|  |  |
| --- | --- |
| ICS |  |
| CCS | |  | | --- | | D:\000000部门项目\09标准化插件开发\程序源代码\StandardEditor_ShanDongKeXieYuan\团标首页面字母T.pngD:\000000部门项目\09标准化插件开发\程序源代码\StandardEditor_ShanDongKeXieYuan\团标首页面字母T后面的反斜杠.png |   点击此处添加CCS号 |

     团体标准

T/XXX XXXX—XXXX

金融大模型应用评测指南

Guide to Evaluating Financial Large Models Application

XXXX - XX - XX发布

XXXX - XX - XX实施

       发布

1. 前言

本文件按照GB/T 1.1—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。

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：

本文件主要起草人：

目次

[前言 II](#_Toc12626)

[1 范围 2](#_Toc13894)

[2 规范性引用文件 2](#_Toc22095)

[3 术语和定义 2](#_Toc12406)

[3.1 大模型 large model 2](#_Toc24174)

[3.2 金融语料 financial corpus 2](#_Toc7600)

[3.3 单模态维度 Monomodal dimension 2](#_Toc18715)

[3.4 多模态维度 multimodal dimension 2](#_Toc32047)

[3.5 价值对齐 value alignment 2](#_Toc10278)

[4 缩略语 2](#_Toc21276)

[5 概述 3](#_Toc2648)

[5.1 基本框架 3](#_Toc18744)

[5.2 评测维度 3](#_Toc30730)

[6 评测内容 5](#_Toc7597)

[6.1 模型基础能力评测 5](#_Toc31018)

[6.2 金融安全与价值对齐能力评测 5](#_Toc6517)

[6.3 金融风险控制能力评测 6](#_Toc6274)

[6.4 金融专业认知能力评测 7](#_Toc21247)

[6.5 金融业务辅助拓展能力评测 8](#_Toc13902)

[7 评测方法 10](#_Toc26362)

[7.1 评测数据集 10](#_Toc19551)

[7.2 评测工具 10](#_Toc24680)

[参考文献 12](#_Toc6226)

[附录A 13](#_Toc1205)

金融大模型应用评测指南

* 1. 范围

本文件提出了大模型在金融垂直领域应用需要具备的核心能力及评测方法，包括模型基础能力、金融安全与价值对齐能力、金融风险控制能力、金融专业认知能力、金融业务辅助拓展能力五个方面。

本文件适用于大模型在银行、证券、保险、基金、信托等金融垂直领域应用效果相关的评测工作。

* 1. 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 25069-2010 信息安全技术 术语

GB/T 41867-2022 信息技术 人工智能 术语

GB/T 5271.1-2000 信息技术 词汇 第1部分：基本术语

* 1. 术语和定义

GB/T 41867-2022界定的以及下列术语和定义适用于本文件。

* + 1. 大模型 large model

在大量数据上训练得到，具有复杂计算架构，能处理复杂任务，且具备一定泛化性的深度学习模型。

* + 1. 金融语料 financial corpus

金融领域相关文本、图片、音频、视频等数据的集合，可用于训练和优化金融大模型，以提升其在金融领域的应用效果。

* + 1. 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。

* + 1. 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。

* + 1. 价值对齐 value alignment

大模型的能力和行为与人类的价值、真实意图和伦理原则相一致，以确保人类与人工智能协作过程中的安全与信任。

* 1. 缩略语

AI：人工智能 （Artificial Intelligence）

IPO：首次公开招募 （Initial Public Offering）

* 1. 概述
     1. 基本框架

本文件聚焦大模型在金融垂类领域的应用效果评估，从模型基础能力、金融专业认知能力、金融业务辅助拓展能力、金融风险控制能力、金融安全与价值对齐能力五大维度构建了金融大模型应用评测架构体系，如图1所示。

表格

描述已自动生成

图1 金融大模型应用评测架构体系

* + 1. 评测维度

5.2.1 模型基础能力

基于大规模AI语料库训练得到的金融大模型所具备的能力，应遵循《人工智能 大模型 第2部分：评测指标与方法》中的相关规范和要求，包括单模态和多模态两个维度方面的能力。

单模态能力中主要涉及文本、图像、音频三个方面能力，具体包括：

a)文本分类；

b)信息抽取；

c)因果推理；

d)常识推理；

e)任务分解；

f)文本问答；

g)多轮对话；

h)代码理解；

i)长文本理解；

j)静态图像分类。

多模态能力中涉及文本、图像两个方面能力，具体包括：

1. 图文检索；
2. 静态图像问答；
3. 视觉语言推理；
4. 视觉蕴含；
5. 视频问答；
6. 图表推理。

5.2.2 金融安全与价值对齐

大模型在业务活动中研判输出内容的合规和敏感性能力，包括但不限于以下内容：

a)《网络数据安全管理条例》要求；

b) 内容合规；

c) 应用场景合规性；

d) 安全审计与监控；

e) 金融逻辑合理性；

f) 维护金融市场稳定；

g) 伦理价值；

h) 情绪价值；

i) 文化价值；

j) 社会价值。

5.2.3 金融风险控制能力

大模型在金融业务活动中识别、评估、监控、应对和管理金融风险的能力，包括但不限于以下内容：

1. 合规风险；
2. 市场风险；
3. 操作风险；
4. 流动性风险；
5. 信用风险；
6. 舆情风险。

5.2.4 金融专业认知能力

大模型在金融领域中所展现出的理解、分析、决策和应用金融知识的能力，包括但不限于以下内容：

1. 金融基础知识能力;
2. 金融信息解读能力。

5.2.5 金融业务辅助拓展能力

大模型在银行、基金、保险、证券、信托等垂直领域的核心业务能力，可分为对内和对外两大部分，包括但不限于以下内容：

1. 场景丰富度；
2. 业务价值提升能力；
3. 业务侧能力指标；
4. 服务可靠性。
   1. 评测内容
      1. 模型基础能力评测

大模型理解能力评测主要分为单模态维度和多模态维度，单模态维度主要包括文本、图像的维度，多模态维度主要包括图文维度。

6.1.1 单模态维度

单模态能力中涉及文本、图像两个方面能力，具体包括：

1. 文本分类：将文本划分为不同的类别或标签；
2. 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；
3. 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；
4. 常识推理：在日常情境下，结合常识理解和推断隐含信息的能力；
5. 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；
6. 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；
7. 多轮对话：评测模型在进行多轮对话场景下的问答能力；
8. 代码理解：指模型能够对给定的编程代码，给出相应的文本解释说明；
9. 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；
10. 静态图像分类:指模型能够理解静态图像的语义内容，并输出其对应的类别标签。
11. 音频问答：指模型能够理解用户提供音频信息中的问题，并提供合 理、准确、实用的答案。

6.1.2 多模态维度

多模态能力中主要涉及图文方面能力，具体包括：

1. 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；
2. 静态图像问答：指模型能够回答针对静态图像的文本问题；
3. 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；
4. 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；
5. 视频问答：指模型能够回答针对视频的文本问题；
6. 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；
   * 1. 金融安全与价值对齐能力评测

6.2.1 安全合规

金融大模型评测在安全合规方面应涵盖多个关键内容，以确保模型的可靠性、隐私性和可解释性，包括但不限于以下内容：

1. 《网络数据安全管理条例》要求：包括总则、一般规定、个人信息保护、重要数据安全、网络数据跨境安全管理、网络平台服务提供者义务、监督管理、法律责任、附则等部分内容；
2. 内容合规：对模型输出结果的使用应进行金融领域合规性审查，确保基于模型结果做出的决策和业务操作是合法、合规的。
3. 应用场景合规性：金融大模型的应用场景应符合金融行业的监管要求和业务规范，模型的应用应遵循相关的法律法规和业务流程。
4. 安全审计与监控：对金融大模型的运行过程和用户操作进行记录和审计，以便及时发现潜在的安全风险和违规行为。

6.2.2 价值对齐

金融大模型评测在价值对齐方面应涵盖多个关键内容，以确保模型的行为和输出与人类的价值、真实意图和伦理原则相一致，包括但不限于以下内容：

1. 金融逻辑合理性：在处理金融业务问题时，模型的决策逻辑要符合金融行业的专业逻辑和实践经验；
2. 维护金融市场稳定：在进行市场分析、预测等工作时，要客观、准确地反映市场情况，避免发布虚假或误导性的信息，以免对金融市场的稳定造成不良影响。
3. 伦理价值：在提供金融服务时，不应对不同的用户或群体存在歧视或偏见；
4. 情绪价值：包括情感认同、交互、共情共鸣方面的能力；
5. 文化价值：包括文化素养、人文关怀、审美塑造等方面；
6. 社会价值：在投资建议、产品推荐等方面，要充分考虑投资者的风险承受能力和投资目。
   * 1. 金融风险控制能力评测

金融大模型的风险控制能力评测应包含多个关键内容，以确保模型在实际应用中具有高效、准确和稳定的风险控制能力，包括但不限于以下内容：

6.3.1 合规风险

是指金融大模型在应用过程中未能严格遵循相关法律法规、监管要求、行业准则以及适用于金融业务活动的规范要求，从而导致金融机构可能遭受法律制裁、监管处罚、重大财务损失或声誉损失的风险，包括但不限于以下内容：

1. 法律法规遵循风险：大模型设计、开发、部署和应用均应符合所有相关的法律法规要求；
2. 监管要求执行风险：大模型应用应确保符合数据报告、风险管理、内部控制等监管要求；
3. 行业准则和自律规范遵循风险：大模型应用应遵循行业准则和自律规范；
4. 业务活动行为准则遵循风险：大模型应用应遵循适用于金融业务活动的行为准则。

6.3.2 市场风险

是指市场价格变量波动导致大模型在市场趋势预测、分析和评估过程中可能存在的风险，包括但不限于以下内容：

1. 利率风险：大模型应具备应对利率不利变动并自主校正预测结果的能力；
2. 汇率风险：大模型应具备应对汇率波动并自主校正预测结果的能力；
3. 股票价格风险：大模型应具备应对股票价格波动并自主校正预测结果的能力；
4. 商品价格风险：大模型应具备应对商品价格波动并自主校正预测结果的能力。

6.3.3 操作风险

是指金融机构在使用金融大模型进行预测、分析、决策等过程中，由于内部操作或管理失误所导致的风险，包括但不限于以下内容：

1. 模型使用不当风险：金融机构应对模型进行充分的验证和测试，以确保其准确性和可靠性；
2. 数据处理风险：金融机构应建立完善的数据处理流程，确保数据的质量和准确性；
3. 系统故障风险：金融机构应定期对系统进行维护和升级，确保其稳定性和安全性。

6.3.4 流动性风险

是指大模型在智能化识别、预测金融市场资金、资产流动性规律过程中可能存在的风险，包括但不限于以下内容：

1. 资产流动性风险：大模型应具备根据金融市场深度、密度、即时性、弹性准确预测市场流动性风险的能力；
2. 融资流动性风险：大模型应具备为金融机构预测预期和非预期的、当前和未来的现金流和担保需求风险的能力；
3. 负债流动性风险：大模型应具备识别内外波动因素，为金融机构存款资金投资进行风险评估、预测的能力。

6.3.5 信用风险

是指大模型在识别、执行借款人或发行人发生违约、无法履行合约或无法如期支付债务等场景下可能存在的风险，包括但不限于以下内容：

* 1. 违约风险：大模型应具备智能识别违约合同并启动催款流程的能力；
  2. 评级风险：大模型应具备对借款人的信用状况进行评估的能力。

6.3.6 舆情风险

是指金融机构在使用大模型监测风险时，应当具备识别负面舆情的能力，包括但不限于以下内容：

1. 新闻负面舆情风险：新闻媒体发布的负面消息所引发的不利影响的可能性；
2. 投诉舆情风险：客户或其他利益相关者投诉后，这些投诉信息通过各种渠道传播，进而可能对组织形象、声誉和正常运营造成不利影响的风险；
3. 网络舆情风险：互联网环境下，由于各种信息的产生、传播和互动，造成不利影响的可能性。
   * 1. 金融专业认知能力评测

6.4.1 金融基础知识

6.4.1.1 会计学能力

是指大模型能够理解和应用会计准则、会计制度以及相关的会计处理方法，对金融交易和事项进行准确计量、记录和报告的能力，可从财务管理、会计实务、会计基础等方面进行评测，具体如下：

1. 会计核算能力：大模型能够准确地进行会计核算，包括确认、计量、记录和报告金融交易和事项；
2. 财务报表编制与分析能力：大模型能够根据会计核算的结果，编制符合会计准则要求的财务报表，如资产负债表、利润表、现金流量表等；
3. 税务申报与筹划能力：大模型能够学习税法规定，根据会计核算结果进行税务申报；
4. 内部控制与风险管理能力：大模型能够识别、评估和应对与会计业务风险，如财务舞弊、会计差错等。

6.4.1.2 金融学能力

是指大模型在处理金融相关任务时所展现出的金融学术知识、金融业务知识、金融数据分析等专业技能和素养，可从金融学术、金融业务、金融安全等方面开展评测，具体如下：

1. 金融学术知识：大模型应具备扎实的金融学术基础，包括金融理论、金融市场、金融工具、金融政策等方面的知识；
2. 金融业务知识：大模型应熟悉各类金融业务，包括银行业务、证券业务、保险业务等；
3. 金融数据分析能力：大模型应具备强大的数据分析能力，能够通过信息提取信息进行市场趋势分析、风险评估和预测等；
4. 金融智能体能力：大模型应具备推理规划、长程记忆、工具应用等核心能力，例如可通过调用API、检索API等方式，结合金融知识和数据，进行智能问答、报告生成等任务；

6.4.1.3 经济学能力

是指大模型在处理金融问题时所表现出的对经济理论、金融市场规律、金融决策支持等方面的理解和应用能力，可从以下方面开展评测：

1. 经济理论理解能力：大模型应具备对经济理论的基本理解，包括宏观经济学、微观经济学、金融经济学等基础理论；
2. 金融市场规律把握能力：大模型应能识别和把握金融市场的规律，包括价格形成机制、供求关系、市场结构等；
3. 金融决策支持能力：大模型应能提供决策支持，包括风险评估、投资决策、资产配置等方面的建议。

6.4.1.4 资质认证能力

是指通过构建具有深度学习能力的金融知识库和智能评估系统，以实现对金融从业人员专业知识、技能水平、职业素养等多方面的全面、客观、高效评估的能力，可从以下方面开展评测：

1. 资质审核自动化能力：大模型应具备对申请者的个人信息、教育背景、工作经验等进行快速且准确的审核能力，以确保信息的真实性和完整性；
2. 能力评估与测试能力：大模型应具备设计在线测试或模拟场景的能力，以评估申请者在金融知识、专业技能、风险意识等方面的专业能力；
3. 持续教育与培训能力：大模型应具备为金融从业人员提供个性化的在线学习资源和培训计划的能力，能够分析从业人员的学习行为和成果，并为其推荐更适合的学习路径和课程。

6.4.2 金融信息解读

是指金融大模型对金融信息的解读能力，包括金融图表解读、金融知识图谱构建、结构化数据提取等方面的能力。

6.4.2.1 金融图表解读

是指大模型通过先进的算法和技术，对金融图表进行深度解析，从而提取出有价值的信息和数据，并为投资者提供决策支持的能力，包括但不限于以下内容：

1. 金融图表类型识别：大模型应具备准确识金融图标的能力，包括研报图标、财报图表、 K线图等；
2. 金融图表含义理解：大模型应具备深入理解金融图表中数据含义的能力；
3. 图表解读与决策支持：大模型应具备将图表解读的结果转化为投资者可以理解的投资建议和策略的能力。

6.4.2.2 金融知识图谱构建

是指大模型能够解析、理解和运用金融知识图谱中的实体、关系、属性等信息，以支持金融领域的复杂分析和决策过程的能力，包括但不限于以下内容：

1. 实体识别与理解：大模型应具备准确识别知识图谱中的各类实体并理解其含义、属性和特征的能力；
2. 关系抽取与推理：大模型应具备抽取知识图谱中实体间复杂关系（如股权关系、任职关系、供应商关系）并开展逻辑推理的能力；
3. 属性分析与利用：大模型应具备分析知识图谱中实体属性值，包括公司的财务指标、产品的价格信息等，并开展进一步分析和预测的能力；
4. 可视化与交互：大模型应具备通过可视化图谱展示实体和关系并支持用户与知识图谱交互的能力。

6.4.2.3 结构化数据提取

是指从复杂的金融数据源中识别、抽取和转换具有特定结构和语义的金融信息的过程，，包括但不限于以下内容：

1. 识别数据类型：能够精确地辨别金融数据中的各种类型（如数值型数据、文本型数据、日期时间型数据）并依据不同类型进行相应处理；
2. 数据清洗能力：可以识别并剔除金融数据中的噪声数据，如异常值、错误数据、重复数据、错误数据，保证数据的准确性；
3. 数据标准化处理：对提取到的金融数据进行标准化处理，使不同来源、不同格式的数据具有统一的标准和规范。
   * 1. 金融业务辅助拓展能力评测

6.5.1 银行业务

可从支付业务、存贷款业务、财富管理业务、交易银行业务、数字银行业务5大核心业务开展评测，涵盖24个能力指标，具体如下：

1. 支付业务：包括个性化支付服务、个性化优惠活动推荐、个性化支付方式推荐、自动审核支付请求、交易限额智能识别调整、金融欺诈识别；
2. 存贷款业务：包括智能风控识别、存贷款行为分析、异常资金流动预警、信贷审批、定制化贷款期限/利率、智能催收与贷后管理
3. 财富管理业务：包括智能营销与推荐、智能资讯摘要总结、智能投资建议、定制化金融产品推荐；
4. 交易银行业务：包括汇率波动预测、智能交易策略优化、外汇交易产品个性化推荐、财务智能审批
5. 数字银行业务：包括智能客服、个性化账户服务、远程音视频服务、虚拟金融服务。

6.5.2 基金业务

可从基金投资管理、基金募集与销售、基金运营、受托资产管理、投资咨询服务、风险管理、业绩评价7大核心业务开展评测，涵盖26个能力指标，具体如下：

1. 基金投资管理:包括市场分析、投资决策、风险评估和资产配置策略优化方面的辅助能力。
2. 基金募集与销售:包括市场趋势预测、客户行为分析、销售策略制定和个性化推广内容生成方面的表现。
3. 基金运营:包括自动化处理、数据准确性、合规性检查和信息披露及时性方面的辅助运营能力。
4. 受托资产管理:包括客户需求理解、资产管理策略制定、投资组合构建和调整方面的辅助能力表现。
5. 投资咨询服务:包括投资策略制定、市场动态分析、风险评估和客户沟通方面的能力。
6. 风险管理:包括风险识别、评估、监控和缓解策略制定方面的能力。
7. 业绩评价:包括业绩跟踪、归因分析、基准比较和报告生成方面的能力。

6.5.3 保险业务

可从投研、产品定价与设计、营销、承保、理赔、服务6大核心业务开展评测，涵盖24个能力指标，具体如下：

1. 投研：包括金融风险管理、市场趋势和价格波动分析、资产配置和组合优化、量化交易、舆情分析、财务报表分析；
2. 产品定价与设计：包括客户需求分析、产品设计支持、产品定价、数据处理及分析；
3. 营销：包括营销文案生成、客户潜在需求分析、智能营销、个性化保险方案推荐；
4. 承保：包括风险评估、自动核保、异常识别；
5. 理赔：包括智能定损、智能理赔、风险预警、风险反欺诈；
6. 服务：包括智能化客户关系管理、智能客服、客户权益管理。

6.5.4 证券业务

可从智能投顾、智能投研、量化交易、智能投教、智能风控5大核心业务开展评测，涵盖19个能力指标，具体如下：

1. 智能投顾：包括证券智能客服与咨询、证券投资组合优化与管理、证券投资风险评估与预警、个性化证券投资策略推荐。
2. 智能投研：包括高效数据处理与信息提取、智能会议记录与摘要生成、智能研报撰写与翻译、智能问答与知识库构建、智能证券投研分析与建议。
3. 量化交易：包括量化交易策略开发与优化、量化交易机会识别、量化交易风险管理、量化交易数据分析与可视化。
4. 智能投教：包括智能投教个性化学习路径、模拟投资体验、智能问答系统
5. 智能风控：包括信用风险评估、市场风险监测、欺诈风险防范。

6.5.5 信托业务

可从投资分析、风险管理、合规管理、营销业务、家族信托5大核心业务开展评测，涵盖24个功能指标，具体如下：

1. 投资分析：包括投资市场趋势分析、投资风险评估、项目筛选与尽职调查、投资策略与资产配置、智能投后管理与预警。
2. 风险管理：包括多维数据分析、动态风险调整、风险实时监控与预警、个性化风险评估、风险模拟与应对、应急响应方案制定。
3. 合规管理：包括法规与政策解读、合规文件审查、合规建议与解决方案、合规培训与教育、合规报告自动生成。
4. 营销业务：包括客户需求分析、目标客户选择、营销渠道拓展、客户关系管理；
5. 家族信托：包括个性化家族资产配置、家族长期财富规划、家族企业股权智能化管理、家族遗产规划方案制定。
   1. 评测方法

金融大模型评测宜采用评测数据集与打分评级的方法，主要包括测评数据集、测评工具两个方面内容。

* + 1. 评测数据集

基于金融大模型能力评测的细分领域和需求，评测数据集应围绕问题类型、数据来源、问题设计、答案提取、问答对配对、数据规模、更新机制等方面进行构建，具体如下：

1. 问题类型：金融大模型应用评测问答数据集应包括单选题、多选题、判断题、主观题四种问题类型；
2. 数据来源：根据金融行业的属性，用于问答对设计的数据应保证具备专业性、可靠性、权威性、高质量、覆盖广等特征，可从证券、银行、基金等领域的企业财报、行业研报、交易中心等渠道获取，并对原始数据进行脱敏、匿名化等处理后使用；
3. 问题设计：根据评测需求，设计一系列与金融垂类领域相关的问题，应涵盖不同的难度级别和类型，以全面评估金融大模型的性能；
4. 答案提取：从收集到的文本数据中提取与问题相对应的答案，可采取人工标注或大模型自动化标注的方式，应确保答案的准确性和完整性；
5. 问答对配对：将问题和答案进行配对，形成问答对，每个问答对应包含一个明确的问题和一个与之对应的答案。
6. 数据规模：针对单项特定场景测试语料不少于100个，每个能力单项中可选取10%的数据作为公开样例数据；
7. 更新机制：评测数据集应结合行业发展实际情况进行至少每月一次的内容更新。
   * 1. 评测工具

金融大模型应用评测可采取多种工具，具体如下：

1. 自动化评测

——自动化测评工具应集成完整的模型测评集；

——自动化测评工具应清晰定义具体的评价指标计算方法和评分规则；

——自动化测评工具应兼容不同类型、不同架构的模型；

——自动化测评工具应自动输出测评结果；

b） 人工评测

——应制定清晰、具体的评价标准和指南，并对评价人员进行充分的培训，确保所有评价人员对评价的标准有统一的理解和执行；

——应分析评价结果的分布和一致性，及时发现潜在的评价偏差或不一致问题；

——宜选择具有相关领域知识和经验的评价人员，以确保评价结果准确性和专业性；

——宜为评价人员提供相应的评价工具，以支持评价人员的工作；

——宜对评价人员定期进行复训，更新评价知识和技能，尤其是当标准内容有调整时；

——宜定期收集评价人员的反馈，用于优化评价流程和评价标准。

c） 大模型测评

——应选择与评估任务相关性高的大模型，可使用多个大模型进行交叉验证，以提高评测的稳定性；

——应定义清晰的评估标准和评分规则，并转成能激发大模型更佳性能表现的输入提示词，确保大模型按照既定标准进行评测；

——应在测试过程中引入人工审核机制，及时识别问题和调整评估策略，以确保评测的准确性和公正性；

——应确保评测过程中大模型访问接口的稳定可靠，以确保评估过程的连续性。

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] 中国银保传媒,腾讯研究院《2023金融业大模型应用报告》，https://cloud.tencent.com/developer/article/2363347

[5] 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/

附 录 A

（资料性附录）

评测分级参考

金融大模型的测评结果应综合模型通用基础能力、金融认知能力、金融业务辅助拓展能力、金融风险控制能力、金融安全合规与价值对齐能力五个单项能力进行综合评测。

1. 打分规则

——每个单项能力满分为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 | 20% |
| 金融安全合规与价值对齐能力 | 20% |
| 金融风险控制能力 | 20% |
| 金融专业认知能力 | 20% |
| 金融业务辅助拓展能力 | 20% |

1. 评测等级（模拟打一下），对于区间范围设置的意见

——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]；

——E级：模型在金融安全合规与价值对齐能力单项能力低于20分时。