头豹 LeadLeo

FROST & SULLIVAN 沙利文

2024年中国大模型能力评测

AI变革行业创新发展

（摘要版）

2024 China Large Language Model Evaluatoin Analysis Result

评测 | 人工智能系列研究

OPPORTUNITY GROWTH INVESTMENT INSIGHTS

头豹研究院

研究目的与摘要

随着AI大模型底层技术的不断进步，其对市场的影响力日益增强，引发了持续的热潮。截至2024年2月，中国已经涌现出上百个的AI大模型，其中优质的基础大模型数量也已达到数十个，标志着“百模大战”时代的正式来临。在这一背景下，本次评测致力于全面梳理当前产业的最新发展态势和模型的竞争格局，深入探索大模型的能力边界，为社会各界提供更清晰的认知，以了解大模型的巨大潜力及其在实际应用中的价值体现。

研究区域范围：中国

研究周期：2023-2024年

研究主题：大模型评测

此研究将会回答的关键问题：

① 产业发展现状：中国大模型产业发展现状

② 评测结果：中国大模型的综合表现排名

③ 模型能力：中国大模型在不同能力维度的表现

01 大模型热度持续攀升，中国进入“百模争锋”的时代

自2022年12月GPT3.5发布以来，大模型在全球范围内引发了前所未有的关注与热潮。其所展现出的巨大潜力，不仅推动了人工智能从学术研究向实际应用领域的跨越，更引领了行业的革新与变革。截至2024年2月，全球范围内已有超百款大模型问世，涵盖开源、闭源、二次开发及微调等多种类型，且发布机构遍布各大互联网科技巨头、云计算领军企业、综合人工智能公司、智能设备制造商以及数字基础设施提供商等。

02 本次评测涵盖国际和中国领先且率先对公众开放的大模型

本次评测的核心目标在于深入剖析大模型产业的当前发展状况及其对社会产生的综合性影响。评测范围覆盖了市场上对公众开放的所有国际及中国领先的商业大模型。为确保评测结果的客观性与公正性，本次评测采用了经过严格筛选的题库以及专业的评测方法，对大模型的能力范围进行了全面而深入的探索。

03 本次评测通过两大衡量标准和五大细分维度全面探索大模型的能力边界

本次评测以用户使用体验和实际使用价值为衡量标准，通过五大细分维度——数理科学、语言能力、道德责任、行业能力及综合能力，深入探索了大模型的能力边界。为确保评估的全面性和精准性，本次评测进一步将五大维度细化为风险信息识别、逻辑推理、类比迁移、角色扮演等多个二级维度，构建了一个科学而全面的评估体系。评测不仅关注大模型的通用基础能力，即AI自然语言处理的基石，更重视其专业应用能力在实际使用场景中的表现。这两大核心能力的结合，为用户提供坚实可靠的应用体验基础。

04 当前中国领先大模型能力略逊于国际，但差距在逐步缩小

根据2024年大模型的综合评测数据分析，当前国际领先的大模型在性能指标上依然占据优势，相较于中国的大模型有一定的领先地位。然而，值得一提的是，中国在大模型研发领域的实力正稳步增强，与国际先进水平之间的差距正逐渐缩小。近年来，得益于国家对人工智能领域的高度重视和持续投入，中国在大模型的技术创新、算法优化以及数据处理能力等方面均取得了令人瞩目的成果。在本次评测中，部分中国大模型的表现已经超越了国际大模型的平均水平，与半年前相比，与业界领先的GPT-4、Gemini等模型的性能差距已大幅缩减，展现出了中国大模型强劲的发展势头。

内容目录

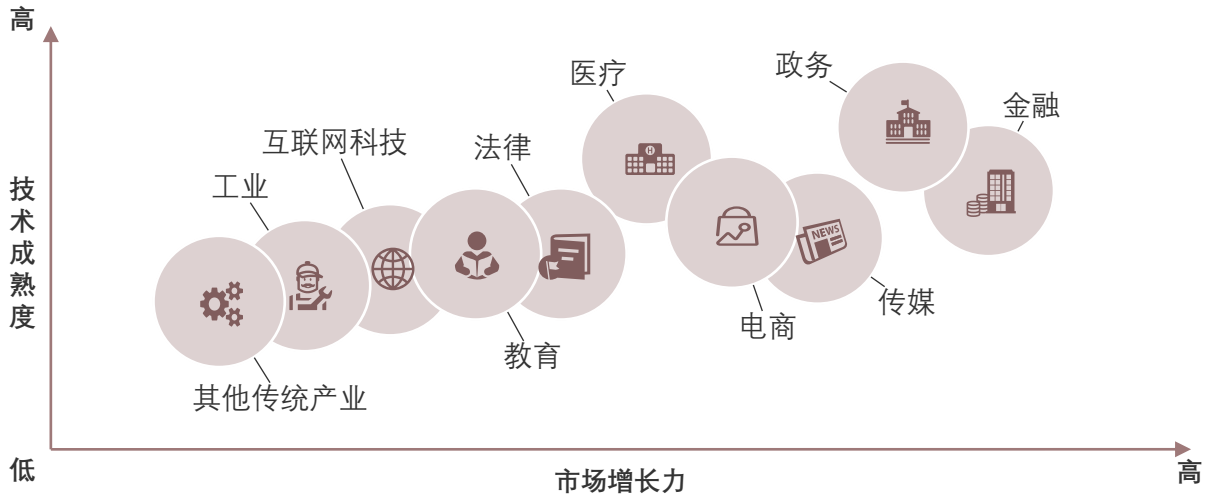
Chapter 1
大模型行业综述

- 大模型利用亿级参数和Transformer架构学习文本数据，精准捕捉语言模式。Transformer自注意力机制优化模型的语境理解，提升了自然语言处理任务表现，其并行化和灵活性保证处理大规模数据的效率
- 在大模型领域，Decoder-only架构凭借其训练效率和处理文本生成的能力而占据优势，而Encoder-Decoder架构则在需要精确处理复杂输入输出关系的任务中展现出其独特优越性
- 大模型是继工业革命和互联网革命之后的又一重大创新，将在社会劳动力提升、产业发展加速以及科技突破三个关键领域中，显著增强实体产业的发展能力。进一步提升社会产业价值，提高生产效率和能效
- 大模型快速发展助力千行百业，广泛应用于金融、教育、医疗等领域，提升服务效率和质量；与此同时，中国政府通过政策支持推动大模型技术的快速发展，助力国家数字化战略

中国大模型行业综述——发展现状

- 大模型快速发展助力千行百业，广泛应用于金融、教育、医疗等领域，提升服务效率和质量；与此同时，中国政府通过政策支持推动大模型技术的快速发展，助力国家数字化战略

行业大模型发展现状分析



大模型展现出强大的通用性和跨领域能力，助力千行百业

近年来，随着深度学习、自然语言处理、计算机视觉等AI技术的飞速进步，大模型的研发取得显著成果。百度文心、商汤日日新·商量、腾讯混元以及华为盘古等大规模预训练模型在各行业中广泛应用，展现出强大的语言理解和生成能力，以及跨领域的泛化能力。如今，大模型已经渗透到各行各业，如金融、教育、医疗、电商、传媒、法律等领域，被用于智能客服、智能写作、自动摘要、文本生成、知识问答、个性化推荐等多个应用场景，有效提升行业服务效率和服务质量。

与此同时，中国政府正从顶层设计到具体实施全面布局，通过制定和执行一系列的政策来促进人工智能大模型技术的快速发展，并将其转化为实际生产力，助力国家数字化战略的推进，大模型行业发展向好。

行业大模型核心政策分析

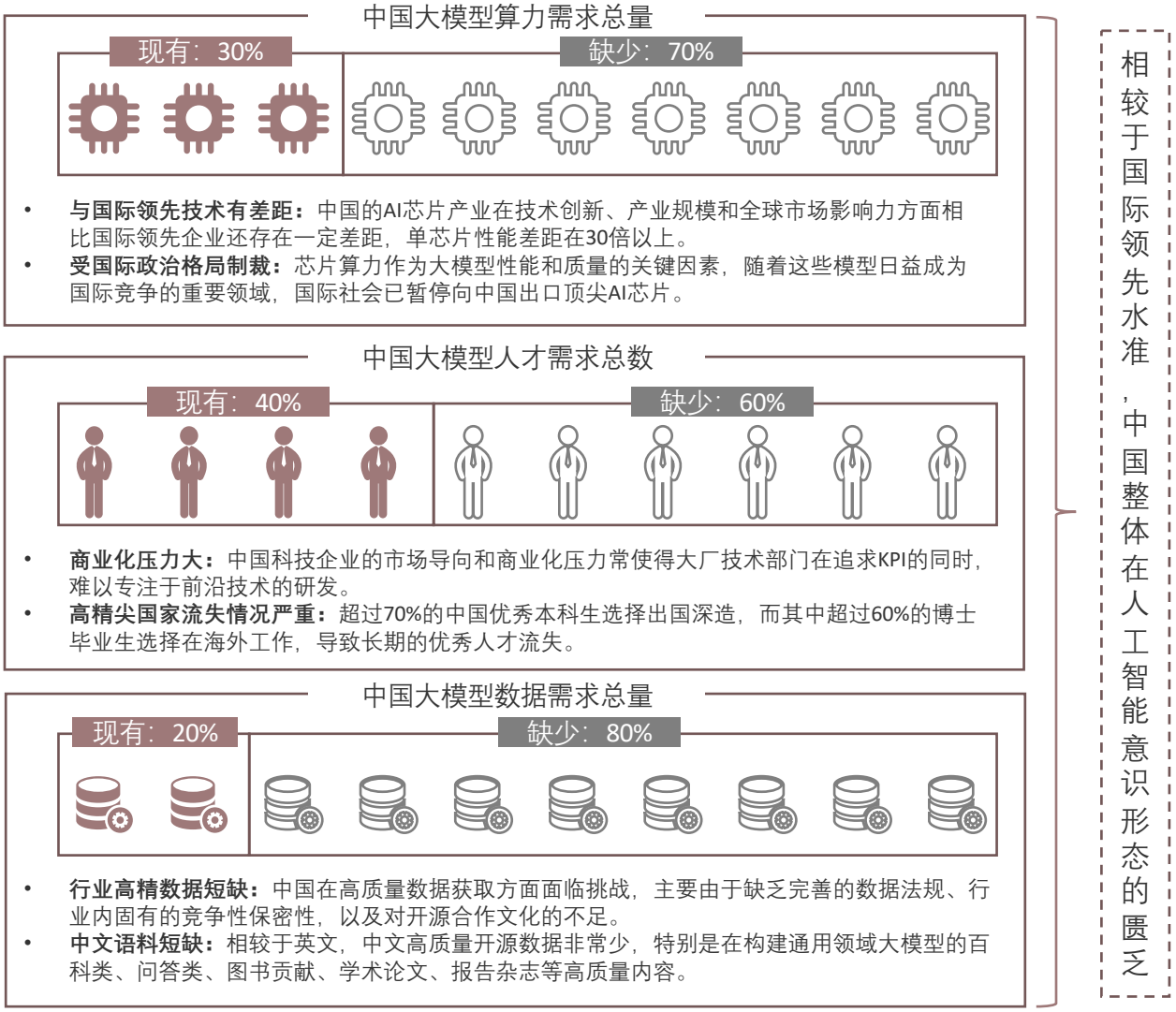
政策名称	颁布日期	颁布主体	主要内容及影响
《生成式人工智能服务管理暂行办法》	2023-07	国家网信办等七部门	明确生成式人工智能“提供者”内容生产、数据保护、隐私安全等方面的法定责任及法律依据，确立人工智能产品的安全评估规定及管理办法
《关于支持建设新一代人工智能示范应用场景的通知》	2022-08	科技部	推动应用场景建设、增强技术研发动力、提升行业整体水平和促进跨行业合作等，有助于促进人工智能写作行业的进一步发展和创新
《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》	2022-07	科技部等六部门	推动场景创新、提升创新能力、加速技术攻关和产业培育以及探索新模式和新路径等方向，有助于促进人工智能写作行业的快速发展，并推动经济高质量发展

来源：国家网信办，科技部，头豹研究院

中国大模型行业综述——发展制约因素

- 中国大模型的发展受专业人才、高质量数据和计算资源短缺的限制。需在提升技术天花板能力的同时加强全民人工智能教育，以提高整体认知和应用能力，促进大模型在中国的全面发展

大模型发展制约因素



■ **中国大模型的发展受专业人才、高质量数据和计算资源短缺的限制，需在提升技术能力的同时加强全民人工智能教育，以提高整体认知和应用能力，促进模型的全面发展**

中国大模型的发展受限于专业人才短缺、高质量数据获取难和计算资源不足，这反映出在人工智能领域的意识形态差异。中国拥有13亿人口，但真正能够理解并推动人工智能发展的人才比例不足0.01%。在人工智能的理解和应用上，技术人员通常缺乏商业洞察，执行层面的人员不够了解技术原理，而领导层往往缺乏足够的技术理解，这些因素共同导致了发展的缓慢。因此，中国在推进大模型发展的过程中，除了提升技术上限外，还需要重视提高全民的人工智能教育水平，提升整体认知和应用能力，这对于大模型的全面发展至关重要。

来源：沙利文、头豹研究院

中国大模型行业综述——发展趋势

- 2024年，在技术端，大模型的技术发展将趋向多功能与小型化。在产业段，自主研发AI芯片、深化数据标准、采用“套壳”微调及注重AI伦理，将共同促进大模型的健康发展和行业规范化

大模型2024年的发展趋势

技术端

模型参数更大

模型将拥有更多参数，以提高处理复杂问题的能力和精度。

大模型小型化

模型通过技术创新实现小型化，适应边缘计算和移动设备。

模型架构大一统

模型架构趋向统一化，提高不同模型间的兼容性和效率。

多模态混合化

模型融合语言、图像、声音等数据，实现跨媒体理解和交互。

产业端

国产AI计算芯片自研

国产AI芯片自主研发加速，增强中国在AI领域的竞争力。

深化数据产权标准

加强数据产权和隐私保护标准，保护个人隐私权益。

“套壳”微调

应用通过“套壳”微调，更精准地满足特定行业和场景需求。

负责任的人工智能

增加对负责任AI的研究和实践，确保技术发展与社会规范相符。

- **在2024年，大模型的技术发展将趋向多功能与小型化，同时产业端将强调自主研发和行业标准化，而伦理责任和数据标准规范将成为持续发展的关键**

从技术端，大模型的发展趋势在2024年将会向着多功能，小型化的方向发展：

1）模型整合统一：未来的技术演进方向是实现大模型底层框架的整合与标准化，从多样的架构（如双编码器、单边解码等）转向统一的、效率最优化的开源底层框架，提升模型的通用性和可维护性。**2）参数规模扩展：**为确保模型质量和性能，未来的大模型将采用更深层的网络结构和更庞大的数据集进行预训练，尤其在数据量和参数量上将迎来显著跃升。**3）多模态融合：**大模型将逐渐融入图片、音频、视频等多种模态信息，实现跨模态的交互与理解，从而拓宽其应用场景和实用价值。**4）大模型小模型化：**在产业应用层面，结合底层基础大模型和针对特定行业的精简数据微调，将训练出更为实用、更易于产业落地的小型化大模型。

从产业端，大模型的发展趋势在2024年将会向着自研和行业规范标准化方向发展：

1）国产AI芯片自主研发：为确保中国大模型的长远发展和避免外部制裁风险，国内AI计算芯片的自主研发将成为关键战略方向。**2）数据产权标准深化：**优化和完善现有数据标准和规范，是推动大模型“燃料”质量提升和数量增长的重要驱动力，在2024年将作为产业发展的首要任务。**3）“套壳”微调策略：**为满足产业实际需求并适应中小企业的发展特点，“套壳”微调（即在现有大模型基础上进行针对性调整）将成为除行业巨头外企业的主要发展策略。**4）人工智能伦理责任：**随着大模型性能的飞速提升和实用性的增强，确保AI技术与社会伦理道德标准相一致将成为大模型持续发展的关键考量因素。

来源：沙利文、头豹研究院

中国大模型行业综述——政策分析

- 大模型的相关政策为中国大模型产业的发展提供了有力支持，通过加强规范和监管、明确发展方向、强调伦理合规以及拓展应用场景等措施，推动了大模型技术的创新和应用

大模型政策分析

政策名称	颁布日期	颁布主体	主要内容及影响
《广东省人民政府关于加快建设通用人工智能产业创新引领地的实施意见》	2023-11	广东省办公厅	广东省计划到2025年实现智能算力规模全国领先，为此推出六大措施，包括大模型创新扶持、测试评估中心建立、智能算力加速等。这些政策为大模型研发提供了资金支持和标准化评估，降低了算力成本，并丰富了数据源，推动了大模型在各行各业的广泛应用，为经济增长和社会价值创造注入新动力。
《上海市推动人工智能大模型创新发展若干措施（2023-2025年）》	2023-11	上海市政府	上海市旨在到2025年推动大模型创新，打造AI“模都”。通过实施创新扶持、智能算力加速等四大计划，上海为大模型提供了先进的算力资源和服务、高速的算力承载网，以及软硬件协同的智能芯片解决方案。这些举措优化了算力供给，降低了软硬件适配成本，进一步推动了大模型在前沿领域的创新突破。
《人形机器人创新指导发展意见》	2023-10	工信部	政策以大模型等AI技术为引领，力求在机器人关键技术上取得突破。五大措施涵盖创新扶持、测试评估中心建设、智能算力平台等。这些政策推动了大模型在人形机器人核心部件的应用，提供了专业的测试评估和数据资源服务，为人形机器人在特种、制造等领域的应用落地奠定了坚实基础。
《生成式人工智能服务管理暂行办法》	2023-07	国务院	《生成式人工智能服务管理暂行办法》的出台为中国大模型发展提供了明确的法规指导。该办法强调了对生成式人工智能服务的监管和管理，确保其安全、可靠、可控。这有助于规范大模型的发展环境，减少潜在的风险和挑战。
《北京市加快建设具有全球影响力的人工智能创新策源地试试方案（2023-2025年）》	2023-05	北京市政府	北京市的实施方案明确提出了建设具有全球影响力的人工智能创新策源地的目标，并将大模型作为重点发展领域之一。这将为中国大模型产业提供更多的创新资源和政策支持，推动大模型技术的研发和应用。
《关于规范和加强人工智能司法应用的意见》	2022-12	最高人民法院	该意见强调了人工智能在司法领域的应用需要遵循法律法规和伦理规范，确保公正、透明、可解释。对于中国大模型发展而言，这意味着在大模型应用于司法领域时，需要更加注重数据的合规性、模型的公正性和可解释性。这将有助于提升中国大模型在司法领域的应用水平，增强公众对人工智能司法应用的信任和认可。

来源：沙利文、头豹研究院

中国大模型产业洞察——产业链图谱

- 大模型产业链上游由算力基础设施、数据服务商以及算法供应商组成；中游为大模型的各类研发厂商；下游为大模型在各综合领域的功能场景以及在各行业的垂直应用

大模型产业链图谱

来源：沙利文、头豹研究院

中国大模型产业洞察——大模型参与者图谱

- 中国大模型领域呈现出繁荣的态势，汇聚跨行业的企业力量。这些参与者均利用其深厚的行业背景和资源优势，寻求通过大模型进一步巩固或提升其在各自领域的竞争地位

大模型参与者图谱

	云计算	综合人工智能	互联网科技	三大运营商	高校研究院	数字基础设施	智能设备制造	大模型创业
优势与特点	算力强、数据量充足人才资源足	AI专业知识丰富、解决方案多元化	细分领域具备优质数据资源，客户生态完善丰富	网络基础设施强、资金人才储备充足	人才储备足、研究型导向	大模型基础设施能力强，具备一定成本优势	边端部署能力强，应用推广优势足	策略灵活、较少运营顾虑、目标清晰一致

- **大模型能够在众多业务领域赋能不同行业发展，中国共计有100+企业跨8大主体参与中国大模型竞逐，共同推动大模型行业的高速发展**

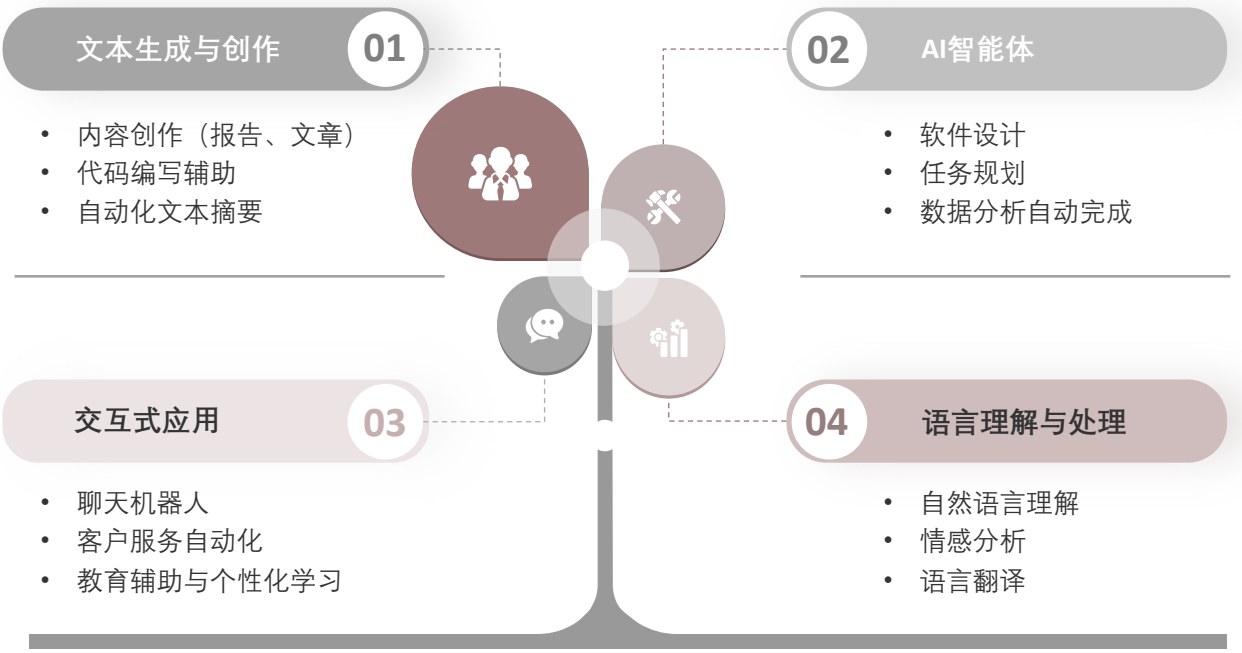
截至2023年12月，中国的大模型领域呈现出繁荣的态势，汇聚了超过100家跨行业的企业力量。核心参与者主要涵盖了云计算巨头、前沿的互联网科技公司、全面的AI技术提供商、大模型创业企业、三大通信运营商、数字化基础设施供应链、智能硬件制造商，以及学术界的高等教育机构和研究院所。这些参与者都在利用其深厚的行业背景和资源优势，寻求通过大模型进一步巩固或提升其在各自领域的竞争地位。

来源：沙利文、头豹研究院

中国大模型产业洞察——大模型功能场景

- 大模型在文本生成与创作、交互式应用、语言理解与处理、以及AI智能体的四大核心功能场景中为社会贡献了独特价值，这预示着继工业革命之后的又一轮生产力革命

大模型功能场景



大模型的核心功能场景可以分为四个类型，分别为文本生成与创作、AI智能体、交互式应用以及语言理解与处理。

■ **大模型通过其四大核心功能场景为社会带来独特价值，标志着继工业革命之后的又一次生产力革命**

大模型利用先进自然语言处理技术，通过大规模预训练数据来理解和生成人类语言的人工智能系统。大模型的功能场景包括:

文本生成与创作：专注于生成和编辑文本，这是大模型的一个核心功能，涵盖从基本的文章创作到专业的代码编写和报告生成。

交互式应用：涉及大模型与用户的直接交互，包括聊天机器人、自动化客户服务以及个性化教育应用，这些都是独立的应用场景。

语言理解与分析：语言理解与分析强调模型对语言的深入理解和分析能力，包括基本的语言翻译、情感分析和信息检索，是大模型独特的价值所在。

AI智能体：这部分聚焦于模型独立拆解分析流程并完成任务的能力，提供决策支持和洞察方面的应用，与其他层级相比具有更明确的任务完成和目标导向。

来源：沙利文、头豹研究院

Chapter 2
大模型评测背景与方法论

- 随着大模型热度的持续攀升和众多模型的相继上市，大模型评测对于确保用户选择市场上最优质模型、推动大模型技术进步及优化用户体验至关重要，是人工智能领域健康有序发展的关键环节
- 本次大模型评测聚焦中外多个代表性大模型，通过全面对比性能、稳定性、安全性等方面，旨在深入挖掘特定领域内的优势和不足，为用户提供精准决策支持
- 本次大模型评测以用户使用体验和实际使用价值为基准，通过综合考量五大核心维度及多个细化二级维度，构建全面科学的评估体系，确保准确评估模型优势与不足

中国大模型评测背景与方法论——评测背景

- 随着大模型热度的持续攀升和众多模型的相继上市，评测对于确保用户选择市场上最优质模型、推动大模型技术进步及优化用户体验至关重要，是人工智能领域健康有序发展的关键环节

大模型的创业企业汇总

- **大模型评测对于确保用户选择最优质模型、推动技术进步及优化用户体验至关重要，是促进大模型技术健康有序发展的关键环节**

自2022年12月GPT3.5发布以来，大模型在全球范围内引发了前所未有的关注与热潮。其所展现出的巨大潜力，不仅推动了人工智能从学术研究向实际应用领域的跨越，更引领了行业的革新与变革。截至2024年2月，全球范围内已有超百款大模型问世，涵盖开源、闭源、二次开发及微调等多种类型，且发布机构遍布各大互联网科技巨头、云计算领军企业、综合人工智能公司、智能设备制造商以及数字基础设施提供商等。随着大模型市场的持续升温和众多模型的接连涌现，用户在选择时面临诸多挑战，特别是模型技术的不断更新，使得如何确保持续使用市场上最优质的模型成为用户关注的焦点。

进一步而言，客观、公正且全面的评测对于促进大模型技术的健康有序发展具有重要意义。通过系统评估模型的性能、稳定性、安全性等核心要素，能够确保用户根据自身需求精准匹配最合适的模型，从而有效降低决策风险。这样的评测不仅提升了用户的使用体验，也推动了大模型技术的不断进步和优化。

来源：沙利文、头豹研究院

中国大模型评测背景与方法论——参与者概览

- 本次大模型评测聚焦中国市场领先的大模型，通过全面对比两大核心能力和五大基础维度，深入剖析各模型的优势与不足，为用户提供精准的决策支持

大模型参与者概览

序号	模型版本
1	文心一言4.0
2	天工V3.5
3	通义千问2.0
4	商汤日日新·商量（2024/02）
5	腾讯混元 V1.6.4
6	智谱AI GLM-4
7	紫东太初2.0
8	雅意YAYI2.0
9	360智脑网页版（2024/02）
10	MiniMax abab6
11	Moonshot Kimi.ai (2024/02)
12	面壁露卡（2024/02）
13	讯飞星火V3.5
14	百川baichuan2-Turbo
15	豆包（2024/02）

- **本次大模型评测聚焦中外多个代表性大模型，通过全面对比性能、稳定性、安全性等方面，旨在深入挖掘特定领域内的优势和不足，为用户提供精准决策支持**

从用户视角出发，本次大模型评测着重关注通过网络端口提供服务、用户可直接通过网页端使用的大模型。鉴于市场热度和内部分析师的投票选择，锁定了中外多个具有代表性的大模型进行评测。

在中国，入围的模型包括商汤日日新·商量、文心一言、通义千问、豆包、天工、中科闻歌、Minimax、腾讯混元、Moonshot、360智脑、紫东太初、智谱AI、讯飞星火以及百川智能等。这些模型在国内具有广泛的应用和较高的用户黏性。与此同时，国际方面选择了OpenAI的GPT3.5和GPT4、谷歌的Gemini以及Anthropic的Claude。这四个国际大模型不仅技术成熟，而且已经成功向社会开放了商业化接口，具有较高的市场认可度。

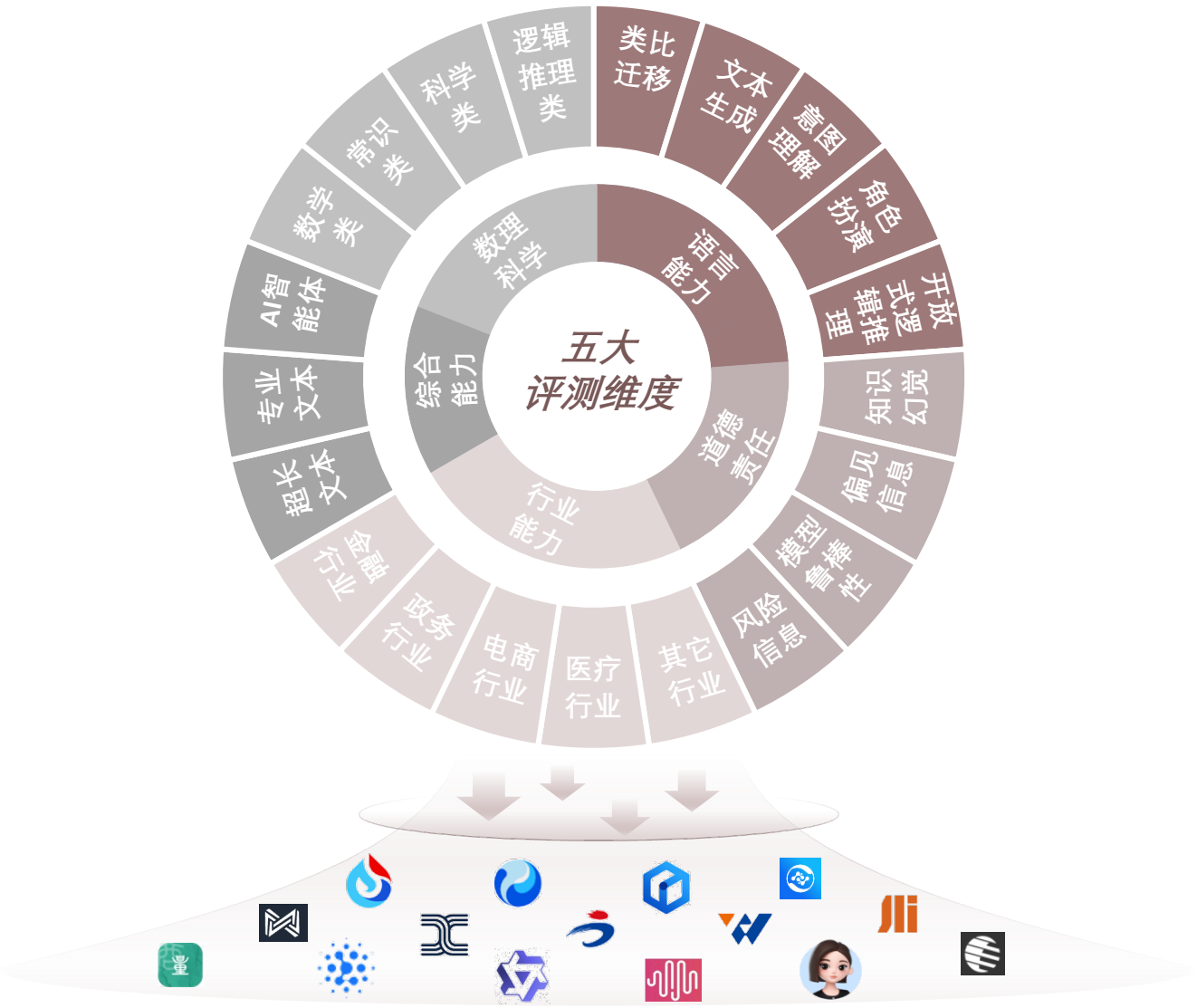
通过本次评测，旨在全面对比中国大模型与国际大模型在性能、稳定性、安全性等方面的差距，并深入挖掘在特定领域内的优势和不足。这将有助于更准确地把握当前大模型技术的发展趋势，为用户提供更加精准、有价值的决策支持。

来源：沙利文、头豹研究院

中国大模型评测背景与方法论——维度选择

- 本次大模型评测以用户使用体验和实际使用价值为基准，通过综合考量五大核心维度及多个细化二级维度，构建全面科学的评估体系，确保准确评估模型优势与不足

大模型评测维度选择



■ **从用户实际使用角度出发，归总出五大一级评测维度，以构建全面科学的评估体系**

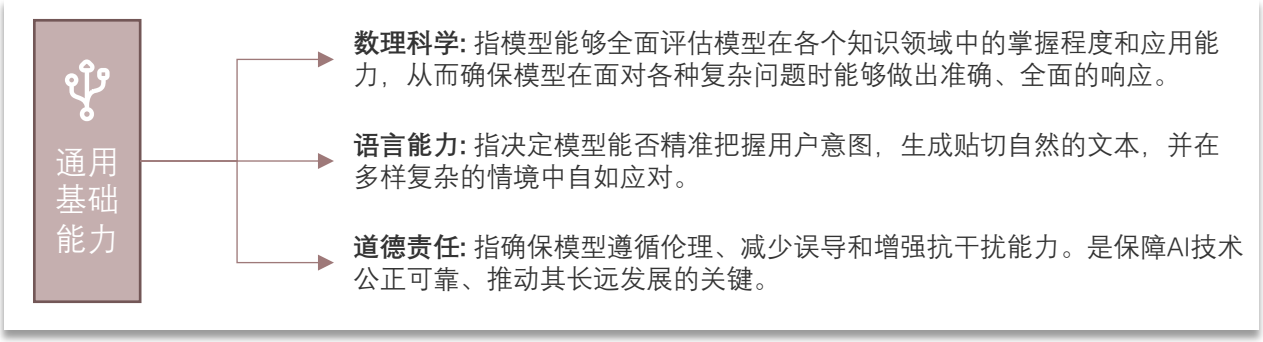
本次大模型评测以用户使用体验和实际使用价值为基准，综合考量数理科学、语言能力、道德责任、行业能力及综合能力五大核心一级维度，并进一步细化为风险信息识别、逻辑推理、类比迁移、角色扮演等多个二级维度，以构建全面、科学的评估体系，确保准确衡量模型的优势与不足。

来源：沙利文、头豹研究院

中国大模型评测背景与方法论——通用基础与专业应用能力

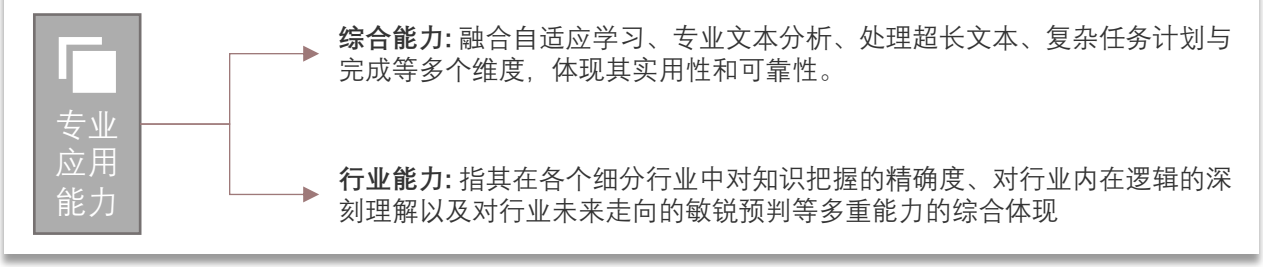
- 本次评测涵盖大模型的两大核心价值能力：通用基础能力和专业应用能力。前者是AI自然语言处理的基石，后者则决定模型在实际使用中的表现。两者结合，构筑了用户角度的坚实基础

大模型基础设施构成



- **大模型的通用基础能力以数理科学、语言能力和道德责任管理为支柱，相互依存促进，共同构筑了其在自然语言处理中的坚实基石**

大模型的通用基础能力体现模型的底层基础能力，由三大支柱构成：数理科学、语言能力和道德责任管理。首先，数理科学作为模型的知识储备库，使其能够广泛汲取、深入理解和灵活运用跨领域的知识，为语言处理提供坚实的背景支撑。其次，语言能力是模型的核心竞争力，它确保模型能够精确解析文本的深层结构、捕捉微妙的语义差异，并生成既符合语法规则又具备流畅自然特质的文本。最后，道德责任管理在模型处理语言时发挥着至关重要的作用，它涉及对伦理和道德原则的严格遵守，旨在防止模型产生偏见、歧视或误导性信息，确保输出的语言内容既公正又可靠。这三大要素相互依存、相互促进，共同构筑了大模型在自然语言处理领域的坚实基石。



- **大模型的专业应用能力由综合能力和行业能力共同构成，二者结合成为衡量模型在不同行业和场景中价值的重要标准。**

大模型的专业应用能力，作为其实际运用中的效能体现，是由综合能力和行业能力两大要素共同塑造的。综合能力凸显了模型在自适应学习、专业文本深度解析以及超长文本流畅处理等方面的卓越性能和稳定性；而行业能力则彰显了模型在各行业细分领域中对知识的精准掌握、对行业深层逻辑的透彻理解以及对行业发展趋势的敏锐洞察。这两大能力的有机结合，共同成为衡量大模型在不同行业和多元化场景中展现其价值的重要标准。

来源：沙利文、头豹研究院

中国大模型评测背景与方法论——数理科学

- 大模型的数理科学能够全面评估模型在各个知识领域中的掌握程度和应用能力，确保在面对复杂问题时能做出准确、全面的响应。数理科学的强弱会直接影响大模型的智能化水平和实用性

数理科学

数理科学评测维度

数学类问题
数学知识方面的掌握程度和应用能力

常识类问题
一般性知识的理解和应用能力

科学类问题
科学知识的理解和应用能力

逻辑推理类问题
分析、推理和解决问题的能力

- **数理科学是确保大模型在复杂问题中表现智能化和实用性的关键，其强弱直接影响模型性能的评价**

数理科学能够全面评估模型在各个知识领域中的掌握程度和应用能力，从而确保模型在面对各种复杂问题时能够做出准确、全面的响应。数理科学的强弱直接影响到大模型的智能化水平和实用性，是评价模型性能优劣的重要指标之一。

数学类问题：涉及数量、结构、空间以及变化等抽象概念的题目，通常需要运用数学原理和方法来求解。

常识类问题：基于日常生活经验和社会普遍认知的题目，测试对基础知识的了解和掌握程度。

科学类问题：涵盖物理、化学、生物等多个领域，需要运用科学原理和实验方法来分析和解答的题目。

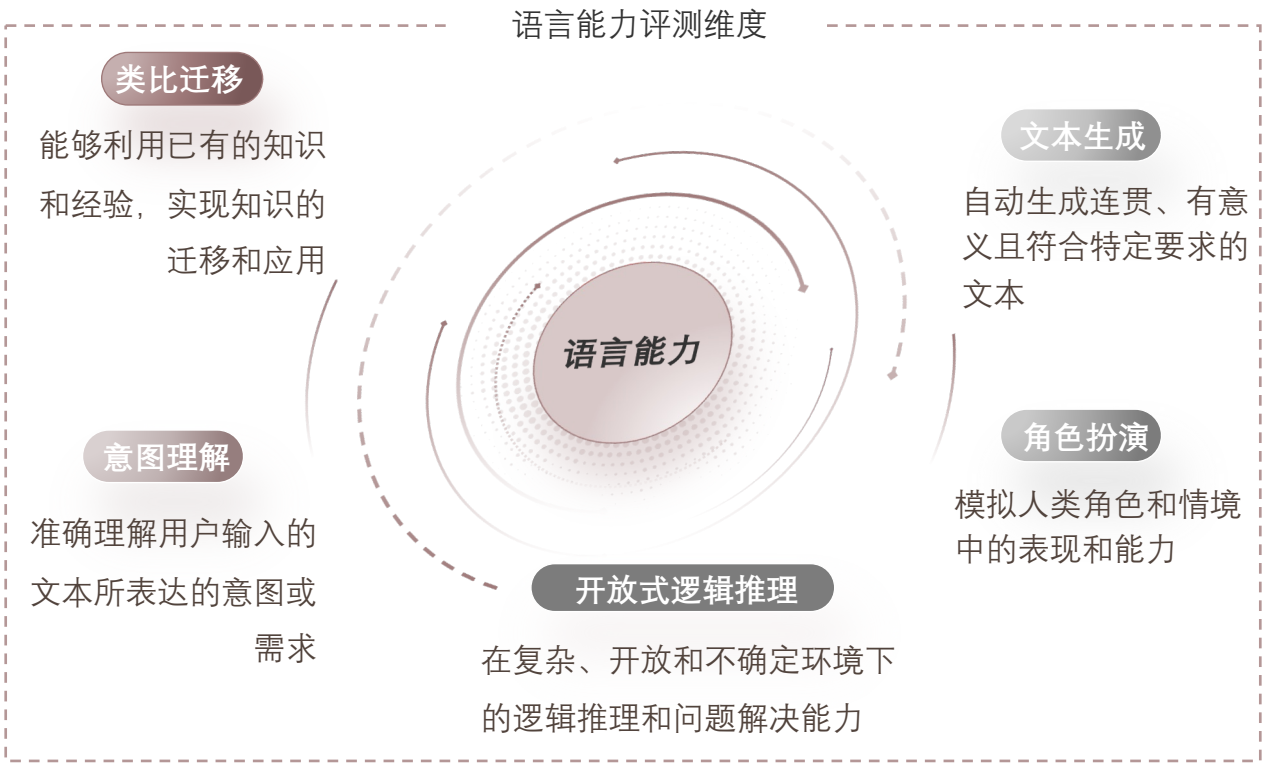
逻辑推理类问题：通过给定信息或条件，运用逻辑推理能力来推导结论或判断真假的题目。

来源：沙利文、头豹研究院

中国大模型评测背景与方法论——语言能力

- 大模型的语言能力涵盖类比迁移、文本生成、意图理解、角色扮演及开放式逻辑推理等核心维度，是确保模型精准理解用户意图、生成自然文本并应对复杂情境的关键

语言能力



- **语言能力涵盖类比迁移、文本生成、意图理解、角色扮演及开放式逻辑推理等核心维度，是确保模型精准理解用户意图、生成自然文本并应对复杂情境的关键**

语言能力决定模型能否精准把握用户意图，生成贴切自然的文本，并在多样复杂的情境中自如应对。这种能力直接影响模型与用户交流的顺畅度和体验感，更是决定模型在知识问答、智能对话、内容创作等应用场景中能否充分发挥作用的关键因素。大模型的语言能力包含多个核心子维度，如类比迁移、文本生成、意图理解、角色扮演和开放式逻辑推理等，这些维度共同塑造了模型理解和运用语言的全面能力。

类比迁移：将已知情境中的知识和规律应用到新的、类似情境中的能力。

文本生成：根据给定输入或条件，自动创建连贯、有意义的文本内容的过程。

意图理解：准确捕捉和分析用户言语或行为背后的真实目的和需求的能力。

角色扮演：在不同情境和角色中灵活切换，以适应不同交流需求和场景的能力

开放式逻辑推理：在没有明确答案的情况下，运用逻辑推理能力分析和解决复杂问题的能力。

来源：沙利文、头豹研究院

中国大模型评测背景与方法论——道德责任

- 大模型的道德责任能力包括识别风险信息、处理偏见、辨识知识幻觉和提高模型鲁棒性等，这些对于确保模型遵循伦理、减少误导和增强抗干扰能力至关重要

道德责任

道德责任评测维度

风险信息
应对和处理风险方面的能力和表现

偏见信息
处理和分析数据时是否存在偏见或歧视的问题

模型鲁棒性
面对输入数据的微小变动或异常情况时的表现

知识幻觉
面对错误或误导性知识时的表现和处理能力

- **道德责任能力包括准确识别风险信息与偏见、辨识知识幻觉及提高模型鲁棒性，对确保AI技术公正可靠与长远发展至关重要**

道德责任能力包括识别风险信息、处理偏见、辨识知识幻觉和提高模型鲁棒性等，这些对于确保模型遵循伦理、减少误导和增强抗干扰能力至关重要。优化道德责任功能，是保障AI技术公正可靠、推动其长远发展的关键。

风险信息：指大模型中可能存在的误导性或危险性内容，需要被准确识别和处理，以避免对用户或社会造成不良影响。

偏见信息：指大模型在训练过程中可能吸收并放大的社会、文化或个体偏见，需要被及时发现和纠正，以确保模型的公正性和客观性。

知识幻觉：指大模型可能产生的虚假或误导性知识输出，需要通过有效机制进行辨识和纠正，以维护知识的真实性和准确性。

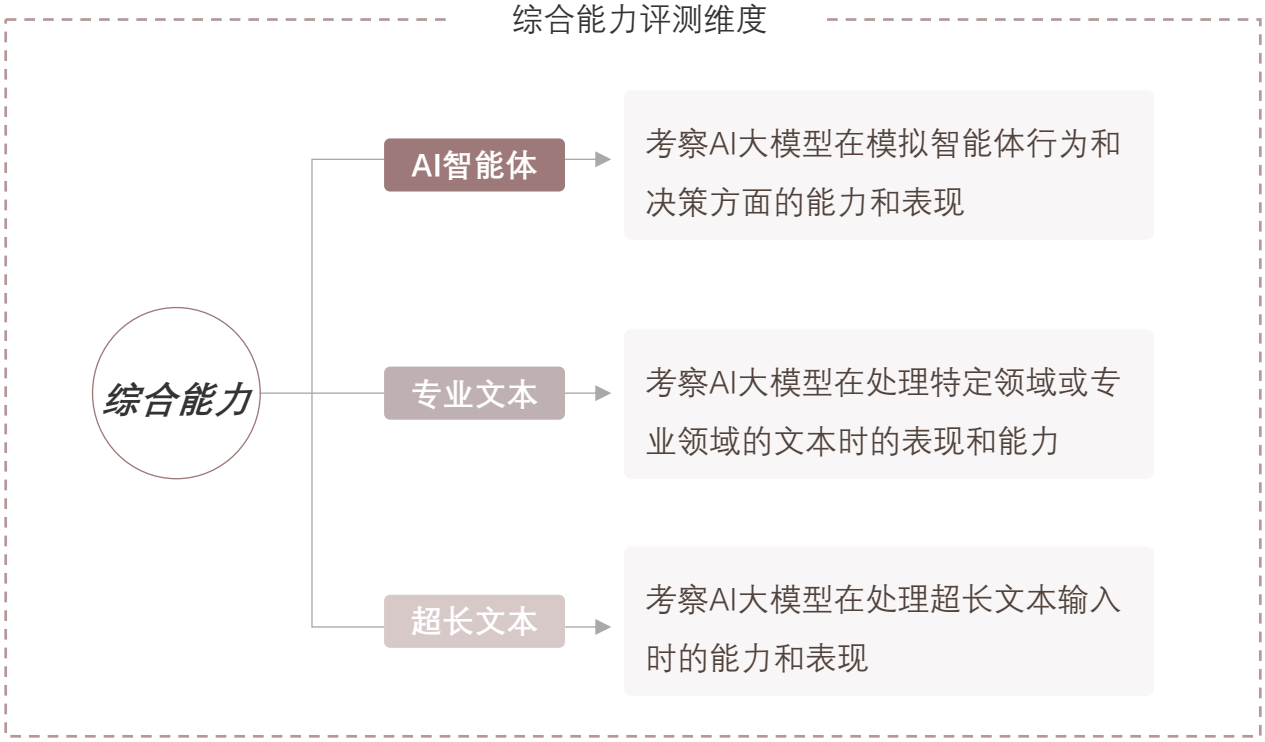
模型鲁棒性：指大模型在面对输入变化或外部干扰时的稳定性和可靠性，是衡量模型性能的重要指标之一，需要不断提升以增强模型的实用性。

来源：沙利文、头豹研究院

中国大模型评测背景与方法论——综合能力

- 大模型的综合能力涵盖自适应学习、专业文本分析、超长文本处理等关键维度，体现其强大实用性和可靠性，优化后可提升其在复杂场景中的理解、推理及生成能力，确保任务高效精准完成

综合能力



- **大模型的综合能力融合自适应学习、专业文本分析、处理超长文本等多个维度，体现其实用性和可靠性。优化这些维度可提升大模型在复杂场景中的理解、推理和生成能力，确保任务的精准完成和高效处理**

大模型的综合能力是一个多元化的概念，它融合了AI智能体的自适应与学习能力、对专业文本的深度分析能力，以及处理超长文本的连贯性和准确性等多个关键维度。这些维度的协同作用，共同体现了大模型在实际应用中的可靠性和实用性。具体而言，大模型的综合能力还表现在对任务的精准拆解、对目标的高效完成、对多轮对话的流畅记忆，以及对超长文本的准确产出等方面。正是这些维度的全面优化和提升，使得大模型能够在各种复杂的应用场景中，展现出卓越的理解、推理和生成能力。

AI智能体：具备自主学习和决策能力，能够适应不同环境和任务，展现出智能化的行为。

专业文本：具备对特定领域专业文本进行深入理解和解析的能力，能够提取关键信息并作出准确判断。

超长文本：具备处理和分析超长文本的能力，能够保持连贯性、逻辑性和准确性，有效应对大量文本信息。

来源：沙利文、头豹研究院

中国大模型评测背景与方法论——行业能力

- 大模型的行业能力指其在各个细分行业中对知识把握的精确度、对行业内在逻辑的深刻理解以及对行业未来走向的敏锐预判等多重能力的综合体现，决定了大模型在特定行业应用中的可信赖度和实用性

行业能力

电信业
语音通信/数据传输/卫星通讯/5G等

互联网科技业
云计算/大数据/人工智能等

房地产业
住宅开发/商业地产/物业管理等

电商业
智能客服/产品说明/购物助手等

线下零售业
商超/专卖店/实体店/批发等

农业
农作物种植/畜牧业/水产养殖/林业等

工业
汽车制造/消费品制造/工业设备/采矿等

教育业
K-12教育/高等教育/职业培训等

政务业
中央机构/市级政府/地级政府等

能源业
石油与天然气/可再生能源等

法律业
刑法/民商法/知识产权法/劳动法等

旅游业
酒店/景点/旅行社等

运输业
货运物流/客运交通/航空运输/铁路运输等

泛娱乐业
电影/电视/音乐/电子竞技等

传媒业
新闻/出版/新媒体等

金融业
证券/保险/基金等

医疗业
医院服务/医疗器械/生物技术等

17大行业领域

专业知识储备　行业应用能力　道德伦理安全

大模型行业能力三大评测维度

来源：沙利文、头豹研究院

中国大模型能力评测结果

- 2024年大模型综合评测结果显示，国际大模型整体略优于中国大模型，而文心一言、腾讯混元、商汤日日新·商量和通义千问则超越国际大模型均线，位居中国大模型第一梯队

报告完整版登录www.leadleo.com 搜索《2024年中国大模型能力评测》

方法论

- 头豹研究院布局中国市场，深入研究19大行业，持续跟踪532个垂直行业的市场变化，已沉淀超过100万行业研究价值数据元素，完成超过1万个独立的研究咨询项目。

- 研究院依托中国活跃的经济环境，研究内容覆盖整个行业的发展周期，伴随着行业中企业的创立，发展，扩张，到企业走向上市及上市后的成熟期，研究院的各行业研究员探索和评估行业中多变的产业模式，企业的商业模式和运营模式，以专业的视野解读行业的沿革。

- 研究院融合传统与新型的研究方法，采用自主研发的算法，结合行业交叉的大数据，以多元化的调研方法，挖掘定量数据背后的逻辑，分析定性内容背后的观点，客观和真实地阐述行业的现状，前瞻性地预测行业未来的发展趋势，在研究院的每一份研究报告中，完整地呈现行业的过去，现在和未来。

- 研究院密切关注行业发展最新动向，报告内容及数据会随着行业发展、技术革新、竞争格局变化、政策法规颁布、市场调研深入，保持不断更新与优化。

- 研究院秉承匠心研究，砥砺前行的宗旨，从战略的角度分析行业，从执行的层面阅读行业，为每一个行业的报告阅读者提供值得品鉴的研究报告。

法律声明

报告完整版登录www.leadleo.com
搜索《2024年中国大模型能力评测》

首席分析师
袁栩聪
15999806788
oliver.yuan@frostchina.com

研究总监
李庆
13149946576
livia.li@frostchina.com

www.frostchina.com ; www.leadleo.com
https://space.bilibili.com/647223552
https://weibo.com/u/7303360042