全省职工职业技能大赛

湖北省人工智能产业职工技能竞赛

AI 大模型赛项技术文件

2025 年 8 月

目　录

一、命题原则

本届 AI 大模型行业应用与开发技能比赛依据人工智能工程技术人员人工智能平台产品实现方向（职业编码：2-02-10-09）国家职业技能标准中级进行命题。

二、比赛形式

比赛采用团体竞赛形式，每支队伍 2 人。理论考试采用纸质试卷笔试答题，实际操作考试分为两阶段限时任务，在规定时间内完成指定任务，并提交相关说明文档，演示视频，答辩材料。

三、比赛内容

比赛内容包括理论考试和实际操作竞赛两部分。理论考试满分为 100 分，占总成绩的 50%；实际操作比赛满分为 100 分，占总成绩的 50%。

（一）理论考试

理论考试时长 90 分钟，题型涉及单选（30 题）、多选（15 题）及应用简答题（4 题）。

主要考核内容：

1. 大模型基础：Transformer 架构核心（如自注意力机制、位置编码作用等），训练与优化（如常用优化器、防过

拟合策略等），模型能力与特性；

2. 高效微调与适配技术：参数高效微调，微调目标与应用（指令微调目的、领域适应与任务适应），解决微调挑战；

3. 推理优化与部署技术：推理加速，轻量化部署，增强与补充技术；

4. 行业相关政策及大模型相关伦理安全：《生成式 AI 服务管理暂行办法》合规要求、开源模型商用授权限制等、透明性与可问责机制等；

理论考试团队总分排名前 15 名的队伍晋级实际操作竞赛。

（二）实际操作比赛

1.竞赛总体设计

本次竞赛采用“双阶段、双赛道”的创新模式，全面、立体地评估参赛者的综合能力。

阶段一：大模型服务化工程部署

竞赛目标： 本阶段旨在考察参赛者在真实硬件环境下，将一个大语言模型部署为稳定、高效、可控的 API 服务的能力。

核心挑战： 在主办方提供的多 GPU 服务器环境中，成功部署一个百亿参数级别的大语言模型，并实现其服务化封装。

考察能力：（1）环境驾驭能力： 对 Linux 服务器环境、

容器化技术（如 Docker）的熟练运用。（2）模型部署能力：掌握主流大模型推理框架（如 vLLM 等），并能进行性能优化配置。（3）服务封装能力： API 接口的设计、实现与调试，确保服务的稳定性与可用性。

阶段二：垂直领域智能体（Agent）应用创新

竞赛目标：基于第一阶段部署的模型服务，结合特定行业数据，构建一个具有实际应用价值的创新智能体。

核心挑战：参赛者需在“Agent 代码开发”与“Agent 平台构建”两条赛道中任选其一，完成一个从 0 到 1 的智能应用搭建。

考察能力：（1）数据洞察与处理能力： 对垂直领域数据（汽车、法律、医学）的理解、清洗与组织。（2）RAG 技术应用能力： 检索增强生成（RAG）方案的设计与实现，构建高质量的知识库。（3）Agent 架构设计能力： 智能体工作流（Workflow）的设计，实现复杂任务的拆解与执行。（4）场景创新与实用性： 应用创意的独特性、解决实际问题的价值以及商业潜力。

2.竞赛题目与数据资源

（1）竞赛题目概览

题目一（部署任务）：“高性能大模型 API 服务搭建”。要求参赛者在规定时间内，在限制网络访问的情况下，将指定的 Qwen3-14B 模型在双 GPU 卡上完成部署，并提供一

组功能完备、性能达标的 API 调用接口。

题目二（应用任务）：“垂直领域 AI 智能体创新应用”。要求参赛者利用题目一的大模型服务，并结合主办方提供的行业领域数据集，开发一款创新应用。

（2）数据资源

主办方将提供三大垂直领域的脱敏数据集作为创新的基础素材，数据格式涵盖 JSON、CSV、JSONL 等。

四、比赛规则

（一）理论考试

时间：90 分钟，迟到 15 分钟不得入场。

规则：线下集中闭卷笔试，所有选手在统一考场完成纸质试卷答题。严禁携带电子设备、参考资料，考场启用全方位防作弊监控系统，交卷后立即离场，不得滞留或交流试题内容。

（二）实际操作考试

时间：10 小时，迟到 15 分钟不得入场。

规则：

1.统一使用赛场机房终端设备，禁用个人设备。

2.赛务组统一分发(Qwen3-14B)模型及行业数据集(汽车/法律/医学），严禁使用其他模型或外部数据集。

3.开发环境需基于组委会预置的 Docker 镜像。

4.团队双人协作完成实操任务以及准备答辩汇报。

五、评分标准

（一）理论考试题型与分值

1.单选题（30 题×1 分）：答对得分，答错不扣分。

2.多选题（15 题×2 分）：全对得 3 分，漏选得 1 分，错选不得分。

3.简答题（4 题×10 分）：按要点评分，逻辑清晰、步骤完整者满分。

（二）实际操作考试题型与分值

1.大模型服务化工程部署，30 分；

2.垂直领域智能体（Agent）应用创新，70 分；

3.评分规则：

阶段一评审：大模型服务化工程部署阶段，评审将以参赛者在规定时间内提交的部署说明文档和 API 调用演示录屏为核心依据，由评委组进行综合评定；

阶段二评审：采取作品演示与答辩结合的方式。评审团将从以下维度综合打分：

（1）技术实现（40%）：架构设计的合理性、技术难点与完成度。

（2）创新性（30%）：应用场景的独创性、解决问题的思路是否新颖。

（3）实用价值（20%）：应用是否能解决真实痛点，是否具有商业或社会价值。

（4）完成度与交互体验（10%）：应用的核心功能是否完整，界面交互是否流畅。

（三）成绩评定

个人成绩排名：按照团队实际操作考试成绩+个人理论考试成绩综合评定

团队成绩排名：按照实际操作考试成绩评定

备注：成绩排名如有同分情况，依次按照理论成绩用时长短、实操部分“大模型服务化工程部署”所用时长、实操部分“垂直领域智能体（Agent）应用创新”所用时长等情况进行二次排名。

六、硬件环境与部署方法

1.硬件环境标准

资源：为每支参赛队伍提供一台专属的高性能云服务器。

核心配置：服务器搭载 2 张 NVIDIA RTX 4090 GPU，并配备 28 核 CPU 以及 180GB 内存，确保满足百亿级模型部署与运行的需求。

2.部署方法框架

操作系统与基础环境： 提供预装了 Ubuntu 操作系统、NVIDIA 驱动及 CUDA 工具包的基础 docker 环境。

阶段一（模型部署）： 推荐（但不限于）使用容器化技术（Docker）进行环境隔离与部署。主办方将提供一个包含主流推理框架（如 vLLM）依赖的基础镜像的容器，参赛者需在此之上完成模型的加载、并行配置与 API 服务的启动。

阶段二（应用开发）： 参赛者可自由选择技术栈。

代码赛道： 可以使用 Langchain、Qwen-agent、LangGraph 等 agent 开发框架构建 agent 应用。

平台赛道：使用主流低代码 Agent 平台（如 Dify、coze 等），参赛者可在此平台上通过图形化界面构建应用。

资源提供： 竞赛所需的大模型文件和三大领域数据集将预置在服务器的指定目录中，供参赛者直接调用。

3.技术支撑

实操比赛前会统一提供《竞赛环境与资源说明手册》，详细说明服务器登录方式、预置资源路径及基础环境信息。