

华中科技大学本科毕业论文

基于大模型的以学生为中心的智能教学平台设计与实现 ——以本科生《电磁场与电磁波》和研究生《学术规范与论文 写作》为例

学生姓名：张三

学 号：2020123456

专 业：计算机科学与技术

学 院：计算机科学与技术学院

指导教师：李四教授

2026 年 5 月

原创性声明

本人声明所呈交的毕业论文是本人在导师指导下进行的研究成果。除文中已经注明引用的内容外，本论文不包含任何他人已经发表或撰写的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：_____

日期：_____

授权使用声明

本人同意华中科技大学保存并向国家有关部门或机构送交本论文的复印件和电子版，允许本论文被查阅和借阅。本人授权华中科技大学将本论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或其他复制手段保存和汇编本论文。

作者签名：_____

日期：_____

摘要

本文面向本科生《电磁场与电磁波》与研究生《学术规范与论文写作》教学中“反馈成本高、个体差异大、过程难追踪”等问题，设计并实现一套以学生为中心的智能教学平台。平台以写作任务为主线，构建“写作提交—智能分析—个性化反馈—教师复核—学习画像更新”的闭环，支持文献综述、课程论文、学位论文与摘要等写作类型的差异化评价与针对性建议。系统采用前后端分离与服务化架构：后端负责课程与写作业务、RBAC 权限控制以及学习事件记录；AI 服务通过 OpenAI-compatible 接口调用大语言模型，提供对话辅导、引导式学习、写作分析、工具调用与 GraphRAG 检索增强能力。为降低幻觉并提升可追溯性，平台针对《学术规范与论文写作》课程，将课程规范、优秀范文片段与常见错误库构建为可检索向量知识库，在生成反馈时注入证据片段并要求引用编号，支持事后审核与复核。为支撑模型定制，本文实现了覆盖云侧与端侧的后训练工具链：在写作场景建立 LoRA/QLoRA 训练评测流程，并在 2026-02-10 完成基于 ms-swift 的端侧 qwen3-0.6B 本机 LoRA 微调与客户端联测闭环。客户端侧采用跨端方案，通过共享 types 与统一 SDK 保持 Web 与移动端的 API 契约一致，并在 UI 层针对不同终端做适配。实践表明，该系统能够以较低的工程成本实现可控、可追溯的写作辅导流程，并为规模化数据驱动的模式迭代提供基础。

关键词：学术写作；学习档案；检索增强生成；GraphRAG；LoRA/QLoRA；跨端统一 SDK

Abstract

This thesis targets graduate-level academic writing instruction, where feedback is costly, student needs are heterogeneous, and progress is difficult to track. We design and implement a student-centric intelligent teaching platform that forms a closed loop: writing submission, automated analysis, personalized feedback, instructor review, and profile updates. The platform is writing-type aware and supports literature reviews, course papers, thesis sections, and abstracts with type-specific rubrics and prompting strategies. We adopt a service-oriented architecture with a Go backend for course and writing workflows, RBAC, and learning-event logging, and a FastAPI AI service that calls an OpenAI-compatible LLM endpoint and exposes chat tutoring, writing analysis, tool calling, and GraphRAG-based retrieval augmentation. To reduce hallucinations and improve traceability, we build a writing knowledge base from course guidelines, exemplars, and common error patterns, retrieve evidence snippets, and require citation markers in generated feedback. For model customization, we provide an end-to-end post-training toolchain across cloud and edge settings, including LoRA/QLoRA pipelines and a local ms-swift LoRA run for Qwen3-0.6B on 2026-02-10 with client integration verification. On the client side, we support both Web and Mobile via shared types and a unified SDK to keep API contracts consistent while adapting UI for each platform. The implementation demonstrates a practical and controllable workflow for academic writing assistance and lays the groundwork for data-driven iterative improvement.

Keywords: academic writing; student profile; retrieval-augmented generation; GraphRAG; LoRA/QLoRA; cross-platform SDK

目录

原创性声明	ii
授权使用声明	iii
摘要	iv
Abstract	v
第一章 绪论	1
1.1 研究背景与意义	1
1.2 国内外研究现状	1
1.3 研究内容与任务要求	2
1.3.1 课题内容	2
1.3.2 课题任务要求	3
1.4 创新点	3
1.5 论文结构	3
第二章 方法与实现	4
2.1 需求分析与总体设计	4
2.2 系统架构	4
2.3 统一契约与跨端共享	5
2.4 学生中心数据模型与学习事件	5
2.5 引导式学习与薄弱点追踪	5
2.6 写作类型感知的智能分析服务	5
2.7 GraphRAG 知识库与检索增强生成	6
2.7.1 学术规范向量知识库构建过程	6
2.8 工具调用与可验证能力	6
2.9 模型后训练与评测管线	7
2.9.1 数据转换与质量校验	7
2.9.2 本机 LoRA 微调配置	7
2.9.3 训练结果与最佳检查点选择	8
2.9.4 部署与客户端联测	8

2.10 系统原型实现与企业微信集成	9
2.11 系统测试与评估	9
第三章 总结与展望	10
附录 A 附录	11
参考文献	12

插图

2.1 平台总体架构示意（占位）	4
----------------------------	---

表格

2.1	端侧微调数据集划分 (2026-02-10)	7
2.2	端侧微调任务类型分布	7
2.3	端侧本机 LoRA 训练配置 (ms-swift)	8
2.4	端侧 LoRA 训练关键指标 (2026-02-10)	8
2.5	端侧模型客户端联测结果 (12 条固定用例)	9

第一章 绪论

1.1 研究背景与意义

研究生课程《学术规范与论文写作》中研究生阶段的专业学术写作是科研训练的重要组成部分，其难点不在于“写出一段文字”，而在于持续地完成选题定位、论证组织、证据使用与学术表达规范的综合训练。以研究生《学术规范与论文写作》课程为例，学生往往需要同时处理语法与措辞准确性、段落结构与逻辑衔接、引用与证据一致性问题；不同学生在语言基础、科研经验与表达习惯上存在显著差异，使得“一次性评分”很难形成有效的改进闭环。写作教学高度依赖过程性反馈：从选题与大纲，到草稿迭代与修改记录，再到教师或助教的具体建议，这些信息共同构成学生能力发展的轨迹。如何把这些过程性信息系统化沉淀，并转化为可执行的个性化建议，是写作课程面临的核心问题之一。

近年来，Transformer 架构^[1]推动大语言模型在自然语言理解与生成方面取得突破，对话式模型（如 ChatGPT）^[2]在解释、改写与建议生成方面表现突出，为写作辅导提供了新的技术手段。然而，大模型在教育场景中仍面临两个突出挑战：其一是建议的可靠性与可追溯性不足，模型可能给出“看似合理但缺乏依据”的写作建议；其二是个性化不足，模型难以在缺少学习过程数据的情况下稳定地贴合学生水平与课程 rubric。检索增强生成（RAG）^[3]与知识图谱^[4]为上述问题提供了可行路径：通过将课程规范、范文片段与常见错误案例构建为可检索的证据库，把生成行为锚定在可引用的片段之上，从而降低幻觉并便于教师复核。

基于此，本文以《学术规范与论文写作》课程为文本任务试点，并以《电磁场与电磁波》课程作为工具型任务的配套验证。基于此，本文以研究生《学术规范与论文写作》与本科生《电磁场与电磁波》为试点场景，提出并实现一套“以学生为中心”的智能教学平台：围绕学生学习过程建立档案与事件记录，结合写作类型感知的评估与反馈策略，并通过 GraphRAG 提供可追溯的证据引用，最终形成“数据—分析—干预—反馈”的闭环。

1.2 国内外研究现状

学术写作研究与写作教学长期关注写作过程与文本产出之间的关系。写作过程模型（如认知过程理论）强调规划、表达与修订在写作中的循环作用，为“过程性反馈”提供了理论依据^[5]。面向研究生的学术写作教学实践则更强调体裁（genre）意

识、论证结构与学术表达规范，例如对文献综述、研究计划与论文各部分的功能定位进行训练^[6-7]。与此同时，自动作文评估与写作反馈系统逐步成熟，能够对语法、篇章结构或整体质量给出评分与提示^[8]，但很多系统仍以“单次评分”为主，难以结合学生历史轨迹给出稳定的个性化建议。

在学习分析与学生建模方面，学习管理系统（LMS）为学习过程数据采集提供了基础，但多数分析仍停留在访问次数、学习时长等行为统计层面。对于写作课程而言，真正具有教学价值的信号往往来自文本本身与修改过程，例如结构性问题的反复出现、论证链条缺失、证据引用不规范等。这类信号需要将文本分析结果与学习事件结合，才能形成可解释的能力画像。

在大语言模型辅助写作方面，对话式模型可以生成润色、改写与建议，但其输出质量与一致性受提示词、上下文与模型能力影响较大。当模型缺少课程材料支撑或未被约束时，容易产生“建议看似合理但难以复核”的问题。检索增强生成（RAG）^[3]通过先检索证据再生成，将答案锚定在外部知识之上，是当前提升可追溯性的主流方向；知识图谱^[4]则可将课程中的概念、规范与示例按关系组织，支持图结构扩展与片段引用。在教育场景中，将 RAG 与课程 rubrics、教师复核机制结合，是实现“可控输出”的关键路径。

1.3 研究内容与任务要求

本文围绕《学术规范与论文写作》课程场景的“过程性数据沉淀与个性化辅导”开展研究，并以《电磁场与电磁波》课程任务验证工具调用可行性，主要内容与任务要求如下。

1.3.1 课题内容

- (1) 以学生为中心的数据模型与学习事件体系设计，实现写作提交、对话辅导、学习时长等过程性数据的可追踪沉淀。
- (2) 写作类型感知的智能分析与反馈策略设计，支持文献综述、课程论文、学位论文、摘要等类型的差异化 rubric 与结构化输出。
- (3) GraphRAG 知识库构建与检索增强生成流程实现，使关键建议绑定证据片段并可引用溯源。
- (4) 面向跨端的系统原型实现（Web/Mobile），抽取共享 types 与统一 SDK，保证多端一致的 API 契约与体验。
- (5) 模型训练与评测管线实现：数据规范、LoRA/QLoRA 微调脚本与回归评测脚本，支撑后续 100k 规模数据的迭代。

1.3.2 课题任务要求

- (1) 掌握大语言模型的基本原理与工程化接入方式，能够通过 OpenAI-compatible 接口完成对话、写作分析与结构化输出。
- (2) 掌握写作课程场景下的数据组织方法，能够将写作样本、反馈与过程性事件统一建模并支持查询与聚合。
- (3) 完成一个可部署的系统原型，覆盖写作提交、分析、反馈与复核的端到端流程，并支持跨端访问。
- (4) 实现可追溯的 RAG 辅导能力：当证据不足时能追问或拒答，并避免编造引用。
- (5) 完成训练与评测工具链的实现与验证，为后续模型微调提供可复现基础。

1.4 创新点

本文工作的创新点主要体现在以下几个方面。

- (1) 面向写作课程构建“学习档案 + 学习事件”的学生中心闭环，强调过程性数据沉淀与纵向追踪，使反馈不仅面向一次作业，更面向能力发展。
- (2) 提出写作类型感知的评估与反馈框架，以类型化 rubric 与结构化输出降低反馈的主观性与随机性，提升建议的可执行性。
- (3) 引入 GraphRAG 形成可追溯的证据引用链路：将课程规范与范文片段构建为知识库，生成反馈时强制引用片段编号，降低幻觉并便于复核。
- (4) 采用跨端共享的 types 与统一 SDK 保障多端一致的 API 契约，将平台能力沉淀为可复用的工程化组件。

1.5 论文结构

本文共分三章：第一章介绍研究背景、相关研究与课题内容；第二章给出需求分析、系统架构与关键模块设计实现，并说明训练与评测管线；第三章总结工作并展望后续改进方向。

第二章 方法与实现

2.1 需求分析与总体设计

平台面向管理员、教师/助教与学生三类主要角色，核心需求可归纳为三条主线：其一是**过程性数据沉淀**，能够围绕写作提交、修改与对话辅导记录学习事件，并形成长期可追踪的学生画像；其二是**写作类型感知的反馈**，针对文献综述、课程论文、学位论文与摘要等不同类型提供差异化 rubric 与结构化建议；其三是**可控与可追溯**，在大模型生成建议时提供证据引用与复核入口，降低幻觉带来的教学风险。

非功能需求方面，系统需要具备可扩展与可维护的工程结构（便于迭代模型与课程模块）、清晰的权限边界（避免越权访问与答案泄露）、以及跨终端一致的调用契约（避免 Web/Mobile 接口分叉）。基于上述需求，本文采用前后端分离与服务化架构，将教学业务、AI 能力与检索索引解耦，通过统一鉴权与模块门控策略保证能力可治理。

2.2 系统架构

系统整体采用“客户端—后端业务—AI 服务—检索/存储”的分层结构：客户端提供 Web 与移动端入口；后端提供业务 API、JWT 鉴权、RBAC 权限与课程模块门控；AI 服务负责提示模板、写作分析与对话能力的编排，并通过 OpenAI-compatible 接口调用上游大模型推理服务；GraphRAG 作为可选组件提供课程知识库检索与引用溯源。整体架构如图 2.1 所示。

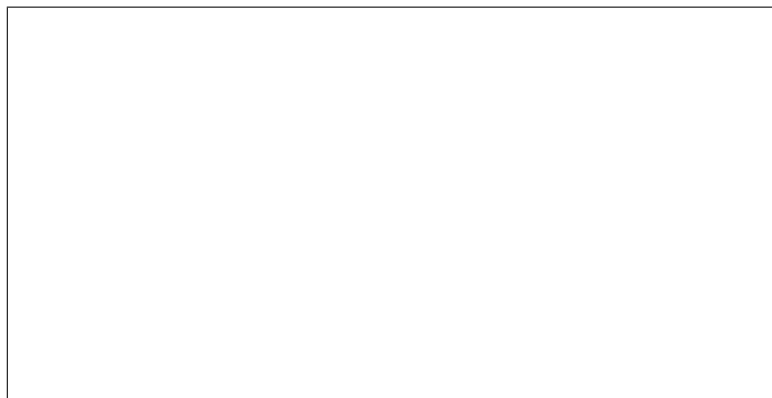


图 2.1: 平台总体架构示意（占位）

2.3 统一契约与跨端共享

为避免跨端开发中出现“同一业务多套接口/字段”的问题，系统采用 Monorepo 组织代码，并抽取共享包沉淀 types 与统一 SDK。共享 SDK 负责：(1) 统一请求层（鉴权头、错误归一、超时与重试策略）；(2) 以类型定义约束前后端契约，减少字段不一致导致的运行时错误；(3) 为 Web/Mobile 提供一致的 API 调用方式，使平台差异主要集中在 UI 与交互层。该设计降低了多端协作成本，并为后续在不同课程模块间复用能力提供基础。

2.4 学生中心数据模型与学习事件

平台以学生为中心组织数据。在课程层面，写作提交被建模为可追踪的业务对象：包含写作类型、标题与内容、提交时间、AI 分析结果与教师反馈等字段；在过程层面，系统记录关键学习事件（如写作提交、写作分析完成、对话辅导、学习时长心跳等），用于后续聚合形成学生画像。画像不仅包含“分数”，更强调可解释的能力维度（例如结构清晰度、证据使用、学术语气与引用规范），从而支持纵向对比与个性化干预。

2.5 引导式学习与薄弱点追踪

除“写作提交—分析—反馈”流程外，平台提供面向过程性辅导的引导式学习能力（guided learning）：系统首先为某一学习主题生成 3-6 步的学习路径（learning path），随后以苏格拉底式提问引导学生逐步完成每一步。例如，在写作课程中，学习主题可围绕 thesis statement、段落结构、证据使用与引用规范等展开，系统会在每轮对话中只提出一个关键问题，并根据学生回答的完整性决定是否进入下一步，从而把复杂能力训练拆解为可管理的阶段。

实现上，AI 服务提供 /v1/chat/guided 端点，使用会话状态（session_id）维护学习目标、当前步骤与路径结构，并在首轮由模型输出 JSON 路径以便前端渲染进度。为将对话信号沉淀为画像特征，系统在每次对话后对助教回复中的纠错与提示语句做轻量检测，提取与写作相关的薄弱点概念（如“逻辑连接”“引用规范”“论点展开”），记录到会话中并可同步到后端学习档案。结合 GraphRAG 时，系统会把检索到的课程规范与示例片段作为证据注入对话上下文，要求回答标注引用编号并在证据不足时追问或拒答，从而提升引导式建议的可追溯性与可复核性。

2.6 写作类型感知的智能分析服务

写作分析服务以“写作类型 + rubric + 结构化输出”为核心。服务端首先识别或校验写作类型，并选择对应的评估维度与权重；随后调用上游大模型生成反馈，并将输出解析为维度评分、优点与改进建议等结构化字段，便于前端展示与教师复核。

与通用润色工具不同，本文更关注“可执行建议”：例如指出段落功能缺失、论证链条不完整、证据不足或引用格式问题，并给出可操作的修改方案。该设计使反馈更贴近课程要求，也更便于后续沉淀高质量标注数据。

2.7 GraphRAG 知识库与检索增强生成

为降低大模型在写作辅导中的幻觉风险，系统引入 GraphRAG 检索增强生成流程^[3]。针对研究生《学术规范与论文写作》课程中常见的“引用格式错误”与“学术不端风险”，本文构建了基于向量检索的课程知识库。构建过程如下：首先，将课程讲义、学校学位论文写作规范及优秀的历年范文进行结构化清洗，并按照“章节—段落”的层级进行切分（Chunking），默认切片大小设定为 1200 字符。其次，采用‘text-embedding-v3’模型将切分后的文本片段转化为高维度向量（Embedding），并存储于 FAISS 向量数据库中。当学生在对话中咨询关于引用规范或格式要求的问题时，系统首先将用户查询转化为向量，通过余弦相似度在向量数据库中检索最相关的 K 个规范条款或范文片段。检索到的片段被作为“证据（Evidence）”注入到大模型的上下文提示词（Prompt）中，并强制要求模型仅依据检索到的证据回答，并在回答末尾标注引用来源编号。这一“向量化—检索—注入—生成”的闭环不仅显著降低了模型的幻觉风险，更模拟了真实的学术问题解决过程——即“查阅规范—理解条款—应用执行”，从而在技术实现的底层逻辑上契合了课程的教学目标。

2.7.1 学术规范向量知识库构建过程

针对《学术规范与论文写作》课程中“建议容易泛化、缺少课程依据”的问题，系统将课程规范资料与论文写作指南按统一流程构建为向量知识库。首先，离线 ingestion 支持‘.md/.markdown/.pdf/.txt’多源文本输入；随后按章节与段落进行切分，并以‘-chunk-chars=1200’作为默认分块上限，将原始文本转换为可检索的知识片段。接着，系统对片段执行 Embedding 向量化（‘api|local|hash|env’，默认模型‘text-embedding-v3’），并将向量写入‘FAISS’向量存储，同时维护图索引中的节点与邻接关系以支持图扩展检索。

在线推理阶段，查询先经过语义检索与关键词/图扩展召回，再将命中的证据片段注入提示词上下文，约束模型在证据范围内生成并标注引用编号。该“原文资料→chunks→embeddings→vector store→检索注入生成”的链路，将回答从“语言模型先验”转为“课程证据驱动”，可显著降低《学术规范与论文写作》辅导中的幻觉建议与领域知识缺失问题，并提升教师复核与过程追溯的可行性。

2.8 工具调用与可验证能力

除写作建议外，教学场景中仍存在需要“可验证计算/查询”的任务，例如对字数、结构要素或格式规则进行检查，或在理工类课程中进行数值计算与仿真。为此，

AI 服务提供工具调用接口，使模型在需要精确结果时可调用外部工具并将结果回注到对话中，再生成解释性回答。工具调用能力本质上为系统提供了“外部可验证执行器”，用于约束模型的自由生成范围，降低“凭空计算/编造规则”的风险。本文在原型中实现了基础工具集合，并预留面向写作场景的扩展空间（如引用格式校验、结构要素检查等）。

2.9 模型后训练与评测管线

为使模型更贴近“端侧低延迟 + 场景化决策”目标，本文在原有训练工具链基础上，于 2026-02-10 完成了端侧课程助手的本机微调实验。该实验采用 ms-swift 作为统一训练与部署框架，围绕课程资源检索、学习追踪、简单问答与复杂推理分流提示四类任务，构建“数据转换-训练-部署-联测”闭环。

2.9.1 数据转换与质量校验

原始样本采用 instruction/input/output 结构。为适配 ms-swift 的对话式监督微调，本文将数据统一转换为 messages 格式 (system/user/assistant)，并在 system 提示中固化端侧行为约束：本地优先、结构化回答、复杂推理返回转云端提示文案。转换后的数据统计如表 2.1 与表 2.2 所示。

表 2.1: 端侧微调数据集划分 (2026-02-10)

数据集	样本数	用途
train	400	LoRA 训练
valid	50	训练中评估
test	50	联测抽样
总计	500	解析成功率 100%

表 2.2: 端侧微调任务类型分布

任务类型	样本数
course_resource	200
learning_tracking	150
simple_qa	100
complex_reasoning	50

2.9.2 本机 LoRA 微调配置

本轮实验固定采用 LoRA (Apple MPS) 而非 QLoRA，以规避本机环境下 bit-sandbytes 的 GPU 能力限制。核心训练参数如表 2.3 所示。

表 2.3: 端侧本机 LoRA 训练配置 (ms-swift)

配置项	取值
训练框架	ms-swift 3.12.4
训练模式	LoRA (Apple MPS, 本机)
基座模型	Qwen3-0.6B-Instruct (HF 本地路径)
model_type	qwen3_nothinking
LoRA 参数	rank=8, alpha=16, dropout=0.05
target_modules	q_proj, v_proj
learning_rate	2×10^{-4}
num_train_epochs	3
per_device_train_batch_size	1
gradient_accumulation_steps	8
max_length	512
logging/save/eval steps	10 / 50 / 50

2.9.3 训练结果与最佳检查点选择

训练日志显示 loss 持续下降，并在第 100 step 达到最优验证指标。关键过程数据如表 2.4 所示。

表 2.4: 端侧 LoRA 训练关键指标 (2026-02-10)

训练节点	train loss	eval loss	eval token_acc
step 50	0.2503	0.2368	0.9529
step 100	0.0334	0.0245	0.9957
step 150	0.0170	-	-

综合 `trainer_state` 与日志记录，最终选择 `checkpoint-100` 作为部署产物 (`best_model_checkpoint`)。训练尾段虽出现本机磁盘临时空间不足告警，但不影响最优 `checkpoint` 的落盘与后续部署验证。

2.9.4 部署与客户端联测

本文将 `checkpoint-100` 通过 `swift deploy` 发布为 OpenAI-compatible 本地服务 (`127.0.0.1:18080`)，并按 “AI Service → Backend → Expo Web” 链路完成联测。系统路由策略固定为 `local_first`，非生产环境云端 `fallback` 关闭。服务健康检查 `GET /health`、`GET /v1/models` 与 `POST /v1/chat/completions` 均返回成功。

上述结果表明，本轮端侧训练/微调链路已实现工程可复现，且在不改动后端公共 API (`/api/v1/ai/chat`) 前提下完成了业务接入验证。ONNX 导出、ANE/NPU 编译优化、INT8 量化后联合评测与端侧功耗 `profiling` 将在后续阶段进行。

表 2.5: 端侧模型客户端联测结果 (12 条固定用例)

用例类型	条数	通过标准	结果
课程资源检索	4	非空回复	4/4
学习追踪	3	非空回复	3/3
简单问答	3	非空回复	3/3
复杂推理提示	2	返回“转发云端”语义	2/2
汇总	12	非空回复成功率 $\geq 90\%$	100% (PASS)

2.10 系统原型实现与企业微信集成

平台前端采用 React + TypeScript 实现 Web 客户端，并提供基于 Expo 的移动
端实现；后端采用 Go/Gin 提供课程、写作与学习事件相关 API，并通过 JWT 与
RBAC 实现权限治理；AI 服务基于 FastAPI，实现对话、写作分析、GraphRAG 与工
具调用等能力，并对接 OpenAI-compatible 上游推理服务。原型部署层面提供 Docker
Compose 配置以便快速启动与验证；对于企业微信等场景，系统预留 OAuth 与组织
对接能力，以支持后续在真实教学流程中落地。

2.11 系统测试与评估

系统测试与评估围绕功能正确性、可追溯性与工程稳定性展开。功能测试验证
写作提交与分析流程、权限校验、学习事件记录与查询等关键链路；可追溯性测试关
注 RAG 模式下的引用一致性与“证据不足时追问/拒答”行为；工程测试关注典型
请求的响应时间与服务稳定性。对于模型效果评估，本文采用“固定回归集 + 案例
分析”的方式进行离线对比，并预留进一步的用户试用与课堂验证方案，用于在真实
课程中评估建议的可采纳性与对学习效果的影响。

第三章 总结与展望

本文围绕研究生专业英文写作教学中“反馈成本高、个体差异大、建议难追溯”等问题，提出并实现了一套以学生为中心的智能教学平台。平台以写作任务为主线，贯通写作提交、智能分析、个性化反馈与教师复核，并将过程性数据沉淀为学习事件与学生画像，从而支持纵向追踪与持续改进。系统在工程上采用分层与服务化设计：后端负责业务与权限治理，AI 服务负责编排对话辅导、引导式学习、写作分析、工具调用与 GraphRAG 检索增强，并通过 OpenAI-compatible 接口对接上游推理服务。为降低幻觉并提升可追溯性，平台将课程规范与范文片段构建为知识库索引，在生成反馈时注入证据片段并要求引用编号。为支撑后续模型定制，本文实现了覆盖云侧与端侧的后训练工具链，并在 2026-02-10 完成端侧课程助手的本机 LoRA 微调、OpenAI-compatible 本地部署、网关链路联调与 Expo Web 客户端联测，验证了训练/微调链路和业务接入链路的可复现性。

后续工作可从以下方面继续推进：一是完善知识抽取与融合策略，引入实体关系抽取与图数据库，提高知识图谱的细粒度与可维护性；二是构建更系统的课程问答评测集，量化准确率、引用一致性与教学效果；三是优化检索与上下文压缩策略，降低响应时延；四是完善企业微信 OAuth 与消息推送能力，提升真实教学场景下的使用体验。

结合写作课程的试点需求，本文进一步的工作重点包括：**(1) 数据层面**持续收集真实写作样本与教师反馈，并引入教师模型生成与筛选的数据蒸馏策略，建立 100k 规模训练集与固定回归集；**(2) 模型层面**在 Qwen3 8B 上开展分阶段 LoRA/QLoRA 微调 (tool/rag/style) 并进行多任务合并训练；**(3) 评价层面**完善 rubric 一致性、引用可追溯性与可采纳性等指标，形成更贴近教学目标的评估体系；**(4) 工程层面**扩展写作场景的可验证工具（如引用格式与结构要素检查），并通过课堂试用与教师复核闭环持续改进系统质量。

附录 A 附录

在此放置补充材料，例如证明、额外实验结果或关键代码片段。

参考文献

- [1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention Is All You Need[C/OL]//Advances in Neural Information Processing Systems: vol. 30. 2017. <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>.
- [2] OpenAI. ChatGPT: Optimizing Language Models for Dialogue[EB/OL]. 2022. <https://openai.com/blog/chatgpt>.
- [3] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks[C/OL]//Advances in Neural Information Processing Systems: vol. 33. 2020. <https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf>.
- [4] HOGAN A, BLOMQVIST E, COCHEZ M, et al. Knowledge Graphs[M/OL]. Morgan & Claypool Publishers, 2020. <https://www.morganclaypool.com/doi/abs/10.2200/S01045ED1V01Y202009DSK016>.
- [5] FLOWER L, HAYES J R. A Cognitive Process Theory of Writing[J]. College Composition and Communication, 1981, 32(4):365-387.
- [6] SWALES J M, FEAK C B. Academic Writing for Graduate Students: Essential Tasks and Skills[M]. 3rd ed. Ann Arbor: University of Michigan Press, 2012.
- [7] HYLAND K. Teaching and Researching Writing[M]. 3rd ed. London: Routledge, 2015.
- [8] SHERMIS M D, BURSTEIN J. Handbook of Automated Essay Evaluation: Current Applications and New Directions[M]. New York: Routledge, 2013.