doi: 10.3969/j.issn.1672-6073.2026.03.002

城轨领域人工智能大模型评测研究与实践

华福才[1]，邵　昕[1]，夏秀江[1]，韩晓艺[1]，王文俊[1]，于松伟[1]，蒋　玥[2]
（1. 北京城建设计发展集团股份有限公司，北京 100037；2. 宁波市轨道交通集团有限公司，宁波 315101）

摘　要：为解决城轨行业的大模型评测标准和方法问题，构建一套针对城轨行业大模型评测体系。该评测体系涵盖评测数据集的构建、评测维度的设计、评测方法的选择及评测指标的制定。通过这一体系，全面、客观地评估城轨大模型在语言理解、知识推理、图文对齐、时空一致性等任务中的表现。同时，设计研发城轨领域大模型专属评测系统，并通过实际应用验证评测体系的有效性。验证结果表明，所构建的评测体系能够为城轨领域大模型的优化与应用提供评测基础，促进大模型在城轨行业的落地与应用。

关键词：城市轨道交通；人工智能；城轨大模型；模型评测；评测体系
中图分类号：U231　　**文献标志码：**A　　**文章编号：**1672-6073(2026)03-0011-08

Research and Practice on the Evaluation of Artificial Intelligence Large Models in Urban Rail Transit

HUA Fucai[1], SHAO Xin[1], XIA Xiujiang[1], HAN Xiaoyi[1], WANG Wenjun[1], YU Songwei[1], JIANG Yue[2]
(1. Beijing Urban Construction Design & Development Group Co., Ltd., Beijing 100037;
2. Ningbo Rail Transit Group Co., Ltd., Ningbo 315101)

Abstract: With the rapid development of large AI models, the demand for their application in urban rail transit is increasing rapidly. However, there is still a lack of evaluation standards and methods for large models specific to the urban rail transit industry. To address this, this paper establishes an evaluation system for large models tailored to the urban rail transit industry. This evaluation system covers the construction of evaluation datasets, the design of evaluation dimensions, the selection of evaluation methods, and the formulation of evaluation metrics. Through this system, the performance of urban rail transit large models for urban rail transit in tasks such as language understanding, knowledge reasoning, image-text alignment, and spatiotemporal consistency can be evaluated comprehensively and objectively. Meanwhile, a dedicated evaluation system for large models for urban rail transit is designed and developed, and the effectiveness of the evaluation system is verified through practical application. The verification results show that the established evaluation system can provide an evaluation basis for the optimization and application of large models in urban rail transit, and promote their implementation and application in the urban rail transit industry.

Keywords: urban rail transit; artificial intelligence; large models for urban rail transit; model evaluation; evaluation system

人工智能依托大规模语言模型(large language models，LLMs)实现技术跃迁，已成为推动传统行业数字化转型的核心引擎。在城市轨道交通领域，随着线网规模持续扩张与运营复杂度不断提升，行车调度、设备运维、

收稿日期：2025-08-15　修回日期：2025-12-09
第一作者：华福才，男，博士，教授级高级工程师，长期从事隧道及地下工程的设计与科研工作，huafc@bjucd.com
基金项目：山东省重点研发计划(重大科技创新工程)项目(2025CXGC010108)
引用格式：华福才，邵昕，夏秀江，等. 城轨领域人工智能大模型评测研究与实践[J]. 都市快轨交通，2026，39(3)：11-18.
Hua Fucai, Shao Xin, Xia Xiujiang, et al. Research and practice on the evaluation of artificial intelligence large models in urban rail transit[J]. Urban rapid rail transit, 2026, 39(3): 11-18.

站务管理等核心环节对智能化技术的需求愈发迫切。2025 年国务院印发《关于深入实施“人工智能+”行动的意见》[1]，推动 AI 与各行业融合，城轨领域青岛地铁等已推出多款行业大模型，但缺乏统一评测标准导致技术路线碎片化、模型可信度难量化，通用大模型评测已形成“能力维度-评测方法-基准数据集”的成熟框架[2-3]，现有研究多将通用能力划分为语言生成、知识记忆、逻辑推理、跨模态理解 4 类，大规模多任务语言理解(massive multitask language understanding，MMLU)、通用语言理解评估(general language understanding evaluation，GLUE)等基准也成为通用模型性能对比的核心依据[4]。但通用评测存在显著局限，指标设计侧重通用能力，未考虑垂直行业的专业约束。数据集以通用文本为主，缺乏真实行业场景数据，难以反映模型实际应用性能，这为垂直领域评测体系的构建提供了研究空间。

在医疗[5]、金融[6]、法律[7]等其他垂直行业，已针对通用评测的局限性，探索形成适配自身需求的大模型评测范式。医疗领域聚焦临床安全与诊断准确性，金融领域围绕风险控制与数据安全，均建立了匹配行业核心需求的评测维度与方法。城轨领域目前尚未形成匹配运营场景核心需求(如安全性、实时性、协同性)的系统化评测框架，缺乏专业数据集与适配的评估方法，且评估方法单一，仅依赖定量指标而忽视定性评估与场景化测试。为此，本文基于城轨行业知识库[8]，构建专属评测体系并落地实践，以解决城轨大模型规模化落地瓶颈为目标，以运营场景为切入点，采用理论构建-工程落地的路径，形成“标准体系框架+技术平台”的完整解决方案，为行业大模型优化与落地提供技术支撑。

1 城轨领域大模型评测体系框架

当前城轨大模型评测存在“体系缺失、行业适配不足”的痛点，即缺乏统一的系统化评测框架，企业自建评估维度零散、指标泛化，通用评测难以覆盖城轨高安全、强实时、多专业协同的核心需求，导致模型性能不可比、落地价值难量化，亟需构建适配行业特性的体系化评测方案。

以《交通强国建设纲要》[9]和《城市轨道交通技术发展纲要建议(2021—2025)》[10]为依据，结合城轨工程智能化需求，明确安全、高效、经济 3 项原则为核心指引，涵盖评测维度、评测指标、评测方法、评测集 4 个核心模块。整体评测体系如图 1 所示，其中城轨大模型评测维度是体系的顶层框架，界定评测的核心方向，为后续方法与指标设计提供指引；城轨大模型评测指标是体系的量化标尺，需将维度要求与方法路径转化为可计算、可对比的具体指标，实现评测结果的标准化呈现；城轨大模型评测方法是体系落地的实施路径，需针对不同评测维度，提出适配城轨场景的评估手段，解决“如何评”的问题；城轨大模型评测数据集是体系构建的核心基础，通过数据集建设明确评测的数据底座。四者层层递进、相互支撑：数据集为维度、方法、指标提供数据验证基础；维度指导方法与指标的设计方向；方法决定指标的计算逻辑；指标则反向验证维度的合理性与方法的有效性。

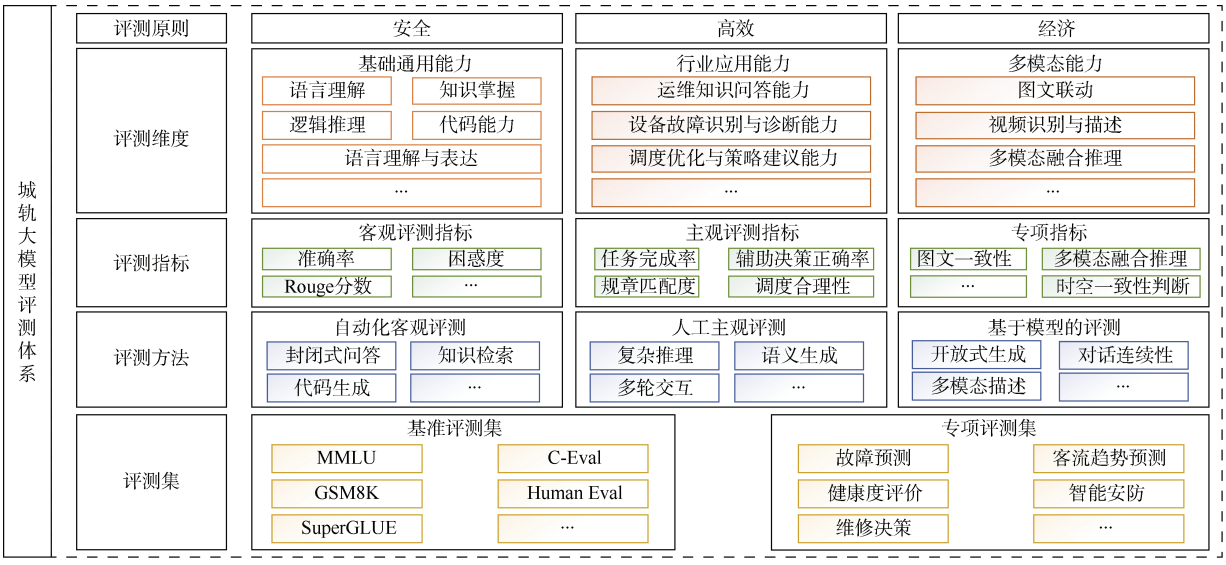

图 1 城轨大模型评测体系

Figure 1 Evaluation system for large models in urban rail transit

本评测体系并非静态架构，而是随城轨行业技术演进与业务场景拓展持续动态迭代的：评测集需适配新场景更新数据底座；评测方法要同步行业技术进展优化评估手段；评测指标则结合实际应用需求迭代量化标尺，以此持续匹配城轨大模型的发展需求，共同构成兼具稳定性与适应性的有机整体，为其后续工程化落地及持续优化提供理论框架支撑。

城轨领域大模型评测体系将“安全为基、高效为要、经济为本”三大原则贯穿于基础通用能力、行业应用能力、多模态能力的全维度评测。从不同层面聚焦三大原则，各有侧重、协同互补：基础通用能力聚焦“底线保障”，在安全维度侧重数据合规性、模型稳定性与容错性，在高效维度关注通用任务处理效率，在经济维度强调通用能力复用对研发成本的降低价值；行业应用能力聚焦“场景落地”，在安全维度覆盖运营安全、行车安全、乘客安全、财务安全等核心场景诉求，在高效维度聚焦运维、调度、能源管控等场景运营效率，在经济维度突出降本增效的实际落地价值；多模态能力聚焦“智能升级”，在安全维度强化多源数据协同决策的安全冗余，在高效维度提升复杂场景下多模态数据融合处理的决策效率，在经济维度通过跨场景智能协同减少资源浪费。三者围绕“安全、高效、经济”核心原则，从基础保障、场景落地、智能升级3个层面形成全方位评测覆盖，既契合城轨行业特殊运营需求，又实现评测维度的逻辑闭环与价值统一。

1.1 城轨大模型评测数据集

在进行城轨行业评测体系研究时，构建城轨专项评测数据集是评估城轨行业相关模型性能的关键基础工作。目前对大模型能力的评估数据集覆盖了语言理解、知识问答、数学推理和代码生成等通用评测基准，整理部分典型的公开数据集如表1所示[11-12]，这些任务适用于广泛的应用领域，但缺乏特定行业背景。

在城市轨道交通智能化发展进程中，面向大模型的行业数据集构建正逐步成为支撑其研发与评估的关键基础。城轨行业数据集是指基于真实系统场景，从列车、轨道、信号、站务等子系统中采集、清洗并结构化的多源数据集合，广泛涵盖设备状态、运维记录、客流信息、环境参数、信号日志等多维度信息[13]，用于模型训练、测试及性能评估。

相较于通用领域数据集，城轨数据集具有更强的行业专属性、实时性与结构复杂性，关键特征为：强场景依赖性，即设备故障与天气、时段等上下文关联紧密；多源异构性，含结构化(如传感器读数)、半结构化(如维护日志)、非结构化(如监控视频)数据；时空连续性，即具有时间序列与空间分布特征，适配列车运行建模、乘客行为预测；高安全隐私要求，即涉及公共交通运行安全，需遵循数据脱敏与合规使用原则。

表1 通用数据集

Table 1 Common datasets

名称	内容	评测维度	数量	评估指标
MMLU	覆盖人文学科、社会科学等共57个主题的评测数据集	知识问答	15 000	Accuracy
C-Eval	涵盖人文社科、理工科等多个学科的多项选择题基准	知识问答	13 948	Accuracy
GSM8K	数据题基准	数学推理	8 500	Accuracy
Human Eval	代码编程，每个题目有函数签名、文档字符串及测试用例，评估模型代码生成能力	代码编程	164	Pass@k
BBH	高难度情景下测试对人类平均认知水准模拟	综合能力	23	Varies
SuperGLUE	评估模型在复杂语言理解和推理任务上的性能	语言理解	8	Varies

为提升模型行业适应性与泛化能力，城轨专项评测数据集构建至关重要：需借鉴公开基准评测集(如语言理解、数学推理)的结构规范与能力维度，结合城轨典型场景及专业任务定制设计。

在构建城轨专项数据集时，以城轨车载传感器预测任务为例，评估模型在故障诊断能力中的表现，每个示例数据都包含不同任务的标注信息或字段说明，包含任务类型、设备类型、设备ID、传感器数据、预测值、正确值，按照示例数据的形式进行数据组织，形成专项数据集。

示例：城轨车载传感器预测任务

```
{
    "task"："sensor_data_prediction"，//任务类型，表示预测
    "sensor"："WheelWearSensor"，//传感器类型
    "vehicle"："Train115"，//车辆编号
    "sensor_reading"：
    {//传感器可读数据，包括多个维度
    "wear_level"：0.18，//磨损程度
    "temperature"：60//温度}，
    "predicted_wear_level"：0.20，//预测数据
    "correct_prediction"：0.20//正确预测值
    }
}
```

通过构建覆盖设备故障诊断、调度优化、智能安防等多类任务的专项数据集，可为大模型的精细化评估提供准确依据，确保评测的准确性和可操作性，促进人工智能在城轨复杂场景中的可靠部署与实用转化。

1.2 城轨大模型评测维度

评测维度系统设计是模型质量控制的前提，也为后续评测指标体系与评估方法构建提供理论依据和操作支撑。为全面科学衡量大模型城轨任务性能，从基础通用、行业应用、多模态扩展能力三方面构建多维度评测体系，聚焦城轨实际运营中的核心诉求与潜在痛点。

通过城轨大模型能力界定评测内容，如表 2 所示，建立能力与评测内容的映射关系，基础支撑能力聚焦通用能力维度的核心评测内容；应用执行能力对应行业应用能力与多模态能力的场景化评测。

表 2 评测维度设计

Table 2 Evaluation dimension design

评测维度	评测内容	评测任务	原则对应
基础通用能力	语言理解	评估命名实体识别、术语理解等自然语言处理基础能力及任务处理效率	高效
	知识掌握	评估模型对城轨基础知识、常识的高效覆盖深度与准确性	高效
	逻辑推理	测试模型在交通逻辑关系构建、因果分析等方面的能力及运算效率	高效+经济
	语言生成与表达	测试模型生成语言的清晰度、准确性、专业性及输出效率	高效
	代码能力	考察模型生成代码的合规性、专业性及开发效率	高效
	数据安全合规	评估模型在数据脱敏、隐私保护、权限管控等方面的合规性	安全
行业应用能力	文档解析与信息抽取能力	面向行业格式文档，测试表格识别、重点提取等任务表现，同时评估敏感运营数据的防护能力及信息抽取效率	安全+高效
	运维知识问答能力	结合行业标准等评估模型对复杂规程的理解与解答能力，同时考察运维操作的安全规范符合性及问答响应效率	安全+高效
	设备故障识别与诊断能力	评估模型故障定位、根因推理等能力，同时验证故障预警的安全时效性、诊断准确性及运维成本优化价值	安全+高效+经济
	调度优化与策略建议能力	考察模型制定列车调度方案的能力，同时评估调度方案的安全冗余设计、运营效率提升效果及能耗成本控制作用	安全+高效+经济
多模态能力	图文联动	测试图像与文本描述的关联能力，同时验证设备故障图像数据的安全脱敏、图文关联效率及故障定位成本优化	安全+高效+经济
	视频识别与描述	评估安防监控视频中异常行为的识别能力，同时考察识别的安全响应及时性、准确率及安防人力成本节约价值	安全+高效+经济
	多模态融合推理	检验多源数据下的综合推理能力，同时评估协同决策的安全风险规避能力、推理效率及跨场景资源协同降本效果	安全+高效+经济

1.3 城轨大模型评测方法

为了全面衡量大语言模型在城市轨道交通场景下的实际应用价值与智能表现，本研究构建了由自动化客观评测、人工主观评测以及基于模型的辅助评测 3 类方法组成的多层次评估体系。三者相辅相成，从准确性、可读性到创新性、可扩展性等多个维度对模型进行系统性质量判断，既满足评测效率的需求，又兼顾复杂任务的语义精度与领域适配性[14]。

自动化客观评测是大模型评估的基础高效方法，依托标准答案对比与规则匹配，适用于封闭式问答、知识检索等任务，以准确率、F1 值、BLEU/ROUGE 分数等为常用指标，虽具有一致性与可重复性、能快速反馈模型迭代性能，但难以覆盖生成型任务的主观质量。人工主观评测作为必要补充，可应对模型复杂推理、语义生成等任务的主观性问题，由行业专家定性打分(评估语言清晰度、回答相关性等)，能贴近业务需求、弥补自动评测的主观判断不足并考核模型合规性，但存在评估成本高、重复性差、难以满足大规模标准化需求的问题。

为平衡效率与质量，引入基于模型的辅助评测方法：通过辅助评分模型(如教师模型、评测模型)对被测模型输出进行语义理解与评分推断，适用于开放式生成、对话连续性等场景，兼具高效性与一定主观语义识别能力，不过其评估准确性受自身性能影响，需通过多模型交叉验证与人工抽查修正。

综上所述，3 类评测方法各具优势，适用对象和任务类型存在显著差异，结合使用才能实现对城轨领域大模型性能的全面、系统评估。未来，在统一评测平台的支持下，构建集自动评分、人工审核与模型辅助融合的评估工作流，将成为推动城轨智能系统高质量发展的关键基础之一。

1.4 城轨大模型评测指标

为确保城轨行业大模型评估的系统性与可操作性，在遵循“安全、高效、经济”3 项原则基础上，构建了覆盖模型关键能力的评测指标体系，该体系由基础通用能力、行业应用能力、专项多模态能力 3 类指标构成，分别对应语言理解与知识能力、城轨场景任务完成表现、复杂交互的跨模态推理能力，覆盖模型基础性能到实际落地能力的全链路评估。

在具体指标设计中，遵循“可测量性、业务关联性、可解释性、可扩展性”的设计要求。可测量性确保评测结果客观量化；业务关联性要求指标应与城轨

核心任务场景(如巡检识别、调度决策、安全检测)紧密结合；可解释性保证指标能反映模型在语义理解、推理认知、跨模态融合等方面的真实差异；可扩展性则支持体系随业务需求拓展新的专业指标。

基于上述原则与设计要求，并结合城轨典型多模态任务特征，提出 3 项具有行业针对性的专项评估指标：图文一致性指标(image-text alignment score，ITA)、多模态融合推理指标(multimodal fusion reasoning score，MRS)与时空一致性判断指标(spatiotemporal consistency score，STC)。

1.4.1 图文一致性得分

在城轨运维过程中，巡检人员往往通过携带终端拍摄现场设备照片，并上传文字描述用于远程诊断。此类信息需输入至大模型进行初步筛查与判断。ITA 指标用于评估模型是否能够准确理解图像中所示设备状态(如开裂、积水、电缆脱落)与文字描述的一致性。ITA 的计算式为：

$$\mathrm{ITA}=\cos\left(f_I(I),f_T(T)\right)=\frac{f_I(I)\cdot f_T(T)}{\left\|f_I(I)\right\|\cdot\left\|f_T(T)\right\|} \tag{1}$$

式中，I 代表输入图像；T 代表图像对应的文字描述；$f_I(\cdot)$将图像映射为语义嵌入向量的图像编码器(如 CLIP、BLIP 等模型)；$f_T(\cdot)$表示将文本映射为嵌入向量的文本编码器；cos(·，·)为余弦相似度函数，输出范围为[–1，1]，越接近 1 表示图文语义越一致。

该指标在设备故障自动预警、远程专家审核等场景中具有广泛应用价值，是模型图文理解与知识联想能力的核心体现。

1.4.2 多模态融合推理得分

城轨调度中心常需处理来自视频监控、实时运营数据、乘客反馈文本等多源信息的融合输入，模型需基于图像、文字、数值数据等多模态输入，判断事件性质并提出调度建议。MRS 的计算式为：

$$\mathrm{MRS}=\frac{1}{N}\sum_{i=1}^{N}I\left(\hat{y}_i=y_i\right) \tag{2}$$

式中，N 代表多模态推理样本总数；$\hat{y}_i$ 表示基于多模态输入给出的第 i 个任务推理结果；y_i 表示人工标注的参考正确答案；$I(\cdot)$为指示函数，判断正确为 1，否则为 0。

MRS 指标衡量模型是否能从多源信号中整合关键特征，并形成有效的推理输出，是验证模型真实决策能力的关键指标。

1.4.3 时空一致性判断能力

在城轨安防监控中，模型需从连续的视频片段中提取时间序列事件(如乘客跌倒、闯入轨道区间)，并结合监控地图标注准确定位其发生时间与空间位置。STC 指标用于衡量模型是否具备事件识别的时间逻辑推断能力与空间位置判断能力。STC 的计算式为：

$$\mathrm{STC}=\lambda\cdot S_{\mathrm{temp}}+(1-\lambda)\cdot S_{\mathrm{spatial}},\ \lambda\in[0,1] \tag{3}$$

式中，λ 表示时间一致性与空间准确性权重系数；S_{temp} 表示时间一致性得分，反映模型输出的事件时间序列与真实顺序的一致程度；S_{spatial} 表示空间定位准确率。

STC 指标用于事故预防、行为溯源、应急响应辅助，是构建智能安防系统的重要基础能力。

城轨领域大模型评测体系主要评测对象是行业大模型，面向的主要用户群体是运营企业，围绕行业生态，也可为技术研发与供应商、港口、铁路等领域运营方提供借鉴。

2 城轨领域大模型评测系统应用实践

以城轨大模型评测体系为核心依据，通过设计、开发城轨大模型评测系统，将理论层面的评测逻辑转化为工程化技术平台，并结合实测实验验证理论体系的有效性与系统的实用性。整体按照系统架构设计、系统功能设计和评测实验结果分析展开，实现对城轨大模型功能准确性、实时性、安全性和鲁棒性等维度的量化评估，为城轨大模型的迭代优化与规模化应用提供实践支撑。

2.1 系统架构设计

城轨领域大模型评测系统架构遵循松耦合、高效、可扩展、灵活原则(见图 2)：通过标准化接口实现模块松耦合，关键模块支持用户自定义(如大模型接入、评测数据集上传、指标定义)，允许灵活切换评测方法，以适配领域内多样化、定制化评测任务；评测任务调度机制保障高效执行。系统提供 Web 端操作界面，支持大模型性能对比分析、评测报告生成；对外提供应用程序编程接口(application programming interface，API)与模型上下文协议(model context protocol，MCP)大模型评测服务。

1) 评测数据层：作为系统基础层，存储管理评测所需各类数据集[15]，含通用、城轨领域数据集(覆盖自然语言处理、行业知识问答等类型)，支持用户自定义数据集。

2) 模型适配层：实现模型元数据管理与多大模型[2]统一接入，经模型适配器无缝集成不同类型、不同框架大模型，支持模型注册、版本管理、能力描述，提供统一接口规范以屏蔽底层模型差异，便于评测任务调用。

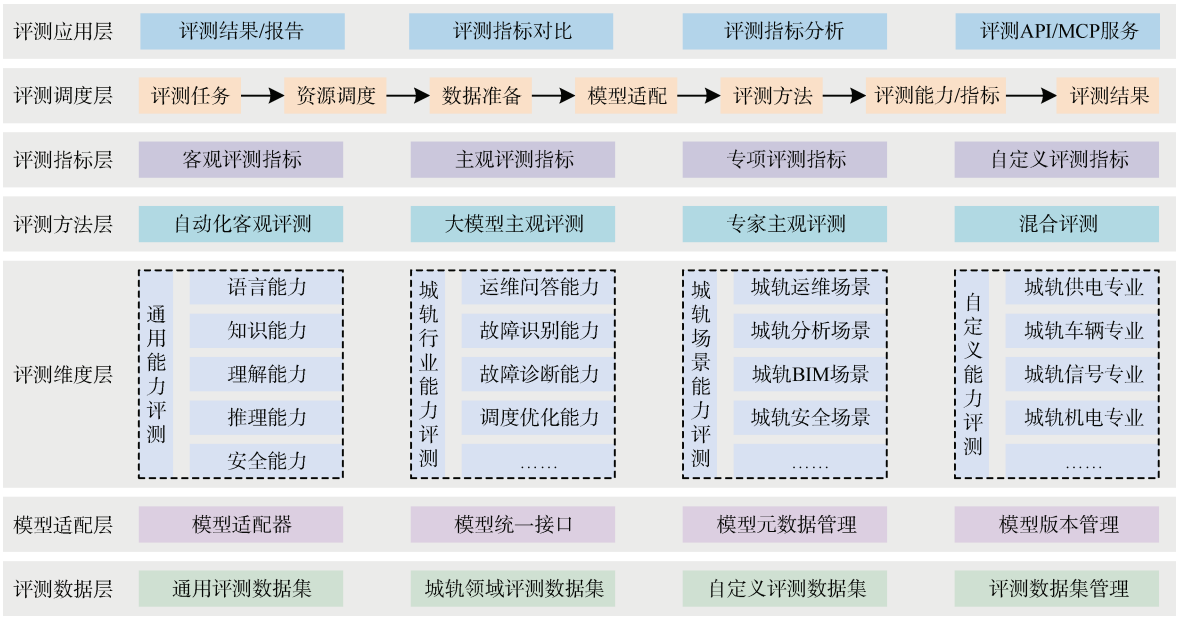

图 2　城轨大模型评测系统架构

Figure 2　Architecture of the evaluation system for large models in urban rail transit

3) 评测维度层[16]：细分模型能力维度(通用、行业、场景、自定义)，评测模型各维度表现，支持多层次、多领域能力评测。

4) 评测方法层[17]：支持自动化客观、大模型主观、专家主观及混合评测，模块化设计便于集成扩展新方法，适配不同大模型与评测任务。

5) 评测指标层：定义管理各类评测指标(系统核心标准库)，支持多维度多类型指标体系(客观、主观、场景、自定义)，满足城轨领域大模型评测需求。

6) 评测调度层：为系统中枢，评测任务[18]提升效率，保障多任务并发时系统稳定。

7) 评测应用层：面向用户提供评测结果可视化、报告自动生成、指标对比分析；通过 API 或 MCP[19]接口开放评测服务，便于其他系统集成，提升服务能力与行业影响力。

2.2　系统功能设计

为实现城轨领域大模型评测系统高效运行与智能化管理，设计覆盖数据、模型、评测、分析及服务全流程的功能模块，引入模块化、自定义设计理念，通过开放式接口支持用户灵活配置功能，提升系统场景适应性与可扩展性。

1) 评测数据层实施评测数据集全生命周期管理，支持导入、导出与版本控制以保障数据可追溯性与一致性；分类管理通用及城轨专用数据集，允许用户上传自定义数据集，结合细粒度数据权限与访问控制确保安全合规。

2) 模型适配层实现多类型、多框架大模型统一接入与高效管理，支持模型注册、元数据维护及版本控制；通过适配器机制屏蔽模型差异、统一接口封装，提供模型能力描述与标签管理，结合状态监控掌握模型实时运行状况。

3) 评测维度层划分通用、城轨行业、城轨场景、自定义 4 大能力维度，分别对应语言、知识、推理等基础能力，运维问答、故障诊断等行业能力，运维、BIM 等场景专项能力，支持能力项灵活配置扩展，适配实际评测需求。

4) 评测方法层管理多类评测手段以适配不同模型与场景，集成自动化客观评测(标准数据集自动打分)、大模型主观评测(AI 判断答案)、专家主观评测(人工打分采集)及混合评测；支持任务自动分发、数据动态分配与评测过程实时监控、中断恢复，可后续集成新算法流程，适配动态需求。

5) 评测指标层负责评测指标全生命周期管理与灵活配置，内置涵盖客观指标(准确率、召回率、推理速度)、主观指标(用户满意度、专家评分)、场景指标(城轨运维、调度)及自定义指标的指标库；支持指标动态增删改查、分组、权重分配与可视化配置，用户可灵活组合调整指标体系；与方法层、能力层联动，实现指标映射与结果归集，保障评测科学性与可追溯性。

6) 评测调度层作为系统核心枢纽，全流程自动化

管理调度评测任务：接收解析任务后，调度分配计算、存储资源，自动完成数据准备(数据集选择、预处理、分发)，调用模型适配模块实现大模型统一接入，再调度评测方法与能力指标模块评估模型表现，最终生成标准化评测报告。过程中实现任务监控、异常处理、日志记录及优先级管理，保障评测高效稳定。

7) 评测应用层承担系统与用户的交互及服务输出，集成评测结果与报告自动生成及可视化展示，支持多模型、多维度指标对比以助力用户掌握模型表现差异；具备评测数据统计、对比分析与可视化解读功能，辅助识别模型优劣势与改进方向；提供标准化评测 API 与 MCP 服务接口，支持外部系统集成调用，实现评测能力服务化输出，提升系统开放性与易用性。

2.3 评测实验分析

为验证城轨大模型在实际场景的性能优势，基于城轨领域大模型评测体系，开展多主体大模型对比评测实验，并与无大模型的传统方案进行基线对照。

2.3.1 评测对象选取

选取研究院、高校、科技公司 3 个机构研发的大模型，依次标记为：模型 M1(研究院研发)、模型 M2(XX 大学研发)、模型 M3(XX 科技公司研发)。

2.3.2 评测流程

根据城轨领域大模型评测体系，设置模型通用基础、城轨专业和综合适配的能力。通用能力参考 MMLU、GLUE 等通用基准，评估语言理解、常识推理等基础性能；城轨领域能力聚焦城轨核心场景，涵盖图文一致性(ITA)、多模态融合推理(MRS)和时空一致性判断(STC)三类核心指标，从不同维度量化模型的城轨适配性；采用加权综合评分法，按通用能力(30%权重)+城轨领域能力(70%权重)计算，评估不同模型的综合适配能力。

2.4 评测结果

各参评模型在各个能力维度下的得分情况及各模型的综合评分如表 3 所示，为模型选择和应用部署提供客观依据。在实际应用中，可根据具体场景需求，重点关注相应的专业指标，实现模型的最优配置。

表 3 评测结果汇总

Table 3 Summary of evaluation results

模型	通用能力	城轨领域能力			综合能力
		ITA	MRS	STC	
模型 M1(研究院)	78.6	80.71	84.36	78.28	80.67
模型 M2(大学)	81.2	74.58	79.99	74.80	78.12
模型 M3(科技公司)	86.1	73.04	75.63	79.38	78.79

评测显示，不同机构模型在城轨领域表现的差异与其模型训练路径和数据来源密切相关。研究院模型在多模态任务上表现突出，得益于领域语料与图像数据的针对性优化；科技公司模型在时空一致性判断上表现较好，反映其底层架构对时间序列信息的建模能力；高校模型虽在通用语义理解方面具有优势，但缺乏针对城轨任务的专项微调，导致专业指标表现相对不足。

本研究依据与业务场景高度相关的专项评测指标。在安全维度中，引入故障诊断错误率、合规性符合率、数据安全防护能力等指标，以验证模型在故障识别、规章遵循及信息安全方面的可靠性。在高效维度中，增加推理速度、任务完成耗时等时效性指标，用于衡量模型在运营信息处理与应急响应中的效率表现。在经济维度中，将模型部署与推理的算力成本、运维成本纳入评估，以刻画模型在全生命周期内的资源利用与成本收益表现，并通过量化结果支撑对模型能力边界、适用场景与优化方向。

总体而言，城轨领域大模型需在“图文一致、多模态融合、时空一致”等多方面形成协同能力，以满足安全、高效、经济三项原则。未来可依托城轨多模态知识库、结合专业语料微调及指令式训练，推动大模型从通用智能向领域智能演进。

3 结论

本文构建的城轨大模型评测体系，是在通用大模型评测框架基础上，结合城轨行业高安全、高效率、多专业协同等业务属性形成的专项评测方案。该体系包括评测数据集、评测维度、评测方法和评测指标等核心内容，并进一步提出图文一致性、多模态融合推理、时空一致性判断等行业专项指标。作为城轨大模型研发与应用的评测支撑，该体系可有效量化不同模型的能力差异，为模型选型、优化迭代和工程化落地提供依据。

未来，随着城轨智能化场景研究的不断深化，可将评测任务拓展至城轨全生命周期场景，持续完善指标体系以适配技术演进与业务新需求；同时加强评测系统与实际业务系统的深度融合，推动评测结果在模型优化、场景适配和运营决策中落地应用，助力城轨大模型价值充分释放。

参考文献

[1] 中华人民共和国国务院. 关于深入实施“人工智能+”行

动的意见[EB/OL]. (2025-08-21)[2025-09-01]. https://www.gov.cn/gongbao/2025/issue_12266/202509/content_7039598.html.

[2] Zhao W X, Zhou K, Li J, et al. A survey of large language models[J]. arXiv preprint arXiv: 2303.18223, 2023, 1(2): 1-124.

[3] Bommasani R, Liang P, Lee T. Holistic evaluation of language models[J]. Annals of the New York academy of sciences, 2023, 1525(1): 140-146.

[4] MEVA D, KUKADIYA H. Performance evaluation of large language models: a comprehensive review[J]. International research journal of computer science, 2025, 12(3): 109-114.

[5] SINGHAL K, AZIZI S, TU T, et al. Large language models encode clinical knowledge[J]. Nature, 2023, 620(7972): 172-180.

[6] Wu S, Irsoy O, Lu S, et al. BloombergGPT: A large language model for finance[P/OL]. arXiv: 2303.17564, 2023[2025-06-01]. https://arxiv.org/abs/2303.17564.

[7] GUHA N, NYARKO J, HO D, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models[J]. Advances in neural information processing systems, 2023, 36: 44123-44279.

[8] 于松伟, 刘巍, 夏秀江, 等. 基于知识图谱的城轨大模型RAG检索增强知识库构建研究[J]. 都市快轨交通, 2025, 38(2): 1-7.
YU Songwei, LIU Wei, XIA Xiujiang, et al. Constructing a retrieval-augmented generation knowledge base for urban rail transit large language models: a knowledge graph-based approach[J]. Urban rapid rail transit, 2025, 38(2): 1-7.

[9] 中共中央, 国务院. 交通强国建设纲要[EB/OL]. (2019-09-19)[2025-06-01]. https://www.gov.cn/zhengce/2019-09/19/content_5431432.htm.

[10] 中国城市轨道交通协会. 城市轨道交通技术发展纲要建议(2021-2025)[R/OL]. (2020-10-21)[2025-06-01]. http://www.camet.org.cn/.

[11] Wang A, Pruksachatkun Y, Nangia N, et al. SuperGLUE: A stickier benchmark for general-purpose language understanding systems[C/OL]//Advances in neural information processing systems. 2019, 32: 3266-3280[2025-06-01]. https://proceedings.neurips.cc/paper/2019/hash/4496bf24925820579a296e4c964599e1-Abstract.html.

[12] TONG X, JIN B, LIN Z, et al. CPSDbench: a large language model evaluation benchmark and baseline for Chinese public security domain[J]. International journal of data science and analytics, 2025, 20(4): 3205-3234.

[13] 韩德志, 韩晓艺, 华福才, 等. 数字孪生与多系统融合的城轨综合管控平台研究与实施[J]. 都市快轨交通, 2025, 38(1): 112-119.
HAN Dezhi, HAN Xiaoyi, HUA Fucai, et al. Research and implementation of an urban rail integrated management and control platform with digital twin and multi-system integration[J]. Urban rapid rail transit, 2025, 38(1): 112-119.

[14] 罗文, 王厚峰. 大语言模型评测综述[J]. 中文信息学报, 2024, 38(1): 1-23.
LUO Wen, WANG Houfeng. Evaluating large language models: A survey of research progress[J]. Journal of Chinese information processing, 2024, 38(1): 1-23.

[15] Liu Y, Cao J, Liu C, et al. Datasets for large language models: A comprehensive survey[J]. Artificial intelligence review, 2025, 58(12): 403.

[16] CHANG Y P, WANG X, WANG J D, et al. A survey on evaluation of large language models[J]. ACM transactions on intelligent systems and technology, 2024, 15(3): 1-45.

[17] Chiang C H, Lee H. Can large language models be an alternative to human evaluations?[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023: 15607-15631.

[18] 赵志泉, 胡蝶, 刘畅, 等. 人文社科领域中文通用大模型性能评测[J]. 图书情报工作, 2024, 68(13): 132-143.
Zhixiao Z, Die H, Chang L, et al.Performance Evaluation of Chinese Universal Large Model in the Field of Humanities and Social Sciences[J]. Library and information service, 2024, 68(13): 132-143.

[19] Hou X, Zhao Y, Wang S, et al. Model context protocol (mcp): Landscape, security threats, and future research directions[J/OL]. ACM Transactions on Software Engineering and Methodology, 2025, 34(2): 1-35. https://doi.org/10.1145/3680490.

（编辑：王艳菊）