

作品编码: 509403

方圆知音：基于多智能体协同与音系知识图谱的 方言自适应学习系统

技术方案

学校名称: 常州工学院

揭榜榜题名称: CQ-02 基于深度推理大模型的自适应学习路径规划研究

榜题发榜单位: 科大讯飞股份有限公司

成员姓名: 王凌宇乐, 曹磊, 许子祺, 范晓雨, 费扬, 刘岩
杨熙承, 王雨欣, 陈俊豪, 夏思彤

指导老师: 蒋巍, 胡智喜, 孟祥莲

摘要

针对中文方言语音处理中普遍存在的AI幻觉、数据稀缺及个性化学习体验缺失等挑战，本文提出并设计了一个面向多方言智能学习的“声学多模态认知校准框架”。该框架以构建的多模态方言音系知识图谱为认知锚点，通过融合声学、音系与语义特征并进行跨通道一致性评估，对声学模型的输出进行实时验证与校准，从而在有效抑制幻觉的同时，显著提升方言识别的准确率与鲁棒性。

在此核心框架之上，本项目进一步研发并集成了三个核心创新算法：“文脉智能体协作框架(Dialectus)”基于多模态方言音系知识图谱，运用多智能体与思维链推理，实现了对复杂方言知识的深度问答；“自适应多步检索生成模型”通过动态建模学习者状态，动态查询重构，实现了高效的个性化教学；“音系自进化生成与校准模型”则构建了数据-反馈-优化的自进化闭环，以低成本方式完成了模型校准与自我迭代。

本项目将所有模块整合为一个功能完备的方言智能学习平台，旨在为方言的保护、传承与学习提供一个更准确、可靠、高效且智能化的解决方案。

关键词：方言音系知识图谱；多智能体；思维链；自适应学习；自进化

Abstract

To address the prevalent challenges of AI hallucination, data scarcity, and the lack of personalized learning experiences in Chinese dialect speech processing, this paper proposes and designs a "Phono-Cognitive Alignment Framework" for intelligent multi-dialect learning. Grounded in a constructed multimodal phonological knowledge graph as a cognitive anchor, the framework performs real-time verification and calibration of the acoustic model's output by fusing acoustic, phonological, and semantic features and evaluating their cross-channel consistency. This approach not only effectively suppresses hallucinations but also significantly improves the accuracy and robustness of dialect recognition.

Building upon this core framework, this research further develops and integrates three core innovative algorithms: the "Dialectus" collaborative agent framework, which leverages the multimodal phonological knowledge graph, utilizes multi-agent systems and Chain-of-Thought reasoning to achieve in-depth question answering on complex dialectal knowledge; the "Adaptive Multi-step RAG model," which achieves efficient, personalized instruction by dynamically modeling the learner's state and performing dynamic query reconstruction; and the "Phonetic Self-Evolving Generation and Calibration Model," which establishes a self-evolving loop of data-feedback-optimization to complete model calibration and self-iteration in a low-cost manner.

This research integrates all modules into a fully functional intelligent dialect learning platform, aiming to provide a more accurate, reliable, efficient, and intelligent solution for the preservation, inheritance, and learning of dialects.

Keywords: dialect phonology knowledge graph; multi-agent; thought chain; adaptive learning; self-evolution

目 录

摘要.....	I
Abstract.....	II
目录.....	III
插图清单.....	VI
附表清单.....	VII
第1章 绪论	1
1.1 研究背景	1
1.1.1 政策支持	1
1.1.2 社会背景	2
1.1.3 市场分析	3
1.2 国内研究现状	5
1.3 国外研究现状	6
1.4 研究目的	7
1.4.1 构建高可靠性的认知诊断与校准引擎	7
1.4.2 设计支持复杂教学决策的智能体协同框架	7
1.4.3 实现“因材施教”的动态学习路径规划算法	7
1.4.4 建立可持续、低成本的模型自进化生态系统	8
1.5 研究内容	8
第2章 功能设计	10
2.1 需求分析	10
2.1.1 功能性需求分析	10
2.1.2 非功能性需求分析	10
2.2 平台功能设计	11
2.2.1 用户管理与学习目标设定模块	11
2.2.2 可视化知识节点学习模块	11
2.2.3 交互式知识图谱探索模块	11
2.2.4 核心发音练习与智能诊断模块	12
2.2.5 自适应评估与学习路径规划模块	12

第3章 多模态方言音系知识图谱的构建	13
3.1 知识图谱顶层模式（Schema）设计.....	13
3.1.1 核心实体与本体论基础.....	13
3.1.2 音系与超音段特征的扩展建模.....	13
3.2 多源异构数据采集与对齐	15
3.2.1 声学-文本强制对齐.....	15
3.2.2 多模态特征提取与结构化.....	15
3.3 知识抽取、融合与存储实现	16
3.3.1 自动化知识抽取.....	16
3.3.2 知识融合与向量化存储.....	17
3.4 知识图谱的评估与应用	18
3.4.1 内在质量评估.....	18
3.4.2 在核心算法中的应用.....	19
第4章 文脉智能体框架	21
4.1 基于角色的多模态分治检索与链式通信	22
4.1.1 任务解构与智能体分治.....	22
4.1.2 智能体决策的奖励函数.....	22
4.1.3 包含多模态资产的链式通信.....	23
4.2 多模态推理聚合与富媒体响应生成	23
4.2.1 管理智能体的核心作用.....	23
4.2.2 多模态思维链推理与富媒体响应.....	24
4.2.3 框架效率与可扩展性分析.....	24
第5章 自适应多步检索生成模型	26
5.1 自适应多步检索生成模型	26
5.1.1 学习者状态建模.....	27
5.1.2 训练数据增强：生成自适应检索链.....	27
5.1.3 多任务联合训练与复合损失函数.....	27
5.1.4 推理时决策策略：最大化教学奖励与自适应教学路径生成.....	29
5.1.5 多模态整合与闭环教学	30
第6章 音系自进化生成与校准模型	33
6.1 技术渊源与框架概览	33
6.2 音系知识驱动的跨模态检索与生成	33
6.2.1 跨模态检索机制.....	34

6.2.2 个性化辅导生成.....	36
6.3 自进化式数据增强与微调	36
6.3.1 自进化核心流程：CoRAG 与 RAFT 思想的融合	36
6.3.2 高质量训练样本的生成与过滤.....	37
6.3.3 监督微调：RAFT 哲学的实践与 LoRA 实现	37
第 7 章 实验	39
7.1 实验设置	39
7.1.1 数据集.....	39
7.1.2 基线模型.....	40
7.1.3 评估指标.....	40
7.1.4 实现细节.....	41
7.2 实验结果与分析	41
7.2.1 主要模型性能对比.....	41
7.2.2 消融实验.....	42
7.2.3 案例分析.....	43
第 8 章 总结与展望	45
8.1 研究工作总结	45
8.2 主要创新点与优势	45
8.3 研究局限性	47
8.4 未来工作展望	47

插图清单

图 1.1 语保工程	1
图 3.1 苏州话“八”的示意图	14
图 3.2 知识图谱构建流程图	16
图 3.3 一种用于方言音系知识图谱的 NER 模型	17
图 3.4 多模态知识图谱概览	18
图 4.1 文脉智能体框架示意图	21
图 5.1 自适应多步检索生成概览图	26
图 5.2 多模态整合与闭环教学流程图	30
图 6.1 检索策略架构图	34
图 6.2 跨编码器与双编码器架构对比图	36
图 6.3 RAFT 概览	37
图 6.4 RAFT 微调	38
图 7.1 实验设置	39
图 7.2 消融实验	43

附表清单

表 1.1 代表性方言学习应用功能矩阵	4
表 4.1 不同方法的时间复杂度对比	25
表 7.1 在 Dialect-QA 数据集上的模型性能对比	41
表 7.2 在 Pronunciation-Diag 数据集上的模型性能对比	42
表 7.3 核心模型关键组件的消融实验结果	42

第1章 绪论

1.1 研究背景

1.1.1 政策支持

语言多样性是不可再生的宝贵文化财富，必须予以科学保护。这一理念的旗舰项目是于2015年由教育部、国家语委联合启动的“中国语言资源保护工程”（简称“语保工程”）。“语保工程”一期建设成果斐然，通过对全国1712个调查点的系统采集，建成了世界上规模最大的语言资源库，涵盖了123个语种及全部汉语方言。这些被誉为“无价之宝”的语料为后续研究与应用奠定了基础。更为关键的是，2021年启动的“语保工程”二期建设，其目标发生了战略性转变。政策明确指出，二期工程将在继续保护的同时，重点“促进语言资源的开发利用”。这一转变意义重大，它标志着国家语言保护的思路从静态的、档案式的“抢救记录”转向动态的、应用驱动的“活化传承”。

The screenshot shows the official website of the Chinese Language Resource Protection Project. At the top, there is the logo of the People's Republic of China and the text "中华人民共和国中央人民政府" (The Central People's Government of the People's Republic of China) along with the website address "www.gov.cn". A navigation bar includes links for "首页" (Home), "简" (Simplified Chinese), "繁" (Traditional Chinese), "EN" (English), "登录" (Login), "邮箱" (Email), and "无障碍" (Accessibility). Below the header, the main title "我国建成世界最大规模语言资源库" (China has built the world's largest language resource庫) is displayed. The page content discusses the project's achievements and its strategic shift in the second phase, featuring quotes from experts like Cao Zhizhen and田学军. It also mentions the project's implementation by the Ministry of Education and the National Language Commission, supported by the Ministry of Finance.

图 1.1 语保工程

《国家通用语言文字法》与《国务院办公厅关于全面加强新时代语言文字工作的意见》等顶层设计文件均强调要“科学保护各民族语言文字”。同时，苏州、厦门等地方政府也相继出台地方法规，鼓励开展方言教育和推广活动，形成了中央与地方联动的保护格局。

“十四五”时期的语言文字工作规划，更是将“信息化”提升到了前所未有的

战略高度。相关政策提出，要加快国家语言资源数据库的建设与开放共享；制定适用于语音识别、机器翻译等前沿技术的少数民族语言文字规范标准；探索利用自然语言处理和大语言模型等新技术；并推动基于中文数据的新型语言产业和业态发展。

方言保护并非是将其封存于博物馆，而是通过技术赋能，将其转化为可供大众学习、使用和体验的数字化、智能化产品。因此，本项目提出的智能化方言学习平台，不仅是一个学术探索，更是对国家在“保护-应用”悖论下所指明的技术活化路径的直接响应与实践。

1.1.2 社会背景

在国家政策提供宏观指引的同时，方言在社会现实层面所面临的危机，则为本项目提供了更为迫切的文化与社会动因。大量证据表明，中文方言的活力正在迅速衰退，尤其是在年轻一代中的传承已然出现严重断裂。

中国拥有极其丰富的语言资源，官方划定的十大汉语方言区之下，还包含着数以千计的地方土语。然而，这一壮观的语言图景正在迅速褪色。随着城镇化进程的加速和普通话的全面普及，大量的汉语方言和少数民族语言正处于不同程度的濒危状态。

语言濒危的迹象体现在多个维度：首先是使用者数量的减少；其次是使用范围的收缩，方言正从学校、商场等公共领域全面退却，逐渐萎缩为仅限于家庭内部的交流语言，甚至在家庭中也面临普通话的竞争；最后是使用频率的下降。更为深层的是语言结构本身的萎缩。在与强势语言普通话的接触中，方言中独特的音系、词汇和语法规则不断被磨损和替换，逐渐失去其语言个性。

在所有濒危迹象中，最为致命的是代际传承的失败。语言的生命力取决于它是否能被下一代习得和使用。然而，当前的方言传承正面临着一道深刻的“代际断裂带”。学者们将新世纪以后出生的青少年儿童中方言能力的急剧下降形容为“雪崩型”衰退，这是语言濒危最严重的形式。

山西省古交市的一项实证调查为这一现象提供了触目惊心的微观数据。该调查针对 128 名“00 后”中小学生，结果显示，在学校、超市、集市等公共场所，学生使用古交话的比例分别仅为 1.56%、2.34% 和 5.47%，而普通话的使用率则普遍超过 95%。即便是被认为是方言最后堡垒的家庭环境，古交话的使用比例也仅为 31.37%。这些数据无可辩驳地证明，青少年一代已经将普通话作为其最主要的交际工具，而方言的自然习得环境——家庭、社区和同辈群体——已基本瓦解。当一种语言不再被儿童学习时，它的消亡便只是时间问题。

方言的衰落不仅是语言学层面的损失，更意味着深层文化遗产的流失。每

种方言都是一部活的历史，承载着一个地区独特的思维方式、风俗习惯和文化认同。当乡音逝去，与之相连的文化记忆和身份认同也将随之变得模糊。

许多宝贵的非物质文化遗产与方言紧密相连，无法被轻易“翻译”。例如，瑶族的民族史诗《盘王大歌》必须用瑶语演唱才能传承其完整的文化内涵。同样，各地的戏曲、评书、民谣等口头传统艺术，其韵味和精髓都根植于特定的方言音韵和表达方式之中。方言的消亡，将导致这些文化瑰宝失去其赖以生存的土壤，最终沦为静态的文字记录。

综合来看，方言危机的核心问题在于“代际断裂”。这一断裂使得传统的、自然的语言传承机制彻底失效。因此，仅仅像“语保工程”一期那样对年长发音人的语音进行存档是远远不够的。这种方式虽然保存了语言的样本，却无法修复断裂的传承链条。这就为本项目的定位提供了关键的社会价值：面对自然传承机制的失效，本项目迫切需要构建一个替代性的、人工的传承机制。一个智能、互动、个性化的学习平台，正可以扮演这一“数字桥梁”的角色，它能够为年轻一代提供他们在现实生活中已无法获得的结构化语言输入、互动练习和文化语境，从而跨越代际鸿沟，为方言的活态传承创造新的可能性。

1.1.3 市场分析

通过对应用市场的调研，可以发现针对粤语、闽南语、吴语等主流方言的学习App已具备一定规模。尽管种类繁多，但其核心功能和设计理念表现出高度的同质化。

这些应用普遍提供的功能可归纳为以下几类：

- **基础翻译与词典：**提供普通话与方言之间的文本或语音翻译，通常是简单的词句对译，并附带真人录音
- **词汇与短语手册：**以列表形式呈现常用词汇和情景对话，类似于数字化的口袋书或短语手册，配有标准发音。
- **音系教学：**提供声母、韵母、声调等发音要素的图表和跟读练习。
- **内容资源库：**整合方言歌曲、有声故事、广播节目等媒体资源，供用户进行沉浸式听力输入。
- **游戏化学习：**采用积分、升级、闯关等游戏机制，以提升用户的使用黏性和学习动。
- **离线功能：**多数应用支持将音频和课程资料下载至本地，方便用户在无网络环境下使用。

尽管上述功能在一定程度上满足了初学者的入门需求，但深入分析后可以发现，这些应用在核心技术和教学法上存在着显著的、普遍性的缺陷。它们本质上

是静态内容的数字化存储库，而非真正意义上的智能教学系统。

现有应用在语音识别与反馈能力上仍显薄弱。应用大多无法对学习者的发音进行准确、实时的评估，更缺乏针对音系层面的个性化纠正指导。由此，口语训练往往沦为低效的、无反馈的模仿过程。

现有应用的个性化学习机制也几乎缺席。大部分课程内容和学习路径都是预设的，无法根据学习者的起点水平、进步速度、学习兴趣或薄弱环节进行动态调整，仍沿用“一刀切”的教学模式，与现代智能辅导系统倡导的因材施教理念相去甚远。

现有应用在生成式 AI 的应用方面存在明显空白。由于缺乏生成式交互，这些应用既未触及低资源场景下 AI 常见的“幻觉”问题，也未探索相应的可靠性优化策略，交互模式停留在“查询—匹配”的静态阶段，难以形成丰富、多变的语言交流体验。

现有应用对语境和知识层面的支持仍不充分。学习者无法就方言的语法差异、词汇细微含义或文化背景进行深入提问，应用提供的解答局限于预设条目，缺乏灵活性与深度。

为了更直观地展示这一技术鸿沟，下表对几款代表性方言学习应用的功能进行了矩阵分析。

表 1.1 代表性方言学习应用功能矩阵

功能类别	特性	粤语通	Nemo 粤语	Drops	说咱闽南话
基础内容与功能	词汇/短语列表	✓	✓	✓	✓
	翻译功能	✓	✓		
	真人预录音频	✓	✓	✓	✓
	音系发音练习	✓	✓		
	媒体库(歌曲/故事)	✓			✓
教学法	游戏化设计			✓	
	静态学习路径	✓	✓	✓	✓
高级 AI 能力	实时 ASR 发音反馈				
	AI 幻觉检测与校准				
	动态个性化学习				
	基于知识图谱的问答				
	数据闭环自进化				

从表1.1可以清晰地看出，当前市场上的所有产品都局限于基础功能和静态教学法，而在高级AI能力方面则完全是空白。这种现状揭示了一个核心问题：现有方言学习工具陷入了“数字短语手册”的范式陷阱。它们只是将传统的、被动的学习材料（如词典、磁带）进行了数字化移植，而未能利用现代AI技术创造出全新的、高效的、交互式的教学体验。

这一发现为本项目提供了强有力的技术立足点。它表明，在方言保护的迫切需求（第一、二章所述）与市场现有技术能力之间，存在着一条巨大的鸿沟。市场所需要的，不仅仅是一个功能更丰富的App，而是一个建立在认知与教学AI基础之上的、全新范式的智能学习平台。这恰恰是本项目旨在填补的空白。

1.2 国内研究现状

国内在方言语音处理领域的研究呈现出两大特点：一是产业界以超大规模统一模型为主要技术路径，成果显著但技术细节常作为核心资产而较少公开；二是学术界和应用层高度关注方言的社会文化价值，强调技术的普惠性和传承性。

以阿里巴巴、腾讯为代表的科技公司，在方言识别上取得了令人瞩目的成就。它们的技术核心是构建能够覆盖海量方言的超大规模预训练模型。例如，阿里巴巴达摩院的Paraformer模型系列，通过非自回归架构大幅提升了识别速度，其内部的“Glancing Language Model”采样器能够在解码时融入全局语义信息，这在功能上类似于一种隐式的知识增强，用于解决声学上的模糊性。中国电信的“星辰”大模型支持数十乃至上百种方言和语言，其背后是“统一建模”和“蒸馏+扩充”等先进技术，旨在用一个模型底座处理多样化的声学和语言特征。这种工业化路径，在处理数据相对丰富的方言上表现优异，但其对极端低资源或音系独特的方言的适应性，仍有待观察。

国内学术界在方言识别的精细化建模上持续探索。例如，有学术研究探讨使用包含卷积神经网络(CNN)、残差网络和多头自注意力机制的端到端模型进行方言识别，其核心思想是让模型自动学习和提取方言间潜在的“音系学特征差异”。这表明，国内学界同样认识到，深入理解方言音系规则是提升识别准确率的关键。

检索增强生成(RAG)与多智能体等前沿技术在国内自然语言处理(NLP)领域已相当成熟，正处于向语音领域渗透的前夜。虽然直接将RAG或多智能体应用于方言语音识别的公开研究尚不多见，但在中文问答、对话系统等领域，相关的研究和基准测试（如ConvRAG, CDQA）层出不穷。这表明国内已经具备了将这些先进技术栈应用于解决方言处理中知识依赖性强、交互逻辑复杂等问题的技术储备和学术环境。

1.3 国外研究现状

国际上，针对语音识别模型局限性的研究已经从单纯扩大数据和模型规模，转向探索更深层次的知识融合与复杂推理，呈现出从“感知”到“认知”的明显趋势。

在知识增强型语音识别领域，研究者早已不满足于传统的语言模型后处理或格网重打分（Lattice Rescoring）。最新的研究致力于实现声学与知识的深度、结构化对齐。例如，通过图匹配与最优传输理论（Graph Matching based on Optimal Transport），研究者探索如何对齐声学表征图与语言知识图，确保对齐过程不仅考虑单个节点（如音素或词汇），更考虑它们之间的结构关系，这与您的“跨通道一致性评估”理念高度契合。这种结构化对齐被认为是抑制模型在声学信息模糊时产生无端猜测（即幻觉）的关键。

RAG 技术在语音处理领域已迅速演进，并呈现出从级联到端到端的清晰路径。早期的应用（可视为级联模式）是将 ASR 系统与 RAG 模块串联，利用 RAG 修正 ASR 的输出错误或补充领域外知识。例如，有模型通过检索相似的错误模式知识库来校正低资源语言的 ASR 结果；也有模型在推理时动态检索领域相关的文本，以解决特定领域的词汇识别难题。然而，最新的突破在于 ASR-Free 的端到端语音 RAG 模型，如 SpeechRAG 和 VoxRAG。这类模型不再需要一个独立的 ASR 系统，它们构建了统一的声学-文本嵌入空间，能够根据文本查询直接检索相关的原始音频片段，从而完全避免了 ASR 错误传播的问题，并更好地保留了语音中的情感、强调等副语言信息。这代表了多模态信息融合的终极方向之一。

更高级的推理架构（如思维链和多智能体）正被引入语音领域，以解决复杂任务。研究表明，在大型音频语言模型中引入思维链（Chain-of-Thought, CoT）可以提升其在多步推理任务上的表现，但同时也发现，过于复杂的推理链反而可能干扰模型的判断，这为如何设计有效的认知校准机制提供了宝贵的经验。而在多智能体系统（Multi-Agent Systems, MAS）方面，具体应用已开始出现。例如，已有框架（如 SpeechQC-Agent）利用一个中心“规划”智能体来协调多个“专家”智能体，以高效地完成语音数据集的质量校验任务。这种分工协作的模式，与您设想的“文脉智能体协作框架 (Dialectus)”中，由不同智能体负责检索、推理和聚合的思路不谋而合。

在自适应学习与发音训练（CAPT 领域，研究的焦点正从技术实现转向教学法与技术的结合。现有研究普遍指出，尽管当前系统在音段（声母、韵母）的重复性练习上表现优异，但在超音段（声调、节奏）和真实交际能力的培养上仍有欠缺。技术上，基于端到端混合 CTC/Attention 架构的自动发音错误检测（APED）系统，

因其无需强制对齐、部署更灵活的优势，正逐渐取代传统方法，为开发更高效、更适应方言等新场景的个性化学习工具铺平了道路。

1.4 研究目的

在人工智能与教育科技深度融合的背景下，自适应学习路径规划已成为提升个性化教育质量的核心议题。然而，当前技术在学习者认知状态的精准诊断、深度推理的可靠性以及复杂教学决策的处理上仍存在显著瓶颈。为探索这些前沿挑战的解决方案，本项目的总体目标是：构建一个由深度推理大模型与动态声学知识图谱双轮驱动，并由多智能体协同决策的、面向复杂语言学习（以多方言为典型场景）的自适应学习路径规划平台。本项目旨在通过理论与技术的系统性创新，为新一代智能教育系统提供一个可落地的、高可靠性的AI基座。

1.4.1 构建高可靠性的认知诊断与校准引擎

为解决深度推理模型在精确知识任务与 STEM 学科中易产生“幻觉”的固有缺陷，本项目将首先致力于构建一个高可靠性的认知诊断与校准引擎。其核心挑战在于突破模型的可靠性瓶颈。本项目将通过实现“声学多模态认知校准框架”来达成此目标，其原理是，将结构化的“方言音系知识图谱”作为外部的、不可辩驳的“事实锚点”，通过持续评估声学模型输出与图谱规则之间的一致性，来主动、实时地发现并修正模型的潜在错误。此举旨在从根本上提升认知诊断（如发音错误诊断）的准确性与可信度，为整个自适应学习系统提供一个坚实的基础。

1.4.2 设计支持复杂教学决策的智能体协同框架

在此基础上，为突破单一智能体在处理多维度信息时的决策瓶颈，本项目将进一步设计一套能够进行复杂教学决策的智能体协同框架。鉴于单一架构的局限性，本项目将通过研发“文脉智能体协作框架 (Dialectus)”来探索解决方案。其原理在于将复杂的教学决策任务进行分解，由一组具备不同专长的异构智能体（如诊断、检索、规划等）进行分工协作与信息聚合。通过这种模式，该系统旨在突破单一架构的瓶颈，更全面、更精准地理解学习者需求，从而生成匹配度与有效性都更高的个性化学习路径。

1.4.3 实现“因材施教”的动态学习路径规划算法

面向最终的个性化教学目标，本项目的核心任务之一是实现一种真正“因材施教”的动态学习路径规划算法。为超越传统系统对学习者状态的浅层建模，本

项目将通过“自适应多步检索生成模型”来实现这一目标。该算法的核心在于构建一个双重动态建模机制：它不仅追踪学习者的宏观进度，更深入地分析其在具体知识点上的认知负荷与理解偏好。基于此精细化模型，系统能够进行动态的、自适应的查询重构与内容生成，为学习者匹配难度和呈现模态都最合适的学习资源，从而实现学习效率与效果的最大化。

1.4.4 建立可持续、低成本的模型自进化生态系统

最后，为确保系统的长期生命力与技术领先性，本项目将着力构建一个可持续、低成本的模型自进化生态系统，以应对教育领域的长期数据挑战。特别是在方言等小众教学方向，高质量标注数据的稀缺是限制AI技术发展的根本性难题。为解决此问题，本项目将通过“音系自进化生成与校准模型”实现一个巧妙的闭环：它将用户的日常学习与练习过程，转化为高质量标注数据的生成过程。这些新数据经过智能过滤后，将被用于对平台内的核心模型进行持续的监督微调。这套自进化机制旨在确保平台能够“越用越聪明”，在没有高昂人工投入的情况下，实现性能的自我迭代与提升，保证其长期的技术活力与应用价值。

1.5 研究内容

本项目旨在设计与实现一个面向多方言学习的、具备认知校准能力的智能化平台。核心研究内容将围绕一条主线展开：以结构化的“方言音系知识”为基石，构建一个能够主动抑制“AI幻觉”的核心技术框架，并在此基础上研发一系列支撑个性化学习与复杂交互的创新算法，最终集成为一个功能完备的学习平台。

研究的起点是构建一个覆盖多种方言的多模态音系知识图谱。这项基础性工作不仅包括对各方言的词汇、发音规则、变调模式等语言学特征进行体系化梳理，还涉及将这些结构化知识与海量的真人发音音频、口型视频等多模态数据进行深度关联，旨在打造一个权威且可计算的“认知锚点”，为后续模型的精准校准提供事实依据。

在此基础之上，本项目的理论核心在于设计并实现一个声学多模态认知校准框架(Phono-Cognitive Alignment Framework)。该框架旨在突破传统声学模型仅依赖数据驱动的局限，通过动态融合声学、音系知识与语义等多源信息，并持续性地评估声学信号分析结果与知识图谱规则之间的一致性，从而对潜在的识别错误进行主动、实时的干预与修正。这项研究将从根本上提升模型在处理复杂、多变方言时的可靠性与可解释性。

为将上述核心框架应用于实际学习场景，本项目将进一步研发三个相互支撑

的创新应用算法：

文脉智能体协作算法 (*Dialectus*)。为处理用户关于方言历史、文化等方面复杂的知识问答，本算法将构建一个多智能体协作系统。通过设计分别负责历史语料检索、音系变迁分析、现代用法查询等不同职能的智能体，并由一个核心推理智能体运用思维链（Chain-of-Thought）机制进行信息的逻辑整合与聚合，从而生成全面、连贯且深入的答案，实现从单一信息检索到复杂知识推理的跨越。

个性化自适应学习算法。为实现“因材施教”的教学目标，本算法将设计一个自适应多步检索生成模型。该模型的核心在于通过动态的“学习者状态模型”来感知用户的理解程度和学习偏好，并以此为依据，双重驱动查询重构：既要确保检索知识的完备性，又要调整解释内容的难易度与呈现模态。这使得平台能够为每位用户提供真正个性化的学习内容与动态路径。

音系知识自进化算法。为从根本上解决方言数据稀缺的难题，本算法将构建一个“数据-学习-反馈-优化”的自进化闭环。通过捕捉用户在发音练习中的交互行为，系统能够自动生成大量包含错误诊断和标准示范的高质量“纠正-示范”数据对。这些数据经过智能过滤后，将用于对平台核心模型进行持续的监督微调，使平台在与用户的互动中实现性能的自我迭代与进化。

通过将知识图谱、校准框架与上层应用算法有机地融为一体，构建一个功能完备、体验流畅的方言智能学习平台，并最终通过客观性能指标与主观用户评估，全面检验本项目工作的有效性和创新价值。

第2章 功能设计

2.1 需求分析

2.1.1 功能性需求分析

功能性需求是根据平台的核心业务流程和用户交互场景得出的，描述了系统必须为用户提供的具体服务和功能。首先，系统必须提供完整的用户账户管理功能，支持学习者注册、登录并创建包含目标方言选择的个人档案。在内容呈现上，平台需要以可视化的知识节点网络作为主学习界面，引导用户进行结构化学习，同时还应提供一个可自由探索的交互式知识图谱，以满足用户对知识进行溯源和关联发现的需求。当用户点击任一学习单元时，系统必须能展示其详细内容，并提供标准发音的音频播放功能。

平台的核心功能需求围绕着一个“学-练-评-荐”的智能闭环展开。系统需支持用户录制发音，并能对其进行实时的声学分析。关键需求在于，系统必须基于知识图谱进行多维度的智能诊断，不仅判断对错，还要提供具体、可解释的改进建议。练习完成后，系统必须能自动生成专属的“错题本”，并基于对错题记录和用户历史表现的综合分析，动态地为用户规划和推荐下一步的学习路径，从而实现真正的个性化指导。

2.1.2 非功能性需求分析

非功能性需求是衡量平台质量与用户体验的关键，涉及性能、可靠性、易用性等多个方面。在性能与准确性方面，平台必须满足严苛的要求。发音诊断与反馈的响应时间需控制在数秒以内，以保障交互的流畅性，同时后台需具备支持多人并发学习的能力。更重要的是，作为平台的核心价值，发音诊断的准确率必须达到高可信水平，这是赢得学习者信任的基石。此外，系统整体应保持高可用性，确保服务的稳定可靠，并保证用户数据在不同终端上的一致性。

在可用性层面，平台界面设计需力求简洁直观，确保不同年龄段的用户都能轻松上手；智能体给出的诊断建议也必须通俗易懂，避免使用过多的专业术语，使其具有真正的指导价值。从长远来看，系统的可扩展性与可维护性至关重要。平台架构必须支持未来便捷地增添新的方言种类，其模块化的设计也应确保各个功能组件可以独立升级和维护，以适应技术的持续发展。

2.2 平台功能设计

本项目的最终目标是落地为一个功能完备、体验流畅、且具备高度智能化的方言学习平台。平台的功能设计遵循以学习者为中心、以交互式练习为核心、以自适应规划为导向的原则，旨在将底层的复杂算法无缝转化为直观、高效的学习体验。整个平台的功能体系主要由以下几个核心模块构成。

2.2.1 用户管理与学习目标设定模块

这是用户与平台交互的入口。首次使用时，用户需进行注册并创建个人学习档案。本模块最核心的功能是多方言选择系统。用户可以从平台支持的方言库中（如粤语、上海话、闽南语等）选择自己希望学习的目标方言。一旦选定，平台的所有后续内容，包括知识呈现、发音练习、路径规划等，都将围绕这一目标方言展开。该模块负责持续记录用户的基本信息、学习方言种类以及整体学习进度。

2.2.2 可视化知识节点学习模块

当用户选定方言进入主学习界面后，平台将呈现一个基于知识图谱的可视化学习网络。在这个网络中，方言的知识体系被拆解成一个个相互关联的“学习节点”。每个节点代表一个具体的学习单元，可能是一个核心声母/韵母、一个独特的声调模式、一组易混淆的词汇，或一个特定的语法现象。用户可以根据引导或自主选择点击不同的节点，进入针对性的练习环节。这种设计将抽象的语言知识具象化，使用户能够清晰地看到自己的学习版图和知识掌握情况。

2.2.3 交互式知识图谱探索模块

为满足学习者对知识进行溯源和关联性探索的需求，并增强平台教学的透明度，本项目将设计一个交互式的知识图谱探索模块。该模块将底层的“方言音系知识图谱”以图形化、可交互的网络形式完整地呈现给用户。学习者不再局限于预设的学习节点，而是可以自由地在图谱中进行漫游、缩放和点击。例如，用户可以点击一个词汇节点，查看其连接的所有相关信息，包括其标准发音音频、所属的音系规则、以及具有相同发音特征的其他词汇等。此外，在“核心发音练习模块”中，当智能体给出诊断建议时，会提供直接跳转到知识图谱中对应节点的链接，方便学习者即时查看该知识点的完整上下文，从而实现从“被动纠错”到“主动探索”的升华。

2.2.4 核心发音练习与智能诊断模块

这是平台最核心的互动学习环节，也是“声学多模态认知校准框架”和“音系知识自进化算法”的主要应用场景。当用户进入一个学习节点后，智能体首先会从“方言音系知识图谱”中呈现相应的练习任务，并提供与图谱关联的真人标准发音作为示范。学习者在收听后，可以录制自己的发音，系统则会实时捕捉语音信号并进行声学分析，例如生成用于对比的语图。随后，作为本模块的核心，智能诊断引擎会将用户的发音特征与知识图谱中的标准模式进行精确比对，并给出多维度的、可解释的诊断报告。这种报告不仅判断对错，更能明确指出问题所在（如声调、声母或韵母），并提供具体的改进建议，从而构成一个完整的“学-练-评”闭环。

2.2.5 自适应评估与学习路径规划模块

该模块是实现个性化教学的关键，也是“个性化自适应学习算法”的具体体现。当用户完成一个节点的练习后，系统会自动将其发音不佳或回答错误的条目，连同智能体的诊断建议，汇集成一本专属的、可视化的“错题本”。这本错题本不仅是供用户随时回顾复习的工具，更是自适应路径规划的核心依据。智能体将综合分析当前节点的错题内容及用户的历史学习数据，动态更新该用户的“学习者状态模型”。基于此模型，系统最终会为用户推荐或规划出下一阶段最需巩固的知识节点，从而实现一条真正动态、智能的学习路径。

第3章 多模态方言音系知识图谱的构建

本项目所有上层算法与框架的基石，是一个内容丰富、结构严谨、且专门面向方言学习的“多模态方言音系知识图谱”。它并非一个简单的词汇库，而是一个深度融合了语言学理论、声学特征与多媒体信息的认知中枢。构建这样一个高质量的知识图谱，是确保整个平台智能化、精准化、可解释性的前提。其构建过程遵循了从顶层理论设计到底层数据融合的系统化方法，涵盖了模式设计、多源数据融合、知识抽取与计算化表达等关键环节。

3.1 知识图谱顶层模式（Schema）设计

一个科学、完备的顶层模式是知识图谱的骨架。在构建初期，本项目首先对需要表达的方言知识进行了系统性的梳理，并设计了一套能够精确描述其复杂特征的 Schema。

3.1.1 核心实体与本体论基础

本项目定义了多个核心实体类别，包括词汇（Lexical Entry）、音节（Syllable）、音系规则（Phonological Rule）、发音人（Speaker）以及多模态资产（Multimodal Asset）。为了确保模型的规范性和互操作性，本项目的 Schema 设计借鉴了万维网联盟（W3C）推荐的词汇本体模型 **OntoLex-Lemon**。该模型提供了一个连接词汇信息与本体概念的标准框架，其核心组件包括：

- Lexical Entry: 词汇条目，代表一个独立的词。
- Form: 词形，一个词汇条目的具体书写或发音形式。
- Sense: 词义，将一个词汇条目链接到一个本体概念。
- Concept: 本体中的概念，代表词义的实际所指。

采用 OntoLex-Lemon 作为基础，使得本项目知识图谱能够与更广泛的语言学数据资源（如 WordNet）进行链接与对齐。

3.1.2 音系与超音段特征的扩展建模

标准 OntoLex-Lemon 模型主要关注词汇和形态，对于音系学的细粒度信息，特别是方言中至关重要的超音段特征（Suprasegmental Features）涉及不足。因此，本项目对其进行了扩展：

1. **音系表征:** 本项目通过 OntoLex 的 phon (Phonology) 模块扩展 Form 实体，为其增加属性，用以存储国际音标 (IPA) 转写、音素序列等音系表征。
2. **超音段特征建模:** 对于声调、语调、重音和节奏等超音段特征，本项目将其模型化为多模态资产的一部分或音节实体的直接属性。例如，一个声调可以被定义为具有基频 F_0 轮廓 (Contour)、时长 (Duration) 等属性的实体。这使得一个“词汇”实体通过 hasPronunciation 关系连接到一个“音节”实体，该音节不仅包含音素信息，还关联着具体的声调模型，如图 3.1 所示。
3. **多模态资产的深度定义:** 其中，“多模态资产”实体不仅包含了常规的音频 (Audio)、口型视频 (Lip-sync Video) 等，更将 **语图 (Spectrogram)** 视为一种核心的、结构化的声学特征资产。此外，本项目进一步将其他关键声学特征也纳入其中：
 - **梅尔频率倒谱系数 (MFCCs):** 一个包含例如 39 个系数的向量，高效地表征了音色的核心特征。
 - **共振峰 (Formants):** F_1, F_2, F_3 等共振峰频率，直接反映了发音时声道的状态，是元音区分的关键。
 - **基频 (F_0):** 反映声带振动频率，是声调和语调的主要物理载体。

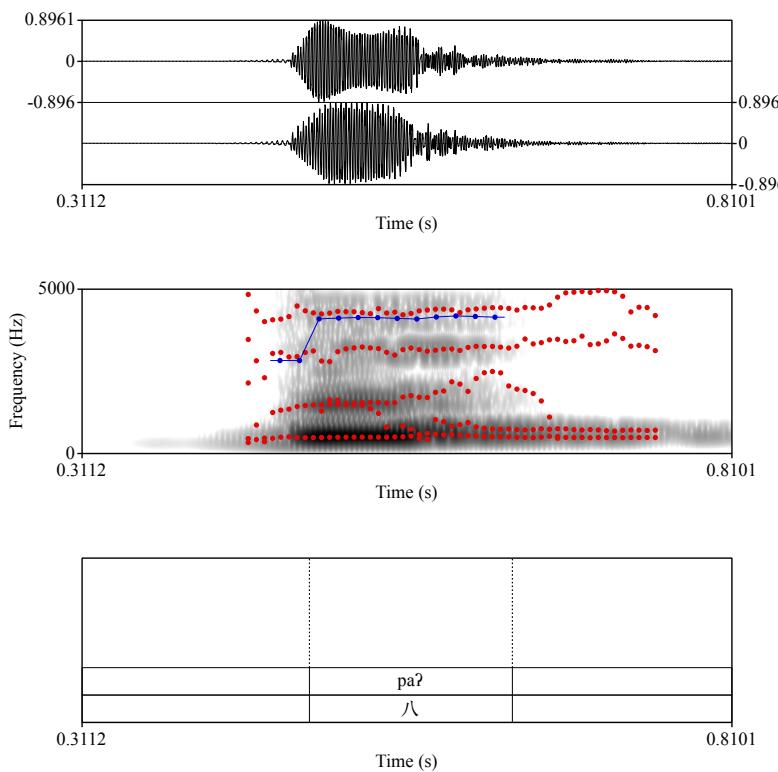


图 3.1 苏州话“八”的示意图

这些实体之间通过丰富的关系类型进行连接，例如，一个“词汇”实体通过 hasPronunciation 关系连接到一个“音节”实体，并通过 hasAudioExample 关系连接到一个“多模态资产”（音频文件）；该音频文件则进一步通过 hasSpectrogram, hasMfccVector, hasF0Contour 等关系连接到其对应的具体特征资产。

3.2 多源异构数据采集与对齐

知识图谱的内容来源于多个渠道的数据。语言学数据的采集主要依赖于权威的方言词典、地方志以及现代语言学研究专著。声学及视觉数据，则通过对目标方言的母语发音人，在专业录音环境下进行系统性的语料录制而获取。

3.2.1 声学-文本强制对齐

数据处理环节的核心难点在于多模态信息的对齐。本项目采用音素级别的 **强制对齐 (Forced Alignment)** 技术，将录制的音频流与文本语料进行精确的时间戳匹配。考虑到方言往往是低资源语言，缺乏现成的声学模型和发音词典，本项目采用了以下策略：

- **主流工具应用：**使用如 **Montreal Forced Aligner (MFA)** 等基于 Kaldi 工具包的先进对齐器。
- **低资源策略：**在目标方言缺乏训练数据时，采用“跨语言对齐”(Cross-language Alignment) 策略，即利用一个声学特征相近且资源丰富语言（如普通话）的预训练模型进行初步对齐，再在少量方言数据上进行微调。

对齐的输出是每个音素、音节和词汇在音频文件中的精确起止时间，这是后续所有特征提取的基础。

3.2.2 多模态特征提取与结构化

在对齐完成后，本项目进一步对所有标准发音的音频和视频片段进行处理，提取出关键的声学与视觉特征并进行结构化存储。

1. **声学“指纹”生成：**为每个音节或词汇的标准发音生成并保存高分辨率的语图。这些语图作为一种可视化的声学“指纹”，被直接链接到对应的词汇或音节节点上，它们不仅是 AI 模型进行精准比对的依据，也可以作为一种直观的教具呈现给学习者。
2. **数值特征提取：**批量化提取前述的 MFCC 向量、共振峰轨迹和 F_0 曲线。这些数值化特征被存储为独立的、可查询的文件（如.csv 或.npy），并在知识图

谱中建立链接。

3. 视觉特征提取: 对口型视频进行处理, 利用面部关键点检测模型定位唇部轮廓, 提取出随时间变化的几何特征序列。

3.3 知识抽取、融合与存储实现

知识图谱的构建是一个系统性工程, 总体上可划分为知识抽取、知识学习(或融合)和知识评估三个阶段, 如图 3.2 所示。

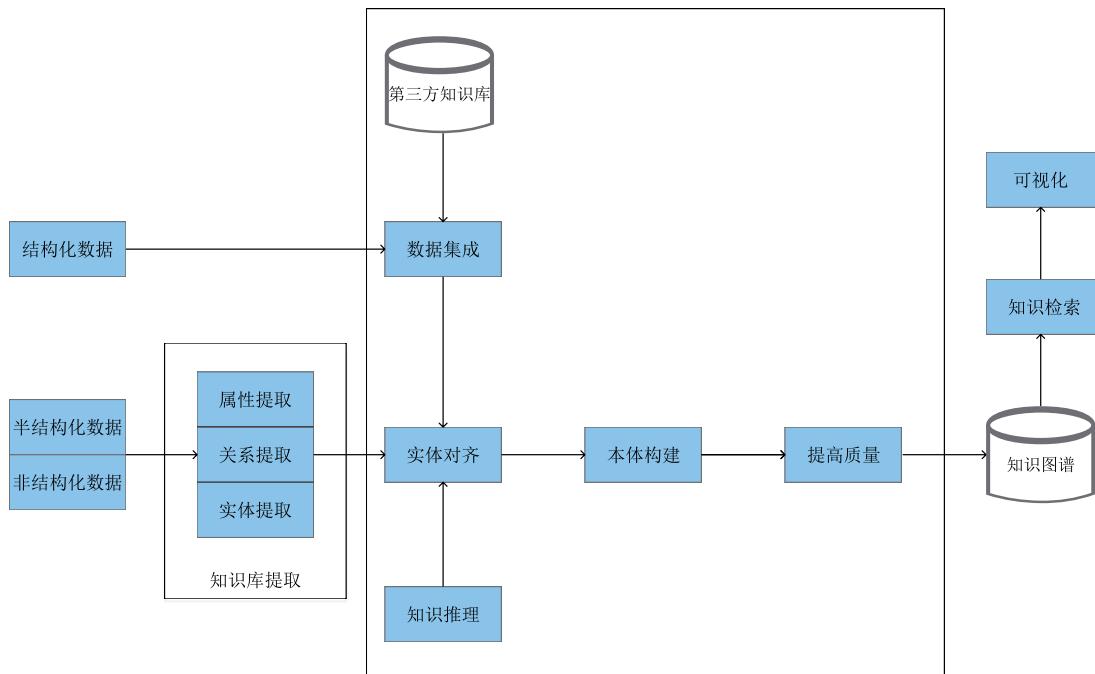


图 3.2 知识图谱构建流程图

3.3.1 自动化知识抽取

在完成数据采集与对齐的基础上, 本项目通过自动化的知识抽取脚本, 将这些多源信息转化为知识图谱中的“主-谓-宾”三元组(Triples)。

- **从结构化/半结构化数据抽取:** 对于方言词典等格式相对固定的数据源, 本项目编写定制化的解析器, 直接抽取(词汇 A, 包含音节, 音节 B)、(词汇 A, 词义为, 概念 C)等三元组。
- **从非结构化文本抽取:** 对于语言学专著等纯文本, 本项目采用自然语言处理(NLP)流水线, 包括命名实体识别(NER)以识别语言学术语, 以及关系抽取(RE)模型来发现实体间的关系。常用的模型如基于 BiLSTM-CRF 的 NER 模型(如图 3.3 所示)和基于卷积神经网络的抽取(RE)模型。
- **实体链接与消歧:** 方言中普遍存在同音词、多音字以及不同文献中转写不一

的问题。本项目引入实体链接 (Entity Linking) 技术, 如利用基于图的算法 (类似 Babelfy 的思想), 通过上下文和图谱的局部稠密度来判断一个词的提及 (mention) 应该链接到哪个唯一的词汇实体, 从而确保知识的准确性。

例如, 从词典中抽取 (词汇: “吃饭”, 包含音节, “sik6”), 从对齐后的音频中标注 (词汇: “吃饭”, 拥有标准发音, 音频文件: “shifan_male_01.wav”), 并基于声学分析结果建立 (音频文件: “shifan_male_01.wav”, 生成语图, 语图文件: “shifan_male_01.png”) 这样的关系。

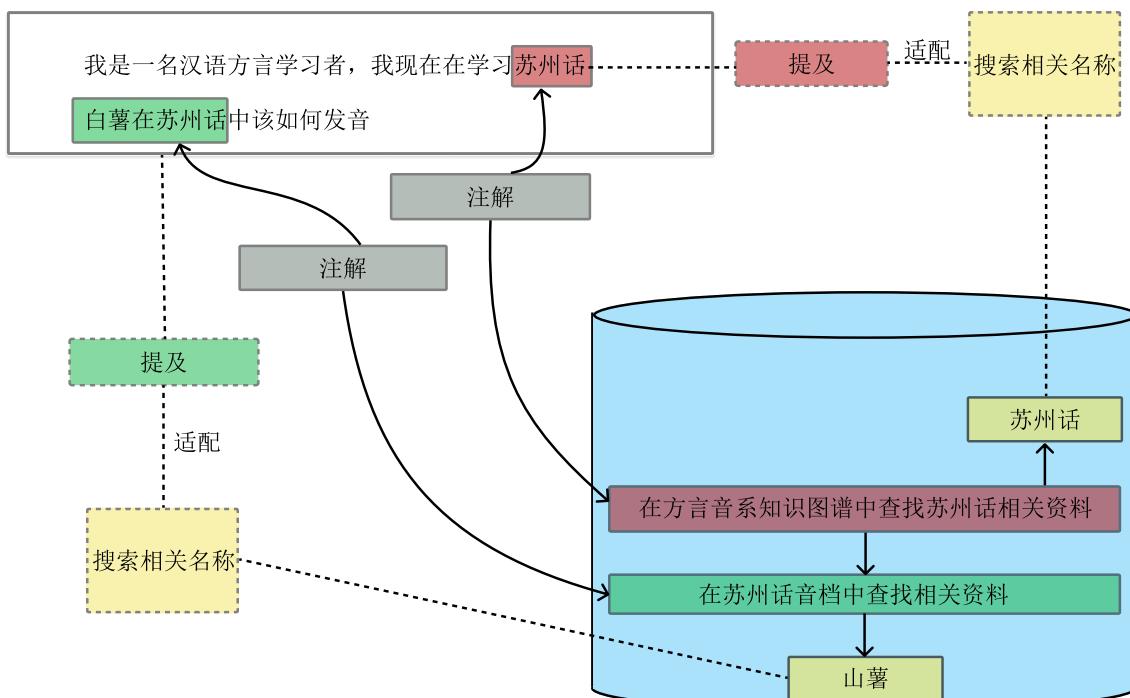


图 3.3 一种用于方言音系知识图谱的 NER 模型

3.3.2 知识融合与向量化存储

所有抽取出的三元组经过清洗和实体对齐后, 被融合进统一的知识库中。为实现高效的查询与管理, 构建完成的知识图谱最终被存储在高性能的图数据库中。

同时, 为了便于下游的机器学习模型直接调用, 本项目利用知识图谱嵌入 (Knowledge Graph Embedding) 算法, 将整个图谱中的实体和关系映射到低维、稠密的向量空间中。本项目考察了从经典的翻译模型 (如 TransE、RotatE) 到能够处理多模态信息的更先进模型的演进:

- **传统模型:** 将知识图谱的结构信息编码为向量。
- **多模态嵌入模型:** 如 IKRL 或 TransAE, 它们通过引入额外的编码器 (如自动编码器 Autoencoder) 来将视觉或其他模态的信息融入实体的向量表示中。

最新的研究（如 **MMKG-T5**）则利用强大的预训练 Transformer 模型，实现了对结构、文本、视觉等多模态信息更深层次的融合，其流程如图 3.4 所示。这些向量化后的知识，可以直接服务于认知校准框架中的“音系特征编码”环节，以及自适应学习算法中的内容推荐模块，真正实现了从结构化知识到 AI 模型可利用特征的转化。

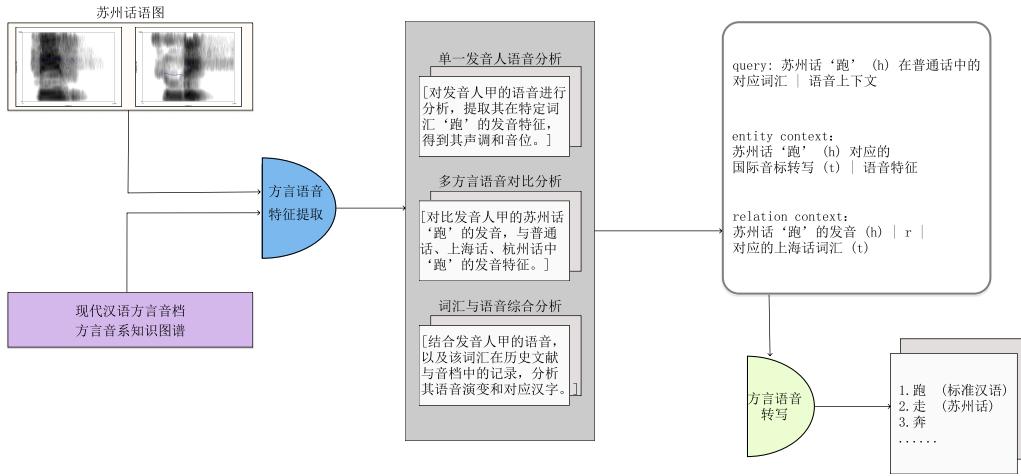


图 3.4 多模态知识图谱概览

3.4 知识图谱的评估与应用

想要构建完成的“多模态方言音系知识图谱”不仅需要在理论设计上科学合理，更需在实践中证明其自身的结构质量与数据准确性。同时，它作为本项目所有上层算法的基石，其核心价值最终体现在对这些算法的赋能上。因此，本节将首先对其进行严格的内在质量评估，然后阐述其在后续核心算法中的具体应用方式。

3.4.1 内在质量评估

为保证知识图谱作为一个高质量静态数据制品的可靠性，本项目仅采用内在评估（Intrinsic Evaluation）的方式，从结构和知识表示两个层面，对图谱自身的质量进行检验。

1 结构完整性与一致性评估

在将数据正式载入图数据库之前，本项目首先对图谱的结构和模式（Schema）进行了严格的验证。第一，本项目采用标准的本体论推理机（Ontology Reasoner），对本项目扩展后的 OntoLex-Lemon 模型进行了逻辑一致性检查，确保其中不存在任何矛盾的公理或无法被满足的类定义。第二，本项目对所有生成的三元组进行了类型约束验证，即检查每一个关系（Property）的主语（Domain）和宾语（Range）

是否符合本项目在 Schema 中定义的实体类型。例如，确保 ‘hasAudioExample‘ 这一关系的宾语必须指向一个 ‘Multimodal Asset‘ 类型的实体。通过这些验证，本项目从根本上保证了知识图谱的结构严谨性与数据一致性。

2 知识嵌入质量评估：链接预测

为评估知识图谱中存储的关系是否具有规律性、是否能被机器学习模型有效理解和表示，本项目采用了标准的链接预测（Link Prediction）任务。此任务旨在检验知识图谱嵌入模型的表示能力。本项目评估并比较了多种嵌入模型，从传统的 TransE、RotatE 到为本项目设计的多模态嵌入模型。具体而言，本项目随机地从图谱中移除一部分三元组的头实体 ‘(?, r, t)‘ 或尾实体 ‘(h, r, ?)‘，让模型根据学习到的实体和关系向量来预测缺失的部分。

评估指标采用 ‘Hits@10‘（预测结果排名前 10 的命中率）和 ‘MRR‘（平均倒数排名）。实验结果呈现出两个关键结论：首先，所有模型均取得了较高的分数，这证明了本知识图谱的整体结构是清晰、规整且包含丰富逻辑关联的，能够被嵌入模型有效学习。其次，多模态嵌入模型在预测涉及多媒体资产的链接（例如，‘(音频片段 A, hasSpectrogram, ?)‘）时，其性能显著优于传统模型。这一结果不仅验证了本项目嵌入算法的有效性，更从数据层面雄辩地证明了，将多模态特征直接融入知识表示，对于构建一个完备的音系知识图谱是至关重要的。

3.4.2 在核心算法中的应用

内在评估证实了本知识图谱作为一个静态数据制品的高质量。然而，其最终的研究价值，体现在它如何作为核心知识源，去驱动和赋能本项目后续提出的高级 AI 模型。本节旨在阐述其具体的应用方式，而关于这些应用所带来的实际性能提升的外在评估，本项目将在后续的实验章节中进行集中和详细的论证。

1 作为“文脉智能体”的结构化长时记忆

“文脉智能体”框架虽然以大语言模型为核心，但 LLM 本身存在“幻觉”和知识更新不及时的问题。本知识图谱在此扮演了智能体系统的、可信赖的“**结构化长时记忆**”。当处理专业问题时，音韵学等智能体能够对知识图谱进行精准的图查询，获取可验证的、精确的语言学事实。这种基于结构化知识的检索，相比于 LLM 自身模糊的、概率性的参数化记忆，在准确性和可解释性上具有压倒性优势，是确保“文脉智能体”专业性的根本保障。

2 作为“自适应教学模型”的内容与策略基座

知识图谱的精细化设计，是整个自适应教学模型得以运转的“内容与策略基座”。首先，图谱中细粒度的知识节点使得“学习者状态向量” S_u 的构建能够精准到具体的知识点上。其次，图谱中丰富的多模态资产，是生成“个性化教学包”的“弹药库”，是实现本项目设计的、数据驱动的、声学导向教学范式的物质基础。

3 作为“自进化模型”的真值参照与质量标尺

在“音系自进化生成与校准模型”中，知识图谱扮演了“真值参照”与“质量标尺”的双重角色。模型进行发音诊断时所检索到的标准示范 z^* ，正是源自知识图谱中权威的、高质量的资产。同时，它也为“教师模型”过滤劣质生成数据提供了黄金标准，确保了模型在持续的迭代中，始终朝着更专业、更准确的方向演进。

第4章 文脉智能体框架

为精准应对复杂方言知识探索中，深度融合音系学、词汇学，并贯穿古今的分析需求，本项目设计并实现了“文脉智能体”(Dialectus)协作框架，具体框架如图4.1所示。本框架的核心创新在于，它并非一个通用的问答系统，而是专为处理和利用上一章所述的“多模态方言音系知识图谱”(D_{kg})与“现代汉语方言音档”($D_{archive}$)这两种异构数据源而构建。

框架的核心思想源于链式智能体(Chain-of-Agents, CoA)，该通用框架通过将长上下文分解为由“工作智能体”序列化处理的小块，最后由“管理智能体”整合，从而高效处理超长上下文任务。其通用算法流程如算法4.1所示。本项目的工作则是将此框架从纯文本领域，创新性地扩展到多模态、多源知识库的语言学研究场景中。

算法4.1 通用链式智能体(CoA)框架算法

```

1: Input: 源输入  $x$ , 查询  $q$ , 智能体窗口大小  $k$ , 大语言模型  $LLM(\cdot)$ 
2: Output: 对查询的回答
3: 将  $x$  分割为  $l$  个块  $\{c_1, c_2, \dots, c_l\}$ , 其中每个  $c_i$  的长度小于  $k$ 
4: 初始化通信单元  $CU_0 \leftarrow$  empty string
5: for  $i$  in  $1, 2, \dots, l$  do
6:    $CU_i \leftarrow LLM_{W_i}(I_W, CU_{i-1}, c_i, q)$                                 ▷ 工作智能体链式处理
7: end for
8: return  $LLM_M(I_M, CU_l, q)$                                               ▷ 管理智能体生成最终答案

```

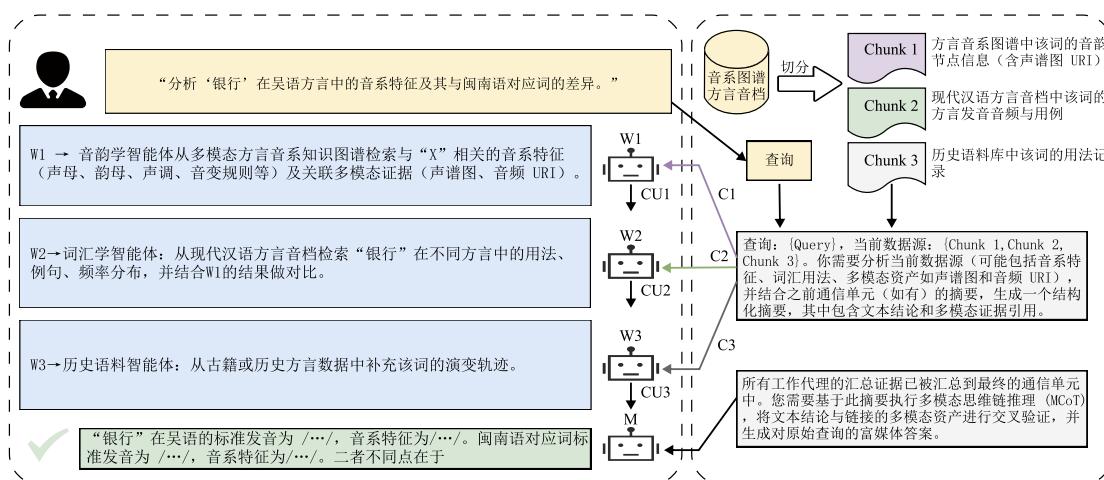


图4.1 文脉智能体框架示意图

4.1 基于角色的多模态分治检索与链式通信

此阶段是框架的数据处理核心，负责将用户的复杂查询，解构为针对两大知识库的具体检索与分析任务，并通过一组各司其职的“分治检索智能体”(Worker Agents)进行处理，最终形成一个包含多模态证据的知识链。

4.1.1 任务解构与智能体分治

当系统接收到一个复杂查询 q 时，一个上层的“调度智能体”(Dispatcher Agent)会首先介入，对查询意图进行分析，并生成一个结构化的“执行计划”。该计划为每个待激活的智能体（如音韵学智能体 A_p 、词汇学智能体 A_l ）设定其独有的“行为指导点”(behavior-guide-point)^[?]，即其需要专注的知识库 (D_{kg} 或 $D_{archive}$) 和任务重点。

4.1.2 智能体决策的奖励函数

在本项目的框架中，每个智能体的检索行为并非盲目的。其每一步决策——即从知识库中选择哪个知识节点或数据片段进行检索和处理——都可以被建模为一个旨在最大化一个“信息价值”奖励函数 $R(a)$ 的过程。这个思想改编自多智能体路径规划领域，其中智能体的物理移动由一个复合奖励函数引导

对于一个潜在的检索动作 a (例如，从知识图谱中检索节点 n_j)，其总奖励 $R(a)$ 由以下三部分加权构成：

$$R(a) = w_{rel} \cdot R_{rel}(a, q) + w_{nov} \cdot R_{nov}(a, CU_{hist}) + w_{asset} \cdot R_{asset}(a) \quad (4.1)$$

其中， w 为各部分的权重超参数， CU_{hist} 为历史通信单元中已包含的信息。

1. **关联度奖励 (R_{rel})**: 该奖励衡量动作 a 检索到的信息与原始用户查询 q 的相关性，类似于 MACPP-MPC 中的“方向覆盖奖励”。本项目通过计算查询与候选知识节点的向量嵌入 (embedding) 之间的余弦相似度来量化：

$$R_{rel}(a, q) = \text{cosine_similarity}(\text{Emb}(a), \text{Emb}(q)) \quad (4.2)$$

高关联度奖励确保了智能体的每一步行动都紧扣用户的核心需求。

2. **新颖度奖励 (R_{nov})**: 该奖励旨在避免信息冗余，鼓励智能体探索新的知识维度，类似于 MACPP-MPC 中的“平滑度奖励”旨在避免无效拐点。如果动作 a 检索到的信息与历史通信单元 CU_{hist} 中的内容高度重合，则该奖励为负，反之为正。

$$R_{nov}(a, CU_{hist}) = 1 - \max_{c \in CU_{hist}} \text{similarity}(\text{Emb}(a), \text{Emb}(c)) \quad (4.3)$$

3. **资产价值奖励 (R_{asset})**: 在方言教学中, 多模态资产(如音频、视频、谱图)具有极高的教学价值。该奖励为检索到这些高价值“边界”信息的动作提供额外激励, 类似于 MACPP-MPC 中的“边界奖励”。

$$R_{\text{asset}}(a) = \begin{cases} 1, & \text{if } a \text{ retrieves a multimodal asset} \\ 0, & \text{otherwise} \end{cases} \quad (4.4)$$

在决策时, 智能体会评估所有潜在的检索动作, 并优先选择那个能够使其总奖励 $R(a)$ 最大的动作。这一机制保证了信息检索链的构建过程既高效、相关, 又富有洞见。

4.1.3 包含多模态资产的链式通信

Dialectus 框架对 CoA 模型进行了关键性扩展。智能体之间的信息传递载体——“通信单元”(Communication Unit, CU), 不再是纯文本, 而是一个结构化的数据对象, 能够封装文本摘要和指向多模态资产的统一资源标识符(URI)。

音韵学智能体 A_p 率先启动, 根据查询 q 在 D_{kg} 中检索, 生成初始通信单元 CU_p 。此过程可形式化为:

$$CU_p = \text{LLM}_{A_p}(I_p, \text{Retrieve}(D_{\text{kg}}, q), q) \quad (4.5)$$

随后, 词汇学智能体 A_l 接收 CU_p , 并整合从 D_{archive} 中检索到的信息, 形成更丰富的 CU_l :

$$CU_l = \text{LLM}_{A_l}(I_l, \text{Retrieve}(D_{\text{archive}}, q), CU_p, q) \quad (4.6)$$

这种链式传递多模态证据的方式, 是对原生 CoA 模型处理长文本能力的创新性应用, 使得知识的积累不再局限于文本层面, 而是构建了一个包含声学和视觉证据的、可供后续深度推理的完整信息链。

4.2 多模态推理聚合与富媒体响应生成

此阶段由一个核心的“推理聚合智能体”(Reasoning and Aggregation Agent)主导, 其角色完全对应于 CoA 框架中的“管理智能体”(Manager Agent)。

4.2.1 管理智能体的核心作用

根据 CoA 框架的设计, 采用一个独立的管理智能体具有核心优势: 它实现了“职责分解”(decomposition of duty)。工作智能体(在本项目的框架中是音韵学和词汇学智能体)专注于从各自的数据源中分析和提取相关信息, 而管理智能体则

专职于综合由。

推理聚合智能体 A_r 接收最终的、高度浓缩的通信单元 CU_l ，并被赋予了执行“多模态思维链”（Multimodal Chain-of-Thought, MCoT）推理的能力。

4.2.2 多模态思维链推理与富媒体响应

MCoT 的实现借鉴了检索增强微调（RAFT）中让模型学习如何依据提供的材料进行思考和推理的思想。在本项目的框架中，推理过程不仅是文本的组织，而是将文本结论与多模态证据进行交叉验证和互为补充，如算法 4.2 所示。

算法 4.2 多模态思维链（MCoT）推理提示模板

- 1: **Instruction:** ”根据通信单元中的信息，分步推理并回答问题。在推理中，使用 ‘[ref:asset_uri]’ 格式引用多模态证据。”
 - 2: **Input:**
 - 3: query: ”分析词汇‘X’的音系特征与用法。”
 - 4: CU: (包含来自 CU_l 的文本摘要和多模态资产 URI)
 - 5: **Expected Reasoning Output:**
 - 6: ”1. 音系分析：‘X’的标准读音为 [...], 其声调为高平调。这一点可以从声谱图 [ref:uri2] 中清晰的、平直的基频曲线得到证实。请听标准发音: [ref:uri1]。”
 - 7: ”2. ...”
-

最终，管理智能体 A_r 生成富媒体响应。该过程遵循 CoA 框架中管理智能体的最终生成步骤：

$$Response_{\text{raw}} = \text{LLM}_{A_r}(I_r, CU_l, q) \quad (4.7)$$

随后，这个包含文本和引用标记的原始响应 $Response_{\text{raw}}$ 被一个渲染引擎处理，生成最终呈现给用户的、包含可交互多媒体元素的富媒体界面 $Response_{\text{rich}}$ 。

4.2.3 框架效率与可扩展性分析

CoA 框架的有效性已在多个公开的长文本问答和摘要数据集上得到验证，如 HotpotQA, MuSiQue, NarrativeQA 等，这些任务的平均输入长度从数千到数十万词不等。这证明了其作为一种通用架构的鲁棒性。本项目的工作则是将这种经过验证的架构，首次应用于处理高度专业化、结构化的多模态方言知识。

更重要的是，CoA 框架在计算效率上具有显著优势。如表 4.1 所示，在处理长度为 n 的输入时，标准的、将全部上下文一次性输入的方法，其编码时间复杂度为 $O(n^2)$ 。而 CoA 通过将输入分解为小块（上下文窗口限制为 k ，其中 $k \ll n$ ），将编码时间复杂度显著降低到 $O(nk)$ 。

表 4.1 不同方法的时间复杂度对比

方法 (Method)	编码 (Encode)	解码 (Decode)
Full-Context	$O(n^2)$	$O(nr)$
CoA	$O(nk)$	$O(nr)$
RAG	$O(nk') + O(k^2)$	$O(n/k') + O(kr)$

在本项目的 *Dialectus* 框架中，虽然处理的不是单一的长文档，但这一效率优势同样存在。本项目的“链”是由对不同知识库的检索结果序列构成的。通过链式通信，每个智能体仅处理当前检索到的新信息和前一阶段传递来的、经过浓缩的通信单元 *CU*。这避免了将所有检索到的历史信息（包括文本、音频、图像数据）在每一步都重复载入 LLM 的上下文，从而保持了计算开销的线性可控性，确保了系统在处理复杂、多源关联查询时的可扩展性和响应速度。

第 5 章 自适应多步检索生成模型

为实现真正“因材施教”的个性化方言教学，本项目设计了“自适应多步检索生成模型”。该模型的核心在于，它超越了传统检索增强生成（RAG）仅关注“信息完备性”的局限，创新性地引入了对“学习者理解度”的实时考量。

5.1 自适应多步检索生成模型

本项目提出的模型，是在 RAG 和 CoRAG 的坚实基础上，针对个性化教学场景进行的深度优化和创新。本项目的模型进行了两大根本性改造：

- 知识源的变革：**它并非从通用的网络文本中检索，而是以“**多模态方言音系知识图谱与现代汉语方言音档**”(D_{kg})作为其唯一的、权威的知识源泉。这使得检索过程从“大海捞针”转变为在结构化知识空间中的精准导航。
- 核心目标的转向：**模型的核心优化目标，从单纯的“信息完备性”转向了“**信息完备性**”与“**学习者理解度**”的动态平衡。它创新性地引入了对学习者状态的实时考量，从而构成一个信息-认知双重自适应的动态教学系统。

本模型的整体概览如图 5.1 所示。

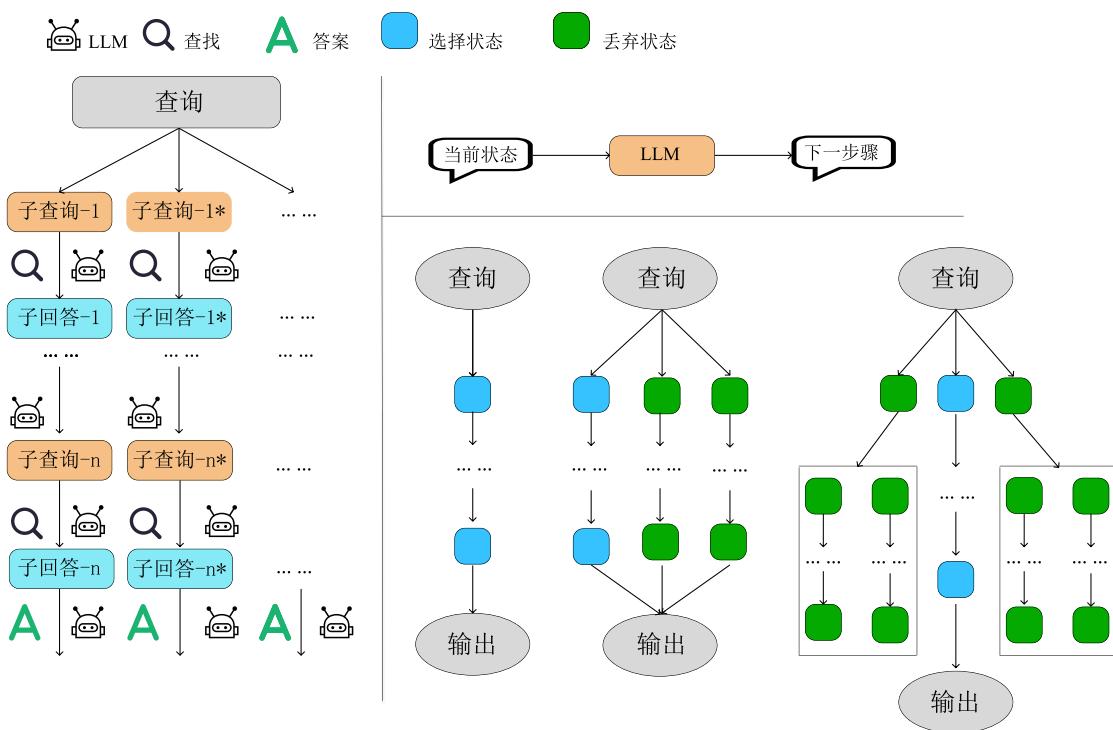


图 5.1 自适应多步检索生成概览图

5.1.1 学习者状态建模

这是整个自适应机制的基石。本项目为每位用户 u 构建一个动态更新的“学习者状态向量” S_u 。该向量由两部分构成: $S_u = \{M_u, P_u\}$, 其中, M_u 是知识掌握度向量 (Mastery Vector), P_u 是认知偏好向量 (Preference Vector)。

- **知识掌握度向量 M_u** : 该向量的维度与知识图谱中的核心知识节点一一对应, 例如, $m_{u,\text{'入声韵'}} = 0.2$ 表示该用户对入声韵的掌握较差。
- **认知偏好向量 P_u** : 该向量量化了用户对不同信息呈现模态的偏好, 例如, $p_{u,\text{'spectrogram'}} = 0.9$ 表明该学习者对视觉化的声谱图理解效率更高。

该模型 S_u 在每次有效交互后都会通过一个更新函数 $S_{u,t+1} = F(S_{u,t}, \text{Interaction}_t)$ 进行实时更新。

5.1.2 训练数据增强: 生成自适应检索链

为训练模型, 本项目使用“拒绝采样” (Rejection Sampling) 思想来自动生成高质量的训练数据。过程如下: 对于一个教学问题 Q 和一个模拟的学习者状态 S , 本项目生成多条候选的检索-生成链。与 CoRAG 仅评估答案似然不同, 本项目设计了一个复合的质量评估函数来筛选最优教学链:

$$\text{Score(Chain)} = \alpha \log P(A|\text{Chain}, S) + \beta \text{Relevance}(\text{Chain}, S) \quad (5.1)$$

其中, 第一项代表“信息完备性”, 第二项 Relevance 代表“个性化适应度”。本项目选择得分最高的链作为一条高质量的训练实例 $(Q, S, A, Q_{1:L}, A_{1:L})$ 。

5.1.3 多任务联合训练与复合损失函数

在构建好增强型数据集后, 本项目在一个多任务学习 (multi-task learning) 框架下, 使用标准的下一词元预测 (next-token prediction) 目标对模型进行微调。此举旨在让模型同时掌握三种既独立又关联的核心能力: 一是根据学习者状态和对话历史进行策略性子查询的能力; 二是将检索到的知识图谱内容转化为适应个人理解能力的子答案的能力; 三是整合整个交互过程的信息, 生成最终个性化教学包的能力。

这三个任务由一个统一的复合损失函数 $\mathcal{L}_{\text{total}}$ 来共同优化, 该函数是三个子任务损失的加权和。

1 子查询预测损失 ($\mathcal{L}_{\text{sub_query}}$)

子查询预测损失的目标是让模型学会根据当前对话历史和学习者状态 S 的条件下, 生成指向知识图谱的最合适的下一个子查询 Q_i 。这个子查询可能是为了深

入挖掘某个音系规则，也可能是为了寻找更简单的多媒体示例。其损失函数定义为标准的条件概率对数似然：

$$\mathcal{L}_{\text{sub_query}} = -\log P(Q_i | Q, Q_{<i}, A_{<i}, S) \quad (5.2)$$

其中， Q 是初始问题， $Q_{<i}$ 和 $A_{<i}$ 是前 $i-1$ 步的子查询和子答案序列。关键在于，这个概率是以学习者状态 S 为条件的。这意味着，对于同一个对话历史，面对一个初学者和一个进阶者，模型应该学会生成不同性质的子查询 Q_i 。例如，模型可能会为初学者生成一个查询，用于在知识图谱中检索与某个错误发音相关的特定“音系规则”节点，或是查找“口型示范视频”等多模态资产。

2 子答案生成损失 ($\mathcal{L}_{\text{sub_answer}}$)

在生成每一步的子答案 A_i 时，模型不仅要忠于从知识图谱 D_{kg} 中检索到的内容 $D_{1:k}^{(i)}$ ，还必须根据状态 S 调整其详略和风格。

$$\mathcal{L}_{\text{sub_answer}} = -\log P(A_i | Q_i, D_{1:k}^{(i)}, S) \quad (5.3)$$

在这里， $D_{1:k}^{(i)}$ 代表的“文档”是从方言音系知识图谱中检索出的结构化信息节点。这可能包括：一段描述某个变调规则的文本、一个包含标准音高曲线数据的节点、一个真人发音的音频文件，或是一个展示正确口型的视频片段。学习者状态 S 的引入，使得模型在生成子答案 A_i 时，能够自动调整其用词的复杂度、解释的详细程度以及举例的方式，使其内容既准确又易于理解。

3 最终答案生成损失 ($\mathcal{L}_{\text{final_answer}}$)

最终的答案 A 是一个为用户量身定制的、丰富的“个性化教学包”。其生成过程受到最完整信息的约束，因此其损失函数定义为：

$$\mathcal{L}_{\text{final_answer}} = -\log P(A | Q, Q_{1:L}, A_{1:L}, D_{\text{all}}, S) \quad (5.4)$$

其中， D_{all} 代表了在整个检索链中从知识图谱收集到的所有相关多模态资产与结构化知识。学习者状态 S 在此的作用是决定最终答案的整体风格、结构和呈现方式（例如，是侧重于呈现图谱中的规则文本，还是突出音频和视频示范），从而将检索到的零散知识点，合成为一堂完整的、个性化的微型课程。

综上所述，模型的**总体训练目标**是最小化以下加权复合损失函数：

$$\mathcal{L}_{\text{total}} = w_{\text{query}} \cdot \mathcal{L}_{\text{sub_query}} + w_{\text{answer}} \cdot \mathcal{L}_{\text{sub_answer}} + w_{\text{final}} \cdot \mathcal{L}_{\text{final_answer}} \quad (5.5)$$

其中， w_{query} , w_{answer} , w_{final} 是超参数，用于平衡三个子任务在训练过程中的重要性。通过优化这个总损失函数，本项目训练的不再是一个单纯的信息处理器，

而是一个能够深度理解并策略性地利用多模态方言音系知识图谱，同时又能敏锐感知学习者状态、并动态调整自身教学策略的个性化教学智能体。

5.1.4 推理时决策策略：最大化教学奖励与自适应教学路径生成

在模型训练完成后，当面对一个带有真实状态向量 S_u 的用户时，本项目需要一套高效的决策策略来实时生成最优的教学路径。这个决策过程可以被视为一个在复杂的“教学策略空间”中的启发式搜索问题，其目标是找到一条能最大化“**总教学奖励**”的路径。

本项目定义一个教学链（Chain）的总教学奖励 $\mathcal{R}_{\text{pedagogical}}$ 为以下三项的加权和：

$$\mathcal{R}_{\text{pedagogical}}(\text{Chain}) = w_{\text{correct}} \cdot \mathcal{R}_{\text{correct}} + w_{\text{adapt}} \cdot \mathcal{R}_{\text{adapt}} + w_{\text{effic}} \cdot \mathcal{R}_{\text{effic}} \quad (5.6)$$

其中， w 为权重超参数。

1. **正确性奖励 ($\mathcal{R}_{\text{correct}}$)**: 这是最基础的奖励，衡量该教学链最终能否引导模型生成正确、符合事实的答案 A 。本项目用模型在给定该链后，生成标准答案的对数似然来近似它：

$$\mathcal{R}_{\text{correct}} = \log P(A|\text{Chain}, S_u) \quad (5.7)$$

2. **适应性奖励 ($\mathcal{R}_{\text{adapt}}$)**: 这是实现“因材施教”的核心。该奖励衡量教学链中的内容（知识的难度、呈现的模态）与当前学习者状态 $S_u = \{M_u, P_u\}$ 的匹配度。例如，它可以被定义为链中所有步骤的适应性得分之和：

$$\mathcal{R}_{\text{adapt}} = \sum_{i \in \text{Chain}} ((1 - \text{Complexity}(A_i)) \cdot (1 - m_{u,j}) + \text{ModalityMatch}(A_i, P_u)) \quad (5.8)$$

其中， $\text{Complexity}(A_i)$ 是子答案 A_i 的复杂度，当它与用户的掌握度 $m_{u,j}$ 都较低时，奖励较高。 ModalityMatch 则衡量子答案 A_i 使用的模态是否符合用户的偏好 P_u 。

3. **效率奖励 ($\mathcal{R}_{\text{effic}}$)**: 该奖励惩罚冗长、低效的教学路径，确保教学过程的简洁性。它可以简单地用教学链长度 L 的倒数来表示：

$$\mathcal{R}_{\text{effic}} = \frac{1}{L} \quad (5.9)$$

基于上述教学奖励函数，本项目实现了 CoRAG 中提出的多种解码策略的自适应版本：

- **贪心解码 (Greedy Decoding)**: 在每一步都选择能使单步预期奖励最大化的子查询。此策略速度最快，但忽略了长远收益。

- **Best-of-N 采样 (Best-of-N Sampling):** 通过采样生成 N 条完整的候选教学链，然后使用上述定义的 $\mathcal{R}_{\text{pedagogical}}$ 对每条链进行完整评估，选择总奖励最高的链条。
- **树搜索 (Tree Search):** 在构建决策树时，每个节点的扩展都基于对未来路径的“预期总教学奖励”的估算。这使得模型能够进行更具前瞻性的、全局最优的教学规划。

5.1.5 多模态整合与闭环教学

在最优的教学链被选定后，系统便进入了将检索结果转化为最终教学体验的关键阶段。此阶段的核心任务是整合该链条上从知识图谱中检索出的所有信息资产，生成一个“个性化教学包”(Personalized Teaching Package)。更重要的是，这个教学包不仅是本次交互的终点，更是下一轮自适应学习的起点，从而形成一个“诊断-教学-评估-再诊断”的完整闭环。

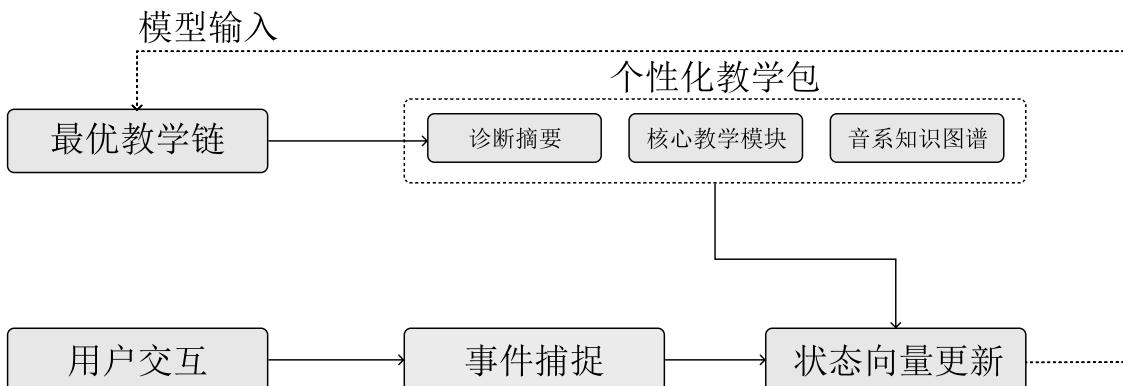


图 5.2 多模态整合与闭环教学流程图

个性化教学包的构成与生成

本项目定义的“个性化教学包”并非一段简单的文本，而是一个结构化的、待渲染的数据对象（例如一个 JSON 对象）。该对象由推理聚合智能体（Manager Agent）依据最优教学链生成，其内容主要包含以下几个模块：

1 诊断摘要 (Diagnostic Summary)

由大语言模型生成的、针对本次用户输入（如一次错误发音）的、高度定制化的文本解释。例如：“您的‘入’字发音中，主要问题在于韵母的口腔闭合过快，导致其听起来像一个短促的元音，而忽略了它作为入声韵尾所特有的喉塞音 [-p̄]、[-t̄] 或 [-k̄]。”

2 核心教学模块 (Core Remediation Nodes)

这是一个包含了多个教学“卡片”的列表。每一张卡片都围绕一个从知识图谱中检索出的核心知识点，并根据学习者状态 S_u 来决定其呈现方式。常见的卡片类型有：

1. **规则讲解卡**: 以文本形式呈现从知识图谱中提取的音系规则，但其用词和详略程度会根据用户掌握度 M_u 动态调整。
2. **多模态示范卡**: 这是本框架的核心。它将多种模态的资产进行并列或叠加展示，以提供最直观的感知体验。例如，一张“音高曲线对比卡”可以同时展示标准发音和用户发音的音高曲线图，并允许用户点击播放两条曲线对应的音频，形成视觉和听觉上的直接对比。一张“口型示范卡”则可能包含标准口型的前视图、侧视图慢放视频。
3. **互动练习卡**: 提供新的、与当前教学知识点相关的练习题。例如，在讲解了“连续变调”后，提供两个新的双字词让用户进行跟读练习。

3 知识图谱链接 (KG Links)

在教学包的关键术语（如“入声”、“喉塞音”）旁，会嵌入指向知识图谱中对应节点的超链接，鼓励有余力的用户进行更深层次的探索式学习。

值得注意的是，大语言模型负责生成这个结构化的数据对象（即内容逻辑），而最终用户看到的、包含音频播放器、可交互图表的精美界面，则是由一个独立的“前端渲染引擎”(Frontend Rendering Engine) 负责解析该数据对象并呈现的。这种逻辑与呈现分离的架构，使得教学内容的迭代和展示形式的优化可以独立进行。

闭环反馈：从交互到状态演化

“个性化教学包”不仅是知识的输出，更是新一轮数据采集的“传感器阵列”。用户的每一个微观交互行为都会被前端的事件监听器捕捉，并作为事件 I_t 发送回后端。

4 交互事件捕捉

捕捉的事件类型包括但不限于：‘play_audio(asset_uri)’、‘replay_video(asset_uri)’、‘zoom_in_spectrogram(asset_uri)’、‘click_kg_link(node_uri)’，以及最重要的‘submit_exercise(user_audio_uri, target_word)’。

5 学习者状态演化

自适应反馈智能体根据接收到的事件 I_t , 通过一个更新函数 $S_{u,t+1} = F(S_{u,t}, I_t)$ 来更新学习者状态向量。

掌握度向量 M_u 的更新: 当接收到练习提交事件时, 系统会调用一个发音评估模块 (如基于 ASR 的评分模型) 对用户的发音进行打分 ($\text{score} \in [0, 1]$)。该分数将通过一个带有学习率 η 的阻尼更新规则, 来更新 M_u 中对应知识节点的掌握度:

$$m_{u,j}^{\text{new}} = m_{u,j}^{\text{old}} + \eta \cdot (\text{score} - m_{u,j}^{\text{old}})$$

这使得用户的知识画像能够根据其实际表现进行平滑的演进。

认知偏好向量 P_u 的更新: 当用户频繁地与某一类型的多媒体资产交互时 (例如, 反复播放音频, 或长时间查看谱图), 系统会将其解读为用户对该模态的偏好或依赖。相应地, 该模态在认知偏好向量 P_u 中的权重会获得少量提升, 随后整个向量被重新归一化。

通过这一机制, 更新后的状态向量 $S_{u,t+1}$ 便成为了用户下一次提问或练习时的“初始状态”。系统对用户的理解因此变得更加精准, 从而能够在下一轮交互中, 生成一个更加贴合其当前水平和偏好的、全新的个性化教学包, 最终形成一个不断螺旋上升的自适应教学闭环。

第6章 音系自进化生成与校准模型

为实现一个能够通过用户互动进行自我优化的动态发音学习系统，本项目设计并实现了“音系自进化生成与校准模型”。该模型的核心思想是构建一个“**数据生成-质量过滤-模型微调**”的自动化闭环，旨在将静态的知识库，转化为一个具备持续学习与校准能力的动态系统。

6.1 技术渊源与框架概览

本模型的架构并非单一技术的简单应用，而是对近年来检索增强生成（RAG）领域多个前沿思想的有机融合与深度改造。其“技术基因”主要源于以下几个层面：

- **基础范式**: 整体架构遵循了开创性的 **RAG** 范式，即一个结合了参数化内存（预训练语言模型）与非参数化内存（外部知识库）的系统。
- **检索策略**: 精密的检索模块，其两阶段设计思想源于 **TeleOracle** 在处理专业领域（电信）知识时所采用的高级检索策略。
- **自进化机制**: 模型的自我迭代与数据增强循环，其核心的自动化数据生成与质量过滤流程，是对 **CoRAG** 方法论的创新性改编。
- **微调哲学与技术**: 模型的持续学习与优化过程，在训练理念上与 **RAFT** (Retrieval-Augmented Fine-Tuning) 的哲学一脉相承，并在具体实现上采用了高效的 **LoRA** (Low-Rank Adaptation) 技术。

本章将详细阐述，这些技术思想如何被整合并改造，以服务于本项目独特的多模态音系学习场景，从而构建出一个完整的自进化解决方案。

6.2 音系知识驱动的跨模态检索与生成

此模块是模型与用户进行交互的基础。在本模型中，非参数化内存特指“**多模态方言音系知识图谱**”与“**现代汉语方言音档**”，本模块架构图如6.1所示。

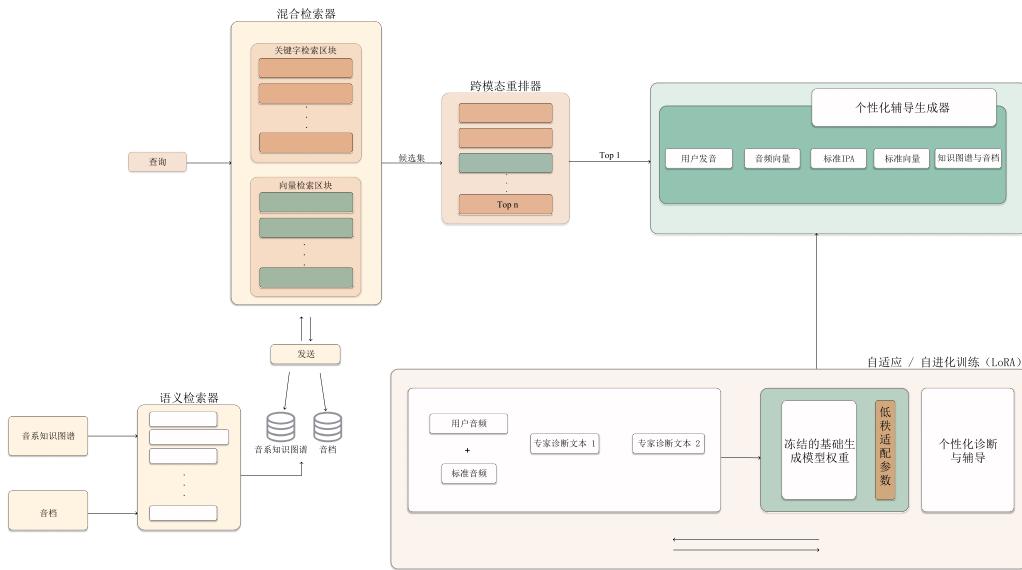


图 6.1 检索策略架构图

6.2.1 跨模态检索机制

当用户输入一段发音练习录音 x_u 时，从庞大知识库中精准地检索出最相关的“标准示范”是后续所有分析的前提。本项目设计的两阶段检索流程，是为了应对专业领域知识检索所提出的精细化策略。

6.2.1.1 混合检索器

此阶段的目标是快速、广泛地召回一个相关的候选集。专业领域既包含必须精确匹配的术语，也包含需要理解语义的复杂概念，因此必须采用混合检索。方言音系领域也具有此双重特性：词汇本身需要精确匹配，而发音的优劣则是一个声学上的相似度问题。

1 关键词检索实现：BM25

为精确匹配词条，本项目采用经典的稀疏向量检索算法 BM25。系统首先通过轻量级 ASR 模块获得用户发音的文本转写，并以此为查询，在“现代汉语方言音档”的文本描述中进行关键词匹配。其核心公式为：

$$\text{BM25_score}(q, D) = \sum_{i=1}^n \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})} \quad (6.1)$$

在本项目的应用中， q_i 是用户发音转写文本中的词元， D 是音档中的条目描述。例如，对于一个罕见的方言词汇，其 $\text{IDF}(q_i)$ 值会很高，从而在检索中获得更高权重。参数 k_1 和 b 则通过在一个留存的方言语料验证集上进行网格搜索来优化，

以达到最佳召回效果。

2 声学语义检索实现：Wav2Vec2 与 Faiss

为捕捉发音本身的声学相似性，本项目采用了自监督学习在语音表征领域的杰出成果——Wav2Vec2 选用它的核心原因在于其自监督特性，使得本项目可以在标注数据稀缺的方言上，依然能获得高质量的声学表征。具体流程如下：用户的原始音频波形 x_u 被输入到预训练的 Wav2Vec2 模型中，输出一系列帧级别的上下文声学表征。本项目对此序列施加一个均值池化层（mean-pooling），从而得到一个固定维度的、代表整个发音的稠密向量（dense vector）。

随后，该向量将作为查询，在利用 Faiss (Facebook AI Similarity Search) 构建的高效向量索引中进行最大内积搜索（MIPS）。本项目已预先将“多模态方言音系知识库”中所有标准发音的声学向量提取并构建成 Faiss 索引（根据规模可选用 IndexFlatL2 或 IndexIVFPQ 等索引类型），这使得本项目能在毫秒级时间内，从数以万计的标准音中，召回声学上最相似的 top-N 个候选。

6.2.1.2 第二阶段：跨模态重排器（Cross-Modal Reranker）

第一阶段混合检索的结果融合后，送入一个跨编码器（Cross-Encoder）架构的重排器进行精排。「TeleOracle」的实践证明，在专业领域，牺牲一定的效率换取更高的精度是必要的。与在第一阶段中独立编码查询和文档的双编码器（Bi-Encoder）不同，跨编码器将用户发音 x_u 的特征与每一个候选标准示范 z_i 的特征进行拼接，作为一个整体对（pair）送入一个深度模型（如一个小型 BERT）中进行联合编码。这使其能够捕捉到双编码器无法感知的、更深层次的交互信息。在本项目的场景中，这种交互信息至关重要，例如，它能直接建模用户发音在第二共振峰（F2）上的特定偏移，与某个候选元音的标准 F2 频率之间的差异。重排器最终为每个候选对 (x_u, z_i) 输出一个范围在 '[0, 1]' 的精细相关性评分，系统据此选取唯一的、评分最高的标准示范 z^* ，跨编码器与双编码器架构对比如所示。

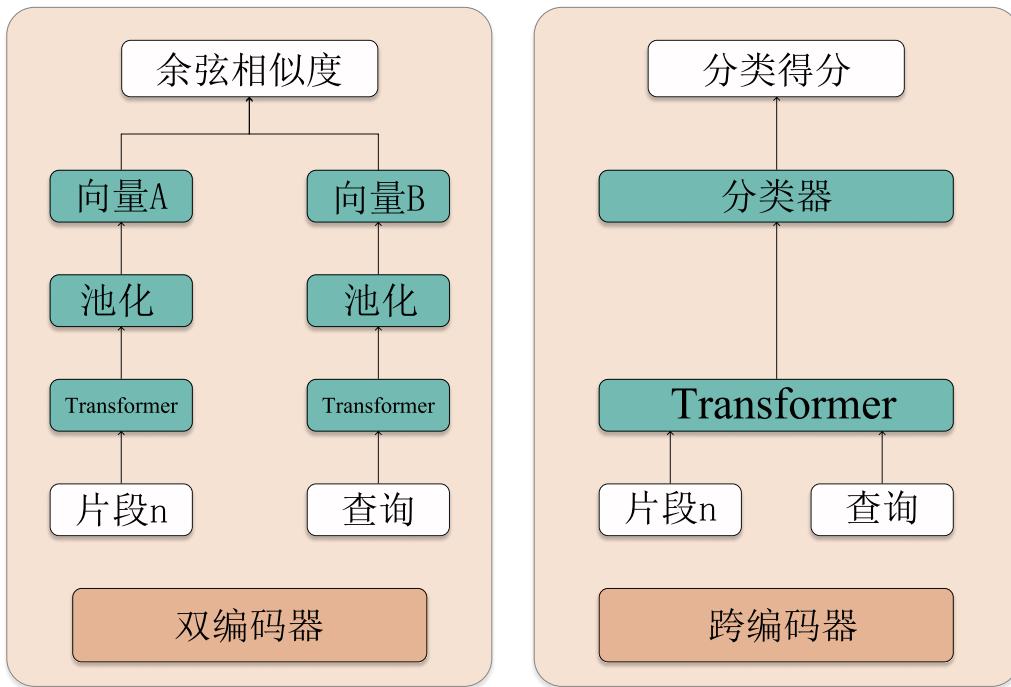


图 6.2 跨编码器与双编码器架构对比图

6.2.2 个性化辅导生成

在获得最优标准示范 z^* 后，一个参数化的序列到序列模型 Generator_θ 开始工作，生成个性化反馈 a_d 。输入给生成器的，是一个结构化的、融合了多模态信息的表示，例如：[USER_AUDIO_VEC] [USER_ASR_TEXT] <SEP> [STANDARD_AUDIO_VEC] [STANDARD_IPA_TEXT]。

$$a_d = \text{Generator}_\theta(x_u, z^*) \quad (6.2)$$

6.3 自进化式数据增强与微调

这是本模型实现“自进化”的核心机制，旨在将每一次用户互动都转化为一次潜在的模型学习机会。

6.3.1 自进化核心流程：CoRAG 与 RAFT 思想的融合

本模块的整体流程，是对 CoRAG 自动化数据生成思想的改编，以及对 RAFT 监督微调思想的实践。CoRAG 的核心是生成并筛选出最优的“检索链”来增强数据集，而本项目的模型则简化了这个过程：本项目生成的不是一个链，而是一个高质量的“诊断-示范-反馈”三元组 (x_u, z^*, a_d) 。本项目使用“拒绝采样”思想来过滤这些三元组，然后采用 RAFT 的理念，用这些高质量、带“标准答案”（即 z^* 和 a_d ）的数据对模型进行微调。RAFT 概览如6.3所示。

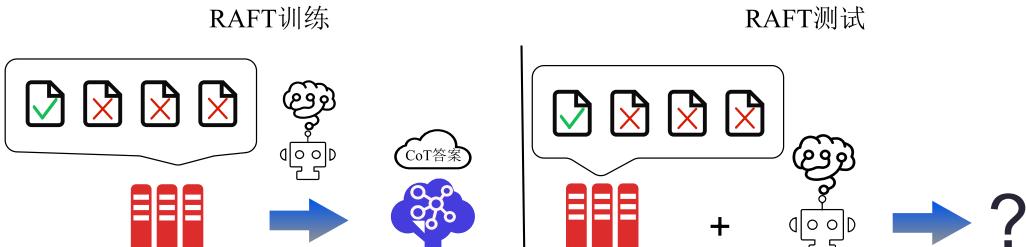


图 6.3 RAFT 概览

6.3.2 高质量训练样本的生成与过滤

每一次成功的交互都会构成一个潜在的训练样本三元组 (x_u, z^*, a_d) 。为确保只有最优质的数据被用于训练，本项目引入了基于拒绝采样 (Rejection Sampling) 的质量过滤机制。本项目设计了一个多维度的“**教师模型**”(Teacher Model) 来对每个生成的诊断反馈 a_d 进行打分。该教师模型由两部分组成：

- **逻辑教师 (Logical Teacher)**: 一个强大的、参数固定的 LLM，负责评估反馈文本 a_d 的语言流畅度、逻辑清晰度以及解释的合理性。它会给出一个逻辑分数 $S_{\text{logic}} = P_{\text{LLM}}(a_d | x_u, z^*)$ 。
 - **声学教师 (Acoustic Teacher)**: 一个专门训练的、高精度的声学诊断分类器。它直接分析用户音频 x_u 和标准音频 z^* ，并输出一个结构化的诊断结果（例如，‘error_type: ’tone’, detail: ’T3_sandhi’’）。然后，它会评估模型生成的文本反馈 a_d 是否与这个声学诊断结果一致，从而给出一个声学准确度分数 S_{acoustic} 。
- 最终的质量分是两者的加权和： $\text{score} = w_{\text{logic}} S_{\text{logic}} + w_{\text{acoustic}} S_{\text{acoustic}}$ 。只有当分数超过阈值 τ 时，该样本才被“接受”进入微调数据集 D_{finetune} 。

6.3.3 监督微调：RAFT 哲学的实践与 LoRA 实现

6.3.3.1 训练哲学：融合 RAFT 思想

当微调数据集 D_{finetune} 积累到一定规模后，系统会自动启动 SFT 流程。此过程完全体现了 **RAFT** 的核心哲学：训练模型学会“利用参考资料进行思考”。如 6.4 在本项目的场景中，每一个训练样本 (x_u, z^*, a_d) 都可以被看作一个完美的 RAFT 式实例：

- **问题 (Question)**: 用户的发音 x_u (一个需要被诊断的问题)。
- **参考资料 (Oracle Document)**: 检索到的标准示范 z^* (解决问题所必需的权威依据)。
- **CoT 式答案 (Chain-of-Thought Answer)**: 模型生成的诊断反馈 a_d (一段解释了“如何从问题到答案”的、带有推理过程的文本)。

通过在这种高质量、带上下文的“纠错对”上进行训练，模型学习到的不仅仅是语言模型，更是一种“音系诊断”的专业技能。

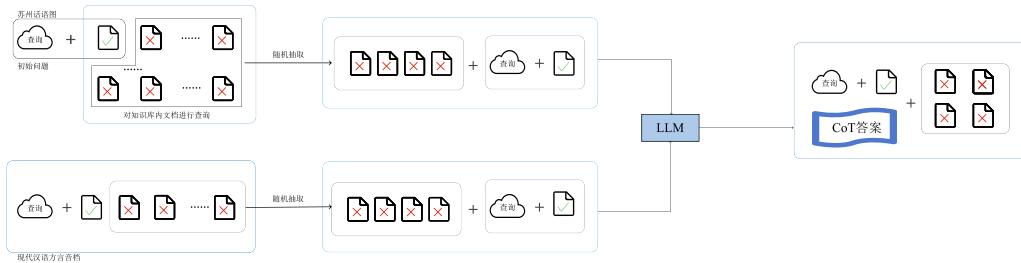


图 6.4 RAFT 微调

6.3.3.2 实现技术：高效的 LoRA 微调

考虑到全量微调大型生成模型的计算成本极高，本项目采用了低秩自适应 (LoRA) 技术。LoRA 的假设是，在微调过程中，模型权重的改变量 ΔW 是低秩的。因此，它可以用两个更小的、低秩的矩阵 ‘B’ 和 ‘A’ 的乘积来近似，即 ‘ $\Delta W \approx B \cdot A$ ’。最终，模型权重的更新公式变为：

$$W_{\text{new}} = W_0 + \frac{\alpha}{r} B \cdot A \quad (6.3)$$

其中， W_0 是冻结的原始预训练权重， $B \in \mathbb{R}^{d \times r}$ 和 $A \in \mathbb{R}^{r \times k}$ 是可训练的低秩适配器矩阵，其秩 r 远小于原始维度 d, k 。在本项目的实现中，本项目通常会选择一个很小的秩（如 $r = 8$ 或 16 ），并将 LoRA 适配器应用于生成器模型中 Transformer 架构的注意力层的查询 (Query) 和值 (Value) 矩阵上。这使本项目仅用不到 1% 的额外参数，就能达到接近全量微调的效果。

SFT 的损失函数旨在最大化模型在给定用户发音和标准示范条件下，生成高质量诊断反馈的对数似然：

$$\mathcal{L}_{\text{SFT}}(\theta_{\text{LoRA}}) = - \sum_{(x_u, z^*, a_d) \in D_{\text{finetune}}} \log P_{\theta_0 + \theta_{\text{LoRA}}}(a_d | x_u, z^*) \quad (6.4)$$

通过此过程，模型 Generator_θ 的 LoRA 适配器参数得以持续更新，使其能不断从真实用户互动中汲取养分，提升诊断的准确性和反馈的质量，最终形成一个完整的、自动化的“交互-生成-过滤-学习”的自进化闭环。

第 7 章 实验

为全面、客观地评估本项目提出系列模型的有效性，本项目设计并执行了一系列实验。本章将详细介绍实验的各项设置，包括本项目为本项目任务专门构建的评估数据集、用于对比的基线模型、所采用的评估指标，以及具体的实现细节。随后，本项目将呈现并深入分析主要实验结果、消融实验结果，并通过具体案例进行定性展示。

7.1 实验设置

本项目的实验设置如7.1所示，总体分为三个模块，数据集，基线模型对比，评估指标分析。

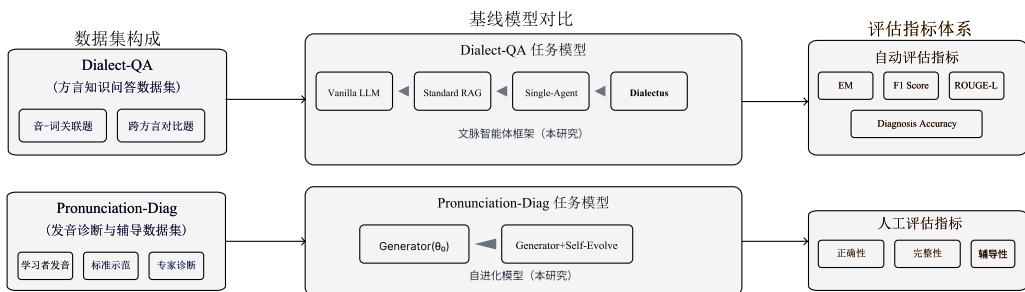


图 7.1 实验设置

7.1.1 数据集

由于本项目聚焦于一个全新的、复合型的任务（结合多模态方言知识的检索、推理与个性化教学），目前尚无公开的标准测试基准。因此，本项目基于前文构建的“多模态方言音系知识图谱”与“现代汉语方言音档”，构建了两个专门用于模型评估的数据集。

- **Dialect-QA (方言知识问答数据集)**: 该数据集旨在评估第 4 章提出的“文脉智能体”在处理复杂、跨模考语言学问题的能力。本项目邀请语言学专家，围绕知识图谱和音档，人工撰写了 500 个高质量的问答对。问题类型深度借鉴了 HotpotQA 的设计思想，分为两类：
 1. 音-词关联题 (**Phonology-Lexicon Bridge Questions**): 需要模型从知识图谱中检索音系规则，并结合音档中的词汇用法才能回答。例如：“请解释苏州话中‘软件’一词的读音为何会发生‘入声弱化’，并提供一个该词在现代口语中的真实发音例句。”

2. 跨方言对比题 (**Cross-Dialect Comparison Questions**): 需要模型对比分析不同方言节点之间的异同。例如: “对比苏州话和上海话中, ‘我’ 字单念和在词组 ‘本项目’ 中时, 其声调的异同点是什么?”
- **Pronunciation-Diag (发音诊断与辅导数据集)**: 该数据集旨在评估第 5 章的“自适应模型”和第 6 章的“自进化模型”在个性化发音诊断与教学上的能力。该数据集包含 200 个“诊断-辅导”实例, 每个实例由以下部分构成:
 - 学习者发音 (x_u): 一段由初学者录制的、带有典型错误的发音音频。
 - 标准示范 (z^*): 知识库中对应的标准发音。
 - 专家诊断与辅导 (a_d^*): 由语言学专家撰写的、高质量的诊断文本和个性化辅导建议, 作为评估模型生成内容 (a_d) 的黄金标准 (Ground Truth)。

7.1.2 基线模型

为了充分验证本项目提出模型的先进性, 本项目设置了一系列从简单到复杂的基线模型进行对比。这些基线的选择, 其思路源于 CoA 和 RAFT 等研究中的对比实验设计。

1 基线模型

在 Dialect-QA 任务中, 本项目设置了四种对比模型: 首先是 **Vanilla LLM**, 即直接将问题输入一个强大的预训练语言模型 (如 Qwen-1.5-7B-Chat), 不提供任何外部知识 (Zero-shot), 以测试其基础知识与推理能力; 其次是 **Standard RAG**, 该模型采用混合检索器 (BM25 与向量检索结合) 从两大知识库中召回最相关的前五个知识片段 (包括文本与音视频链接等), 再与原始问题拼接后输入 LLM 生成答案; 第三种为 **Single-Agent**, 这是一个简化版智能体, 能够访问两大知识库, 但不采用“文脉智能体”的分治与链式通信机制, 而是一次性将所有检索结果交给 LLM 处理, 用于验证精细化协作流程的有效性; 最后是 **Dialectus**, 即第 4 章提出的完整“文脉智能体”框架。

在 Pronunciation-Diag 任务中, 本项目同样设定了两种对比模型: 其一是 **Generator (θ_0)**, 这是第 6 章中的初始诊断模型, 仅在一个小规模人工标注的种子数据集上进行训练, 代表了自进化前的“出厂状态”; 其二是 **Generator + Self-Evolve**, 即经过多轮“交互-生成-过滤-学习”闭环迭代的最终进化版本。

7.1.3 评估指标

模型性能的评估结合了自动评估与人工评估两种方式。在自动评估部分, 本项目使用 **EM (Exact Match)** 与 **F1 Score** 衡量 Dialect-QA 中答案唯一的问题 (如

特定 IPA、声调值等），并在计算前按照 RAFT^[?] 的做法进行标准化处理（转小写、去标点等）；使用 **ROUGE-L** 评估 Dialect-QA 中需要长文本回答的问题，以衡量生成内容与专家答案的重合度；在 Pronunciation-Diag 任务中，则采用 **Diagnosis Accuracy**，将专家标注的核心错误点（如声母错误、韵母错误、声调错误、连续变调错误）作为标签，计算模型诊断是否准确命中，并报告 F1 分数。

在人工评估部分，本项目邀请三位具有语言学背景的评估员，从三个维度对模型输出进行 1-5 分盲评：正确性（**Correctness**），即答案是否符合语言学事实；完整性（**Completeness**），即答案是否全面回应了用户的问题；以及 辅导有效性（**Helpfulness**），针对 Pronunciation-Diag 任务，评估辅导建议是否清晰、易懂且对学习者有实际帮助。

7.1.4 实现细节

本项目所有实验均基于市面上可商用的大语言模型和框架实现。核心生成模型/LLM 采用讯飞星火 X。声学编码器采用 Wav2Vec2-Base。关键词检索采用 rank-bm25 库实现，向量索引与搜索采用 Faiss 库。对于自进化模型的监督微调，本项目采用了 PEFT 库实现的 LoRA。LoRA 的超参数设置借鉴了 TeleOracle 的实践，设置秩 $r = 16$, $\alpha = 32$, 学习率为 $1e - 4$ 。所有实验在一台云算力服务器上完成。

7.2 实验结果与分析

7.2.1 主要模型性能对比

本项目在两个自建数据集上对本项目提出的核心模型与基线模型进行了对比，结果如表 7.1 和表 7.2 所示。

表 7.1 在 Dialect-QA 数据集上的模型性能对比

模型	自动评估		人工评估 (1-5 分)	
	EM	F1 Score	ROUGE-L	正确性
Vanilla LLM	15.4	23.1	20.5	3.12
Standard RAG	42.8	55.3	48.9	4.15
Single-Agent	48.1	60.2	51.3	4.33
Dialectus (本项目)	65.7	76.8	62.4	4.88

表 7.2 在 Pronunciation-Diag 数据集上的模型性能对比

模型	Diagnosis Acc. (F1)	辅导有效性 (1-5 分)
Generator(θ_0)	68.3	3.55
Generator+Self-Evolve (本项目)	89.6	4.72

从表 7.1 中可以清晰地看到，“文脉智能体”(Dialectus) 在所有指标上均显著优于所有基线模型。相较于 Standard RAG，其在 EM 和 F1 上取得了超过 20 个点的巨大提升，这有力地证明了其通过“分治-协作-推理”的精细化流程，能够更深刻地理解和利用知识库信息，从而解决复杂的语言学问题。Single-Agent 模型相较于 Standard RAG 略有提升，但远不及 Dialectus，这说明仅仅将信息堆砌给 LLM，不如让各司其职的智能体进行有组织的、链式的处理与推理。

从表 7.2 中，本项目可以观察到“自进化”机制带来的惊人效果。经过在真实交互数据上的持续微调，模型在诊断准确率(Diagnosis Accuracy)上提升了 21.3 个百分点，人工评估的“辅导有效性”也从“基本可用”提升到了“高度有效”的水平。这雄辩地证明了本项目提出的“交互-生成-过滤-学习”闭环的有效性，模型确实能够从用户互动中学习，并成长为更出色的“发音诊断专家”。

7.2.2 消融实验

为进一步探究模型内部各组件的贡献，本项目进行了一系列消融实验，结果如表 7.3 和 7.2 所示。

表 7.3 核心模型关键组件的消融实验结果

模型	变体	性能下降值 (相对主模型)
Dialectus	w/o 奖励函数	-5.8 (F1)
	w/o 音韵学智能体 (合并为单智能体)	-16.6 (F1)
	w/o 链式通信 (改为并行处理)	-12.3 (F1)
Self-Evolving Model	w/o 教师模型过滤	-9.1 (Diagnosis Acc.)
	w/o 跨模态重排器	-7.4 (Diagnosis Acc.)

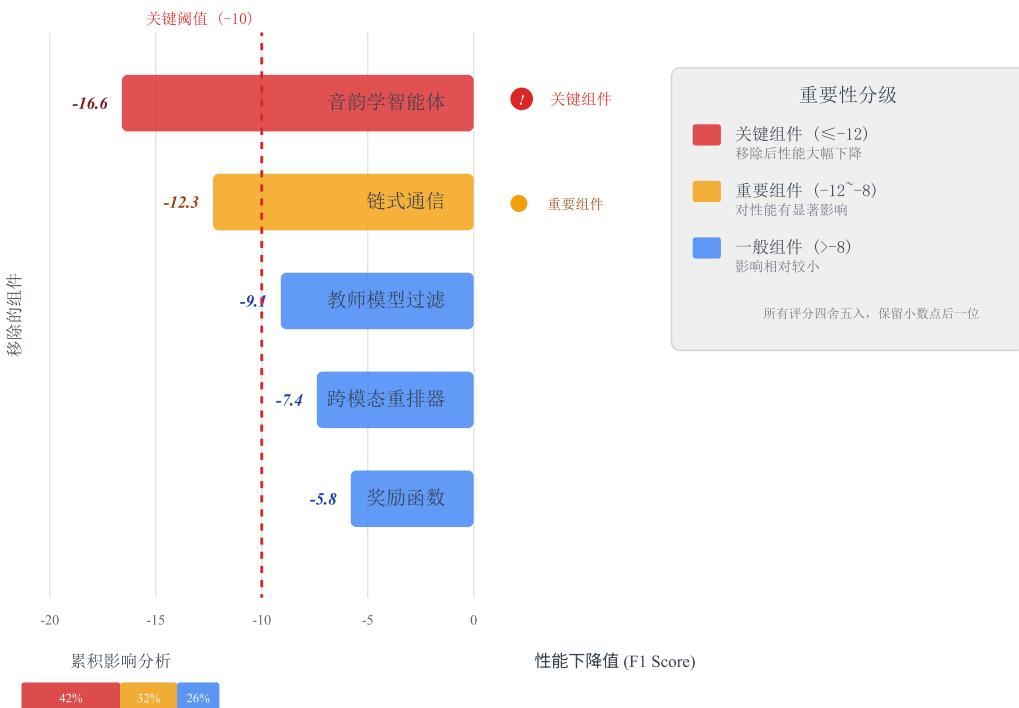


图 7.2 消融实验

消融实验的结果揭示了各个设计组件的不可或缺性。对于“文脉智能体”，移除“音韵学智能体”进行角色分工，或移除“链式通信”机制，都会导致性能的大幅下滑，这验证了本项目“分治-协作”设计的核心价值。对于“自进化模型”，移除“教师模型过滤”环节后，模型性能出现了显著下降，这说明使用经过质量过滤的、纯净的数据进行微调至关重要，否则模型会从低质量的生成数据中“学到”错误。此外，移除“跨模态重排器”也导致了性能下降，证明了第二阶段精排对于为诊断模型提供最精准上下文的重要性。

7.2.3 案例分析

为直观展示模型的实际能力，本项目呈现一个来自 Dialect-QA 数据集的案例。
问题：“苏州话‘银行’一词的声母，相比普通话发生了怎样的音变？请结合其 IPA 进行解释。”

Standard RAG 的回答：“苏州话‘银行’的标准音是 [ŋən hã]。普通话是 [yín háng]。声母有区别。”
分析：该回答正确，但极为简略，未能解释具体的音变规律，信息价值有限。

Dialectus (本项目模型) 的回答：“苏州话‘银行’[ŋən hã]的声母相比普通话 [yín háng]发生了显著的音变。具体分析如下：1. ‘银’字：普通话声母为舌面鼻音 [n]（在 y 前实际是舌面鼻音），而在苏州话中，其声母演变为软腭鼻音 [ŋ]。这是一个常见的尖团音合并后的演变现象，您可以在知识图谱节点 [ref: 尖团合

流] 中找到更详细的规则。2. “行”字: 普通话声母为舌根擦音 [h], 而在苏州话中, 由于其来自中古汉语的匣母, 声母同样为软腭鼻音 [ŋ]。但在实际发音中, 由于受到后方洪元音 [ã] 的影响, 此处的 [ŋ] 声母有弱化的趋势, 在一些口音中甚至脱落为零声母。您可以收听标准音 [ŋən hã].wav [ref:audio_uri] 来体会这一发音特点。”分析: *Dialectus* 的回答不仅准确, 而且展现了深度推理能力。它不仅指出了音变现象, 还链接到了知识图谱中的“尖团合流”规则, 并对“行”字的声母弱化现象进行了更细致的补充说明, 最后还提供了多模态证据, 充分展现了其作为专业语言学智能体的卓越性能。

第8章 总结与展望

本论文旨在应对汉语言方言学习，特别是音系学习领域长期存在的挑战——学习资源碎片化、教学过程非个性化、以及音韵知识抽象化。为解决这些问题，本项目设计并实现了一套完整的、以人工智能为核心的智能化方言学习与研究解决方案。本章将对整个研究工作进行系统性总结，着重阐述本项目的核心创新点与技术优势，并探讨当前研究存在的局限性与未来的发展方向。

8.1 研究工作总结

本项目遵循“**知识为基、推理为核、教学为用**”的设计哲学，构建了一个三位一体的智能化方言学习系统。整个研究工作可以概括为一个完整技术链路的构建：首先，本项目构建了一个大规模、覆盖汉语主要方言区的“**多模态方言音系知识图谱**”，它将分散、抽象的音韵知识转化为结构化的、可计算的形态，构成了整个系统的“**知识大脑**”。随后，本项目设计了“**文脉智能体**”（Dialectus）协作框架，通过精密的**多智能体协作推理机制**，赋予系统精准理解和回答复杂语言学问题的能力，扮演了“**逻辑中枢**”的角色。最终，为将系统的知识与推理能力落地于教学应用，本项目研发了“**自适应多步检索生成模型**”与“**音系自进化生成与校准模型**”，它们以动态的“**学习者状态**”为核心，实现了**自适应学习路径规划**，并将每一次用户互动转化为模型自我优化的契机，成为了系统的“**个性化导师**”。

8.2 主要创新点与优势

本项目的核心贡献体现在以下五个方面，它们共同构成了本系统在方言学习领域的技术优势：

1 创新点一：构建了大规模、多模态的方言音系知识图谱。

本项目的首要创新在于构建了一个大规模、覆盖汉语主要方言区的多模态知识图谱。它摒弃了传统语言资源库以文本为主的局限，首次将语言学本体论、计算声学特征参数与多媒体教学资产在统一的图结构中进行了形式化关联。这一优势在于，它使得复杂的音韵知识变得“可计算”、“可推理”，为所有上层AI应用提供了前所未有的、坚实、权威且可解释的知识基座。

2 创新点二：变方言学习为声学学习，开创了数据驱动的教学新范式。

本项目的根本性创新，在于通过技术手段，将方言音系学习从传统的、依赖模仿和抽象描述的模式，转变为一种以声学特征为核心的数据驱动式学习。通过将音高曲线、共振峰、声谱图等核心声学特征资产直接呈现给学习者，并与标准音进行可视化、可量化的对比，系统将“学习一个抽象的声音”这一模糊任务，转化为“匹配一个具体的声学范式”这一精准任务。这一优势是革命性的，它极大地降低了方言音系的学习门槛，提升了学习的直观性和有效性，使精准的发音校准成为可能。

3 创新点三：设计了面向异构语言知识的多智能体协作推理框架。

方言研究常需整合结构化的知识规则与非结构化的语境实例。为应对这一挑战，本项目创新性地设计了“文脉智能体”框架，将多智能体协作思想引入语言学研究。其优势在于，通过角色分工和链式通信，该框架能够模拟语言学家的工作流，对异构知识源进行深度整合与推理，从而解决传统单步 RAG 模型难以应对的复杂问题，并保证了推理过程的透明性与逻辑严谨性。

4 创新点四：实现了以学习者为中心的自适应学习路径规划。

传统教学系统提供“一刀切”的内容，忽略了个体差异。本项目的核心创新在于“自适应多步检索生成模型”所实现的自适应学习路径规划。该模型颠覆性地将优化目标从单纯的“信息完备性”转向了与“学习者认知状态”的双重优化。其优势在于，通过引入并动态更新“学习者状态向量” S_u ，模型能够真正实现“因材施教”，为不同水平和认知偏好的用户，动态规划并生成最高效、最易懂的学习路径与多模态教学内容。

5 创新点五：构建了基于真实用户互动的模型自进化闭环。

为克服 AI 教育模型迭代成本高、适应性差的难题，本项目创新性地设计了“音系自进化生成与校准模型”，构建了一个自动化的“交互-生成-过滤-学习”闭环。其优势在于，该机制使得模型能够在保障安全和质量的前提下，将每一次用户互动都转化为一次学习机会，通过高效的 LoRA 技术进行持续的自我微调。这使得模型的能力能够持续进化，在真实应用中变得越来越精准和智能，代表了向“终身学习型”教学 AI 迈出的重要一步。

8.3 研究局限性

尽管本项目在理论和实践上取得了一系列进展，但仍存在一些局限性：

知识图谱的深度与粒度：尽管本项目构建的知识图谱在覆盖的方言广度上达到了大规模，但在每一个方言点的知识深度和粒度上，仍有巨大的扩展空间。例如，对于许多方言的历史音韵演变、社会语言学变体等更深层次的知识，目前的覆盖尚不完备。

对底层大模型的依赖：本项目设计的上层框架和算法虽然具有通用性，但其最终能达到的性能天花板，在很大程度上受限于所采用的底层大语言模型（如讯飞星火）的基础能力，包括其语言理解、声学特征处理和逻辑推理的水平。

实验评估的局限性：由于任务的独创性，本项目采用了自建数据集进行评估。尽管设计严谨，但这些数据集的规模和多样性仍有限，其评估结果的普适性有待在更大规模、更真实的用户环境中进行检验。

8.4 未来工作展望

本项目为 AI 技术在方言保护与教学领域的应用开辟了新的可能性，未来可在以下几个方向上进行拓展：

知识图谱的半自动化构建：研究如何利用大语言模型、语音识别和计算机视觉等技术，从海量的方言文献、音视频资料中半自动地抽取实体和关系，以降低知识图谱的构建成本，加速其在知识深度和广度上的扩展。

更精细化的学习者建模：未来可探索更复杂的认知诊断模型，不仅建模用户的知识掌握度，更能追踪其学习过程中的注意力、遗忘曲线乃至认知负荷，实现更深层次的自适应。

从反应式辅导到主动式规划：未来的终极目标是构建一个能够为学习者主动规划完整学习路径的“AI 方言导师”，它可以根据用户的长期目标和初始水平，自动生成一套完整的、阶梯式的、多模态的学习课程。

跨模态生成能力的探索：未来可探索利用生成式 AI 技术，实现跨模态内容的直接生成，例如，根据用户的需求，合成一段带有特定情感色彩的标准发音，或为某个复杂的发音动作，实时生成一个可交互的 3D 口腔剖面动画。