



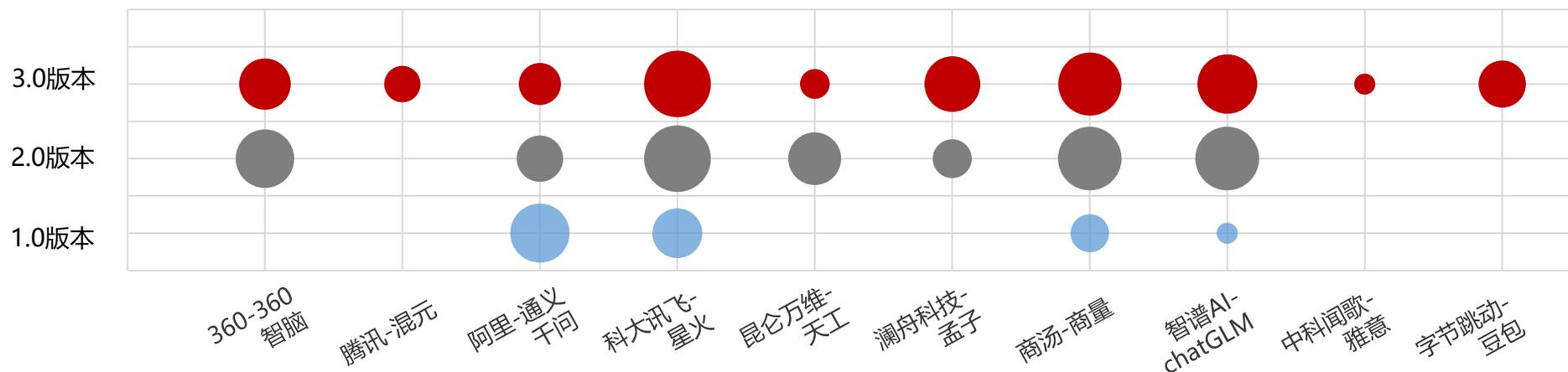
# 人工智能大模型体验报告3.0

新华社研究院中国企业发展研究中心

2023年11月

# 报告回顾：大模型产品市场竞争激烈，技术不断进步

## 三次测评概览



注：气泡越大代表实际测评得分越高，分数越靠前；空白表示该版本下未测试。

从1.0版本到3.0版本的测评过程中，参与厂商的分数变化反映了大模型产品竞争的激烈和技术的进步。可以明显观察到一些厂商在持续进步，而一些新的厂商也在逐渐崭露头角。

- **在1.0版本中（5月测评）**，ChatGPT系列占据领先地位，其中ChatGPT4表现最佳，紧随其后的是ChatGPT3.5。百度文心一言位于整体第三位置，展现出了较强的实力。这一阶段的科大讯飞星火和阿里通义千问也进入了整体前五，国内前三，表现不俗。
- **到了2.0版本（8月测评）**，专注于国内市场，科大讯飞星火实现了跨越式发展，整体表现亮眼。百度文心一言、商汤商量与智谱CHATGLM也取得优异的成绩。同时，360智脑和昆仑万维天工也进入了前五。这些变化展示了厂商之间的激烈竞争和技术进步的速度。
- **到了3.0版本（11月测评）**，科大讯飞星火依旧保持领先，商汤商量紧随其后，稳定发挥。值得一提的是，澜舟科技孟子取得了显著进步，字节跳动豆包作为新面孔，在这次评估中崭露头角。

纵观这三个版本的发展，可以看出一些规律。首先是竞争日益激烈，厂商之间的得分变动频繁，这体现了大模型产品市场的活跃和竞争激烈程度。其次，技术进步明显，从1.0到3.0版本，各个厂商的产品在性能和功能上都有显著提升。最后，新面孔的涌现也展示了这个行业的活力和吸引力，新的厂商不断加入竞争，推动了大模型产品的发展。

# 目录

01

## 大模型产品测评综述

---

- 大模型产品现状与进程
- 3.0版本大模型测评规则

02

## 大模型厂商整体测评

---

- 3.0版本大模型综合指数
- 3.0版本测评细分维度指数及评述
- 测评题目展示

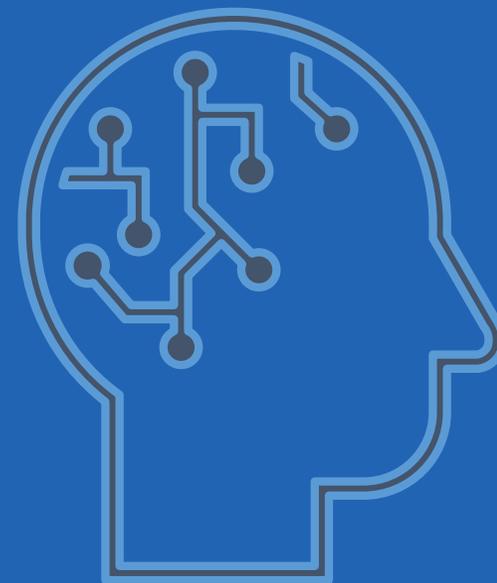
03

## 厂商最佳实践案例

---

- 厂商优秀案例展示

# 01.大模型产品测评综述



# 迭代风潮之下，大模型产品如何迅速适应并引领市场变革

1. 大模型更新迭代速度日新月异，OpenAI 2022年11月发布了GPT3.5，2022年3月就发布了GPT4，参数规模成倍增长，2023年6月百度发布了文心一言3.5，仅仅4个月之后百度又发布了文心一言4.0，基础模型全面升级，其他厂商的产品也在不断升级迭代，大模型的发展速度可谓“日新月异”。
2. 虽然大模型的性能不断提升，但是大模型在快速发展的同时也面临着一些问题——
  - 不稳定：在GPT4上线初期，有大量用户反馈大模型的回答质量有所下降，尤其在程序生成方面，GPT4生成的代码时常出现错误。OpenAI开发者推广大使Logan Kilpatrick也坦言，由于大模型本身存在不稳定性，因此对于相通的提示词，大模型存在回答前后不一致的情况。
  - 幻觉：大模型仍然具有一定局限性，不是完全可靠的，会出现“幻觉”事实并犯推理错误。
  - 安全：安全是重中之重。在模型训练、模型线上推理服务、模型安全测试、模型训练过程对齐、模型生成内容等方面的安全合规能力需要持续加强。
3. 市场需求也在发生变化，ChatGPT上线之初更多的人还是把它当成对话工具，但是不久ChatGPT的功能就得到深度挖掘，如今各个领域的内容创作，以及和行业细分领域深度结合的应用越来越多，市场对于大模型的需求也随着用户所在行业的变化而不断发生变化。

# 历时半年，大模型测评体系再次升级，助力持续发展之路

从历史视角观察，测评1.0与2.0阶段代表着大模型产品的初期评测阶段，主要基于小规模问题集进行评测。这种评测方式在初期对于了解大模型产品的基本性能具有一定的价值。然而，随着时间的推移和技术的进步，课题组发现在经过这两次测评后，众多厂商开始根据测评题目进行定向优化。通过针对测评数据集特定调整，以获取相对较好的排名。

此次推出的大模型测评，其核心目标是获取客观、真实的产品结果，并希望通过评测向客户反馈产品真实体验状况，并推动产品的实际进步。针对目前市场上出现的问题，研究团队认为，亟须对评测方式进行优化，确保评测能够全面反馈产品现状。为确保评测能够全面反馈产品现状，课题组实施了以下策略来调整和优化评测体系：

## 评估维度立体化

多维度全面考察，立体展现模型性能，避免单一指标偏颇。从数据准确性、模型稳定性、到算法效率，全方位、多角度深入评估，实现真实且全面的性能展现。

## 主观评价融合客观结果

既注重客观数据衡量的精准性，又兼顾用户主观感受的真实性，以用户为中心，让产品的优质体验真正落地，助推质量提升。

## 指标随市场情况动态变动

随着市场环境的变迁和技术进步，及时对评估指标进行更新与调整，确保评估体系始终保持与市场需求的紧密关联。