2025 一带一路暨金砖国家技能发展与技术创新大赛

【首届 DeepSeek 大模型及行业 AI 应用开发赛项】

BRICS2025-ST-034

样题（选拔赛）

金砖国家工商理事会技能发展、应用技术与创新中方工作组
一带一路暨金砖国家技能发展与技术创新大赛组委会
竞赛技术委员会专家组制定
2025 年 5 月

DeepSeek 大模型及行业 AI 应用开发赛项选拔赛样题

项目包括以下内容：

1. 参加比赛的形式
2. 项目
3. 项目模块和所需时间
4. 试题详情

竞赛时间：3 小时

1. 参加比赛的形式

双人团队参与

2. 项目

1) 理论测试

2) 行业 AI 系统开发

只有场地或平台无法完成并经技能竞赛经理批准时，才能更改测试项目和标准。

如果竞争对手不遵守要求或使自己和/或其他竞争对手面临危险，则可能会将其从竞争中移除。

测试项目将根据随机抽签按顺序完成。当参赛者完成模块时，结果将进行评分。

3. 项目阶段和所需时间

阶段和时间总结在表 1 中

序号	阶段名称	阶段完成时间
1	第一阶段：理论测试	60 分钟
2	第二阶段:行业 AI 系统开发	120 分钟

表 1.项目阶段列表

4.试题详情

第一阶段：理论测试

单选题（50 题）

1. 在 DeepSeek 大模型开发中，以下哪种 Python 库常用于数据处理和分析？（ ）
A. NumPy
B. Matplotlib
C. OpenCV
D. TensorFlow

2. 以下哪项属于 RAG（检索增强生成）的核心步骤？（ ）
A. 直接输入文本生成答案
B. 先检索相关信息，再结合检索结果生成答案
C. 仅对输入文本进行格式转换
D. 只对输入文本进行关键词提取

3. 若要存储和查询结构化数据，以下哪种数据库系统较为合适？（ ）
A. MariaDB
B. Milvus
C. 都不适合
D. 都适合

4. 数据采集方法中，通过编写程序自动获取网页数据的方式是？（ ）
A. 传感器技术
B. 网络爬虫
C. 问卷调查
D. 人工录入

5. 以下哪种算法属于模型训练中的优化算法？（ ）
A. 随机森林
B. 决策树
C. Adam
D. K-Means

6. 在深度学习中，Transformer 架构的核心组件是？（ ）
A. 卷积层
B. 循环层
C. 注意力机制

D. 全连接层

7. 若要对图像进行处理和计算机视觉的基础操作，Python 中常用的库是？（ ）
A. NumPy
B. Pandas
C. Matplotlib
D. OpenCV

8. 以下哪个工具常用于大模型的集成和部署？（ ）
A. Anaconda
B. Streamlit
C. Ollama
D. Milvus

9. DeepSeek 大模型对硬件资源的需求，以下说法正确的是？（ ）
A. 仅依赖 CPU
B. 仅依赖 GPU
C. 主要依赖 GPU，对 CPU 也有一定需求
D. 对硬件资源无特殊要求

10. 使用 Anaconda 创建独立虚拟环境的主要目的是？（ ）
A. 提高计算机性能
B. 隔离不同项目的 Python 环境和包依赖
C. 加速模型训练
D. 优化数据存储

11. 以下关于 DeepSeek 与 GPT 系列模型在架构上的关键区别描述，最准确的是？（ ）
A. DeepSeek 采用纯 CNN 架构，GPT 使用 Transformer
B. DeepSeek 基于多模态融合架构，GPT 仅处理文本
C. DeepSeek 优化了 Transformer 的多头注意力机制，在长序列处理效率上更优
D. DeepSeek 不依赖预训练，GPT 必须经过大规模预训练

12. 在使用 RAG 进行文档问答系统开发时，以下哪个操作最可能导致检索结果相关性下降？（ ）
A. 对文档进行关键词提取后构建索引
B. 将检索向量维度从 768 维压缩至 384 维
C. 未对用户输入问题进行停用词过滤
D. 采用 TF-IDF 算法替代 BM25 算法进行检索

13. 当利用 Python 进行数据采集时，以下哪种场景适合使用 Selenium 而非网络爬虫框架？（ ）

A. 爬取动态加载的网页数据
B. 爬取静态 HTML 页面数据
C. 爬取 JSON 格式 API 数据
D. 批量下载 CSV 文件

14. 在模型训练过程中，以下哪种超参数调整方式会导致模型出现“梯度消失”问题？（ ）
A. 增大学习率
B. 减小 batch_size
C. 加深神经网络层数且未使用残差连接
D. 增加 Dropout 概率

15. 关于 Milvus 与 Faiss 在向量检索中的应用，以下说法正确的是？（ ）
A. Milvus 适合大规模向量数据的分布式存储与检索，Faiss 侧重单机高效计算
B. Faiss 仅支持 CPU 计算，Milvus 仅支持 GPU 计算
C. Milvus 的数据索引类型比 Faiss 更丰富
D. Faiss 的安装与部署比 Milvus 更简便

16. 在使用 Langchain 构建知识图谱问答系统时，以下哪个组件用于解析自然语言问题为 SPARQL 查询语句？（ ）
A. LLMChain
B. SQLDatabaseChain
C. GraphCypherRetriever
D. NLCYPHERChain

17. 当对非结构化文本数据进行清洗时，以下哪项操作会破坏文本语义完整性？（ ）
A. 去除 HTML 标签
B. 统一文本大小写
C. 使用正则表达式删除特殊符号时误删缩写词中的点号
D. 进行停用词过滤

18. 在 DeepSeek 大模型部署过程中，以下哪种硬件配置组合能实现最佳性价比？（ ）
A. 高端 CPU+普通 GPU
B. 低功耗 CPU+专业级 AI 加速卡
C. 多核 CPU+高性能 GPU
D. 仅使用多核 CPU

19. 关于 Anaconda 虚拟环境与 Docker 容器，以下描述正确的是？（ ）
A. Anaconda 虚拟环境用于隔离 Python 运行环境，Docker 容器用于隔离操作系统环境

B. Anaconda 虚拟环境只能管理 Python 包，Docker 容器只能管理系统级软件
C. Anaconda 虚拟环境无法在不同操作系统间迁移，Docker 容器也不能
D. Anaconda 虚拟环境与 Docker 容器的创建和管理命令完全相同

20. 在使用 PyTorch 训练神经网络时，以下哪个函数用于计算交叉熵损失？（ ）
A. nn.MSELoss( )
B. nn.L1Loss( )
C. nn.CrossEntropyLoss( )
D. nn.SmoothL1Loss( )

21.在使用 PyTorch 搭建一个简单的卷积神经网络（CNN）进行图像分类时，以下代码片段缺少关键层定义。请选择正确的代码补全空白处，实现一个包含卷积层、池化层和全连接层的网络结构。(　)

```
import torch
import torch.nn as nn

class SimpleCNN(nn.Module):
    def __init__(self):
        super(SimpleCNN, self)._init_()
        self.conv1 =nn.Conv2d(3, 16, kernel_size=3, padding=1)
        self.pool = nn.MaxPool2d(2, 2)
        # 补全此处代码
        self.fc1 = nn.Linear(_____, 128)
        self.fc2 = nn.Linear(128, 10)

    def forward(self, x):
        x = self.pool(torch.relu(self.conv1(x)))
        x = x.view(-1, _____)
        x = torch.relu(self.fc1(x))
        x = self.fc2(x)
        return x
```

A. 16 * 16 * 16 和 16 * 16 * 16
B. 16 * 8 * 8 和 16 * 8 * 8
C. 16 * 32 * 32 和 16 * 32 * 32
D. 16 * 4 * 4 和 16 * 4 * 4

22.下表对 Milvus 和 MariaDB 在 AI 应用中的特性进行了对比，其中描述错误的一项是？（ ）

对比项目	Milvus	MariaDB

数据类型支持	主要存储向量数据，适合非结构化数据的特征表示	存储结构化数据，如关系型表格
查询方式	基于向量相似度检索	基于 SQL 语句进行数据查询
性能优势	在大规模向量检索场景下效率高	在事务处理和结构化数据查询方面表现出色
数据更新频率	适合频繁更新向量数据	不适合频繁更新数据

A. 数据类型支持
B. 查询方式
C. 性能优势
D. 数据更新频率

23.在开发一个结合 DeepSeek 大模型和计算机视觉技术的智能安防系统时，需要实现对监控视频中的异常行为进行检测，并通过 DeepSeek 生成处理建议。以下技术流程排列正确的是？（ ）
① 使用 OpenCV 对监控视频进行逐帧处理，提取图像特征
② 将异常行为相关的文本描述输入 DeepSeek 大模型
③ 利用深度学习目标检测模型（如 YOLO）对图像进行分析，识别异常行为
④ DeepSeek 根据输入文本生成相应的处理建议
⑤ 将检测到的异常行为转换为文本描述
A. ①③⑤②④
B. ①⑤③②④
C. ③①⑤②④
D. ③⑤①②④

24. 在使用随机梯度下降(SGD)算法训练神经网络模型时，学习率(learning rate)参数设置过大，可能会导致以下哪种情况？（ ）
A. 模型收敛速度变慢，训练时间显著增加
B. 模型在训练过程中出现振荡，无法收敛到最优解
C. 模型容易陷入局部最优解，难以找到全局最优
D. 模型参数更新幅度极小，几乎不进行学习

25. 某公司希望利用 Langchain、DeepSeek 和 Milvus 开发一个智能法律咨询系统，以下关于各技术分工描述错误的是？（ ）
A. Langchain 用于构建与用户的交互逻辑，解析用户问题并调用 DeepSeek
B. DeepSeek 作为核心语言模型，根据用户问题和检索信息生成法律解答
C. Milvus 存储法律条文的向量表示，用于快速检索相关法律知识
D. Langchain 负责存储和管理法律条文的结构化数据

26. 关于 Bert 和 Transformer 的关系，以及 DeepSeek 对其架构的应用，以下说法正确的是？（ ）
A. Bert 完全继承了 Transformer 的编码器和解码器结构，DeepSeek 重新设计了全

新架构

B. Bert 仅使用了 Transformer 的编码器部分，DeepSeek 在 Transformer 架构基础上进行了优化和改进

C. Bert 对 Transformer 架构进行了大幅修改，DeepSeek 沿用了 Bert 的架构

D. Transformer 是 Bert 的简化版本，DeepSeek 与 Transformer 架构无关

27. 若要部署一个 DeepSeek 大模型，同时满足高并发的推理请求，以下硬件资源配置方案最合理的是？（ ）

A. 普通办公 CPU，搭配大容量内存

B. 多核高性能 CPU，不配置 GPU

C. 多核高性能 CPU+多块专业级 GPU，搭配高速存储设备

D. 低功耗 CPU+单块普通 GPU

28. 以下是使用 scikit-learn 库进行数据挖掘和模型训练的流程描述，其中存在错误的一项是？（ ）

① 收集原始数据，可能包括结构化和非结构化数据

② 直接将原始数据输入模型进行训练

③ 使用数据预处理技术，如数据清洗、去重、标准化等

④ 划分训练集和测试集

⑤ 选择合适的机器学习模型，如决策树、支持向量机等

⑥ 使用训练集数据训练模型，并在测试集上评估模型性能

A. ①

B. ②

C. ③

D. ④

29. 针对“利用 AI 技术实现电商平台商品评论的情感分析和热点提取”这一需求，以下技术方案中最全面且合理的是？（ ）

A. 使用网络爬虫采集评论数据，用 Python 的字符串处理函数进行简单清洗，直接输入 DeepSeek 模型获取结果

B. 通过数据库查询获取评论数据，使用 Pandas 进行数据清洗，利用 scikit-learn 库的情感分析算法进行分析，不使用大模型

C. 使用网络爬虫采集评论数据，利用 Langchain 构建交互逻辑，结合 DeepSeek 大模型进行情感分析和热点提取，同时利用 RAG 技术检索相关知识辅助分析

D. 仅依靠人工阅读评论进行情感判断和热点提取

30. 下列关于“生成式 AI”与“判别式 AI”的概念描述，正确的是？（ ）

A. 生成式 AI 只能生成文本，判别式 AI 只能进行分类任务

B. 生成式 AI 通过学习数据分布生成新内容，判别式 AI 用于判断输入数据的类别或属性

C. DeepSeek 仅属于判别式 AI 模型

D. 判别式 AI 在训练时不需要标注数据

31. 在 AI 应用开发中，“数据增强”与下列哪个概念的关联最为紧密？（ ）
A. 模型推理速度
B. 数据隐私保护
C. 模型泛化能力
D. 硬件资源利用率

32. 以下哪个场景最适合应用“迁移学习”概念？（ ）
A. 基于 DeepSeek 开发全新领域的智能客服系统，已有少量该领域对话数据
B. 对大量标准化的电商商品图片进行分类
C. 用固定的模板生成公司月度报表
D. 实时统计网页访问流量数据

33. 关于“Transformer 架构中的注意力机制”，下列说法错误的是？（ ）
A. 注意力机制使模型在处理序列数据时能关注到重要部分
B. 多头注意力机制通过多个“头”从不同角度捕捉数据特征
C. 注意力机制会增加模型计算量，导致训练速度变慢
D. 注意力机制仅在 DeepSeek 模型训练阶段起作用，推理阶段不参与

34. 下列对“结构化数据”和“非结构化数据”的描述，正确的是？（ ）
A. 结构化数据只能用关系型数据库存储，非结构化数据只能用文件系统存储
B. 图片和音频属于结构化数据，表格数据属于非结构化数据
C. 结构化数据有明确的数据模式，非结构化数据没有固定格式
D. 非结构化数据无法进行数据分析，结构化数据才能分析

35. “超参数调优”与以下哪个 AI 开发环节直接相关？（ ）
A. 数据采集工具选择
B. 模型性能优化
C. 数据标注规范制定
D. 硬件设备采购

36. 在“利用 AI 技术对古籍文献进行语义检索”场景中，最关键的概念应用是？（ ）
A. 图像识别
B. 语音合成
C. 自然语言处理与向量检索
D. 强化学习

37. “监督学习”和“无监督学习”的主要区别在于？（ ）
A. 监督学习需要标注数据，无监督学习不需要
B. 监督学习只能用于分类任务，无监督学习只能用于聚类任务

C. 监督学习模型训练速度比无监督学习快
D. 无监督学习能自动发现数据中的模式，监督学习不能

38. 关于“大模型的涌现能力”，下列说法正确的是？（ ）
A. 涌现能力是所有 AI 模型天然具备的特性
B. 模型参数达到一定规模后，无需训练就能产生涌现能力
C. 涌现能力表现为模型在训练过程中未专门学习的任务上展现出良好性能
D. 涌现能力仅体现在语言生成任务中

39. DeepSeek 大模型首次公开亮相的年份是？（ ）
A. 2021 年
B. 2022 年
C. 2023 年
D. 2024 年

40. Milvus 向量数据库默认使用的向量索引类型是？（ ）
A. FLAT
B. IVF_FLAT
C. HNSW
D. SQ8

41. 在 scikit - learn 的决策树模型中，min_samples_split 参数的含义是？（ ）
A. 叶节点的最小样本数
B. 进行划分时所需的最小样本数
C. 树的最大深度
D. 随机数种子

42. Bert 模型中，基础版本的隐藏层数量是多少？（ ）
A. 6 层
B. 12 层
C. 18 层
D. 24 层

43. Anaconda 创建虚拟环境时，默认使用的 Python 版本是？（ ）
A. Python 3.6
B. Python 3.7
C. Python 3.8
D. 与系统默认 Python 版本一致

44. Adam 优化算法结合了哪两种优化算法的优点？（ ）
A. 随机梯度下降（SGD）和 AdaGrad
B. AdaGrad 和 RMSProp

C. RMSProp 和随机梯度下降（SGD）
D. AdaDelta 和 AdaGrad

45. Streamlit 库主要用于实现 AI 应用的哪方面功能？（ ）
A. 数据采集
B. 模型训练
C. 可视化界面搭建
D. 数据标注

46. RAG 在 AI 领域的全称是？（ ）
A. Retrieval - Augmented Generation
B. Recurrent Attention Generation
C. Randomized Algorithm Generation
D. Reinforcement Adaptive Generation

47. DeepSeek 大模型的核心优势之一是？（ ）
A. 低能耗运行
B. 专注于图像识别
C. 多模态处理能力突出
D. 训练速度极快

48. 在 Python 的 Pandas 库中，用于读取 CSV 文件的函数是？（ ）
A. read_excel( )
B. read_csv( )
C. read_sql( )
D. read_json( )

49. 某企业要开发一个智能文档处理系统，需对大量合同文档进行语义检索与关键信息提取，以下工具组合中最合理的是？（ ）
A. MariaDB + OpenCV + Langchain
B. Milvus + DeepSeek + Langchain
C. MySQL + TensorFlow + Streamlit
D. PostgreSQL + Scikit-learn + Ollama

50. 在训练一个大规模图像分类模型时，分别使用批量梯度下降（BGD）和随机梯度下降（SGD）算法，以下关于两者在该场景下表现的分析，正确的是？（ ）
A. BGD 在训练初期收敛速度比 SGD 快
B. SGD 更容易陷入局部最优解
C. BGD 在内存占用上明显低于 SGD
D. SGD 更适合在有限内存资源下训练大规模数据集

多选题（25）

1. 以下关于 Python 在 DeepSeek 大模型开发中的应用，说法正确的有（ ）
A. 利用 NumPy 库进行高效的数值计算，为模型训练提供数据支持
B. 使用 Pandas 库对训练数据进行清洗、转换和分析，提升数据质量
C. 通过 Matplotlib 库绘制模型训练过程中的损失函数曲线，直观展示训练效果
D. 借助 OpenCV 库实现大模型的核心算法逻辑

2. 在运用 Langchain 进行 AI 应用开发时，可实现的功能包括（ ）
A. 构建与 DeepSeek 等大模型的交互接口，优化 Prompt 输入
B. 对采集到的网络爬虫数据进行实时清洗
C. 整合多种数据源，实现基于 RAG 的检索增强生成功能
D. 自动生成复杂的数据库 SQL 查询语句

3. 关于数据库系统在 AI 应用中的选择与使用，下列说法正确的是（ ）
A. MariaDB 适合存储结构化的用户信息、交易记录等数据
B. Milvus 可用于存储图像、文本等数据的特征向量，实现快速相似性检索
C. 当需要频繁进行事务处理时，选择 Milvus 比 MariaDB 更合适
D. 在搭建知识图谱问答系统时，可同时使用 MariaDB 和 Milvus，发挥各自优势

4. 数据采集过程中，常用的工具和方法有（ ）
A. 使用 Scrapy 框架编写网络爬虫，采集网页上的公开数据
B. 通过传感器设备获取温度、湿度等物理环境数据
C. 设计在线问卷调查，利用问卷星等平台收集用户反馈数据
D. 直接从数据库中复制其他企业的商业数据用于分析

5. 数据预处理对于 AI 模型训练至关重要，以下属于数据预处理操作的有（ ）
A. 对采集到的文本数据进行分词、去除停用词处理
B. 将不同格式的图像数据统一调整为 RGB 格式，并缩放至固定尺寸
C. 对数值型数据进行归一化或标准化处理，使数据处于相同尺度
D. 为提高模型训练速度，直接删除数据集中的所有缺失值记录

6. 深度学习基础理论中，常见的神经网络架构有（ ）
A. Transformer 架构，广泛应用于自然语言处理和计算机视觉领域
B. Bert 架构，基于双向 Transformer 编码器，在语言理解任务中表现出色
C. CNN（卷积神经网络）架构，特别适合处理具有网格结构的数据，如图像、音频
D. RNN（循环神经网络）架构，能有效处理序列数据，但存在梯度消失问题

7. 在模型训练过程中，以下哪些方法可以用于超参数调优？（ ）
A. 交叉验证，将数据集划分为多个子集，多次训练评估模型，选择最优超参数
B. 网格搜索，通过穷举指定超参数的所有可能组合，找到最佳参数配置
C. 随机搜索，在超参数空间中随机采样进行训练，适用于超参数较多的情况
D. 遗传算法，模拟生物进化过程，通过选择、交叉、变异操作优化超参数

8. 以下机器学习库中，可用于构建和训练机器学习模型的有（ ）
A. scikit-learn，提供丰富的机器学习算法，如线性回归、决策树等
B. TensorFlow，功能强大的深度学习框架，支持构建各种复杂神经网络模型
C. PyTorch，以动态计算图为特色，广泛应用于深度学习研究和开发
D. Pandas，主要用于数据处理和分析，不能用于模型构建

9. 关于 DeepSeek、通义千问等大模型对硬件资源的需求，说法正确的有（ ）
A. 大模型训练和推理过程对 GPU 计算能力要求极高，需要高性能 GPU 加速
B. CPU 在大模型应用中仅起到辅助作用，可有可无
C. 大模型运行时需要足够的内存来存储模型参数和中间计算结果
D. 硬盘的读写速度会影响大模型训练数据的加载效率，进而影响训练速度

10. 以下关于大模型集成工具的描述，正确的有（ ）
A. Streamlit 可以快速搭建大模型应用的 Web 界面，方便展示模型输出结果
B. OpenWebui 提供可视化的操作界面，便于用户与大模型进行交互
C. Cherry Studio 支持多种大模型的集成开发，提供丰富的开发插件和工具
D. 这些集成工具只能用于 DeepSeek 大模型，无法集成其他大模型

11. 关于 Prompt 工程在 DeepSeek 大模型应用中的作用，以下表述正确的有（ ）
A. 合理设计 Prompt 可以引导模型生成更符合预期的答案
B. 通过调整 Prompt 的结构和内容，能提升模型的推理能力
C. 在处理多轮对话场景时，Prompt 的连续性和逻辑性至关重要
D. 简单重复的 Prompt 也能让模型产生多样化的优质输出

12. 在使用 RAG（检索增强生成）技术时，可能涉及的操作包括（ ）
A. 对海量文档数据进行向量化处理，构建向量索引
B. 根据用户问题在向量索引中检索相关信息
C. 将检索到的信息与用户问题整合后输入 DeepSeek 等大模型
D. 对大模型生成的答案进行后处理，如格式调整、内容校验

13. 以下关于数据清洗中去重操作的说法，正确的有（ ）
A. 对于结构化数据，可以通过指定唯一键字段来识别和删除重复记录
B. 在处理文本数据时，语义重复但表述不同的内容也需要进行去重
C. 去重操作可能会导致数据量减少，影响模型训练的样本多样性
D. 去重后的数据一定能提高模型的训练效果和泛化能力

14. 在深度学习模型训练中，优化算法的选择会影响训练效果，以下关于优化算法的描述，正确的有（ ）
A. 批量梯度下降（BGD）每次更新使用全部样本，训练过程更稳定，但计算量大
B. 随机梯度下降（SGD）每次更新仅使用一个样本，收敛速度快，但容易产生振荡
C. Adam 优化算法结合了 AdaGrad 和 RMSProp 的优点，适用于多种场景

D. 不同的优化算法对模型的超参数设置要求相同

15. 关于 Scikit-learn 库在机器学习中的应用，下列说法正确的是（ ）
A. 可以使用 Scikit-learn 的 train_test_split 函数划分训练集和测试集
B. 利用其中的 LinearRegression 类可以构建线性回归模型
C. GridSearchCV 类可用于对模型的超参数进行网格搜索和交叉验证
D. Scikit-learn 仅适用于传统机器学习算法，无法与深度学习结合

16. 在计算机视觉任务中，OpenCV 库的常见应用包括（ ）
A. 读取和显示图像、视频文件
B. 对图像进行滤波、边缘检测等预处理操作
C. 实现目标检测算法，如 Haar 特征级联检测
D. 训练深度学习的图像分类模型

17. 若要搭建一个基于 DeepSeek 的智能客服系统，需要考虑的方面有（ ）
A. 设计合适的对话流程和交互逻辑
B. 准备高质量的问答语料，用于微调或优化 Prompt
C. 部署服务器并配置足够的硬件资源，保障系统响应速度
D. 开发用户界面，提供友好的交互体验

18. 在大模型运行环境方面，以下描述正确的有（ ）
A. Ollama 可以方便地部署和运行多种大语言模型
B. DeepSeek 运行时对硬件的要求会随着模型规模增大而提高
C. 通义千问只能在阿里云的特定环境中运行
D. Dify 提供了大模型应用开发和部署的一站式解决方案

19. 当在 AI 应用开发中涉及多模态数据时，需要处理的内容包括（ ）
A. 对不同模态的数据进行采集和预处理
B. 设计合适的方法融合多模态数据，如早期融合、晚期融合等
C. 针对多模态数据训练或选择合适的模型
D. 评估多模态模型的性能，考虑不同模态的贡献度

20. 关于 DeepSeek 大模型与传统机器学习模型的区别，以下说法正确的是（ ）
A. DeepSeek 基于深度学习架构，能够自动学习数据特征，而传统机器学习模型多依赖人工特征工程
B. DeepSeek 可处理大规模数据和复杂任务，传统机器学习模型在处理海量数据时效率较低
C. DeepSeek 通过大量数据预训练获得通用能力，传统机器学习模型通常针对特定任务训练
D. DeepSeek 的模型参数数量远多于传统机器学习模型，需要更强的计算资源支持

21. 在使用 Langchain 构建 AI 应用时，其提供的实用功能模块包括（ ）

A. PromptTemplate，用于创建结构化、可复用的 Prompt 模板
B. ConversationBufferMemory，可存储对话历史，实现多轮对话的上下文感知
C. SQLDatabaseChain，方便与关系型数据库交互，执行数据查询和操作
D. AgentExecutor，能根据任务需求动态调用不同工具和模型，实现复杂任务处理

22. 对于数据采集过程中的网络爬虫技术，以下描述正确的有（ ）
A. 爬虫需要遵循网站的 robots.txt 协议，避免非法采集数据
B. 可以使用 Selenium 模拟浏览器行为，解决动态网页数据采集问题
C. 为了提高采集效率，可以设置多个线程或进程并发抓取数据
D. 采集到的数据需要进行合法性和完整性检查，防止无效数据干扰后续处理

23. 在数据预处理环节，针对文本数据的清洗操作通常包括（ ）
A. 去除文本中的 HTML 标签、特殊字符和乱码
B. 将文本统一转换为小写，消除大小写差异带来的影响
C. 进行词形还原（Lemmatization）或词干提取（Stemming），规范词汇形态
D. 使用词袋模型（Bag of Words）将文本转换为向量形式

24. 在深度学习模型的训练过程中，以下哪些方法可以用于防止过拟合？（ ）
A. 增加训练数据量
B. 使用正则化方法，如 L1 和 L2 正则化
C. 采用 Dropout 技术
D. 减小神经网络的规模

25. 关于自然语言处理中的词向量表示，以下说法正确的是（ ）。
A. 词向量是将单词映射到低维向量空间的一种表示方法
B. 词向量可以捕捉单词之间的语义和句法关系
C. Word2Vec 是一种常用的词向量生成模型
D. 不同的词向量模型生成的词向量在维度和含义上可能不同

判断题（25）

1. 使用 Python 的 Pandas 库的`read_csv()`函数读取 CSV 格式数据集时，会自动处理数据中的缺失值。（ ）
2. 在使用 Langchain 优化与 DeepSeek 的交互时，`load_memory`函数可以直接加载外部知识库，增强模型回答的专业性。（ ）
3. MariaDB 作为关系型数据库，不适合存储 AI 模型训练过程中的超参数配置记录。（ ）
4. 分布式爬虫通过多台机器协同工作，能够提高数据采集的速度和效率。（ ）
5. 在数据预处理中，对非结构化文本数据去除停用词后，可直接将剩余词语作为模型输入。（ ）
6. L1 正则化通过约束参数的平方和，防止深度学习模型过拟合。（ ）

7. PyTorch 采用静态计算图，在模型部署方面比 TensorFlow 更具优势。（ ）

8. 在部署 DeepSeek 大模型时，增加内存容量可以确保模型参数和中间计算结果能够快速存取。（ ）

9. 使用`conda env export`命令将 Anaconda 虚拟环境配置导出为 YAML 文件后，在不同操作系统之间迁移虚拟环境无需重新安装任何依赖包。（ ）

10. 在 Streamlit 开发大模型应用界面时，`st.cache`装饰器可对所有函数进行缓存，无任何限制条件。（ ）

11. 随机梯度下降（SGD）每次更新使用全部样本，训练过程更稳定，但计算量大。（ ）

12. Milvus 在处理大规模非结构化数据特征向量时，检索效率比关系型数据库低。（ ）

13. 数据增强技术只能通过增加数据量来提高模型泛化能力，不能作为一种正则化手段。（ ）

14. 在使用 Scrapy 框架编写网络爬虫时，`Item`类用于定义要采集的数据结构。（ ）

15. 通过调整 Prompt 的结构和内容，无法提升 DeepSeek 大模型的推理能力。（ ）

16. 当 AI 项目需要频繁进行事务操作时，Milvus 比 MariaDB 更适合作为数据库选型。（ ）

17. 词嵌入技术（如 Word2Vec、BERT Embeddings）不能将非结构化文本转换为向量表示。（ ）

18. 在深度学习模型训练中，学习率设置过小会导致模型在训练时无法收敛，出现振荡现象。（ ）

19. 仅复制 Anaconda 虚拟环境文件夹，就能保证在其他机器上正常使用该虚拟环境。（ ）

20. 在设计推荐系统时，不能同时使用关系型数据库和 Milvus 向量数据库。（ ）

21. DeepSeek 大模型只能处理文本数据，无法应用于图像、音频等多模态场景。（ ）

22. 在使用 RAG（检索增强生成）技术时，不需要对文档数据进行向量化处理。（ ）

23. 批量梯度下降（BGD）算法在训练大规模数据集时，比随机梯度下降（SGD）算法收敛速度更快。（ ）

24. 使用 Scikit-learn 库进行机器学习模型训练时，不需要对数据进行标准化处理。（ ）

25. 在 AI 应用开发中，数据标注的准确性对模型训练结果没有影响。（ ）

第二阶段：行业 AI 系统开发

一. AI 聊天助手开发

1、背景与需求

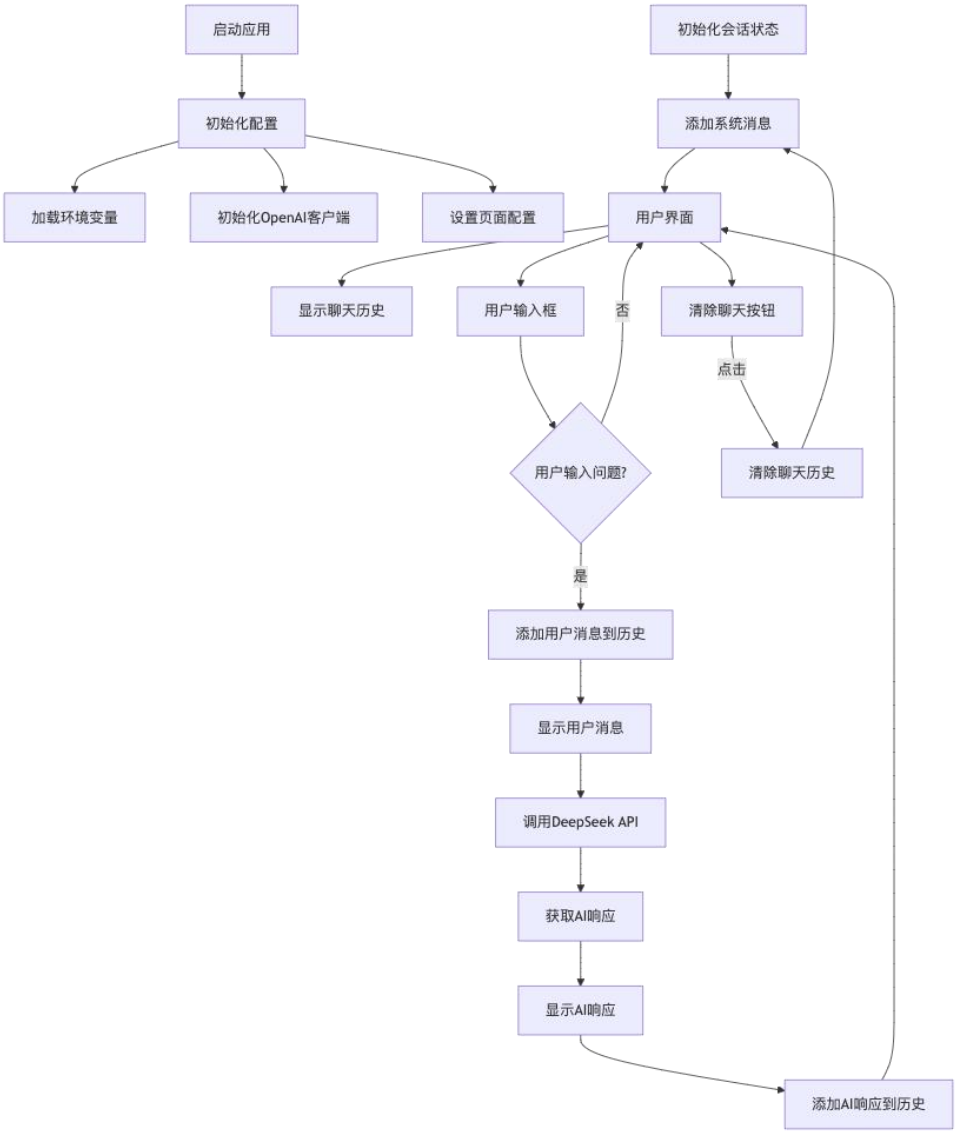
在数字化浪潮席卷全球的今天，人工智能技术正在深刻改变着人们的生活方

式。某知名科技公司"未来智能"正在开发新一代智能客服系统，旨在为全球用户提供更智能、更人性化的服务体验。作为公司的技术负责人，你被委以重任：开发一个基于 DeepSeek 大语言模型的智能聊天助手原型。这个助手不仅要能够理解用户的意图，还要能够提供专业、准确、富有同理心的回答。在开发过程中，你需要考虑多轮对话的连贯性、响应速度、用户体验等多个维度，确保这个聊天助手能够真正帮助用户解决问题，提升服务体验。你的目标是打造一个既智能又温暖的 AI 助手，让它成为连接用户与服务的桥梁，为用户带来全新的交互体验。

2、技术要点

使用 Streamlit 构建直观的用户界面
集成 DeepSeek API 实现智能对话
实现多轮对话的上下文管理
优化响应速度和用户体验
确保对话的连贯性和准确性

3、DeepSeek 聊天助手流程图



4、主要流程说明

初始化阶段

加载环境变量

初始化 OpenAI 客户端（配置 DeepSeek API）

设置 Streamlit 页面配置

初始化会话状态，添加系统消息

用户界面

显示聊天历史记录

提供用户输入框

显示清除聊天按钮

对话流程

用户输入问题

将用户消息添加到历史记录

调用 DeepSeek API 获取响应

显示 AI 响应
将 AI 响应添加到历史记录
清除功能
点击清除按钮
重置聊天历史（保留系统消息）
返回初始状态

5、DeepSeek 聊天助手效果图

6、题目要求 ：

请你在实操环境的 IDE 中调试下面代码，运行成功后，在聊天助手中输入：“我是 xxx，请你以我的名字写首藏头诗”；把包括回复的截图提交作答；

```
import streamlit as st
import os
from dotenv import load_dotenv

# 加载环境变量
load_dotenv()

# 初始化 OpenAI 客户端
client = OpenAI(
    api_key=,
    base_url=" "
)
```

```
    # 设置页面配置
    st.set_page_config(
        page_title="DeepSeek 聊天助手",
        page_icon=" ",
        layout="wide"
    )

    # 初始化会话状态
    if "messages" not in st.session_state:
        st.session_state.messages = [
            {"role": "system", "content": "你是一个有帮助的AI助手"}
        ]

    # 页面标题
    st.title("  DeepSeek 聊天助手")


    # 显示聊天历史



    # 用户输入
    if prompt := st.chat_input("请输入您的问题"):
        # 添加用户消息到历史
        st.session_state.messages.append({"role": "user", "content":
prompt})
        with st.chat_message("user"):
            st.markdown(prompt)

        # 获取 AI 响应
        with st.chat_message("assistant"):
            with st.spinner("思考中..."):
                response = client.chat.completions.create(
                    model="deepseek-chat",
                    messages=st.session_state.messages,
                    temperature=0.7,
                    max_tokens=1000,
                    stream=False
                )
                ai_response = response.choices[0].message.content
                st.markdown(ai_response)
```

```
        # 添加 AI 响应到历史
        st.session_state.messages.append({"role":       "assistant",
"content": ai_response})

    # 添加清除聊天按钮
    if st.button("清除聊天历史"):
        st.session_state.messages = [
            {"role": "system", "content": "你是一个有帮助的AI助手"}
        ]
```

二、青少年零花钱管理应用开发

1、背景

在当今数字化时代，培养青少年的理财意识和消费习惯变得越来越重要。零花钱是孩子们接触金钱管理的第一步，通过合理记录和分析零花钱的使用情况，可以帮助他们建立正确的消费观念和理财意识。然而，目前市面上针对青少年的零花钱管理工具往往过于复杂或不够直观，无法满足他们的实际需求。

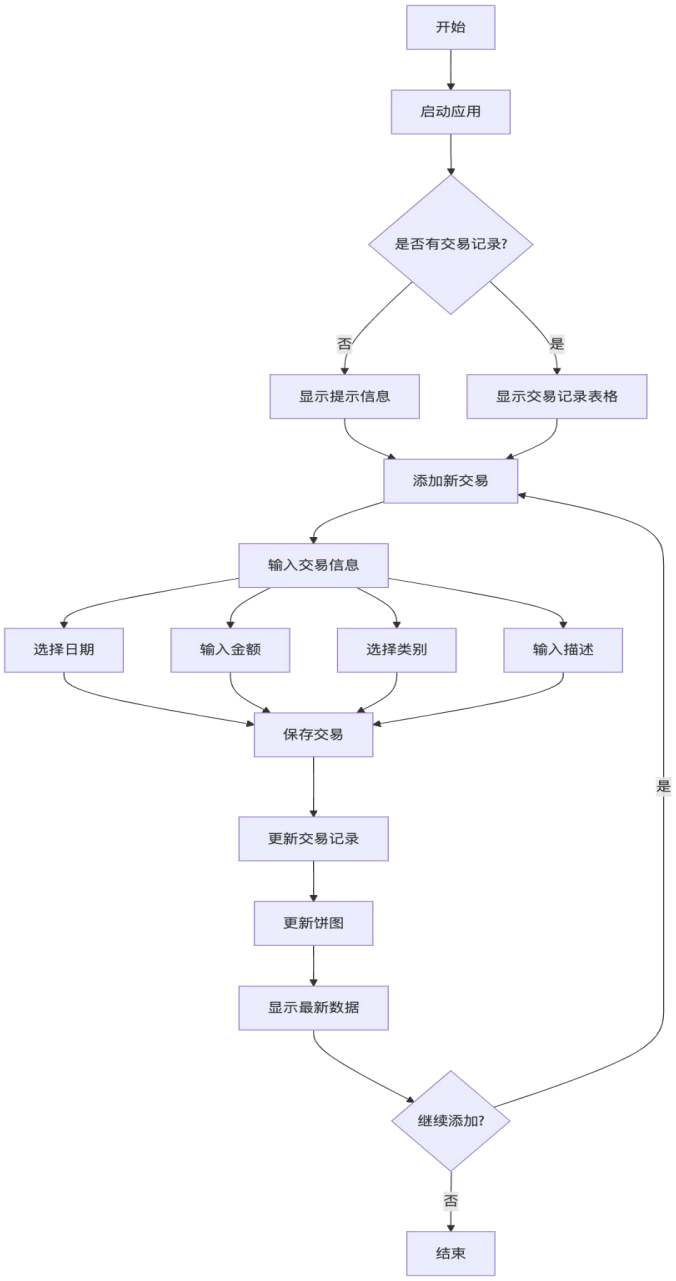
2、需求

请基于Streamlit框架开发一个简单易用的零花钱管理应用，要求：
实现基本的零花钱记录功能，包括：
日期
金额
类别
描述等信息
提供直观的数据可视化展示，帮助用户了解自己的支出分布
界面设计要简洁友好，适合青少年使用
代码结构清晰，注释完整
考虑用户体验，添加适当的交互反馈

3、参考技术栈

前端框架：Streamlit
数据处理：Pandas
数据可视化：Plotly
开发语言：Python

4、零花钱管理应用的流程图：



5、主要功能流程说明：

应用启动流程：

应用启动

检查是否有交易记录

根据检查结果显示相应界面

数据录入流程：

添加新交易

输入交易信息（日期、金额、类别、描述）

保存交易数据

数据展示流程：

更新交易记录表格

更新支出类别分布饼图
显示最新数据
循环流程：
用户可以选择继续添加新交易
或结束使用应用

6、零花钱分析师效果图

7、题目要求：

请你在实操环境的 IDE 中调试下面代码，运行成功后，添加类数据：
1).文具 10 元；
2).零食 10 元；把包括增加的交易记录与支出分析图表（饼图）截图提交；

```
import streamlit as st
import pandas as pd
import plotly.express as px
from datetime import datetime

# 设置页面配置
st.set_page_config(
    page_title="零花钱分析师",
    page_icon=" ",
    layout="wide"
)

# 设置页面标题
st.title("  零花钱分析师")
st.markdown("### 记录你的零花钱使用情况")
```

```
    # 初始化会话状态
    if 'transactions' not in st.session_state:


    # 侧边栏 - 添加新交易
    with st.sidebar:
        st.header("添加新交易")

        date = st. ("日期", datetime.now())
        amount = st ("金额", min_value=0.0, step=0.01)
        category = st.selectbox(
            "类别",
            ["零食", "文具", "玩具", "游戏", "其他"]
        )
        description = st.text_input("描述")

        if st.button("添加交易"):
            new_transaction = {
                "日期": date,
                "金额": amount,
                "类别": category,
                "描述": description
            }
            st.session_state.transactions.append(new_transaction)
           # 主界面
    if st.session_state.transactions:
        # 转换为 DataFrame
        df = pd.DataFrame( )

        # 显示交易记录
        st.subheader("交易记录")
        st.dataframe(df)

        # 可视化
        st.subheader("支出分析图表")

        # 按类别统计
        fig_category = px.pie(
            df,
            values='金额',
            names='类别',
            title='支出类别分布'
```

```
        )
        st.plotly_chart(fig_category)

    else:
        st.info("还没有添加任何交易记录,请在左侧添加你的第一笔交易!
")

    # 添加页脚
    st.markdown("---")

    st.markdown("###    使用提示")

    st.markdown("""
    1. 在左侧添加你的每一笔支出
    2. 查看图表了解你的支出类别分布
    """)
```