

T/SAIAS²⁰²⁵

上海市人工智能行业协会团体标准

T/SAIAS 029—2025

人工智能 高等教育语料库建设导则

Artificial Intelligence—Guidelines for Construction of Higher Education Corpus

2025-07-24 发布

2025-07-25 实施

目 次

前 言 III

1 范围 1

2 规范性引用文件 1

3 术语和定义 1

4 高等教育语料库数据来源 1

 4.1 概述 1

 4.2 教学资源 1

 4.3 教学过程 2

 4.4 教学管理 2

 4.5 学术文献 3

5 高等教育语料库管理机制 3

 5.1 数据存储机制 3

 5.2 数据管理机制 3

 5.3 数据更新机制 3

6 高等教育语料库建设要求 3

 6.1 数据资源要求 3

 6.2 质量要求 3

 6.3 价值对齐 3

 6.4 数据安全 3

附录 A（资料性） 高等教育语料库中教学视频语料表达示例 4

附录 B（资料性） 高等教育语料库中教学视频语料存储示例 7

参 考 文 献 8

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这些专利的责任。

本文件由上海市人工智能行业协会提出并归口。

本文件起草单位：清华大学、上海库帕思科技有限公司、北京慕华信息科技有限公司、上海市教育考试院、上海交通大学、复旦大学、同济大学、上海市人工智能行业协会、上海华东电信研究院、上海阶跃星辰智能科技有限公司、华东理工大学、东华大学、上海开放大学、记忆张量（上海）科技有限公司、上海商汤智能科技有限公司、联通（上海）产业互联网有限公司、中国电信股份有限公司上海分公司、上海师范大学、上海戏剧学院、上海第二工业大学、上海师范大学天华学院、上海南湖职业技术学院、上海海事大学、联通数据智能有限公司、国创智造科技（上海）有限公司、国家工业信息安全发展研究中心人工智能所、中文在线集团股份有限公司、云南师范大学。

本文件主要起草人：山栋明、王帅国、黄海清、孔令和、杜婧、徐燊、施佳樑、王长冕、胡茂密、钟俊浩、周云、陈春丽、张峰、卫志华、肖仰华、张宁、张凯伦、王玮、李婧超、陆喆、李璟、朱燕民、靳文瑞、张立文、应方立、李灵光、孙雯、刘雨桐、陈巧慧、王娜、路少卿、唐波、沈雯、赵春昊、饶雪、常永波、孙丹诚、邓思文、郭汉杰、魏飞、贺仁驹、曹艳平、肖麒、刘东、折震霄、谢菲尔、金常飞、蒋思航、彭莉、张朕伟、王薇、包洋、陈紫娟、何美瑄、李建华、肖君、熊飞宇、沈滨、刘俊毅、姜淞译、罗勇军、李志宇、沈滨、堵炜炜、虞祝豪、朱莉琴、杨文恺、郑更河、华静、宋佳琪、王文东、袁非牛、沈鸣、王倩、黄志立、张博锋、陈方疏、凌旭峰、李忠唐、朱建柳、周一峰、于棋峰、高欣佳、王昊、胡立轩、陈亚乐、贺仁龙、郑茂宽、李卫、纪婷婷、谭晓坤、刘虹、陈紫娟。

本文件首次制定。

首期执行单位：清华大学、上海库帕思科技有限公司、北京慕华信息科技有限公司、上海市教育考试院、上海交通大学、复旦大学、同济大学、上海市人工智能行业协会、上海华东电信研究院、上海阶跃星辰智能科技有限公司、华东理工大学、东华大学、上海开放大学、记忆张量（上海）科技有限公司、上海商汤智能科技有限公司、联通（上海）产业互联网有限公司、中国电信股份有限公司上海分公司、上海师范大学、上海戏剧学院、上海第二工业大学、上海师范大学天华学院、上海南湖职业技术学院、上海海事大学、联通数据智能有限公司、国创智造科技（上海）有限公司、国家工业信息安全发展研究中心人工智能所、中文在线集团股份有限公司、云南师范大学。

本文件版权归上海市人工智能行业协会所有。未经许可，不得擅自复制、转载、抄袭、改编、汇编、翻译或将本标准用于其他任何商业目的。

人工智能 高等教育语料库建设导则

1 范围

本文件给出了高等教育语料库的数据来源，并规定了其管理机制和建设要求。
本文件适用于生产、使用语料的相关企业、机构或组织构建高等教育领域的语料库。

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 45574—2025 数据安全技术 敏感个人信息处理安全要求

GB/T 45652—2025 网络安全技术 生成式人工智能预训练和优化训练数据安全规范

T/SAIAS 015—2024 语料库建设导则

3 术语和定义

T/SAIAS 015—2024界定的以及下列术语和定义适用于本文件。

3.1

高等教育语料库 higher education corpus

由依据一定抽样方法收集的高等教育领域的语料所构成的电子数据库。

4 高等教育语料库数据来源

4.1 概述

高等教育语料库数据来源及范围包括教学资源、教学过程、教学管理和学术文献四类，其中：

- 教学资源包括各类大纲、教材、教案等主要由文字、图片组成的多模态信息数据；
- 教学过程包括教学过程中教师和学生的互动、交流等主要由文字、语音、视频等组成的多模态信息数据；
- 教学管理包括教学教研管理过程中涉及的各类统计数据、评估数据等主要由文字组成的信息数据；
- 学术文献包括各类文献著作等主要由文字、图片组成的多模态信息数据。

4.2 教学资源

4.2.1 教学大纲

高等教育语料库中的教学大纲数据包括但不限于：

- 课程基本信息包括课程名称、代码、性质、学分与学时、适用专业等信息；
- 课程目标包括知识目标、能力目标、素质目标等信息；
- 课程内容与教学要求包括课程内容大纲和教学要求等信息；
- 课程考核与评价包括考核方式、考核内容与标准等信息；
- 教材及参考资料包括课程所选用的主要教材的名称、作者、出版社、出版年份等详细信息。

4.2.2 教材

高等教育语料库中的教材数据包括但不限于：

- 基础知识包括基本概念与原理、学科发展历程等信息；
- 专业知识包括专业核心知识、前沿知识与拓展阅读知识等信息；
- 实践应用包括案例分析、实验与实践指导等信息；

- d) 辅助学习资料包括学习目标与要求、思考与练习题、参考文献与推荐读物等信息；
- e) 教材语料的其他资料还包括插图与图标、附录等信息。

4.2.3 教案

高等教育语料库中的教案数据包括但不限于：

- a) 教学基本信息包括课程名称、授课教师、授课对象、授课时间、授课地点等信息；
- b) 教学目标包括知识目标、能力目标、情感目标等信息；
- c) 情感目标涉及对学生情感态度和价值观的引导；
- d) 教学方法包括讲授、讨论、演示、练习、探究等过程信息；
- e) 教学过程包括导入新课、讲授新课、课堂练习、课堂小结、布置作业等过程信息；
- f) 教学资源包括教材、参考书籍、多媒体资源、网络资源等信息：
 - 教材信息包括教材名称、版本、作者等；
 - 参考书籍信息包括可供学生进一步学习和参考的其他书籍名称、作者、出版社等；
 - 多媒体资源包括教学课件、视频、音频、图片、动画等；
 - 网络资源包括与本节课内容相关的网站、在线课程、学术数据库等。
- g) 教学反思包括教学效果、教学方法、教学过程和改进措施等信息。

4.2.4 课程讲义

高等教育语料库中的课程讲义数据包括但不限于：

- a) 高等教育语料库中的课程讲义数据包括课程基本信息、课程概述、教学内容；
- b) 课程概述包括课程目标、课程性质、课程内容简介等信息；
- c) 教学内容包括章节主题、知识点讲解、案例分析、图标与数据等信息。

4.2.5 教学课件

高等教育语料库中的教学课件数据包括但不限于图表、图片、简洁的文字说明等。

4.2.6 课程作业

高等教育语料库中的课程作业数据包括但不限于教学的重点和难点，以及对学生知识掌握程度的要求等。

4.2.7 试题

高等教育语料库中的试题数据包括但不限于题目及其具有推理步骤的解题思维链数据。

4.3 教学过程

4.3.1 教学行为记录

高等教育语料库中的教学行为记录数据包括但不限于直播音视频、音视频讲稿、板书、课件等。

4.3.2 师生互动记录

高等教育语料库中的师生互动记录数据包括但不限于包括弹幕、投稿、答题等。其中答题信息包括客观题和主观题，需要包含题目和学生答题情况和教师评语等反馈学习成效的信息。

4.3.3 AI 教学过程数据

高等教育语料库中的AI教学过程数据包括但不限于教师与大模型交互的过程数据、学生与大模型交互的过程数据。

4.4 教学管理

4.4.1 教学数据分析

高等教育语料库中的教学数据分析数据包括但不限于教学管理者、教师、学生等主体的数据分析结果。

4.4.2 教学过程评估

高等教育语料库中的教学过程评估数据包括但不限于教学质量、学生反馈、多维评价信息等。

4.5 学术文献

高等教育语料库中的学术文献数据包括但不限于期刊论文、学位论文其他学术著作等。

5 高等教育语料库管理机制

5.1 数据存储机制

a) 文件系统存储：将语料以文本文件的形式存储在文件系统中，可以按照不同的类别、格式等进行分类存储。

b) 数据库存储：应根据语料数据特性选择相应存储数据方式。

注1：如关系型数据库适合存储结构化的语料信息，如标注后的语料，能方便地进行数据查询和管理；非关系型数据库适合处理半结构化或非结构化的语料，如原始的文本、音频、视频等；图存储方式适合具有复杂关系、深度连接和多跳查询需求的数据（如思维链数据等）。

注2：高等教育语料库中教学视频语料表达示例和存储示例可参考附录A和附录B。

5.2 数据管理机制

a) 元数据管理：对话料的元数据进行管理，包括语料的来源、采集时间、学科标签、文本类型、语言种类等信息。通过建立元数据索引，可以方便地对语料进行分类、检索和筛选。

b) 用户权限管理：根据用户的不同角色和需求，设置不同的访问权限，宜通过用户权限管理保证语料库的安全性和数据的完整性。

注：如管理员具有对话料库的完全控制权，包括数据的添加、删除、修改等操作；普通用户可能只有读取语料的权限。

5.3 数据更新机制

宜按照课程学年的时间周期定期更新数据。

注：如春/秋季学期或每年，对话料库进行更新，以保证语料库的时效性和丰富性。

6 高等教育语料建设要求

6.1 数据资源要求

高等教育语料的数据要求应符合T/SAIAS 015 第5章的规定。

6.2 质量要求

高等教育语料的质量要求应符合T/SAIAS 015 6.3.1的规定。

6.3 价值对齐

高等教育语料的价值对齐应符合T/SAIAS 015 第9章的规定。

6.4 数据安全

高等教育语料库的数据安全应符合T/SAIAS 015 第10章的规定。

高等教育语料库的敏感个人信息处理应符合GB/T 45574—2025的规定。

高等教育语料库的数据处理活动的安全要求应符合GB/T 45652—2025第5章、第6章规定。

附录 A
(资料性)
高等教育语料库中教学视频语料表达示例

A.1 基本数据信息

围绕教学视频的语料信息包括课程信息、视频信息、字幕信息、课件信息和题目信息，见表A.1。

表 A.1 高等教育语料数据类别

语料信息类别	用途
题目信息	课程标题/题目
课程数据信息	课程相关的系统性描述记录
视频信息	与课程相关的视频信息的记录
字幕信息	视频关联字幕
课件信息	与课程相关的课件信息（如 PPT、文字笔记、图片笔记等）
试题信息	与课程相关的试题信息

A.2 语料表达

A.2.1 课程数据信息

课程数据信息应包括课程名称、开课时间、结课时间、开课机构、教师团队、课程介绍、教学时长、学习投入、参考教材、详细介绍、课程ID、章节信息，见表A.2。

表 A.2 课程数据信息表达

基本信息	字段名	说明
课程名称	course_name	课程资源包名称
开课时间	start_time	课程的开始时间
结课时间	end_time	课程的结课时间
开课机构	organizations	课程开设的学校/单位
教师团队	teachers	课程主讲介绍
课程介绍	short_intro	简要介绍课程内容
教学时长	course_duration	实际上课的时长
学习投入	learning_duration	学生需要投入的时间
参考教材	reference_materials	需要哪些课程教材作为参考
详细介绍	long_intro	关于课程的详细介绍
课程 ID	course_id	课程对应的唯一 ID
章节信息	resource_tree	课程里的章节信息，包括章节 ID、标题、序号

A.2.2 视频信息

视频基本信息应包括章节信息、单元信息、授课老师、视频时长、视频大小、分辨率、视频ID等，见表A.3。

表 A.3 视频信息表达

基本信息	字段名	说明
章 ID	chapter_id	视频对应的章 ID
章标题	chapter_title	视频所在章的标题
章序号	chapter_order	视频所在章在所有章里的序号
节 ID	section_id	视频对应的节 ID
节标题	section_title	视频所在节的标题
节序号	section_order	视频所在节在当前章里的节序号
单元 ID	leaf_id	视频对应的学习单元 ID
单元标题	leaf_title	视频所在学习单元的标题
单元序号	leaf_order	视频所在学习单元在当前节里的单元序号
授课老师	teacher_name	授课老师姓名
视频时长	duration	视频实际时长
视频大小	size	视频文件大小
视频格式	format	视频文件的格式，例如:mp4
分辨率	resolution	视频分辨率，例如：720P、1080P 等
视频 ID	video_id	视频在资源的唯一标识
视频地址	url	视频的播放地址链接

A.2.3 字幕信息

字幕信息应包括视频ID、文件名称、大小、格式等，见表4。

表 A.4 字幕信息表达

基本信息	字段名	说明
视频 ID	video_id	字幕文件所关联视频的 ID
文件路径	file_path	字幕文件路径
大小	file_size	文件大小
格式	format	例如：srt

A.2.4 课件信息

课件信息应包括单元信息、文件名称、大小、格式、页码、课件截图list、标注、备注、课件ID等，见表A.5。

表 A.5 课件信息表达

名称	字段名	说明
课件 ID	slides_id	课件在资源的唯一标识
文件路径	file_path	课件文件路径
大小	file_size	文件大小
格式	format	例如：pptx 等

A.2.4.1 试题信息

试题信息应包括题目ID、题型、题干、答案、解析、难度等，见表A.6。

表 A.6 试题信息表达

名称	字段名	说明
题目 ID	problem_id	题目资源唯一编码
题型	type	包括：单选题、多选题、填空题、判断题、投票题、主观题（别名）
题干	body	包括：题干、选项
答案	answer	题目的正确答案
解析	explanation	对题目和答案的整体分析
难度	difficulty	题目难度

A.2.4.2 单页课件信息

单页课件信息应包括课件ID、课件页码、视频ID、章ID、章标题、章序号、节ID、节标题、节序号、单元ID、单元标题、单元序号、开始时间、结束时间、讲稿、课件内容、单页课件截图、页面类型等，见表A.7。

表 A.7 单页课件信息表达

名称	字段名	说明
课件 ID	slides_id	在资源的唯一标识
课件页码	slide_index	单页课件的页码
视频 ID	video_id	课件所对应视频的 ID
章 ID	chapter_id	课件对应的章 ID
章标题	chapter_title	课件所在章的标题
章序号	chapter_order	课件所在章在所有章里的序号
节 ID	section_id	课件对应的节 ID
节标题	section_title	课件所在节的标题
节序号	section_order	课件所在节在当前章里的节序号
单元 ID	leaf_id	课件对应的学习单元 ID
单元标题	leaf_title	课件所在学习单元的标题
单元序号	leaf_order	课件所在学习单元在当前节里的单元序号
开始时间	start_time	单页课件对应视频的开始时间
结束时间	end_time	单页课件对应视频的结束时间
讲稿	lecture	单页课件对应时间内视频字幕的内容
课件内容	content	单页课件的内容，包括标题，内容（文字、表格、图片、公式），重点标注、页面布局
单页课件截图	snapshot	单页课件的截图
页面类型	type	单页课件的类型：标题页、目录页、内容页、致谢页

附录 B
(资料性)
高等教育语料库中教学视频语料存储示例

B.1 教学视频语料的数据关联形式

教学视频语料的存储形式由时间轴及一系列对应时间轴的数据片段有序组合而成，见图B.1。

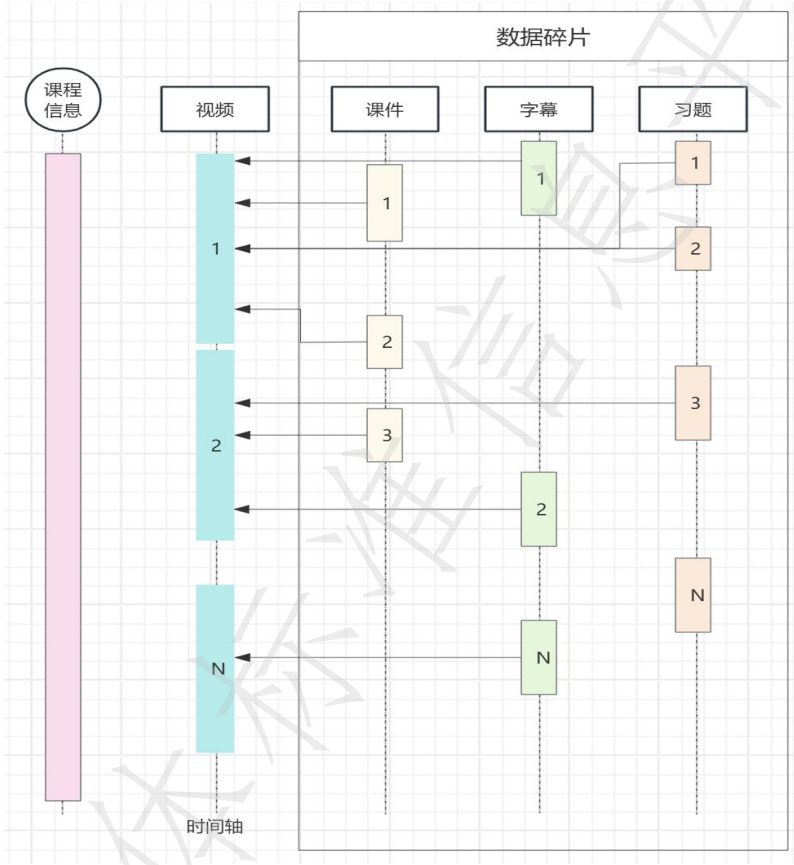


图 B.1 教学视频语料数据关联形式

B.2 教学视频语料数据集存储

多个B.1所述的教学视频语料数据存储在一起形成语料数据集，见图B.2。



图 B.2 视频语料库数据集存储方式

参 考 文 献

- [1] 中国分类主题词表（第二版）
 - [2] 中国图书馆分类法
 - [3] 中国档案分类法
 - [4] 普通高等学校本科专业目录（2024年）
-