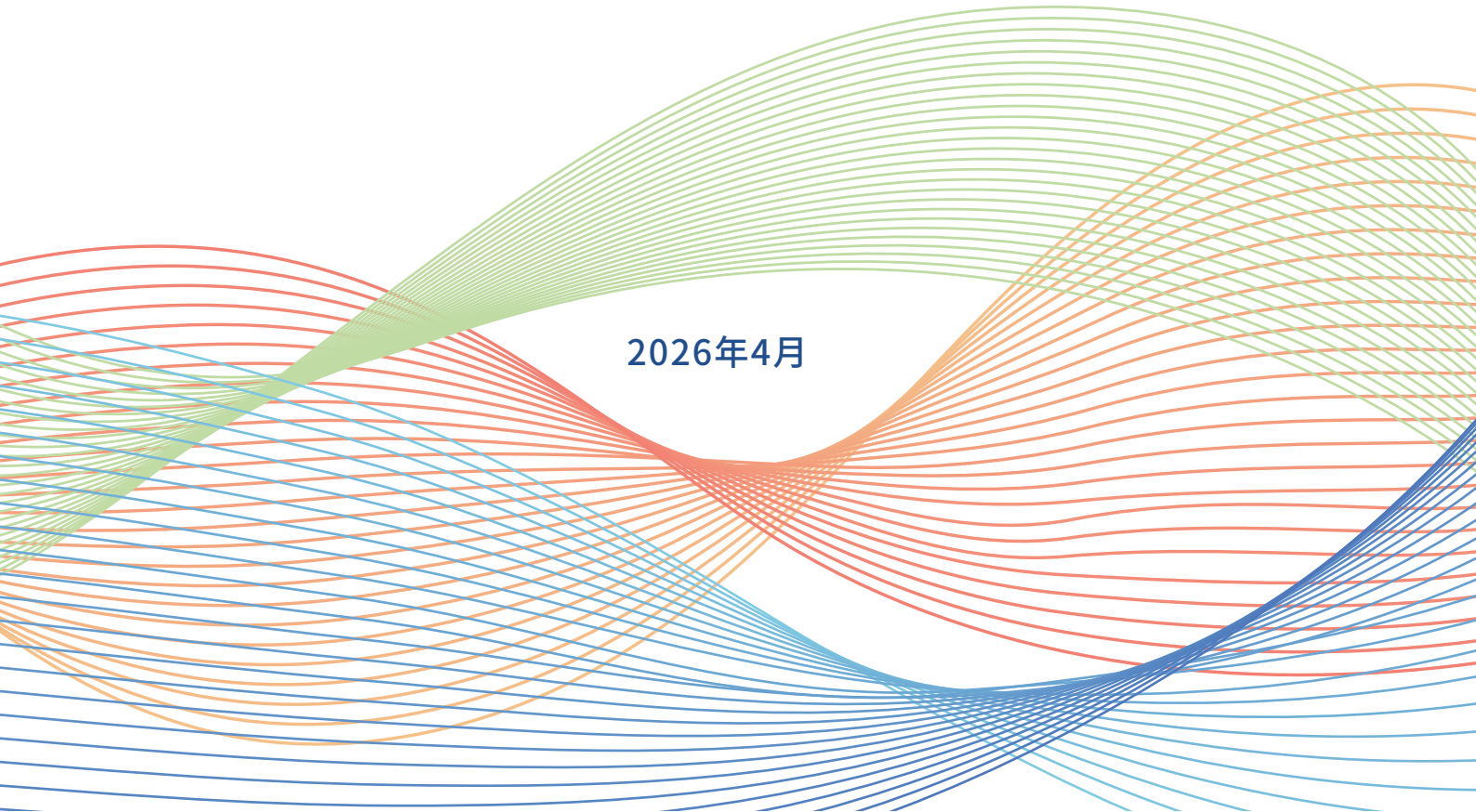
世界互联网大会智库合作计划系列成果

认知对齐·场景深耕·生态协同：AI评测未来核心范式与路径

中国电信北京研究院
中国电信国际有限公司



2026年4月


世界互联网大会
World Internet
Conference


中国电信
CHINA TELECOM
研究院
CTRI

工作组

组长

杨明川，中国电信北京研究院 大数据与人工智能研究所

副组长

王峰，中国电信北京研究院 大数据与人工智能研究所

张园，中国电信北京研究院 大数据与人工智能研究所

林建辉，中国电信国际有限公司 云中台部

工作组成员

丁鹏，中国电信北京研究院 大数据与人工智能研究所

赵君，中国电信北京研究院 大数据与人工智能研究所

刘倩，中国电信北京研究院 大数据与人工智能研究所

郑秋宏，中国电信北京研究院 大数据与人工智能研究所

王禹乔，中国电信北京研究院 大数据与人工智能研究所

赵艺涵，中国电信北京研究院 大数据与人工智能研究所

联系邮箱

zhengqh@chinatelecom.cn

前言

在人工智能技术向通用化、规模化、产业化深度演进的背景下，AI评测已从单一技术验证工具升级为影响全球技术竞争、产业布局与治理规则的核心基础设施。本文立足全球视野，结合全球前沿理论创新与实践，提出未来AI评测的三大核心趋势：以“认知论+”为核心的智能本质对齐、从通用基准到垂直场景的深度渗透、以及平台化支撑下的多元协同治理。本文将系统剖析各趋势的理论逻辑、全球实践路径与产业核心价值，引入全球典型案例，为全球政策制定者、研究机构、产业界提供具有前瞻性与可操作性的智库参考，推动AI评测向更科学、更实用、更具治理效能的方向发展。

目录

一、AI评测的全球战略定位与演进逻辑

（一）AI评测的定义和内涵

AI评测并非孤立存在的评估形式，其核心体系由早期大模型评测逐步拓展演进而来，现已形成覆盖大模型、智能体、AI应用系统及具身智能等多类AI形态的综合性评估范畴。从定义来看，AI评测是依托科学的理论框架、标准化的指标体系与系统化的技术方法，对各类AI系统的能力边界、性能表现、场景适配性、安全风险等核心维度进行量化评估与质性研判的综合性活动。AI评测的价值并非局限于大众熟知的榜单排名形式，其更大核心价值在于通过构建专业化的评测能力、研发标准化的评测工具，将评测深度融入AI研发与生产的全流程，既为研发优化提供精准的方向指引，也为安全风险排查筑牢防线，最终为AI系统的产业选型、监管治理提供客观可信的决策依据，成为连接AI技术供给与产业需求的关键桥梁。

从分类体系来看，当前成熟的AI评测体系已形成多维度的划分标准：按评估对象可分为通用大模型评测、行业大模型评测、AI智能体评测、多模态AI系统评测、具身智能系统评测五大类；按评估生命周期可分为研发期前置评测、上线前合规评测、运行期持续评测三大阶段；按评估核心维度可分为能力评测、安全评测、合规评测、能效评测、公平性评测五大方向，形成了全维度、全周期的立体化评测内涵。

从内涵维度进一步拆解，AI评测的核心价值体系包含三层核心要义：其一，技术维度的“性能度量”，聚焦模型及系统的准确率、响应速度、非幻觉率、鲁棒性等通用技术指标，这是评测体系的核心基础；其二，产业维度的“价值适配”，重点关注AI系统在具体行业场景中的问答准确性、知识检索能力、内容生成质量等行业场景指标，是实现评测与产业需求的深度绑定；其三，治理维度的“风险防控”，涵盖意识形态对齐、隐私保护、伦理合规等核心要求，是评测为AI技术的安全规范发展筑牢的底线。随着人工智能技术的向多形态、全场景演进，AI评测的内涵已从单一技术维度的性能验证，全面拓展为覆盖“技术-产业-治理”的全链条综合评估体系。

（二）AI评测成为全球竞争与治理的核心枢纽

当前，人工智能技术正处于加速迭代、全域渗透的关键发展阶段，其演进趋势不仅决定了AI评测的价值边界，更推动着评测体系的持续升级，具体呈现三大核心趋势：一是从单一任务静态测试转向融合认知科学的动态适配性评估，大模型的参数规模持续扩大、能力边界不断拓展，逐步具备跨领域、跨模态、多任务的通用处理能力，对“智能本质”的评估需求日益迫切；二是从技术研发向产业规模化落地演进，AI技术深度渗透政务、工业等千行百业，场景化适配能力成为衡量技术价值的核心标尺；三是从创新突破向规范治理并重演进，各国纷纷出台AI治理政策，安全可信、伦理合规成为AI技术落地的前置条件，推动AI评测强化风险防控维度的核心作用。

人工智能技术的持续演进与应用场景的不断拓展，使得 AI评测具备了持续迭代的价值基础，不仅成为支撑技术创新与产业发展的重要支撑，更逐步成长为支撑全球 AI 技术创新、产业落地与治理规范的核心基础设施，成为全球科技竞争的“隐形战场”与治理规则的“制定基础”（如图1所示）。在技术层面，它决定着AI研发的方向与效率，引导全球创新资源的分配；在产业层面，统一、可信的评测标准是打破市场信息壁垒、降低技术落地成本的关键；在治理层面，评测体系是将伦理原则、安全要求转化为可操作指标的核心载体，直接影响全球AI治理规则的话语权分配。从欧盟《人工智能法案》将合规性评估作为高风险AI准入条件，到各国推动本土化AI治理框架，将安全与能力评测体系作为核心配套措施，无不印证其战略重要性。

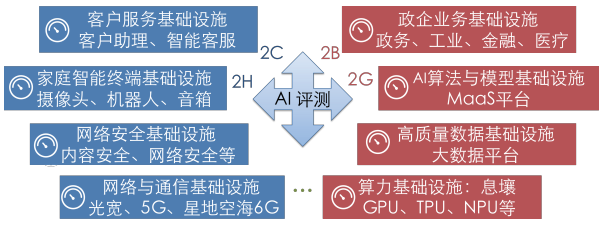


图1 AI评测基础设施

（三）AI评测从技术验证到生态赋能演进

大模型诞生后，AI 评测的发展已历经三个阶段：第一阶段聚焦大模型单一核心性能（如语言生成准确率、知识问答正确率），解决大模型“能否用”的基础问题；第二阶段转向大模型通用能力评估（如多模态理解、复杂推理等），回应“好不好用”的核心需求；当前正进入第三阶段，核心是解决大模型“如何安全、公平、高效地规模化落地”的问题，评测维度从大模型技术性能延伸至认知本质、场景适配、治理合规等多个层面，呈现出“理论深度化、场景精细化、生态协同化”的鲜明特征，如图 2 所示。

第一阶段为 2022-2023 年，是大模型诞生初期的单一性能评测阶段，核心以大语言模型的基础性能评测为主，标志性事件为 GPT-3 的发布（2022 年）及早期大模型的快速迭代。这一阶段大模型刚诞生，技术尚未成熟，评测核心聚焦大模型单一任务的基础性能指标，如语言生成的流畅度、知识问答的准确率、文本分类的召回率等，解决“大模型能否完成基础语言任务”的核心问题。评测对象以早期通用大语言模型为主，评测形式仍以静态数据集的离线测试为核心，整体处于大模型技术验证的辅助工具定位，尚未涉及复杂认知能力与场景适配性的评估。

第二阶段为 2024-2025 年，是大模型迭代期的通用用能力评测阶段，核心以大语言模型、多模态大模型的通用能力评测为主，标志性事件为 OpenAI 的 o1 系列、DeepSeek-R1 等新一代大模型的发布，以及 MMLU、AIME、GPQA 等通用评测基准的广泛应用与升级。这一阶段大模型逐步具备跨领域、多模态的通用处理能力，AI 评测从单一任务性能测试转向多任务通用能力评估，覆盖语言理解、逻辑推理、多模态融合、知识储备、上下文连贯性等多个通用维度，上下文连贯性等多个通用维度，核心解决“大模型通用能力强弱”的评估问题，榜单排名成为这一阶段的主流呈现形式。但随着大模型技术的快速演进，这一阶段的评测体系逐步暴露核心缺陷：全球主流的通用评测基准中，有一定比例已被纳入主流大模型的训练数据，导致测试结果严重失真，大模型“刷榜高分”与“实际应用低分”的矛盾日益突出，推动大模型评测体系进入全新演进阶段。

第三阶段为 2025 年至今，是大模型规模化落地期的全栈式全生命周期评测阶段，核心特征是 AI 评测从单一工具向生态基础设施的全面升级，也是本文核心研究的阶段。这一阶段的大模型已从技术迭代转向规模化产业落地，AI 评测突破了传统的“性能测试”定位，全面围绕大模型的全生命周期与全应用链条展开，向认知本质对齐、垂直场景深耕、生态协同治理三大方向演进。评测对象从基础大模型拓展至面向行业的大模型微调版本、AI 智能体、多模态大模型应用、具身智能（基于大模型驱动）等全链条，评测周期拓展至研发、微调、部署、运营的全生命周期，评测价值从技术验证延伸至产业赋能与治理支撑，成为大模型规模化健康发展的核心基础设施。此处“全栈式”核心指代覆盖“基础大模型 - 行业微调模型 - 大模型应用”的全层级，与“全生命周期”形成“空间 + 时间”的双重维度，完善大模型评测体系的核心内涵。

图2 AI评测发展阶段

二、趋势一：认知对齐——“认知论+”重构AI评测的理论根基

（一）趋势内涵：从“测性能”到“测智能”的本质跃迁

传统AI评测以任务完成度为核心指标，本质是衡量模型的“模式匹配能力”，而非真正的“智能水平”。随着大模型向通用智能演进，“黑箱”问题与能力评估局限性日益凸显——模型可能在基准测试中取得高分，却缺乏常识推理、因果判断等核心认知能力。李德毅院士提出，人类认知的四种基本模式——记忆驱动的经验模式、知识驱动的推理模式、联想驱动的创造模式和假说驱动的发现模式——构成了认知形式化的基础框架，这一形式化为新一代人工智能系统架构提供了理论支撑。通过将人类认知模式抽象为可计算结构，认知形式化使机器能够模拟经验积累、逻辑推理、创造性联想和假设验证等过程，从而实现从计算智能向记忆智能的扩展。

这一趋势的底层逻辑是：AI的终极价值在于模拟并辅助人类智能，其评测体系必须以人类认知为参照系。认知的形式化为构建可交互、会学习和自成长的新一代人工智能系统架构奠定了基础”，只有基于认知科学的评测，才能真正揭示AI系统的能力边界与潜在风险，实现从“测性能”转向“测智能”的本质跃迁。

从认知科学的核心理论来看，这一跃迁的核心必要性源于人类认知的双过程理论（卡尼曼提出的系统1“快速直觉”和系统2“缓慢理性”）：当前主流AI评测体系，大多聚焦于测试模型的“系统1能力”——即基于训练数据的直觉性、自动化模式匹配，而缺乏对“系统2能力”——即慢思考、逻辑推理、因果判断、反思修正的高阶认知能力的有效评估。这也是导致模型“高分低能”的核心原因：模型能够通过模式匹配完成静态测试集的任务，却无法在真实复杂场景中完成需要逻辑推理、因果判断的任务。而“认知论+”评测范式的核心，就是以人类认知机制为核心参照系，重构AI评测的理论根基，实现从“结果导向的性能测试”向“过程+结果双重导向的智能本质评估”的根本性转变。

（二）全球实践：认知科学与AI评测的融合探索

全球顶尖机构已纷纷布局这一方向，形成多元化探索路径，通过理论框架创新，构建基于认知科学的可量化评测体系：

图灵奖得主 Yoshua Bengio 联合斯坦福大学、MIT、加州大学伯克利分校等全球 29 所顶尖研究机构的学者共同完成研究，2025 年 10 月发布论文《A Definition of AGI》，首次建立了可量化的 AGI 评测框架，将 AGI 定义为“在认知多样性与熟练度上，媲美或超过受过良好教育的成年人的 AI”，借鉴心理学权威的 CHC（Cattell-Horn-Carroll）认知能力理论，将通用智能系统拆解为常识与知识、读写能力、数学能力、临场推理、工作记忆等十大核心认知领域，实现对 AI 认知能力的模块化、可量化评估。基于该框架对 GPT-4（2023）和 GPT-5（2025）进行了测试，结果显示：GPT-4 的 AGI 得分为 27%，GPT-5 的 AGI 得分为 57%，首次通过标准化认知框架量化了当前通用大模型与人类认知能力的核心差距。

佐治亚理工学院科学家 Anna A. Ivanova 团队于 2025 年在《Nature Human Behaviour》（Nature 子刊）发表论文《How to evaluate the cognitive abilities of LLMs》，提出涵盖语言理解、工作记忆、注意力控制、因果推理、类比推理、心智理论、元认知、常识推理、道德推理、创造力、问题解决、决策制定、空间认知、数字认知的 14 种认知能力评测方法论，强调通过模拟人类认知过程设计更稳健的评估方案，精准解释 AI 认知能力评估研究的核心结果。

德国人工智能研究中心（DFKI，德国顶级的人工智能研究机构）Nghia Duong-Trung 博士在 ECTEL 2024 会议发表论文《BloomLLM: Large Language Models Based Question Generation Combining Supervised Fine-Tuning and Bloom》，介绍了基于布鲁姆认知目标分类体系，即“记忆、理解、应用、分析、综合、评价、创造”进行大模型微调的方法，是将认知科学方法论与大模型结合的一种探索。

在中国，2025 年 2 月，中国电信研究院联合上海创新算法研究院 Cell 子刊《Patterns》发表论文《Attention Heads of Large Language Models》，借鉴认知神经科学方法，提出了创新性的“知识回忆 - 上下文识别 - 潜在推理 - 表达准备”四阶段的类人脑认

知框架，将 LLM 推理过程与人类推理机制进行了对齐，如图 3 所示。南方科技大学刘泉影博士在 Cell Press 合作期刊 The Innovation 发表论文《Promoting interactions between cognitive science and large language models》提出借鉴认知科学的方法论来评估 LLMs 的智能水平、心理状态、道德水平等，为 LLMs 的多维度智能评估提供理论和方法论。

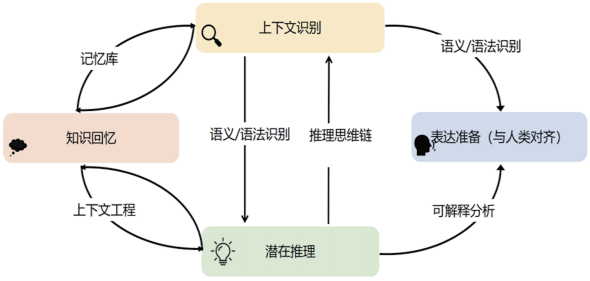


图3 类人脑认知论框架

（三）核心价值：破解通用智能评估的根本性难题

针对这一趋势，中国电信研究院提出了“认知论+”评测范式，其核心是将认知科学理论与AI评测深度融合，通过对齐人类认知机制，推动评测从“结果导向”转向“过程+结果”双重导向，对大模型“输入-处理-输出”全链路的认知能力进行评估，实现对AI智能本质的精准度量。

“认知论+”评测范式的价值在于三个层面：一是解决“黑箱”问题，通过模拟人类认知过程，让AI决策逻辑可解释、可追溯；二是提升评测的前瞻性，基于认知规律预判AI在复杂场景中的表现，而非依赖静态数据集；三是统一通用智能的评估标尺，为全球AGI发展提供可对比、可协同的度量体系。

从产业落地与安全治理的维度来看，“认知论+”评测范式的核心价值进一步延伸：在产业层面，基于认知对齐优化后的AI模型，能够更好地适配复杂的行业业务场景，尤其是需要逻辑推理、因果判断、创造性解决问题的高价值场景，大幅提升AI应用的业务价值与用户满意度。在安全治理层面，认知对齐是AI安全对齐的核心基础，只有实现了AI与人类认知机制的对齐，才能从根本上实现AI与人类价值观、伦理规范的对齐，防范AI的灾难性风险。

三、趋势二：场景深耕——从通用基准到垂直领域的精准渗透

（一）趋势内涵：产业落地倒逼评测的场景化转型

企业级AI应用的落地率在逐年提升，但难以持续为企业产生可量化的业务价值，核心瓶颈在于缺乏适配行业场景的全流程评测体系。当前AI评测面临的核心痛点之一，是通用基准与产业实际需求的脱节——模型在MMLU、GLUE等通用基准上的高分，往往难以转化为具体行业的实用价值。随着AI技术从“实验室走向产业界”，评测体系必须向垂直场景深度渗透，针对不同行业的业务逻辑、数据特征、安全要求，构建精细化、定制化的评估框架。这一趋势的核心是“以评测促落地”，通过场景化评测打通AI技术与产业价值之间的“最后一公里”。

场景化评测的核心本质，是从“以模型为中心”的通用性能测试，转向“以业务为中心”的全维度价值评估，其核心遵循四大原则：一是业务导向原则，所有评测指标均围绕行业核心业务目标设计，而非通用技术指标；二是真实环境原则，评测数据与测试场景完全还原行业真实业务环境，包含噪声数据、边缘案例、突发状况等真实场景中的复杂因素；三是全生命周期原则，评测贯穿AI应用的需求设计、模型训练、上线部署、运营维护全流程，而非上线前的一次性验收；四是风险前置原则，在评测过程中提前识别场景中的合规风险、安全风险、业务风险，实现风险的早发现、早处置。

（二）全球实践：行业定制化评测的多元探索

全球已在多个高价值领域形成成熟实践，印证了场景化评测的必要性：

对于高风险领域，如：医疗场景中，评测聚焦诊断准确率、跨设备泛化能力、临床可解释性，如美国FDA对医疗AI的审批要求包含多中心临床试验数据支持的特异性、敏感性指标；金融领域，反欺诈模型评测不仅关注预测准确率，更强调公平性（避免对特定群体的歧视）、对抗性鲁棒性（抵御欺诈手段迭代）与监管合规性（满足反洗钱、数据隐私要求）。

对于垂直行业场景，如：客户服务领域，中国电信针对客服行业构建评测体系，围绕话术推荐、报告生成、话务小结等业务场景，设计针对性评测指标，实现评测与业务流程深度嵌入，助力行业场景下AI应用生产效率提升；政务场景中，针对社保、户籍咨询等需求，评测融入“用户意图理解-政策知识推理-合规表达”全流程指标，确保AI应用既高效又安全。此外，由中国电信、中国移动、中国电子技术标准化研究院联合牵头的“慧聚智评”大模型评测工作组在行业大模型评测领域已形成规模效应，建立起了电力、物流、石化、交通等行业的体系化评测标准，联合多家相关企业正开展评测实施工作，如：2025年已面向5家电力行业开展评测，深度聚合生态力量。

对于未来产业布局，在具身智能、自动驾驶、低空经济等新兴领域，场景化评测提前卡位标准制定。如自动驾驶评测已从封闭场地测试转向真实道路的“野外”评估，涵盖恶劣天气、突发状况等复杂场景；中国电信针对人形机器人、智能无人机等载体，构建“环境感知-路径规划-应急决策”场景化评测指标，为未来产业发展奠定基础。

（三）核心价值：加速AI产业的规模化落地

场景化评测的核心价值在于“精准匹配”——为企业提供贴合自身需求的技术选型依据，降低试错成本；为开发者提供明确的优化方向，避免技术研发与市场需求脱节；为监管机构提供针对特定领域的风险评估工具，实现“精准治理”。从全球范围看，场景化评测已成为AI产业成熟度的重要标志，其普及程度直接决定着AI技术向千行百业渗透的速度与深度。

从产业供需两端的底层逻辑来看，场景化评测的核心价值，在于打破了通用评测体系与产业真实需求之间的结构性错位，推动AI评测从“技术本位”向“价值本位”的根本性转型。对于技术需求方而言，场景化评

测通过锚定产业真实业务目标构建评估标尺，消解了通用技术指标与实际业务价值之间的信息差，让市场主体能够摆脱“唯参数论”“唯榜单论”的选型误区，建立起以业务价值为核心的选型决策体系，从根源上降低AI技术落地的试错成本与决策成本，同时也为不同规模、不同数字化基础的市场主体提供了平等的技术选型参考依据，为AI技术的普惠化应用扫清了认知与决策壁垒。

对于技术供给方而言，场景化评测将模糊、分散的产业需求，转化为可量化、可落地的技术优化目标，彻底打破了技术研发与市场需求脱节的行业痛点。这一评测体系能够引导研发资源从单纯的参数规模竞赛、通用榜单刷分，转向能够真正解决产业实际痛点的技术创新，实现研发效率与产业价值的双向提升。同时，场景化评测为AI技术迭代建立了清晰的、贴合产业需求的反馈闭环，让技术优化能够始终围绕产业真实需求展开，推动AI技术从实验室创新向产业实用化的持续、稳定演进。

在治理层面，场景化评测是实现AI分级分类精准治理的核心载体，推动AI治理从“一刀切”的通用规则，向适配不同领域特征的精细化治理转型。通过贴合不同行业业务逻辑、风险等级、合规要求构建的差异化评估体系，场景化评测能够将宏观的伦理原则、安全规范、合规要求，转化为可操作、可验证、可追溯的具体评估指标，既能够精准识别不同场景下AI系统的潜在风险，牢牢守住安全发展的底线，又能够避免过度监管对技术创新的制约，实现AI产业发展与安全的动态平衡，为全球AI治理体系的落地执行提供了坚实的工具支撑。

从产业发展的全局维度来看，场景化评测的成熟度，是AI产业从技术探索期进入规模化落地期的核心标志。当AI评测体系完成从通用基准向垂直场景的深度渗透，意味着AI技术与实体经济的融合进入了体系化、标准化的新阶段。场景化评测能够为AI技术在千行百业的落地建立统一的行业共识与价值标尺，打破不同行业、不同主体之间的技术壁垒与信息壁垒，加速AI技术从单点应用向全产业链、全业务流程渗透的进程。其在各行业的普及深度与标准化程度，不仅决定了AI技术向实体经济渗透的速度与广度，更从根本上决定了 AI 产业整体的发展质量与可持续增长潜力。

四、趋势三：生态协同——平台化支撑与治理化升级的双重驱动

（一）趋势内涵：从单一工具到协同生态的体系进化

AI评测的复杂性日益提升，单一机构、单一工具已难以满足全产业链的需求。未来评测体系将呈现“平台化支撑+治理化升级”的双重特征：平台化解决“易用性、规模化”问题，通过一站式、端到端工具降低评测门槛，覆盖从普通用户到研发者、监管者的全群体；治理化解决“公信力、统一性”问题，通过建立透明、独立、可问责的治理框架，确保评测结果的客观性与权威性。二者协同构建开放、协同的全球AI评测生态。这一趋势的底层逻辑是：AI评测已成为全球治理的“事实上的工具”，其自身必须具备完善的治理体系；同时，产业界对评测的规模化、低成本需求，推动着工具的平台化整合。

从产业链分工来看，AI评测生态已形成清晰的六大核心环节：一是评测理论与标准研究机构，负责构建AI评测的理论框架与标准体系；二是评测数据集提供商，负责构建标准化、场景化的评测数据集；三是评测工具与平台开发商，负责研发评测工具、搭建评测平台，实现评测的工程化、规模化落地；四是第三方独立评测机构，负责提供客观、公正的第三方评测服务，出具评测报告；五是产业应用方，包括政府、企业等AI系统的使用方，是评测服务的核心需求方；六是监管机构，负责制定评测的监管规则，规范评测行业的发展。只有六大环节协同发力，才能构建完整、健康、可持续的AI评测生态，而平台化是生态协同的核心载体，治理化是生态协同的核心保障，二者相辅相成，缺一不可。

（二）全球实践：平台建设与治理框架的并行推进

在平台化实践方面，国际上，OpenAI推出了Evals开源评估框架，可为研究者和开发者提供了一套标准化的评估任务和架构，以便比较不同的 LLMs 各方面的性能；Google（Vertex AI）、Amazon（Bedrock）等企业推出内置评测工具的MaaS平台，实现模型部署与评测的一体化。在中国，除了司南等主打以榜单构建影响力的平台外，逐渐催生了以专攻AI全栈评测能力而打造的独立平台，如：中国电信“天罡”AI评测平台以“一站式、端到端、可视化”为核心，跟踪适配全球主流大模型、自建评测数据集，覆盖超200个通用模型、40个行业模型、100个细分场景，实现从普通用户（可视化操作）到研发者（私有化部署或API接口）、监管者（安全合规专项模块）的全群体适配，降低评测使用门槛，如图4、图5所示。

图4 天罡AI评测平台通用大模型榜单

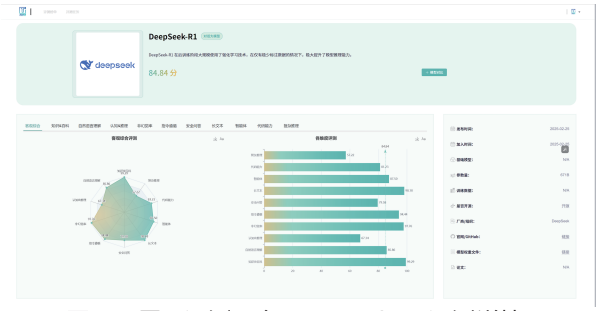


图5 天罡AI评测平台DeepSeek-R1评测详情

在治理化探索方面，全球正围绕“谁来评估评估者”的核心问题，构建评测治理框架。欧盟《AI法案》明确要求对高风险AI评测的机构需具备独立性与透明度；学术研究将“评测卡”机制延伸至评测领域，倡导评测/模型发布方披露设计原理、数据构成、潜在偏见等信息；世界经济论坛（WEF）推动建立全球性AI评测治理框架，促进跨国家、跨行业的标准互认。其中，在中国的AI发展土壤下还需培育适合本土化发展的治理方案，如：中国电信在评测实践中，将意识形态、价值观对齐纳入核心指标，形成符合我国治理要求的合规评测

模块，为治理化提供了本土实践样本。

（三）核心价值：构建可信、普惠的全球AI生态

平台化与治理化的协同，将实现两大价值：一是普惠性，通过低门槛工具让更多企业、机构享受到专业评测服务，缩小“数字鸿沟”；二是公信力，通过透明、独立的治理框架，提升评测结果的全球认可度，打破“基准操纵”“各说各话”的碎片化局面。最终，构建一个开放协同、可信可靠的全球AI评测生态，为AI技术的健康发展提供体系保障。

从全球AI治理的维度来看，生态协同的核心价值在于，通过全球统一的评测标准与治理框架，降低AI技术跨国流动的合规壁垒，促进全球AI产业的协同创新与健康发展。AI技术是全球性的技术，其治理也需要全球协同，而AI评测是全球AI治理落地的核心抓手，只有实现全球评测标准的互认，才能实现全球AI治理规则的协同，避免形成“数字壁垒”与“治理分裂”。从AI安全的维度来看，生态协同的评测体系，能够汇聚全球的智慧与资源，更早地发现前沿AI系统的潜在风险，构建全球统一的AI安全防线，防范AI技术的滥用与灾难性风险，保障AI技术的安全可控发展。从普惠发展的维度来看，平台化的评测工具，能够大幅降低AI评测的技术门槛与成本，让中小企业、发展中国家能够平等地享受到专业的评测服务，缩小发达国家与发展中国家、大型企业与中小企业之间的AI能力鸿沟，推动AI技术的普惠应用，实现“智能向善”的核心目标。

五、全球AI评测发展的挑战与建议

（一）AI评测面临的核心挑战

1.评测技术层面

认知评测技术瓶颈：对常识推理、因果判断、创造力等高阶认知能力的量化评估方法仍不成熟，现有技术难以全面衡量 AI 的智能本质。

前沿技术评测能力滞后：AGI、具身智能、多模态 AI 智能体、端侧 AI 等前沿技术加速迭代，而对应的评测理论、方法、工具的发展相对滞后，出现了 “技术先行、评测滞后” 的局面，无法为前沿技术的安全可控发展提供有效支撑。

2.场景落地层面

场景化数据壁垒：垂直行业的真实数据往往涉及隐私或商业秘密，难以共享，制约了场景化评测的发展。

评测与研发全流程融合不足：当前多数企业的 AI 评测仍停留在上线前的 “一次性验收” 阶段，未将评测融入 AI 研发、部署、运营的全生命周期，导致研发过程中的风险无法提前发现，模型优化缺乏精准的方向指引，出现 “评测与研发两张皮” 的问题。

3.生态协同层面

标准碎片化：全球范围内缺乏统一的评测标准，不同机构、国家的基准差异较大，导致评测结果难以互认，增加了企业合规成本与技术跨国流动难度。

评测的公平性与普惠性不足：当前全球主流的评测基准、数据集大多基于英语语言与西方文化背景，对非英语语言、发展中国家的本土场景适配性严重不足，导致评测结果存在系统性偏见，加剧了全球 AI 发展的数字鸿沟。

治理体系不完善：评测机构的独立性、透明度、问责制缺乏全球统一的规范，可能导致评测结果被滥用，影响其公信力。

（二）AI评测发展建议

1.对政策制定者：推动标准互认与治理协同

（1）加强国际合作，主导或参与全球AI评测标准制定，推动基于认知科学的核心指标互认，打破标准碎片化格局。

（2）加大对AI评测核心技术研发的资金支持，如设立专项科研基金，支持认知对齐评测、场景化评测、AI安全评测、具身智能等新场景的核心评测技术的研发，突破技术瓶颈。

（3）在保障数据安全与隐私的前提下，推动公共数据与行业高质量数据集的开放共享，为场景化评测提供数据支撑。

（4）培育第三方独立评测机构，建立完善的评测机构资质认证体系、监管规范与退出机制，规范评测行业的发展，提升评测行业的整体专业水平与公信力。

（5）建立评测机构的监管框架，明确独立性、透明度要求，确保评测过程可追溯、可监督。

2.对产业界：拥抱场景化与平台化，践行负责任评测

（1）积极参与场景化评测标准共建，将自身业务经验转化为行业通用指标，推动评测与产业需求深度绑定。

（2）加强与研究机构、高校的产学研用协同合作，联合开展AI评测理论研究、技术研发与标准制定，推动评测技术的创新与落地。

（3）依托平台化工具开展全生命周期评测，将评测融入AI产品的设计、开发、部署全流程，提前规避风险。

（4）建立企业内部的AI评测治理体系，明确评测的责任部门、流程规范、质量标准，将AI评测作为AI产品上线的强制前置条件，落实企业主体责任。

（5）坚持透明化原则，公开自身AI产品的评测方法与结果，避免“刷榜”等投机行为，共同维护评测生态的公信力。

3.对研究机构：深化跨学科研究，突破技术瓶颈

（1）聚焦前沿技术的评测研究，针对AGI、具身智能、AI智能体等前沿技术，提前布局评测理论与方法的研究，实现评测与技术创新的同步发展。

(2)加强AI评测与认知科学、心理学、伦理学、法学的跨学科融合,探索高阶认知能力的量化评估方法,完善“认知论+”评测范式。

(3)研发动态化、交互式评测技术,模拟真实世界的复杂性与不确定性,弥合“实验室性能”与“生产效果”的差距。

(4)关注评测的社会影响,研究公平性、普惠性相关指标,推动AI技术向“智能向善”方向发展。

(5)加强国际学术交流与合作,深度参与全球AI评测理论与技术的研究,推动全球评测理论与方法的协同创新。

4.对第三方评测机构:坚守独立公正,提升专业能力

(1)坚守独立性、客观性、公正性的核心原则,建立完善的内部治理体系与利益冲突隔离机制,杜绝与评测对象存在利益关联,确保评测结果的公信力。

(2)持续加强自身的技术能力建设,跟进AI技术的迭代演进,不断提升对前沿AI技术的评测能力,为产业界与监管机构提供专业、权威、全面的评测服务。

(3)严格遵守全球各国的AI监管法规与行业标准,规范评测流程,完整记录评测全流程信息,确保评测过程可追溯、可审计、可复现。

5.对国际组织:推动全球协同,促进普惠发展

(1)发挥国际组织的协调作用,搭建全球AI评测标准协同平台,推动全球AI评测标准的统一与互认,解决标准碎片化的核心问题。

(2)建立全球AI评测资源共享平台,推动评测数据集、工具、方法、算力资源的全球共享,缩小发达国家与发展中国家之间的AI评测能力鸿沟。

(3)推动全球AI评测治理的协同,制定全球统一的AI评测治理原则与指引,规范全球AI评测行业的发展,提升AI评测的全球公信力。

六、结论

未来AI评测的发展将围绕“认知对齐、场景深耕、生态协同”三大核心趋势展开，这不仅是技术演进的必然结果，更是全球AI产业健康发展与有效治理的客观需求。“认知论+”范式强化了评测的理论根基，实现从“测性能”到“测智能”的跃迁；场景化评测打通了技术与产业的连接，加速AI规模化落地；平台化与治理化协同构建了可信、普惠的生态体系，为全球AI发展提供保障。

从长远来看，AI评测的发展，将深刻影响全球AI技术的发展方向、产业格局与治理规则。在通用人工智能加速演进的背景下，AI评测不再是AI技术的“配套工具”，而是决定AI技术安全可控发展、规模化产业落地的核心基础设施，是全球AI科技竞争与治理话语权争夺的核心制高点。

面对全球竞争与治理挑战，需要各国、各机构秉持开放协同的态度，共同推动评测标准互认、技术创新与治理完善。全球机构的实践，已为全球AI评测提供了“认知对齐+场景落地+平台支撑”的初阶方案。未来，随着三大趋势的深入演进，通过持续深化“认知论+”评测范式的理论创新，推动场景化评测的全行业普及，构建开放协同的评测生态，AI评测将真正成为全球AI技术创新的“指南针”、产业落地的“度量衡”与治理规则的“基石”，为构建人机共存的智能时代提供坚实支撑。