



『弈衡』人工智能大模型 评测平台白皮书

(2024 年)

发布单位：中移智库

编制单位：中国移动通信研究院

目 录

- 第一章 引言 1
- 第二章 人工智能大模型评测需求 3
 - 2.1 应用需求 3
 - 2.2 评测内容 4
- 第三章 业界大模型评测平台现状 6
 - 3.1 业界典型大模型评测平台 6
 - 3.2 各平台的优劣势分析 8
- 第四章 中国移动“弈衡”大模型评测平台 10
 - 4.1 设计原则和思路 10
 - 4.2 平台整体架构 11
 - 4.3 平台特色及创新点 12
 - 4.4 平台主要功能 13
 - 4.5 平台主要优势 16
 - 4.6 成功案例与应用实践 17
- 第五章 未来发展展望 19

第一章 引言

人工智能技术的迅猛发展带来了AI大模型的广泛应用，这些模型在自然语言处理、图像识别、数据分析等领域展现出巨大潜力。然而，随着模型规模和复杂性的增加，如何发现模型的长短板并开展针对性优化，如何在实际应用场景中选取合适的模型，已成为大模型评测领域亟待解决的新挑战。目前的人工评测方法存在效率低下、评测组织有效性不足、结果管理无序和评测体系更新缓慢等问题，这些问题制约了AI技术的健康发展和创新应用。

为了解决这些挑战，我们确立了构建一个高效、系统化且可信赖的AI大模型评测平台的愿景。该平台将通过自动化和智能化技术，提升评测效率，确保流程的组织性和管理的有序性，并支持评测体系的持续拓展和演进。具体愿景如下：

1. 提高评测执行成效

为了提升AI大模型的评测成效，评测平台必须适应模型规模和复杂性的增加，能够细致分析和处理大量的测试案例，并考虑到多样化的数据集和应用场景。同时，由于不同应用领域对AI大模型的评测标准和方法有着不同的需求，评测平台应提供多样化的评测方式以满足这些需求。此外，通过提升自动化程度，可以降低人工测试成本，实现常态化的评测。自动化评测平台通过预设的流程和算法能够快速评估模型性能，这不仅加快了研发和迭代过程，而且也便于集成最新的评测技术，进而缩短整个评测周期。

2. 加强评测全面客观性

为了加强AI大模型评测组织的效率和有效性，首先需要实现评测流程的标准化，这有助于减少人为差异和错误，同时模块化设计确保了评测流程的定制性和灵活性。其次，确保评测结果的一致性和可重复性至关重要，这可以通过消除人为不一致性来实现，从而保证每次评测都能得到相同的结果，并且评测过程可以无限次地重复。此外，集成多样化的评测体系是提高评测全面性的关键，这不仅包括准确率，还应涵盖伦理考量等多个维度。最后，提供动态和实时的反馈机制，可以帮助开发者迅速了解模型性能，从而实现快速迭代和优化。通过这些措施，可以显著提升评测组织的效率和有效性，为AI大模型的持续改进和创新提供坚实的基础。

3. 增强结果管理规范

为使AI大模型评测平台的结果管理更加规范，首先，平台应能够自动记录和存储评测结果，这不仅能减少错误率，而且确保了数据的完整性和可追溯性。其次，它配备强大的查询

功能，使得用户可以轻松地查询和比较不同评测周期的结果，从而提高结果分析的效率，并快速识别出模型性能的变化趋势。第三，平台支持对模型的表现进行长期跟踪与评估，使用户能够全面评估模型的稳定性和泛化能力，同时监测性能退化，及时发现并解决潜在问题。最后，集成工具使用户能够更直观地理解复杂数据，并通过图表等形式把握关键信息，从而辅助决策过程。综合这些特性，AI大模型评测平台为结果管理提供了一个有序、高效且易于操作的平台，极大地促进了模型的深入分析和持续改进。

4. 保障评测体系演进性

AI大模型评测平台致力于保障评测体系的持续演进，首先，平台能够快速响应市场变化，及时吸收最新的研究成果和技术突破，以此来更新评测框架，确保评测体系的时效性和前沿性。其次，它支持持续的研发创新，可通过全面而深入的评估，精准地识别出潜在的问题和改进空间，为模型的迭代和优化提供指导，从而激发创新思路。第三，平台促进技术标准迭代，通过建立统一的评测标准和流程，减少不同技术之间的差异，推动整个行业的标准化和规范化，降低技术应用的门槛。通过这些措施，AI大模型评测平台不仅保障了评测体系的持续改进，也为AI技术的健康发展和广泛应用奠定了坚实的基础。

通过实现这些愿景，我们旨在构建一个能够适应AI大模型技术快速发展的评测平台，为AI技术的持续进步和应用提供坚实的支撑。我们号召整个产业链参与进来，共同构建一个开放、协作的环境，加速AI技术的创新和应用。一是快速响应市场变化，及时整合最新的研究成果和技术突破，确保我们的评测框架始终处于行业前沿。二是支持持续的研发创新，全面评估AI模型的性能，发现问题，为模型的迭代和优化提供科学指导，激发更多的创新思路。三是致力于促进技术标准的迭代，建立统一的评测标准和流程，减少技术差异，推动整个行业的标准化和规范化，降低技术应用的门槛。通过产业链的共同努力，我们不仅能够保障评测体系的持续演进，还能推动AI技术更广泛、更高效的应用，为社会的发展贡献更大的价值。

第二章 人工智能大模型评测需求

AI大模型以其庞大的参数数量、卓越的学习能力和广泛的应用潜力，正在引领AI技术的新一轮革命。AI大模型的参数规模是其最显著的特点之一。庞大的参数量使得模型能够捕捉和学习数据中的细微特征和复杂关系，从而在各种任务上实现优异的性能，不仅能够理解自然语言、识别图像和视频内容等，还能在复杂的策略游戏中与人类玩家竞争，甚至超越人类的表现。但在某些领域中，模型的准确性和安全性还有待提升。大模型的“黑箱”特性使得理解和解释其行为变得困难，这对于需要高度透明度和可解释性的应用场景构成挑战。

2.1 应用需求

随着大模型技术在实际应用中的广泛应用，如何全面、准确地评估这些模型的性能、安全性、合规性以及其对社会的影响，成为一个迫切需要解决的问题。在大模型研发和应用过程中，面临着一系列挑战，在技术验证、质量控制、风险管理以及合规性等多个层面提出评测需求。

一是在技术验证方面，人工智能大模型评测验证及应用是企业引入大模型以支持其业务流程的关键环节。首先需要验证模型的性能是否能够满足特定应用场景的技术要求，这包括但不限于模型的预测准确性、处理速度、以及在特定数据集上的表现。技术验证是确保大模型能够在实际应用中发挥预期作用的基础步骤，对于企业来说，这一步是不可或缺的，它使企业可以更有信心地将大模型集成到其业务流程中，从而实现技术升级和业务增长。

二是在质量控制方面，大模型的输出质量直接关系到其应用的效果和企业的声誉。因此，通过定期的评测来确保大模型的输出质量和决策的准确性变得尤为重要。质量控制旨在减少或消除可能由模型引起的错误信息传播，保障用户能够获得可靠和准确的服务。这对于提升用户信任和满意度至关重要，同时也有助于企业及时发现并修正模型的潜在问题。

三是在风险管理方面，大模型可能带来的安全风险是业界关注的焦点。这包括但不限于数据偏见、性别或种族歧视、隐私泄露等问题。风险管理需求推动对大模型进行深入评测的需要，以便及时发现并解决这些问题，确保模型的公正性和安全性。随着AI技术的广泛应用，这些风险管理措施变得尤为重要，它们有助于构建一个更加安全、公正的AI生态系统。

四是在合法合规方面，随着各国对AI技术的法律法规日益完善，确保大模型的应用符合法律法规和伦理标准成为企业必须面临的挑战。合规性需求促使企业必须对大模型进行评测，以验证其是否遵守相关的法律法规，从而避免潜在的法律风险。这不仅是法律的要求，也是

企业社会责任的体现，有助于提升企业的社会形象和公众信任。

2.2 评测内容

在人工智能领域，大模型的评测是一个多维度、多目标的复杂过程，它涉及到模型的功能、性能、安全性、可解释性、可靠性、经济性等多个方面。评测内容的广泛性要求我们不仅要关注模型在特定任务上的表现，还要考虑其在不同环境和条件下的稳定性和泛化能力。随着AI技术的快速发展，大模型在各个领域的应用日益广泛，对大模型的评测需求也日益增长。

在功能性方面，大模型评测集中在三个核心领域：任务特定性能、多任务能力以及交互能力。对于任务特定性能，大模型在执行如语言翻译或图像识别等特定任务时，必须展现出出色的表现。这涉及到模型的准确度、处理速度以及对不同数据集的适应性。多任务能力则强调大模型在广泛任务类型上的泛化能力，评测时需综合考量模型在多样化任务上的表现，包括其对未知任务的适应性和表现。此外，交互能力评测关注于大模型与用户交互的流畅性和有效性，包括模型对用户输入的理解、回应的自然性，以及对话上下文的关联性，这些都是提升用户体验的关键因素。综合这些功能性需求的评测，可以全面地评估大模型在实际应用中的性能和潜力。

在性能方面，大模型的评测集中在准确度、效率和可扩展性等关键维度。准确度是评测的核心，它直接决定模型输出的可靠性，尤其在医疗和法律等高风险领域，准确度的水平直接影响着决策的质量。效率则涉及模型处理输入并提供输出的速度，以及在整个计算过程中所消耗的计算资源和能源。一个高效率的模型能够在维持性能标准的同时减少用户的等待时间，并有效降低运行成本。可扩展性是评测模型是否能够适应未来业务需求的变化，例如处理更大数据集或更复杂的任务。这要求模型具备通过增加资源投入或参数调整来提升性能的能力，从而保障其长期的实用价值和适应性。综合这些性能指标的评测，可以确保大模型在满足当前需求的同时，也具备未来技术发展和业务扩展的潜力。

在安全性方面，大模型的评测覆盖数据隐私、系统安全和伦理合规这三个基本而关键的领域。数据隐私强调在处理敏感信息时，模型必须实施有效的保护措施以防止数据泄露，评测过程中需要严格检查模型是否具备健全的隐私保护机制。系统安全则关注模型对于恶意攻击的防御能力，包括对抗性攻击和模型滥用等风险，确保模型在面对这些潜在威胁时的安全性和稳定性。伦理合规要求模型在设计和应用过程中遵循社会伦理标准和法律法规，防止产生歧视或不公平的结果，确保技术应用的社会影响正面且负责任。综合这些安全性指标的评测，旨在为大模型提供一个安全、可靠且稳定的运行环境，保障其在实际部署和应用中的安全性和可信赖性。

在可靠性方面，大模型的评测着重于鲁棒性和一致性两个关键指标。鲁棒性指的是模型对于输入噪声和异常情况的抵抗能力，这确保即使在面对不完美或有敌意的输入数据时，模型也能够维持其性能，稳定地输出结果。而一致性则评价模型在不同时间点、处理不同输入时是否能提供连贯且一致的响应，这对于建立用户对模型的信任至关重要。如果模型的输出表现不稳定，或者在类似情境下产生大相径庭的结果，可能会导致用户对模型的不信任，影响其长期的可用性和用户满意度。因此，一个高度可靠和稳定的大模型，需要在各种条件下都能提供可预期且一致的性能表现。

第三章 业界大模型评测平台现状

针对上述大模型评测需求,诸多科研机构 and 头部公司已经开展了广泛的相关研究与实践。我们调研了业界典型大模型评测平台,充分掌握平台能力以及它们在实际应用中的表现,并分析了这些平台的优势和不足。

3.1 业界典型大模型评测平台

- **OpenCompass**

OpenCompass是由上海AI实验室精心打造的一个开源、全面且高效的大模型评测平台。该平台的评测框架完全开源,使用户能够自由访问和修改评测代码,以满足不同的评测需求,同时保证评测过程的透明性和可复现性,便于学术界和工业界的研究人员进行结果的验证和比较。OpenCompass支持包括零样本、小样本和思维链在内的多样化评测方式,构建一个多维度的评估体系,全面反映模型在知识、语言、理解、推理等方面的能力。

为应对大规模参数量模型的评测挑战,OpenCompass提供分布式评测方案,支持单机或集群上的并行处理,显著提高评测效率。通过巧妙的任务分割与合并策略,实现计算负载的均衡,充分利用所有可用的计算资源,使得在资源充足时,千亿参数量级模型的全面评测可以在极短时间内完成。此外,OpenCompass不仅支持开源模型的评测,还为开发者提供简洁的模型接口,方便接入API模型,已经支持OpenAI、Claude、PaLM等众多API模型的评测。

OpenCompass的设计灵活性允许用户通过预留接口轻松添加新的评测数据集和模型,自定义数据划分策略,甚至接入新的集群管理后端,极大地拓展了评测模型的类型和范围。作为一个公开的评测平台,OpenCompass向技术社区全面开放,支持的数据集及各数据集的多版本提示词都可以供用户一键下载,确保了评测结果的可复现性。

在多模态能力评测方面,OpenCompass覆盖从感知到认知能力的多个细粒度评估维度,如目标检测、文字识别、动作识别、图像理解、关系推理等,为多模态大模型提供全面的能力评估。同时,OpenCompass提供公开的评测榜单,降低评测门槛,鼓励技术社区提交评测申请,评测结果将实时更新在榜单上,为用户提供直观的性能比较。

● AgentBench

AgentBench是一个创新的多维演进基准测试平台，专为评估大语言模型（LLMs）在复杂的多轮次开放式生成环境中的推理和决策能力而设计。该平台通过模拟真实世界的场景和任务，提供一个全面的性能评估框架，旨在推动大语言模型在理解、推理、学习和适应性方面的进步。AgentBench的评估体系是多维的，它不仅关注模型的语言生成能力，还包括对模型在复杂情境下的推理、决策、学习能力和适应性等方面的综合评估。这种多维的评估方法可以更准确地反映模型在现实世界任务中的性能，为研究者和开发者提供一个更为全面的模型能力视图。

AgentBench模拟多轮次的开放式对话环境，考察在连续的交互中保持上下文的连贯性和逻辑的一致性。这种环境对模型的记忆、注意力和长期策略规划能力提出更高的要求，是对模型综合能力的考验。推理和决策是AgentBench评估的核心。平台设计一系列任务，旨在挑战模型的推理能力，包括归纳推理、演绎推理、因果推理等。同时，AgentBench还评估模型在面对复杂决策时的表现，包括风险评估、成本效益分析和多目标优化等。

AgentBench提供高度的定制化和扩展性，允许研究者和开发者根据自己的需求设计和添加新的评估任务。这种灵活性使得AgentBench可以适应不同的研究领域和应用需求，促进评测方法的创新和发展。AgentBench鼓励社区合作和共享。平台的设计允许不同的研究团队和开发者共享他们的评估任务和模型，从而形成一个开放和协作的研究环境。这种合作精神有助于推动大语言模型评测技术的发展，并加速模型性能的提升。

● OpenEval

OpenEval是一个专注于中文大模型评估的多维度、全面、开放的评测平台。该平台由大模型基准评测专家委员会研制开发，旨在探索前沿模型评测方法，建立大模型评测标准与协议，并建设高质量的大模型评测基准数据。通过综合评估大模型的知识、能力、对齐及安全级别，OpenEval发布详细的评测报告，以助推中文大模型能力与安全协同发展。OpenEval平台从四大维度出发进行评估，包括知识能力、价值对齐、安全可控和专业领域。每个维度下又细分为多个子维度，并收录相应的数据集。这种细分不仅有助于全面评估模型性能，也进行针对性的优化和改进。

OpenEval平台提供自动化的评测流程，支持本地评测和平台评测两种方式。用户可以在本地完成模型推理后，生成待评测文件并在平台上提交，以获得评测结果。此外，平台还支持API评测方式，用户提交模型API信息后，结果将以邮件形式发送。OpenEval平台收录多种数据集，如GaoKao-bench、TGEA和CSNLI等，这些数据集覆盖高考题目、文本错误检测、自然语言推理等多个领域，为大模型的评估提供丰富的基准。

● 百度大模型内容安全评测平台

百度大模型评测平台是一个专为大语言模型（LLMs）设计的综合性评测系统，提供内容安全等多维度的评测服务。该平台的建立和运作严格遵守国家法律法规，如《网络安全法》和《互联网信息服务管理办法》，确保大模型输出的内容合法合规，避免不当内容的产生，保护用户隐私安全，防止模型被滥用。内容安全评测在AI领域尤为重要，因为大语言模型强大的生成能力可能带来包括色情、暴力、歧视或虚假信息在内的风险，对社会和个人造成不良影响。百度大模型评测平台通过标准化的评测流程，确保模型的安全性和可靠性。

平台的评测服务基于科学严谨的流程设计，结合自动化工具和专业团队，对模型内容进行深度审查。这一流程涉及数据预处理、规则匹配、深度分析、人工审核以及结果反馈等环节，确保评测的全面性和准确性。此外，评测不仅限于内容安全，还涵盖性能、功能及应用场景适应性等，构成一个全面的模型评测体系，使开发者能够全方位了解模型表现。

3.2 各平台的优劣势分析

业界的大模型评测平台在多个方面展现出显著的优势，这些优势共同推动大语言模型评测技术的发展，并为用户提供高效、可靠的服务。

首先，这些平台在评测的全面性方面做得非常出色。多数平台在评估模型时，不仅重视其语言生成能力，还综合考虑了模型在复杂情境中的推理、决策、学习以及适应等多方面的能力。例如，百度大模型评测平台依据国家相关法规和标准，包括《网络安全法》和《互联网信息服务管理办法》，进行安全评测，确保大模型的输出内容合法合规，避免生成违法违规信息，保护用户隐私，防止模型被用于不良目的。这种全面的评估方式为研究人员和开发者提供一个更为全面的模型能力视图，有助于他们深入理解模型的性能。

其次，多样化的评测方式是这些平台的另一大特色。平台如AgentBench支持零样本评测、小样本评测和思维链评测等，能够更全面地评估模型性能。零样本评测衡量模型在未经特定任务训练时，面对新任务的理解与执行能力；小样本评测则检验模型在仅提供少量样本后，对新任务的快速学习与适应水平；而思维链评测通过模仿人类思维过程，对模型处理复杂问题的推理及解决能力进行评估。这种多样化的评测方式使得平台能够适应不同类型的模型和应用场景，为用户提供更加精准的评测服务。

另外，分布式高效评测是这些平台的显著特征。一些平台如OpenCompass提供分布式评测方案，支持在单机或集群上并行处理计算任务，显著提高评测效率。通过分割大任务、合并小任务等策略，实现计算负载的均衡，充分利用所有可用的计算资源。在资源充足的情况下，OpenCompass能够在短时间内完成千亿参数量级模型的全面评测。这种分布式的评

测方案不仅大幅提升评测效率，也使得大规模的评测任务变得可行。

开源可复现性是这些平台的另一个重要优势。平台如OpenCompass开源评测框架，允许用户自由地访问和修改评测代码，以适应不同的评测需求。开源的特性也确保评测过程的透明性和可复现性，允许学术界和工业界的研究人员对评测结果进行验证和比较。这种开源的精神极大地促进评测技术的发展，并且增加平台的可信度。

平台在发展过程中也面临着一系列挑战和局限性，这些问题的存在可能会影响到评测的准确性和泛化性。

首先，缺乏统一的评测标准是业界平台面临的一个重要问题。不同的评测平台可能会采用各自的评测标准和方法，这导致不同平台之间的评测结果难以进行横向比较。由于缺少一个共通的参考框架，研究者和开发者可能难以准确判断模型在不同平台下的相对性能。这种标准的不一致性也可能阻碍跨平台、跨模型的公平比较和性能分析，限制学术交流和知识共享。

其次，评测平台自身的安全性和公正性问题也不容忽视。评测平台需要保证其评测过程的安全性，避免恶意攻击和不正当竞争行为。因此，评测平台需要建立严格的安全机制和审计流程，确保评测的公正性和透明性，从而为用户提供可信的评测服务。

数据集局限性是另一个挑战。一些评测平台可能受限于特定的数据集，这使得它们难以全面评估模型在多样化场景下的表现。不同的应用场景和语言环境需要不同的数据集支持，如果平台的数据集覆盖范围有限，就可能无法充分测试模型在特定场景下的性能。因此，评测平台需要不断扩大和丰富其数据集，以适应不同模型和应用的评测需求。

技术门槛是评测平台面临的另一个问题。对于非专业人士来说，使用某些评测平台可能存在一定的技术门槛。复杂的操作界面、繁琐的评测流程和专业的术语可能会阻碍非专业用户的有效使用。为了提高平台的普及率和可用性，评测平台需要提供更加友好的用户界面和简化的操作指南，降低用户的使用难度。

最后，分布式评测虽然能够提高评测效率，但其对计算资源的高需求也可能限制部分用户的使用。分布式评测需要大量的计算资源支持，这对于一些资源受限的用户来说可能是一个不小的负担。此外，分布式评测的设置和维护也需要专业的技术支持，这可能会增加用户的操作成本。

第四章 中国移动“弈衡”大模型评测平台

为了全面解决人工智能大模型评测在技术验证、质量控制、风险管理和合规性等多个层面上的需求，同时克服当前评测过程中存在的多种局限性问题，我们提出构建一个高效智能的大模型评测平台。该平台能够确保评测流程的标准化、安全性、公正性，并提升数据集的多样性和代表性。同时，通过降低技术门槛和优化分布式评测系统的维护，我们致力于提供一个更加高效、经济、且易于接入的评测平台。

4.1 设计原则和思路

在“弈衡”大模型评测平台构建过程中，我们采取一系列设计原则和实现思路，确保平台的高效性、用户友好性和未来的可持续发展。设计原则主要包括以下三点。

一是强化“智能化自动化”的原则。通过自动化流程，我们显著降低了对人工干预的需求。同时，采用人工智能和机器学习技术不仅提升了评测的效率和精确度，也使得平台能够根据历史数据不断学习和优化评测策略，以实现评测过程的持续改进和自我完善。

二是保障“灵活可扩展性”的原则。“模块化设计”让我们能将复杂的评测流程分解为独立的功能模块，不仅提升了平台的灵活性，也便于未来的扩展和升级。每个模块都能独立运作，同时与其他模块有效协同，保证了平台的长期适用性和前瞻性。

三是坚持“交互体验设计”的原则。这意味着每项设计和功能决策都以用户体验为核心。为了确保用户能够迅速地掌握平台的使用并高效执行评测任务，我们提供了直观易操作的用户界面和流程。此外，通过详尽的帮助文档和在线客服支持，我们进一步降低了用户的学习门槛。

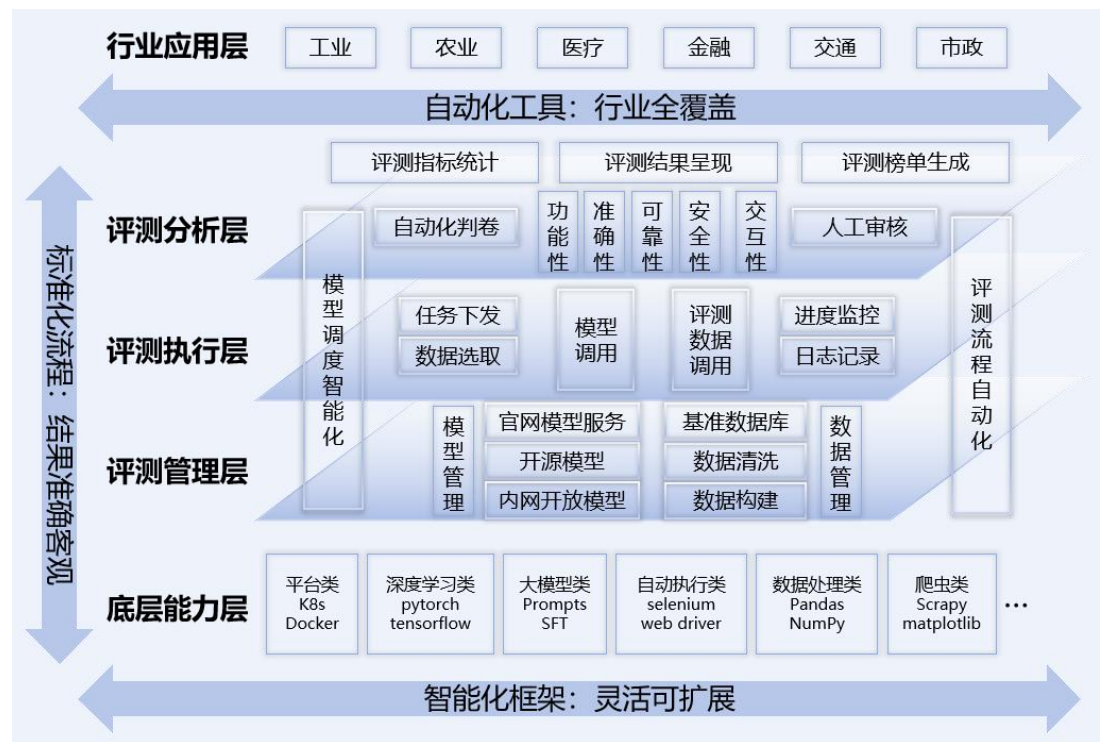
基于以上原则，提出以下设计思路。一是“全流程自动化”，利用了人工智能与机器学习技术，实现了从数据整理到评测报告生成的自动化，同时通过智能分析提升了评测的准确性。二是“模块化架构”，增强了平台的灵活性和可扩展性，使平台功能可分解为独立却能协同工作的模块。三是专注于“交互体验设计”，通过深入研究用户习惯与需求，打造出简洁而直观的用户界面和交互流程。

通过这些原则和思路的指导，我们可以有效地构建一个高效智能的大模型评测平台，旨在满足在技术验证、质量控制、风险管理、合规性等多个层面的评测需求。同时该平台也致力于克服当前评测过程中面临的诸多局限性，提供标准化、公正、安全且易于操作的评测服

务，从而推动大模型技术的持续创新和应用拓展。

4.2 平台整体架构

“弈衡”大模型评测平台整体架构采用分层与模块化的设计理念，主要由底层能力层、评测管理层、评测执行层、结果分析层和行业应用层五大部分构成，确保了评测过程的高效性、准确性和安全性。平台整体架构如图所示：



底层能力层是整个评测平台的基石，确保了平台的架构稳定和高效运行。平台本身具备强大的计算能力和稳定的架构，能够支撑起大规模数据处理和复杂模型的运算需求。该层融合了容器化编排技术，深度学习和大模型技术，以及自动执行、数据处理和爬虫技术，这些能力共同构建了一个高效、智能的评测环境，为大模型的性能评估提供了坚实的后盾。

评测管理层由数据管理层和模型管理层共同组成。数据管理层专注于数据的存储、清洗和构建，通过高可靠性数据库系统支持大规模数据集的存储与快速检索，自动去除噪声和异常值，并提供数据版本控制与访问权限管理，确保数据安全与完整性。模型管理层则支持模型接入、版本控制和安全性维护，支持多种大模型接入，记录模型变更历史，便于性能比较和模型重用，提升评测效率。

评测执行层作为自动化评测流程的关键，负责任务分发、进度监控和结果收集。它智能调度评测任务，提供实时进度监控，确保任务顺利执行，并具备异常检测与处理能力，保障任务连续性和稳定性。此外，执行层详细记录评测过程日志，便于问题诊断和策略优化。

评测分析层则深入分析数据，提供全面评估。它进行数据预处理，利用AI技术自动判定和分析评测结果，评估模型性能指标，并提供数据可视化工具帮助用户理解模型性能。同时，该层严格遵守数据安全和隐私保护标准，确保评测活动的安全性和合规性。通过这些层次的协同工作，平台为大模型的评测提供了一个坚实、高效和智能的环境。

行业应用层作为大模型评测的应用领域，覆盖了工业、农业、医疗、交通、市政等多个关键行业领域。该层利用评测平台提供的全面评估能力，针对各行业特点定制评测方案，充分挖掘模型优劣势，确保大模型在实际应用中发挥出最大价值。同时，平台不断收集各行业数据，持续优化评测算法，以满足日益复杂的行业需求。

4.3 平台特色及创新点

● 评测平台主要特色

1. 对标国际一流的大模型评测体系

本平台构建了多维度的“2-4-6”评测体系，该体系将评估场景划分为基础任务和应用任务，明确评测指标、评测数据等四项主要评测要素，并制定涵盖准确性、可靠性、安全性等六大维度 50 余个评测指标。在数据集方面，基于中国移动人工智能领域评测数据积累，充分融合斯坦福、微软等国内外权威评测数据基准，构建全面客观的通用大模型专属评测数据集，涵盖科学常识、逻辑推理、阅读理解、程序编写等两百多类典型场景，确保充分评估大模型综合能力。对标业界主流大模型评测体系，在分档一致性和排名一致性评估均处于业界一流水平。

2. 基于大模型技术的智能化结果评判

基于大模型技术的智能化结果评判是“弈衡”平台的核心优势之一。平台的智能化评判不仅减少了人工干预，还提高了评测的一致性和可复现性。平台利用“大模型评测大模型”的理念，通过对主流开源大模型的评判能力进行针对性调优，集成自动化评测结果判别能力，提升评测结果判别准确性。“弈衡”平台还采用了多模型融合技术，结合了多个模型的评估结果来提供综合的评判。这种方法利用了不同模型之间的互补性，提高了评测结果的鲁棒性和可信度。

3. 基于自组织的端到端自动化流程

“弈衡”大模型评测平台通过自组织技术实现了评测流程的全面自动化，从而显著提升了评测工作的效率。在整个评测流程中，平台能够自动部署评测模型，根据评测任务匹配针对性的评测数据集，并独立执行整个评测任务，无需人工干预。任务完成后，系统不仅可以

自动统计和汇总结果，还能进行深入的性能分析，利用智能化评判算法评估模型在各项评测指标上的表现，并生成详细的分析报告。此外，平台还能自动检测和处理评测过程中的异常情况，确保流程的连续性和稳定性。实时进度跟踪功能和详尽的日志记录使得用户能够随时了解评测状态，同时便于问题诊断和策略优化。整体而言，“弈衡”平台的自动化全流程设计极大地简化了用户的操作复杂度，让用户可以更专注于模型优化和决策分析，而不必关注评测的具体执行细节。

● 平台创新点

1. 用大模型测试大模型

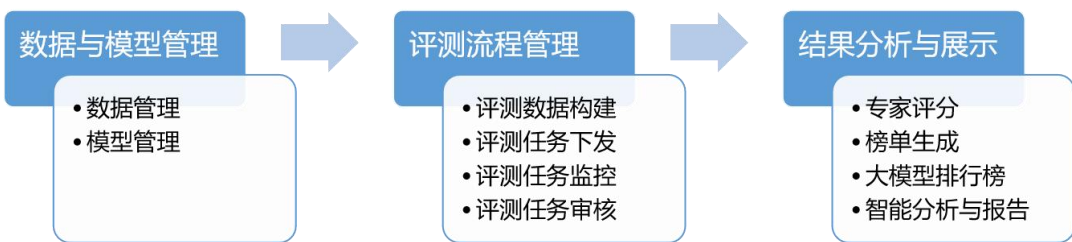
“弈衡”平台的一个显著创新之处在于其采用了“大模型测试大模型”的评测方法，即利用经过针对性训练的大模型来测试和评估其他大模型的性能。通过针对性调优主流开源大模型的评判能力，提升平台评测结果的准确性，同时集成了自动化评测结果判别能力，减少了人为因素的干扰，大幅提升评测结果判定智能化比例。相较于传统的以小模型或固定规则为基础的测试方法，随着经验的积累，评测大模型能够不断优化评测策略，适应多样化的大模型和不断演进的评测需求，为大模型的迭代优化提供了更加深入的洞察。

2. 自动化智能评测引擎

“弈衡”平台的另一个创新点是其实现了一套全自动化智能评测引擎。从用户上传模型和数据开始，到最终评测报告的生成，整个过程无需人工干预，平台可以独立完成。这一流程包括但不限于任务的自动下发、实时进度监控、结果的自动收集与分析，以及评测报告的智能生成。全自动化的评测流程不仅极大地提升了评测的效率，减少了人为错误的可能性，而且也使得评测结果更加客观和一致。

4.4 平台主要功能

“弈衡”大模型评测平台的功能逻辑图如下所示。“弈衡”平台的功能覆盖了从数据和模型管理到评测流程的自动化，再到结果的分析与展示，形成了一个完整的评测系统。每个类别内部的功能相互支持，共同构成了一个高效、自动化的AI大模型评测解决方案。



● 平台首页

下图为“弈衡”大模型评测平台访问入口。

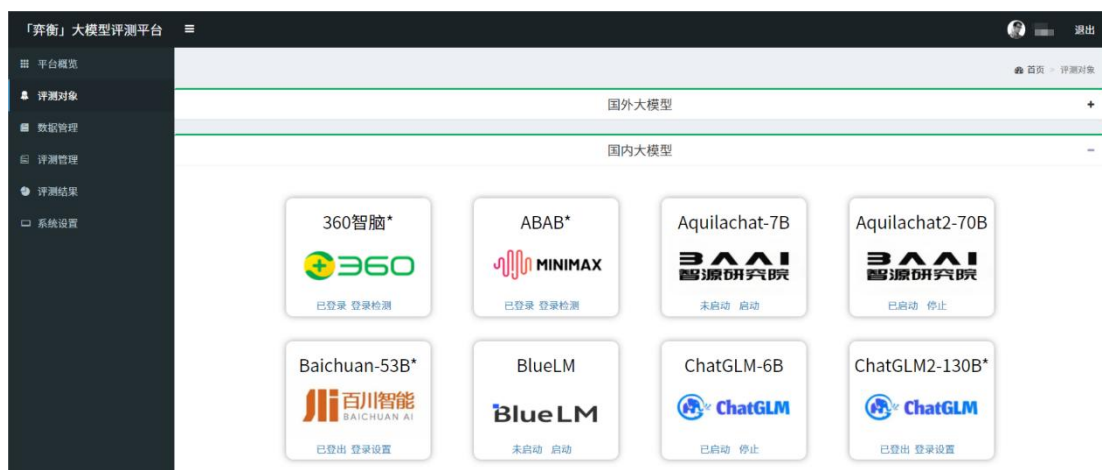


● 数据管理

数据管理提供标准化的数据存储、访问和预处理功能，包括数据清洗、去重、去噪和异常值处理等功能，以帮助用户更好地构建和选择数据集，确保数据集的质量，从而更好地评测大模型。

● 模型管理

模型管理提供了全面的模型接入支持，能够实现自动化模型配置。它不仅涵盖了对官网发布模型的接入，同时也兼容部署的开源模型，确保了平台的开放性和灵活性。通过一站式的启停管理，用户可以轻松地对模型进行启动、暂停或终止操作，从而优化资源利用率并提高工作效率。



● 评测数据构建

评测数据构建允许用户根据特定评测任务自主设计数据集和选择评测指标。这一功能通过提供数据预处理和多样化的指标模板，适应多变的评测需求，确保了评测的针对性和准确性，支持从基础到复杂的多场景应用，增强了评测的灵活性和实用性。

● 评测任务下发

评测任务下发是AI大模型评测平台的高效自动化特性之一。用户无需深入了解不同模型的接口细节，只需在平台上选定评测对象和相应的数据集，即可通过一键式操作快速下发评测任务。这种设计极大地简化了评测流程，减少了人工设置和干预，显著提升了评测工作的效率和准确性。同时，自动化的下发机制确保了评测任务的一致性和可复现性，为大规模和高频次的模型评测提供了强有力的支持。

● 评测任务监控

评测任务监控为大模型评测提供了直观的进度跟踪能力。用户可以通过用户界面实时监控评测任务的执行状态，包括当前的进度、已处理的数据量等。这种透明化的监控机制不仅增强了用户对评测流程的掌控感，而且有助于及时发现并解决评测过程中可能出现的问题，从而确保评测任务的顺利进行。

● 评测任务审核

评测任务审核功能允许专业人员对自动生成的评测结果进行人工核查，以确保评测结果的准确性。通过自动判卷和人工核查双重保障机制，提供更为真实和有效的评测结果。人工审核的介入，尤其是在处理复杂或关键性的评测任务时，为评测的精确性和权威性提供了额外的保障，增强了评测结果的可信度和实用性。

● 专家评分

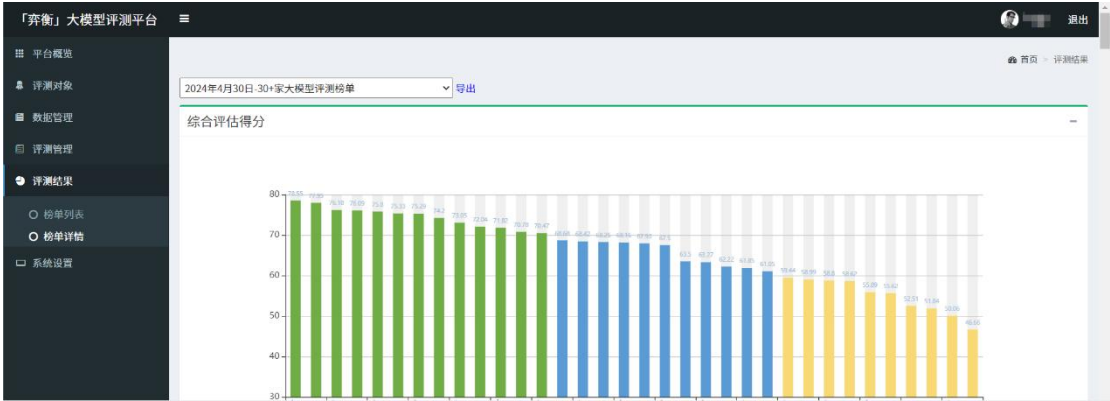
“弈衡”平台支持专家评分，对于模型主观评价问题，允许领域内的专家对模型的性能进行打分和评价。这一功能结合了人工智能的自动化评测和专家的深入洞察，提供了更为全面和权威的评测结果。

● 榜单生成

榜单生成功能允许用户通过简单的操作，一键生成模型性能排名榜单。平台根据模型的评测结果和专家评分，自动整理并展示每个模型在不同性能指标上的表现，从而为用户提供了一个快速便捷了解模型性能水平的途径。这种自动化的榜单生成方式很大程度上节省了用户整理数据的时间，极大地提升了用户体验。

● 大模型排行榜

大模型排行榜是“弈衡”平台的一项直观展示工具，它通过图形化界面清晰地展示各个模型的综合排名。用户可以迅速了解不同模型在关键性能指标上的相对位置，从而对模型的整体表现有一个直观的认识。这种图形化的展示方式不仅使复杂的数据信息变得易于消化，而且帮助用户快速识别顶尖模型和发现性能瓶颈，为模型选择和优化提供了有力的视觉支持。



● 智能分析与报告

智能分析与报告功能通过人工智能技术，对评测数据深度挖掘，精准捕捉并归纳模型能力，自动化编制评测报告。报告不仅呈现了模型的性能指标，还提供了模型间的比较排名，为用户提供了一个全方位的视角，以全面评估和比较不同模型的性能表现。

4.5 平台主要优势

"弈衡"大模型评测平台凭借其核心优势为用户提供了一个全面、高效、智能的评测解决方案，这些优势包括创新的"2-4-6"多维度评测体系，业界领先的自动化评测能力，用户友好的"一键测试"功能，高度的可拓展性等。

首先，平台的核心自研“2-4-6”多维度“弈衡”评测体系已达业界一流水平。采用“基于一致性的大模型评测体系量化评估方法”对标业界大模型评测体系，结果表明，在分档一

致性和排名一致性两类指标上，中国移动“弈衡”大模型评测体系，已经与国际主流评测体系如微软、斯坦福等实现了对标，其整体能力已达到国际一流水平。

其次，“弈衡”大模型评测平台具备“一键测试”的功能，显著提升了用户体验，该功能简化了人工操作流程，用户能够轻松地上传模型，并等待评测自动化执行完成即可。这一创新设计极大地降低了用户执行模型评测的难度，使得非技术背景的用户也能便捷地完成专业级别的评测。此外，平台可自动化生成详尽的评测报告，为用户提供模型功能和性能情况的全面视图，使得用户能够迅速掌握模型的强项与改进空间。

平台的另一大特色是其高度的可拓展性，评测模型、指标等内容均可灵活配置、调整。“弈衡”允许用户根据具体的评测需求，轻松添加或修改评测模型和指标，无需复杂的编程或系统调整。这种设计哲学赋予了平台强大的适应性，能够快速响应技术进步和市场变化，支持新兴模型和指标的集成。同时，开放的 API 接口进一步增强了平台的可拓展性，鼓励第三方开发者和服务提供商参与到平台的生态建设中，共同推动大模型评测技术的发展和應用。

与传统手动评测相比，“弈衡”平台显著提升了效率，将大模型评测周期从三周以上缩短至一周以内，并以其出色的实时响应能力，确保评判平均响应时间控制在 5 秒以内，满足了用户对实时性的要求。平台的自动化评测成功率稳定在 90% 以上，证明了其高可靠性。通过容器化技术和资源调度优化，平台提升了硬件资源利用率并降低了运营成本，同时已成功适配 50 余款业界主流大模型，展现了强大的兼容性。

4.6 成功案例与应用实践

“弈衡”大模型评测平台在不同层面上展现了其深远的影响力和专业评测能力，具体体现在如下几个方面：

在国家层面，中国移动已多次为政府部委提供大模型评测支撑，通过专业的评测，深入分析了人工智能的发展趋势和风险，获得高度认可。在政务领域，该平台与北京市大数据中心合作，参与政务大模型能力评测，科学支撑了北京市政务应用的决策过程，受到相关部门和客户的一致好评。为国家部委等递交近 10 篇专报，得到各级领导批示。

在产业层面，基于“弈衡”大模型评测平台，开展多领域行业大模型评测，根据不同行业的特点和需求，“弈衡”平台能够提供定制化的评测方案。同时能够对不同领域的大模型进行跨领域的性能比较，帮助企业识别各模型在特定任务上的优势和局限性，从而做出更合适的技术选型决策。根据评测结果，企业可以了解到模型在实际应用中的表现，并据此进行模型的优化和迭代，提升模型的实用性和有效性。此外通过成立行业学会大模型评测工作组、

举办全国性大模型评测竞赛等多种方式，推动建立完善的评测体系，研发高效的审核评估工具，助力国产大模型技术成熟，促进产业生态繁荣。

第五章 未来发展展望

展望未来，“弈衡”大模型评测平台将致力于深化、扩展评测能力，以适应不断演进的大模型技术。

在评测需求的扩展与深化方面，随着大模型技术的不断进步，平台将面临从传统文本处理向多模态领域扩展的挑战。这意味着评测需求将不限于语言模型的文本理解与生成能力，还将涵盖图像、视频、音频等多种模态的交互与融合能力。为了全面评估模型在这些领域的的能力，平台必须开发新的评测指标和方法，构建相应的多模态数据集，并设计能够模拟真实应用场景的复杂评测环境。

在评测技术的创新与应用方面，“弈衡”平台将持续升级评测平台架构，优化资源管理，并增强并发处理能力，以高效应对大规模的评测任务。同时，进一步改进资源分配策略，提升运算效率，并加速整个评测流程。此外，随着评测领域的扩展，通过各行业数据不断学习和优化评测策略，增强评测大模型的评判能力，通过模型调优提升判别的准确性和稳定性，实现对大模型全方位智能化评测分析。

在评测数据构建与共享方面，“弈衡”大模型评测平台将着重于构建高质量、多样化的评测数据集。这些数据集将涵盖多个模态，如：文本、图像、音频和视频，并注重数据的真实性和代表性。平台还将积极推动数据的共享与开放，鼓励研究者们共同贡献和使用这些数据集，以促进大模型评测领域的快速发展。通过数据的共享与协作，我们将共同推动评测技术的进步，为大模型的持续优化提供有力支持。

未来“弈衡”大模型评测平台将进一步构建并完善大模型评测生态，并推动各方深度合作。为此，“弈衡”平台将倾力打造一个开放且包容的评测生态系统。积极携手学术界、工业界、政策制定者及其他相关利益方，共同研讨并制定权威的评测标准，研发先进的评测工具，并汇聚多元化的数据集，进一步提升评测的公信力与实用性，促进大模型技术的稳健进步，从而为人工智能技术的广泛落地奠定扎实的基础。