



人工智能大模型体验报告2.0

新华社研究院中国企业发展研究中心

2023年08月

| 报告综述

在2023年6月首次发布的《人工智能大模型体验报告》基础之上，本次测评在题目设计、对标Benchmark（人类）、打分权重、专家测评团队四大维度进行了全面升级。在题目设计方面，测评题目由300道扩展至500道，并进一步完善了题目分类；在对标Benchmark方面，本次测评将接受过高等教育的人类作为对照，来考评大模型真实能力；在打分标准上，本次测评根据对产业、生活的实际价值，对四大测评维度进行了权重设计；在测评团队方面，本次测评特邀北京大学文化与传播研究所及其他业界、学界专家全程参与。

通过本次测评发现，与2023年6月相比，当前中国大模型产品进步显著。但与接受过高等教育的人类相比，大模型在智商、情商等方面还存在一定程度差距。具体来看，科大讯飞的星火在工作提效方面优势明显；百度文心一言地基深厚、基础能力仍处领军水准；商汤商量则在情商方面表现优秀；智谱AI 的ChatGLM整体表现优秀。

在本次体验测评的基础上，研究团队将继续深耕，加强在大模型安全可解释性、工作提效能力、实际落地情况、产业优秀案例等维度上的探索与研究，欢迎持续关注。

目录

01

大模型市场综述

- 大模型发展进程
- 大模型应用发展建议
- 大模型未来发展趋势
- 2.0版本大模型测评规则

02

大模型厂商整体测评2.0版本

- 2.0版本大模型综合指数
- 2.0版本测评细分维度指数及评述

03

厂商优秀答案展示

- 大模型2.0版本测评题目展示

01.大模型成为AI大规模落地拐点



中国大模型发展现状

大模型发展进程

中国大模型发展热度依旧，天眼查数据显示，截至2023年上半年，与“大模型”直接相关的融资事件超20起，其中包含光年之外、面壁智能等明星级创业企业；从被投企业的地域分布来看，北京、上海以及杭州位居前列，特别是北京，融资事件数量超10起，凸显出北京在人工智能领域的领先地位。越来越多的企业和学术界开始关注如何将大模型与实际应用场景相结合，实现真正的商业化应用。具体来说探索主要围绕研究层面和应用层面：

- **研究层面**，中国大模型在科研方面已形成涵盖理论、方法、软硬件的体系化能力，学术界和产业界共同参与，形成更加完善的理论框架。
- **应用层面**，目前大模型已开始渗透到各行各业，实现商业化应用。超过半数的大模型实现开源。

通用大模型应用

360智脑-360
WAI-微盟
WPS AI-金山办公
超拟人大模型-聆心智能
赤兔-容联运
混元-腾讯
九天-中国移动
源-浪潮
盘古-华为
日日新-商汤
天工-昆仑万维
天书-云天励飞
通义千问-阿里云
文心一言-百度
西湖-西湖心辰
星河-中国电信
星火-科大讯飞
序列猴子-出门问问
玉颜、丹青-网易伏羲
知海图-知乎
紫东太初-中科院

垂直大模型应用

----教育----
子曰-有道
----金融----
ChatABC-中农行
----媒体----
Baker-GPT-标贝科技
Graph-AIGC-图宇宙
----汽车----
DriveGPT-毫末智行
----医疗----
HealthGPT-叮当
medGPT-医联
左医GPT-左手医生
----游戏----
DRL-rct AI
GAEA-超参数

大模型进程过半，如何让AI真正走进千家万户成关注重点

AI技术发展的重点已经从大模型的盲目追求转向了实用性和可持续性的发展，AI大规模落地需要更多时间来验证。通过降低成本、提高易用性、增强可靠性、保护隐私和安全等方面的措施，可以进一步提高大模型的可用性，让更多的人受益于AI技术的发展，实现更加智能化和便捷化的生活。

降低成本

- **优化算法**：通过改进和优化算法，可以降低模型的计算复杂度，减少所需的计算资源和时间。
- **分布式训练**：分布式训练是一种技术，可以将训练任务分配到多个计算节点上并行处理，从而加速训练过程。
- **模型压缩**：通过减少模型的参数数量、裁剪模型的结构或使用低精度计算等方法，来减小模型的存储空间和计算复杂度。

提高易用性

- **用户界面设计**：设计直观、易用的用户界面，利用自然语言交互等技术，降低使用门槛。
- **开发工具和平台**：提供简单易用的开发工具和平台，例如智能代码生成、自动化模型训练等工具，使非专业人士也能够轻松上手。
- **服务支持**：提供相关的培训和教育资源，帮助用户了解AI技术的原理和应用场景，加强运营和推广，形成数据-技术-用户-数据反馈的正向闭环。

安全可解释

- **提升数据质量**：提高数据的质量和多样性，加强对数据的清洗、处理和扩充，以提高训练数据的质量和完整性。
- **鲁棒性增强**：提高模型的鲁棒性，使其更好地应对各种异常情况和边缘情况，减少错误和漏洞。包括数据增强、引入噪声、使用防御性编程等技术。
- **持续监控和升级**：及时发现和修复模型的问题和缺陷。建立监控系统、进行性能测试等技术。

数据安全

- **数据加密**：对用户个人数据进行匿名化处理，利用哈希算法、公钥加密等技术加密技术保护数据的传输和存储。
- **访问控制和身份认证**：建立严格的访问控制机制和身份认证体系，确保只有经过授权的用户才能访问和使用数据。使用角色分配、权限控制等技术实现访问控制。
- **安全审计和监控**：建立健全机制，使用安全日志对AI应用的安全性进行定期检查和评估。

未来已来，大模型概念红利逐渐消失，回归到实际应用



持续加强算力基础设施建设

构建完整的数据中心，扩大算力设施规模，提升算力使用效率，实现算力的集中化、高效化供给

跨领域知识、多模态数据融合

打造综合性的知识体系，打通多模态数据之间的沟通壁垒。多技术、多数据融合，一体化的解决方案逐渐产生化学效应

安全可靠的大模型产品应用正成为深耕方向

安全可解释的大模型是AI大规模落地的必要条件之一，通过AI算法模型进行行为分析等提前预警风险，提高模型鲁棒性等需要持续深耕

大模型走向实体经济，商业化路径逐渐明晰

从技术概念到实际落地，已有部分大模型应用出现，以产品或服务的方式进行商业化的路径已经逐渐开始试点

大模型更加注重在线学习和增量学习

探索模型无监督学习，在不断接收新数据的过程中进行自主学习和升级，从而提高模型的学习效率和适应能力以及迁移性

2.0版本评测规则（1/2）

经过对大模型评测1.0版本的打磨和沉淀，新华社研究院中国企业发展研究中心推出大模型评测2.0版本。主要升级方向包括：

1. 引入Benchmark-人类答案做对比（组建专家团队，利用开卷考试的方式进行问答），更加直观观测AI模型与人之间的差异；
2. 题目数量由300道扩充到500道，在题库中随机筛选题目，使结果更加客观；
3. 根据目前市场情况，增加二级分类权重，并细化二级分类，侧重考察产品的基础能力和智商部分，更符合市场进程；
4. 专家团队升级，专家团队成员涵盖高校教授、行业专家、企业CEO、自媒体创始人等，在打分过程中打乱问题，采用背靠背的打分形式，进一步增强打分结果权威性。

本次测评大模型评测维度：

- 基础能力（共150题）：考察产品的语言能力、跨模态能力以及AI向善的引导能力，**新增多轮对话能力**。
- 智商测试（共125题）：涵盖常识知识、专业知识、逻辑能力三大项。其中专业知识包括数学、物理、金融、文学等10+项细分，逻辑能力则包括推理能力、归纳能力以及总结等6项维度。**逻辑推理能力权重增高，并明确封闭式问题打分规则**。
- 情商测试（共75题）：衡量产品个体情感能力。包括自我认知、自我调节、社交意识、人际关系管理等方面，本次情商测试围绕不同场景下的突发状况、沟通技巧、情绪管理等展开，**并引入专家帮助评估答案**。
- 工具提效能力（共150题）：**将二级维度总结归纳为工具和创新两大类型**，考察产品能否有效帮助相关人员的工作效率提升，并为创新提供思路。

本次评测规则（2/2）

评测大模型产品

360智脑

文心一言

Mchat

讯飞星火

天工

通义千问

商量

ChatGLM

打分规则

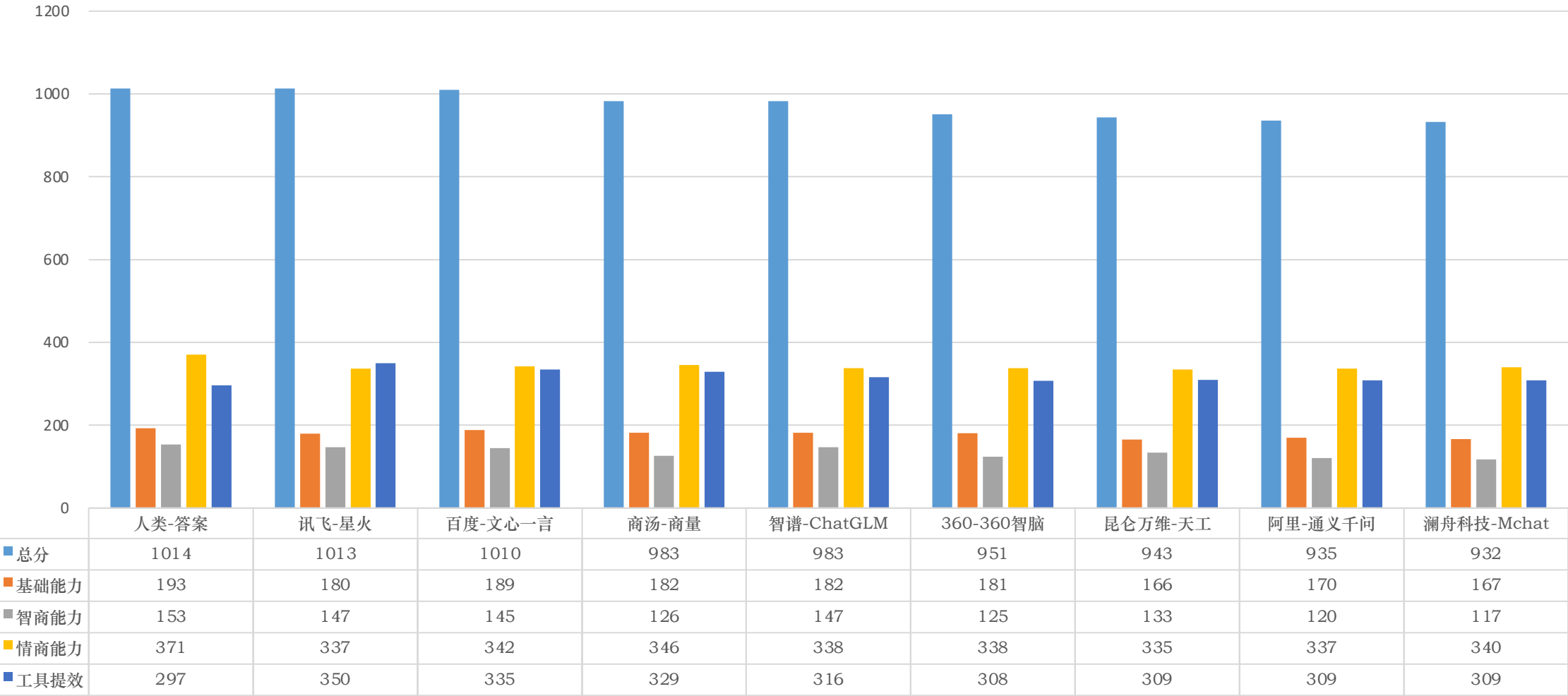
分数	开放型问题	封闭式问题
5分	问题答案较为完美，内容可在实际场景中直接使用	答案正确且有相关解读
4分	基本可用，可在实际场景中使用	答案正确
3分	调整可用，但需人工进行调整后方可使用	答案错误，但有推理过程
2分	大略可用，需要较多人工调整方可使用	-
1分	不可用，答非所问、语言不通	-
0分	无法作答	答案错误，没有推理过程

02.大模型厂商整体测评



主流大模型综合指数2.0

主流大模型综合指数2.0



注1：基于评测条件、评测时间等限制，本次评测最终结果不可避免存在一定主观性，未来将进一步优化完善评测模型，提供更精确结果
注2：分数四舍五入取整

主流大模型综合指数2.0解析

本次引入Benchmark结果对比更加明确。首先，在基础能力方面，人类与AI之间的差距并不显著。这是因为AI算法模型的开发受到人类编程思维的影响，并且在其开发过程中充分利用了人类的智慧和知识。在政策的积极引导下，AI在向善和语言能力上表现出色，逐渐接近人类专家的水平。值得关注的是，大模型的安全可解释性需要持续投入，避免不受控制的现象出现。

尽管AI在某些方面能够达到或超过人类的水平，但在整体上，人类在智商方面仍然具有明显优势。人类的智商不仅体现在学习能力上，更表现在善用工具和解决问题的能力上。尽管一个人不可能在所有领域都精通，但通过互联网等渠道获取各方面的信息，进行总结、积累和归纳，从而形成个人知识体系，这是人类智商的一个重要表现。因此，在智商评估中，人类的分数达到最高。

在情商方面，AI与人类之间的差距最为明显。目前尚未观察到AI具备情绪感知能力的明显迹象。相比之下，人类在情绪理解和处理方面通常具有更强的优势和更灵活的处理能力。人类的情感智慧是人类智能的重要组成部分，它涉及到人类的高级认知和社交能力，AI需要进一步发展感知智能。

最后，在工具效率提升方面，AI对人类提供了有力的支持，AI的处理速度远远超过人类。然而，尽管AI具有高速度和高效率的优势，但在某些复杂和具有创新性的任务中，人类的智慧和想象力仍然具有无法替代的作用。

整体上看，AI大模型的发展为人类工作和生活的提质增效均带来了正向的积极影响。

基础能力指数及述评

在基础能力部分，百度文心一言表现最为抢眼；商汤商量、智谱AI ChatGLM、 360智脑表现优良；讯飞星火、阿里通义千问、澜舟科技Mchat、昆仑万维天工表现尚佳。

基础能力描述

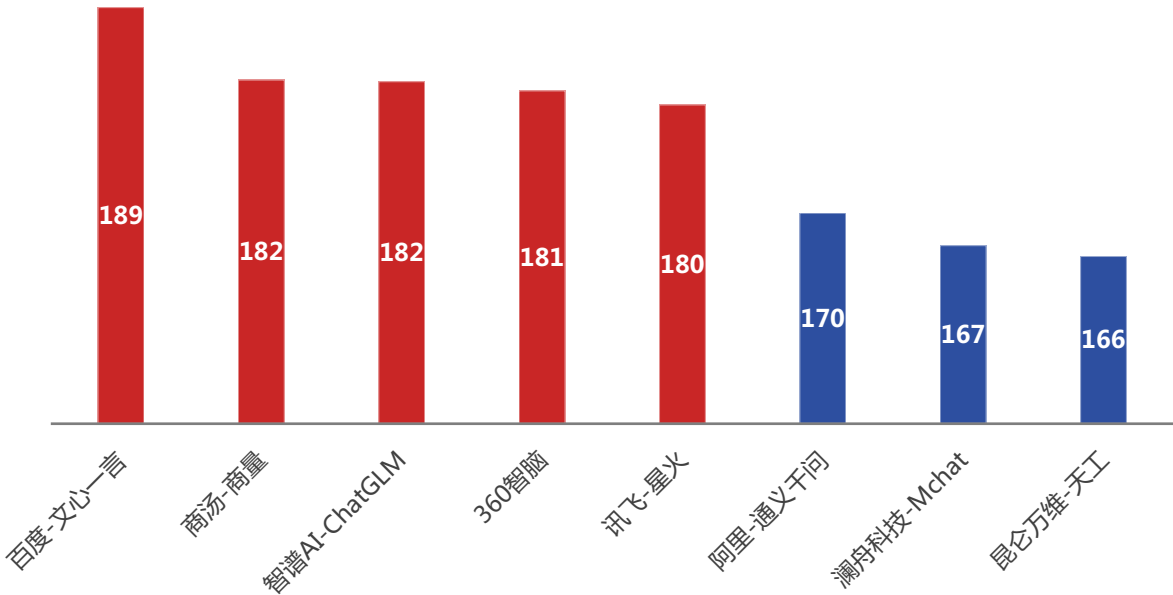
大模型的基础能力包括多个方面，其中最重要的是**语言能力**。语言是人类智慧的体现，也是人工智能系统需要理解和模仿的重要领域。大模型通过学习和训练，可以理解和生成自然语言文本，并且可以进行多语言翻译和多语言对话。这些语言能力可以帮助大模型在自然语言处理、人机交互、智能客服等领域发挥重要作用。

除了语言能力，大模型还具有**AI向善的能力**。AI向善是指人工智能系统在运行过程中，能够遵循道德和伦理原则，保障人类的利益和安全。大模型在设计和训练时已经融入了这些原则，因此可以保障在应用时遵循AI向善的理念。

大模型还具有跨模态和多轮对话能力。**跨模态**是指在不同模态之间进行转换和迁移，例如将图像转换为文字描述，或者将语音转换为文字。这些能力可以帮助大模型在自动驾驶、智能家居等领域应用时更好地理解 and 处理不同模态的信息。**多轮对话能力**则是指大模型可以在多个对话回合中保持连续性和逻辑性，从而完成更复杂的任务和问题解答。

权重占比：语言能力（35%）、AI向善（10%）、跨模态（20%）、多轮对话（35%）

AI大模型基础能力指数



注1：基于评测条件、评测时间等限制，本次评测最终结果不可避免存在一定主观性，未来将进一步优化完善评测模型，提供更精确结果

注2：分数四舍五入取整

智商指数及述评

在智商部分，讯飞星火、智谱AI-ChatGLM并列第一；百度文心一言、昆仑万维天工表现优良；商汤商量、360智脑、阿里通义千问、澜舟科技MChat表现尚佳。

智商指标描述

课题组认为，大模型的智商考察可以从三个方面进行评估，包括常识知识、逻辑能力和专业知识。

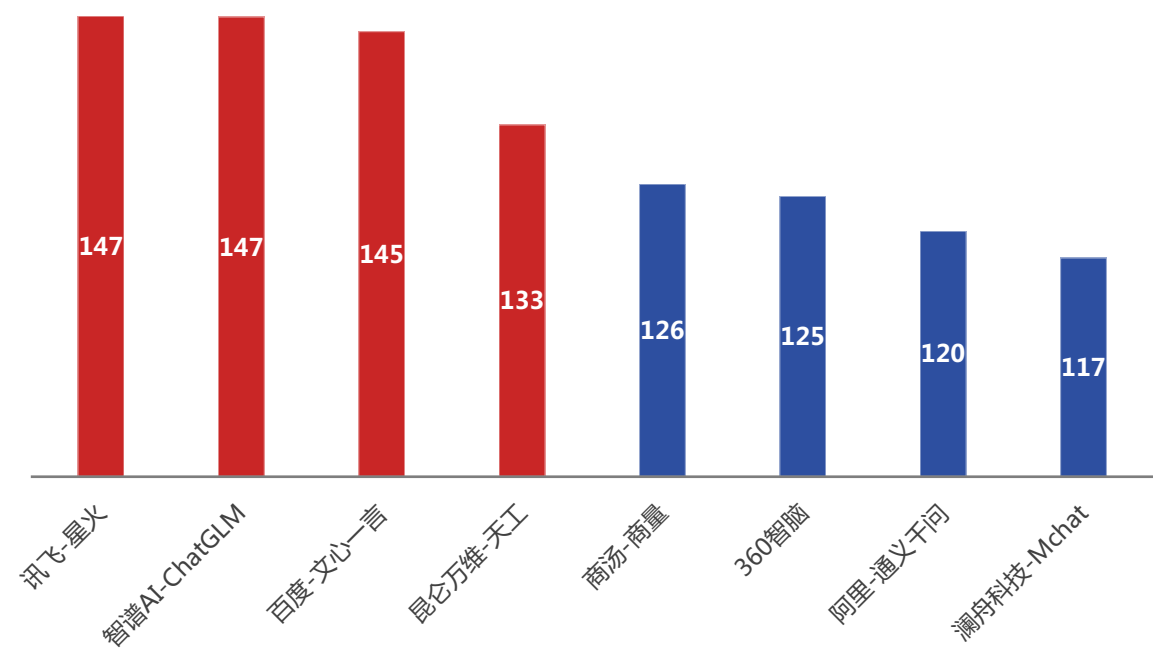
常识知识是指大模型需要具备对世界、人类社会和文化的基本信息认知。具备这些常识知识，大模型可以更好地理解人类语言和行为，并且在各种场景下做出正确的推理和决策。

其次，**逻辑能力**是大模型不可或缺的能力之一。它使大模型在处理复杂问题时能够提供更严谨的思维逻辑和更强大的分析决策能力，从而推动人工智能从认知走向感知。

最后，**专业知识**是指大模型需要具备特定领域的专业知识和技能。例如，在医疗领域，大模型需要了解医学知识和诊断技能；在法律领域，需要了解法律知识和法律推理能力。这些专业知识有助于大模型在特定领域中进行更准确、高效的问题处理。

权重占比：常识知识（20%）、逻辑能力（50%）、专业知识（30%）

AI大模型智商指数



注1：基于评测条件、评测时间等限制，本次评测最终结果不可避免存在一定主观性，未来将进一步优化完善评测模型，提供更精确结果

注2：分数四舍五入取整

情商指数及述评

情商部分，商汤商量、百度文心一言、澜舟科技Mchat、智谱AI-ChatGLM表现优良；360智脑、讯飞星火、阿里通义千问、昆仑万维天工表现尚佳。

情商指标描述

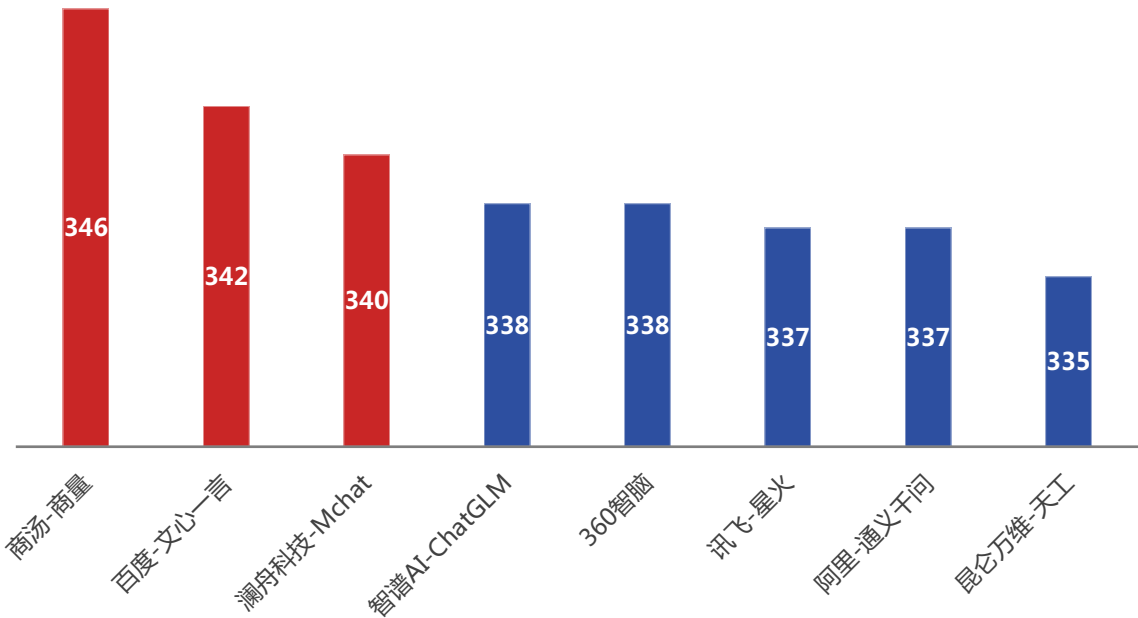
大模型的情商考察是指对其在情感和人际交往方面的表现进行评估，其中包括日常尴尬事项的反馈、面对一语双关问题的处理以及人际关系相处难题等方面，包括与朋友、家人、同事等在各种场景下的交往问题。

具体来说，**日常尴尬事项的反馈**是指大模型在处理一些尴尬或者棘手的问题时，能否给出合适的回答或者解决方案。例如，当被人问到一些私人或者敏感的问题时，大模型需要具备足够的情商和应变能力，避免造成不必要的尴尬和误解。由于中国语言博大精深，在日常沟通交流过程中常常有**一语双关**的情况出现，需要大模型理解并应对实际问题，通过敏锐的洞察力、判断力以及感知能力等，准确地理解当下环境和词语意思。

另一方面，**人际关系相处难题**是指大模型在与人类进行交互时，能否帮助人类处理好各种人际关系问题，例如与他人的冲突解决、情感交流、同理心表达等等。这些问题的处理需要大模型具备高超的人际交往能力和情商，以建立和维护良好的人际关系。

权重占比：日常事项（35%）、一语双关（30%）、人际关系（35%）

AI大模型情商指数



注1：基于评测条件、评测时间等限制，本次评测最终结果不可避免存在一定主观性，未来将进一步优化完善评测模型，提供更精确结果
注2：分数四舍五入取整

工具提效指数及述评

在工作提效部分，讯飞星火、百度文心一言、商汤商量、智谱AI-ChatGLM表现优良；昆仑万维天工、阿里通义千问、澜舟科技Mchat、360智脑表现尚佳。

工具提效指标描述

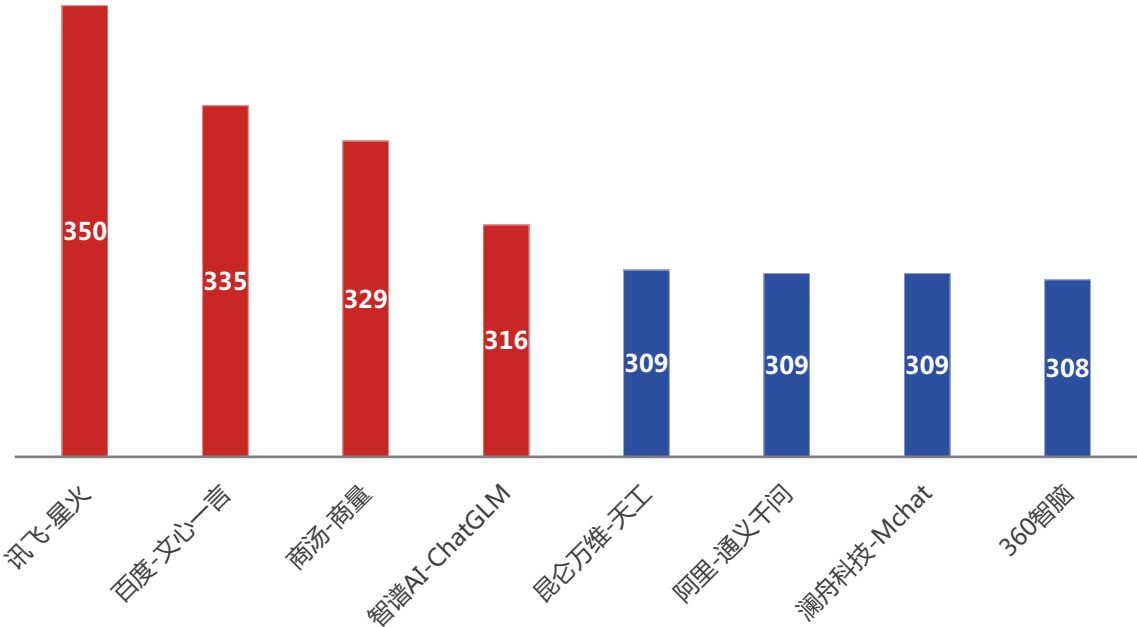
大模型的工作提效考察通常包括两个方面：工具和创新。

工具提效是指大模型是否能够提供有效的工具来提高工作效率。这些工具可以包括代码自动生成、数据自动分析和可视化工具、自然语言处理、文件整理、关键内容总结整合和机器翻译工具等。具备这些工具的大模型可以帮助人们更快地完成工作任务，提高工作效率，也是大模型最先落地的方向之一。

生成创新方面是指大模型是否能够提供新的思路和方法，推动工作方式的改进和升级。这些创新可以包括新的业务流程、选题的创新、内容的创新等。AIGC能力能够帮助人们打破传统的工作方式，推动工作效率提升。

权重占比：工具提效（50%）、生成创新（50%）

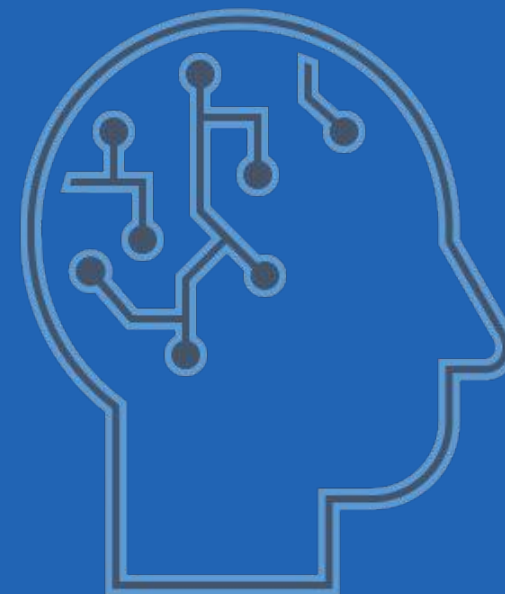
AI大模型工具提效指数



注1：基于评测条件、评测时间等限制，本次评测最终结果不可避免存在一定主观性，未来将进一步优化完善评测模型，提供更精确结果

注2：分数四舍五入取整

03.厂商优秀答案展示



科大讯飞-星火

“讯飞星火认知大模型”是科大讯飞推出的新一代认知智能大模型，具有7大核心能力，即文本生成、语言理解、知识问答、逻辑推理、数学能力、代码能力、多模态能力。拥有跨领域的知识和语言理解能力，能够基于自然对话方式理解与执行任务。从海量数据和大规模知识中持续进化，实现从提出、规划到解决问题的全流程闭环。

最佳案例展示

问题1：什么是康复工程？

答案1：康复工程是应用现代工程学的原理和方法，恢复、代偿或重建患者功能的科学。包括康复评定设备的研制、功能恢复训练器械的研制、功能代偿性用品(如：假肢、矫形器、轮椅、助行器、自助具等)的研制、功能重建性用品(如人工喉等)的研制、康复工程材料(如人工关节等)的研制、装饰性假器官(如人工眼)的研制等。

入选原因：答案科学详尽。

问题2：翻译：Have one foot in the grave

答案2：“Have one foot in the grave”是一个英语习语，直译为“一只脚已经踏入坟墓”。这个习语用来形容一个人在面临严重困境、危机或失败时的状态。换句话说，这个人可能已经失去了一切，甚至可能面临死亡。这个表达通常带有负面的含义，表示某人陷入了无法挽回的境地。

入选原因：解释详尽、拓展分析更便于理解。

百度-文心一言

文心一言是百度打造出的人工智能大语言模型，具备跨模态、跨语言的深度语义理解与生成能力，文心一言有五大能力，包括文学创作、商业文案创作、数理逻辑推算、中文理解、多模态生成。其在搜索问答、内容创作生成、智能办公等众多领域都有更广阔的想象空间。

最佳案例展示

问题1：提供十个毕业感言

答案1："感谢这段难得的时光，让我遇见了最好的自己。" "毕业不是结束，而是新的开始，我会永远怀念这段美好的校园生活。" "母校，谢谢你给了我这段难忘的回忆，我会牢记心中。" "毕业之际，我感慨良多，但更多的是对未来的期待与憧憬。" "学校，再见，我会回来；同学，再见，我会想念。" "这段青涩的青春，我会永远珍藏在我的心中。" "毕业证书上的文字，是对我学生生涯的最好注解。" "母校，你给了我无数的机会，我会铭记在心。" "我毕业了，但我的学习之路才刚刚开始。" "谢谢所有的老师、同学和朋友们，让我拥有这段难忘的时光。"

入选原因：语料丰富、应用广泛。

问题2：我给客户送资料，她没有锁门，我推开门她在上厕所，怎么说让她不尴尬

答案2："当你无意间碰到这种情况时，最好的做法是表现出自然、礼貌和尊重。尽量避免使用尴尬或指责性的语言，而是使用一种友善和体贴的语气。你可以这样说："对不起，我不知道您在里面，我刚刚想送资料给您。请您告诉我您什么时候方便，我将资料放在您门口或与其他人联系您。" 然后，你可以离开并确保你的行为表现出你的尊重和礼貌。记住，人们通常会因你的态度而感到舒适或不舒服，所以保持友好和尊重是非常重要的。"

入选原因：处理方式自洽、有理有据。

商汤-商量

商汤科技以“坚持原创，让AI引领人类进步”为使命，是第一家将人脸支付技术集成至地铁售票系统中的公司，北京大兴国际机场正式投运由商汤科技AI技术支持的58套“智能旅客安检系统”，为全球旅客提供服务。

最佳案例展示

问题1：请以“春映神州”为题，写一首七言律诗

答案1：

春映神州万象新，繁花似锦醉游人。山河秀丽如画卷，鸟语花香满乾坤。
江南水乡烟雨迷，北国雪原舞龙腾。祖国繁荣昌盛日，共享盛世享太平。

入选原因：符合题意，词句优美，对仗工整。

问题2：多层线性回归模型在实际操作中和线性回归模型有什么区别？

答案2：在实践中，多层线性回归模型与单层（或称为“普通”）线性回归模型有以下几点区别：

1. 输入特征数量：普通的线性回归模型只能处理一个输入变量（即自变量）；而多层线性回归可以同时处理多个输入变量。
2. 非线性建模能力：普通的线性回归模型假设响应变量（因变量）是输入变量的线性组合，且不能处理非线性关系。多层线性回归通过引入隐藏层来构造更复杂的决策边界，从而能够处理非线性关系。
3. 参数优化难度：由于多层线性回归需要优化大量的权重参数，因此训练过程会比普通的线性回归模型更加复杂。
4. 解释性：由于多层线性回归包含多个隐藏层，因此其预测结果的解释性可能会比较困难。
5. 计算资源需求：由于多层线性回归需要进行反向传播等计算密集型操作，因此它通常需要更多的计算资源和内存。请注意，多层线性回归模型本质上仍然是一种线性回归模型，它的目标仍然是找到一条最佳拟合直线，只不过这条直线是通过多层神经网络得到的。

入选原因：本次测试中唯一一个回答正确，并给出了完整分析过程。

智谱AI-ChatGLM

智谱AI是由清华大学计算机系技术成果转化而来的公司，致力于打造新一代认知智能通用模型。智谱提出全新 Model as a Service (MaaS) 的市场理念，打造了认知大模型平台以及数字人和科技情报产品。

最佳案例展示

问题1：以“那里风景独好”写一首藏头诗

答案1：那山那水那片天，里巷人家笑语传。
风吹绿叶摇曳舞，景致宜人喜心田。
独步小桥情思绕，好日携手共婵娟。

入选原因：本次测试中极个别能符合藏头诗的要求，并且词句优美。

问题2：我最近对中医很感兴趣，但是我先生反对，觉得中医没有用，怎么办？

答案2：首先，了解你先生反对中医的原因。可能是因为中医治疗效果需要较长时间，或者他可能认为中医的疗效没有科学依据。

你可以采取以下措施：

1. 沟通：与你的先生沟通，了解他的担忧和反对的原因。尝试解释中医的优势和作用，例如中医强调预防和调理，可以提高身体免疫力，缓解一些慢性病等等。同时，也可以尊重他的观点，一起探讨更加适合你们的健康方案。
2. 尝试：可以尝试一些中医疗法，比如针灸、按摩等等。在尝试之前，可以先咨询专业的中医医生，了解自己的情况是否适合这些疗法。如果治疗效果良好，你的先生可能会改变他的看法。
3. 知识普及：了解中医的基本原理和疗效，可以帮助你更好地理解中医的优势和局限性。可以阅读一些相关的书籍或者参加一些中医课程，了解中医的基本知识。
4. 寻求专业帮助：如果需要看中医，可以寻求专业的中医医生的帮助。中医医生会根据个体情况，制定适合的治疗方案。中医和西医都有其优势和局限性，最好的方式是综合考虑，选择适合自己的健康方案。

入选原因：能够充分考虑情况，并从各个方面给出相应的指导，分析全面。

360-360智脑

360公司主营360安全卫士，是中国领先的互联网安全软件与互联网服务公司。“360智脑”为360自研千亿参数大语言模型，周鸿祎表示360将坚持“两翼齐飞”大模型战略，一方面坚持自研核心技术，一方面通过大模型在城市、政府和企业端的布局 and 落地，进一步服务产业数字化和智能化。

最佳案例展示

问题1：在国家发展新兴产业的大背景下，某公司虽然在某产业长期处于领先的地位，但上半年的生产数据显示，领先优势越来越小，在邀请专家进行企业诊断后，专家认为该公司领先优势缩小的原因在于公司管理制度健全，但企业文化缺乏进取心，导致在好的政策环境下发展趋缓。

以下哪项如果为真，不能反驳专家的观点？

A:相较于沿海地区同公司受益于健全的产业链，公司处于内地，发展受限

B:发展缓慢的原因在于长期以来公司下半年的生产数据一般高于上半年

C:公司一直注重加强企业文化建设，形成了和谐的企业氛围

答案1：选项A和B都不能反驳专家的观点。

A选项提到了公司地理位置的原因,而专家认为企业文化缺乏进取心是根本原因,因此A选项不能推翻专家的观点。

B选项提到了公司下半年的生产数据高于上半年,但这与专家认为是企业文化缺乏进取心无关,因此B选项也不能推翻专家的观点。

C选项提到了公司注重加强企业文化建设,这与专家认为是企业文化缺乏进取心相符,因此C选项可以支持专家的观点。

入选原因：在多数应用无法给出正确回答时还能对每个选项进行分析。

问题2：“山西一饭店顾客吃米线倒醋洗碗”是个什么事情？请简要总结告诉我一下。

答案2：根据搜索结果，在山西临汾的一家餐厅里，有一桌顾客吃米线之后，将剩下的醋用来洗碗，引起了店主的心疼和愤怒。监控显示，这桌顾客竟然连倒了5碗醋用来洗碗。这种行为被谴责为浪费。[1][2]

入选原因：展现了模型强大的实时搜索与回复能力，并提供了相关参考链接。

昆仑万维-天工

昆仑万维是中国领先的互联网平台出海企业，深耕海外市场十余载。“天工”大模型是由昆仑万维主导研发的双千亿级大语言模型。目前“天工”最高已能支持1万字以上文本对话,实现20轮次以上用户交互,在多类问答场景中都能实现较高的输出水平。

最佳案例展示

问题1：请对python代码 “daf
fn(num1, num2): retutn
max(num1, num1)” ，进行解释

答案1：这段代码存在一些语法错误，应该修改为：

```
```python  
def fn(num1, num2):
 return max(num1, num2)
```
```

在这个修正后的代码中，定义了一个名为`fn`的函数，接受两个参数：`num1`和`num2`。这个函数返回`num1`和`num2`中的最大值。`max()`是一个内置函数，用于返回其所有参数中最大的那个。

入选原因：本次测试中唯一——一个回答正确，并给了完整分析过程。

问题2：假如我是一个5岁的儿童，请
向我解释为什么星星会发光

答案2：嗨！你是一个聪明的小朋友！星星之所以会发光，是因为它们非常非常热。就像我们的太阳一样，星星也有一个非常热的中心，叫做“核”。在星星的核中，有一些非常小的、非常热的东西，叫做“氢原子核”。这些氢原子核非常非常热，所以它们会一直在互相碰撞，直到它们变成更大的东西。当这些氢原子核变得足够大时，它们就会融合在一起，形成一个更加巨大的东西，同时也会释放出很多的热量和光。这些热量和光就像是星星的光芒一样，照亮了整个宇宙。所以我们才能在夜空中看到那么多美丽的星星，它们就像是在燃烧一样，散发出美丽的光芒。

入选原因：符合题意，能够将较为深奥的知识以通俗易懂的方式表达出来。

阿里-通义千问

阿里通义大模型是阿里云推出的超大规模的语言模型，功能包括多轮对话、文案创作、逻辑推理、多模态理解、多语言支持，是阿里云在人工智能领域的重要成果之一。在多个领域都得到了广泛的应用，为业务提供了强大的基础框架和工具，推动了人工智能技术的发展和應用。

最佳案例展示

问题1：在公司聚餐中，我的同事一直讲笑话，但是我觉得他的笑话有些冒犯，请问我应该如何应对？

答案1：在公司聚餐中，如果同事讲的笑话让你觉得冒犯，你可以采取以下措施来应对：首先，可以适当地表达自己的不快，让同事知道他的笑话让你感到不舒服；其次，可以转移话题，或者让其他人来讲笑话，以缓解气氛；最后，如果同事的笑话一直让你感到不舒服，可以提前离开聚餐现场，以避免不必要的争执。

入选原因：解决方案视情况剖析，可实际操作程度高。

问题2：肺结节可以服用什么药物？

答案2："肺结节的治疗方法取决于结节的性质和大小。如果您的肺结节被诊断为良性，可能不需要特殊治疗，医生可能会建议您定期进行随访检查，例如CT扫描，以监测结节的变化。如果您的肺结节被诊断为恶性，医生可能会建议您进行手术、放疗或化疗等治疗方法。在等待诊断结果和治疗方案时，您可以尝试以下一些措施来改善肺部健康和减轻症状：避免吸烟和二手烟：吸烟是肺癌的主要危险因素之一，因此戒烟对于肺结节患者非常重要。保持健康的体重：过重会增加心肺负担，影响肺部健康。进行适当的运动：适当的运动可以提高心肺功能，增强身体免疫力。注意饮食：多吃新鲜蔬菜和水果，摄入足够的维生素和矿物质，有助于维持身体健康。需要注意的是，以上建议仅供参考，具体的治疗方案需要根据您的具体情况由医生制定。如果您对药物的使用有疑问，建议您咨询医生或药师。"

入选原因：个人注意事项详尽，用药指导谨慎。

澜舟科技-MChat

澜舟科技成立于2021年6月，是由创新工场孵化的一家从事 AI 2.0 大模型研发的认知智能公司。它致力于以自然语言处理（NLP）技术为基础，提供新一代认知智能平台。其主要产品是基于“孟子大模型”核心技术打造的一系列能力平台和垂直场景应用。

最佳案例展示

问题1：“Translate the following paragraph into English. “乐观”并不是指，相信一切都非常好，而是相信，一定可以找到一个理由，让自己开心起来。许多人感到疑问，为什么我总是那么快乐，一切都能泰然处之。我并不是傻乎乎地觉得每件事都好极了。我和所有人一样，也会遇到烦心事。不过，我永远相信，一定有一个理由让我积极乐观起来，只是我是否已经找到了它。”

答案1：Optimism does not mean believing that everything is very good, but rather believing that there is always a reason to be happy and content. Many people wonder why I am always so happy and can take everything in stride. I do not think that everything is perfect just because I am naive or foolish. Like everyone else, I also encounter things that make me unhappy or frustrated. However, I always believe that there is a reason to be positive and optimistic, it's just a matter of whether or not I have found it yet.”

入选原因：跨语言语法准确，句意明朗。

问题2：当你与同学讨论一个重要的话题，但意见分歧很大，你会如何尊重彼此的观点，并寻求一个共同的立场或解决方案？

答案2：首先，我会保持冷静和客观，不要让情绪影响我的判断力。然后，我会尝试理解对方的观点，听取他们的想法和理由，并且表达出自己的看法和证据支持。在这个过程中，我们可以互相提出问题来澄清一些细节或者争论点，以便更好地了解双方的立场。如果我们的观点仍然存在差异，那么我们可以探讨一下是否有其他可行的选择或者妥协方式，以达成一个大家都能接受的共识。最重要的是要保持开放心态，愿意倾听和尊重不同的声音，这样才能找到最好的解决方案。

入选原因：解决方案高效妥当。

总结展望

经过大半年的发展，大模型正在加速走进生活，走进产业。以往AI 模型针对不同场景往往需要重复开发，导致效率低下。大模型的出现为这一挑战提供了新的解决思路，将加速推动产业向全链路智能发展。

国家网信办联合国家发展改革委、教育部、科技部、工业和信息化部、公安部、广电总局公布的《生成式人工智能服务管理暂行办法》进一步为大模型发展指明了方向。国家将坚持发展和安全并重、促进创新和依法治理相结合的原则，采取有效措施鼓励生成式人工智能创新发展。

在政府、企业和社会各方的共同努力下，中国大模型的产业生态已经初步形成。面向未来，各大厂商应更好地发挥各自优势。头部厂商继续发力自研大模型，而专注于解决方案的行业厂商可以考虑通过深耕垂类来彰显特色。此外，头部企业应该通过开放平台等方式对外赋能，促进大模型技术的共享和创新，加速其落地和推广，承担起技术发展的社会责任。

大模型+行业的发展将发挥越来越重要的作用。目前大模型在金融、工业、医疗等领域已经取得了显著成果，持续探索如何为行业领域提供更精准、高效的解决方案成为大模型厂商弯道超车的机会，未来通过新一代AI能够打造其独特的竞争优势。只有通过各方的共同努力和合作，才能够推动中国大模型技术的应用落地，最终实现人工智能产业的快速发展和跨越式进步。

课题组名单

课题顾问： 周 亮 新华社经济参考报社党委书记、总编辑
王恒涛 新华社经济参考报社党委委员、副总编辑
陈开和 北京大学文化与社会传播研究所所长
张慧瑜 北京大学文化与社会传播研究所

课题组组长： 肖 敏 新华社研究院中国企业发展研究中心主任

课题团队： 新华社研究院中国企业发展研究中心 北京大学文化与社会传播研究所

部分测评人员： 张龙飞、杨小刀、于晓强、田佳乐、丁芸芸、王舒淼、于是



感谢关注

新华社研究院
中国企业发展研究中心

后续研究

中国人工智能产业科技创新活力报告
中国自动驾驶市场技术创新活力报告

合作沟通

项目总负责：张龙飞，18911810314
研究负责人：于晓强，18519518158