2025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） 主观题

以下案例应用题题目请在文件"C:\KS\人工智能-答题纸E卷.docx"中作答!

（一） 简答题(本大题包含2小题，每题5分，共10分)

1.随着人工智能向大模型发展，今年DeepSeek、Qwen、InternLM等开源项目引领了行业的快速发展，请写出大模型生态里面的强化学习、推理引擎、训练框架、Agent等5个优质开源项目。(5分)

请在答题纸作答! 此处答题一律无效!

2. 今年年初DeepSeek的爆火，除了领先的技术以外，还举行了开源周，公开了大量的核心技术，促进了整个行业的突破。持续学习和读懂前沿技术是行业从业的必备能力，开源周有一篇关于DeepEP的介绍（见C:\素材\），请结合附件写出MoE的功能和技术亮点。(5分)

请在答题纸作答! 此处答题一律无效!

（二）基础操作题(本大题包含2小题，每题15分，共30分)

1.在业务实际操作中，需要对环境和服务器进行部署，例如安装基本的PyTorch、MindSpore等AI框架，安装基本的sklearn、transformer、Matplotlib等库包。请回答如下问题:

(1) 请完成如下两个任务，填写在答题纸中:(5分)

①使用pip命令安装最新版的sklearn库;

②检查sklearn库的版本号。

请在答题纸作答! 此处答题一律无效!

(2) 应某企业应用需求，需要在其公司新购置安装Windows 10操作系统的服务器上安装Qwen，目前已安装了Ollama，请接着完成如下安装Qwen（Qwen3:1.7b）的任务，该任务包括三个步骤，请将命令写在答题纸中。(5分)

①创建和安装Qwen模型 ;

②运行模型;

③检查包括Qwen在内的各个模型版本。

请在答题纸作答! 此处答题一律无效!

(3) 在Linux操作系统环境中,日常需要完成Docker容器的日志排查，AI训练脚本的运行等工作。请撰写代码命令满足如下要求。(5分)

场景：AI训练脚本运行报错 Permission denied，模型文件 /data/models/resnet50.pth 无法读取。

要求：

①检查文件权限（仅需显示权限位，如 -rw-r--r--）；

②为 ai_user 用户组添加读写权限（不改变其他用户权限）；

③检查权限修改结果（输出修改后权限）。

请在答题纸作答！ 此处答题一律无效！

2.接下来公司研发部要在服务器上开展AI的研发，例如利用深度神经网络对图像进行训练和分类预测，或者利用预训练模型进行分类预测等等，请补充代码完成以下任务。(15分，每空3分)

（1）在给定的一个预训练ResNet50模型上，对输入的图像（image.jpg）进行目标分类，并输出预测类别。以下为核心代码，请补充空白处。

```
from torchvision import models, transforms
from PIL import Image
import torch

model = models.resnet50(pretrained=True)
model.eval()

transform = transforms.Compose([
    __________①__________, #图像大小统一为 32x32
    transforms.CenterCrop(224),
    __________②__________, #图像转为Tensor
    transforms.Normalize(mean=[0.485, 0.456, 0.406],
                         std=[0.229, 0.224, 0.225])
])

img = Image.open("__________③__________") #打开待处理图像
img_t = transform(img).unsqueeze(0)

with torch.no_grad():
    output = model(img_t)
    _, predicted = torch.max(output, 1)

print(f"Predicted class index: {predicted.item()}")
```

请在答题纸作答！ 此处答题一律无效！

（2）使用 Hugging Face Transformers 库加载 bert-base-chinese 模型，实现中文句子情感分类（正向/负向）。以下是关键流程和代码，请补充空白处。

关键流程：

(1) 加载分词器与模型

(2) 对输入文本进行分词编码

(3) 前向推理并获取 logits

(4)取最大概率类别作为预测结果

关键代码：

```
from transformers import BertTokenizer, BertForSequenceClassification
import torch

tokenizer = BertTokenizer.from_pretrained('bert-base-chinese')
model = BertForSequenceClassification.from_pretrained('bert-base-chinese', num_labels=2)

inputs = tokenizer("这家餐厅真不错", return_tensors="pt")
outputs = model(inputs)
pred = ____________④____________ (outputs.logits, dim=1)
print(pred. ____________⑤____________)  # 0-负向，1-正向
```

请在答题纸作答！ 此处答题一律无效！

（三）实践应用题(本大题包含2小题，每题15分，共30分)

1.近年来，随着人工智能技术在医疗健康领域的快速发展，AI诊断辅助系统已成为提升诊疗效率与准确性的关键工具。某三甲医院计划部署一套AI诊断辅助系统，旨在实现医学影像智能分析、电子病历自然语言处理及跨科室数据安全共享，以支持医生进行精准诊断。该系统需处理CT、MRI等医学影像，理解病历文本内容，并在保障患者隐私的前提下整合多源医疗数据。

你作为项目组的AI架构师，请结合医疗AI的技术体系与实际落地挑战，完成以下问题。

（1）请列举支撑"AI诊断辅助系统"所需的核心AI技术（不少于3项），并简要说明其作用。(5分)

请在答题纸作答！ 此处答题一律无效！

（2）请描述该系统的辅助诊断流程（如：从接收患者影像和病历到生成诊断建议的全过程），并标出关键AI模块或步骤。(5分)

请在答题纸作答！ 此处答题一律无效！

（3）该系统在肺部CT影像分析中出现"误诊肺结节为正常肺部"问题，请分析原因并提出相应的改进策略。(5分)

请在答题纸作答！ 此处答题一律无效！

2.随着普通话的普及，方言正在逐步消失，但是目前老年人在生活中大多数时间仍使用方言。为了方言的保护以为在政务、公共服务等场景上为老人提供更贴心的服务，各地都积极打造独特的方言库和相关数字人。

上海拥有独特的沪语体系，上海大学发布沪语大模型"小沪2.0"，以助力方言保护和政务服务，同时上海大学还将发布"基于沪语模型的创新创业大赛"的项目，鼓励更多有志青年加入到这个项目中来，实现成果转化。

下面我们将利用"小沪2.0"这个基础模型以及RAG技术对养老院的陪伴系统进行改造。基于该背景，完成以下题目。（15分）

（1）请设计一套基于陪伴系统完整的RAG（检索增强生成）系统搭建流程，需重点考虑沪语特性（如方言词汇、语法差异、语音文本转换），明确各环节的具体方案。（5分）

请在答题纸作答！ 此处答题一律无效!

（2） 请写出沪语文本预处理与向量入库的具体操作。（需包含方言词汇预处理、文本分块、沪语适配的 Embedding 模型调用）（5分）

请在答题纸作答！ 此处答题一律无效!

（3）为了服务的便捷性，需针对系统响应时间和准确率进行提升，提出 3 项针对性优化策略，并说明其原理。（5分）

请在答题纸作答！ 此处答题一律无效!

（四）场景设计与行业应用题(本大题包含1小题，每题30分，共30分)

2025世界人工智能大会暨人工智能全球治理高级别会议以"智能时代 同球共济"为主题，设置会议论坛、展览展示、赛事评奖、应用体验、创新孵化五大板块，全景呈现AI技术前沿、产业趋势与全球治理的最新实践。大会聚焦国际化、高端化、年轻化、专业化，着力打造"产业创新策源地、垂类应用示范地、产业生态聚集地、创新创业首选地、治理合作先行地"五大高地，推动大会迈向更高专业度、国际化和创新性。

那随着我们的步伐来探索AI时代的奥秘!

1.首先来到展区，我们首先看到昇腾384超节点Atlas 900 A3 SuperPoD，其通过三大技术优势，入选WAIC 2025镇馆之宝。它通过高速互联总线，突破互联瓶颈，让超节点像一台计算机一样工作。相比传统集群，主要有以下3大优势：

① 超大带宽：超节点内任意两个AI处理器之间通信带宽，相较于传统架构提升15倍，超节点内单跳时延降低10倍。

② 超低时延：昇腾超节点支持全局内存统一编址，具备更高效的内存语义通信能力。通过更低时延指令级内存语义通信，可满足大模型训练/推理中的小包通信需求，提升专家网络小包数据传输及离散随机访存通信效率。昇腾384超节点Atlas 900 SuperPoD是业界唯一突破Decode时延15ms的方案，满足实时深度思考下的用户体验需求。

③超强性能：经过实际测试，在昇腾超节点集群上，LLaMA3等千亿稠密模型训练性能可达传统集群的2.5倍以上。在通信占比更高的Qwen、DeepSeek等多模态、MoE模型上，可以达到3倍以上的提升。

在大模型的分布式训练中，带宽性能一直是一个难点。数据并行（Data Parallelism）和模型并行（Model Parallelism）是大模型分布式训练的两种核心策略，请写出两者在通信需

求上的核心差异。(5分)

请在答题纸作答! 此处答题一律无效!

2.下面我们来到商汤展区。现场有大量可爱的毛绒玩具。相比之前的玩具，依托大模型的能力，让其对话更活灵活现，使孩子多个贴心的玩伴。该产品应用到了大模型垂直领域微调，请结合儿童陪伴玩具的对话模型优化（如儿童口语化、短语化较多等特点），写出模型微调流程和注意点。(5分)

请在答题纸作答! 此处答题一律无效!

3.随着参观的持续，发现现场除了多种多样的大模型应用以外，今年开源的氛围大幅提升，除了DeepSeek、Qwen、InternLM等大模型开源以外，国内AI框架--昇思MindSpore持续开源开放，兼容主流推理引擎vLLM、SGLang等三方生态，也联手启智、书生、魔乐等开源社区纷纷打造多元化活动，为开发者打造最佳贡献成长之路。其中，推理引擎vLLM作为今年开源社区最佳的项目之一，其依托独特的技术，极大的增加了推理性能。请写出vLLM的PagedAttention的核心思想与内存优化机制。(5分)

请在答题纸作答! 此处答题一律无效!

4.现场除了大模型推理以外，MoE模型的比例远超去年同期。请对比DeepSeek-MoE的"细粒度专家"与Mixtral的"标准专家"之间的设计差异，并说明各自的适用场景。(5分)

请在答题纸作答! 此处答题一律无效!

5.来到展览馆二楼，现场聚集了上百家具身智能的企业，如智元、宇树等企业。请问如何设计RL框架，使灵巧手完成"抓取→扫码→放置→打结"的序列任务？ (5分)

请在答题纸作答! 此处答题一律无效!

6.离开世博展览馆，我们来到科学计算（AI4S）相关论坛。无论是去年诺奖的归属还是近年来基于昇思MindSpore打造的商飞"东方.翼风"大模型、南方电网"驭电"大模型荣获WAIC SAIL奖，都表明在通往AGI的路线上离不开AI4S。请说出AI4S大模型和传统NLP大模型的区别。(5分)

请在答题纸作答! 此处答题一律无效!

（五）附加测试题

智能体（Agent）具备自主决策能力、能与环境交互并完成任务的AI系统，是大模型产业落地重要途径之一，掌握其搭建能力也是考量大模型素养提升的重要标志。因此特增设附加测试。

上海，作为全球文化枢纽，每年汇聚海量旅游活动、国际级演唱会及盛大漫展（如上海

国际漫展），吸引数百万游客。这些精彩纷呈的文化盛宴，生动彰显城市创新活力与包容精神，使其成为世界级文化中心地位和中外文化交流的璀璨窗口，持续引领全球文化潮流。

漫展等二次元活动广受青少年的热爱。需要您通过Agent Flow开发平台（智算网址：hpc5000.shu.edu.cn）所提供的平台搭建智能体，可以实现一句"帮我看看最近2个月上海二次元活动的情况"，获得活动信息，报名方式，并生成交通出行提示，天气预告等信息。帮助二次元爱好者更好的参与活动。

具体要求：

1. 创建智能体组件：

新建一个名为"漫展助手"的智能体应用。根据你的分析，在应用中创建出对应的各个组件节点（如：语言交互、交通规划、行程输出、天气提醒等）；

2. 配置各组件的提示词；

3. 连接工作流：

4. 测试与验证：

完成工作流搭建后，使用以下用户请求进行测试：

在会话中输入"帮我看看最近2个月上海二次元活动的情况"。

在测试时，智能体从上述网址中成功获得活动信息，报名方式，并生成交通出行提示，天气预告等信息。

5.需提交的成果

建立Agent.docx文件，包含两张清晰的屏幕截图，要求如下：

① 展示你搭建完成的最终工作流的截图。

② 测试后返回信息的截图。

此文件需要存放在C:\KS\目录下。