附件：竞赛实操样题

生成式人工智能系统应用员

竞

赛

样

题

（职工组）

大赛组委会技术工作组

2026年9月

选手须知

1.实操比赛时间240分钟，其中最后15分钟为成果整理、归档与设备恢复时间，其得分计入任务八职业素养与安全规范之中。

2.除有说明外，不限制各任务的作答顺序；建议按任务编号依次完成，遇到长时间训练或批量推理任务时，可先完成其他可独立作答的任务。比赛过程中，若发生危及设备或人身安全的事故，立即停止比赛。

3.选手对比赛过程中需裁判确认的部分，应当先举手示意，等待裁判人员前来处理。

4.请务必阅读各任务的重要提示与提交规范。

5.参赛选手在竞赛过程中，不得使用U盘等移动存储设备，不得访问赛场未开放的外部网络与在线大模型服务。

6.选手在竞赛过程中应遵守相关的规章制度和安全守则，如有违反，则按照相关规定在竞赛的总成绩中扣除相应分值。

7.选手在比赛开始前，应认真对照工位清单检查设备与竞赛平台状态，确认后开始比赛。

8.所有实验应使用统一的输入、参数和评价口径；不得通过手工填写指标、修改输入数据或伪造模型输出替代真实过程。结果报告中的样本数、参数、指标、图表和文字结论必须彼此一致，并能够追溯到输入数据或运行记录。

9.不得删除、覆盖、重命名原始素材、模型权重和题目文件；临时文件应与正式提交成果分开管理。

10.遇到依赖缺失、显存不足、运行中断或结果异常时，应优先保留现场信息，再采用合理的降级或调整方案，并在结果说明中写明原因、调整内容和对结果的影响，不得隐藏失败过程或只保留有利结果。

11.需要裁判验收的各项任务，任务完成后裁判只验收1次，请根据

赛题说明，确认完成后再提请裁判验收。

12.选手严禁携带任何通讯、存储设备及技术资料，如有发现将取消其竞赛资格。选手擅自离开本参赛队工位、与其他工位的选手交流或在赛场大声喧哗，严重影响赛场秩序的，将取消其竞赛资格。

13.选手必须及时保存自己编写的程序及材料，防止意外断电及其他情况造成程序或资料的丢失；离场前应确保计算机处于开机状态。

14.所有成果统一保存至指定的结果目录，按题号分别建立子目录，文件名应清楚表达内容；成果文件中不得出现学校、企业、姓名等与选手身份有关的信息。

15.赛场提供的任何物品，不得带离赛场。

16.任务配分表

序号	一级指标	配分	说明
1	任务一 开发环境构建与模型准备	10	过程评分＋结果评分
2	任务二 数据处理与指令微调预处理	15	结果评分
3	任务三 大语言模型 LoRA 微调训练	20	过程评分＋结果评分
4	任务四 模型评估	10	结果评分
5	任务五 模型推理与提示工程	15	结果评分
6	任务六 模型量化与部署	10	结果评分
7	任务七 检索增强生成（RAG）	15	结果评分
8	任务八 职业素养与安全规范	5	过程评分
合计		100	

任务背景说明

生成式人工智能正在从模型训练和实验验证走向数据治理、知识增强、业务应用和工程部署。一个完整的应用项目不仅要能够生成文本，还要能够在有限资源下稳定运行，能够通过数据和指标证明效果，能够将实验过程整理为可复现、可审查、可交付的成果。

本样题以一个端到端生成式人工智能应用项目为载体，要求选手综合运用环境准备、数据预处理、参数高效微调、推理对比、提示工程、检索增强、综合评估、量化与并发部署等能力，完成从资源准备到部署分析的完整流程。

作答要点	执行说明
作答顺序	建议按照任务编号完成；遇到长时间训练或推理任务时，可先完成其他可独立作答的任务。
过程留痕	保留关键参数、运行日志、样本统计、指标结果和异常说明，使成果能够复核。
文件管理	按任务分别建立结果目录，文件名应清楚表达内容，避免覆盖原始数据和模型资源。
结果表达	图表、表格、文字结论和实际运行结果必须相互一致，结论应有数据依据。
异常处理	出现资源不足、依赖缺失或运行中断时，先记录现象，再说明处理方式和对结果的影响。

任务一 开发环境构建与模型准备

情境：生成式人工智能应用运行环境搭建

任务背景：生成式人工智能应用需要稳定的运行环境和完整的模型资源。考生首先要确认 Python、深度学习框架、计算设备和项目依赖可用，并验证主模型与文本向量模型能够被正确加载。

任务目标：完成运行环境检查、依赖验证、模型资源确认和基础推理验证，形成环境准备结果。

输入材料：

项目运行环境及依赖安装介质；主语言模型和文本向量模型的本地资源；项目目录中的资源配置说明。

一.运行环境与依赖核验

1. 检查 Python 版本、操作系统、深度学习框架版本、计算设备状态及可用显存。

2. 验证模型加载、参数高效训练、数据处理、向量检索和可视化相关依赖。

3. 检查主模型和向量模型的关键配置文件，确认资源完整可用。

二.模型资源验证

检查主模型和向量模型的关键配置文件，确认资源完整可用，并记录模型结构关键参数、向量维度与最大输入长度。

1. 完成主模型加载和短文本生成验证，记录模型状态、设备信息和基础输出。

三.基础推理与资源释放

完成资源释放，确保后续任务可在干净状态下继续执行。

绘制环境信息图（含依赖版本与显存占用），保存为图片文件；

四.提交成果

环境与依赖检查结果（结构化文件）；模型资源验证记录；基础推理结果与环境信息图；任务一成果说明。

Python	PyTorch	CUDA 可用	GPU	GPU 显存 (GB)	平台
3.11.13	2.2.2+cu121	可用	NVIDIA TITAN Xp	5.9	Linux

评价关注点：环境信息是否完整，模型资源是否核验，基础推理是否真实完成，输出材料是否清晰可复核。

任务二 数据处理与指令微调预处理

情境：从对话样本构建训练数据

任务背景：高质量训练数据是生成式模型应用的基础。考生需要将结

构化对话数据转换为模型训练所需的输入张量，并通过长度分析和标签设计降低无效训练与信息截断风险。

任务目标：完成数据字段分析、质量检查、长度分布分析、分词处理和监督微调标签构造。

输入材料：

项目提供的领域对话数据、分词器和预处理工具；非结构化客服对话文本、标准指令微调字段要求及原有领域训练数据。

一.数据探索与质量检查

1. 统计数据规模、字段名称、字段类型及代表性样本，说明各字段在训练中的作用。

2. 检查缺失值、空字符串、空白字符和异常样本，删除目标回答为空的无效记录。

二.长度分析与最大序列长度确定

1. 统计指令、回答和组合文本的长度分布，使用分位数方法确定合理的最大序列长度并说明依据。

2. 将指令和回答组织为统一对话格式，完成分词、有效位标记和序列截断处理。

3. 构造监督微调标签：指令部分不参与损失计算，回答部分作为学习目标；通过样本检查确认标签位置正确。

三.指令微调样本构造

1. 完成批量预处理，保留字段结构、样本数量和长度统计。

四.客服数据清洗与联合训练集构建

补充处理非结构化客服对话，将业务文本转换为可训练的统一指令微调样本，并与原有训练数据按规则合并。

输入材料：非结构化客服对话文本；标准指令微调 JSONL 字段要

求；原有领域训练数据及合并规则。

1. 识别客户与客服角色、对话分隔方式和可能出现的多种标记形式。

2. 解析问答对，清除表情符号、连续特殊字符、重复标点和多余空白。

3. 按问题去重，设置合理的最小/最大长度阈值并说明过滤依据。

4. 将清洗问答转换为统一的指令、输入和输出字段，逐行输出为 JSONL。

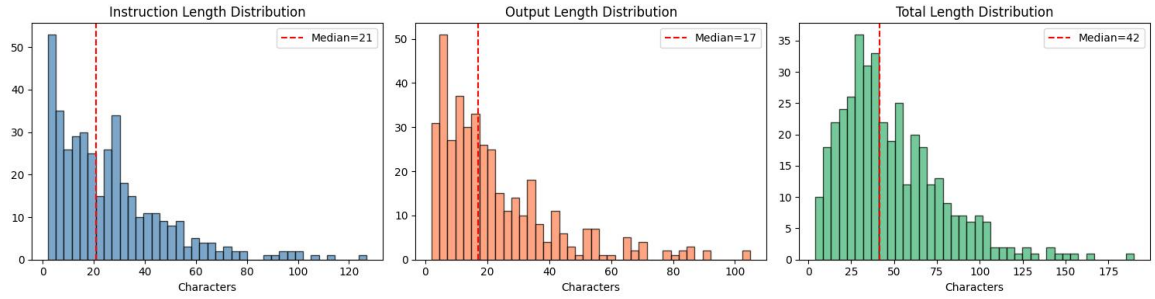
5. 逐行校验 JSONL，统计总行数、有效行数、字段完整率和错误类型。

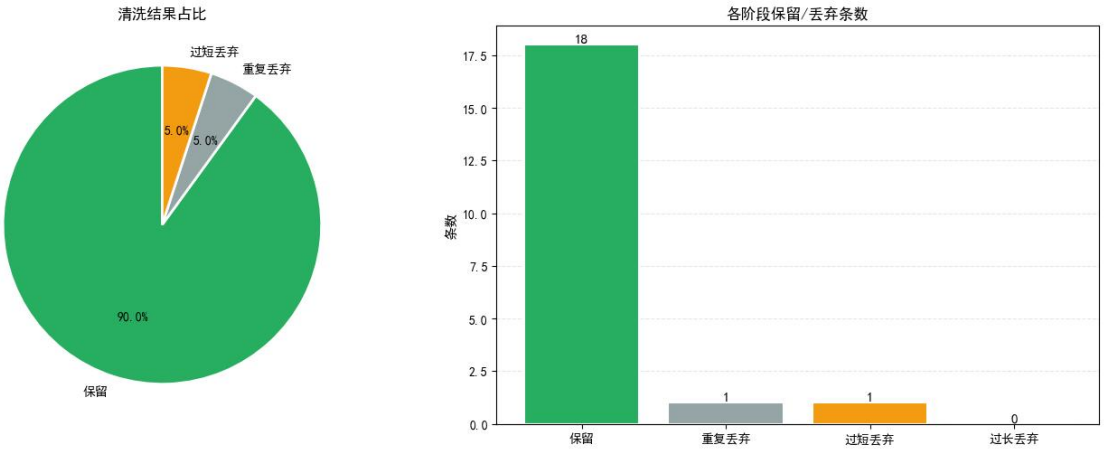
6. 按指定规则与原有训练数据合并，避免字段不一致和数据污染。

7. 生成清洗前后数量变化、长度变化和噪声处理效果图。

五. 提交成果

数据质量统计结果；长度分布图；预处理样本结构及标签检查记录；批量预处理结果说明；清洗后的客服 JSONL 数据；格式校验报告；清洗流水线统计和可视化图；合并训练集及数据质量说明。





评价关注点：数据清洗规则是否合理，最大长度选择是否有依据，标签掩码是否正确，JSONL 是否合法，合并数据是否可训练，处理结果能否直接供后续训练使用。

任务三 大语言模型 LoRA 微调训练

任务背景：在有限显存条件下，需要让通用语言模型适应特定对话风格。考生应使用参数高效微调方法，仅更新少量适配参数，同时通过训练记录判断模型是否稳定收敛。

任务目标：完成量化基座准备、适配器配置、训练参数设置、模型训练、权重保存和训练过程分析。

一. 基座加载与适配器配置

1. 采用适合设备资源的基座加载方式，说明低比特加载、梯度检查点和梯度累积等策略的作用。

2. 配置适配器的秩、缩放系数、随机失活比例及目标层范围，说明参数选择理由。

二. 训练超参数设置与训练执行

1. 设置训练轮数、有效批量、学习率、日志频率、保存策略和随机种子。

2. 启动训练并记录训练损失、学习率、梯度范数、训练步数和运行时间。

三. 产物保存与训练过程分析

保存可供后续加载的适配器权重、分词器配置和训练相关配置。

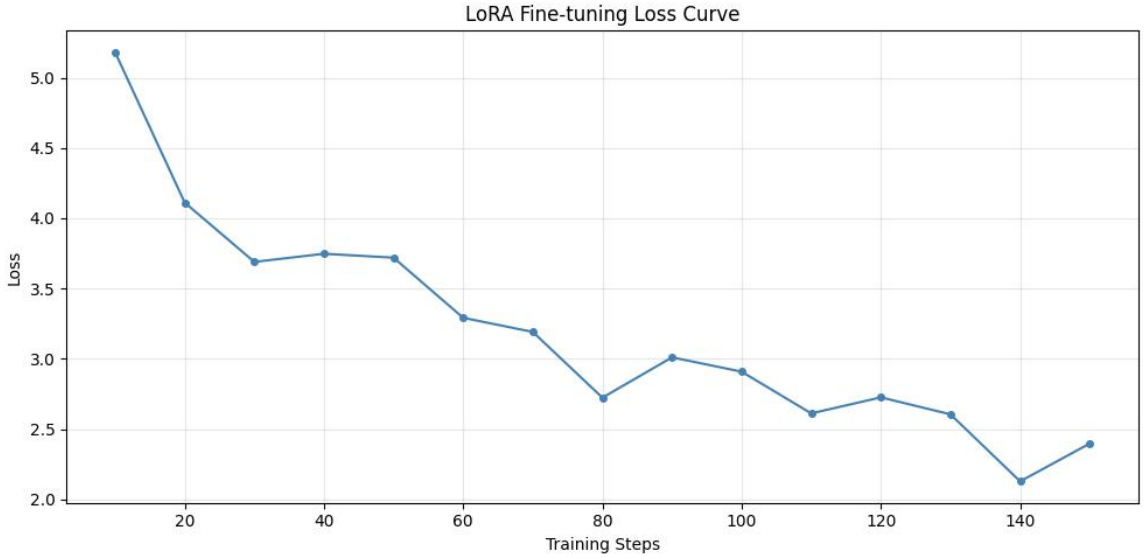
保存训练历史，并绘制训练损失曲线、学习率曲线与梯度范数监控图；分析收敛情况，判断是否存在欠拟合或过拟合迹象，说明判断依据。

四. 微调后推理验证

使用不少于 3 个代表性问题进行微调后推理验证，比较输出是否体现目标数据的语言风格，记录输入、输出与生成参数。

五. 提交成果

适配器权重及相关配置；训练历史与损失/学习率/梯度监控图；微调后推理测试结果；训练参数与结果说明。



评价关注点：训练配置是否合理，参数高效微调是否正确，训练曲线是否真实完整，模型成果是否可加载复用。

任务四 模型评估

任务背景：完整系统需要同时接受自动指标和人工检视。考生应构造兼顾风格与知识的评测集，对比微调模型与检索增强模型，并形成可读、可复核的综合报告。

任务目标：完成评测集构建、批量生成、文本指标计算、分组统计、

人工检视和综合报告生成。

一.评测集构建

1. 将评测样本划分为风格类和知识类，保证评测数据与训练数据有明确边界。

二.批量生成与指标计算

1. 对每条样本生成微调模型回答和检索增强回答，保留问题、参考答案和两种结果。

2. 计算 BLEU 系列和 ROUGE 系列指标，并按评测类型和模型模式分别汇总平均值。

3. 对典型样本进行人工检视，指出自动指标与实际体验可能不一致的情况。

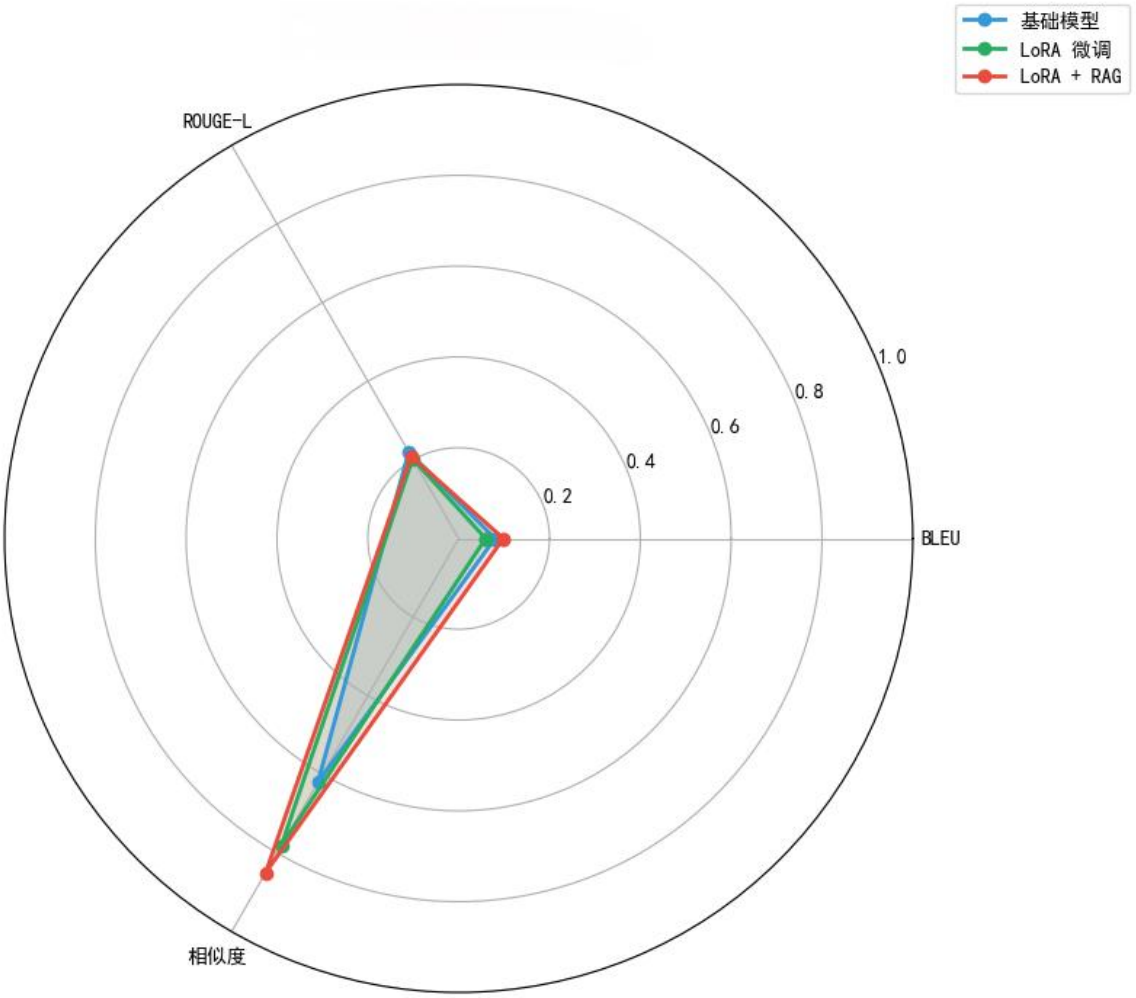
三.人工检视与综合报告

1. 生成 Markdown 或 HTML 综合报告，至少包含实验配置、结果表、典型样本、图表和结论。

2. 区分风格能力、知识准确性、检索贡献和可能的失败模式。

四.提交成果

完整评测原始数据；BLEU/ROUGE 汇总表；典型样本人工检视记录；Markdown/HTML 综合评估报告及图表。



评价关注点：评测集设计是否合理，指标计算是否统一，报告是否可复核，结论是否避免以单一指标替代全面评价。

任务五 模型推理与提示工程

任务背景：模型训练完成后，需要通过统一测试集比较基础模型和微调模型的输出差异。考生应控制变量、统一生成参数，并用表格和图形呈现实验结论。

任务目标：完成测试集设计、双模型推理、输出对比、结果记录和效果分析。

本任务同时覆盖双模型推理效果对比，以及零样本、少样本、分步思考和结构化输出等提示策略验证。

一. 基座模型与微调模型推理效果对比

1. 设计覆盖目标风格、常识问答、开放问答和边界场景的测试问题。

2. 使用相同问题、输入模板和生成约束，分别获得基础模型和微调模型回答。

3. 记录问题、两种回答、生成参数和异常情况，形成逐条对比表。

4. 从语言风格、回答完整性、事实一致性、重复性和安全性等维度分析差异。

5. 生成对比图和结构化报告，结论必须引用实际输出，不得只写预设结论。

二. 提示工程与结构化输出控制

1. 为同一业务问题分别设计零样本和少样本提示，比较分类结果的一致性。

2. 针对需要推理的任务设计分步思考提示，检查最终结论和过程表达。

3. 设计结构化输出要求，明确字段、类型、取值范围和不得输出的额外内容。

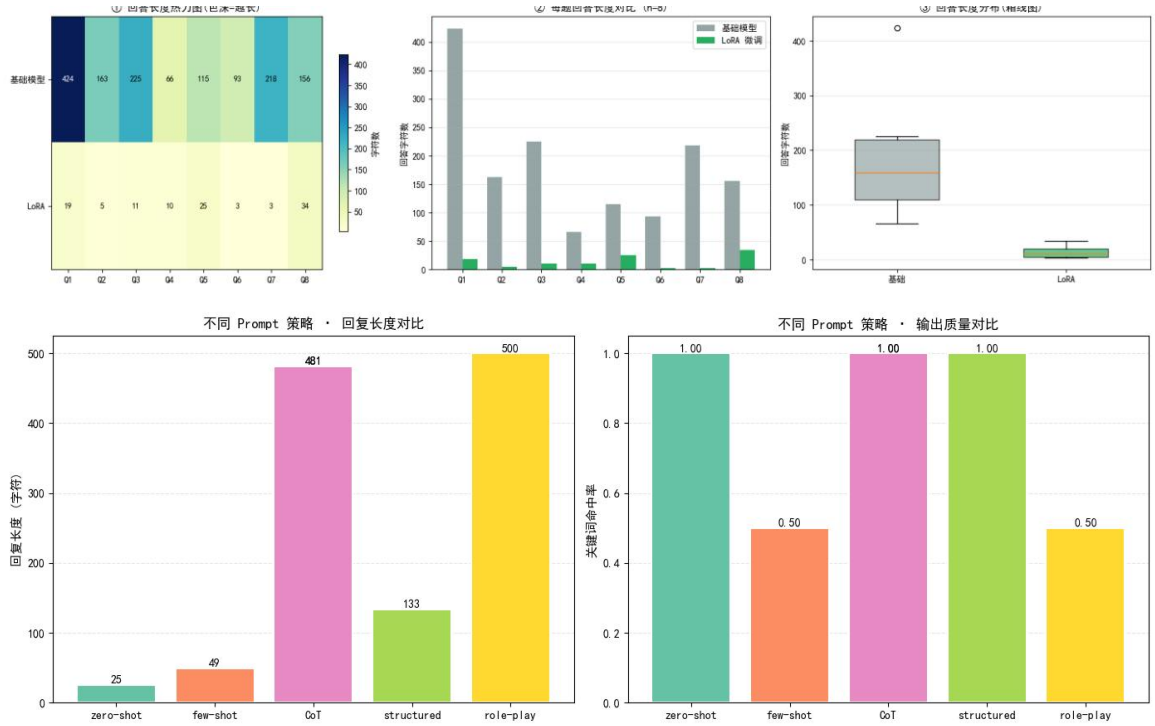
4. 对模型结果进行格式解析，统计结构化输出成功率并记录失败样本。

5. 统一测试不同策略的输出长度、关键词命中情况、结构化成功率和可读性。

6. 形成提示策略选择建议，说明不同场景的适用范围和风险。

三. 提交成果

双模型推理记录与逐条对比表；推理效果对比图与评分统计；各类提示策略说明与典型输入输出记录；结构化解析结果及成功率统计；策略对比图、原始数据和结论说明。



评价关注点：测试集是否有代表性，变量控制是否一致，提示是否清晰可执行，结构化约束是否有效，比较指标是否与任务目标一致。

任务六 模型量化与部署

任务背景：生成式模型部署时不仅要关注单次回答质量，还要关注单位时间内可处理的请求数量。考生需要在同一模型和同一请求集下，对比串行、多线程和批量推理方案。

任务目标：完成统一请求集构造、三种推理模式基准测试、不同精度量化测试、吞吐/延迟统计、批大小扫描和部署建议。

一.并发推理与吞吐优化

在同一模型和同一请求集下，对比串行、多线程和批量推理方案，分析吞吐、延迟、批大小和资源约束。

1. 构造覆盖多个业务类别的统一请求集，保持各模式输入一致。

2. 完成串行推理，记录请求数、总耗时、平均延迟、吞吐量和回答摘要。

3. 完成多线程推理，说明线程并发与单设备实际计算之间的关系，

记录相同指标。

4. 完成批量推理，处理批量填充、批大小和生成结果对应关系，记录相同指标。

5. 计算 QPS 或等价吞吐指标，生成三种模式对比图。

6. 扫描多个批大小，分析吞吐、显存和稳定性变化，提出部署建议。

二.不同精度量化对比测试

1. 以统一提示和生成参数完成半精度基线测试，记录平均延迟、生成速度、峰值显存和示例输出。

2. 完成 8-bit 测试，使用相同指标并计算相对于基线的显存变化。

3. 完成 4-bit NF4 测试，说明双重量化和计算精度配置的作用。

4. 比较三种精度的显存、速度、输出长度和样例质量，避免仅根据理论比例下结论。

5. 生成显存/速度对比图和节省率统计，说明不同硬件条件下的选择建议。

6. 结合四位基座与适配器训练方式，说明参数高效微调的适用条件及潜在风险。

三.部署成果与清单

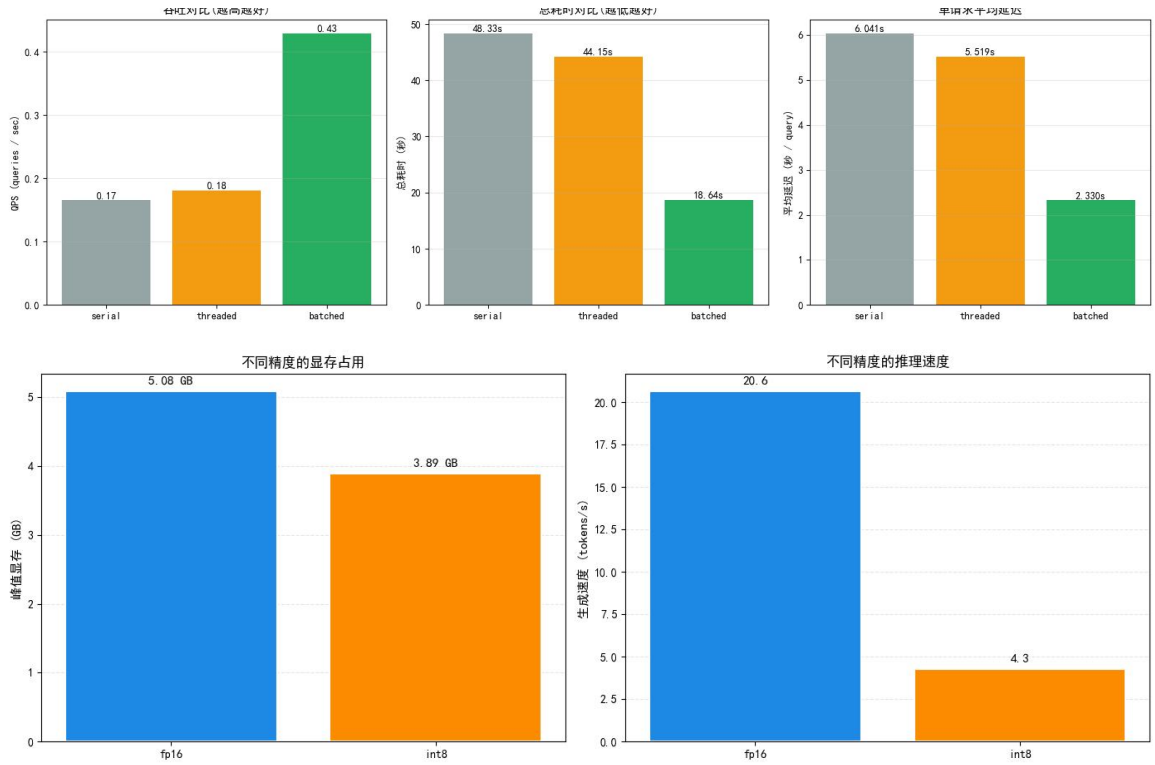
整理并保存基准测试、批大小扫描、量化测试和部署配置，生成 manifest 成果清单。

记录运行清理、显存释放和异常处理过程，给出不同硬件条件下的部署建议。

四.提交成果

三种精度的原始测试结果；串行、多线程和批量推理基准测试结果；QPS、总耗时、平均延迟及显存/速度对比图；批大小扫描结果；合并结果

数据与量化配置说明；成果清单（manifest）；部署权衡分析、参数高效微调拓展说明及运行清理记录。



评价关注点：测试变量是否统一，吞吐与延迟计算是否正确，批量生成结果是否对应，量化配置是否符合要求，显存/速度/质量数据是否完整，结论是否体现实际资源权衡。

任务七 检索增强生成（RAG）

任务背景：单纯依靠模型参数回答专业或领域问题，容易出现知识遗漏和事实编造。考生需要构建检索增强生成流程，把外部知识库内容检索出来并组织到生成上下文中。

任务目标：完成知识库加载、文本切分、向量化、相似度检索、上下文组装和端到端问答验证。

输入材料：

领域知识库文本、文本向量模型、已完成的生成模型或微调模型，以及检索结果和问答结果保存模板。

一.知识库加载与切分

1. 读取并检查知识库文本，说明文本切分粒度、重叠策略和切分结果数量。

二. 向量化与索引构建

1. 使用向量模型将知识片段转换为向量，建立可查询的向量存储结构。

三. 语义检索与上下文组装

1. 设计查询问题，检查检索结果的相关性、排序和返回数量，保留典型样例。

2. 将检索知识片段与用户问题组织为生成上下文，完成端到端回答。

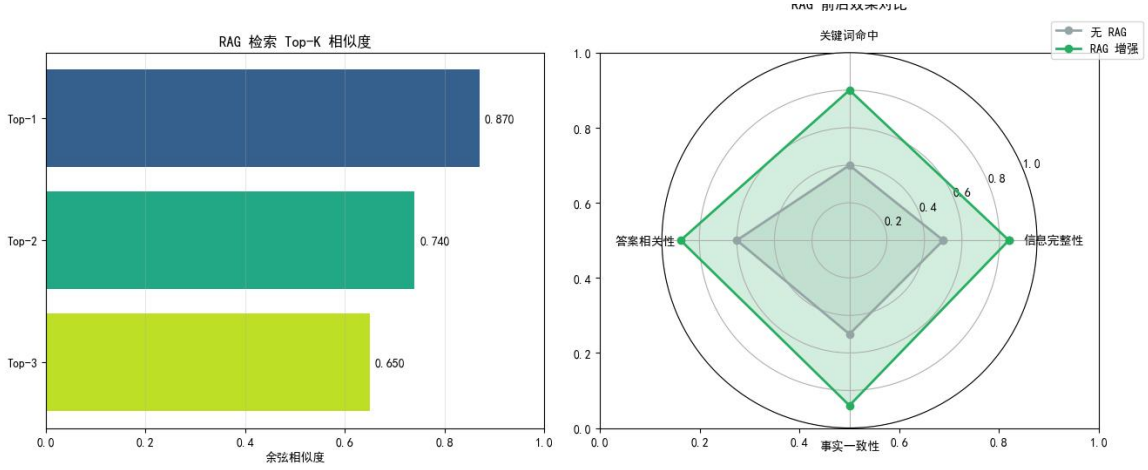
四. 效果对比与成果保存

1. 比较纯模型回答和检索增强回答，分析检索对事实性、完整性和可解释性的影响。

2. 保存知识库处理信息、检索记录、问答结果和可视化材料。

五. 提交成果

知识库切分与索引结果说明；检索样例及相关性分析；检索增强问答记录；纯模型与检索增强对比图及报告。



评价关注点：知识库处理是否规范，检索结果是否相关，上下文是否正确传递，RAG 对比结论是否有数据支撑。

任务八 职业素养与安全规范

一.设备与环境恢复

选手根据赛题要求完成计算资源释放、临时文件清理和工位环境恢复（以实际赛题为准）。

二.成果归档与提交规范

按任务编号建立结果目录，文件名称、目录层级和格式符合项目要求；

报告中的参数、样本数量、指标和图表必须能够从运行结果或原始数据追溯得到；

提交前完成目录核对、文件可读性检查、报告与图表一致性检查和结果复现检查；

成果材料中不得出现学校、企业、姓名等与选手身份有关的信息。

三.安全意识与职业素养

严格遵循相关职业素养要求及安全规范，安全文明参赛；操作规范；工位整洁；着装规范；资料归档完整；诚信守法，尊重知识产权，遵守数据安全与生成式人工智能合规要求。

四.统一提交与过程规范

每项任务提交规定的结果文件、图表和说明材料，文件名称、目录层级和格式应符合项目要求。

说明材料应包含实验目的、处理过程、关键参数、结果摘要、问题分析和结论，不以零散截图代替完整说明。

图表应有标题、坐标或指标名称、必要的图例和结论说明；表格中的数值应保留合理精度并标明单位。

不得修改原始数据、模型权重和题目材料来制造预期结果，不得用手工填写的表格替代算法输出。

运行过程中合理管理显存、内存和临时文件；出现异常时保留错误信息并说明处理过程。

涉及个人信息、人脸数据或业务文本时，仅使用题目提供的材料并做好数据保护。

五.提交成果

设备恢复与归档确认记录；任务目录、文件可读性、报告与图表一致性及结果复现检查记录。

评价关注点：设备与环境是否恢复，成果归档是否完整，数据与结论是否可追溯，安全规范、职业素养和诚信要求是否落实。

任务编号	主要提交成果	格式要求
任务一	环境依赖清单、模型验证记录、基础推理结果、环境信息图	JSON/MD/PNG
任务二	数据质量统计、长度分布图、预处理数据集及标签检查记录	JSON/JSONL/PNG/MD
任务三	适配器权重及配置、训练历史、损失/学习率/梯度曲线、推理验证记录	权重目录/JSON/PNG/MD
任务四	评测原始数据、BLEU/ROUGE 汇总表、人工检视记录、综合评估报告	CSV/JSON/MD 或 HTML/PNG
任务五	双模型对比表与对比图、提示策略说明、结构化解析结果与成功率统计	CSV/JSON/PNG/MD
任务六	各精度测试结果、对比图、manifest 清单、权衡分析	JSON/CSV/PNG/MD
任务七	知识库切分与索引说明、检索记录、RAG 问答结果、对比图与报告	JSON/PNG/MD
任务八	设备恢复与归档确认记录	现场评分