

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 II 类 (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

本考试分为两个部分：

- 第一部分（120分钟）：需完成客观题、基础操作题及实践应用题，其中“（一）基础操作题”和“（二）实践应用题”请在文件“C:\KS\答题纸E.docx”中作答。
- 第二部分（30分钟）：需完成综合题，按题目要求保存。

注意：第一部分考试结束并提交答卷后，方可进入第二部分考试，请合理安排答题时间。

第一部分

（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2道小题，共35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！

第二部分

(三) 综合题 (本大题3道小题, 共30分)

以下题目请在文件夹“C:\KS”下按题号创建对应代码工程目录:【第一题 AI辅助代码修复】、【第二题 AI辅助程序设计】、【第三题 AI智能体开发】, 完成后将3个文件夹制作成1个压缩包。

1.AI 辅助代码修复 (10 分, 每小题 5 分)

以下代码存在硬编码密钥、明文打印敏感信息、缺乏异常处理、文件 IO 泄漏敏感数据、无超时设置以及难以测试等严重问题。请在 VS Code + Cursor 中修复以下代码, 保持“两次调用: 初稿-审校”的行为。

要求: API 密钥从环境变量读取且不得打印/写日志; 统一封装 HTTP 客户端并设置超时; 使用依赖注入便于单元测试; 非 2xx 状态码和响应结构错误抛出明确异常; ResearchAgent 只负责编排。

```
API_KEY = "sk-demo"
URL = "http://127.0.0.1:8000/v1/chat/completions"
def ask(q):
    print(API_KEY, q)
    r = requests.post(
        URL,
        headers={"Authorization": "Bearer " + API_KEY},
        json={"messages": [{"role": "user", "content": q}]}
    )
    if r.status_code != 200:
        return ""
    return r.json()["choices"][0]["message"]["content"]

def research(q):
    draft = ask(q)
    final = ask("请审校: " + draft)
    open("log.txt", "a").write(API_KEY + final)
    return final
```

- (1) 提交修复代码, 以满足上述要求;
- (2) 请考虑该修复代码在 API 密钥安全性、HTTP 响应问题、环境变量异常处理等的测试需求。编写测试代码, 提供至少 2 个测试点代码。

2.AI 辅助程序设计 (10 分, 每小题 5 分)

你负责维护一个开源 Python 库, 用户报告了一个功能缺陷: “当输入包含特殊字符 (如 emoji) 时, 函数 calculate_hash() 会抛出 UnicodeDecodeError”。请使用 AI Coding 工具链 (API+Cursor+VSCode) 完成以下任务:

- (1) 在 VS Code 中用 Cursor 设计一个消除上述缺陷的程序;
- (2) 编写两个测试用例代码 (一个正常输入、一个 emoji 输入), 需经验证后无异常。

3. AI 智能体开发 (10 分, 每小题 2 分)

【背景材料】

随着分级诊疗和智慧医疗服务的推进, 社区卫生服务中心与二三级医院共同构成了城市就医网络。中老年患者就医时常遇到不知道挂哪个科室、没号白跑一趟、忘记带齐证件材料、

取药找不到医保定点药房等问题。某健康服务平台计划开发“社区就医智能陪诊助手”，根据患者家属描述的症状、所在位置和就诊时间，自动完成科室分诊建议、号源查询、路线规划和就诊提醒，帮助患者一次性带齐材料、少走弯路。

【用户任务】

“我父亲这两天膝盖疼，想明天上午去医院看看，位置由系统提供。请帮我判断应该挂什么科室，查一下附近医院明天上午有没有号，规划从家到医院的路线，并提醒我就诊需要带哪些材料。看完病如果需要取药，请帮我查一下医院附近的医保定点药房怎么走。如果明天有高温或暴雨等天气预警，也请提前提醒我。”

【考试环境与 API】

(1) 考试平台提供的基础 API/MCP

API/工具	主要功能	主要输入	主要返回
天气查询	查询指定位置、指定日期的天气	location、date	天气状况、温度、降雨概率、高温/暴雨等预警
地图查询	查询两地之间的路线、距离与耗时	origin、destination、departure_time	交通方式、距离、预计耗时、预计到达时间
搜索引擎	检索公开网络信息	query	相关网页标题、摘要、链接

(2) 考生须自行开发并准备 Mock 数据的业务 API

API	主要功能	主要输入	主要返回
department_guide	根据症状描述提供科室分诊建议（症状关键词→科室的规则映射）	symptom	推荐科室、分诊说明、就诊提示（需携带材料、注意事项）
hospital_search	查询附近医院及指定科室号源	location、department、date、time_range	医院名称、距离、号源状态（有号/已满/停诊）、可约时段、医保定点标志
pharmacy_query	查询医院附近医保定点药房	location、radius	药房名称、距离、是否医保定点、营业时间

说明：业务 API 由考生自行开发，Mock 数据须覆盖正常与异常情形（如首选医院号源“已满”或“停诊”），用于检验智能体的决策与调整能力。各 API 统一返回 JSON 数据，智能体应以 API 实际返回结果为依据，不得由大模型虚构科室建议、号源状态、距离或天气信息。

【操作要求】

(1) 智能体任务规划

根据用户需求设计合理的任务执行流。流程至少应包括：分析患者症状、就诊时间、出发位置和取药等需求；调用自行开发的分诊 API 确定目标科室；查询附近不少于 2 家医院的科室号源；调用地图查询从家到医院的路线；查询医院附近医保定点药房；调用天气查询获取明天天气与预警；综合生成就诊计划。要求说明各 API 的调用目的及其基本调用顺序。

(2) 业务 API 开发与 Mock 数据准备

① 使用 Python 轻量框架（如 Flask/FastAPI）实现 department_guide、hospital_search、pharmacy_query 三个业务 API，要求接口设计规范清晰，参数请求明确，数据响应统一采用标准 JSON 格式；

② Mock 数据字段完整、取值合理，且至少包含以下异常情形：首选医院目标时段号源“已满”或“停诊”；

③ 业务 API 能被智能体实际调用（本地 HTTP 服务或函数封装均可），智能体不得绕过 API 直接读取 Mock 数据文件。

(3) API 调用与信息获取

① 调用自行开发的 department_guide 接口获取推荐科室及就诊提示；

- ② 调用自行开发的 `hospital_search` 接口查询不少于 2 家医院明天上午的号源；
- ③ 对最终选定的医院调用地图查询从家出发的路线；
- ④ 调用自行开发的 `pharmacy_query` 接口查询选定医院附近的医保定点药房；
- ⑤ 调用天气查询获取明天天气及预警信息；调用搜索引擎检索症状相关的公开就诊常识（如“膝盖疼挂什么科”），与分诊 API 的结果相互印证。

（4）智能决策与结果调整

- ① 优先推荐“有号”、距离较近且为医保定点的医院，并说明选择理由；
- ② 若首选医院号源“已满”或“停诊”，应自动改选其他候选医院并说明原因；
- ③ 结合可约时段和地图返回的路程耗时，倒推建议从家出发的时间；
- ④ 就诊提醒必须包含 `department_guide` 返回的需携带材料信息（如身份证、医保卡、既往病历等）；
- ⑤ 若天气查询返回高温或暴雨预警，应增加相应的防暑、防雨和错峰出行提醒。

（5）最终结果输出

生成结构清晰的就医计划，至少包含：推荐科室及理由、选定医院及预约时段、建议出发时间与路线、就诊携带材料清单、取药药房信息（名称、距离、是否为医保定点）以及天气/预警提醒。建议格式如下：

环节	安排	地点	说明
出发	08:XX 从家出发	家	根据预约时段和地图路线耗时确定出发时间和路线，同时考虑天气情况。
就诊	上午时段，××科室	××医院	号源状态、医保定点及选择理由。
携带材料	身份证、医保卡、既往病历等	——	依据分诊 API 返回的就诊提示生成。
取药	医保定点药房	医院附近	名称、距离、营业时间。
返程	就诊结束后	——	考虑返程路线，同时考虑天气情况。

【考生提交内容】

- （1）智能体源代码或工作流文件；
- （2）业务 API 源代码及 Mock 数据文件；
- （3）System Prompt 或核心任务提示词（应约束智能体不虚构医疗与号源信息，并声明分诊建议不替代医生诊断）；
- （4）API 调用记录（覆盖自行开发的业务 API 与平台基础 API，至少包含 API 名称、主要输入参数、返回结果和基于返回结果的决策）；
- （5）智能体最终生成的“就医陪诊计划”。

【备注】

- （1）本环节考试平台仅提供天气查询、地图查询、搜索引擎三类基础 API/MCP，并提供大模型调用接口；
- （2）题干涉及的业务 API 由考生自行开发并准备 Mock 数据；Mock 数据仅作为业务数据源，天气、路线、公开网络信息等必须来自平台基础 API，不得由大模型自行编造；
- （3）可使用 Python requests、Function Calling、MCP 或考试平台提供的智能体开发环境，不限定具体 Agent 框架；
- （4）不要求开发图形界面，不要求实现真实挂号、支付功能；分诊建议仅作为就医参考，不替代专业医疗诊断；
- （5）不以 API 调用次数作为评分依据，重点考查“任务规划-API 开发-工具调用-信息判断-结果生成”的完整智能体工作流程。