ModelArts

模型评测

文档版本 01

发布日期 2026-05-20

华为云计算技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为云计算技术有限公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为云计算技术有限公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为云计算技术有限公司

地址： 贵州省贵安新区黔中大道交兴功路华为云数据中心 邮编：550029

网址： https://www.huaweicloud.com/

目 录

1 实时对比

使用场景

在AI工程化落地过程中，面对众多的基础大模型和微调版本，如何选择“最合适”的模型是关键难题。模型实时对比功能提供了一个直观的对比平台，允许用户在完全一致的输入条件下，对不同模型进行横向评测。

核心作用与价值：

- **基础大模型选型：**在项目初期，对比如DeepSeek3、Qwen3、GLM4.5等不同架构模型的表现，快速锁定适合业务场景的基模。
- **微调效果验证：**将“未微调的原生模型”与“微调后的模型”做同步对比，直观验证微调是否成功注入了领域知识，或是否存在能力退化。
- **参数策略调优（A/B Testing）：**对比同一模型在不同超参数（如Temperature、Top-P）下的输出差异，寻找最佳推理配置。

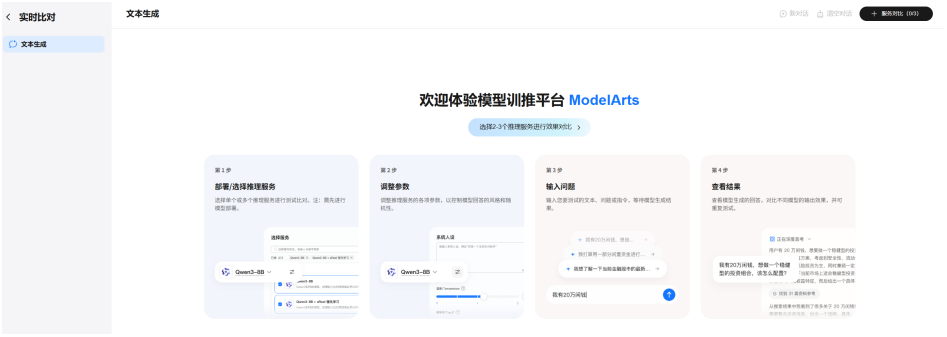
约束限制

- **区域使用限制：**仅“西南-贵阳一”区域的新版控制台支持。
- **模型类型限制：**当前仅支持大语言模型领域的文本生成类模型对比，暂不支持其他领域模型对比。
- **数量限制：**为了保证前端渲染性能及便于人眼比对，单次对比任务最多支持 3 个模型同时进行。
- **超时限制：**在实时比对任务运行过程中，如果某个比对模型因为思考或者性能原因超过5分钟未结束，该模型窗口会提示超时而中断回答。

操作步骤

1. **选择实时比对功能。**前往**ModelArts管理控制台**，选择“模型评测 > 实时对比”，进入“实时比对”操作页面。如**图1-1**所示。右上角功能说明如下：
   - **新对话：**清空当前对话，开始一个新的对话。
   - **清空对话：**清空当前对话内容上下文。后续对话不受上文对话影响。
   - **服务比对：**开始一个对话服务，可以选择1~3个模型对比。

图 1-1 实时比对



说明

实时对比有多个入口，除了从控制台左侧直接选择外，也可通过如下入口进入：

1. 在**ModelArts管理控制台**“模型推理 > 在线推理”的服务列表页，单击右侧操作列的“实时比对”，进入“实时比对”页面。

   或在“模型推理 > 在线推理”的服务列表页单击服务名称，进入服务详情页，单击右上角的“实时比对”，进入“实时比对”页面。

2. **选择对比模型。**单击右上角“服务对比”按钮，弹出“实时比对|选择服务”对话框。选择目标模型，如果已经部署的模型为目标对比模型，则选择对应的模型即可，如**图1-2**所示。如果目标模型不存在，需要先部署要对比的模型或选择对话框推荐模型“一键部署”，如**图1-3**所示。一键部署可参考**推理入门：一键完成Qwen3-32B模型部署**完成模型部署。选好模型后，进入模型对话工作区，即可开启对话，如**图1-4**所示。

图 1-2 选择目标模型

图 1-3 无目标模型

图 1-4 选好模型后的对话工作区

3. **调整模型各项参数**。设置不同参数，可使模型输出在随机性、最大生成长度等维度的输出不同。在实时比对时为保证模型对比控制变量单一，请保持模型配置参数一致。参数设置可参考**模型运行参数配置**。

图 1-5 模型参数配置

如果对比模型支持深度思考模式，可以打开深度思考开关。模型在给出答案前，会将深度思考结果打印在对话框。如图1-6所示。

图 1-6 模型深度思考过程

4. **输入问题。**输入框中撰写您的测试提示词（Prompt）。按“Enter”键发送，按“Shift+Enter”键换行。
   - 系统会将这一条Prompt同时发送给所有选中的模型。
   - 支持单轮问答测试，也支持在当前上下文中进行多轮对话测试。

   图 1-7 输入测试提示词对话框

5. 查看各模型的输出结果。系统将以分栏视图（Side-by-Side）的形式，并行展示各模型的生成内容，您可以直观地通过肉眼比对文本的逻辑性、格式规范度以及语义准确性。如图1-8所示。

   图 1-8 大模型输出结果比对

6. 切换模型服务。如果在模型对比过程中需要切换不同模型。可在模型对话框中选择新模型，即可切换模型。

指标说明

除了主观的文本内容比对外，您还可以通过结果面板上的数据标签查看技术指标，以辅助量化评估。具体指标可以参考表1-1。

表 1-1 模型指标

指标类型	指标名称	指标说明
性能指标	总耗时	完成整个回答所需的总时间。耗时越短，说明模型输出的性能越强。
	思考时间	针对思考模型
	首字延迟 (TTFT-Time To First Token)	从用户单击“发送”按钮开始，到屏幕上出现AI回复的“第一个字（Token）”所花费的时间。TTFT越低代表模型响应速度越快。
	每个Token耗时（TPOT-Time Per Output Token）	当第一个字出来后，后续输出字符出现时，平均生成每个字需要的时间。TPOT越低代表模型后续输出越快，越流畅。
消耗指标	消耗Token	显示本次问答的Input Tokens（输入量）和Output Tokens（输出量），用于预估调用成本。

模型运行参数配置

在调用大模型时，经常会遇到模型回答问题和预想结果有较大差异的问题。您可以通过调整“解码参数”来控制模型生成的随机性和创造力。简单来说，这些参数决定了模型是像严谨的科学家一样回答，还是像浪漫的诗人一样创作。表1-2说明参数配置示例。

表 1-2 模型核心参数

参数名	作用	示例	推荐调试顺序
Temperature	控制整体随机性。数值越大，发散性越强；数值越小，答案越确定。 ● 低温度 (0.1)：模型极其保守，总是选择概率最高的那个选项。适合标准答案明确的场景。 ● 高温度 (0.9)：模型变得兴奋，愿意尝试概率较低的选项。适合需要创意的场景，但容易一本正经地胡说八道（幻觉）	prompt：请用“天空”造句。 ● Temperature = 0.1 (严谨) – **结果：**天空是蓝色的，飘着几朵白云。 – **特点：**准确、平淡、每次运行结果几乎一样。 ● Temperature = 0.9 (发散) – **结果：**天空宛如一块被打翻的蓝莓果酱，星辰在其中沉浮。 – **特点：**生动、多变、每次运行结果差异大。	优先调整
Top_P	动态截取概率最高的词。数值越大，可选词汇越丰富（但也可能越生僻）。Top_P不看**数量**，看累计概率。模型会按概率高低排序，把概率加起来达到P值（如0.9）的词留下来，剩下的丢掉。	● Top_P = 0.1：只取最头部、最稳的几个词。 ● Top_P = 0.9：允许更多长尾词汇进入候选池，词汇更丰富。 Top_P是动态的。如果下一组词都很确定，候选池就小；如果下一组词都很模糊，候选池就大。这比Top-K更智能。	配合Temperature微调
Top-K	强制保留排名前K个词。数值越大，保留的候选词越多。	● Top-K = 1：贪婪解码，每次只选第1名（效果等同于极低温度）。 ● Top-K = 50：主要用于防止模型生成极低概率的乱码。	辅助参数（通常保持默认或较大值）

以下是配置参数的典型场景，请根据使用场景配置不同参数。

表 1-3 模型参数配置典型场景

业务场景	建议配置	期望效果	典型应用
代码生成 数学解题	Temp: 0.0 - 0.2 Top_P: 0.1	**极度精确** 拒绝随机性，保证代码逻辑正确，语法严谨。	辅助编程、SQL生成、逻辑推理

业务场景	建议配置	期望效果	典型应用
知识问答 客服	Temp: 0.3 - 0.5 Top_P: 0.7	**稳定且自然** 事实准确，但语言组织比机器人更像人类。	智能客服、RAG文档问答
文案创作 闲聊	Temp: 0.7 - 0.9 Top_P: 0.9	**丰富多样** 词汇量大，句式多变，富有创意。	营销文案、小说续写、角色扮演
头脑风暴	Temp: 1.0+ Top_P: 0.95	**天马行空** 跳出常规逻辑，寻找意外的灵感（需人工筛选）。	创意构思、起名

2 模型评测

2.1 模型评测功能说明

模型评测功能介绍

模型评测是测试和衡量大模型在现实世界情境中表现如何的过程，是了解大模型性能的关键。

效果优秀的模型需要保证模型拥有良好的泛化能力，即模型不仅要在已给定的数据（训练数据）上表现良好，还要能够在未见过的数据上也达到类似的效果。为了实现这一目标，模型评测是必不可少的环节。

在ModelArts模型的开发流程中，模型评测是在完成模型训练后，对尚未投入使用的模型做多方位评测，只有经过评测后的模型，才能部署上线并使用。是模型开发工具链的关键环节。

为什么需要模型评测

模型评测能够帮助用户识别模型的优缺点，确保其在实际应用中的有效性，能够胜任特定任务并满足相关要求。

在收集评估数据集时，必须保持数据集的独立性和随机性，确保收集到的数据能够代表现实世界的样本数据。这有助于避免对评估结果产生偏见，从而更准确地反映模型在不同场景下的表现。通过使用评估数据集对模型进行评估，开发者可以了解模型的优缺点，从而找到优化方向。

模型评测对开发者的核心价值：

- **验证训练效果**：衡量微调/增量预训练后模型的能力提升程度。
- **发现优化方向**：定位模型在特定任务上的薄弱环节，指导后续迭代。
- **支撑部署决策**：以量化指标判断模型是否达到上线标准。
- **对比模型选型**：在多个候选模型中选择最适合业务场景的版本。
- **满足合规要求**：提供模型能力的量化证据，支持审计与合规。

模型评测场景

模型评测主要考验模型的知识记忆能力和文本理解能力。具体可分为通用能力和行业能力。以下将分别介绍**通用能力评测**和行**业能力评测**的使用场景。

通用能力评测

通用能力：主要包含通用领域的数据集评测任务，如文本分类、逻辑推理、情感分析、问答系统等任务。

典型场景：

- 文本分类准确率评测。
- 逻辑推理能力评测。
- 情感分析正确率评测。
- 阅读理解与问答系统评测。
- 文本摘要质量评测。
- 机器翻译流畅度评测。

推荐数据集来源： ModelArts提供了开源评测集的管理功能，便于用户能够方便使用开源数据集，对相关大模型做更加精准高效的评测。

行业能力评测

行业能力：主要包含特定领域的数据集评测任务，如金融实体识别、金融文本分类、催收意图识别等任务。

典型场景：

- 金融行业：实体识别、合同条款分类、风控意图识别。
- 医疗行业：医学问答、病历摘要、药物信息抽取。

推荐数据集来源： 创建特定评测集：如需评测模型的领域知识能力，可以使用同源数据集构建实体识别、文本分类或内容生成等评测集，精确率、召回率和F-score作为评测指标。

模型评测类型

ModelArts提供了功能强大的模型评测功能。支持人工评测、自动评测两种评测模式。

自动评测

自动评测：包含"基于规则"、"基于大模型"两种规则。

基于规则（相似度/准确率）自动对模型生成的回答进行评测。用户可使用评测模板中预置的专业数据集进行评测，或者自定义评测数据集进行评测。

适用范围： 有明确标准答案的封闭式任务，如分类、实体识别、选择题问答等。

运行方式： 系统自动将模型输出与评测数据集中的参考答案进行比对，基于相似度算法或准确率规则计算评测得分。

基于大模型，使用大模型对被评估模型的生成结果进行自动化打分，适用于开放性或复杂问答场景，包含评分模式与对比模式。

适用范围： 没有唯一标准答案的开放式任务，如创意写作、开放式问答、对话生成等。

两种子模式：如**表2-1**所示。

表 2-1 基于大模型评测的子模式

子模式	描述	典型用途
评分模式	使用裁判大模型对被测模型的生成结果进行多维度评分。	评估单个模型的生成质量。
对比模式	使用裁判大模型同时对比两个模型的输出，给出优劣判断。	模型A/B选型对比。

人工评测

人工评测：通过人工创建的评测数据集和评测指标项对模型生成的回答进行评测，评测时需要人工基于创建好的评测项对模型回答进行打分，评测完成后会基于打分结果生成评测报告。

适用范围： 需要人类主观判断的场景，如回答的风格、语气、专业性、安全性等难以用自动化规则衡量的维度。

运行方式： 在人工评测页面对每条数据进行评估并打分，直到所有数据评估完成后，单击“提交”，提交评估结果。

2.2 创建模型评测任务

前提条件

- 已注册华为账号并开通华为云，进行了实名认证，且在使用ModelArts前检查账号状态，账号不能处于欠费或冻结状态。
  - **注册华为账号并开通华为云**
  - **进行实名认证**
- 配置委托访问授权

  ModelArts使用过程中涉及到OBS等服务交互，首次使用ModelArts需要用户配置委托授权，允许访问这些依赖服务。

计费说明

数据连接计费涉及到OBS计费，具体可参考**数据管理计费项**。

约束限制

- 仅“西南-贵阳一”区域的新版控制台支持。
- 仅支持大语言模型支持模型评测。

创建大模型自动评测任务（基于规则）

创建大模型自动评测任务步骤如下：

1. 前往**ModelArts管理控制台**。
2. 在控制台左侧导航栏选择“模型评测 > 评测任务”，在评测任务工作区左上角选择“自动评测”页签后，选择“创建”，如**图2-1**所示。

图 2-1 创建自动评测

3. 在“创建自动评测任务”页面，参考**表2-2**完成部署参数设置。

表 2-2 大模型自动评测任务参数说明（基于规则）

参数分类	参数名称	参数说明
基本信息	任务名称	评测任务名称。任务名称字段要求输入以中文、字母**开头**，以中文、字母、数字**结尾**，长度2~32的字符。只允许输入中文、字母、数字、中划线、下划线字符。
	描述	填写评测任务描述，该字段可选。
评测对象	评测类型	当前仅支持“文本生成”类型。
	添加服务	选择部署至ModelArts平台的模型进行评测。单次最多可评测10个模型。
评测配置	评测规则	选择“基于规则”：基于规则自动打分，即基于相似度/准确率进行打分，对比模型预测结果与标注数据的差异，适合标准选择题或简单问答场景。
	评测数据集	• 预置评测集：使用预置的专业数据集进行评测。 • 自定义评测集：由用户指定评测指标（F1分数、准确率、BLEU、Rouge）并上传评测数据集进行评测。选择“自定义评测集”时需要上传待评测数据集。 （上传单个.jsonl文件，文件大小不超过10M，最大1000条）
	评测结果存储位置	模型评测结果的存储位置。

4. 参数填写完成后，单击“立即创建”，返回至“评测任务 > 自动评测”页面。
5. 当状态为“已完成”时，可以单击操作列“评测报告”，在“评测报告”页面，可以查看评测任务的评测报告和详情。

创建大模型自动评测任务（基于大模型）

创建大模型自动评测任务步骤如下：

1. 前往**ModelArts管理控制台**。

2. 在控制台左侧导航栏选择“模型评测 > 评测任务”，在评测任务工作区左上角选择“自动评测”页签后，选择“创建”，如图2-2所示。

图 2-2 创建自动评测

3. 在“创建自动评测任务”页面，参考表2-3完成部署参数设置。

表 2-3 大模型自动评测任务参数说明（基于大模型）

参数分类	参数名称	参数说明
基本信息	任务名称	评测任务名称。任务名称字段要求输入以中文、字母**开头**，以中文、字母、数字**结尾**，长度2~32的字符。只允许输入中文、字母、数字、中划线、下划线字符。
	描述	填写评测任务描述，该字段可选。
评测对象	评测类型	当前仅支持“文本生成”类型。
	添加服务	选择部署至ModelArts平台的模型进行评测。单次最多可评测10个模型。
评测配置	评测规则	选择“基于大模型”。
	选择模式	• 评分模式：裁判模型将根据设置的评分标准对模型推理结果自动进行打分。 • 对比模式：模型将对比每个模型服务和基准模型服务的表现，选择win、lose、tie展示对比结果， 对比模式下服务必须选择2个及以上。
	评测数据集	• 预置评测集：使用预置的专业数据集进行评测。最多只能添加一个预置评测集。 • 自定义评测集：由用户指定评测指标（F1分数、准确率、BLEU、Rouge）并上传评测数据集进行评测。选择“自定义评测集”时需要上传待评测数据集。（上传单个.jsonl文件，文件大小不超过10M，最大1000条）
	评测结果存储位置	模型评测结果的存储位置，选择OBS存储评测结果。

参数分类	参数名称	参数说明
裁判员配置	裁判模型	已部署服务：选择已部署至ModelArts平台的模型进行评测。
	打分规则	打分Prompt模板可以选择预置，也可以选择自定义。 预置Prompt不支持修改。 创建自定义Prompt模板需要在“编辑自定义规则”右侧对话框选择“新建”，根据页面输入名称，人设，任务描述，是否包含问题，是否包含参考答案，打分策略，评测指标，最终单击“保存模板”。

4. 参数填写完成后，单击“立即创建”，返回至“评测任务 > 自动评测”页面。单租户可以创建的最大评测任务数为2000。
5. 当状态为“已完成”时，可以单击操作列“评测报告”，在“评测报告”页面，可以查看评测任务的评测报告和详情。

创建大模型人工评测任务

创建大模型人工评测任务步骤如下：

1. 前往**ModelArts管理控制台**。
2. 在控制台左侧导航栏选择“模型评测 > 评测任务”，在评测任务工作区左上角选择“人工评测”页签后，选择“创建”，如**图2-4**所示。

图 2-3 创建人工评测

3. 在“创建人工评测任务”页面，参考**表2-4**完成部署参数设置。

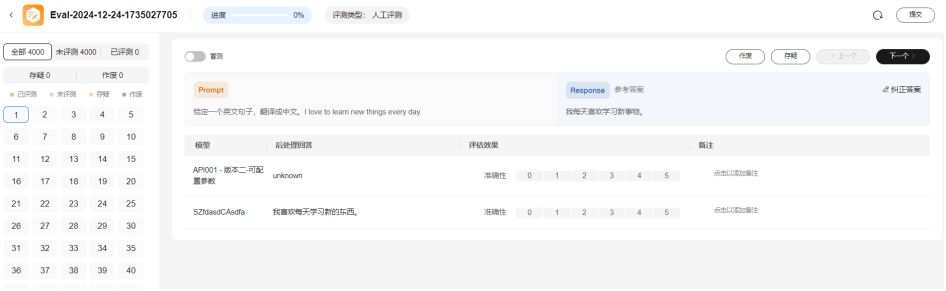
表 2-4 大模型人工评测任务参数说明

参数分类	参数名称	参数说明
基本信息	任务名称	评测任务名称。任务名称字段要求输入以中文、字母**开头**，以中文、字母、数字**结尾**，长度2~32的字符。只允许输入中文、字母、数字、中划线、下划线字符。
	描述	填写评测任务描述，该字段可选。
评测对象	评测类型	当前仅支持“文本生成”类型。

参数分类	参数名称	参数说明
	添加服务	选择部署至ModelArts平台的模型进行评测。单次最多可评测10个模型。
评测配置	评测指标	由用户自定义评测指标并填写评测标准，最多支持同时添加6个指标。
	评测数据集	待评测的数据集。
	是否开启盲测	开启盲测后，人工打分时将看不到模型名称，且多个模型的排列顺序是打乱的。
	评测人员	只有配置的人员才能对该评测任务评分，且必须所有配置的人员都对所有case评分后才能生成评测报告。
	评测结果存储位置	模型评测结果的存储位置。

4. 参数填写完成后，单击“立即创建”，返回至“评测任务 > 人工评测”页面。
5. 当状态为“待评测”时，可以单击操作列“在线评测”进入评测页面。
6. 依据页面提示完成评测，全部数据评测完成后单击“提交”。
   - 单击“存疑”或者“作废”进行用例的存疑或作废，如果取消存疑或作废，单击“取消存疑”或“取消作废”进行处理。
   - 给用例的所有评估指标打分，单击“保存并下一个”，可保存分数并切换到下一个用例
   - 单击“上一个”，可以回到上一个用户重新打分
   - 单击备注下方“单击以添加备注”，可以进行新增备注。
   - 评测页面，长按鼠标左键选中需要标记的文本内容，单击“标记”可以标记成重点内容。

   图 2-4 人工评测



7. 返回“评测平台 > 评测任务 > 人工评测”页面，单击操作列“评测报告”查看模型评测结果。

   评测完成之后，进入人工评测列表页面，单击“人工复核”，进行复核评测，复核完成之后，单击“提交”，提交评估结果。

2.3 查看模型评测报告

评测任务创建成功后，可以查看大模型评测任务报告，具体步骤如下：

1. 前往**ModelArts管理控制台**。
2. 在控制台左侧导航栏中选择“模型评测 > 评测任务”。
3. 单击操作列“评测报告”，在“评测报告”页面，可以查看评测任务的基本信息及评测概览。

   其中，各评测指标说明详见**大模型评测指标说明**。
4. 导出评测报告。
   a. 在“评测报告 > 服务结果分析”页面，单击“导出”，可选择需要导出的评测报告，单击“确定”。
   b. 单击右侧“导出记录”，可查看导出的任务ID，单击操作列“下载”，可将评测报告下载到本地。

大模型评测指标说明

大模型支持自动评测与人工评测，各指标说明如**表2-5**、**表2-6**、**表2-7**、**表2-8**。

表 2-5 大模型自动评测指标说明-不使用预置评测集

评测指标（自动评测-自定义评测集）	指标说明
准确率	正确预测(标注与预测完全匹配)的样本数与总样本数的比例。分数越高表示模型正确预测的样本比例越高，模型的效果越好。
F1分数	精确率和召回率的调和平均数，分数越高表示模型在这两个指标上表现越好，即模型在精确率和召回率之间取得了更好的平衡。
BLEU-1	模型生成句子与实际句子在单字层面的匹配度，分数越高，表示模型效果越好。
BLEU-2	模型生成句子与实际句子在中词组层面的匹配度，分数越高，表示模型效果越好。
BLEU-4	模型生成结果和实际句子的加权平均精确率，分数越高，表示模型效果越好。
ROUGE-1	将模型生成结果和标注结果按1-gram拆分后，计算出的召回率（n-gram指一个语句内连续的n个单词组成的片段），分数越高，表示模型效果越好。
ROUGE-2	将模型生成结果和标注结果按2-gram拆分后，计算出的召回率（n-gram指一个语句内连续的n个单词组成的片段），分数越高，表示模型效果越好。

评测指标（自动评测-自定义评测集）	指标说明
ROUGE-L	将模型生成结果和标注结果按最长公共子序列（longest-gram）拆分后，计算出的召回率，分数越高，表示模型效果越好。

表 2-6 大模型自动评测指标说明-使用预置评测集

评测指标（自动评测-使用评测模板）	指标说明
评测得分	每个数据集上的得分为模型在当前数据集上的通过率；评测能力项中如果有多个数据集则按照数据量的大小计算通过率的加权平均数。
综合能力	综合能力是计算所有数据集通过率的加权平均数。

表 2-7 大模型人工评测指标说明

评测指标（人工评测）	指标说明
准确性	模型生成答案正确且无事实性错误。
average	模型生成句子与实际句子基于评估指标得到的评分后，统计平均得分。
goodcase	模型生成句子与实际句子基于评估指标得到的评分后，统计得分为5分的占比。
badcase	模型生成句子与实际句子基于评估指标得到的评分后，统计得分1分及以下的占比。
用户自定义的指标	由用户定义的指标，如有用性、逻辑性、安全性等。

表 2-8 大模型自动评测指标说明

模型类型	评测指标（自动评测-基于规则-基于大模型）	指标说明
大模型	裁判员模型打分	数据集中每个用例，裁判员模型给的评分值。
	平均值	数据集中所有用例得分的平均值。
	中位数	数据集中所有用例得分的中位数。
	标准差	数据集中所有用例得分的标准差。

模型类型	评测指标（自动评测-基于规则-基于大模型）	指标说明
	win	统计所有对比模型中，性能指标（需提前明确 “优” 的定义，如准确率高为优、误差低为优）优于基准模型的模型数量。
	lose	统计所有对比模型中，性能指标劣于基准模型的模型数量。
	tie	统计所有对比模型中，性能指标与基准模型完全持平（无优劣差异）的模型数量。
	分位	(win+tie)/(lose+tie)
	分位（剔除tie_bad）	剔除tie_bad所得分位，(win+tie_good)/(lose+tie_good)
	分位（剔除tie_good）	剔除tie_good所得分位，(win+tie_bad)/(lose+tie_bad)

2.4 管理模型评测

2.4.1 预置评测集与评测模板

预置评测集

功能介绍

预置评测集是一组经过精心设计、标注和标准化的数据样本，专门用于测试、评估和量化人工智能模型在特定任务上的表现。

本次支持的预置评测集名称列表，和评测集描述见**表1 预置评测集列表**。

表 2-9 预置评测集列表

名称	评测集描述
MMLU-Pro	MMLU是人工智能领域最有影响力的大模型测评基准之一，涵盖了基础数学、计算机科学、法律、历史等57项子任务，用于评测大模型的世界知识和问题解决能力。
GPQA_Diamond	GPQA_Diamond是一个由生物学、物理学和化学领域专家编写并验证的多选问答数据集，包含448个极其困难的问题。该数据集设计用于评估人工智能系统在跨学科问题上的表现，尤其针对非专家领域的问题（如物理学家回答化学问题）进行测试。

名称	评测集描述
BoolQ	BoolQ是一个专为“是/否”类型问题设计的问答数据集，数据集中的问题均源自真实的查询场景，未经任何特定引导，因此更加贴近真实世界的复杂性和多样性。
AGIEval	AGIEval基准包含了多种高质量的官方入学考试、资格考试、高级竞赛，如法学院入学考试（LSAT）、大学入学考试（如中国高考和美国SAT）、数学竞赛以及律师资格考试等，在数据集的构建上，AGIEval剔除了主观题，只保留客观题（如选择题和填空题）。这些考试和竞赛不仅具有官方认可的标准，而且能够全面考察模型认知能力、知识掌握程度以及推理能力。
C-Eval	C-Eval是一个全面的中文基础数据集，涵盖了52个不同的学科和四个难度级别，用于评测大模型中文理解能力。
GSM8K	GSM8K是由OpenAI发布的大模型数据推理能力评测基准。一个由8.5K小学数学问题组成的数据集，可以评测大模型的数学推理运算能力。
MathBench	对大语言模型的数学能力进行全面评估，涵盖理论概念理解和应用问题解决两方面
ARC Challenge	ARC Challenge是一个逻辑推理和问题解决的数据集，包含了来自不同领域的问题，用于评测模型的高级推理能力。
BBH	BBH是一个包含204项任务的大型语言模型评测数据集，涵盖了语言学、儿童发展、常识推理、社会偏见、软件开发等多个领域，用于评测模型在处理困难任务时的表现。
CMMLU	CMMLU是MMLU的中文版本，涵盖了人文学科、法律、工程、数学等多个通用领域的知识，用于评测模型在中文领域的多学科知识。
OpenFinData	OpenFinData是由东方财富与上海人工智能实验室联合发布的开源金融评测数据集。该数据集代表了最真实的产业场景需求，是目前场景最全、专业性最深的金融评测数据集。它基于东方财富实际金融业务的多样化丰富场景，旨在为金融科技领域的研究者和开发者提供一个高质量的数据资源。
FinEval	FinEval金融行业评测基准依据定量的基本方法，通过长期客观调研总结和严格的人工筛选，利用多项选择题、主客观简答题、推理规划和检索问答等超过26000道多种与实际应用场景高度一致的题型，包括了金融学术知识、金融行业知识、金融安全知识、金融智能体、金融多模态和金融严谨性，旨在全方位检验大模型在金融行业的综合应用能力。
MedMCQA	一个大规模的多项选择题问答（MCQA）数据集，旨在解决现实世界中的医学入学考试问题。

名称	评测集描述
PubMedQA	PubMedQA是一个从PubMed摘要中收集的新型生物医学问答（QA）数据集。PubMedQA的任务是利用相应的摘要来回答“是/否/可能”形式的研究问题（例如：术前使用他汀类药物是否能减少冠状动脉旁路移植术后房颤的发生？）。每个PubMedQA实例由以下四部分组成：（1）一个问题，该问题要么来自现有的研究文章标题，要么基于标题衍生而来；（2）一个上下文，即相应的摘要，但不包含结论部分；（3）一个长答案，即摘要的结论部分，通常也回答了研究问题；（4）一个“是/否/可能”的答案，用于总结结论。PubMedQA是首个需要基于生物医学研究文本（尤其是其中的定量内容）进行推理才能回答问题的问答数据集。

2.4.2 管理评测任务

在评测任务列表中，任务创建者可以对任务进行克隆（复制评测任务）、启动（重启评测任务）和删除操作。

1. 前往**ModelArts管理控制台**。
2. 在左侧导航栏中选择“模型评测 > 评测任务”，可进行如下操作：
   - 克隆。单击操作列的“克隆”，可以复制当前状态为“已完成”评测任务。
   - 启动。单击操作列的“启动”，可以重启状态为“已停止”的评测任务。
   - 删除。单击操作列的“更多 > 删除”，可以删除当前状态为“已完成”的不需要的评测任务。