附件1

ICS 35.240.01
CCS R07

T/SDASTC

团　　体　　标　　准

T/SDASTC XXXX—XXXX

智能旅游大模型定制化服务系统技术要求

Technical requirements for customized service system of intelligent tourism large language model

（征求意见稿）

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

山东科技咨询协会　　发 布

目　　次

前言 II
1　范围 1
2　规范性引用文件 1
3　术语和定义 1
4　缩略语 2
5　系统框架 2
6　功能要求 3
7　性能要求 6
8　可靠性要求 6
9　维护性要求 6
10　兼容性要求 7
11　安全性要求 7
12　可扩展性要求 7

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由山东科技咨询协会提出并归口。

本文件起草单位：

本文件主要起草人：

智能旅游大模型定制化服务系统技术要求

1 范围

本文件提出了智能旅游大模型定制化服务系统的系统框架，规定了功能、性能、可靠性、维护性、兼容性、安全性和可扩展性要求。

本文件适用于文旅行业基于大语言模型构建的定制化智能旅游服务系统及相关解决方案的设计、研发、测试和部署。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 30225 旅游景区智慧化运营管理要求

GB/T 35273 信息安全技术 个人信息安全规范

GB 45438 网络安全技术 人工智能生成合成内容标识方法

GB/T 45654 网络安全技术 生成式人工智能服务安全基本要求

GB/T 45674 网络安全技术 生成式人工智能数据标注安全规范

GB/T 47863 生成式人工智能技术应用社会影响 服务提供者合规管理指南

3 术语和定义

下列术语和定义适用于本文件。

3.1

智能旅游大模型定制化服务系统 intelligent tourism large model customized service system

基于大语言模型技术，面向旅游服务场景，通过用户画像分析、需求理解和内容生成能力，为不同出行人群提供个性化行程规划、智能景点讲解、旅行文案生成、POI推荐等定制化服务的软件系统。

3.2

大语言模型（LLM） large language model

基于深度学习技术，以Transformer架构为基础，通过大规模文本语料训练而成的、具有自然语言理解和生成能力的人工神经网络模型，通常参数量达到数十亿及以上规模。

3.3

定制化服务 customized service

根据用户的个性化需求（如出行人群、预算水平、兴趣偏好、时间约束等），通过智能算法自动生成差异化的旅游服务内容和方案的服务模式。

3.4

兴趣点推荐（POI 推荐） point of interest recommendation

基于用户画像和位置信息，对餐厅、酒店、景区、购物场所、娱乐设施等地理兴趣点进行智能筛选和个性化推荐的技术与服务。

3.5

多模态交互 multimodal interaction

支持文本、图像、语音等多种信息模态的输入与输出，实现人机之间多通道、多维度信息交互的技术方式。

3.6

AI 导游 AI tour guide

基于大语言模型和知识库构建的、能够模拟专业导游进行景点讲解、行程引导和知识问答的智能服务系统。

3.7

微调 fine-tuning

在预训练模型的基础上，使用特定领域或任务的训练数据对模型参数进行调整，以适应特定应用场景需求的技术过程。

3.8

低秩自适应（LoRA）low-rank adaptation;

一种高效模型微调方法，通过在原始权重矩阵中注入可训练的低秩分解矩阵，仅更新少量参数即可实现模型领域适配的技术。

4 缩略语

下列缩略语适用于本文件。

AES：高级加密标准（Advanced Encryption Standard）
API：应用程序编程接口（Application Programming Interface）
GPU：图形处理器（Graphics Processing Unit）
GQA：分组查询注意力（Grouped Query Attention）
HTTPS：超文本传输安全协议（Hypertext Transfer Protocol Secure）
LLM：大语言模型（Large Language Model）
LoRA：低秩自适应（Low-Rank Adaptation）
OTA 在线旅行社（Online Travel Agency）
POI 兴趣点（Point of Interest）
REST 表述性状态转移（Representational State Transfer）
RoPE 旋转位置编码（Rotary Position Embedding）
SwiGLU Swish门控线性单元（Swish-Gated Linear Unit）
TLS 传输层安全协议（Transport Layer Security）

5 系统框架

5.1 概述

智能旅游大模型定制化服务系统采用分层架构，由基础设施层、数据资源层、模型层、服务层和应用层构成，各层之间通过标准化接口交互；安全保护与运维保障能力贯穿各层，为系统持续稳定运行提供支撑。系统框架见图1。



图1　智能旅游大模型定制化服务系统框架

5.2　基础设施层

基础设施层为系统提供计算资源、存储资源和网络资源，包括计算设备、人工智能加速卡、分布式存储和网络链路，支撑模型训练、微调、推理和服务运行，并具备资源弹性伸缩能力。

5.3　数据资源层

数据资源层为系统提供训练语料、旅游知识、兴趣点数据和交互数据等资源，涵盖数据采集、数据构成、数据质量、数据存储和数据更新等环节，是模型领域适配和生成内容质量的基础。

5.4　模型层

模型层承载大语言模型的训练、微调和推理，包括基座模型、领域微调模型、多模态交互处理能力和模型评测，向上层提供文本理解、内容生成和图像识别等能力。

5.5　服务层

服务层将模型能力封装为标准化服务接口，提供接口调用、流式输出、鉴权审计和服务编排等功能，支持应用层和多类终端按需调用。

5.6　应用层

应用层面向终端用户提供个性化行程规划、智能景点讲解、多风格旅行文案创作、旅游避雷指南、兴趣点智能推荐和拍照识景导游等服务，并支持多形态终端接入。

6　功能要求

6.1　基础设施层

6.1.1　基础设施层功能要求包括：

a) 应提供图形处理器等人工智能加速计算资源，单卡显存不低于 24 GB，满足模型训练与推理需要；
b) 应提供分布式存储，可用容量不低于 10 TB，读写性能满足大规模训练数据加载需要；
c) 应支持资源弹性伸缩，根据业务负载自动扩缩容；
d) 网络带宽应不低于 1 Gbit/s，并支持 HTTPS/TLS 1.3 加密传输。

6.1.2 网络基础设施应符合 GB/T 30225 的相关要求。

6.2 数据资源层

6.2.1 数据采集

6.2.1.1 训练数据应来源于合法合规渠道，包括已获授权的公开数据接口、公开旅游文本和自建模板引擎等。

6.2.1.2 训练数据采集应符合 GB/T 45654 的要求。

6.2.2 数据构成

6.2.2.1 训练数据总量应不少于 1 亿条，数据格式应统一。

6.2.2.2 训练数据应覆盖不少于 500 个城市、5000 个景点兴趣点的信息。

6.2.2.3 训练数据应包含以下核心数据域，各域占比合理分布：

a) 行程规划数据：覆盖不同天数（1 天～14 天）、预算（500 元～10000 元以上）、出行人群（不少于 8 类）、兴趣偏好（不少于 9 类）的全维度行程方案。
b) 旅游避雷数据：涵盖不少于 15 类旅游陷阱的识别和应对信息。
c) 景点讲解数据：覆盖 4A 及以上景区、世界遗产、全国重点文物保护单位，包含历史文化、建筑民俗、自然地理等多维度讲解内容。
d) 导游脚本数据：融合名人轶事、民间传说、建筑典故等不少于 6 类故事元素。
e) POI 推荐数据：涵盖餐饮、住宿、购物、娱乐、户外 5 大品类，基于真实地图 POI 数据。
f) 旅行文案数据：包含不少于 5 种文体风格的创作样本。

6.2.2.4 训练数据质量要求应符合 GB/T 45654 的要求。

6.2.3 数据质量

6.2.3.1 应建立数据去重机制，采用内容哈希去重等方法消除重复数据。

6.2.3.2 应建立数据质量管控体系，包含质量检查规则。

6.2.3.3 门票价格、开放时间、交通费用等关键信息应经过验证，不确定性信息应注明建议核实来源。

6.2.4 数据存储

6.2.4.1 训练数据和模型权重应实施分级访问控制，仅授权人员可访问。

6.2.4.2 数据存储应采用 AES-256 或同等强度的加密算法。

6.2.5 数据更新

6.2.5.1 应建立训练数据定期更新机制，其中兴趣点信息更新周期应不超过 6 个月。

6.3 模型层

6.3.1 模型架构

6.3.1.1 模型宜采用 Transformer Decoder-only 架构，采用其他架构的，应具备等效的文本理解、知识推理与长文本生成能力。模型应基于已完成生成式人工智能服务备案的开源基座模型进行文旅领域定制。

6.3.1.2 模型总参数量应不低于 80 亿（8B），具备文本理解、知识推理和长文本生成能力。

6.3.1.3 模型应支持高效的中文分词处理，中文旅游文本的 Token 利用率应优于同规模英文优化的模型。

6.3.1.4 模型应支持序列长度扩展，适应超长行程规划和深度讲解场景。

6.3.1.5 模型应支持分组查询注意力（GQA）、旋转位置编码（RoPE）、SwiGLU 激活函数等优化技术。

6.3.2 微调方法

6.3.2.1 应采用 LoRA 高效微调技术，在保持基座模型泛化能力的前提下实现文旅领域适配。
6.3.2.2 LoRA 微调参数比例不应超过模型总参数量的 1%。
6.3.2.3 LoRA 微调参数配置中，秩值应不低于 8，缩放系数应为秩值的 2 倍。
6.3.2.4 微调应基于经过筛选的高质量数据，筛选后数据量应不少于 30 万条。
6.3.2.5 微调后的模型在文旅领域核心任务上的表现应优于未微调的基座模型，对比评测方法见 13.3。

6.3.3 多模态处理能力

6.3.3.1 系统应具备多模态交互处理能力，支持文本模态的输入与输出，能够处理行程规划、景点讲解、文案创作、避雷问答、POI 推荐等多类文旅任务；
6.3.3.2 系统宜支持视觉模态的输入，具备景点照片识别能力，包括：

a) 景点名称和类别自动识别；
b) 古建筑年代、风格流派识别；
c) 自然地貌类型（山水、丹霞、喀斯特等）识别。

6.3.3.3 视觉模态下典型景点场景的识别准确率应不低于 90%。
6.3.3.4 应支持基于视觉输入的追问式多轮对话交互。

6.4 服务层

6.4.1 API 接口应采用 RESTful 架构，支持 HTTP/HTTPS 协议。
6.4.2 应支持流式输出接口，按 Token 逐步返回生成内容。
6.4.3 应支持 API Key 认证和 IP 白名单双重鉴权机制。
6.4.4 请求日志应保存不少于 30 天，支持安全审计和追溯。
6.4.5 接口应支持各类编程语言调用。
6.4.6 应支持通过接口集成至第三方应用系统。

6.5 应用层

6.5.1 个性化行程规划模块

6.5.1.1 应支持多维度约束条件输入，包括但不限于：出发地、目的地、出行天数（1 天～14 天）、预算范围、出行人群类型、兴趣偏好、出行季节等。
6.5.1.2 应支持不少于 8 类出行人群的定制化行程方案：包括单人、情侣、亲子、长辈、朋友、商务、研学、团建等。
6.5.1.3 应支持多档预算水平的费用控制。
6.5.1.4 应输出结构化的逐日旅行行程规划安排，包含景点、时间、交通、餐饮、费用估算等要素。

6.5.2 智能景点讲解模块

6.5.2.1 应支持通过景点名称、关键词或照片等方式获取景点讲解信息。
6.5.2.2 讲解内容应涵盖历史文化、建筑民俗、自然地理等多维度信息。
6.5.2.3 每条讲解应融入不少于 1 个历史典故或文化故事。
6.5.2.4 应支持 AI 导游追问式多轮对话交互，如“这个建筑是什么朝代的？”“有什么名人来过？”等深度交互。
6.5.2.5 景点讲解覆盖范围应不少于 5000 个景点。

6.5.3 多风格旅行文案创作模块

6.5.3.1 应支持不少于 6 种文风的一键生成，包括小红书笔记、朋友圈种草、深度游记、旅行诗歌、实用攻略、公众号推文等。
6.5.3.2 各文风输出应符合对应平台的格式规范，如朋友圈文案不超过 80 字、小红书笔记 100～200 字、深度游记不超过 800 字。
6.5.3.3 应支持根据景点特征和用户偏好自动生成适配文案内容。

6.5.4 旅游避雷指南模块

6.5.4.1 应覆盖不少于 15 类旅游陷阱的识别和应对建议，包括但不限于：景区套路、交通踩坑、餐饮被宰、购物陷阱、门票陷阱、住宿陷阱、季节性风险、安全警示等。
6.5.4.2 应支持基于目的地自动匹配相关避雷信息。
6.5.4.3 避雷信息应具有实用性和时效性，关键价格信息应注明建议核实来源。

6.5.5 POI 智能推荐模块

6.5.5.1 应支持餐厅、酒店、娱乐、购物、户外 5 大品类的个性化 POI 推荐。
6.5.5.2 POI 推荐结果应包含名称、位置、人均消费、评分、交通指引等关键信息。
6.5.5.3 POI 推荐应基于真实地图 POI 数据，具备地理真实性和信息可靠性。
6.5.5.4 POI 推荐引擎应支持根据用户偏好和位置信息进行动态排序。

6.5.6 拍照识景导游模块

6.5.6.1 应支持通过拍照或上传图片的方式识别景点。
6.5.6.2 识别后应自动生成该景点的 AI 导游专业讲解内容。
6.5.6.3 应支持识别后的追问式多轮对话。

7 性能要求

7.1 响应时间

7.1.1 首 Token 响应延迟应不超过 300 ms。
7.1.2 云端部署环境下，流式输出速率应不低于 50 tokens/s；本地部署环境下（消费级加速快），流式输出速率应不低于 15 tokens/s。
7.1.3 平均响应延迟应不超过 3 s。

7.2 并发能力

7.2.1 单张加速卡（显存不低于 80 GB 或同等性能）应支持不少于 8 路并发推理请求。
7.2.2 系统应支持根据业务负载水平扩展，通过增加计算节点（如 GPU 节点）提升并发处理能力。

7.3 模型输出质量

7.3.1 生成内容经质量检查后的合格率应不低于 99%。
7.3.2 旅游场景下的事实准确性应不低于 75%。
7.3.3 旅游幻觉率（生成虚假旅游信息的比例）应不超过 1%。

8 可靠性要求

8.1 服务可用性

8.1.1 云端服务可用性应不低于 99.5%。
8.1.2 API 响应成功率应不低于 99.5%。

8.2 故障恢复

8.2.1 应制定应急预案，确保服务中断后能在 30 min 内恢复。
8.2.2 应建立数据备份和灾备恢复机制，备份周期应不超过 24 h。

9 维护性要求

9.1 运行监控

9.1.1 应建立系统监控体系，实时监控 GPU 利用率、内存使用率、请求队列长度等关键指标。
9.1.2 应建立异常告警机制，对服务中断、性能下降、安全事件等进行实时告警。

9.2 版本管理

9.2.1 应建立模型版本管理和热更新机制，支持模型迭代升级不影响在线服务。
9.2.2 应建立数据版本管理机制，支持训练数据的可追溯性。

10 兼容性要求

10.1 软件兼容性

10.1.1 接口格式应兼容主流大语言模型服务接口格式，降低系统集成成本。
10.1.2 云端部署应兼容主流公有云平台，并具备弹性伸缩能力。
10.1.3 本地化部署应兼容主流推理框架，支持 INT4、INT8、FP16 等量化精度。

10.2 硬件兼容性

10.2.1 应兼容显存不低于 12 GB 的图形处理器用于推理，显存不低于 80 GB 的人工智能加速卡用于训练。
10.2.2 宜兼容统一内存不低于 16 GB 的 ARM 架构计算平台用于推理。
10.2.3 宜支持国产人工智能加速卡。

10.3 终端兼容性

10.3.1 Web 端应支持主流浏览器，兼容 PC 端和移动端响应式布局。
10.3.2 宜支持微信小程序、H5 等移动端接入。
10.3.3 宜支持景区自助终端、触摸屏等大屏设备接入。

11 安全性要求

11.1 数据安全

11.1.1 数据传输应采用 HTTPS/TLS 1.3 全链路加密。
11.1.2 数据存储应采用 AES-256 或同等强度加密算法。
11.1.3 训练数据和模型权重应实施分级访问控制。
11.1.4 数据安全应符合 GB/T 45654 和 GB/T 45674 的相关规定。

11.2 内容安全

11.2.1 应内置敏感词过滤和违规内容检测机制。
11.2.2 生成内容应经过安全合规检查，不得生成违法违规信息。
11.2.3 人工智能生成的合成内容应按照 GB 45438 的要求进行标识。
11.2.4 应建立内容安全事件应急响应机制。
11.2.5 模型生成内容安全性应符合 GB/T 45654 的要求。

11.3 用户隐私保护

11.3.1 用户上传的景点照片应在实时推理后即时清除，不落盘存储。
11.3.2 应提供对话历史一键清除功能。
11.3.3 用户个人信息的收集和使用应符合 GB/T 35273 的要求。不应采集用户位置、设备指纹、浏览行为等非必要隐私信息。
11.3.4 应明示告知用户数据处理方式和隐私政策。

11.4 合规管理

11.4.1 提供生成式人工智能服务前，应按照国家规定完成算法备案和安全评估。
11.4.2 社会影响评估应符合 GB/T 47863 的要求。

12 可扩展性要求

12.1 应支持模型参数的升级。
12.2 应支持功能模块的灵活扩展和组合。
12.3 应支持训练数据的增量更新。