

蓝桥杯全国大学生软件和信息技术大赛组委会

第十八届蓝桥杯全国大学生软件和信息技术大赛

人工智能赛（AI 模型应用开发科目）

竞赛规则及说明

一、参赛对象

具有正式全日制学籍并且符合相关科目报名要求的普通本科院校、高职高专院校学生（以报名时状态为准），以个人为单位参加比赛。

每名选手可配置一名指导教师，每名选手的指导教师最多一名，同一名指导教师可指导多名选手。全国选拔赛和全国总决赛比赛后，指导教师原则上不能更改。

二、组别设置

该科目设置大学组。

三、竞赛赛制

本届大赛采用校赛、全国选拔赛、全国总决赛三级竞赛体系。

校赛：由参赛院校自行组织并进行评审。

全国选拔赛：由大赛组委会统一组织，比赛时长为 4 小时。

全国总决赛：由大赛组委会统一组织，比赛时长为 4 小时。

详细赛程安排以大赛组委会公布的信息为准。

四、竞赛形式

1. 全国选拔赛、全国总决赛均采用封闭、限时方式举办。选手以个人为单位，独立进行作答。

2. 采取线上比赛的形式。选手需访问互联网，通过大赛组委会指定的比赛系统进行答题。

3. 选手将答案提交到比赛系统中，超过比赛时间将无法提交。

五、参赛选手机器环境

1. 在线编程环境（组委会提供统一开发环境，无需选手自行安装）

序号	环境项目	统一配置或说明
1	操作系统	Ubuntu 22.04 LTS 或组委会当届公布的统一 Linux 镜像

2	开发界面	浏览器版 Visual Studio Code 或同等在线开发界面
3	Python	Python 3.10，具体补丁版本以镜像清单为准
4	计算资源	基础配额为 2 核 CPU、4 GB 内存；基础题不依赖 GPU
5	AI 技术栈	每场比赛从 PaddlePaddle、MindSpore、ModelScope 支持的国产 AI 生态中确定一套官方基线技术栈
6	Web 服务	FastAPI、Uvicorn、Pydantic 或赛题指定的同类组件
7	数据处理	NumPy、Pandas、PyYAML、JSON 等常用组件
8	知识检索	FAISS 或统一镜像中提供的同类本地检索组件
9	测试工具	Pytest、HTTP 客户端、本地自检脚本和接口 Schema 检查工具

同一场次的比赛镜像、基础模型、向量模型、依赖版本、资源配额保持一致。具体软件版本、模型校验值和镜像校验值以比赛系统发布的环境为准。

竞赛配套在线编程环境体验（<https://www.lanqiao.cn/courses/67125>）。

2. 环境限制

1. 禁止下载外部模型、安装未经允许的软件包或调用远程服务。
2. 基础模型、向量模型和正式评分数据由平台只读挂载，选手不得替换、修改或重新提交。
3. 服务启动时间、单次请求时限、CPU、内存、并发数和提交包大小均按题面规定执行。

六、赛题形式

赛题为场景实战题（AI 模型应用开发项目实操）。每场比赛设置一个统一业务背景，具体业务场景、任务数量和技术资源以正式比赛赛题为准。

赛题提供一个已经预训练、能够基础运行但尚不能完全满足业务要求的轻量 AI 模型，以及必要的向量模型、业务资料、项目模板、接口规范、公开测试样例等。选手依据任务说明持续完善同一个项目，最终形成可运行的 AI 应用。

赛题按照企业 AI 应用落地过程设置五个阶段：

1. 模型部署与接口调用。
2. 模型配置、Prompt 与参数优化。
3. 数据增强、RAG 与模型适配。
4. 轻量级推理优化。
5. AI 应用开发与工程规范。

所有题目均提供完整题面 PDF 文档及必要资源包。题面文档将说明业务背景、任务目标、文件目录、接口协议、提交格式。选手应认真阅读题意，结合所给资源完成代码、配置、Prompt、索引和应用开发。

特别说明：在无明确说明的情况下，不得随意更改基础资源中的文件名称、文件夹名称、服务端口、接口路径及其存放结构，否则可能影响得分。

Prompt、RAG、规则、数据增强和轻量适配均可作为合理的模型应用优化路线。LoRA 等轻量适配不设为默认必做项，不单独按照训练轮次、Loss 曲线或是否提交微调权重给分。评分以最终业务效果、性能和工程交付结果为准。

七、比赛范围

本次比赛主要考查参赛者利用已有 AI 模型解决实际业务问题的工程应用能力，要求选手在指定比赛环境内完成以下工作任务：

1. 模型部署：环境检查、模型加载、设备与精度配置、服务启动、健康检查和基础推理调用。
2. 模型配置与参数优化：系统 Prompt、任务模板、生成参数、输出约束、指令遵循、回答边界和结构化输出。
3. 数据增强与模型适配：业务文档解析、清洗、切分、索引、检索增强生成（RAG）、来源标注、同义问法、口语表达、错别字和规则适配。
4. 轻量级模型优化：缓存、批处理、量化、检索复用、生成长度控制、连接复用和合理并发，在效果达到最低阈值的前提下降低时延与资源占用。
5. AI 应用开发：标准 API、简单 Web 页面、业务流程集成、异常输入处理、合理拒答、配置管理、运行日志和统一打包交付。

赛题不以从零训练大模型、大规模参数训练、原创算法研究、底层算子开发、推理引擎研发或论文复现为主要考查内容。

八、答案提交

选手只有在比赛时间内提交的答案内容可以用于评测，比赛结束后的任何提交均无效。选手应使用比赛指定网页提交答案，其他方式的提交不作为评测依据。

选手在比赛系统中可重复提交，以最后一次有效提交为准。比赛过程中不向选手显示隐藏测试结果和正式得分，选手应使用公开样例和本地自检工具自行调试。

本科目只提交一个统一项目包。参考目录结构如下，正式比赛以题面要求为准：

```
submission/
manifest.json
app.py
src/
config/
  model.yaml
  generation.yaml
prompts/
  system.txt
artifacts/
  index/
  document_manifest.json
web/
logs/
  env_check.json
  app.jsonl
README.md
```

manifest.json 用于声明服务启动入口、端口、接口路径、使用框架和各成果文件位置。基础模型、向量模型和隐藏测试数据不包含在提交包中。

选手须仔细阅读并严格遵守赛题指定的目录、文件名、编码、接口和数据格式。提交包不得包含外部模型、密钥、远程服务地址、未经允许的可执行文件或指向提交目录之外的软链接。

平台在提交时自动检查目录、文件大小、危险文件、软链接和清单完整性，并生成文件校验值和提交回执。

九、学习资源及样题

1. 赛项资源

免费开放供参赛选手备赛使用，包含模拟题、真题解析、课程等资料。链接：<https://www.lanqiao.cn/paths/3361/>。

2. 样题

样题详见文档附录。

样题仅用于说明本科目的完整题目形态、任务要求、提交规范。正式比赛会更换业务场景、模型、数据、公开样例和性能阈值。

十、评分

1. 试题分值及任务分布

模块	分值
模块一：模型部署与接口调用	20 分
模块二：模型配置、Prompt 与参数优化	15 分
模块三：数据增强、RAG 与模型适配	25 分
模块四：轻量级推理优化	15 分

模块五：AI 应用开发与工程规范	25 分
合计	100 分

2. 题型设计及评分方法

试题按照以上五个模块进行命题。每个模块设置独立评分点和部分得分条件，单条测试用例错误或超时只影响对应评分点，不影响整份作品评测。

（1）运行状态类：评分系统按照 manifest.json 在隔离容器中启动选手服务，检查进程、端口、模型加载状态、启动时限和健康检查结果。

（2）接口协议类：评分系统向规定接口发送公开或隐藏请求，检查 HTTP 状态码、JSON Schema、字段类型、字段约束、批量顺序和错误响应。

（3）业务效果类：评分系统使用隐藏测试检查关键事实、数值条件、来源文档、输出结构、同义改写、错别字、组合条件和无答案拒答。生成式答案以关键事实和来源为主，不采用单一全文完全匹配，也不调用外部大模型进行裁判。

业务答案评分原则如下：

$$\begin{aligned} \text{业务答案得分} = & \\ & 0.50 * \text{关键事实槽位得分} \\ & + 0.20 * \text{来源文档命中得分} \\ & + 0.15 * \text{输出结构得分} \\ & + 0.10 * \text{关键词与数值条件得分} \\ & + 0.05 * \text{本地语义相似度辅助得分} \end{aligned}$$

语义相似度使用组委会统一提供的本地模型，仅作参考。

（4）性能类：评分系统在统一预热和资源配额下采集平均响应时延、P95 响应时延、峰值内存和并发成功率。性能得分以业务效果达到题面规定的最低阈值为前提；返回空结果、固定答案或关闭必要检索所获得的性能提升不计分。

（5）工程结构类：评分系统自动检查目录、配置、日志、页面固定元素、编码、绝对路径、危险文件和提交清单。

服务启动失败时，评分系统仍可检查 Prompt、配置、文档清单、索引元数据和日志格式等可独立验证的静态成果。页面失败不影响已经通过的 API 和业务效果得分；性能项失败不影响已经取得的功能得分。

十一、奖项设置及评选办法

1. 全国选拔赛

全国选拔赛设立一、二、三等奖，原则上各奖项的获奖比例为 10%、15%、25%，总获奖比例不超过 50%。获奖比例仅作为参考，组委会将根据赛题难易程度及整体答题情况，制定

各奖项获奖最低分数线，未达到获奖最低分数线者不得奖。全国选拔赛一等奖选手获得全国总决赛参赛资格。

2. 全国总决赛

全国总决赛设立一、二、三等奖，原则上各奖项的获奖比例为 10%、25%、40%，总获奖比例不超过 75%。获奖比例仅作为参考，组委会将根据赛题难易程度及整体答题情况，制定各奖项获奖最低分数线，未达到获奖最低分数线者不得奖。

十二、奖项公示

全国选拔赛及全国总决赛评审完成后，大赛组委会将在大赛官网开放奖项公示。参赛选手若对奖项有异议，可在 3 个工作日内按照大赛组委会相关要求提出复核申请。

十三、监督反馈

详见《蓝桥杯全国大学生软件和信息技术大赛章程》。

十四、其他注意事项

1. 选手必须符合参赛资格，不得弄虚作假。资格审查中一旦发现问题，取消其报名资格；竞赛过程中发现问题，取消竞赛资格；竞赛后发现问题，取消竞赛奖项，收回获奖证书及奖品，并按大赛规定处理。

2. 参赛选手应严格遵守蓝桥杯大赛诚信参赛管理办法（办法链接：<https://dasai.lanqiao.cn/notices/844/>），服从大赛组委会的指挥和安排，爱护竞赛场地的设备。未尽事宜请参照组委会在大赛官方网站公布的通知、章程、比赛管理办法及相关要求并遵照执行。

3. 选手不允许调用外部商业模型 API、远程计算服务或本机以外的资源，不允许使用 U 盘等外接存储设备导入资料。

4. 赛项全部技术评分点采用机器自动评分。选手应特别注意答案项目包的目录、文件名、接口和内容格式，不得随意添加影响评分系统运行的内容。

5. 禁止通过固定问题匹配、隐藏信息探测、评分接口攻击或其他方式规避真实模型调用。隐藏测试将使用同义改写、错别字、数值变化和组合条件识别硬编码。

6. 禁止替换组委会提供的基础模型和向量模型，禁止使用外部模型、外部 API、远程计算资源或未经允许的软件。

7. 通过比赛系统发放赛题、基础模型、项目模板、业务资料和公开测试样例，并回收选手提交的答案项目包。

8. 赛题采用统一业务场景下的项目制任务。选手在同一个项目中依次完成模型部署、配置与参数优化、数据增强与模型适配、轻量级模型优化和 AI 应用开发，最终交付一个可运行、可调用、可测试的 AI 应用。

9. 选手可自主安排各任务的完成顺序，但所有任务成果须集成在同一服务入口、同一目录和同一提交包中。

10. 选手在比赛时间内可重复提交，以截止时间前最后一次有效提交作为正式评分版本。超过比赛时间将无法提交。



附录【样题仅作题型参考，不提供配套数据、评分】

竞赛说明（样题）

一、题目背景

“云学在线教育平台”为本样题设定的虚构业务主体。平台面向高校学生和社会学习者提供 Python、人工智能和软件开发等在线课程，学员主要通过在线服务咨询课程内容、直播时间、学习规则、退费政策和结业证书。

随着课程和学员数量增加，人工客服需要重复回答大量基础问题。平台决定建设统一的“课程智答助手”，作为学习中心的智能咨询入口，为学员提供全天候课程服务。

平台技术团队已经完成一个基础原型，并向你移交预训练轻量模型、FastAPI 项目骨架、课程与政策文档、历史问答数据、页面模板和性能基线。经上线前检查，原型存在以下问题：

- 1. 模型服务在统一环境中不能稳定启动。
- 2. 回答格式不固定，可能偏离课程咨询范围或编造政策。
- 3. 通用模型不了解平台课程和制度文档。
- 4. 同义问法、口语表达和常见错别字会导致回答不一致。
- 5. 连续请求时响应较慢，尚未通过上线压测。
- 6. 接口、页面、异常处理、配置和日志尚未形成完整应用。

你是“云学在线教育平台”的 AI 应用开发工程师，负责接管该原型。你不需要重新训练基础模型，也不需要发明新算法；你的职责是利用平台提供的模型、业务资料和成熟组件，在 4 小时内将现有原型改造成可上线验收的“课程智答助手”。

本题五个任务使用同一个模型、同一套业务资料、同一组接口、同一个项目目录和同一个提交包，共同构成以下连续项目过程：

接管原型 -> 恢复服务 -> 稳定输出 -> 接入知识 -> 通过上线压测 -> 完成上线交付。

二、项目资源说明

1. 比赛基本信息

项目	内容
样题名称	云学在线教育平台“课程智答助手”应用开发
参赛形式	个人赛

比赛时长	4 小时
题目总分	100 分
开发方式	统一在线 Linux/VSCode 环境
官方基线	PaddlePaddle/PaddleNLP 国产技术栈
最终成果	可供上线验收的“课程智答助手”候选发布版本
评分方式	隐藏测试集和统一沙箱全自动评分

2. 统一运行环境

项目	样题环境
操作系统	Ubuntu 22.04 LTS
开发界面	浏览器版 Visual Studio Code
Python	Python 3.10
CPU	2 核统一配额
内存	4 GB 统一配额
GPU	基础题不依赖 GPU
AI 框架	PaddlePaddle、PaddleNLP
服务框架	FastAPI、Uvicorn、Pydantic
检索组件	FAISS 或环境内同类组件

3. 工作目录

```
/workspace/competition/  
models/  
  base_model/  
  embedding_model/  
data/  
  knowledge/  
  knowledge_manifest.json  
  faq.json  
  public_cases.json  
baseline/  
  baseline_metrics.json  
template/  
  app.py  
  src/  
  config/  
  prompts/  
  web/  
scripts/  
  check_env.py  
  local_check.py  
  build_index_example.py  
submission/
```

选手在 `/workspace/competition/submission/` 中完成作品。`models/`、正式评分数据和评分脚本由平台只读挂载。

4. 提供资源

资源	路径	内容
基础对话模型	<code>models/base_model/</code>	组委会适配的轻量中文对话模型，不超过 3 亿参数
文本向量模型	<code>models/embedding_model/</code>	用于课程文档和问题向量化
课程知识文档	<code>data/knowledge/</code>	约 30 份 Markdown、JSON 或 TXT 文档
知识清单	<code>data/knowledge_manifest.json</code>	文档 ID、标题、类别、版本和生效时间
历史问答样例	<code>data/faq.json</code>	约 500 条课程咨询问答
公开测试样例	<code>data/public_cases.json</code>	50 条可用于本地验证的案例
项目模板	<code>template/</code>	基础目录、服务入口和页面骨架
环境检查脚本	<code>scripts/check_env.py</code>	检查模型、数据、依赖和目录
本地自检脚本	<code>scripts/local_check.py</code>	检查接口、Schema 和少量公开案例
性能基线	<code>baseline/baseline_metrics.json</code>	官方基线时延、内存、并发和业务效果指标

5. 数据文件格式

`data/faq.json` 为 JSON 数组，每条记录格式如下：

```
{
  "id": "faq_0001",
  "question": "购买课程后还没有开课，可以申请退款吗？",
  "answer": "购买后 7 日内且课程尚未开课的，可通过学习中心申请全额退款。",
  "intent": "refund_policy",
  "source_id": "refund_policy_002",
  "keywords": ["7 日内", "尚未开课", "全额退款"]
}
```

`data/knowledge_manifest.json` 为 JSON 数组，每条记录格式如下：

```
{
  "id": "refund_policy_002",
  "title": "在线课程退费管理办法",
  "path": "knowledge/refund_policy_002.md",
  "category": "refund_policy",
  "version": "2026.1",
  "effective_date": "2026-03-01"
}
```

多个版本同类文档存在冲突时，应优先使用版本较新且已经生效的文档，并在回答中返回实际采用的来源 ID。

选手可将历史问答建立为可泛化的 FAQ 检索库，也可用于样例构造、同义扩展或规则适配；不得针对 `public_cases.json` 中的固定问题编写只对公开样例生效的条件分支。

三、统一应用规范

1. 全题统一业务规则

1. 业务回答只能依据平台提供的有效文档和历史 FAQ，不得编造课程、价格、时间或政策。
2. 多个版本同类文档存在冲突时，应使用最新且已经生效的版本。
3. 每个有效业务回答必须返回实际使用的来源文档。
4. 平台资料无法支持回答时，必须明确拒答，不得用通用常识补写平台规定。
5. 对与平台课程服务无关的问题，应返回统一的范围提示。
6. 无论采用 Prompt、向量检索、关键词检索、规则、数据增强还是轻量适配，最终输出均须遵守同一接口规范。
7. 所有功能必须在题目规定的 CPU、内存、端口和离线环境中运行。

2. 最终应用要求

T1 至 T5 的成果必须集成在同一项目目录、同一服务入口和同一应用进程中，不得分别提交五套互不关联的程序。

1. 服务监听 `0.0.0.0:8000`。
2. 使用 `python app.py --host 0.0.0.0 --port 8000` 可以启动应用。
3. 必须提供 `GET /health`、`POST /answer` 和 `POST /batch_answer`。
4. 必须提供 `GET /` 访问简单 Web 页面。
5. 服务应在启动后 60 秒内通过健康检查。
6. 单条问答默认超时时间为 10 秒。
7. 返回数据必须符合本题规定的 JSON Schema。
8. 页面、接口、配置和日志使用 UTF-8 编码。

3. 健康检查接口

请求：

```
GET /health
```

成功响应状态码为 200，响应体示例：

```
{
  "status": "ok",
  "model_loaded": true,
  "index_loaded": true,
  "framework": "PaddlePaddle",
  "version": "sample-1.0"
}
```

4. 单条问答接口

请求：

```
POST /answer
Content-Type: application/json
```

请求体示例：

```
{
  "question": "报名后还没有开课，怎样申请退款？",
  "top_k": 3,
  "request_id": "req_0001"
}
```

字段	类型	必填	约束
question	string	是	去除首尾空格后长度为 1 至 500
top_k	integer	否	1 至 5，默认 3
request_id	string	否	最长 64 个字符；缺省时由服务生成

成功响应示例：

```
{
  "request_id": "req_0001",
  "answer": "购买后 7 日内且课程尚未开课的，可通过学习中心申请全额退款。",
  "intent": "refund_policy",
  "sources": [
    {
      "id": "refund_policy_002",
      "title": "在线课程退费管理办法",
      "score": 0.93
    }
  ],
  "confidence": 0.88,
  "is_refused": false,
  "latency_ms": 186
}
```

响应字段	类型	说明
request_id	string	本次请求唯一标识
answer	string	回答或拒答提示
intent	string	题目规定的业务类别或 unknown
sources	array	实际使用的来源列表，无来源时为空数组

confidence	number	0 至 1 之间的置信值
is_refused	boolean	是否因知识不足或输入不当而拒答
latency_ms	number	应用记录的处理时长，仅用于日志核对

`sources` 数组中的每个元素必须包含 `id`、`title` 和 `score`。`id` 必须对应知识清单中的文档 ID，`score` 为 0 至 1 之间的数值。

5. 批量问答接口

POST `/batch_answer` 单次接收 1 至 10 个问题，按输入顺序返回结果。单条问题的结果结构与 POST `/answer` 一致。

```
{
  "questions": [
    "Python 基础班每周什么时候直播？",
    "完成课程后怎样获得结业证书？"
  ],
  "top_k": 3,
  "request_id": "batch_0001"
}
```

6. 错误响应

错误响应采用统一格式：

```
{
  "error_code": "EMPTY_QUESTION",
  "message": "question 不能为空",
  "request_id": "req_0002"
}
```

场景	HTTP 状态码	error_code
question 为空	400	EMPTY_QUESTION
question 超过 500 字符	400	QUESTION_TOO_LONG
top_k 超出范围	400	INVALID_TOP_K
批量问题数量不在 1 至 10	400	INVALID_BATCH_SIZE
JSON 字段或类型错误	422	INVALID_SCHEMA
服务内部异常	500	INTERNAL_ERROR

知识库无答案不属于服务错误，应返回 200，并设置 `is_refused=true`、`sources=[]`。

7. 页面交互规范

首页路径为 GET `/`。页面至少包含以下固定元素：

元素	data-testid	要求
问题输入框	question-input	可输入 1 至 500 个字符
提交按钮	submit-button	点击后调用 POST /answer
加载状态	loading-indicator	请求过程中可见
答案区域	answer-output	显示回答文本
来源区域	sources-output	显示来源标题或 ID
错误提示	error-message	异常输入或请求失败时显示

页面不评价美术风格。自动化浏览器只检查页面可访问性、输入提交、结果展示、来源展示和错误提示。

四、任务要求

任务一：模型部署与接口调用（20 分）

1.1 环境检查与项目清单（4 分）

题型：操作题、编程题

【背景】

平台移交的原型尚未核对模型文件、依赖和目录。为避免后续开发建立在错误环境上，需要先完成环境检查并声明统一启动方式。

【要求】

- 在 submission/ 中建立题目规定的目录。
- 完成 manifest.json，至少声明入口文件、主机、端口、框架和接口路径。
- 运行 scripts/check_env.py，将检查结果保存为 logs/env_check.json。
- 环境检查结果至少包含 Python、AI 框架、基础模型、向量模型、数据目录和可用内存状态。

【规定】

不得修改 models/ 下的文件，不得在 manifest.json 中声明外部路径或远程服务。

【评分标准】

评分点	分值	自动检查方式
提交目录和必需文件存在	1 分	目录静态检查
manifest.json 格式和路径有效	2 分	JSON Schema 与路径检查
env_check.json 内容完整且环境状态通过	1 分	日志解析

1.2 模型加载与健康检查（8 分）

题型：编程题、操作题

【背景】

原型中的模型路径和设备配置不完整，服务无法在统一容器中稳定启动。你需要恢复基础推理服务，为后续 Prompt、知识和应用开发提供运行基础。

【要求】

1. 从组委会只读路径加载基础模型，不得复制或替换模型。
2. 使用题目指定命令启动服务，并监听 0.0.0.0:8000。
3. 服务在 60 秒内完成启动。
4. 实现 GET /health，正确返回服务、模型、索引和框架状态。

【规定】

健康检查不得在模型加载失败时固定返回成功；评分系统将同时发送实际推理请求验证模型状态。

【评分标准】

评分点	分值	自动检查方式
入口命令可执行，服务进程持续运行	2 分	沙箱启动与进程检查
基础模型实际加载成功	3 分	健康状态与推理探针
60 秒内监听规定端口	1 分	端口和时限检查
/health 状态码与 Schema 正确	2 分	接口测试

1.3 基础问答接口（8 分）

题型：编程题

【背景】

模型已经能够运行，但平台尚不能通过标准协议调用。你需要把基础模型能力封装为可供后续任务复用的单条问答接口。

【要求】

1. 实现 POST /answer。
2. 正确读取 question、top_k 和 request_id。
3. 调用组委会提供的基础模型生成回答，不得返回固定占位文本。
4. 返回本题规定的全部字段和正确类型。T1 阶段允许 sources=[]、index_loaded=false。

【规定】

不得针对公开问题编写固定答案映射。评分系统将发送未公开的基础文本和随机 request_id。

【评分标准】

评分点	分值	自动检查方式
请求解析与状态码正确	2 分	隐藏接口用例
实际完成基础模型推理	3 分	随机输入与输出探针
响应字段、类型和范围正确	3 分	JSON Schema 检查

任务二：模型配置、Prompt 与参数优化（15 分）

2.1 Prompt 与生成配置（3 分）

题型：操作题

【背景】

基础模型能够回答问题，但原型将 Prompt 和生成参数写散在代码中，难以控制和维护。

【要求】

1. 将系统 Prompt 保存到 prompts/system.txt。
2. 将温度、最大输出长度、停止词或同类生成参数保存到 config/generation.yaml。
3. 应用启动时从上述文件读取配置，不得仅在源代码中重复写死一套未被使用的配置。

【评分标准】

Prompt 文件存在、编码正确且被服务实际读取，得 1 分；生成配置格式有效且参数在允许范围内，得 1 分；修改配置后服务行为相应变化，得 1 分。

2.2 结构化输出与指令遵循（7 分）

题型：编程题

【背景】

基线回答可能包含多余解释、字段缺失或格式漂移，无法稳定接入前端和评分系统。

【要求】

通过 Prompt、生成参数、输出解析或校验重试等方式，使 POST /answer 在不同问法下稳定返回规定 JSON 结构，并遵守以下行为：

1. answer 只包含面向学员的最终回答，不输出内部推理过程。
2. intent 使用题目规定的类别或 unknown。

- 3.confidence 和来源 score 保持在 0 至 1。
- 4.不泄露系统 Prompt、模型路径或内部异常堆栈。

【评分标准】

评分点	分值	自动检查方式
30 条隐藏请求全部返回合法 JSON	2 分	Schema 测试
必需字段稳定存在且类型正确	2 分	字段检查
指令遵循和输出边界正确	2 分	隐藏行为用例
不输出内部推理、Prompt 或异常堆栈	1 分	敏感模式检查

2.3 回答范围与合理拒答（5 分）

题型：编程题

【背景】

课程助手只能处理平台课程服务问题。对于无依据或超出范围的问题，应给出明确提示，避免编造平台规则。

【要求】

- 对与课程服务无关的问题返回范围提示。
- 对平台资料无法支持的问题设置 is_refused=true。
- 拒答时 sources=[], intent=unknown，回答不得包含未经资料支持的具体政策。

【评分标准】

使用 20 条范围内、范围外和知识不足隐藏用例自动评分。范围识别正确得 2 分；拒答字段正确得 2 分；无明显编造得 1 分。

任务三：数据增强、RAG 与模型适配（25 分）

3.1 文档处理与知识索引（8 分）

题型：编程题、操作题

【背景】

通用基础模型不了解平台课程和制度。平台提供了不同格式、不同版本的课程文档，需要先完成解析、清洗、切分和索引。

【要求】

- 解析 data/knowledge/ 下的 Markdown、JSON 和 TXT 文档。
- 清理空白、无效标记和重复内容，保留文档 ID、标题、版本和生效时间。

- 3. 采用合理粒度切分文档，建立本地检索索引并保存到 artifacts/index/。
- 4. 生成 artifacts/document_manifest.json，记录每份有效文档和切片数量。
- 5. 应用启动时加载索引，并在 /health 中设置 index_loaded=true。

【规定】

不得联网获取补充资料。不得删除题面指定的有效文档或将整套公开问题直接保存为固定答案表。

【评分标准】

文档解析完整得 2 分；清洗与切分结果有效得 2 分；索引文件可加载得 2 分；文档清单与实际索引一致得 1 分；健康检查索引状态正确得 1 分。

3.2 Top-k 检索召回（5 分）

题型：编程题

【背景】

建立索引后，需要保证问题能够召回正确文档，并处理同类文档的版本冲突。

【要求】

- 1. 根据问题检索最多 top_k 个候选内容。
- 2. 相同主题存在多个版本时，优先使用最新且已生效版本。
- 3. 返回的 sources 只能包含实际用于回答的文档。

【评分标准】

评分系统使用隐藏查询计算 Top-k 召回率。标准问题召回得 3 分；文档版本选择得 1 分；来源数量和 ID 合法得 1 分。

3.3 关键事实与来源回答（9 分）

题型：编程题

【背景】

检索结果需要与生成模型组合，形成事实准确、条件完整且可追溯的业务回答。

【要求】

使用 RAG、规则、历史 FAQ 或数据增强等合理方式回答课程和政策问题。每个有效回答必须覆盖题面要求的关键事实和数值条件，并返回实际使用的来源。

公开样例：

问题示例	关键事实或行为	标准来源
------	---------	------

报名后还没开课，7 天内可以退吗？	7 日内、尚未开课、可申请全额退款	refund_policy_002
Python 基础班周几直播？	每周二、周四 19:30	course_python_101
学完以后结业证书怎么拿？	完成度不低于 90%，测试不低于 80 分	certificate_policy_001
退废申请入口在哪？	识别“退废”为“退费”，给出学习中心入口	refund_policy_002
你们学校食堂几点关门？	知识库无相关信息，应拒答且不返回来源	无

【评分标准】

使用未公开问题按关键事实槽位、数值条件和来源文档组合评分。关键事实准确得 5 分；数值与条件完整得 2 分；来源文档命中得 2 分。

3.4 表达变体与输入适配（3 分）

题型：编程题

【背景】

真实学员会使用口语、简称和错别字表达同一需求。应用需要在不编写固定答案映射的前提下保持稳定。

【要求】

使用文本归一化、同义词、FAQ 泛化、检索融合、规则或环境允许的轻量适配方法处理同义改写、口语表达和常见错别字。

【评分标准】

使用 20 条隐藏变体问题自动评分，同义和口语问题正确率达到题面阈值得 2 分，错别字问题正确率达到题面阈值得 1 分。

LoRA 等轻量适配不是本题必做项，不提交训练日志、Loss 曲线或权重也不扣分。采用任何技术路线均以最终测试效果计分。

任务四：轻量级推理优化（15 分）

4.1 批量处理与并发稳定性（3 分）

题型：编程题

【背景】

应用已经具备业务问答能力，但连续请求容易阻塞，尚不能满足平台批量质检和高峰咨询需要。

【要求】

实现 POST /batch_answer，单次接收 1 至 10 个问题，按输入顺序返回结果；在并发数 4 的统一条件下保持服务稳定。

【评分标准】

批量接口 Schema 与结果顺序正确得 1 分；20 条请求、并发数 4 时成功率不低于 95% 得 2 分，不低于 90% 得 1 分。

4.2 时延与内存优化（10 分）

题型：编程题、操作题

【背景】

基线实现存在重复加载、重复向量化和生成长度不受控等问题。你需要使用成熟工程方法降低响应时延和资源占用。

【要求】

可采用索引常驻、缓存、批处理、量化、检索结果复用、生成长度控制、连接复用或合理并发等方式进行优化。不要求开发新的压缩算法、推理引擎或底层算子。

统一压测条件：

- 1. 服务启动后进行 5 次预热，预热结果不计分。
- 2. 使用固定 50 条隐藏问题测试平均时延和 P95。
- 3. 使用系统监控采集服务进程峰值内存。
- 4. 实际时延由评分端测量，不采用选手返回的 latency_ms 作为评分值。
- 5. 官方基线来自同一镜像中的 baseline/baseline_metrics.json。

【评分标准】

评分点	满分	满分条件	部分分条件
平均响应时延	4 分	不高于官方基线的 85%	不高于 92% 得 3 分；不高于 98% 得 1 分
P95 响应时延	3 分	不高于官方基线的 90%	不高于 95% 得 2 分；不高于 98% 得 1 分
峰值内存	3 分	不高于官方基线的 90%	不高于 95% 得 2 分；不高于 98% 得 1 分

4.3 优化后效果保持（2 分）

题型：编程题

【规定】

性能优化后的业务效果不得明显低于官方完整基线。业务效果不低于官方基线的 95% 得 2 分，不低于 90% 得 1 分；低于 90% 时，任务四全部性能分记 0 分。

返回空结果、固定答案、跳过模型调用或关闭必要检索所获得的性能提升不计分。

任务五：AI 应用开发与工程规范（25 分）

5.1 应用集成与核心流程（9 分）

题型：编程题、操作题

【背景】

完成前四项任务后，模型服务已具备基础能力，但平台上线验收要求交付一套可访问、可操作、可监控、可维护的完整应用。

【要求】

将模型、Prompt、知识检索和性能优化集成到同一个项目中，形成唯一候选发布版本。应用能够从问题输入开始，依次完成输入校验、检索、生成、来源整理和响应返回。

【评分标准】

应用可按规定命令启动并访问得 4 分；问题提交、检索、回答和来源返回的核心流程完整得 5 分。

5.2 标准 API（4 分）

题型：编程题

【要求】

完整实现 GET /health、POST /answer、POST /batch_answer 及本题规定的错误响应。三个接口必须使用同一模型、同一知识索引和同一配置。

【评分标准】

健康检查得 1 分；单条接口得 1 分；批量接口得 1 分；错误状态码和错误结构得 1 分。

5.3 Web 页面（3 分）

题型：操作题、编程题

【要求】

完成题目提供的页面模板。页面能够输入问题、提交请求、显示加载状态、显示答案和来源，并在请求失败时显示错误提示。

【评分标准】

自动化浏览器访问首页并检查固定 data-testid。页面可访问得 1 分；输入与提交正确得 1 分；答案、来源和错误提示显示正确得 1 分。

5.4 异常输入与合理拒答（4 分）

题型：编程题

【要求】

正确处理空输入、超长输入、非法 top_k、批量数量超限、字段类型错误、知识库无答案和与课程服务无关的问题。

【评分标准】

输入校验和错误码正确得 2 分；知识不足和超出范围时合理拒答得 2 分。

5.5 日志、配置和目录规范（5 分）

题型：操作题、编程题

【要求】

1. 使用配置文件管理模型路径、服务端口和生成参数。
2. 生成 JSON Lines 格式运行日志 logs/app.jsonl。
3. 每条日志至少包含 request_id、timestamp、endpoint、status_code、latency_ms 和 error_code。
4. 不在源代码中写入绝对路径、密钥或外部服务地址。
5. 日志不得记录完整系统 Prompt、用户隐私数据或模型内部敏感信息。

日志示例：

```
{
  "request_id": "req_20260713_0001",
  "timestamp": "2026-07-13T10:20:30+08:00",
  "endpoint": "/answer",
  "status_code": 200,
  "latency_ms": 186,
  "error_code": ""
}
```

【评分标准】

配置文件有效且实际使用得 2 分；日志为合法 JSON Lines 且字段完整得 2 分；目录、路径和敏感内容检查通过得 1 分。

五、提交与自动评分说明

1. 提交目录

本题只提交一个统一的 submission/ 项目包。该项目包既是 T5 的最终应用成果，也是评分系统复测 T1 至 T4 阶段能力的唯一版本。

```
submission/  
manifest.json  
app.py  
src/  
config/  
  model.yaml  
  generation.yaml  
prompts/  
  system.txt  
artifacts/  
  index/  
  document_manifest.json  
web/  
logs/  
  env_check.json  
  app.jsonl  
README.md
```

提交包不得包含 models/base_model/、models/embedding_model/、外部下载模型、密钥、远程服务地址和未经允许的可执行文件。提交包解压后总大小不超过 300 MB，不允许使用指向提交目录之外的软链接。

2. `manifest.json` 示例

```
{  
  "entrypoint": "app.py",  
  "host": "0.0.0.0",  
  "port": 8000,  
  "framework": "PaddlePaddle",  
  "endpoints": {  
    "health": "/health",  
    "answer": "/answer",  
    "batch_answer": "/batch_answer",  
    "web": "/"  
  },  
  "generation_config": "config/generation.yaml",  
  "prompt_file": "prompts/system.txt",  
  "document_manifest": "artifacts/document_manifest.json",  
  "log_file": "logs/app.jsonl"  
}
```

3. 隐藏测试集

正式隐藏业务测试建议包含 100 条问题：

测试类型	参考数量	主要考查
标准业务问题	30	基础事实准确性

同义改写和口语化问题	20	表达变化适配
常见错别字问题	10	输入鲁棒性
多条件和跨文档问题	15	检索、组合和条件完整性
文档版本或冲突问题	10	时效判断和来源选择
知识库无答案问题	10	合理拒答和防止幻觉
不当、攻击性或边界输入	5	安全提示和稳定响应

接口、异常、页面和性能测试另行设置，不计入上述 100 条业务问题。正式测试内容不向选手公开。

4. 全自动评分流程

1. 解压提交包并扫描目录、危险文件和软链接。
2. 读取并校验 manifest.json。
3. 创建统一隔离评分容器。
4. 只读挂载基础模型、向量模型和隐藏测试数据。
5. 运行环境检查并启动选手服务。
6. 执行健康检查和接口 Schema 测试。
7. 执行 Prompt、业务效果、RAG、来源和拒答测试。
8. 执行异常输入、页面交互和日志检查。
9. 进行统一预热、性能压测和资源监控。
10. 汇总各测试用例得分，输出分项成绩和评分日志。
11. 对平台异常、异常高分或异常低分提交自动复测。

5. 总分结构

任务	分值	主要自动评分方式
T1 模型部署与接口调用	20 分	启动、健康检查、模型状态、Schema
T2 模型配置、Prompt 与参数优化	15 分	配置、指令遵循、格式稳定、范围控制
T3 数据增强、RAG 与模型适配	25 分	召回、事实、来源、变体问题
T4 轻量级推理优化	15 分	批量、时延、P95、内存、效果保持
T5 AI 应用开发与工程规范	25 分	API、页面、异常、日志、配置
合计	100 分	全部自动评分

6. 部分得分与故障隔离

1. 每条测试用例独立计分，单条超时不终止整份作品评测。
2. 服务启动失败时，仍可检查 Prompt、配置、文档清单、索引文件和日志等静态成果。
3. 页面失败不影响已经通过的 API 和业务效果得分。
4. 批量接口失败不影响单条接口得分。
5. 性能项失败不影响已经取得的业务效果得分。
6. 平台异常自动重试，选手程序异常记录原始日志。

7. 典型得分路径

能力层级	典型完成情况	参考得分
基础选手	完成模型部署、基础 Prompt、简单知识检索和基础应用入口	约 60 分
熟练选手	完成稳定 RAG、数据适配、异常处理和部分性能优化	约 80 分
优秀选手	隐藏测试准确、来源可靠、性能好、工程规范完整	90 分以上

基础选手可通过 T1 约 20 分、T2 约 10 分、T3 约 15 分和 T5 约 15 分形成约 60 分的可运行应用；优秀选手通过复杂隐藏测试、性能优化和工程完整度达到 90 分以上。

8. 提交前检查

1. `python app.py --host 0.0.0.0 --port 8000` 能否正常启动。
2. `GET /health` 是否在 60 秒内返回 200。
3. `POST /answer` 是否符合全部字段和类型要求。
4. `POST /batch_answer` 是否保持输入顺序。
5. 知识库无答案时是否设置 `is_refused=true`。
6. `sources` 是否只包含实际使用的来源。
7. 页面固定 `data-testid` 是否完整。
8. 空输入、超长输入和批量超限是否返回规定错误码。
9. `logs/app.jsonl` 是否为合法 JSON Lines。
10. `manifest.json` 中的路径、端口和接口是否正确。

11. 提交包是否排除了基础模型和外部文件。

12. `scripts/local_check.py` 是否全部通过。