

2026 年上海市高等学校信息技术水平考试试卷

四级 人工智能 (A 场)

(本试卷考试时间 150 分钟)

一、单选题 (本大题 15 道小题 , 每小题 1 分 , 共 15 分), 从下面题目给出的 A、B、C、D 四个可供选择的答案中选择一个正确答案。

1. _____是语音处理中常用的声学特征。

- A. 像素值
- B. 梅尔频率倒谱系数
- C. 词向量
- D. 句法树

2. 某团队在评估一款大模型在医疗长文档数据检索中的表现时,发现当输入文档长度约为 128K token 时,无论文档多长,当关键病例信息位于文档开头或结尾时,模型提取信息的准确率较高,但当关键信息位于文档中部时,准确率明显下降。这一现象最有可能说明_____。

- A. 模型的上下文窗口存在截断机制,中部信息被自动丢弃
- B. 模型的注意力机制对长文本中部的信息处理能力较弱
- C. 模型对文档中部的内容存在严重偏见,拒绝回答
- D. 文档中部的编码质量通常低于开头和结尾

3. 某在线教育公司上线了一款智能答疑助手。最初,它只能被动地等待用户提问后给出回答;后来,公司按照“判断问题类型→检索知识库→生成回答”的固定步骤自动执行答疑流程;最近,公司给答疑助手设定了“提升用户课程完成率”的目标,助手会自主规划先推送学习提醒、再生成个性化练习并跟踪完成情况。该应用依次经历的形态演进是_____。

- A. 工作流→聊天机器人→智能体
- B. 聊天机器人→工作流→智能体
- C. 聊天机器人→智能体→工作流
- D. 智能体→工作流→聊天机器人

4. 某钢铁企业利用“CIS Digital AI 金睛”视觉大模型对高炉火焰进行实时监测, AI 算法能全天候精准识别喷溅、溢渣等情况,准确率高达 99%以上。这一应用主要属于计算机视觉的_____任务。

- A. 图像生成
- B. 目标检测与识别
- C. 三维重建
- D. 风格迁移

5. 多跳检索主要用于解决需要跨_____查询的复杂问题。

- A. 单个文档
- B. 多个文档
- C. 单个句子

D. 单个词语

6. _____推理引擎以 Continuous Batching 和 PagedAttention 为核心优化技术。

- A. TensorFlow Serving
- B. vLLM
- C. ONNX Runtime
- D. PyTorch Distributed

7. 大模型推理服务采用流式输出 (Streaming) 可以_____。

- A. 减少模型参数量
- B. 降低首 token 延迟, 实现逐 token 响应
- C. 增加训练数据吞吐量
- D. 提高模型推理精度

8. 机器学习中, _____算法属于有监督学习。

- A. K-Means 聚类
- B. 线性回归
- C. PCA 降维
- D. 关联规则挖掘

9. 某金融风控团队曾构建基于规则的专家系统来审批贷款, 但规则一旦增多, 新增规则可能与旧规则发生冲突, 导致审批结果不稳定。通过引入机器学习模型, 利用历史借贷数据训练分类器, _____是其核心优势。

- A. 模型可以直接读取人类专家的思维过程
- B. 模型通过数据学习特征, 减少了人工编码知识的成本与冲突问题
- C. 模型完全不需要人工干预
- D. 模型保证零错误率

10. 在自然语言处理的预训练阶段, 工程师利用训练好的 Word2Vec 模型进行词语类比推理测试, 发现“国王-男人+女人”的结果最接近“女王”, 而“柏林-德国+意大利”的结果最接近“罗马”。这一现象主要说明 Word2Vec 生成的稠密词向量具备_____的能力。

- A. 对语义关系和语法关系进行向量运算
- B. 根据动态上下文实时改变词义
- C. 精准识别所有训练语料中的命名实体
- D. 彻底消除多义词在不同语境下的歧义

11. 某自动驾驶团队在采集道路图像数据时, 发现数据集中晴天白天的样本占 80%, 而雨天、夜间等场景样本很少。为了提升模型在恶劣天气下的鲁棒性, _____的数据管理策略最合理。

- A. 删除所有晴天白天的样本, 只保留恶劣天气样本
- B. 在训练时对雨天和夜间样本进行数据增强, 并补充相关采集
- C. 直接使用原始数据训练, 不进行任何处理
- D. 将晴天白天样本复制多份, 以增加总数

12. _____不是数据标注的常见类型。

- A. 图像分类
- B. 实体识别
- C. 语音转写
- D. 数据清洗

13. 某银行客服机器人在回答客户关于理财产品的问题时，需先从内部知识库中检索相关资料，再基于检索结果生成回答。这种将内部知识库与生成模型相结合的技术是_____。

- A. RAG
- B. Function Calling
- C. ReAct
- D. Skills

14. 某在线教育平台通过大模型动态调整学习路径，学生完课率提高 25%，这体现了大模型在_____行业的应用。

- A. 金融
- B. 医疗
- C. 教育
- D. 零售

15. 在卷积神经网络中，池化层（Pooling Layer）可以_____。

- A. 增加网络深度
- B. 降低特征图空间维度
- C. 引入非线性激活
- D. 防止梯度消失

二、多选题（本大题 10 道小题，每小题 2 分，共 20 分），从下面题目给出的 A、B、C、D 四个可供选择的答案中选择所有正确答案。

1. 评测 RAG 系统时，_____是常用的指标。

- A. 召回率
- B. 答案忠实度
- C. 响应时间
- D. 模型参数量

2. _____属于参数高效微调（PEFT）方法。

- A. LoRA
- B. Prefix Tuning
- C. P-Tuning v2
- D. Full Fine-tuning

3. _____是人工智能的三大要素。

- A. 数据
- B. 算法
- C. 算力

D. 应用

4. _____属于自然语言处理的基础任务。

- A. 分词
- B. 词性标注
- C. 命名实体识别
- D. 图像分类

5. Transformer 编码器模块常见的组成包括_____。

- A. 多头自注意力
- B. 掩码注意力
- C. 前馈网络
- D. 残差连接与层归一化

6. _____应用于生成式人工智能的前沿领域。

- A. 扩散模型
- B. 生成对抗网络
- C. 多模态学习
- D. 线性回归

7. 构建负责任的人工智能系统时，应纳入的治理措施包括_____。

- A. 分群公平性评估
- B. 隐私影响评估
- C. 模型与数据版本追踪
- D. 风险事件处置与复核流程

8. 某医疗 AI 公司用大模型开发辅助诊断系统，输入患者症状描述后，模型生成可能的疾病列表。开发团队发现，对于常见病（如感冒）模型表现良好，但对于罕见病（如某种基因突变导致的代谢异常），往往给出错误或模糊的判断。同时，模型在回答“这种病应该挂什么科”时，有时给出“神经内科”和“内分泌科”两个不同答案，且都相当自信。关于该模型能力的分析，_____是正确的。

- A. 对常见病的诊断建议，体现了模型在文本生成和问答上的优势，即基于训练数据中的高频信息进行组织输出
- B. 对罕见病的误判，说明模型难以“创造”训练数据分布之外的新医学知识，这是其固有局限
- C. 对同一问题给出不同但都自信的答案，反映了模型输出存在不确定性，且校准能力不足
- D. 该模型在医学领域完全不可用，因为校准问题无法通过任何工程手段缓解

9. 某电商运营团队需要每天为 5000 个商品生成详情页文案，要求文案风格统一、可批量产出，并自动写入商品数据库。团队决定调用大模型 API 而非使用网页对话工具，_____是正确的。

- A. API 支持编程调用，可通过脚本批量处理 5000 个商品文案
- B. 网页对话可以保持长时间会话，更适合批量生成
- C. API 可以在系统提示词中固化文案风格要求，保证一致性

D. 网页对话的上下文窗口更大，生成质量更高

10. 某金融公司需要高频调用 AI 进行合规文本审核，并要求每次调用成本可核算、审计日志可追溯。_____是正确的。

- A. API 调用可以精确计量每次请求的 token 数和费用
- B. 网页对话也能提供详细的调用量和费用报表
- C. API 调用日志可记录每次请求的时间、内容和响应，便于审计
- D. 网页对话的历史记录无法自动导出，审计困难

三、是非题（本大题 15 道小题，每小题 1 分，共 15 分）。

1. 多 Agent 系统只要增加 Agent 数量，就一定比单 Agent 更可靠且成本更低。
2. 某大型企业在排产调度中引入深度强化学习（DRL）技术，表明其可用于解决复杂的优化决策问题。
3. 智能计算系统的定义与组成中不包括硬件部分。
4. 强化学习的核心是让智能体通过与环境交互获得奖励信号来学习最优策略。
5. 某公司计划部署一个用于文档摘要的 Transformer 模型。根据 Chinchilla 的结论，为了在给定算力预算下获得最佳性能，应该优先增加模型参数量，而不是增加训练数据量。
6. 人工智能的起源可以追溯到 20 世纪 50 年代的达特茅斯会议。
7. 联邦学习要求所有训练数据必须集中存储在一个中心服务器上。
8. 在 Gradio 中，ChatInterface 是 Interface 的一个特殊变体，专门为聊天场景设计，可自动处理对话历史、用户输入和模型输出的循环，因此相比直接用 Interface 搭建对话界面，ChatInterface 的代码量更少。
9. 某金融企业使用大模型构建投研助手，用户询问“今天央行最新公布的 LPR 利率是多少”，模型因训练数据截止于 2025 年而无法回答。为解决这个问题，可以在 API 调用前，先通过 RAG 系统从官网检索最新 LPR 数据，并将其放入提示词上下文。
10. 在 BIG-Bench 基准测试中，有超过 90% 的任务出现“涌现能力”，在改用连续度量后依然表现出明显的突然跃升，说明这些任务的涌现现象与度量选择无关。
11. RNN 能够处理变长的序列输入，但存在梯度消失的问题。
12. 雅砻江水风光一体化智慧运营大模型实现了 60 天超长周期流域水风光综合功率预报。这一案例说明大模型在能源领域具有应用价值。
13. 在调用大模型 API 时，如果服务器返回 HTTP 状态码 401，通常表示客户端发送的请求参数格式有误。

14. 在构建基于向量的知识库时，文档切分、向量化及建立索引是核心步骤。

15. 前向传播是指从输入层到输出层逐层计算网络输出的过程。

四、操作题

以下所有题目请在文件“C:\KS\答题纸E.docx”中作答！

（一）基础操作题（本大题5道小题，共35分）

1. 当前AI Coding和AI开发助手领域生态快速发展。请写出至少3个你了解的热门AI Coding/Agent/工具相关的名称（不限于Skill、规则文件、插件、框架、协议等），并至少对其其中1个简要说明其功能。（5分）

请在答题纸作答！此处答题一律无效！

2. 你负责部署一个RAG应用，需要将向量数据库（如Milvus）和推理服务（如vLLM）运行在同一台Linux服务器上。但服务器磁盘空间不足，你发现“/var/log”目录下日志文件占用了大量空间。请给出清理日志的Linux命令（至少两条），并解释在清理时需要注意什么以避免影响正在运行的服务。（5分）

请在答题纸作答！此处答题一律无效！

3. 论文《Training Language Models to Follow Instructions with Human Feedback》中提出了基于人类反馈的强化学习（RLHF）方法。请结合“C:\素材”中论文原文有关内容，阐述大模型在人类对齐方面的发展方向，并说明RLHF中奖励模型的作用。（5分）

请在答题纸作答！此处答题一律无效！

4. 阅读下面基于PyTorch实现图像分类模型训练的Python代码片段，该代码使用SGD优化器完成模型训练。代码中存在3处空缺，请从PyTorch训练循环基础API的角度，补全①②③位置的代码，实现完整的“前向计算-损失计算-反向传播-参数更新”训练流程。（5分）

```
model.train()
optimizer = torch.optim.SGD(model.parameters(), lr=1e-3)
for images, labels in train_loader:
    # 模拟随机增强
    images=torchvision.transforms.RandomHorizontalFlip(p=0.5)(images)
    # 梯度清零
    ①
    outputs = model(images)
    # 反向传播、梯度计算
    loss = torch.nn.functional.cross_entropy(outputs, labels) ②
    # 参数更新
    ③
```

请在答题纸作答！此处答题一律无效！

5. 某城市河道管理部门部署了AI巡检系统。巡检摄像头定时拍摄河面图像，系统识别塑料瓶、树枝等漂浮物，并生成清理建议。

请根据以下技术流程，回答相关问题。（15分）

河面图像输入→[A]定位漂浮物→[B]识别漂浮物类别→[C]生成清理建议→发送给工作人员

（1）步骤[A]通常使用什么计算机视觉技术？（3分）

请在答题纸作答！此处答题一律无效！

（2）步骤[B]中，模型主要根据图像中的哪些特征区分“塑料瓶”和“树枝”？请列举两类。（3分）

请在答题纸作答！此处答题一律无效！

（3）假设使用Python调用图像识别API，补全以下代码。（6分）

```
import requests
url = "https://api.example.com/v1/image/classify"
headers = {"Authorization": "Bearer your_api_key"}
files = {"image": open("river.jpg", "rb")}
data = {"model": "river-object-v1"}
response = requests.post(
    url, files=①, data=②, headers=③
)
print(response.json()["result"]["label"])
```

请在答题纸作答！此处答题一律无效！

（4）当模型识别置信度低于60%时，系统不直接生成清理工单，而是转交工作人员确认。这体现了AI应用中的哪项治理原则？（3分）

请在答题纸作答！此处答题一律无效！

（二）实践应用题（本大题3道小题，共65分）

上海正加快提升国际金融中心能级。依托股票、债券、期货等全品类国家级金融要素市场与国内外头部金融机构的深度集聚，上海已构建起开放领先、机构云集的现代金融产业体系，并在资管科技与绿色金融等前沿领域确立核心竞争优势。

伴随人工智能特别是大模型技术快速迭代，其已成为上海金融业数字化升级的重要引擎，全面赋能投研分析、智能风控、信贷服务、财富管理、合规审计等关键业务环节。立足国际金融中心政策与场景优势，上海联动金融要素市场、持牌机构、科研院所与科技企业，布局金融垂类基座大模型研发，实现财报智能解析、舆情研判、信贷风险识别，有效压缩业务处理时长、降低运营成本。与此同时，金融大模型评测体系、中试创新平台陆续落地，为智能投顾、普惠信贷、跨境贸易金融等业务提供支撑。浦东、徐汇等产业高地搭建产学研协同创新载体，推动大模型技术从算法原型走向业务验证与规模化产业落地，为上海建设更高水平的国际金融中心注入数字动能。

请结合这一背景，完成以下题目：

1.你是大模型工程师小沪，所在团队负责研发金融智能投研系统智投（IntelliInvest）。系统面向内部业务部门、投资顾问及终端客户，支撑个股投研分析、基金产品解读、公告舆情解析、合规审核、风险研判、智能问答等金融服务，日常承载数千次用户咨询，高并发峰值流量波动大，同时需满足金融行业强隔离、强合规、高稳定、低延迟的监管与业务要求。系

统迭代落地过程中，团队开展多轮技术研讨，同事观点会产生误导，需你进行甄别，结合金融业务特性完成企业级工程方案设计。（15分，每小题5分）

（1）业务现象：智投系统日常日咨询量达数千次，业务峰值并发压力大，当前所有用户请求均完整执行“向量检索-重排-大模型生成”全链路，导致线上服务推理成本居高不下、问答响应延迟偏高、GPU算力利用率波动剧烈，资源浪费严重。

部门研讨发言（存在误导观点，答题需甄别）

小李：全程使用大模型效果最准，不需要做任何性能节流优化。

王工：服务成本高、延迟大，直接扩容GPU、增加机器堆叠硬件即可解决。

小陈：语义缓存容易缓存过期数据、旧费率规则，坚决不用。

老张：连续批处理开发实现复杂，业务收益有限，没必要落地。

任务要求：请基于多层Serving服务架构，为智投系统设计包含语义缓存、模型路由、连续批处理、投机解码的全维度服务优化方案，梳理核心设计要点，并列出方案对应的核心观测指标。（本题聚焦服务侧调度与解码层优化，无需重复阐述量化相关内容。）

请在答题纸作答！此处答题一律无效！

（2）业务现象：智投系统同时服务公司多个业务部门、合作机构及终端客户，存储大量部门涉密资料、合作合同、产品检验报告、专属投研数据，需严格按照部门、客户实现数据隔离，集团内部各分支机构资料相互不可见，严禁跨客户越权查询、泄露业务数据。

部门研讨发言（存在误导观点，答题需甄别）

小李：所有数据统一存入一个向量库，架构简单、维护成本最低，无需分库隔离。

王工：权限控制无需改造架构，仅在提示词中增加“仅查看自身权限数据”即可。

小陈：检索阶段无需做权限拦截，等待召回结果后再过滤，不会影响业务。

老张：图检索场景使用频次低，不需要做权限校验管控。

任务要求：设计适配金融智投业务的检索权限与租户隔离方案，说明在关键词、向量、图混合召回链路中如何逐层施加权限约束，同时解决重排序阶段越权片段混入，避免答案被越权数据污染的核心问题。

请在答题纸作答！此处答题一律无效！

（3）业务现象：金融行业具备强监管、高合规要求，智投知识库存储大量敏感数据，包含员工隐私信息、客户信息、内部投研资料、未公开风控数据、涉密合同等内容。数据直接入库、检索生成回答极易引发信息泄露，触碰金融合规红线，存在重大业务风险。

部门研讨发言（存在误导观点，答题需甄别）

小李：敏感数据仅做加密存储，即可完全保障信息安全。

王工：在提示词中增加“禁止泄露敏感信息”约束，就能规避泄露风险。

小陈：数据脱敏会损耗文本语义、降低问答质量，无需做脱敏处理。

老张：安全审计非核心业务，可延后落地。

任务要求：结合金融强监管合规要求，为智投系统设计入库脱敏、查询水印、全链路审计完整合规机制，实现系统在严守合规底线的前提下稳定可用，构建金融大模型应用的安全防护体系。

请在答题纸作答！此处答题一律无效！

2.当前上海国际金融中心正加速推进金融数字化转型，金融大模型已广泛应用于财报智能解析、舆情风险研判、智能投研、合规审计、普惠金融咨询、跨境金融服务等核心场景。金融业务具备长文本财报研报输入、线上高并发咨询服务、低延迟交易辅助、高精度风控推

理、垂类场景定制迭代等典型特征，对大模型推理性能、资源开销、服务时延、微调效果有着极高的工程要求。请结合金融大模型落地实践，完成以下问题。（20分，每小题5分）

（1）金融大模型线上业务流量以流式持续抵达为主，涵盖实时智能投询、舆情极速研判等短吞吐请求，以及长篇财报解析、批量合规文稿审核等长文本请求，请求长短混杂、并发波动大，对推理算力利用率和服务时延稳定性要求极高。企业部署大模型推理服务时，主流采用静态批处理与连续批处理两种组批方式。请结合金融落地场景作答：

① 对比静态批处理与连续批处理，说明二者在请求组批逻辑、显存占用、KV Cache管理上的核心差异；（3分）

② 阐述PagedAttention分页KV管理机制的核心优势，说明其如何支撑金融场景下“动态增删请求”的高效批处理调度。（2分）

请在答题纸作答！此处答题一律无效！

（2）金融智能问答、投研文本摘要、合规报告生成等流式生成场景，对逐Token输出延迟极为敏感，TPOT指标直接决定用户体验与业务SLA。为优化解码阶段推理效率、降低生成时延，技术团队计划引入投机解码技术。请结合金融大模型线上业务特性回答：

① 简述投机解码的基本原理，投机解码能够有效提升TPOT（每Token延迟）的核心原因；（3分）

② 说明投机解码的适用边界与性能代价，列举具体场景与文本特征，阐述在什么情况下启用该技术反而会导致推理速度变慢、额外开销增加。（2分）

请在答题纸作答！此处答题一律无效！

（3）金融机构落地大模型推理服务，需兼顾高并发稳定性、低时延推理、合规可运维、量化适配、业务快速迭代等核心诉求，行业主流推理框架包含vLLM、TGI、SGLang、TensorRT-LLM。请结合金融业务高可用、高可靠、强合规的落地要求完成选型分析：

① 金融场景下，大模型推理框架选型应重点考察哪些核心维度；（3分）

② 对比说明vLLM、TGI、SGLang、TensorRT-LLM四者在产品定位、性能表现、技术取舍、适配能力上的核心差异。（2分）

请在答题纸作答！此处答题一律无效！

（4）金融大模型线上高并发场景下，长短请求混发现象突出，长耗时的批量合规核验、长篇研报分析请求易占用大量GPU资源，严重拖慢理财咨询、实时风险预警等低延时短请求，造成服务卡顿、时延飙升、SLA不达标等问题。请结合金融业务严苛的时延与稳定性要求，设计合理的请求调度方案：

① 简述FCFS、优先级调度、抢占式调度三种策略的核心特点，并分析各自在金融大模型推理场景中存在的问题；（3分）

② 线上运维过程中，需重点监控哪些核心指标，以此精准判断系统负载水平、科学决策服务限流与算力弹性扩容。（2分）

请在答题纸作答！此处答题一律无效！

3.金融智能投研Agent（智投）面向理财客户、投研人员、投资顾问，依托基金合同、招股书、年度财报、行业研报、行情数据、票据财报图表等资料，提供产品规则解读、个股基本面分析、市场动态研判、投资答疑、报告解读等服务。金融文档层级严谨、条款交叉引用多、结构化表格多、金融专有名词密集，对知识库切分、检索策略、多模态解析、工具调度能力要求极高，请结合业务场景完成以下题目。（30分，每小题10分）

（1）金融文档知识切分策略与向量检索基础

① 智投系统需批量接入数百份基金合同、招股说明书、上市公司年报等标准化金融长文档。若采用所有文档作为单一Chunk入库，会出现精准问答错位问题，例如用户询问“本基金赎回费怎么算”，系统错误召回“申购费率”整章内容，同时文档内跨章节的关联条款、补充说明被完全切断，问答准确性大幅下降。请结合金融合规文档特征，设计适配智投业务的知识切分策略：如何在保证检索精准度的同时保留上下文完整性？针对文档中高频出现的费率数据表、跨章节交叉引用条款、多级层级标题分别如何处理？请说明金融场景下合理的切片重叠（overlap）取值思路与依据。（5分）

请在答题纸作答！此处答题一律无效！

② 智投系统依托向量检索实现行业研报、财报文本的语义问答，支持用户自由提问行业趋势、公司经营、业务变化等语义类问题。团队在落地过程中，对Embedding模型选型、向量维度配置、距离度量方式、向量归一化操作、元数据过滤策略存在不确定性，容易出现语义匹配不准、新旧研报混杂、文档类型错乱等问题。请结合投研业务要求作答：金融垂类RAG场景下如何科学选型Embedding模型？金融向量检索通常采用何种距离度量方式？为什么必须做向量归一化处理？如何在不破坏原有语义检索效果的前提下，实现按研报发布日期、文档类型（年报/季报/研报/公告）的精准过滤筛选？（5分）

请在答题纸作答！此处答题一律无效！

（2）混合检索方案设计与重排序机制

① 在个股投研问答场景中，用户精准提问“600519（贵州茅台）最新动销情况”，纯向量检索对股票代码、债券代码、ISIN编码、金融专有名词辨识度极低，无法精准定位个股相关资料，大量召回无关的白酒行业综述、宏观行业分析内容，导致答非所问。为解决金融专属实体检索失效问题，请设计落地的混合检索方案：为什么金融精准问答场景必须补充关键词检索？向量相似度分数与关键词匹配分数如何做结果融合？请给出一套可直接落地的双路检索权重分配思路与排序逻辑。（5分）

请在答题纸作答！此处答题一律无效！

② 为提升投研问答召回质量，团队初期采用“向量粗排召回100条片段全部送入大模型生成答案”的方案，存在两大严重问题：一是输入上下文超长，大模型推理Token成本极高、响应延迟大；二是大量无关研报片段、过期公告内容混入上下文，干扰模型判断，导致答案失真、论据混杂。请结合智投业务说明重排序模块的完整机制：重排序模块的输入与输出分别是什么？详细说明检索链路从“向量粗排100条”筛选至“精排Top5/Top10”的完整流程；同时列出金融RAG场景下用于评估重排序效果的核心离线评测指标。（5分）

请在答题纸作答！此处答题一律无效！

（3）Agent工具调用治理与多模态解析规范

① 智投Agent需要具备自主能力，可根据用户问题自动调用行情查询、研报检索、财报解析等工具，实现“用户提问、模型自主决策、自动取证作答”。但线上落地极易出现工具调用参数错误、重复调用、无限循环调用、无效调用等问题，严重影响服务稳定性。请针对金融智投单Agent架构作答：各类金融工具如何通过Function Calling完成能力暴露与参数定义？智投Agent的ReAct思考-行动循环机制如何在投研场景落地？需要设计哪些明确的终止条件，彻底规避工具死循环、无效调用问题？（5分）

请在答题纸作答！此处答题一律无效！

② 智投业务支持用户自主上传财报截图、经营数据图表、票据影像、业绩趋势图等资料进行问答。若直接将图像输入多模态VLM模型，极易出现趋势判断颠倒、营收利润金额读错、数据维度混淆等幻觉问题，可信度无法满足投研参考标准。请设计金融多模态解析落地规范：明确哪些金融图文场景优先使用OCR、哪些场景必须使用VLM视觉大模型？为保证智投输出答案有据可依、可溯源、低幻觉，请设计至少两条落地机制，杜绝图表解读编造、错读问题。（5分）

请在答题纸作答！此处答题一律无效！

上海市教育考试院
版权所有