大语言模型在教育信息化中的实践：规范、框架与应用

徐刚[1]，刘志鹏[2]，冯骐[2]，沈富可[3]
（1.华东师范大学软件工程学院，上海 200062；2.华东师范大学信息化治理办公室，上海 200062；
3.华东师范大学计算机科学与技术学院，上海 200062）

摘　要：为了解决大语言模型在高等教育领域落地中方案复杂、模型深度绑定、可复用性差等问题，提出了一套大语言模型在教育信息化中的应用规范，从通用对话、检索增强和智能体3个角度给出了API的规范设计，并在此基础上给出了一套技术框架。结合华东师范大学的3个具体案例展示了大语言模型如何深度融入校园生活，提升教学质量和学习体验。研究结果表明，基于规范性框架的应用开发具备良好的交互性和扩展性，底层能力高效复用有效提升了应用的开发效率。最后，总结了大语言模型在华东师范大学的应用情况，并对未来发展趋势进行了展望，强调了与教育场景融合的重要性，为推动人工智能技术在教育领域的广泛应用提供了参考依据。

关键词：大语言模型；教育信息化；规范；框架；应用；人工智能
中图分类号：TP18
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2024249

Practical application of large language models in educational informatics: specification, framework, and applications

XU Gang[1], LIU Zhipeng[2], FENG Qi[2], SHEN Fuke[3]
1. Software Engineering Institute, East China Normal University, Shanghai 200062, China
2. Information Technology Services, East China Normal University, Shanghai 200062, China
3. School of Computer Science and Technology, East China Normal University, Shanghai 200062, China

Abstract: To address the issues of complex implementation, deep model coupling, and poor reusability in deploying large language models (LLMs) in higher education, a set of application standards for LLMs in educational informatics was proposed. The standards covered API interface designs from three perspectives: general conversation, retrieval enhancement, and intelligent agents. Based on these standards, a technical implementation framework was provided. The paper demonstrated how large language models could be deeply integrated into campus life to improve teaching quality and learning experiences through three specific cases at East China Normal University (ECNU). The results show that applications developed based on a standardized framework possess good interactivity and extensibility, and the efficient reuse of underlying capabilities significantly improves development efficiency. Finally, the paper summarized the application scenarios of LLMs at ECNU and looked forward to future trends, emphasizing the importance of integrating with educational contexts. This provided a reference for promoting the widespread application of artificial intelligence technology in the field of education.

Keywords: large language model, educational informatic, specification, framework, applications, AI

收稿日期：2024–10–21
通信作者：冯骐，qfeng@admin.ecnu,edu.cn
基金项目：中国高等教育学会高等教育科学研究规划课题基金资助项目（No.20XX0301）
Foundation Item： Research Planning Projects for Scientific Studies in Higher Education by the China Association of Higher Education(No.20XX0301)

0 引言

近年来，随着深度学习和人工智能技术的飞速发展，大语言模型逐渐崭露头角，刘挺[1]指出，人工智能经历了四次高潮，其中第三次高潮是由深度学习推动的，而2022年11月OpenAI发布的由大语言模型支持的ChatGPT-3.5标志着第四次高潮的到来。大语言模型不仅在文本分类、问答系统等任务上表现出色，随着进一步的整合应用，它在政务[2]、安防[3]图书情报[4-5]金融审计[6-7]教育[8-9]，科研[10]等行业产生了深远的影响。

特别地，在教育信息化领域中，邬贺铨[11]提出“模型即服务”，认为大语言模型应该融入云平台，构建基于人工智能的基础设施为数智化转型赋能。张春红等[12]认为大语言模型在教育问答系统中有非常大的价值和影响。贾积有等[13]提出了一套针对教育访谈文本处理的提示词框架，通过生成高质量的摘要式总结，有效提高教育科学研究的效率和质量。汪张龙[14]探讨了大语言模型在试题命制、交互式语言测试、智能化评阅卷、基于考试数据的教育评价及智能化考试管理与服务五大场景中的作用。Dinh等[15]构建了一套专为大学相关问题设计的聊天机器人系统，在提供自动化服务、优化客户体验方面取得了积极效果。周杰等[16]通过在教育语料库上进行预训练和系统提示词工程及指令微调，实现了个性化、公平且富有同情心的智能教育服务。这些工作体现了大语言模型在教育信息化领域中的巨大潜力和应用价值。

然而，目前针对大语言模型在教育信息化领域中的研究都是基于具体场景设计的专用方案，其技术架构与所选用的大语言模型深度绑定，难以解耦，可复用性不佳，其中的一些方案对算力要求偏高，不符合大部分学校的实际资源能力，不利于在更大范围内进行推广和复制。因此，给出一种大语言模型在教育信息化领域的规范设计和技术框架，有非常重要的研究意义和现实价值。

本文的研究工作主要如下。

1) 提出了一套大语言模型在教育信息化领域的规范设计。从通用对话、检索增加和智能体3个维度阐述了大语言模型在应用时所应遵循的交互接口规范，并给出了具体的接口请求和响应示例。

2) 给出了一套大语言模型在教育信息化领域的技术框架。技术框架采用模块化的设计，将整个架构分为模型适配器、RAG适配器、代理网关、控制中心、前端应用、智能体编排平台和开放平台等模块，并描述了模块之间的交互逻辑和实现的阶段优先级。框架同时考虑了安全平面和可观测性平面的设计方案，并描述了各个模块接入安全平面和可观测性平面的交互规范。

3) 结合华东师范大学的3个具体案例——基于意图理解的精准检索“校园百事通”、数据驱动的课程助手“师大AI助教”以及数字人赋能的微课合成工具“微课工坊”，展示了教育大模型如何深度融入校园生活，提升教学质量和学习体验。研究结果表明，基于规范性设计的应用开发具备良好的交互性和扩展性，底层能力高效复用有效提升了应用的开发效率。

1 规范设计

在考虑教育信息化场景下的大语言模型的规范设计时，本文仅考虑应用程序接口（API）的规范，而不必讨论模型本身的部署要求。这使学校在应用大语言模型时，既可以选择基于算力加载本地模型，也可以选择以软件即服务（SaaS）形式调用外部的大语言模型服务。只要能通过某种形式，将其封装为统一的API规范即可。考虑到基于OpenAI规范下的生态已经颇具规模，本文所设计的API规范，应当在兼容OpenAI的接口基础上，再适当扩展以支持更多场景。根据大语言模型的应用场景，在API层面本文只需要定义3个核心场景下的接口规范：通用对话、检索增强和智能体。

1.1 通用对话

出于OpenAI的兼容要求，大语言模型的对话接口应该保持/v1/chat/completions作为接口地址后缀，前缀则可以根据实际需要自行定义，例如，https://{domain}/api/v1/chat/completions和https://{domain}/v1/chat/completions都是符合规范的接口地址。

在接口请求中，至少需要支持4个参数，如表1所示。

表1 请求参数

参数	含义
message	形如 [{"role": "user", "content": "你好呀"}] 的对话数组
stream	是否开启流失输出
model	区分模型
temperature	模型温度值，决定输出的随机性

图 1 给出了一个请求的示例。

```
###chat 接口###
Send Request
POST https://[redacted]/open/api/v1/chat/completi
Content-Type: application/json
Authorization: Bearer [redacted]51840df381217

{
    "messages": [{
        "role": "system",
        "content": "你是华东师范大学教育大模型"
    }, {
        "role": "user",
        "content": "你好呀"
    }
    ],
    "stream": false,
    "model": "ChatECNU",
    "temperature":0.5
}
```

图 1　请求示例

在接口响应中，至少需要支持 5 个参数，如表 2 所示。

表 2　响应参数

参数	含义
ID	形如 chatcmpl-xxxx 的唯一 ID
object	固定为 chat.completion
created	时间戳，秒，例如 1723480728
choices	形如 [{"index":0,"message":{"role":"assistant","content":"你好，有什么我可以帮你的吗？"},"finish_reason":"stop"}] 的回答响应数组。当采取流式响应时，"finish_reason":"stop" 代表输出结束。
usage	形如 {"prompt_tokens":9,"completion_tokens":7,"total_tokens":16} 的 token 消耗统计

非流式（stream=false）响应示例如图 2 所示。

在开启流式输出时，响应应符合 W3C server-sent events 的规范[17]，即每一项数据更新应该包含一个 data:字段，每一行必须以换行符(\n)结束，并且所有事件数据必须编码为 UTF-8。流式响应示例如图 3 所示。

只要符合上述规范，则接口符合 OpenAI 的接口兼容要求。在此基础上，本文针对一些典型场景进行了扩充定义。

```
HTTP/1.1 200 OK
Content-Type: text/plain; charset=utf-8
Content-Length: 290
Connection: close
Date: Mon, 12 Aug 2024 16:57:30 GMT

{
  "id": "chatcmpl-201fea16c82c48709261d0a2626a3a88",
  "object": "chat.completion",
  "created": 1723481850,
  "choices": [
    {
      "index": 0,
      "message": {
        "role": "assistant",
        "content": "你好！很高兴和你交流。"
      },
      "finish_reason": "stop"
    }
  ],
  "usage": {
    "prompt_tokens": 8,
    "completion_tokens": 6,
    "total_tokens": 14
  }
}
```

图 2　非流式的响应示例

```
HTTP/1.1 200 OK
Content-Type: text/event-stream
Transfer-Encoding: chunked
Connection: close
Date: Mon, 12 Aug 2024 16:43:11 GMT
Cache-Control: no-cache

data: {"id":"chatcmpl-2dd381143c4245b78f2eda470d904f0e","object":"chat.completion.chunk","created":1723480991,"model":"SparkDesk","choices":[{"index":0,"delta":{"content":"你好"}}]}

data: {"id":"chatcmpl-d55079be2fe640e1a44212f119f201f6","object":"chat.completion.chunk","created":1723480992,"model":"SparkDesk","choices":[{"index":0,"delta":{"content":"！有什么"}}]}

data: {"id":"chatcmpl-6eeaafe56b794e4aa456d84824c3343d","object":"chat.completion.chunk","created":1723480992,"model":"SparkDesk","choices":[{"index":0,"delta":{"content":"我可以帮助你的"}}]}

data: {"id":"chatcmpl-3b59f3f78fe7470396f201a51b17e0d9","object":"chat.completion.chunk","created":1723480992,"model":"SparkDesk","choices":[{"index":0,"delta":{"content":"吗？"},"finish_reason":"stop"}]}

data: [DONE]
```

图 3　流式响应示例

1) 会话同步

OpenAI 的接口规范本身是无状态的，然而在实际应用中，如果需要长期记忆会话，或者多个终

端上同步会话状态，则本文需要增加一个会话参数conversationid，通过它来长期存储会话状态，并使跨终端使用时，会话状态可以借助后端存储的conversationid实现同步。

2) 意图分类

在许多问答类场景里，本文需要先理解用户的提问意图，再针对性地引流到专用的模型或者专用的知识库做响应，以期获得更好的回答效果。Dinh等[15]训练了一个专用的深度学习网络实现意图判断器，这种模式过于复杂和高耦合，不利于复制和推广。

本文提出一种简化的实现方式。首先，实现若干个大语言模型服务，每个模型服务绑定了某个专用的场景，所有的模型暴露的API都必须符合上述接口规范。本文称这些模型的API为渠道，同时本文实现了一个统一的代理网关，以实现对不同渠道的路由选择。

其次，本文选择一个支持Function Call的模型，定义若干个意图理解函数，尽可能详细地描述函数的description以提高意图理解的准确性。同时本文扩充了function call的请求参数，增加了渠道ID，并基于此ID将请求整体路由到对应的渠道，实现不需要二次训练即可文本化配置的意图分类器。

以信息化服务咨询和图书馆服务咨询的意图分类为例，基于Function Call的意图分类器的工作流如图4所示。

3) 多模态

多模态[18]交互是指系统能够理解和处理多种类型的信息请求和响应的能力，包括但不限于文本、图像、音频和视频等形式。为了支持多模态功能，本文需要对现有的API进行扩展，将content扩展为一个object数组，并允许在一次对话中，传入多个不同类型的object。每个object中以type字段区分输入的类型，例如text、image、audio或video等。对于非文本类型的输入，应该通过上传文件到特定的服务后，以可引用的URL形式输入。

在模型响应上，应使用Markdown格式进行包装，这样可以方便地展示包括图像在内的多种格式的内容，同时保持良好的可读性和兼容性。一种引入图片时的多模态请求和响应示例如表3所示。

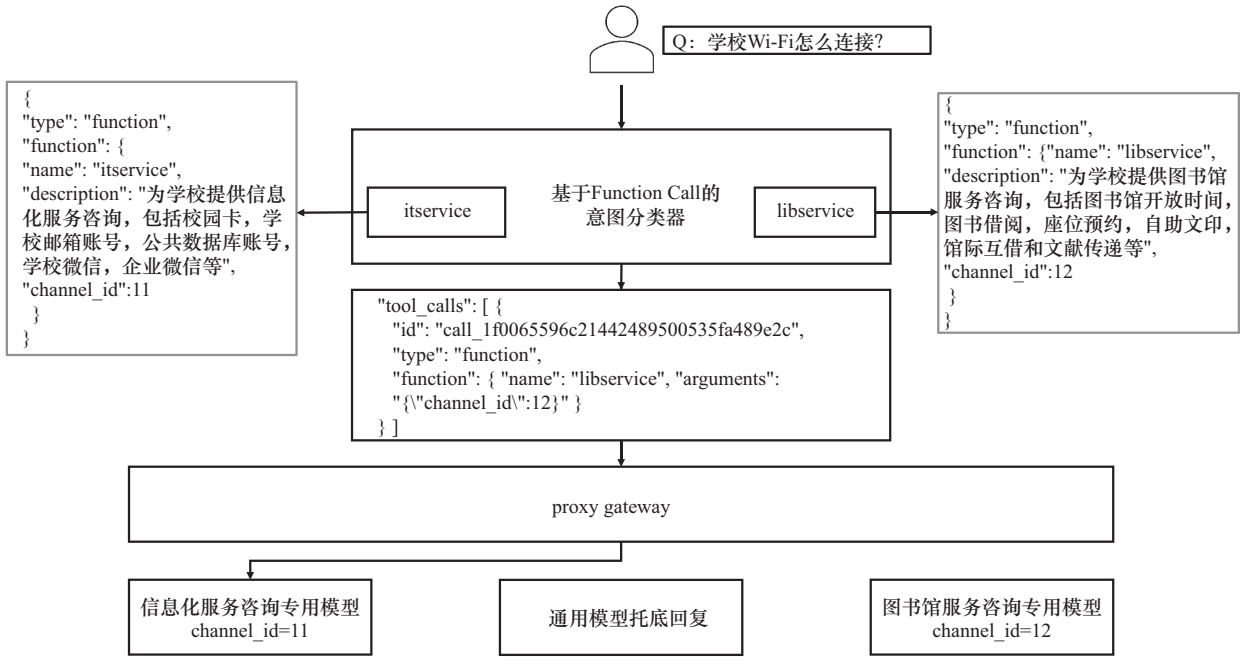


图4 基于Function Call的意图分类器的工作流

表3 多模态的请求和响应示例

类型	示例
request	"content":[{"type":"text","value":"请解释一下这张图的意思"},{"type":"image","value":"https://example.com/image.png"}]
responset	"message":{"role":"assistant","content":"这张图展示的是双缝干涉图案。\n"}

4) 格式化输出

在很多场景里，希望模型的输出按照特定格式呈现，例如指定输出格式为 json。尽管通过 prompts 能够让模型以期望的格式响应内容，但很多时候模型的响应还会混杂一些额外的文本，这是当前基于大语言模型的输出进行自动化处理时所面临的重要难点之一。

表 4 给出了一个通过 prompts 引导模型输出 json 格式的样例和可能的模型输出。

表 4　prompts 引导的 json 格式化样例与输出

prompts	响应	是否符合预期
## system prompt 你必须以 json 格式输出响应，并将 json 部分的内容以 markdown 文本块引用的形式输出，且标注这个文本块是 json 例如 ```json { "colum":"hello" } ``` ## user question 列出 3 种水果和他们的常见价格，价格单位是元/公斤	```json { "fruits": [{ "name": "苹果", "price": 5.0 }, { "name": "香蕉", "price": 3.0 }, { "name": "橙子", "price": 4.5 }] } ``` 好的，这是你想要的 3 种水果 ```json { "fruits": [{ "name": "苹果", "price": 5.0 }, { "name": "香蕉", "price": 3.0 }, { "name": "橙子", "price": 4.5 }] } ```	是 否

显然当出现第 2 种情况时，模型在输出 json 之前的“废话”会使自动化处理失效，而仅依赖 prompts 消除这些“废话”并不总是奏效的。

本文提出一种 API 字段扩充的方案，在请求中增加 response_format 字段，允许用户指定期望的输出格式，例如 json 或者 xml。此时，模型应仅输出文本引用的部分，从而确保输出的内容严格符合结构化规范。

这个方案不需要模型本身支持严格的 json 输出，如果模型本身缺失这方面的能力，通过 API 响应时对数据进行清洗过滤，也能够实现相同的效果。约定 API 的规范而不绑定具体的内部实现，将使规范更容易推广和应用。

1.2　检索增强

在提供具体事实、政策细节或操作指南等信息时，保证信息的准确性和来源的权威性至关重要。此时本文需要引入检索增强（RAG）机制[19]，允许模型通过检索文档获得准确的外部知识，并在回答时给出具体的文档内容引用。RAG 不仅能够提高答案的准确性和可信度，还能帮助用户验证信息来源，从而增强用户对模型的信任。为了能够支持文档引用部分的响应，本文进一步扩展 API 的响应参数，加入了 doc_reference 字段，字段内应包含引用文档的索引 ID（index_id）、标题（title）、文档 ID（doc_id）、文档名称（doc_name）和引用文本（text），使用户可以直接访问文档来源，验证信息的真实性。图 5 展示了一个带文档引用的响应示例。

图 5　带文档引用的响应示例

1.3　智能体

智能体[20]是一种综合了多种技术手段的高级应用形态，它结合了提示词工程、检索增强、搜索

引擎整合等技术，并通过精心设计的工作流编排，使智能体能够在特定任务中给出更佳的表现。它不仅需要能够理解和处理复杂的任务，还需要能够灵活地与现有系统和服务进行集成。

经过定义后的智能体仍然应该遵循前文所述的所有API规范。这意味着智能体的接口设计既要保持与OpenAI兼容的格式，又要能够区分不同的智能体实例。为此，可以通过在URL中加入智能体的标识符（agentid）来实现这一点，例如采用 ttps://{domain}/{agentid}/api/v1/chat/completions的形式。这种方式可以在标识agent的同时，保持OpenAI的接口地址兼容。

2　技术框架

基于第1节给出的规范设计，本文建立了一套大语言模型应用的技术框架，框架遵循模块化设计，并对每个模块的功能给出了清晰的定义，这对于后续的工程实现或者产品选型起到了重要的支撑作用。代理网关、控制中心和前端应用是整个框架的核心部分，需要优先完成。其余模块则是扩展能力，可以分期逐步实现和完善。整体技术框架如图6所示。

2.1　代理网关

代理网关是整个技术框架的核心组件之一，它类似于API网关[21]，但区别在于其主要职责是对不同模型进行适配和路由。通过模型适配器（model adapter）对各个模型进行适配接入，屏蔽底层模型协议差异，提供统一的规范化API。此外，通过RAG适配器（RAG adapter）对不同的RAG解决方案进行适配，同样屏蔽底层RAG实现的差异，提供符合前文所述的规范化RAG接口。在此架构下的模型和RAG被定义为“渠道”，不同的渠道通过代理网关实现统一的路由策略处理，并根据实际需求动态切换响应渠道，从而实现多模型的融合。

在响应请求时，代理网关会先与控制中心联动，校验用户携带的短期令牌，并获取用户的角色以进一步获取角色对应的策略。在策略确定了路由的渠道后，代理网关会调用对应的适配器将请求转发到特定的渠道进行响应。

在执行渠道的路由策略时，代理网关首先会控制中心下发的路由策略进行转发，例如基于Model的策略匹配、基于APP的策略匹配、基于意图的策略匹配等。特别地，当基于意图进行策略路由

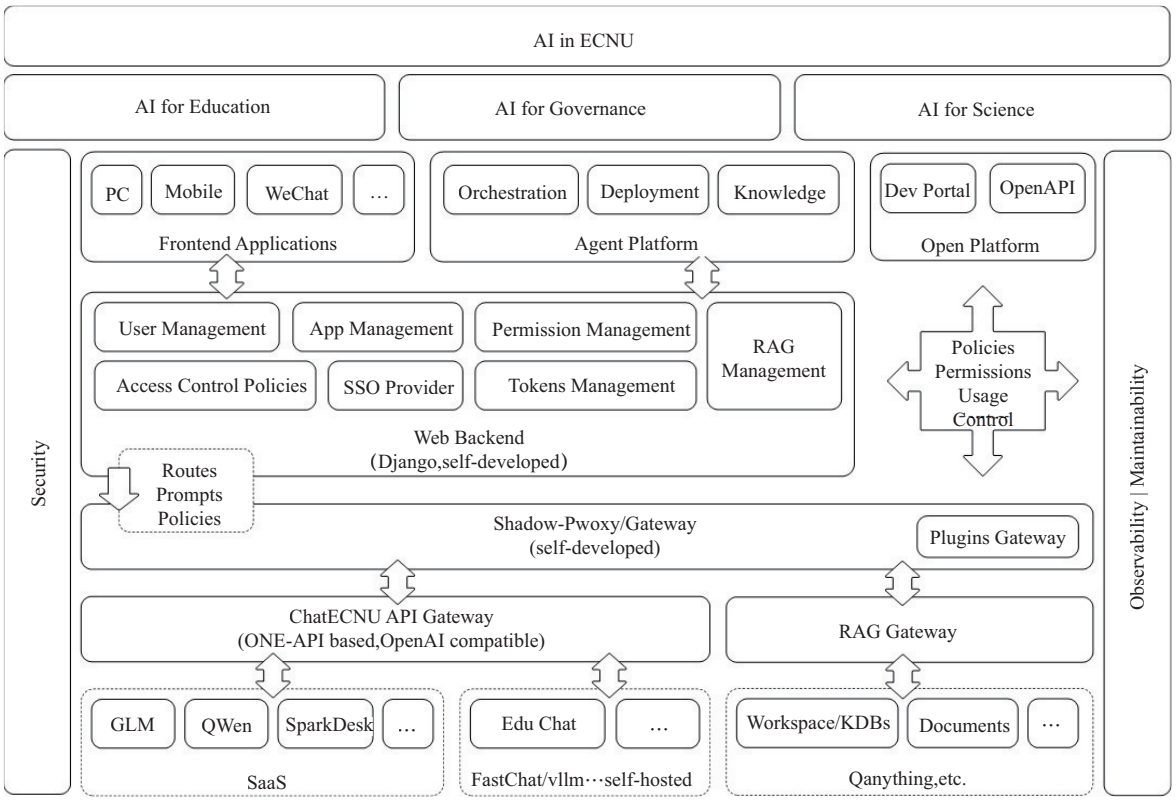


图6　技术框架

时，会根据前文所述的意图分类器执行意图理解，并转发给所命中的渠道。

当请求包含前文所述的 conversationid 时，代理网关会根据存储的会话记录填充上下文，从而支持长期记忆和跨终端的会话同步。

2.2　控制中心

控制中心是技术框架中的另一个核心组件，它是整个框架中的管理控制平面，将身份认证、权限管理、令牌分发、应用管理等统一进行管理。

1) 身份认证。SSO 组件负责实现统一认证的集成，SSO 组件支持主流的认证协议，如 CAS、SAML2、OAuth2 等。

2) 权限管理。定义不同角色和用户组的权限，并通过策略绑定到不同的应用上。本文采取基于策略的动态角色分配模型[22]，从而在不需要批量同步用户的情况下，实现动态的角色分发。基于策略的动态角色分配如表 5 所示。

3) 令牌分发。用户在通过统一身份认证后，控制中心将给这个用户颁发一个令牌，并将用户的基本信息和所绑定的角色以令牌为键值进行缓存。当前端应用使用这个令牌访问代理网关请求大语言模型的响应时，代理网关将通过令牌向控制中心查询令牌所绑定的用户信息和角色，并匹配策略确定其访问应用的权限范围。

4) 应用管理。基于不同的 system prompts，集合知识库或者编排等构建的适配某些特定场景的大语言模型服务，本文称之为框架内的“应用”。不同的应用显然应该有权限上的差异化区分。本文将应用权限分为访问（Access）和管理（Manage）两类，并通过角色标签的形式构建权限关系，从而实现动态灵活的访问应用控制权限。应用权限示例如表 6 所示。

表 6　　应用权限示例

应用	Access 权限	Manage 权限
AI 助教 001	teacherCourse001, studentCourse001, assistantCourse001, manager	teacherCourse001, assistantCourse001, manager
信息化服务助手	default, manager	nic, manager
教案助手	staff, manager	Manager

2.3　前端应用

前端应用是用户与大语言模型交互的直接界面，也是技术框架中的核心组成部分。前端的设计旨在提供丰富、直观的用户体验，并支持多模态内容的加载和展示，确保用户能够便捷地获取所需信息和服务。前端实现时需要重点支持的部分如下。

1) Markdown 支持。由于大语言模型的内容输出还是以文本为主，输出的内容呈现效果取决于对文本标记的渲染效果。因此只要能够支持 Markdown 及其常见的扩展语法，就能够在前端优雅地展示各种格式，如加粗、斜体、列表、链接等。

特别地，对于数学公式，可以通过 Markdown Latex 扩展组件予以支持；对于流程图，可以通过 Markdown Mermaid 扩展组件予以支持；对于图像、视频等多媒体的加载，也可以通过

表 5　　基于策略的动态角色分配

角色	定义语法	角色含义
default	userId != ''	默认角色
staff	userType == '教职工'	教职工
manager	userId in ['001', '002']	管理员
nic	department='信息化治理办公室'	信息办
teacherCourse001	courseid = '001' and courseType='teacher'	001 课的授课老师
studentCourse001	coursed='001' and courseType='student'	001 课的上课学生
assistantCourse001	coursed='001' and courseType='assistant'	001 课的助教

Markdown 引用 URL 的形式进行渲染。通过 Markdown 的丰富能力，能够把纯文本的模型输出渲染出“多模态”的前端呈现效果，对用户体验将有很好的提升。

2) 弹窗式的文档引用。为了增强信息的可信度，前端应用还支持弹窗式的文档引用功能。当用户收到带有文档引用的响应时，可以通过点击文档链接在弹窗中查看原始文档的内容，从而验证信息的来源和准确性。弹窗式的文档引用示例如图 7 所示。

2.4 智能体编排平台

智能体编排平台是技术框架中后期的拓展能力部分，它负责管理和编排智能体（Agent），使它们能够执行复杂的任务。智能体是一种高度集成的应用形态，它们结合了提示词工程、检索增强、搜索引擎整合等技术，并通过精心设计的工作流编排，以适应特定的任务需求。智能体编排平台应该包含以下功能。

1) Agent 管理。平台提供了一套完整的生命周期管理工具，包括智能体的创建、配置、启动、监控和更新。管理员可以轻松地定义智能体的行为逻辑和触发条件，并通过平台提供的界面进行配置。

2) 编排协作。智能体之间可以协同工作，完成更复杂的任务。编排平台允许开发者定义智能体间的交互逻辑，例如传递数据、触发事件或同步状态，从而形成一个整体的解决方案。

3) 知识库管理。每个智能体都可以关联一个或多个知识库，这些知识库包含了智能体执行任务所需的上下文信息和数据。平台支持知识库的创建、更新和版本控制，确保智能体能够访问到最新和最准确的信息。

4) 定制化工作流。平台支持定制化的工作流设计，允许用户根据特定场景的需求，灵活地配置智能体的执行流程。这使智能体能够在各种复杂的环境中提供精准的服务

2.5 开放平台

开放平台是技术框架走向生态成熟、协作共建的重要组成机制，它面向开发者提供了一系列工具和服务，使他们能够利用大语言模型的能力开发新的应用和服务。通过吸引更多的开发者参与到大语言模型应用的开发中，逐步形成应用生态和正反馈的循环。一个良好的开放平台应该具备以下能力。

1) 原生接口开放。开放平台提供了原生的 API，允许开发者直接调用大语言模型的功能。这

图 7 弹窗式的文档引用示例

些接口遵循规范化的设计，确保与其他系统和服务可无缝集成。

2) 嵌入式开放。除了传统的API，开放平台还支持嵌入式开放模式，允许开发者将大语言模型的能力直接嵌入自己的产品和服务中，使第三方应用的大模型接入更快捷和高效。

3) 开发者门户。开发者门户为开发者提供了全面的管理工具，包括应用注册、配置管理、访问控制和监控等功能。开发者可以通过管理平台轻松地管理自己的应用和服务。

4) 开放市场。开放市场是一个集中展示和分发基于大语言模型的应用和服务的平台。开发者可以将自己的作品发布到市场上，在通过学校审核后即可被其他师生访问使用。

结合这一机制还可以进一步举办大语言模型的开发大赛，鼓励校园内的开发者们开发并发布自己的大模型应用，学校投入孵化其中的优秀创意项目，充分挖掘校内师生的优秀创意。

2.6　安全与可观测性

安全与可观测性平面是整个技术框架中的保障组件，旨在确保系统的安全性、稳定性和可维护性。

安全方面，除了常规的信息系统安全防护以外，还需要额外考虑大语言模型本身内容输出上的安全性。一方面，需要通过前期的测试验证，确保接入的模型安全可靠；另一方面，需要一些外置的检测机制作为安全备份。因此，需要引入敏感词过滤器，在必要的时候对大语言模型的输出进行过滤，以确保模型的输出内容安全合规。

可观测性方面，所有的模块均应该以兼容Prometheus的形式暴露到metric接口，使所有兼容Prometheus生态的可观测性工具都可以采集到这些指标并构建可观测性和告警体系。代理网关的可观测性指标示例如图8所示。

3　场景应用

在第2节技术框架设计理念下，华东师范大学开发了学校的大语言模型平台ChatECNU，除了提供基本的大语言模型能力以外，本文更看重大语言模型与具体的教育场景相结合，为教育信息化赋能更大的价值。本文在客服咨询、课堂教学、课程制作3个教育场景进行了大语言模型的应用探索，取得了比较好的效果。

3.1　基于意图理解的精准检索：校园百事通

利用大语言模型优秀的自然语言对话能力来实现客服咨询机器人，是最常见的大语言模型应用场景之一。为了让客服机器人的答复更准确，本文通过1.2节中描述的检索增强方案，引入外部知识库，并在响应中以文档引用的方式向用户展示回复的文档来源。

然而，校园业务的咨询范畴非常广泛，过分混杂的知识库有时候会降低RAG的检索精度。例如在面向本科生和研究生的学籍管理、学习生活等场景内，文档内容在形式上有一定的相似性，但实质业务要求上并不一致。

本文通过1.1节中提出的基于Function Call的意图分类器，实现了一个基于意图理解的RAG类应用：校园百事通。通过意图分类器，本文能够根

```
# HELP proxy_gateway_requests_total Total number of requests processed by the proxy gateway.
# TYPE proxy_gateway_requests_total counter
proxy_gateway_requests_total 15000

# HELP proxy_gateway_requests_inprogress Number of requests currently being processed by the proxy gateway.
# TYPE proxy_gateway_requests_inprogress gauge
proxy_gateway_requests_inprogress 5

# HELP proxy_gateway_errors_total Total number of errors encountered by the proxy gateway.
# TYPE proxy_gateway_errors_total counter
proxy_gateway_errors_total 100

# HELP proxy_gateway_routes_total Total number of requests routed to different channels by the proxy gateway.
# TYPE proxy_gateway_routes_total counter
proxy_gateway_routes_total{channel="sparkdesk"} 2000
proxy_gateway_routes_total{channel="qwen2"} 1500
proxy_gateway_routes_total{channel="chatglm"} 1000
proxy_gateway_routes_total{channel="deepseek"} 500
```

图8　代理网关的可观测性指标示例

据用户提出的问题动态加载不同的知识库，以减少知识之间的干扰，提高检索的准确性。同时基于Function Call的意图分类器不需要额外训练，普通管理员也能够通过Function定义描述的方式对分类器进行优化调整，使整体方案非常灵活和轻量。图9展示了校园百事通的运行流程。

3.2　数据驱动的智能课程助手：师大AI助教

课程的AI助教也是大语言模型在教育信息化中落地的重要场景之一。然而，开发AI助教并不简单地等价于开发一个RAG类应用。本文发现这里至少存在3个难点：1) 课程教材大量的公式和表格难以被大模型理解；2) AI助教的应用权限与选课/退课情况相关，动态选课/退课导致权限变动频繁，权限配置的人工维护成本太高；3) AI助教答疑仅与学生交互，与授课老师关联度小，难以对课堂教学提供更多的能力支持。

本文针对这3个难点，提出了以数据驱动的AI助教开发理念。1) 在课程教材方面，使用了PDF-to-Mardown模型来识别教材中的公式、表格等内容，并结合人工干预和调整，将课程教材内容全部转换为Markdown，显著提高了模型对教材内容的理解能力。2) 在权限方面，与教务系统进行了数据集成对接，实时获取了课程的选课名单，并自动构建AI助教的访问权限，大大降低了人工的权限维护负担；3) 在教学反馈方面，利用AI助教的日志数据，再次调用大语言模型的能力提取学生提问的关键词和摘要，并生成课程反馈报告，为授课教师提供教学改进的建议和支持。图10展示了师大AI助教的运行流程。

3.3　数字人赋能的微课合成工具：微课工坊

在线教学[23]已经成为当前高校课程教学中的重要部分，然而在实践中虽然高校都建立自己的在线教学平台，平台的使用率却不是非常理想。本文在调研后发现，很多老师并非没有开设线上课程的意愿，但往往受限于课程视频录制的门槛偏高，自身精力有限而作罢。在已经开设了在线课程的内容中，也普遍存在教学内容已经更新但是课程视频未及时更新的情况，这同样是因为课程视频的录制对老师而言是一种负担。

因此本文结合数字人技术与大语言模型技术，老师只需要上传PPT，一键即可合成基于PPT内容理解的微课视频。显著降低了课程视频的制作门槛，提升了老师开设在线课程的意愿，为在线教学的推广提供了重要支持。

整个微课视频的合成过程分为4个步骤。

1) 内容理解。平台接收到老师上传的PPT后，首先进行内容解析，将PPT的页面元素、备注、文字等解析成每一页的内容大纲，并通过大语言模型重新组织后生成发言脚本。

2) 语音合成。根据第一步所生成的发言脚本，使用语音合成技术合成语音，获得每一页的播放时长。

3) 数字人合成。根据获得的语音音频，采用

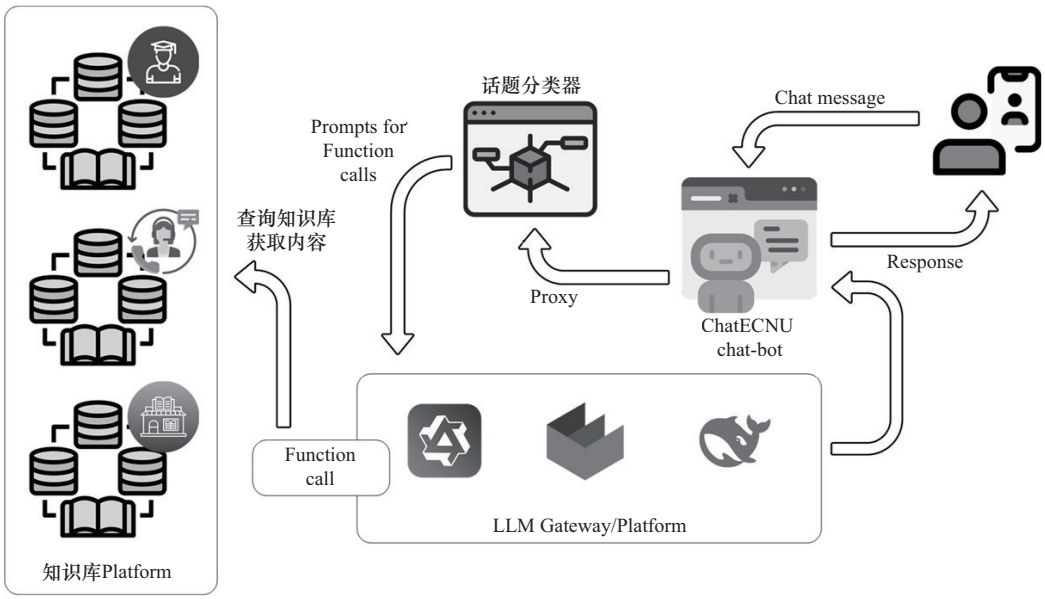


图9　校园百事通的运行流程

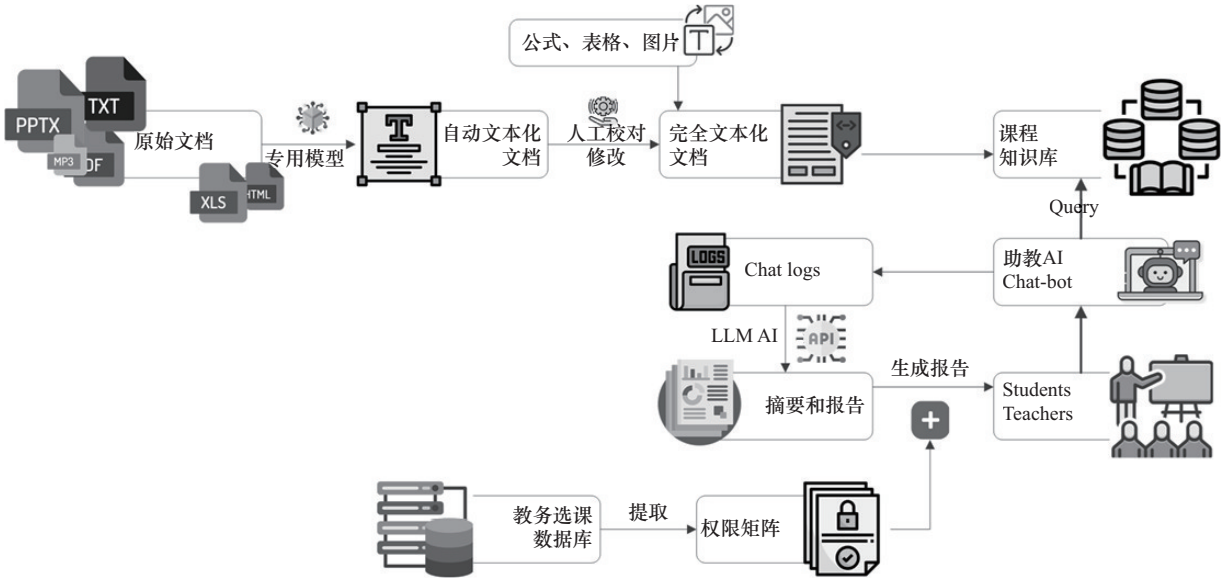


图 10　师大 AI 助教的运行流程

微表情合成的方案，通过静态的用户照片生成带有微表情的数字人形象，并和语音部分对齐唇型，使数字人的播报体验更真实。

4) 微课合成。将合成的数字人形象、语音和 PPT 页面整体合成渲染，生成最终的微课视频。

图 11 展示了微课工坊的合成技术路径。

在整个过程中，老师也可以部分介入修改发言脚本以使其更符合课程教学的要求，总之微课工坊大大降低了微课视频的制作门槛，制作课程视频对教师而言不再是一种负担，老师开设在线课程的意愿得到了明显的增强。

4　结束语

本文提出了一套大语言模型在教育信息化中的应用规范与技术框架，旨在解决高等教育领域中大语言模型应用复杂、模型深度绑定以及可复用性差等问题。

华东师范大学的 3 个案例表明，基于这套规范性框架的应用开发具备良好的交互性和扩展性，底层能力的高效复用显著提高了应用开发的效率。

本文预计大语言模型将在教育信息化领域发挥越来越重要的作用，在这个过程中与教育场景的深度融合是关键。本文的研究成果为推动人工智能技术在教育领域的广泛应用提供了参考依据，具有一

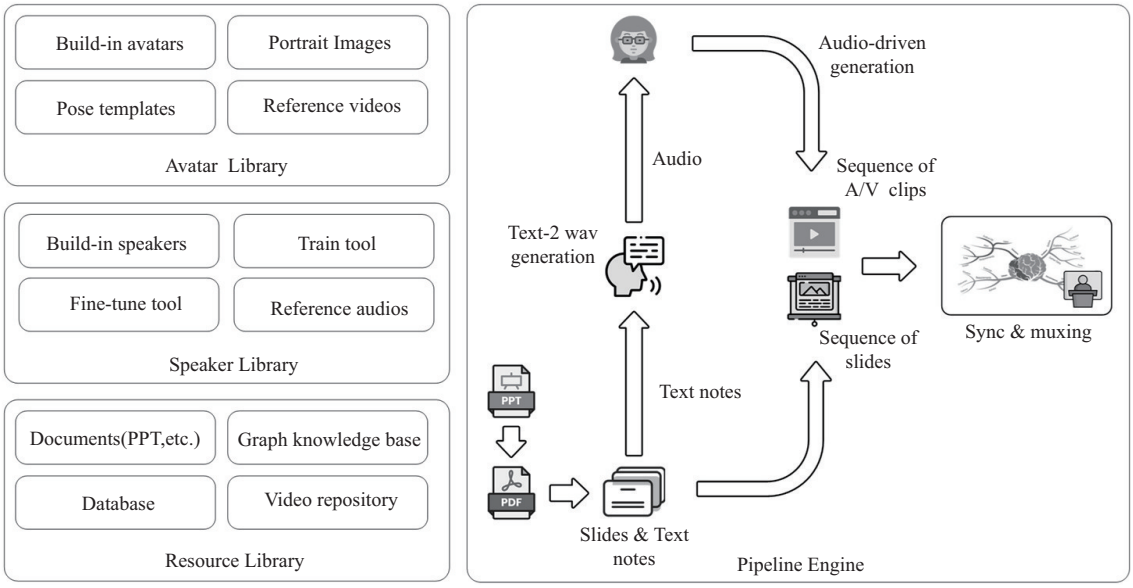


图 11　微课工坊的合成技术路径

定的指导意义。

参考文献：

[1] 刘挺. 从ChatGPT谈大语言模型及其应用[J]. 语言战略研究, 2023, 8(5): 14-18.
LIU T. A look into large language models and its applications from the perspective of ChatGPT[J]. Chinese Journal of Language Policy and Planning, 2023, 8(5): 14-18.

[2] 王昀, 胡珉, 塔娜, 等. 大语言模型及其在政务领域的应用[J]. 清华大学学报(自然科学版), 2024, 64(4): 649-658.
WANG Y, HU M, TA N, et al. Large language models and their application in government affairs[J]. Journal of Tsinghua University (Science and Technology), 2024, 64(4): 649-658.

[3] 李夏风, 傅小龙. 大模型在安防领域的实践应用: 以云从科技从容大模型安防领域实践应用为例[J]. 中国安防, 2023(9):49-53.
LI X F, FU X L. Practical application of large model in security field—taking Yuncong technology as an example[J]. China Security & Protection, 2023(9):49-53.

[4] 符荣鑫, 杨小华. AIGC语言模型分析及其高校图书馆应用场景研究[J]. 农业图书情报学报, 2023, 35(7): 27-38.
FU R X, YANG X H. Analysis of AIGC language models and application scenarios in university libraries[J]. Journal of Library and Information Science in Agriculture, 2023, 35(7): 27-38.

[5] LI Z Y, CHEN Y F, ZHANG X L, et al. BookGPT: a general framework for book recommendation empowered by large language model[J]. arXiv Preprint, arXiv: 2305.15673, 2023.

[6] 吴武清, 赵煜东, 赵越, 等. GPT等大语言模型在会计与审计中的应用[J]. 国际商务财会, 2023(22): 81-87.
WU W Q, ZHAO Y D, ZHAO Y, et al. Application of GPT and other large language models in accounting and auditing[J]. Finance and Accounting for International Commerce, 2023(22): 81-87.

[7] HUANG A H, WANG H, YANG Y. FinBERT: a large language model for extracting information from financial text[J]. Contemporary Accounting Research, 2023, 40(2): 806-841.

[8] 卢宇, 余京蕾, 陈鹏鹤, 等. 多模态大模型的教育应用研究与展望[J]. 电化教育研究, 2023, 44(6): 38-44.
LU Y, YU J L, CHEN P H, et al. Study and prospect of the applications of large multimodal models in education[J]. e-Education research, 2023, 44(6): 38-44.

[9] LUO Y W, YANG Y. Large language model and domain-specific model collaboration for smart education[J]. Frontiers of Information Technology & Electronic Engineering, 2024, 25(3): 333-341.

[10] WANG L, MA C, FENG X Y, et al. A survey on large language model based autonomous agents[J]. Frontiers of Computer Science, 2024, 18(6): 186345.

[11] 邬贺铨. 大模型融入云平台，信息化走向数智化[J]. 重庆邮电大学学报(自然科学版), 2024, 36(1): 1-8.
WU H Q. Integration of large-scale models and cloud, transition from informatization to digital intelligence[J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2024, 36(1): 1-8.

[12] 张春红, 杜龙飞, 朱新宁, 等. 基于大语言模型的教育问答系统研究[J]. 北京邮电大学学报(社会科学版), 2023, 25(6): 79-88.
ZHANG C H, DU L F, ZHU X N, et al. Educational question-answering systems based on large language model[J]. Journal of Beijing University of Posts and Telecommunications (Social Sciences Edition), 2023, 25(6): 79-88.

[13] 贾积有, 王光迪. 应用大语言模型快速有效分析教育访谈文本[J]. 中国教育信息化, 2023, 29(12): 34-41.
JIA J Y, WANG G D. Analyzing educational interview texts using large language model[J]. Chinese Journal of ICT in Education, 2023, 29(12): 34-41.

[14] 汪张龙. 认知智能大模型加速教育考试数字化转型[J]. 中国考试, 2023(8): 11-18.
WANG Z L. Digital transformation of educational examinations with cognitive intelligence models[J]. Journal of China Examinations, 2023(8): 11-18.

[15] DINH H, TRAN T K. EduChat: an AI-based chatbot for university-related information using a hybrid approach[J]. Applied Sciences, 2023, 13(22): 12446.

[16] DAN Y H, LEI Z K, GU Y Y, et al. EduChat: a large-scale language model-based chatbot system for intelligent education[J]. arXiv Preprint, arXiv: 2308.02773, 2023.

[17] 吴晓东, 王鹏. Html5的通信机制及效率的研究[J]. 长春理工大学学报(自然科学版), 2011, 34(4): 159-163.
WU X D, WANG P. Research on communication mechanism and efficiency of Html5[J]. Journal of Changchun University of Science and Technology (Natural Science Edition), 2011, 34(4): 159-163.

[18] 赵朝阳, 朱贵波, 王金桥. ChatGPT给语言大模型带来的启示和多模态大模型新的发展思路[J]. 数据分析与知识发现, 2023, 7(3): 26-35.
ZHAO C Y, ZHU G B, WANG J Q. The inspiration brought by ChatGPT to LLM and the new development ideas of multi-modal large model[J]. Data Analysis and Knowledge Discovery, 2023, 7(3): 26-35.

[19] SIRIWARDHANA S, WEERASEKERA R, WEN E, et al. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering[J]. Transactions of the Association for Computational Linguistics, 2023, 11: 1-17.

[20] NI B, BUEHLER M J. MechAgents: large language model multi-agent collaborations can solve mechanics problems, generate new data, and integrate knowledge[J]. Extreme Mechanics Letters, 2024, 67: 102131.

[21] 冯骐, 沈富可. 高校能力开放平台中的API网关设计与实现[J]. 中国教育信息化, 2021, 27(3): 61-66.
FENG Q, SHEN F K. Design and implementation of API gateway in university capacity open platform[J]. The Chinese Journal of ICT in Education, 2021, 27(3): 61-66.

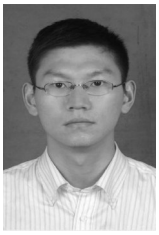
[22] 冯骐, 马晨辉. 基于策略的动态角色分配模型及应用[J]. 计算机与数字工程, 2024, 52(1): 75-80.
FENG Q, MA C H. Policy-based dynamic role assignment model and application[J]. Computer & Digital Engineering, 2024, 52(1): 75-80.

[23] 张鹏高, 冯骐, 罗兰. 中国高等在线教育发展现状探究[J]. 中国教育信息化, 2016, 22(1): 18-21.
ZHANG P G, FENG Q, LUO L. On the development of higher online education in China[J]. The Chinese Journal of ICT in Education, 2016, 22(1): 18-21.

[作者简介]

徐刚（1983– ），男，江苏常熟人，华东师范大学工程师，主要研究方向智能系统、软件工程等。

刘志鹏（1980– ），男，山东文登人，华东师范大学助理研究员，主要研究方向为信息化理论与方法、教育数据共享与可视化、数据安全等。



冯骐（1989– ），男，上海人，华东师范大学高级工程师，主要研究方向为教育数据治理、教育联邦认证、人工智能等。

沈富可（1967– ），男，山东莱西人，博士，华东师范大学教授级高级工程师，主要研究方向为教育信息化治理、智能网络等。