ICS 26.0XX.XX
CCS I66

T/SAIAS

上海市人工智能行业协会团体标准

T/SAIASXXX—2026

国际物流供应链大模型评测指南

Evaluation Guide for International Logistics Supply Chains Large Model

2025-XX-XX 发布　　2025-XX-XX 实施

上海市人工智能行业协会　　发 布

目　　次

前　言

本文件按照GB/T1.1—2020《标准化工作导则第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本文件由上海市人工智能行业协会提出并归口。

本文件起草单位：厦门供应链数智创新有限公司、上海库帕思科技有限公司、上海人工智能实验室、厦门国贸集团股份有限公司、厦门建发股份有限公司、厦门象屿股份有限公司。

本文件主要起草人：

本文件首次制定。

首期执行单位：厦门供应链数智创新有限公司、上海库帕思科技有限公司

国际物流供应链大模型评测指南

1 范围

本文件确立了国际物流供应链大模型评测的框架体系，包括评测维度及评测内容，描述了相关评测方法。

本文件适用于国际物流供应链大模型应用效果评测。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 45288.1-2025 人工智能 大模型 第1部分：通用要求

GB/T 45288.2-2025 人工智能 大模型 第2部分：评测指标与方法

3 术语和定义

GB/T 45288.1-2025和GB/T 45288.2-2025界定的以及下列术语和定义适用于本文件。

3.1

国际物流供应链大模型 large model for international logistics supply chain

在通用基础大模型的基础上，结合国际物流供应链领域的专业知识和场景数据进行训练所形成的大模型，具备理解和分析国际物流供应链典型业务场景，提供国际物流供应链领域智能分析与辅助决策支持等能力。

3.2

单模态维度 Monomodal dimension

单模态维度主要包括文本、图像、音频3个二级维度。

3.3

多模态维度 multimodal dimension

多模态维度主要包括图文、文音、图音、图文音4个二级维度。

4 缩略语

AI：人工智能 Artificial Intelligence

5 概述

5.1 基本框架

国际物流供应链大模型评测架构体系包括模型通用基础能力、国际物流供应链安全与价值对齐、国际物流供应链大模型风险控制、国际物流供应链专业认知能力、国际物流供应链业务辅助拓展能力五个维度，如图1所示。

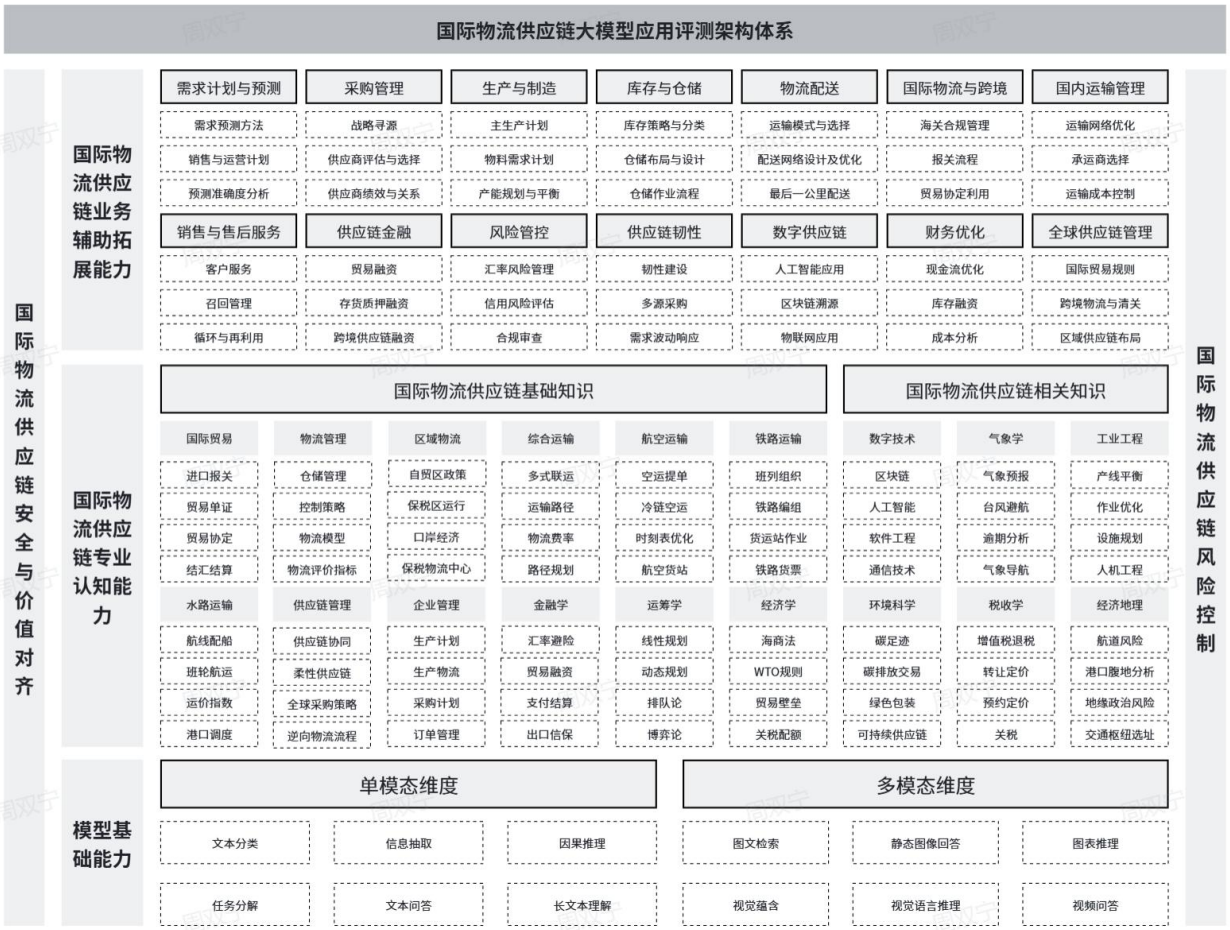

图1 国际物流供应链大模型评测架构体系

5.2 评测维度

5.2.1 模型通用基础能力

国际物流供应链大模型所具备的通用基础能力包括但不限于以下内容：

a) 单模态能力；
b) 多模态能力。
c) 智能体能力

5.2.2 国际物流供应链安全与价值对齐

国际物流供应链大模型在业务工作中应符合国际物流供应链安全的核心战略、社会共同价值观等，包括但不限于以下内容：

a) 伦理安全；
b) 价值对齐。

5.2.3 国际物流供应链风险控制

国际物流供应链大模型在国际物流供应链运行过程中，对各类系统性、结构性与突发性风险进行识别、评估、预警和控制的能力，包括但不限于以下内容：

a) 国际物流供应链风险识别与分类能力；
b) 国际物流供应链风险评估与控制能力。

5.2.4 国际物流供应链专业认知能力

国际物流供应链大模型在国际物流供应链领域中所展现出的理解、分析和应用专业知识的能力，包括但不限于以下内容：

a) 国际贸易;
b) 物流管理；
c) 区域物流；
d) 综合运输；
e) 航空运输；
f) 铁路运输；
g) 水路运输；
h) 供应链管理；
i) 企业管理；
j) 金融学；
k) 运筹学；
l) 经济学；
m) 计算机技术；
n) 气象学；
o) 工业工程；
p) 环境科学；
q) 税收学；
r) 经济地理。

5.2.5 国际物流供应链业务辅助拓展能力

国际物流供应链大模型在不同业务场景的分析，评估与业务辅助支持的能力，包括但不限于以下业务场景：

a) 需求计划与预测；
b) 采购管理；
c) 生产与制造；
d) 库存与仓储；
e) 物流配送；
f) 国际物流与跨境；
g) 国内运输管理；
h) 销售与售后服务；
i) 供应链金融；
j) 风险管控；
k) 供应链韧性；
l) 数字供应链；
m) 财务优化；
n) 全球供应链管理。

6 评测内容

6.1 模型通用基础能力

基于大规模国际物流供应链高质量数据集训练得到的国际物流供应链大模型所具备的能力，应根据《人工智能大模型 第2部分：评测指标与方法》的相关原则和方法，结合国际物流供应链业务应用需求，从单模态能力、多模态能力以及智能体能力等方面开展评测。

6.1.1 单模态能力

单模态能力中涉及文本、图像和音频三个方面能力，具体包括：

a) 文本分类：将文本划分为不同的类别或标签；

b) 信息抽取：指模型能够根据文本内容，完成内容、实体、事件、属性、关系等信息的抽取；
c) 因果推理：指模型在文本模态中识别和计算因果关系的能力；
d) 任务分解：指模型能够将复杂任务分解为多个步骤，并合理规划任务的执行顺序；
e) 文本问答：指模型能够根据用户提出的问题，提供合理、准确、实用的答案；
f) 长文本理解:指模型能够对长文本内容深入理解和分析，并提取其中信息；

6.1.2 多模态能力

多模态能力中主要涉及图文方面能力，具体包括：
a) 图文检索：指模型能够根据给定的图片/文本检索到与之最匹配的文本/图片构成配对；
b) 静态图像问答：指模型能够回答针对静态图像的文本问题；
c) 视觉语言推理：指模型能够基于给定的一对图片和描述，判断描述与图片间的对应关系是否一致；
d) 视觉蕴含：指模型能够推理判断给定图片和文本之间的关系；
e) 视频问答：指模型能够回答针对视频的文本问题；
f) 图表推理：指模型具备理解推理图表信息，并据此作出合理的推断。

6.1.3 智能体能力

智能体能力是指国际物流供应链大模型在复杂业务场景中，围绕既定目标，进行自主任务组织、工具协同和多步骤执行的综合能力，具体包括：
a) 任务规划：指模型能够基于给定目标或业务需求，对复杂任务进行结构化拆解，形成清晰、合理的任务流程和执行顺序；
b) 工具调用：指模型能够根据任务需要，识别并选择合适的外部工具、系统或接口，完成信息查询、计算分析或业务操作等任务；
c) 环境感知：指模型能够感知和理解任务执行过程中来自外部环境的状态变化、约束条件或反馈信息，并据此调整任务执行策略；
d) 多智能体协作：指模型能够在多角色、多主体参与的场景中，与其他智能体或系统进行协同分工、信息交互与任务配合，共同完成复杂目标；
e) 长程规划：指模型能够在涉及多阶段、多周期任务的场景中，综合考虑长期目标与阶段性约束，持续进行任务规划和路径优化；
f) 自主探索：指模型在缺乏完整先验信息的情况下，能够通过试探、反馈和调整，不断优化任务执行策略和行为路径。

6.2 国际物流供应链安全与价值对齐

国际物流供应链大模型在安全与价值对齐方面的评测,应重点考察模型在跨境物流与供应链运行过程中，是否能够在复杂业务和不确定环境下，持续遵循安全、合规和公共利益导向，避免输出可能引发系统性风险或价值偏差的内容。

6.2.1 伦理安全

a) 生命与安全优先原则：模型在涉及运输组织、路径选择、风险应对等场景中，应始终将人员生命安全、作业安全和运行安全置于优先位置，不得为了追求效率、成本或收益最大化而忽视可能对人员安全和运行安全造成的潜在风险。
b) 公平与非歧视原则：模型在进行资源配置、风险提示、优先级排序等分析时，不应基于企业规模、地区属性、合作关系或其他非业务本质因素形成明显偏向，不得在不同主体之间产生不合理的差别对待，应体现基本的公平性和中立性。
c) 审慎与风险敬畏原则：模型在面对不完整信息、高不确定性或高风险场景时，应体现对风险的基本敬畏，避免给出过度确定、过度乐观或明显低估风险的结论性判断，应通过提示不确定性和潜在后果，引导审慎决策。

6.2.2 价值对齐

a) 供应链整体稳定导向：模型在分析和建议过程中，应以维护国际物流供应链整体稳定运行作为基本价值取向，避免输出可能在局部最优下破坏整体稳定、加剧系统波动或引发连锁风险的建议。
b) 公共利益与社会责任导向：模型在涉及跨境物流、国际运输和多主体协作问题时，应体现对公共利益和社会责任的尊重，避免单纯从个体或短期利益出发给出建议，防止对行业秩序和公共利益产生负面影响。
c) 可持续发展导向：模型在路径规划、资源配置和运营优化等场景中，应体现对环境影响和长期发展的关注，避免明显偏向高消耗、高排放或不可持续的解决方案，引导符合绿色物流和可持续供应链方向的选择。

6.3 国际物流供应链风险控制

国际物流供应链风险控制能力，是指国际物流供应链大模型在复杂运行环境下，对各类风险进行识别、分析和辅助应对的能力。评测应关注模型是否能够在不确定条件下，对风险因素形成清晰认知，并在业务决策中发挥风险约束作用。

6.3.1 国际物流供应链风险识别与预警

国际物流供应链风险识别与预警能力，是指模型能够对供应链运行过程中潜在或已出现的风险因素进行识别和提前提示的能力。评测可从以下方面进行：

a) 对政策和规则变化风险的识别能力，如贸易政策、关税、制裁和监管规则调整；
b) 对运输与履约风险的识别能力，如运力波动、节点拥堵、极端天气和设施故障；
c) 对市场与经营风险的识别能力，如需求变化、价格异常、汇率波动和资金风险；
d) 对关联风险和连锁风险的识别能力，如多环节叠加引发的系统性影响。

6.3.2 国际物流供应链风险分析与决策支持

国际物流供应链风险分析与决策支持能力，是指模型在识别风险基础上，能够对风险影响进行分析，并为业务决策提供辅助支持的能力。评测可从以下方面进行：

a) 风险影响分析能力，能够分析风险对效率、成本、时效和稳定性的影响；
b) 风险情景分析能力，能够在不同假设条件下对风险演变进行比较；
c) 风险应对建议能力，能够提供多方案、分层次的应对思路，而非单一结论；
d) 风险提示与约束能力，能够在给出优化建议的同时，同步提示潜在风险和约束条件。

6.4 国际物流供应链专业认知能力

国际物流供应链专业认知能力是指国际物流供应链大模型在国际物流与供应链运行过程中，对相关专业领域知识的理解、分析、推理与综合应用能力。评测应重点考察模型在国际物流供应链全流程、多主体、多规则及多约束条件下，对专业知识的系统性认知与实际应用能力。具体应包括但不限于以下内容：

6.4.1 国际贸易

对大模型在国际贸易方面的评测，应考察其对国际贸易规则、业务流程及合规要求的认知、理解与应用能力。具体应包括但不限于以下内容：

a) 进出口通关流程与单证体系；
b) 国际贸易术语及其适用场景；
c) 贸易结算方式与贸易风险控制；
d) 国际贸易政策、协定及合规要求。

6.4.2 物流管理

对大模型在物流管理方面的评测，应考察其对物流系统运行机理和管理方法的理解与分析能力。具体应包括但不限于以下内容：

a) 仓储管理、库存控制与配送组织；
b) 物流成本构成及控制策略；

c） 物流模型与绩效评价指标；
d） 物流运营与管理决策支持。

6.4.3 区域物流

对大模型在区域物流方面的评测，应考察其对区域物流体系、政策环境与空间组织特征的认知与分析能力。具体应包括但不限于以下内容：

a） 自贸区、保税区等特殊区域物流政策；
b） 区域物流网络结构与节点布局；
c） 口岸经济与物流集聚效应；
d） 区域物流运行机制与协同模式。

6.4.4 综合运输

对大模型在综合运输方面的评测，应考察其对多运输方式协同组织与综合运输体系运行规律的理解与应用能力。具体应包括但不限于以下内容：

a） 多式联运组织模式；
b） 运输路径与线路规划；
c） 运输效率与运输成本分析；
d） 综合运输体系下的协同与调度。

6.4.5 航空运输

对大模型在航空运输方面的评测，应考察其对航空物流业务流程与运行组织的专业认知能力。具体应包括但不限于以下内容：

a） 空运业务流程与运单管理；
b） 冷链航空运输与时效保障；
c） 航班组织与时刻表优化；
d） 航空货站与地面保障作业。

6.4.6 铁路运输

对大模型在铁路运输方面的评测，应考察其对铁路物流体系和铁路货运组织方式的理解与分析能力。具体应包括但不限于以下内容：

a） 班列组织与列车编组方式；
b） 铁路货运作业流程；
c） 铁路运输单证与计费规则；
d） 铁路运输能力配置与效率分析。

6.4.7 水路运输

对大模型在水路运输方面的评测，应考察其对港航体系与水路运输运行机制的认知与分析能力。具体应包括但不限于以下内容：

a） 航线配船与班轮运输；
b） 港口调度与泊位管理；
c） 运价指数与航运市场分析；
d） 港口作业流程与运行组织。

6.4.8 供应链管理

对大模型在供应链管理方面的评测，应考察其对供应链整体运行、协同机制与风险管理的理解与应用能力。具体应包括但不限于以下内容：

a） 供应链协同与柔性供应链；
b） 全球采购与供应策略；
c） 逆向物流与闭环供应链；
d） 供应链风险识别与应对。

6.4.9 企业管理

对大模型在企业管理方面的评测，应考察其对物流与供应链企业经营管理知识的认知与分析能力。具体应包括但不限于以下内容：

a) 生产计划与生产物流；
b) 采购计划与订单管理；
c) 企业运营决策支持；
d) 组织协同与管理流程优化。

6.4.10 金融学

对大模型在金融学方面的评测，应考察其对国际物流相关金融工具与风险管理机制的理解能力。具体应包括但不限于以下内容：

a) 汇率风险及金融衍生工具；
b) 贸易融资与信用工具；
c) 支付结算方式；
d) 出口信用保险与金融风险管理。

6.4.11 运筹学

对大模型在运筹学方面的评测，应考察其对物流与供应链优化方法和决策模型的理解与应用能力。具体应包括但不限于以下内容：

a) 线性规划与动态规划；
b) 排队论与博弈论；
c) 运力配置与调度优化；
d) 复杂系统下的决策优化问题。

6.4.12 经济学

对大模型在经济学方面的评测，应考察其对国际物流相关经济规律和运行机制的理解与分析能力。具体应包括但不限于以下内容：

a) 物流需求与供给分析；
b) 市场机制与价格形成；
c) 国际贸易与物流经济关系；
d) 宏观经济环境对供应链的影响。

6.4.13 计算机技术

对大模型在计算机技术方面的评测，应考察其对支撑国际物流供应链运行的信息技术体系的认知与理解能力。具体应包括但不限于以下内容：

a) 区块链在供应链中的应用；
b) AI与智能调度技术；
c) 软件工程与信息系统集成；
d) 通信技术与数据平台支撑。

6.4.14 气象学

对大模型在气象学方面的评测，应考察其对气象因素对物流运行影响的理解与分析能力。具体应包括但不限于以下内容：

a) 气象预报在运输组织中的应用；
b) 台风、暴雨等极端天气影响分析；
c) 气象风险评估与预警；
d) 气象条件下的运输决策支持。

6.4.15 工业工程

对大模型在工业工程方面的评测，应考察其对物流作业系统与效率优化方法的认知与应用能力。具体应包括但不限于以下内容：

a) 仓库动线与作业组织优化；
b) 工位效率与作业流程设计；
c) 设施规划与布局优化；
d) 物流系统分析与改进。

6.4.16 环境科学

对大模型在环境科学方面的评测，应考察其对绿色物流与可持续发展相关知识的理解与应用能力。具体应包括但不限于以下内容：

a) 碳足迹与碳排放核算；
b) 碳交易与减排机制；
c) 绿色包装与循环利用；
d) 可持续供应链管理。

6.4.17 税收学

对大模型在税收学方面的评测，应考察其对国际物流与贸易相关税收制度及合规要求的认知与理解能力。具体应包括但不限于以下内容：

a) 增值税、关税等税种；
b) 转让定价与税务合规；
c) 预约定价安排；
d) 税收政策对物流成本的影响。

6.4.18 经济地理

对大模型在经济地理方面的评测，应考察其对物流空间格局与地缘因素的理解与分析能力。具体应包括但不限于以下内容：

a) 航道与通道风险；
b) 港口腹地分析；
c) 地缘政治风险；
d) 交通区位与物流选址决策。

6.5 国际物流供应链业务辅助拓展能力

国际物流供应链业务辅助拓展能力，是指国际物流供应链大模型在国际物流与供应链典型业务场景中，对业务问题进行分析、评估并提供辅助支持的能力。评测应重点考察模型在各业务环节中对业务逻辑的理解程度、分析深度以及对实际业务决策的辅助效果。

6.5.1 需求计划与预测

需求计划与预测能力的评测，应考察模型对需求预测方法、业务数据和预测结果应用场景的理解能力，具体包括以下方面：

a) 对历史需求数据、市场信息和业务背景进行分析的能力；
b) 对需求变化趋势进行预测并识别异常波动的能力；
c) 对预测结果不确定性和偏差来源进行解释的能力；
d) 对需求预测结果在生产、采购和库存等后续决策中的辅助支持能力。

6.5.2 采购管理

采购管理能力的评测，应考察模型对采购策略、寻源方式和供应商管理的分析能力，具体包括以下方面：

a) 对采购战略和采购计划匹配关系的理解能力；
b) 对供应商评价、选择及绩效表现进行分析的能力；
c) 对采购成本、交期及相关风险因素进行综合分析的能力；

d) 对采购决策提供多方案辅助建议的能力。

6.5.3 生产与制造

生产与制造能力的评测，应考察模型对生产计划和供应链协同关系的理解能力，具体包括以下方面：

a) 对主生产计划与物料需求计划之间关系的理解能力；
b) 对产能规划、产需平衡及约束条件的分析能力；
c) 对生产物流与库存协同关系的辅助分析能力。

6.5.4 库存与仓储

库存与仓储能力的评测，应考察模型对库存策略和仓储运营支持能力，具体包括以下方面：

a) 对库存策略与库存分类方法的理解能力；
b) 对仓储布局与设施配置问题的分析能力；
c) 对仓储作业流程及运行效率的分析能力；
d) 对库存风险（积压、短缺等）的识别与提示能力。

6.5.5 物流配送

物流配送能力的评测，应考察模型对配送组织和履约执行支持能力，具体包括以下方面：

a) 对运输模式选择和配送网络设计的分析能力；
b) 对配送路径优化和网络协同问题的分析能力；
c) 对末端配送和“最后一公里”问题的识别能力；
d) 对物流配送方案优化的辅助支持能力。

6.5.6 国际物流与跨境

国际物流与跨境能力的评测，应考察模型对跨境物流流程和规则环境的理解能力，具体包括以下方面：

a) 对跨境运输、通关流程和单证体系的理解能力；
b) 对国际贸易规则和跨境约束条件的分析能力；
c) 对跨境物流风险因素的识别与提示能力；
d) 对国际物流方案优化的辅助支持能力。

6.5.7 国内运输管理

国内运输管理能力的评测，应考察模型对运输组织和运力管理的分析能力，具体包括以下方面：

a) 对运输网络结构和网络优化问题的分析能力；
b) 对承运商选择和运输资源配置的分析能力；
c) 对运输成本控制和运行效率提升问题的分析能力。

6.5.8 销售与售后服务

销售与售后服务能力的评测，应考察模型对订单履约和客户服务支持能力，具体包括以下方面：

a) 对客户服务流程和服务能力的分析能力；
b) 对异常订单、退换货及召回管理问题的识别能力；
c) 对循环利用和再利用相关业务的理解能力；
d) 对销售与售后服务改进的辅助支持能力。

6.5.9 供应链金融

供应链金融能力的评测，应考察模型对资金流和金融风险分析支持能力，具体包括以下方面：

a) 对贸易融资模式和结算方式的理解能力；
b) 对存货质押融资和跨境供应链金融业务的分析能力；
c) 对资金占用和现金流风险的识别能力；
d) 对供应链金融方案的辅助支持能力。

6.5.10 风险管控

风险管控能力的评测，应考察模型对业务风险识别和管控支持能力，具体包括以下方面：

a) 对经营风险、运输风险和金融风险的识别能力；
b) 对风险影响进行分析和分级的能力；
c) 对风险应对措施提供辅助建议的能力；
d) 对风险决策过程提供约束和提示的能力。

6.5.11 供应链韧性

供应链韧性能力的评测，应考察模型对供应链稳定性和恢复能力的支持程度，具体包括以下方面：

a) 对供应链脆弱环节的识别能力；
b) 对多源采购和替代方案的分析能力；
c) 对需求波动和中断情景的响应分析能力；
d) 对提升供应链韧性的辅助决策能力。

6.5.12 数字供应链

数字供应链能力的评测，应考察模型对数字技术与供应链融合应用的支持能力，具体包括以下方面：

a) 对AI在供应链中的应用场景理解能力；
b) 对区块链溯源机制的认知与分析能力；
c) 对物联网在物流中的应用理解能力；
d) 对数字供应链建设方案的辅助支持能力。

6.5.13 财务优化

财务优化能力的评测，应考察模型对财务结构和资金效率分析支持能力，具体包括以下方面：

a) 对现金流优化问题的分析能力；
b) 对库存融资和成本结构的理解能力；
c) 对成本分析和财务风险的识别能力；
d) 对财务优化决策的辅助支持能力。

6.5.14 全球供应链管理

全球供应链管理能力的评测，应考察模型对跨区域、跨主体供应链协同管理的支持能力，具体包括以下方面：

a) 对全球供应链结构和运行特征的理解能力；
b) 对区域差异和地缘因素影响的分析能力；
c) 对全球供应链布局调整方案的辅助分析能力；
d) 对全球供应链管理决策的支持能力。

7 评测方法

7.1 评测数据集

评测数据集应满足以下要求：

合规性和密级要求：数据收集过程应遵循适用的法律法规、跨境数据合规及行业相关保密要求，并保护个人信息与企业商业秘密。国际物流供应链业务数据通常涉及企业经营信息、贸易单证、客户信息、运力与节点运行数据等，具有较强的敏感性，应建立分级分类、脱敏处理、最小必要使用与访问控制机制，确保数据全生命周期安全可控。不同敏感等级的数据应遵循差异化的处理、存储、使用和共享规范，并设置相应的审计机制，确保评测过程可追溯、可核查。

评测指标完备：为每个评测指标构建满足相应数量的数据集。评测问答数据集应包括单选题、多选题、判断题、材料分析题四种问题类型。

时效性：数据集结合开源数据集和自建数据集,定期更新维护。数据集应建立更新、维护和质量评估机制，确保数据长期可用，反映最新的国际物流供应链应用场景和行业需求。

可用性：数据集格式和接口符合广泛的标准,以便于获取和使用。资源数据应以规定文件格式之一的形式存在，不符合的需采取措施进行格式转换。

多样性和代表性：涵盖国际物流供应链及相关行业的不同专业知识、业务场景等,以确保数据能覆盖不同的使用情况。

7.2 评测环境

7.2.1 软硬件环境搭建

应根据待测模型的实际性能要求，搭建配套的软硬件平台，包括通用计算芯片、AI计算加速芯片、计算服务器、存储服务器、通信网络、云服务、容器/虚拟化等。

7.2.2 部署方式

测试框架可部署在单一服务器上进行少样本测试，也可部署在集群中进行大数据量测试。应将评测环境部署在指定的测试环境中。

7.2.3 算力配置

在模型微调或评测阶段，应综合考虑模型参数大小、训练数据规模、预期训练时长等多方因素进行适当的算力配置。

7.3 评测工具

7.3.1 自动化评测功能

应集成全面的测试集，覆盖国际物流供应链大模型专业知识和业务能力的各个维度。

应支持灵活扩展功能，根据需求及时更新扩展评测数据集。

应具备确定明确的评价指标计算方法和评分规则，并根据业务需求，对评价指标体系进行迭代和更新。

应能够自动生成并输出评测结果，提高评测效率。

7.3.2 人工评测功能

应为评测人员提供相应的工具链平台，可辅助评测人员校核自动化评测结果，并可支持评测人员对模型的回答进行人工打分。

应能分析评测结果的分布和一致性，及时发现评测人员潜在的评测偏差或不一致问题。

7.4 评测实施

7.4.1 自动化评测

在评测数据集中应构建出相应的参考答案。

在自动化评测脚本中应清晰定义具体的评测指标计算方法和评分规则。

7.4.2 人工评测

应制定清晰、具体的评测标准和指南,并对评测人员进行充分的培训,确保所有评测人员对评测的标准有统一的理解和执行。

应选择具有国际物流供应链专领域知识和经验的评测人员,以确保评测结果的准确性和专业性。

宜对评测人员定期进行复训,更新评测知识和技能,尤其是当标准内容有调整时。

宜定期收集评测人员的反馈,用于优化评测流程和评测标准。

7.5 评测结果评估

7.5.1 总体要求

评测结果的评估应建立一套科学、量化且可操作的指标体系。

数据质量评估应重点关注数据的准确性、一致性、完整性和及时性。

模型有效性评估：模型应能理解业务需求，针对国际物流供应链领域典型的应用场景，生成可执行的策略。

应建立反馈循环机制，以迭代优化语料和模型质量，强调语料和大模型服务于实战场景。

7.5.2 评分与等级划分

评分规则宜与评测目标、数据集类型（客观题/主观题）、评测方式（自动/人工）相匹配，并保持可复现、可解释、可审计。

综合得分宜由各评测维度得分按权重汇总得到，权重设置应体现国际物流供应链应用的关键关注点，并可根据应用场景进行调整。

评测等级划分宜基于综合得分区间或分位阈值进行设置，并结合实际应用门槛、风险容忍度与对比基线进行校准。

评分细则及等级划分示例可参见附录A。

附 录 A
（资料性）
评分规则与评测等级设置示例

A.1 模型评分规则示例（100分制）

能力满分为100分，其中单选题占40分、多选题占30分、判断题占20分、主观题占10分，题目以随机的方式从评测数据集中抽取；

模型的综合得分满分为100分，由每个单项以加权平均的方式得到模型的最终综合得分，其中各单项能力权重如下表：

表A.1 模型测评单项能力权重表

能力测试维度	权重
模型通用基础能力	10%
国际物流供应链安全与价值对齐	10%
国际物流供应链风险控制	20%
国际物流供应链专业认知能力	40%
国际物流供应链业务辅助拓展能力	20%

A.2 评测等级设置示例

——A级：综合得分区间在[80,100]；
——B级：综合得分区间在[60,80)；
——C级：综合得分区间在[50,60)；
——D级：综合得分区间在[0,50]；

参 考 文 献

[1] GB/T 25069-2010 信息安全技术 术语
[2] GB/T 41867-2022 信息技术 人工智能 术语
[3] GB/T 5271.1-2000 信息技术 词汇 第1部分：基本术语
[4] 中国信息通信研究院,《大模型基准测试体系研究报告(2024年)》, http://www.caict.ac.cn/