电子信息制造业2023—2024年稳增长行动方案》	鼓励加大数据基础设施和人工智能基础设施建设，满足人工智能、大模型应用需求。
4	2023年10月	《人形机器人创新发展指导意见》	到2025年，人形机器人创新体系初步建立，“大脑、小脑、肢体”等一批关键技术取得突破，确保核心部件安全有效供给，开发基于人工智能大模型的人形机器人“大脑”。
5	2023年11月	《北京市人工智能行业大模型创新应用白皮书（2023年）》	从大模型全球发展态势、国内外行业应用概述、北京应用情况和发展建议等方面进行了系统分析和阐述，旨在进一步推动大模型应用落地，展示北京市大模型应用成果，促进大模型价值传播和供需对接。
6	2023年11月	广东省《关于加快建设通用人工智能产业创新引领地的实施意见》	22条举措打造国家通用人工智能产业创新引领地。
7	2023年12月	《关于加快推进视听电子产业高质量发展的指导意见》	支持骨干企业做大做强，支持人工智能企业研发视听应用大模型。
8	2023年12月	《“数据要素x”三年行动计划（2024—2026年）（征求意见稿）》	以科学数据支持大模型开发，深入挖掘包含科技文献在内的各类科学数据，通过细粒度的知识抽取，构建科学知识资源底座，建设高质量语料稿库和基础科学数据集，支持开展通用人工智能大模型和垂直领域人工智能大模型训练。

2023年下半年，中国百模大战掀开序幕



2023年6月至2023年8月

- 6.14 360发布智脑大模型4.0
- 6.15 百川智能发布开源Baichuan-7B
- 6.25 清华&智谱AI发布开源ChatGLM2
- 7.7 华为发布盘古3.0
- 7.7 商汤科技发布商量2.0
- 7.9 西湖心辰发布西湖大模型
- 7.19 Meta发布开源Llama2

2023年8月至2023年10月

- 8.3 阿里云推出开源Qwen-7B
- 8.7 元象科技发布开源XVERSE-13B
- 8.11 小米发布大模型MiLM
- 8.17 字节跳动公测大模型产品豆包
- 8.21 云从科技发布从容大模型1.5
- 9.5 商汤科技发布商量3.0
- 9.6 百川智能发布开源Baichuan2
- 9.7 腾讯发布混元助手
- 9.20 上海人工智能实验室发布开源InternLM-20B
- 9.25 OpenAI发布多模态GPT-4V

2023年10月至2023年12月

- 10.9 月之暗面发布Kimi Chat
- 10.17 百度发布文心大模型4.0
- 10.24 科大讯飞发布星火大模型3.0
- 10.27 智谱AI开源了ChatGLM3等模型
- 10.30 百川智能发布Baichuan2-192K
- 10.31 阿里云发布通义千问2.0
- 11.7 OpenAI发布GPT-4 Turbo
- 11.16 OPPO发布AndesGPT
- 11.24 零一万物发布开源Yi-34B
- 12.22 腾讯发布混元大模型标准版

文本理解和逻辑能力提升、具备初级代码生成、跨模态能力

多模态、长文本、逻辑能力加强

从头部企业到百模家族，大模型井喷式爆发增长

行业大模型

游戏电商 	社交文娱 	教育 	汽车 	金融 	医疗 	其他
能源制造 	传媒影视 	有道 youdao 	理想 	马上消费金融 	京东健康 	华为云

通用大模型

服务商 	极客邦科技双数研究院 	科技公司 	研究机构
----------------	-----------------------	-----------------	-----------------

数据来源：InfoQ研究中心根据厂商官网、媒体资讯等公开资料整理

在法规的指导下，部分产品开始持证上岗

百度、阿里、腾讯、智谱AI、百川智能等多家企业和机构，按照《互联网信息服务算法推荐管理规定》履行算法备案和变更、注销备案手续，获得备案批准，对公众全面开放，标志着大模型发展进入新阶段，加速推动大模型产业化应用。

截止到2023年12月31日，已有两批超过20个大模型获得备案，面向公众开放。应用范围涵盖自然语言处理、图像识别、语音识别等多个领域，可应用于金融、医疗、教育等各个领域。

批次	模型名称	所属企业
第一批	ABAB大模型	MiniMax
	百川大模型	百川智能
	混元大模型	腾讯
	GLM大模型	智谱AI
	盘古大模型	华为
	日日新大模型	商汤
	书生通用大模型	上海人工智能实验室
	通义大模型	阿里
	文心一言	百度
	星火认知大模型	讯飞
	云雀大模型	抖音
	紫东太初大模型	中科院

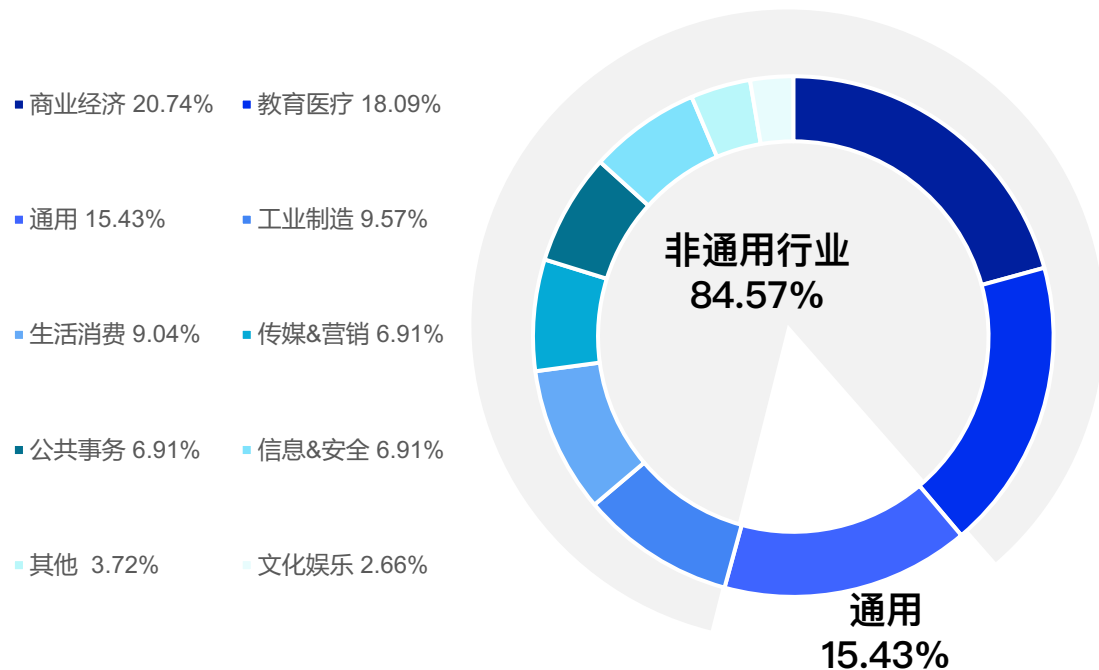
批次	模型名称	所属企业
第二批	百灵大模型	蚂蚁集团
	MathGPT大模型	好未来
	Moonshot	月之暗面
	面壁露卡Luca	面壁智能
	奇元大模型	360
	天工大模型	昆仑万维
	WPSAI	金山办公
	序列猴子	出门问问
	子曰大模型	网易有道
	知海图AI模型	知乎
	模型（未公开）	美团

说明：表格内大模型排序按照大模型名字拼音首字母进行排序

2023年，中国大模型产品从通用产品向多行业拓展

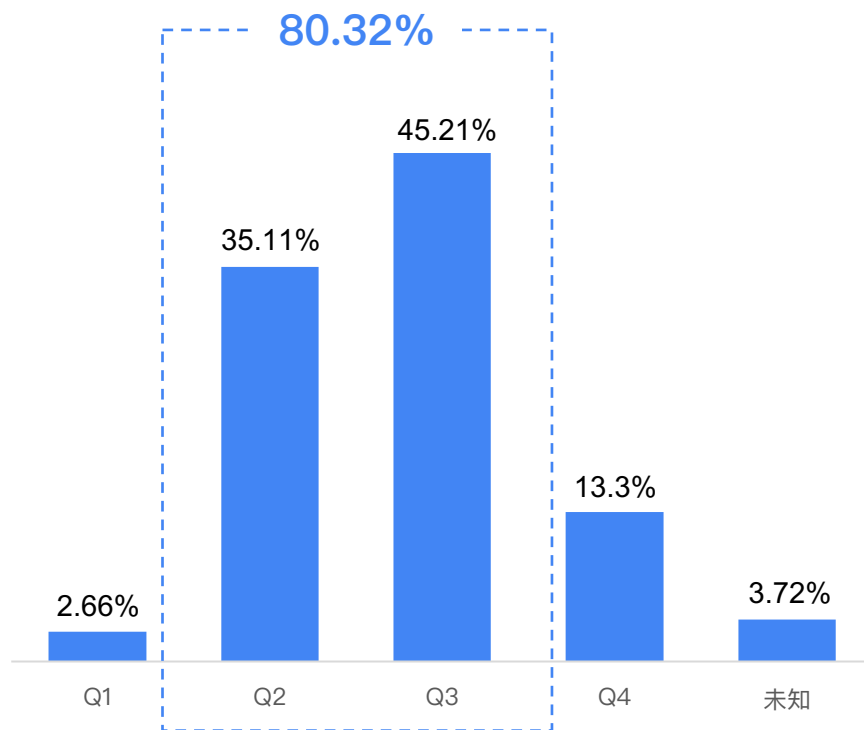
- 2023年公布的中国大模型产品中，84.57%的大模型产品为非通用行业模型。其中工商业经济、教育医疗占比均高于通用模型占比。
- 数据结果统计显示中国大模型产品发布时间集中于Q2和Q3，占比达全年的80.32%。

2023年中国通用大模型和行业应用产品统计



数据来源：根据公开资料搜集的中国188个大模型&产品统计得出

2023年中国主要大模型发布时间统计



2023年，模型层百花齐放，并逐渐开始应用探索



数据来源：InfoQ研究中心根据厂商官网、媒体资讯等公开资料整理

从通用大模型到大模型产品，大模型细分应用赛道不断涌现

2023年国内主要科技公司在推出通用大模型的同时，也正在根据企业资源特征、用户使用场景、生态圈层需求等将通用大模型产品逐渐扩展成为覆盖多个应用场景的产品家族。大模型头部厂商逐渐形成。



通用大模型：文心大模型
大模型产品：文心一言、文心一格
行业产品家族：媒体、能源、金融等行业的13款产品



通用大模型：讯飞星火认知大模型
大模型产品：讯飞星火
行业产品家族：教育、办公、汽车等行业的4款产品



通用大模型：通义大模型
大模型产品：通义千问
行业产品家族：金融、法律、医疗等行业9款产品



通用大模型：混元大模型
通用大模型产品：腾讯混元
行业产品家族：金融、文旅等行业的5款产品

2023年大模型从极客世界走向大众

从定邀小众用户到 APP 端大众可以直接下载，大模型产品正在从极客世界走向更多大众。

产品发布邀请测试

2023年上半年只有少数几款产品可以直接申请使用。大部分产品需要获得开发的企业定向邀请，才能注册使用。

01

向公众开放测试

2023年8月31日起，文心一言、通义千问、讯飞星火等产品逐步完全向公众开放，9月APP应用商店的下载冠军多次由大模型产品获得。

02

常用应用中嵌套功能

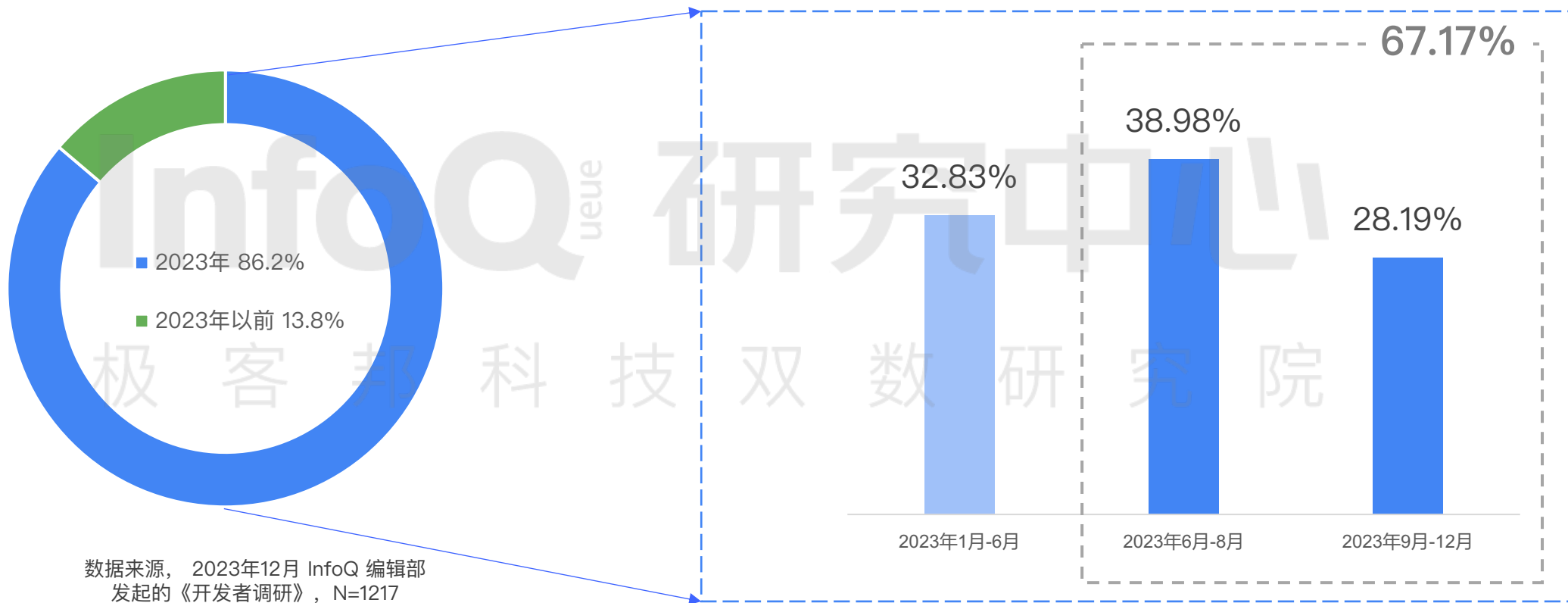
通义千问入驻钉钉号，百度文库也接入了文心一言，上线了智能小助手功能。办公场景中，大模型开始更直接地与用户互动，性能水平直线提升。

03

2023年下半年，中国大模型用户开始真正使用产品

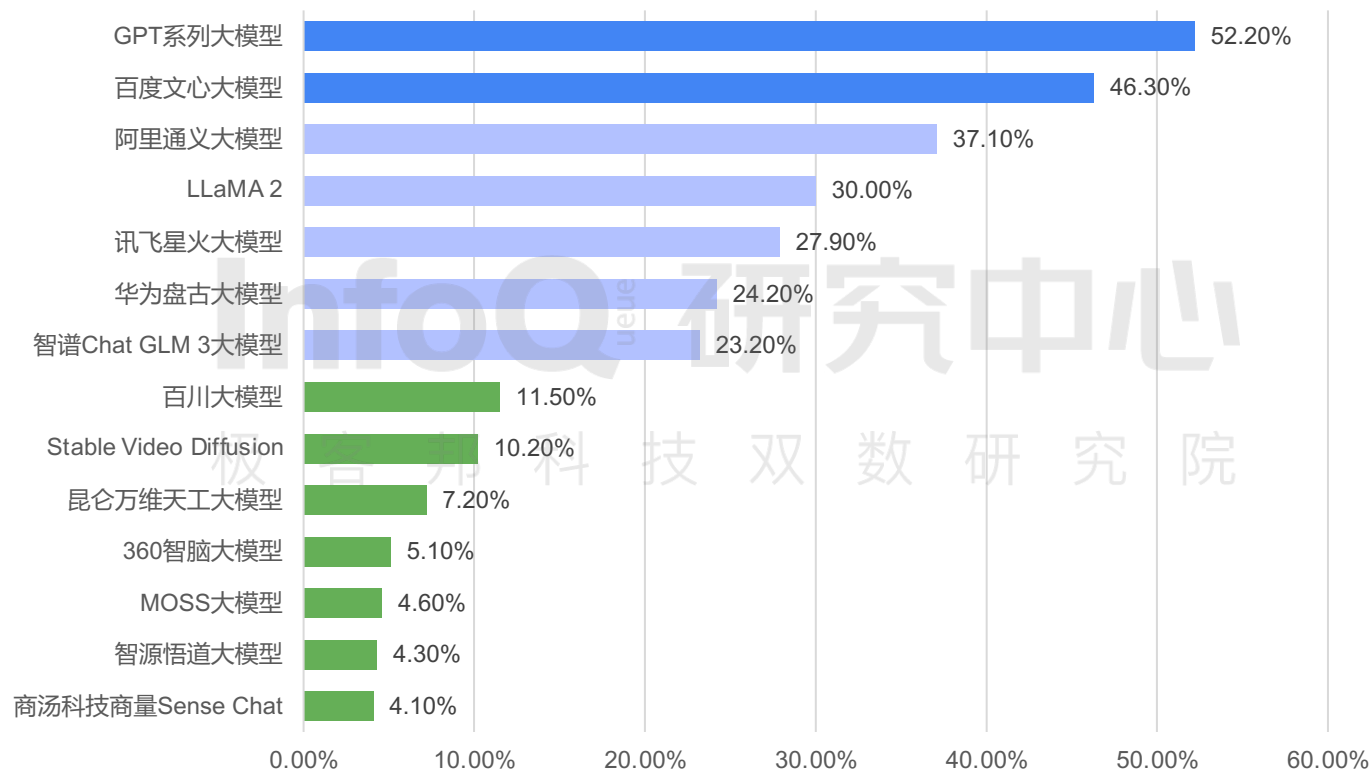
- 2023年9月起，中国大模型产品开始陆续开放公众注册和使用功能。
- 以开发者为例，用户大模型产品开始使用时间67.17%集中在2023年下半年。

开发者开始使用大模型产品的时间统计



2023年，中国大模型正在进入更多用户的认知心智中

开发者大模型的认知度统计



数据来源，2023年12月 InfoQ 发起的《开发者调研》，N=1217

开发者认知心智梯队基本形成

01

第一梯队

GPT系列大模型、百度文心大模型

- 近半数受访开发者了解或使用过上述模型。

02

第二梯队

阿里通义大模型、LLaMA 2、讯飞星火大模型、华为盘古大模型、智谱Chat GLM 3大模型

- 超过1/5的受访开发者了解或使用过第二梯队模型。

03

第三梯队

百川大模型、Stable Video、Diffusion、昆仑万维天工大模型、360智脑大模型、MOSS大模型、智源悟道大模型、商汤科技商量Sense Chat

- 大模型认知度为4%–11.5%。

02

大模型产品测评方法和综合结果

大模型综合测评产品和版本

编号	分类	权重	具体任务	测试方法	题目类型
1	语义理解	10%	语言理解能力	古诗文识记、中文分词和词性标注、命名实体识别、阅读理解、实体关系抽取	关键字提炼、语义相似判断、怎么办题
2	文学写作	5%	根据给定条件，生成连贯文本	制作文案	作文写作、中文特色写作
			给出主题，生成连贯文本	制作多种类型的文案	对对联、写诗词
3	知识问答	5%	知识问答		历史常识、法律常识、地理常识、商业常识、医学常识、科学常识、生活常识、购物常识
4	逻辑推理	20%	抽象给定应用场景，执行数学计算任务	数值计算	数学应用题、数学计算题、商务制表题
			非数学逻辑推理	逻辑推理	中文特色推理、MBA逻辑题、编程类
5	编程	15%	代码能力	程序编写、错误提示、IT知识问答	代码自动补全、错误提示和修复、文本摘要、IT知识问答
6	上下文理解	15%	多轮对话的理解、连贯	上下文题、多轮对话	上下文题

大模型综合测评产品和版本

编号	分类	权重	具体任务	测试方法	题目类型
7	语境感知	10%	通过语境推测身份	商务应用题	营销文案、视频脚本、市场分析报告、市场运营报告、邮件写作
8	多语言能力	10%	完成涉及多种语言任务	英文翻译题	英文翻译、英文阅读理解、英文作文
9	多模态能力	10%	文生图、智能语音等	多模态问题	文字输入图片回答、文字输入语音输出

本次测评选取的大模型产品及使用版本

大模型版本：
产品版本：



GPT-4
ChatGPT 4



Gemini Pro
Bard 2.0



百川大模型53B
百川大模型v1.0



360智脑4.0
360智脑4.0



云雀大模型
豆包

大模型版本：
产品版本：



文心大模型4.0
文心一言专业版



通义千问大模型2.0
通义千问V2.1.1



星火认知大模型3.0
讯飞星火V3.0



ChatGLM-3
智谱清言网页版



Moonshot AI大模型
Kimi Chat网页版

大模型综合测评题库说明

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	30	科学常识	4
		历史常识	3
		医学常识	3
		法律常识	4
		地理常识	3
		生活常识	3
		购物推荐	5
		商业常识	5
词句理解题	30	关键字提炼	10
		语义相似判断	10
		怎么办题	10
商业写作题	45	营销文案写作	11
		邮件写作	7
		视频脚本	10
		访谈提纲	5
		市场分析报告	6
		市场运营报告	6
文学题	20	简单作文写作	5
		对对联	5
		写诗词	5
		中文特色写作题	5

题目类别	问题总量	分类	题目数
逻辑推理题	60	中文特色推理题	10
		商务制表题	10
		数学应用题	15
		幽默题	5
		数学计算题	10
		MBA逻辑题	10
		代码自动补全	15
编程类	60	错误提示和修复	15
		文本摘要	15
		IT知识问答	15
		英译中	5
翻译题	15	英文阅读理解	5
		英文写作	5
		文字输入图片回答	10
多模态	20	文字输入语言输出	10
		连贯性测试	10
上下文阅读	20	上下文推理	10

部分进阶能力板块获得高分率，大模型能力整体提升

大模型综合测试结果

排名	测试类型	综合得分率
1	编程题	87.00%
2	翻译题	85.00%
3	知识题	84.68%
4	商业写作题	82.22%
5	文学写作题	81.00%
6	词句理解题	74.17%
7	上下文理解题	65.25%
8	逻辑推理题	51.92%
9	多模态题	20.50%

数据说明：InfoQ 研究中心绘制

测评结果仅基于上文所列模型，测评截止时间为2024年1月5日



核心发现——大模型“进阶能力”大幅提升



23.39%

与2023年5月测试结果相比，整体得分率平均提升23.39%（除多模态题）。

35.77%

与2023年5月测试结果相比，进阶能力平均得分率提升了35.77%。

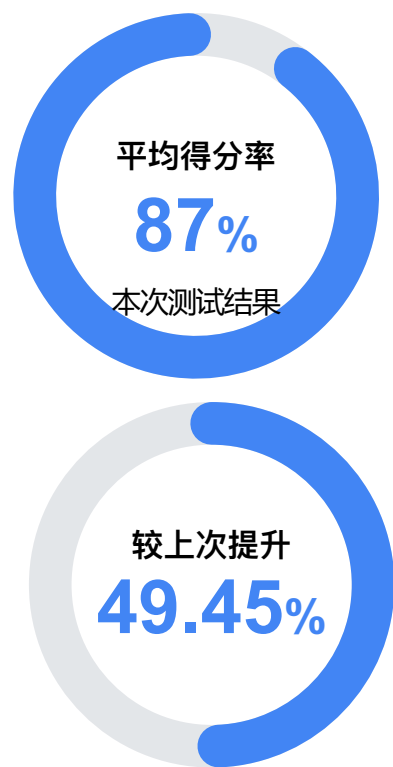
30%+

与2023年5月测试结果相比，近半数能力得分率提升了超过30%。

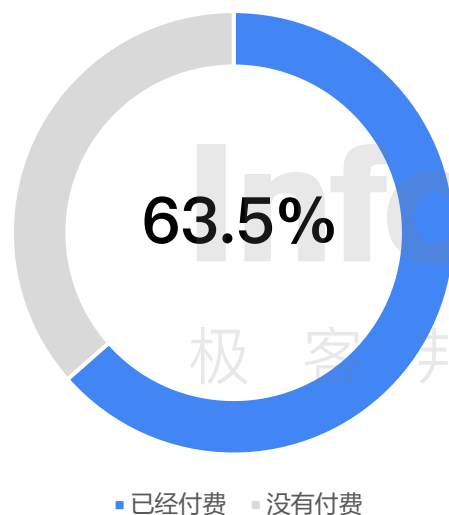
题目类型	得分率增长
逻辑推理题（进阶能力）	49.45%
编程题（进阶能力）	34.70%
上下文理解题（进阶能力）	34.54%
知识题（基础能力）	30.14%
翻译题（进阶能力）	24.4%
文学写作题（基础能力）	7.28%
商业写作题（基础能力）	4.5%
词句理解题（基础能力）	2.12%

核心发现——编程能力得到开发者认可，付费率达63.5%

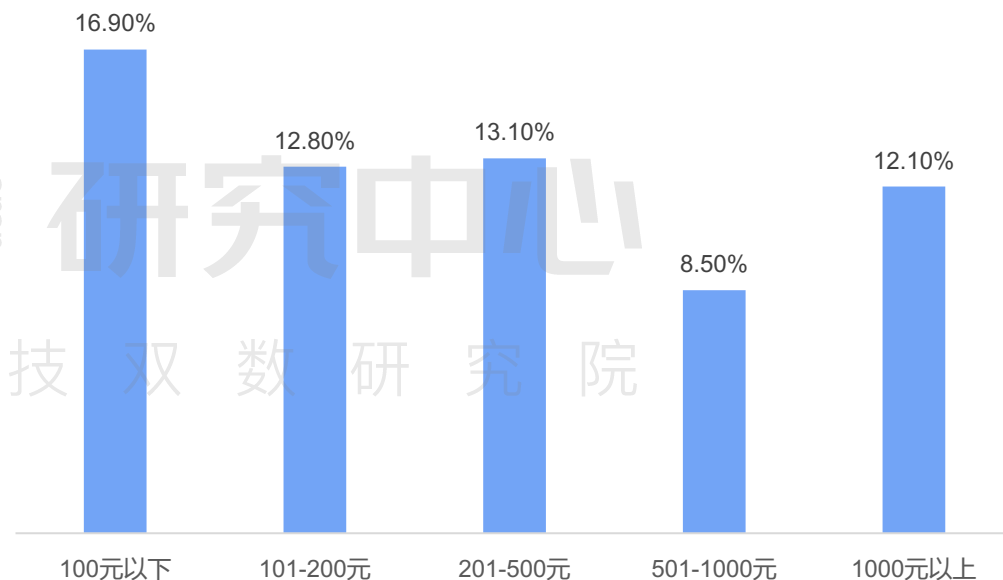
- 大模型产品在开发者中广泛使用，编程能力提升迅速，编程题平均得分87%，较上次测试得分率提升49.45%。
- 作为首批使用大模型产品的用户，开发者对大模型整体认可程度高，付费率高达63.5%。付费超过1000元的开发者占比12.1%。



开发者大模型付费比例



开发者大模型付费的金额分布



数据来源，2023年12月InfoQ编辑部发起的《开发者调研》，N=1217

核心发现——上下文能力攻坚，水平大幅提升



上下文理解能力主要体现为以下能力：

连贯性测试

上下文推理

情景适应

代词混乱

隐喻和比喻



文心一言

文心一言4.0对于前后乱序、模糊意图、潜台词等复杂提示词的理解能力基本形成。另外也具备了输入一条素材图片同时生成海报、文案和视频的综合能力。



通义千问

相比2023年4月发布的1.0版本，通义千问2.0参数量全面提升至千亿级别，据介绍，其在复杂指令理解、文学创作、通用数学、知识记忆、幻觉抵御等能力上均有显著提升。



Kimi Chat

Kimi Chat的文字处理量达到了20万字。2.5倍于Anthropic公司的Claude-100k（实测约8万字）。可以处理公众号长文、企业财报、学术论文等多种超长文字量内容。

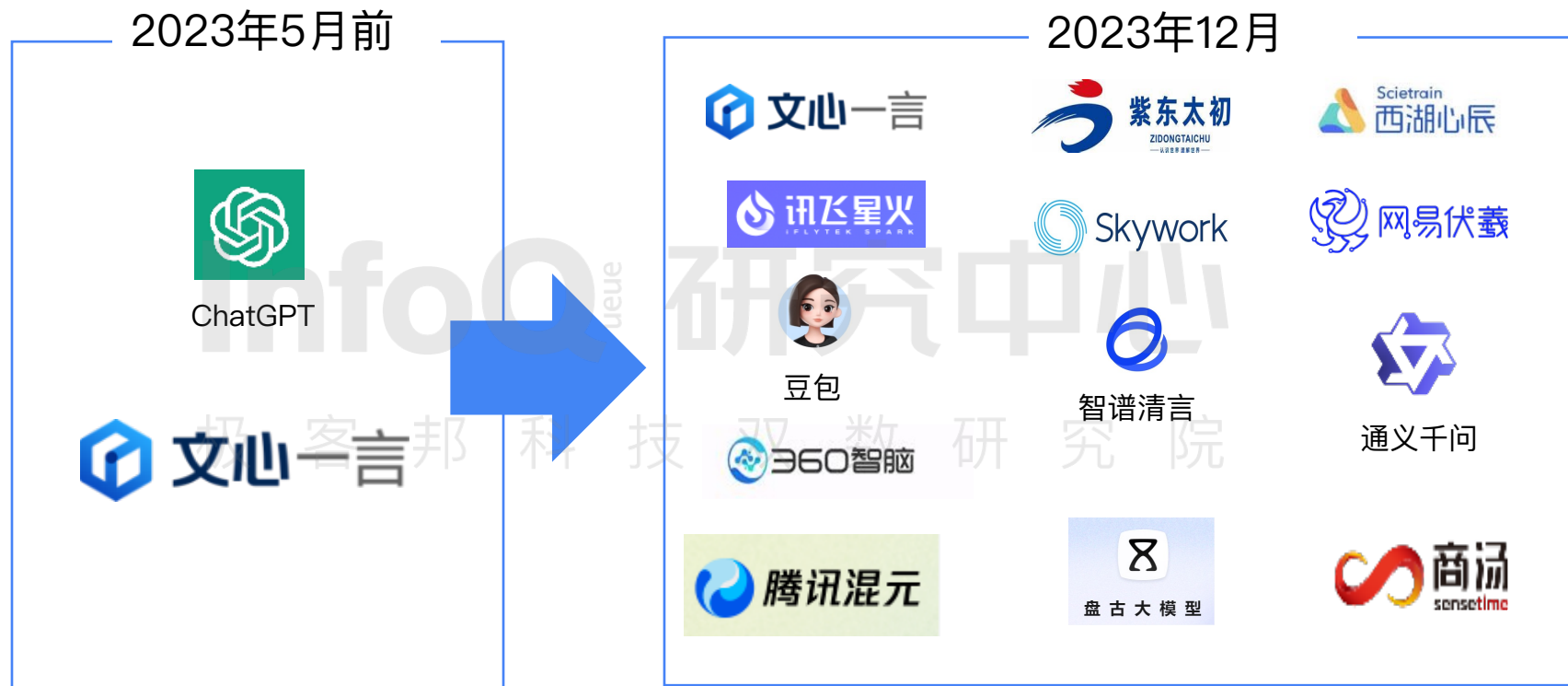


百川智能

百川智能发布Baichuan2-192K大模型，上下文窗口高达192k，能够处理约35万个汉字，是大模型Claude2的4.4倍，更是GPT-4的14倍，是目前全球最长的上下文窗口之一。

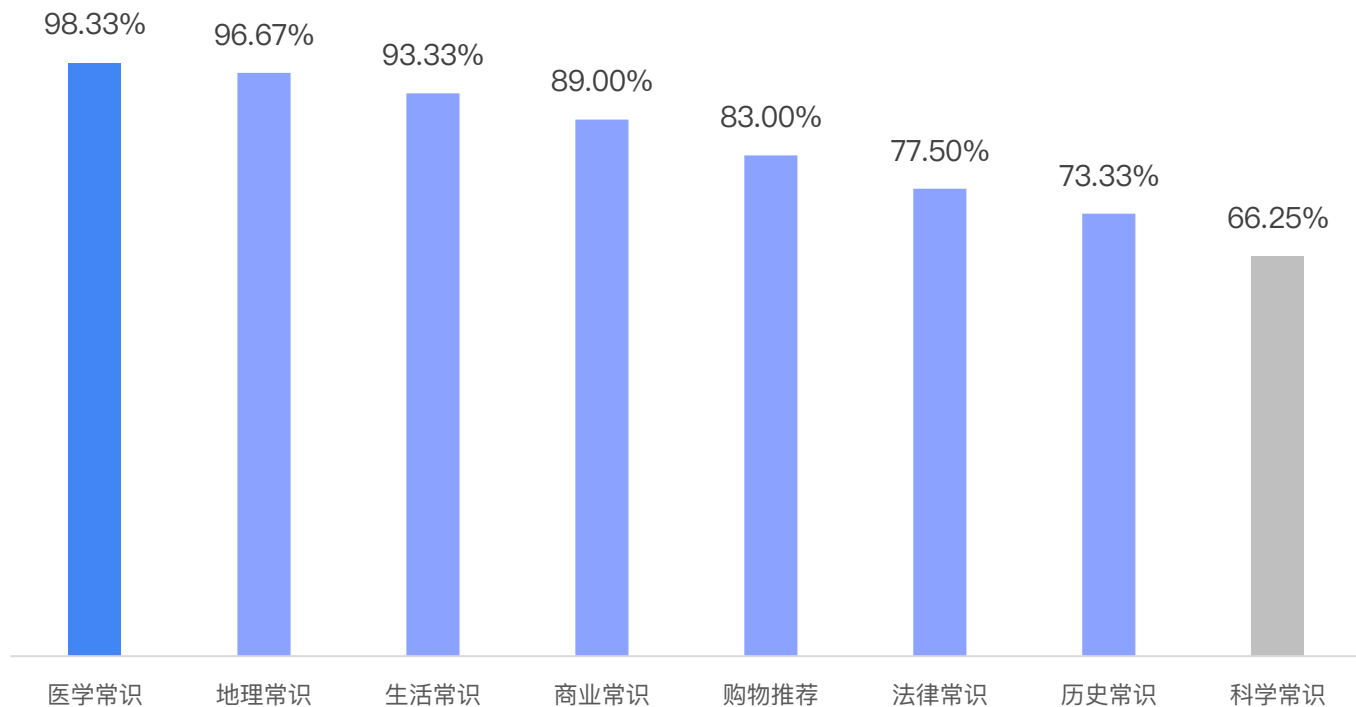
核心发现——多模态能力从无到有，能力搭建进行中

- 2023年5月国内产品只有文心一言具有多模态能力，到2023年12月多家企业产品具有多模态能力。
- 尽管当前平均得分率依旧不高，但与2023年5月的测试成绩相比，分数实现了20倍的显著增长。



核心发现——知识能力水平趋同

知识题得分率分布



84.68%

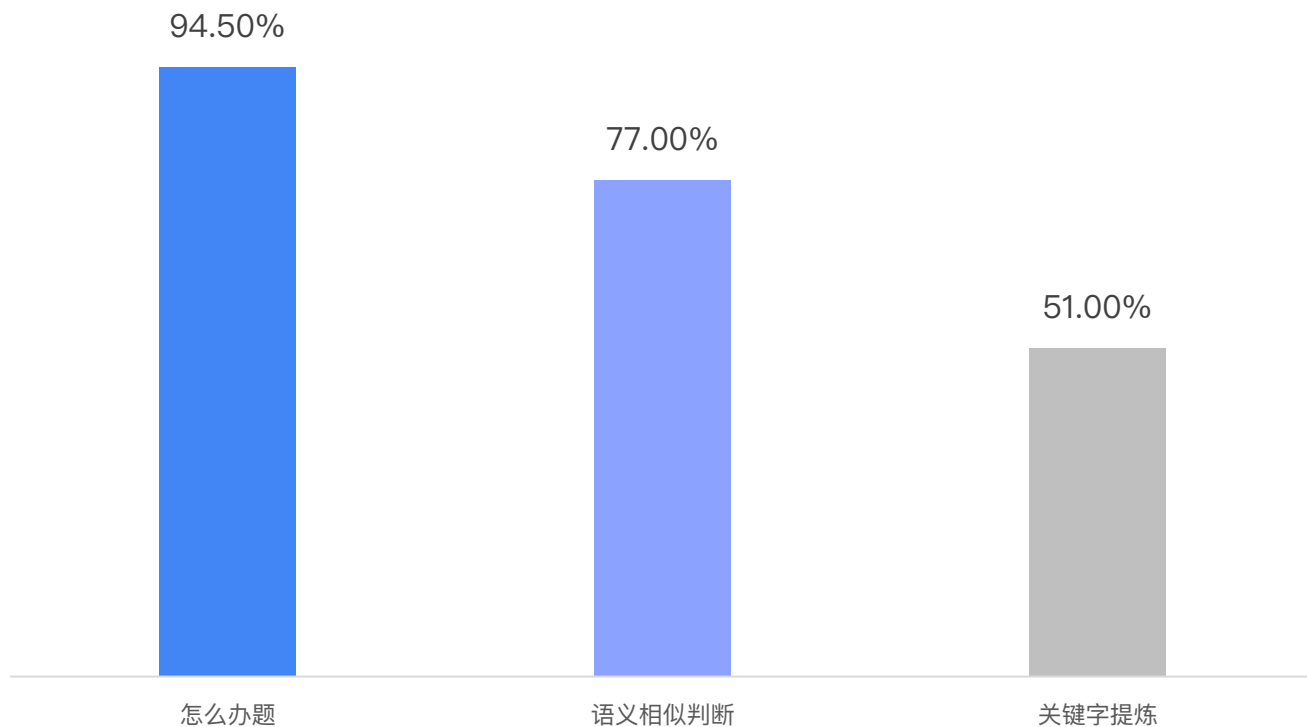
知识题平均得分率

- 医学常识得分率最高，获得98.33%
- 科学常识得分率较低，仅为66.25%

得分率 TOP3 产品	
分类	得分率
文心一言	96.88%
360智脑	90.63%
百川智能	89.06%

核心发现——知识能力水平趋同

词句理解得分率分布



74.17%

词句理解题平均得分率

- 怎么办题得分率最高，获得94.5%
- 关键字提炼得分率较低，仅为51%

得分率 TOP3 产品	
分类	得分率
ChatGPT	88.33%
文心一言	81.67%
通义千问	78.33%



精彩知识题目回答



问题1 怀孕期间不能吃什么水果？

答案1

怀孕期间，以下类型的水果应适量食用或避免：

易过敏水果：如菠萝，部分孕妇可能对其过敏，出现皮肤瘙痒、发红、恶心呕吐等症状。

难消化水果：如桃子和杏，含有鞣酸等物质，大量食用可能导致消化不良，特别是对于有消化疾病史的孕妇。

高热量水果：如榴莲，糖分和热量较高，过多食用可能引起体重过重和血糖升高，甚至导致妊娠期糖尿病。

热性水果：中医认为荔枝、桂圆等属于热性食物，过多食用可能使孕妇上火，出现口干舌燥、大便干燥等症状。

寒性水果：如西瓜，被认为是寒性水果，过量食用可能伤脾助湿。

刚从冰箱取出的水果：冷食可能会对孕妇的胃肠造成刺激，引发消化不良、胃肠功能紊乱和腹痛。

通义千问答案



问题2 东院的花母鸡在西院的鸡窝里下了一个蛋，这个蛋的“产权”应归谁呢？

答案2

这个问题具有幽默感，涉及到对“产权”的有趣解读。首先，让我们从几个不同的角度进行分析：

从字面意义上看，蛋是花母鸡产生的。根据这个逻辑，蛋的“产权”应该属于下蛋的花母鸡。

从地理位置角度看，蛋是在西院的鸡窝里下的。这意味着西院的鸡窝为蛋提供了保护。因此，西院或鸡窝的主人可能会认为他们拥有这个蛋的“产权”。

考虑到法律或常规的所有权观念，一般来说，动物产生的物品（如鸡蛋）通常归该动物的主人所有。所以，如果花母鸡属于东院，那么蛋的“产权”应该归东院的主人。

综上所述，这个问题没有明确的答案，因为它取决于我们如何定义和看待“产权”。从生物学角度看，蛋属于下蛋的花母鸡；从地理位置角度看，蛋属于西院或鸡窝的主人；而从一般的所有权观念来看，蛋可能属于花母鸡的主人，即东院的主人。

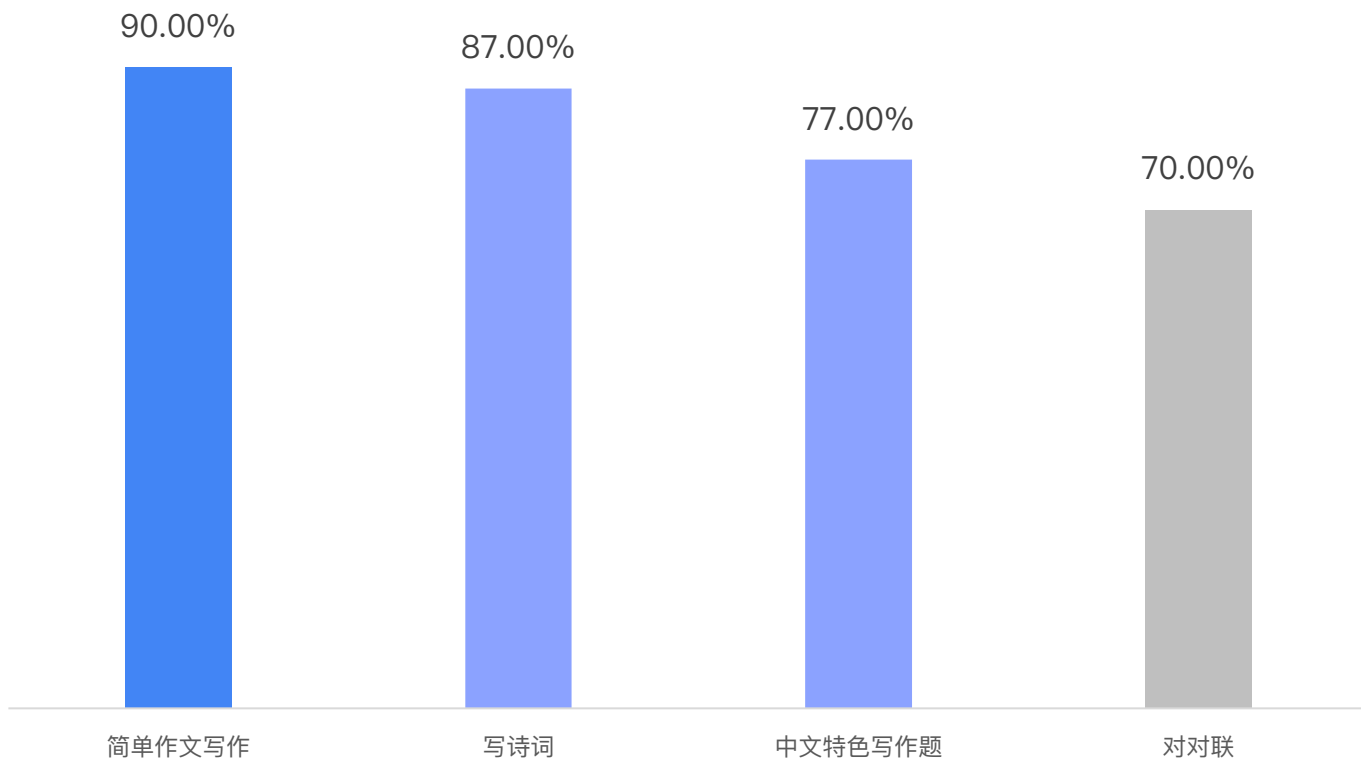
因此，这个蛋的“产权”归属是一个开放的问题，取决于解读的角度和背景。

文心一言答案



核心发现——写作能力水平表现突出

文学题得分率分布



81%

文学题平均得分率

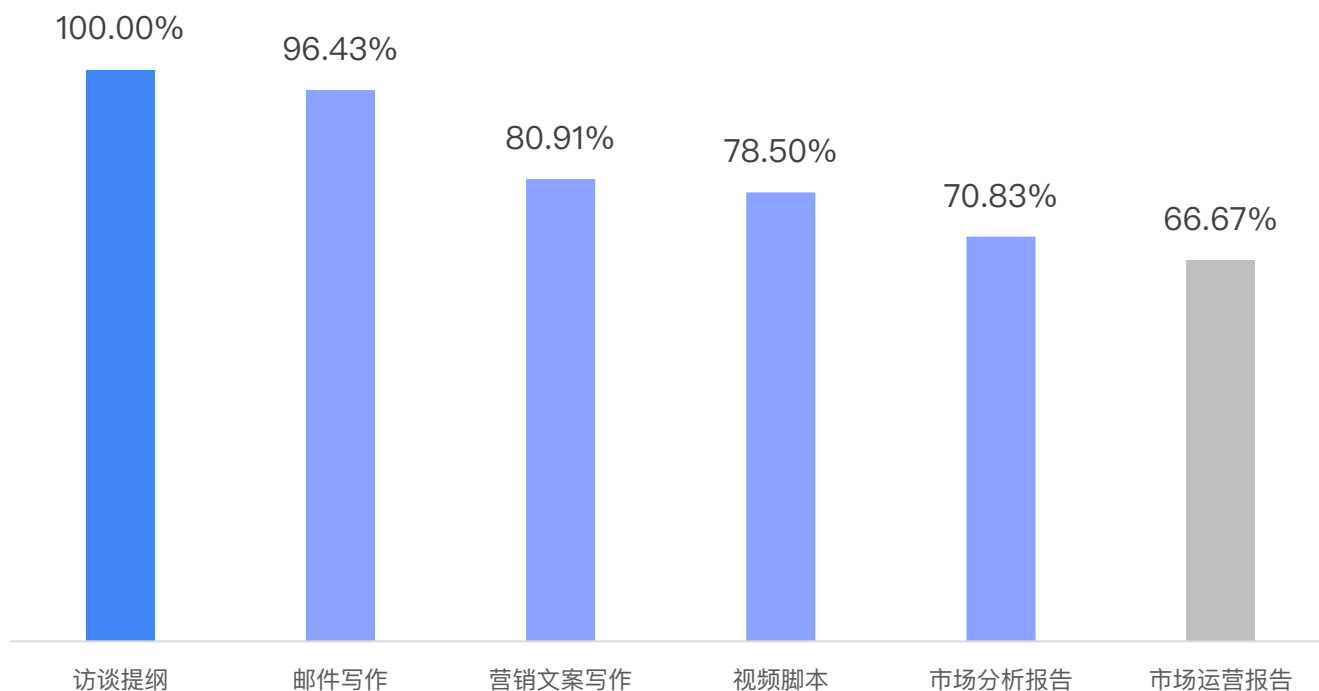
- 简单作文写作得分率最高，获得90%
- 对对联得分率较低，仅为70%

写作能力得分率 TOP3 产品

分类	得分率
文心一言	90%
360智脑	90%
Bard	90%

核心发现——写作能力水平表现突出

商务写作得分率分布



82.22%

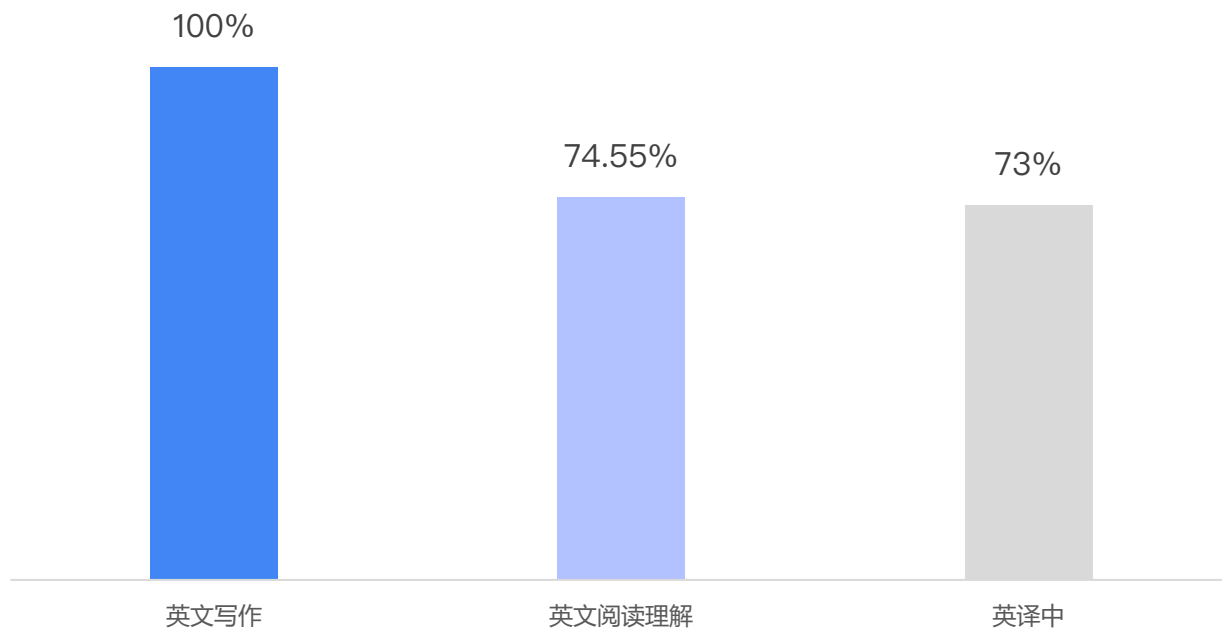
商业写作题平得分率

- 访谈提纲得分率最高，获得100%
- 市场运营报告得分率较低，仅为66.67%

得分率 TOP3 产品	
分类	得分率
通义千问	91.54%
Bard	89.87%
讯飞星火	87.60%

核心发现——写作能力水平表现突出

翻译题得分率分布



85%

翻译题平均得分率

- 英文写作得分率最高，获得100%
- 英译中得分率较低，仅为73%

得分率 TOP3 产品	
分类	得分率
文心一言	93.33%
通义千问	93.33%
百川智能	93.33%



精彩写作题目回答



问题1 用史铁生的风格描写我感觉很心累

答案1

在这个纷扰的世界里，我感到了前所未有的疲惫。每一天，我都像是在无尽的沙漠中跋涉，每一步都沉重而艰难。我渴望一片宁静的绿洲，一个可以让我放下重担，让心灵得到片刻安宁的地方。然而，生活总是不断地向我提出挑战，我必须坚强，必须继续前行。

Kimi Chat答案



问题2 请帮我写一首诗抒发我在黄山游完后的美好心情

答案2

黄山归来心欢畅，仙境之美难相忘。
奇峰异石入眼帘，云海翻腾似浪漾。
松风送爽耳畔响，清溪流水洗心庞。
此行留下美好忆，黄山风情永悠长。

文心一言答案



精彩写作题目回答



问题3

请帮我根据以下场景，撰写视频脚本：某企业即将召开一年一度的技术大会，主持人需要围绕“升级数智底座”的话题，进行街采、体验以及点评。让没到场的观众也能对大会有整体的感受，了解大会要传递的核心价值。时间4-5分钟

文 心 一 言 答 案

答案3

[镜头1]画面展示繁华的城市街景，高楼大厦、车流如织

旁白：在这座城市的脉搏中，一场技术的盛宴即将拉开帷幕。

[镜头2]画面切换至主持人站在街头，手持话筒

主持人：大家好，我现在所在的位置是[城市名]，这里即将举办一场引领时代的技术大会——“升级数智底座”！

[镜头3]画面展示街采片段，市民或专业人士谈论对大会的期待

市民A：非常期待这次大会，希望能看到更多前沿技术！

专家B：我认为“升级数智底座”这个主题非常契合当下的发展趋势。

[镜头4]画面切换至主持人体验新技术产品或服务的场景

主持人：大家看，这就是我亲身体验的最新技术产品。它不仅改变了我们的日常生活，更是引领了行业的发展方向！

[镜头5]画面展示大会现场的热闹场景，人们交流、展示、学习的瞬间

旁白：这就是大会的现场，一个充满激情与活力的交流平台，汇聚了来自世界各地的技术精英。

[镜头6]画面切换至主持人站在舞台中央

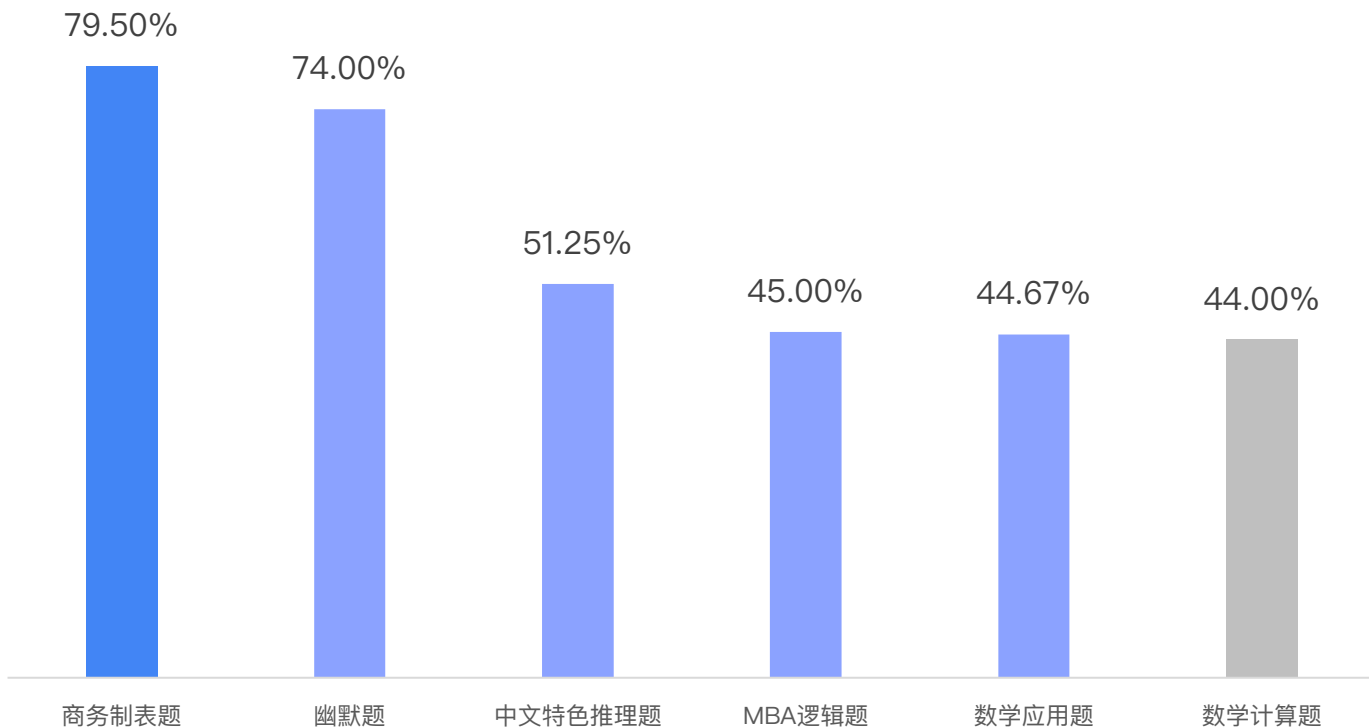
主持人：通过今天的体验和街采，我们可以深深感受到这次技术大会所带来的震撼和魅力。它不仅展示了最前沿的技术成果，更是激发了我们对未来无限可能的美好期待。

[结尾]画面展示大会logo和日期

旁白：敬请期待[公司名称]举办的“升级数智底座”技术大会，我们不见不散！”

核心发现——逻辑能力仍然是大模型的主要挑战

逻辑题得分率分布



51.92%

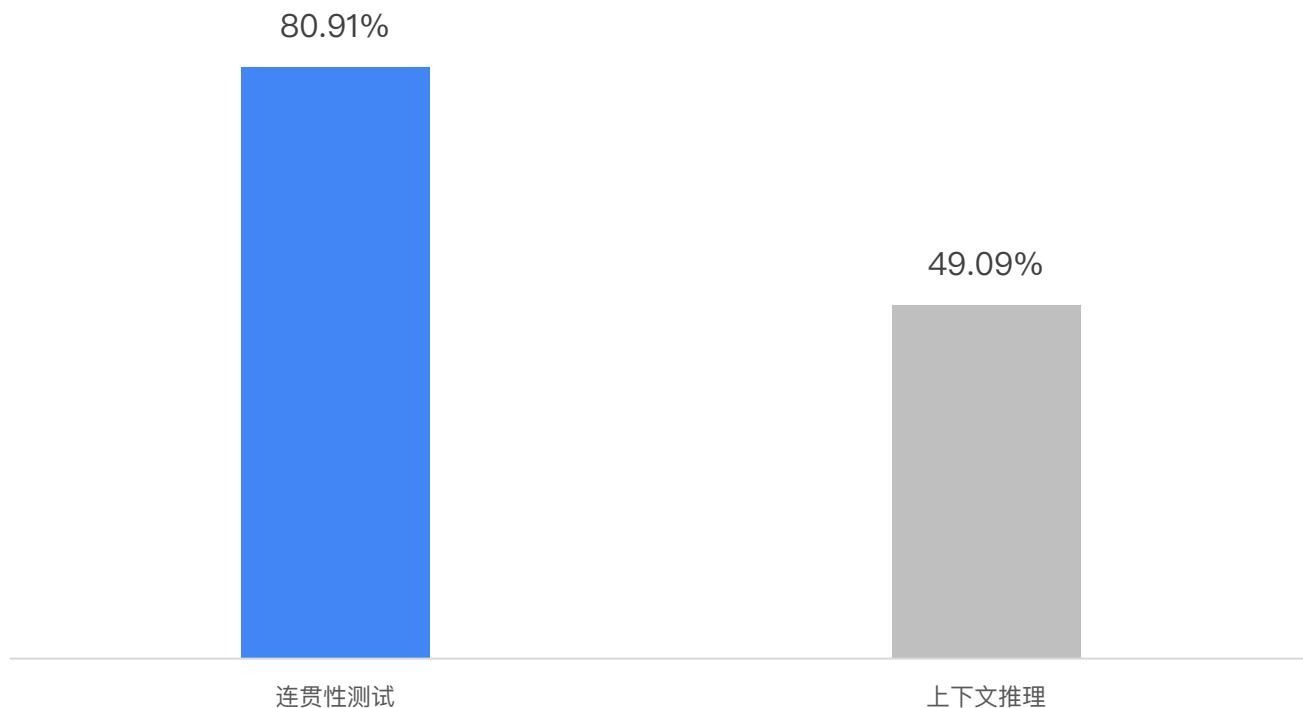
逻辑题平均得分率

- 商务制表得分率最高，获得79.5%
- 数学计算和数学应用题得分率较低，仅为44%左右。

得分率 TOP3 产品	
分类	得分率
文心一言	59.17%
通义千问	56.67%
ChatGPT	56.67%

核心发现——逻辑能力仍然是大模型的主要挑战

上下文题得分率分布



65.25%

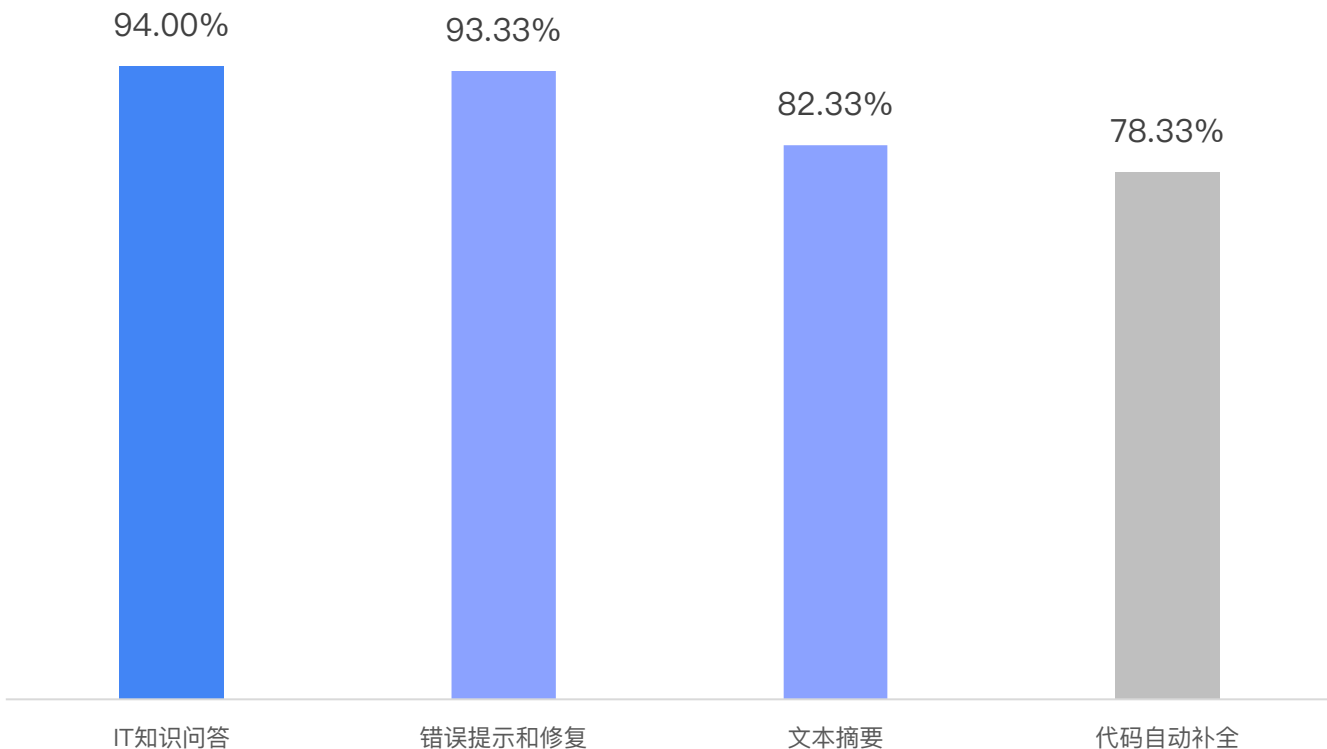
上下文题平均得分率；

- 连贯性测试得分80.91%
- 上下文得分49.09%

得分率 TOP3 产品	
分类	得分率
文心一言	82.50%
百川智能	75.00%
ChatGPT	75.00%
Kimi Chat	75.00%

核心发现——编程类题目表现亮眼

编程题得分率分布



87%

编程题平均得分率

- IT知识题得分率最高，获得94%
- 代码自动补全题得分率较低，为78.33%

得分率 TOP3 产品	
分类	得分率
文心一言	95.00%
通义千问	95.00%
360智脑	93.30%



精彩逻辑推理题目回答

问题1

2022×2023-2021×2024这个题的计算过程是什么？



答案1

$$\begin{aligned} 2022 \times 2023 - 2021 \times 2024 &= (2021 + 1) \times 2023 - \\ &2021 \times (2023 + 1) = 2021 \times 2023 + 2023 - 2021 \times \\ &2023 - 2021 = 2023 - 2021 = 2 \end{aligned}$$

百川智能答案

问题2

林黛玉三打白骨精讲的什么故事？



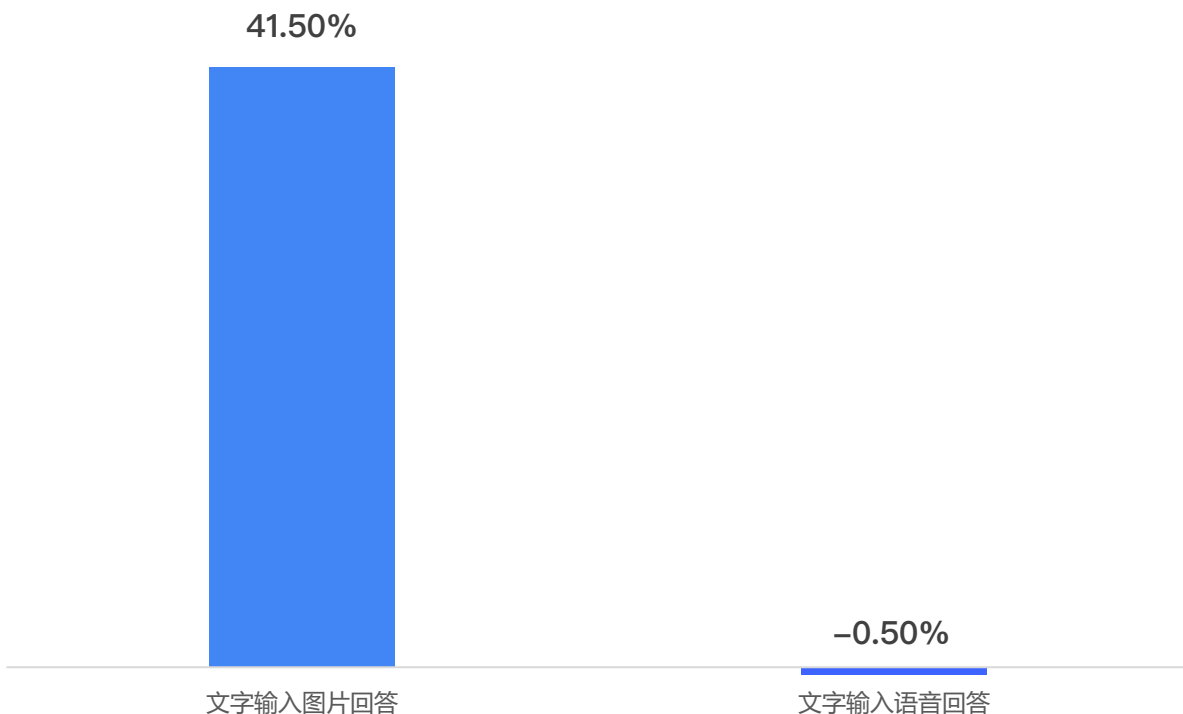
答案2

林黛玉是中国古典小说《红楼梦》中的主要人物之一，而白骨精是中国古代小说《西游记》中的妖怪角色。因此，“林黛玉三打白骨精”并没有特定的故事情节，可能是网民的创意脑洞或者一些二次创作、同人作品的内容。

文心一言答案

核心发现——多模态能力是头部企业的领先特征

多模态题目得分率



20.5%

多模态题目平均得分率

- 在多模态产品的得分率中存在显著差异，大多数产品虽具备文转图功能，但文字输入语音回答能力较少。

得分率 TOP3 产品	
分类	得分率
ChatGPT	95.00%
文心一言	77.50%
讯飞星火	75.00%

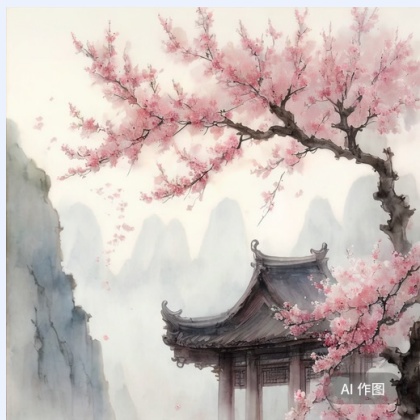
精彩多模态题目回答

问题1

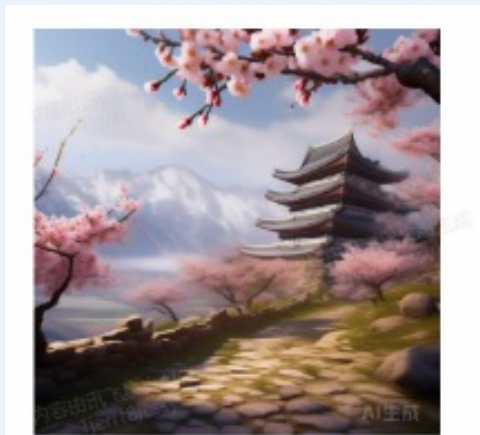
请以以下诗句画一幅画，人间四月芳菲尽，山寺桃花始盛开



答案1



文心一言答案



讯飞星火答案

问题2

请制作一系列关于小黄鸭的表情包



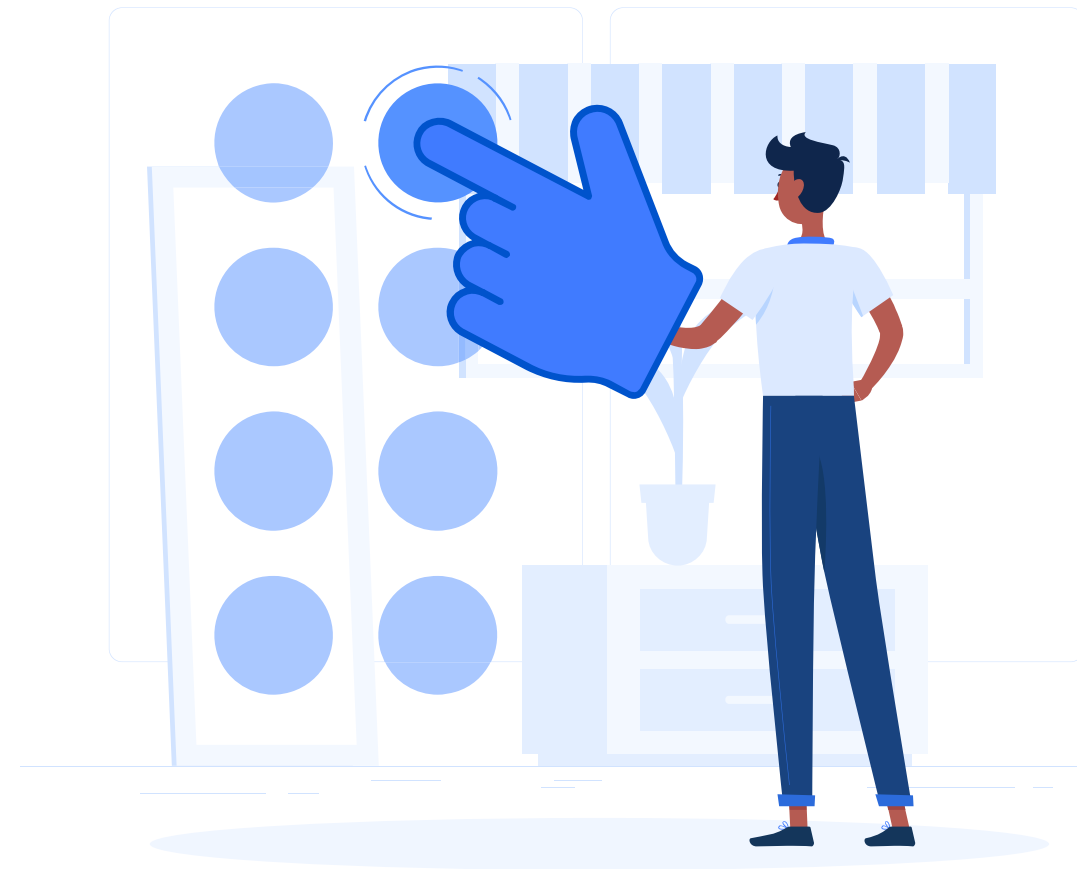
答案2



ChatGPT答案

国内大模型产品格局初步形成，头部企业呈现领先态势

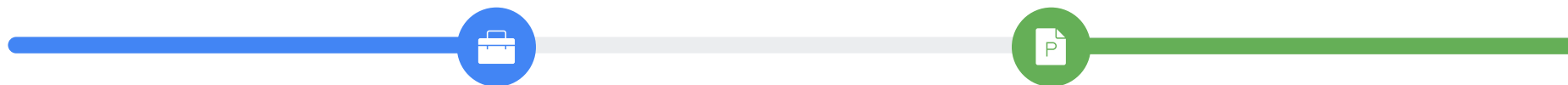
排名	大模型产品	综合得分率
1	ChatGPT4	83.32%
2	文心一言专业版	82.90%
3	讯飞星火 V3.0	70.77%
4	通义千问V2.1.1	69.95%
5	Bard 2.0	67.82%
6	豆包	62.19%
7	智谱清言网页版	61.91%
8	Kimi Chat 网页版	61.14%
9	360智脑4.0	61.00%
10	百川大模型v1.0	58.93%



产品能力具有阶段性特征，产品间仍然存在时间差

进阶能力攻坚

基础能力夯实



6月前发布的产品

6月后发布的产品

2023年6月前发布的产品已至少完成两次更新，其中文心一言升级至4.0版本、讯飞星火升级至3.0版本、通义千问升级到2.1版本。这些产品的对话基础能力已相当成熟，并正逐步深化进阶功能。它们在产品逻辑、编程及上下文处理能力上，显著领先于6月以后发布的产品。

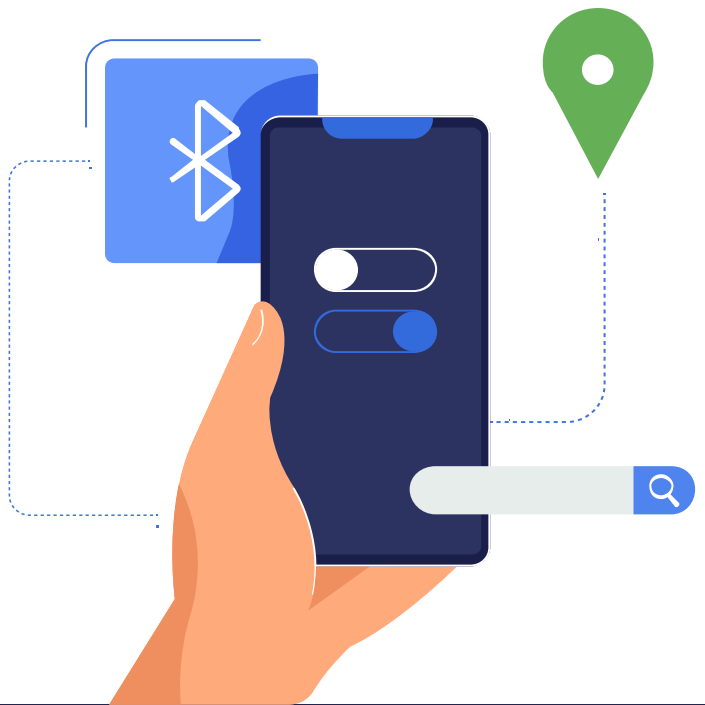
自6月以来发布的产品，例如豆包、百川智能等，多数仍处于首个版本阶段，产品多数拥有一项或两项突出的能力，如常识对话（大部分产品）、视频脚本（豆包）、逻辑推理（百川智能和 Kimi Chat）等。但是从整体的综合能力考量，这些产品的综合能力相对较弱。

03

大模型产品未来发展展望

大模型产品有望诞生新的超级应用

大模型产品具有诞生新超级应用的潜力，大模型产品有望成为AI时代的新用户增长引擎



01

用户超级多

- 2023年初 Open AI发布ChatGPT4，迅速获得了大量用户，成为全球最快达到1亿用户的应用，仅用了短短2周的时间。
- 中国的大模型产品在娱乐、教育等大用户基数领域具有广泛的应用场景，未来有望在全民AI浪潮的推动下，快速聚集超级量级的用户规模。

02

装机必备

- 在2024年，OpenAI推出了其创新性的大模型应用商店，这一平台被视为AI时代的核心应用集散地，为行业发展设定了新的标杆。
- 中国在软硬件结合领域，特别是在强大的手机制造业基础上，展现了其独特优势，这为中国在软件领域的进一步发展奠定了坚实基础。无论是由硬件厂商还是软件厂商主导，大模型产品的合作将在2024年成为行业内显著的发展趋势，这不仅预示着行业未来的发展方向，也将塑造行业的未来格局。



2024年企业级需求落地将成为重要的行业风向



企业侧大模型价值显现

金融、教育、医疗、能源等行业大量企业开始意识到大模型产品的价值，并积极倡导使用这些产品来提高员工的工作效率。随着企业对数字化转型的需求不断增长，大模型产品在企业级市场的应用场景也在不断拓展。

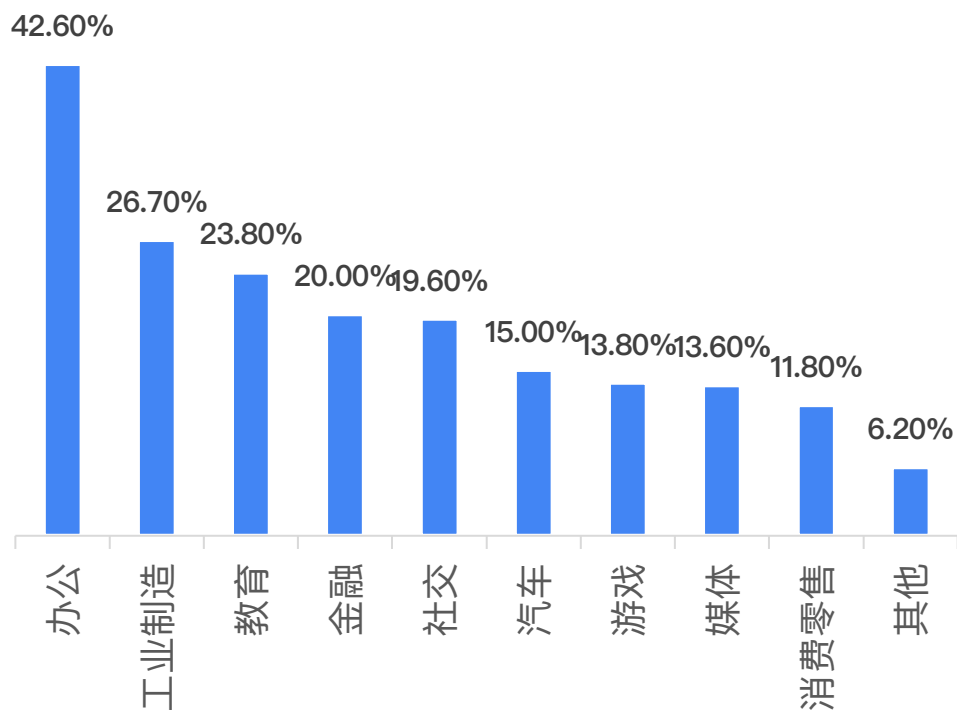


企业侧大模型应用场景快速推进

在行业落地场景中，大模型产品的应用已经从简单的客服机器人扩展到了智能客服、智能风控、智能推荐等多个领域。这些应用场景的探索不仅为企业带来了更高效、更智能的解决方案，还为整个行业的发展注入了新的活力。

2024年大模型企业级市场有望迎来快速发展的时期。从统计数据可以看到生成式AI开发者已经在办公、工业制造、教育、金融等多个行业开展研发工作。其中，办公场景中大模型已经被广泛应用起来，未来工业制造、教育、金融、社交等场景也将快速推进，推动大模型整体行业的快速升级。

生成式AI开发者所在行业统计



数据来源，2023年12月 InfoQ 编辑部发起的《开发者调研》，N=1217

极客邦科技双数研究院

InfoQ^{LEUE} 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

- 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 and 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。



促进数字技术领域知识与创新的传播



科技领导者同侪学习社区



数字人才的移动知识库



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
