WORKING MANUAL · CHINA AI POLICY DATABASE

中国政府人工智能政策库建设工作指南

检索、下载、梳理、入库、核验的全流程手册
供智能体按步骤执行、供人工逐项核验

L1 中央	**66 个机关单元**
L2 省级	**31 省＋兵团**
L3 第三层	**333 地级＋86 区县＋32 特殊**
时间范围	**不设起始年份**

v1.0（草案，待确认） · 编制日期 2026-11-14
配套表单：《中国政府人工智能政策库建设-工作表单.xlsx》（29 张工作表）
项目文件夹：人工智能的政策分析

这本手册怎么用

手册把政策库建设拆成名录、检索、下载存证、梳理、入库、核验六段工作，每段写清楚输入、步骤、检查点和输出。智能体照着做，人工照着查，负责人照着验收。

三类读者，三条阅读路径

读者	先读	执行时对照
负责人	第 1 章旧库诊断、第 2 章、第 9.6—9.9 节（执行主体）、第 12 章执行计划	表单 25 验收门、27 进度看板、28 执行环境检查
执行主体（Codex 及其操作者）	第 3—9 章，附录 A 提示词、附录 B 建库语句；执行包中的 AGENTS.md 与 rules.json	表单 02—06 名录、08 词表、09 路由、14 字段字典、24 自检规则、26 提示词、28 执行环境检查
核验员	第 3 章边界、第 10 章核验与抽检（特别是 10.7）	表单 11 相关性、13 判例、19 抽检方案、20 抽检记录、21 问题单

和以往文件的关系

项目文件夹里已有的《AI政务服务央地政策分工–研究架构与操作指南》第 6 章和《方案乙–地级检索执行清单_197单元》是上一轮的采集方案：只覆盖 7 个省和 4 个直辖市的第三层，检索词限定在“政务、政务服务”上，时间限定在 2017 年以后。本手册替代那部分内容，范围扩大到全部层级、全部领域、不设起始年份。197 个单元的清单仍然有效，它们是本手册第三层 451 个单元的子集，已做的检索记录可以直接迁入新表单。

研究问题（扩散模式与分工模式）怎么提、编码表怎么定，不在本手册范围内。本手册只负责交付一个边界清楚、证据可查、完整性可以估计的政策文本库，并把条款切分好，供后续“政策工具挖掘”直接使用。

手册里的几个约定

- “检索单元”是检索、记录和验收的最小单位：中央按机关划分，省级和第三层按政府划分。
- “主库”是研究分析的分母，只收 A、B 级政策文件；辅库、线索库、登记表各有用途，但不计入主库数量。
- 凡标“待核”的网址、文号和日期，都要在开工前（验收门 G0）由执行者打开官方来源核实。
- 多智能体工作由 Codex 执行，Claude 只承担独立复核等少量工作（第 9.6 节）。执行时以执行包为准，对话记录只作背景。
- 表单中的浅黄色单元格是执行时要填写的列。

目录

第 1 章　旧库诊断：673 个文件哪里不对 …… 5
1.1　覆盖：分布由检索进度决定，不由政策实际决定 …… 5
1.2　内容：主题被“政务服务”框住了 …… 6
1.3　形式和种类：文本类型混杂，残缺件不少 …… 6
1.4　可追溯性：文件和清单对不上账 …… 7
1.5　时间：2017 年以前基本没有 …… 7

第 2 章　目标、原则与总体流程 …… 8
2.1　建成什么样的库 …… 8
2.2　六条工作原则 …… 8
2.3　总体流程 …… 8
2.4　与朱旭峰“NLP—LLM 混合协同”工作流的对照 …… 9

第 3 章　边界：什么进库 …… 11
3.1　什么算“人工智能政策”：相关性四级 …… 11
3.2　邻近领域怎么处理 …… 12
3.3　什么算“政策文件”：文种与去向 …… 12
3.4　算一份还是几份 …… 13
3.5　时间边界 …… 13
3.6　判例 …… 14

第 4 章　谁的政策：层级、单元与发文主体 …… 15
4.1　四个层级 …… 15
4.2　检索单元总表 …… 15
4.3　每个单元要查哪些机关 …… 16
4.4　特殊情形 …… 17

第 5 章　检索：怎么找全 …… 18
5.1　来源矩阵 …… 18
5.2　检索词表 …… 18
5.3　检索路由 …… 19
5.4　单元检索步骤（A2 执行） …… 20
5.5　停止规则与“0 份”结论 …… 20
5.6　检索日志要记什么 …… 21
5.7　站点访问规范与人机交接 …… 21

第 6 章　下载与存证 …… 22
6.1　原件优先级 …… 22
6.2　一份文件要存什么 …… 22
6.3　命名与目录 …… 22
6.4　完整性检查 …… 23
6.5　格式转换与 OCR …… 23

第 7 章　梳理：元数据、研判、去重与关联 …… 25
7.1　元数据从哪里取 …… 25
7.2　关键字段的规则 …… 25
7.3　相关性研判 …… 25
7.4　去重与归并 …… 26
7.5　关系 …… 26
7.6　条款切分（为工具挖掘做准备） …… 27

第 8 章　数据库 28
8.1　表结构 28
8.2　编号规则 29
8.3　版本、冻结与增量 29
第 9 章　智能体作业规程 30
9.1　角色分工 30
9.2　工作包与并发 30
9.3　状态机 31
9.4　各智能体作业卡 31
9.5　自检规则 33
9.6　执行主体换成 Codex：哪些要跟着变 34
9.7　执行包 34
9.8　脚本做什么，模型做什么 35
9.9　交接与续跑 35
第 10 章　人工核验与抽检 37
10.1　三道关 37
10.2　抽样方案 37
10.3　批次判定 38
10.4　召回评估：怎么知道漏了多少 38
10.5　一致性与仲裁 38
10.6　返工与问题单 39
10.7　核验员操作要点 39
第 11 章　旧库迁移 40
11.1　建议去向汇总 40
11.2　迁移步骤 40
第 12 章　执行计划 42
12.1　阶段与验收门 42
12.2　试点设计 42
12.3　人员与工作量 43
12.4　主要风险与应对 43
12.5　交付物清单 43
附录 A　提示词模板 45
附录 B　SQLite 建表语句 48
附录 C　字段字典 49
附录 D　配套表单说明 52

第 1 章

旧库诊断：673 个文件哪里不对

在动手重建之前，先把现有的 673 个文件逐个读了一遍（抽取全文、计算哈希、按标题和正文特征分类）。下面的数字都来自这次机器诊断，逐文件结果在表单“23 旧库诊断与迁移”。机器判断会有误差，但问题的方向是清楚的。

13	129	84	21
中央文件，最早一份是 2022 年	直辖市区县文件被归在“地级”	份正文里没有任何人工智能核心词	份是“通知壳”，附件正文缺失

1.1 覆盖：分布由检索进度决定，不由政策实际决定

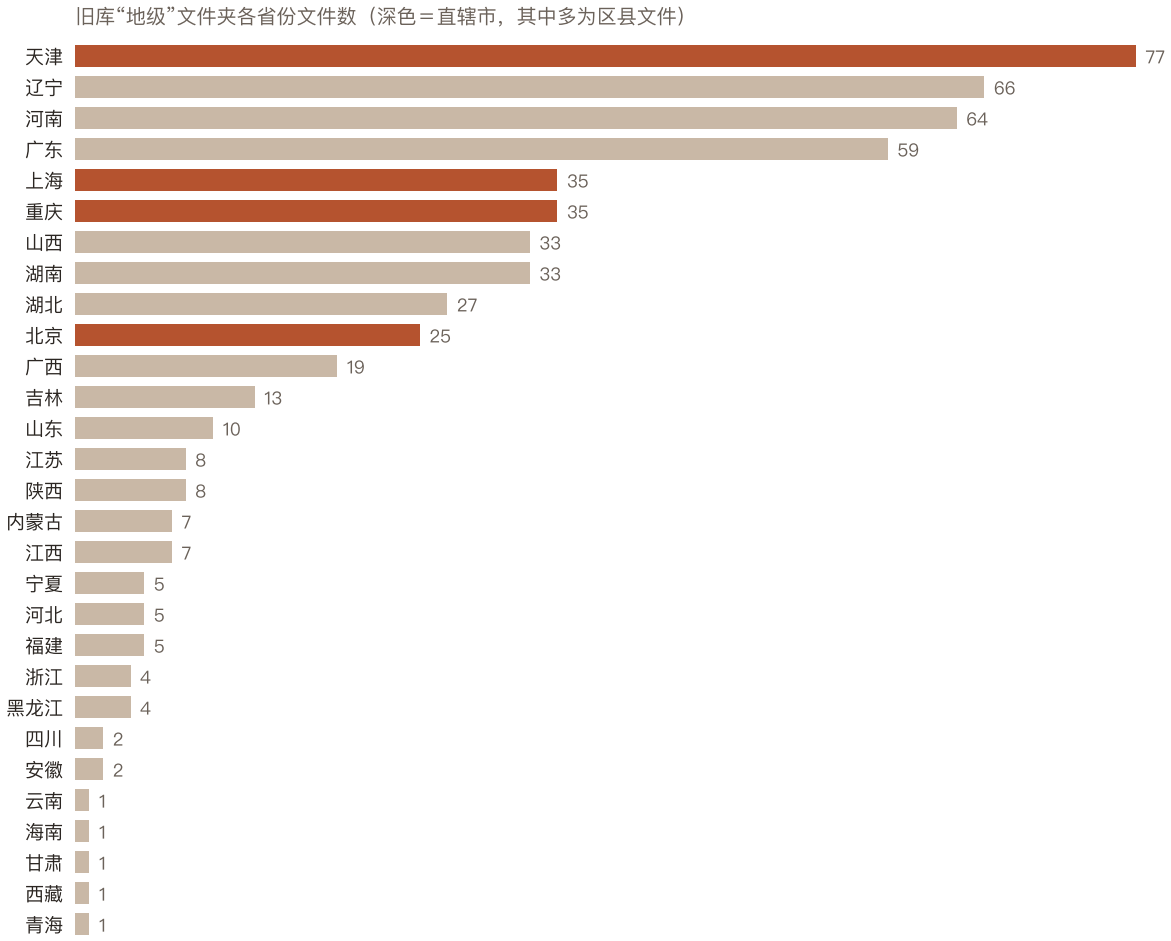


图 1　旧库第三层文件的省份分布：多少取决于检索进度，而非发文实际

- **中央层几乎空白。**只有 13 份，全部是 2022 年以后的文件，而且集中在场景、药监、医疗等少数主题。《新一代人工智能发展规划》（2017）、算法推荐、深度合成、生成式人工智能等管理规定，一份都没有。中央是央地分工研究的起点，这一层缺失，后面的纵向比较无从谈起。
- **省级只有 102 份，覆盖 27 个省份。**西藏、青海、宁夏、新疆没有省级文件，多数省份只有 1—5 份。

- **第三层分布极不均衡。**按文件名匹配，地级文件只落在大约 121 个地级单元上，其余 200 多个单元一份都没有；天津 77 份、辽宁 66 份、河南 64 份，而云南、甘肃、海南、青海、西藏各 1 份。这反映的是谁检索得多，不是哪里发文多。
- **层级混淆。**“地级”文件夹里的北京、天津、上海、重庆文件共 172 份，其中约 129 份是区县政府或区县部门的文件，被当作地级处理。直辖市的区县可以作为第三层收录（见第 4 章），但必须单独标记。

1.2 内容：主题被“政务服务”框住了

上一轮的检索式是“人工智能类词 AND 政务类词”，结果收进来的是一批数字政府、政务云、12345 热线、免申即享的文件，其中很多与人工智能关系不大；而产业发展、科技研发、教育、医疗、算法治理这些人工智能政策的主体部分，大量没有收进来。按正文里人工智能核心词的出现情况做机器预判：

机器预判	判据	份数	占比	按新口径大致去向
A?	标题含人工智能、大模型、智能等词	125	19%	多数可进主库，需核对
B?	标题不含，正文核心词≥10 次	112	17%	部分为 B 级，部分只是反复列举
C?	正文核心词 1—9 次	333	49%	多为 C 级，只登记
X?	正文没有任何核心词	84	12%	排除
待 OCR	无文字层，无法判断	19	3%	OCR 后再判

也就是说，按本手册的口径，旧库里可能只有三分之一左右能进主库。剩下的不是没用，而是应该进登记表、辅库或排除，并且需要换一种检索方式把真正缺的文件补回来。

1.3 形式和种类：文本类型混杂，残缺件不少

问题	份数	说明
政府工作报告	14	不是政策文件，应进辅库
年度报告、情况报告、工作总结	27	法治政府建设报告、信息公开年报等，应排除
新闻、动态、解读、会议报道、提案答复	27	机器初判，应进线索库并反查原文
通知壳（附件缺失）	21	只有印发通知，政策正文没下下来；北大法宝件显示“附件预览 无相关内容”的有 16 份
无文字层	19	扫描件或图片，无法检索和统计
文件名截断	68	北大法宝导出件，从文件名看不出是哪份
完全重复 / 疑似同题重复	2 组 / 3 组	同一文件多次下载

格式方面，PDF 404 个、DOC 155 个、DOCX 113 个、WPS 1 个，没有统一的正文文本；北大法宝导出件 222 个，其中 68 个文件名被截断（“关于印发《东莞市数字…”），无法从文件名判断是哪份文件。

1.4 可追溯性：文件和清单对不上账

- 除北大法宝件自带的法宝链接外，其余文件都没有记录官方来源网址和获取时间，无法回到原处核对，也无法证明文件是官方版本。
- 汇总表是 652 条，口头说 653 份，文件夹里是 673 个文件，三者没有逐条对应关系。
- 没有检索记录。某个城市没有文件，无法区分是“没发”还是“没查”。
- 有 1 份文件放错省份（白银市的文件在“宁夏”文件夹），2 组完全相同的重复文件、3 组疑似同题重复。

1.5 时间：2017 年以前基本没有

从正文能识别出日期的文件中，2017 年以前的不到 10 份（其中还有正文里引用的旧日期），2020 年以前的约 40 份。一方面是检索时限定了 2017 年起，另一方面是早年的地方文件很多已从网站下线，只靠门户检索找不到，需要政府公报目录和法规数据库。

> **结论**
>
> 不在旧库上修补，按本手册的流程从名录开始重建。旧库有两个用处：一是作为“种子”，迁移时逐个核对、补全来源；二是作为召回测试的已知项，检验新流程能不能把它们找回来（第 10.4 节、第 11 章）。

第 2 章

目标、原则与总体流程

2.1 建成什么样的库

1. **全层级、全领域。**中央、省级、第三层（地级市及同位置单元）全部检索；人工智能的产业、科技、应用、治理、人才、基础设施各领域都收，用标签切出政务服务等子集。
2. **每一条都能回到官方原件。**记录官方链接、获取时间、原件、网页快照和哈希值。
3. **完整性可以估计。**每个检索单元都有检索日志，0 份也有记录；用已知项回收和双路独立检索估计召回率，而不是只说“尽量找全”。
4. **可复现、可增量。**口径、词表、提示词、模型版本都有版本号；冻结后按月增量，不推倒重来。

2.2 六条工作原则

原则	含义与做法
先名录，后检索	先列出全部检索单元和每个单元的机关、站点，再开始检索。分布必须由名录决定，不能由谁先查到什么决定。
没有记录就等于没做	每一次检索都登记，包括 0 命中；每一个判定都附原文证据。
官方原件优先	商业数据库和搜索引擎用来查漏，下载以官网原件为准；北大法宝件必须检查附件。
边界写成规则和判例	“算不算人工智能政策”“算不算政策文件”“算谁的”都有成文规则；遇到新情形由负责人裁决并追加判例。
机器做量，人做判断	智能体负责检索、下载、抽取、初判和自检；口径裁决、待定、0 份复核、抽检由人完成。分类框架由研究者定，模型只执行。
按批验收	每个省、中央、试点各为一批；批次抽检不合格整批返工，不做“差不多就行”的放行。

2.3 总体流程

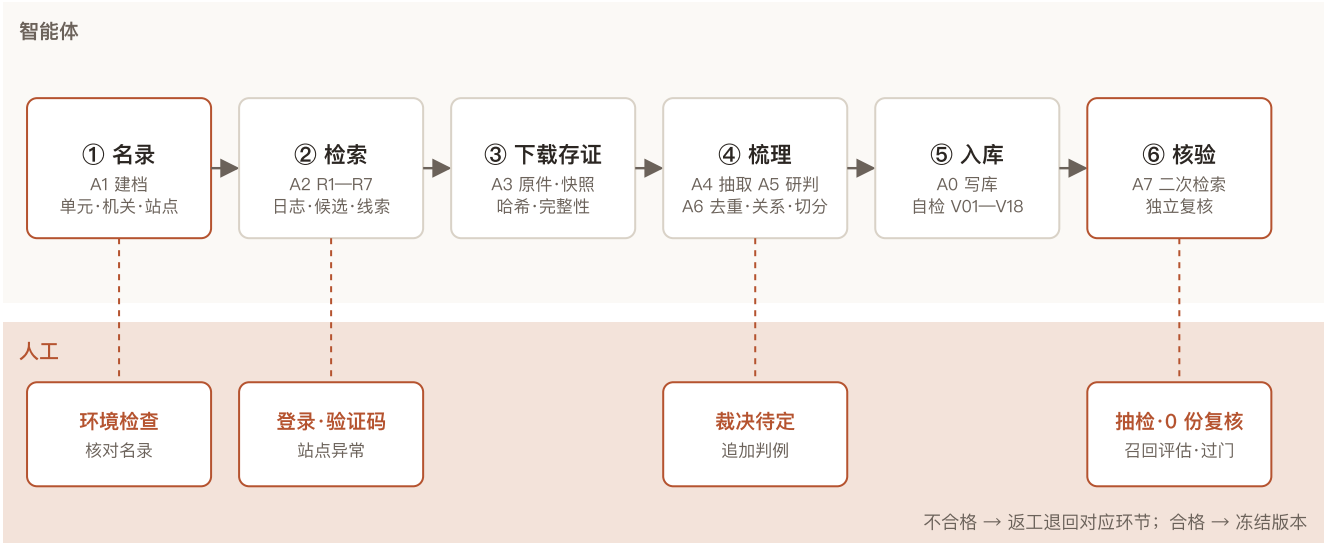


图 2　总体流程：六段工作，人工在四处介入

流程分六段。①名录：建立检索单元、机关和站点清单。②检索：每个单元按 R1—R7 路由检索，登记日志，产出候选。③下载存证：取官方原件、附件和快照，做完整性检查。④梳理：抽取元数据、判定相关性和去向、去重、识别关系、切分条款。⑤入库：写入 SQLite，跑自检规则。⑥核验：二次独立检索、抽检、召回估计，过验收门后冻结。人工在四处介入：开工前完成执行环境检查、核对名录；检索、下载中处理登录、验证码和站点异常；研判中裁决待定并追加判例；核验阶段做抽检、0 份复核和召回评估，签字过门。

2.4 与朱旭峰“NLP—LLM 混合协同”工作流的对照

朱旭峰团队的新能源汽车政策工具研究，把工作流分成 4 个模块、12 个步骤，前三步（多源采集、原始记录建库、文件筛选）对应本手册的范围；后面的分块、召回、抽取、分类、面板构建属于工具挖掘阶段。下表列出本手册沿用了什么、在哪些地方补强。

环节	朱旭峰工作流（报告所述）	本手册的做法
采集来源	北大法宝、地方政府网站、政务微信公众号多源采集	沿用；补充部门网站、政府公报、法律法规数据库、上级文件反查。搜索引擎和新媒体只作线索，下载以官网原件为准
采集组织	JSON 记录组织、SQLite 存储	沿用 SQLite；以检索单元为组织单位（549 个），每次检索留日志，0 份也有记录，分布由名录决定
文件筛选	文种类型纳入标准；政策类别与标题相关性判定	文种分八个去向；相关性分 A/B/C/X 四级，以“是否有以人工智能为对象的独立任务”判定，必须附原文证据
元数据与去重	补齐主体、层级、地区、年份；同一文件多来源去重	40 余个字段的字典；来源冲突的取值顺序；三级去重；八类文件关系（转发、依据等直接服务扩散研究）
证据链	报告未提及	官方链接、获取时间、原件、网页快照、SHA-256
质量控制	专家校准提示词；多模型交叉复核；小样本人工金标准，算准确率、F1	沿用；增加 18 条程序化自检、盲填式独立复核、按批抽检与接受规则、κ 一致性门槛

环节	朱旭峰工作流（报告所述）	本手册的做法
完整性评估	报告未提及召回评估	已知项回收；双路独立检索的捕获—再捕获估计
时间与版本	固定研究窗口（2009—2023）	不设起始年份；版本冻结＋按月增量；口径、提示词、模型版本可追溯
衔接下一阶段	Step 3 按公文结构与语义约束分块	条款切分在本阶段完成，直接交给工具挖掘

朱旭峰工作流内容依据项目文件夹中《朱旭峰–数智政策科学–报告学习手册》第 06—07 节、第 13—14 节（方法卡 A、B）。

第 3 章

边界：什么进库

一份文本能否进主库，要过三道判断：它是不是人工智能政策（相关性）、它是不是政策文件（文种与去向）、它算一份还是几份（计数单位）。三道判断都写成规则，规则覆盖不了的情形进判例。

3.1 什么算“人工智能政策”：相关性四级

等级	名称	定义	处理	例子
A	主题型	文件以人工智能或其具体技术形态为对象：标题含核心词；或正文三分之一以上篇幅针对人工智能；或为人工智能专门设立的规划、条例、方案、清单、征集通知、名单。	全文收集，进主库	《××市推动人工智能产业发展三年行动计划》；《××省“人工智能+”行动实施方案》；《生成式人工智能服务管理暂行办法》
B	专项型	文件主题不是人工智能，但含有以人工智能为对象的独立章、节、条或段落；或设置了以人工智能为对象的具体任务、项目、指标、资金、责任单位。	全文收集，进主库	数字政府规划中设“推进人工智能大模型政务应用”一节；医疗卫生规划中有一条专讲人工智能辅助诊断
C	提及型	人工智能只出现在列举或修饰中（如“运用大数据、人工智能等技术手段”），没有以它为对象的独立任务。	只登记元数据、链接和命中句，不下载原件，不进主库	营商环境方案中一句“推广人工智能等新技术应用”
X	不相关	没有核心词；或“智能”“智慧”只指一般信息化，未涉及人工智能技术。	排除，保留检索记录	只讲政务云、电子证照的管理办法；智慧城市规划中没有人工智能内容

判定规则

1. 判定以文件正文为准，不看新闻稿、解读和标题党摘要。
2. B 级的关键是“以人工智能为对象的独立任务”，不是出现次数。机器统计的命中次数只用于排序和初筛。
3. 印发通知加附件视为一份文件，按附件内容判定。
4. 邻近领域（智能制造、智能网联汽车、机器人、算力、数据要素、智慧城市、数字政府）文件：正文出现核心词且满足 B 级条件的才进主库；否则按 C 或 X 处理，在“邻近领域”字段标明。
5. 拿不准时标“待定”，写明理由，交人工裁决；不得为了进度直接排除。
6. 研判必须给出证据：逐字摘录 1—3 处原文句子及其所在章节或条款编号。没有证据的 A、B 判定视为无效。

> **为什么不按出现次数划线**
>
> 一份 3 万字的数字政府规划可能出现 20 次“人工智能”，全是列举；一份 2 千字的征集通知可能只出现 6 次，却是专门为人工智能场景发的。次数只能用于机器排序和初筛（旧库诊断里就是这样用的），等级必须看“有没有以人工智能为对象的独立任务”。

3.2 邻近领域怎么处理

智能制造、智能网联汽车、机器人、算力、数据要素、智慧城市、数字政府这几类文件，和人工智能边界最模糊。处理办法是：检索时用 K2 词把它们找出来，研判时看正文有没有 K1 核心词，以及有没有以人工智能为对象的独立任务。

情形	判定	说明
正文有 K1 词且有独立任务	B 级，进主库，填“邻近领域”字段	如算力行动计划里有“建设智算中心支撑大模型训练”一节
正文有 K1 词但只是列举	C 级，登记	
正文没有 K1 词	X 级，排除，填“邻近领域”字段	检索日志保留，方便以后扩大口径时回查
技术上属于人工智能应用、但文中不用人工智能词（如“自动驾驶系统”）	待定，人工裁决	负责人裁决后写入判例；试点阶段必须对自动驾驶、人脸识别两类先定规则

3.3 什么算“政策文件”：文种与去向

旧库的第二类问题是文本类型混杂。本手册把检索到的文本分到八个去向，只有“主库”三类计入研究分母。

类别	典型形式	去向	说明
法律法规规章	法律、行政法规、地方性法规、自治条例、部门规章、地方政府规章、人大常委会决定	主库	含修正、修订版本；废止信息记入时效字段
党内法规与党委文件	党委、党委办公厅（室）印发的意见、方案、规定，党委与政府联合发文	主库	党委单独发文或与政府联合发文都收
规范性文件与政策文件	意见、实施意见、通知（印发型）、决定、规划、纲要、方案、行动计划、工作要点、办法、细则、指南、指引、若干措施、若干政策、清单、目录	主库	最常见的一类
实施类文件	征集通知、申报指南、评选/认定通知、名单公示、试点批复、场景清单发布、项目资金指南	主库（实施类）	研究场景扩散的重要证据，不要当噪声丢掉
转发文件	“关于转发××的通知”	主库（转发）	单独成条，并与被转发文件建立“转发”关系；转发本身就是纵向扩散的记录
标准与技术规范	国家标准、行业标准、地方标准（政府部门发布或批准）	辅库·标准	团体标准、企业标准不收
政府工作报告	各级政府工作报告	辅库·政府工作报告	v1.0 只收提到人工智能的报告；全量建库作为后续独立任务
司法文件	司法解释、指导意见、典型案例发布通知	辅库·司法	法院、检察院文件不进主库
解读与问答	政策解读、答记者问、图解	线索库（挂到原文件）	不独立计数；与原文件建立“解读”关系
新闻与动态	新闻报道、工作动态、会议报道、领导讲话、媒体文章	线索库	只用来反查官方原文

类别	典型形式	去向	说明
报告与总结	年度工作报告、工作总结、法治政府建设报告、政府信息公开年度报告、统计公报	排除	不是政策文件；旧库中有 27 份
答复与内部文书	建议提案答复、请示、批复（未公开政策内容的）、会议纪要	排除（可入线索库）	个别答复里有政策线索，可登记到线索库
采购与招标	采购意向、招标公告、中标公告	排除	对研究政策落地有用，但属于另一套数据，不混入政策库

法定公文文种 15 种（《党政机关公文处理工作条例》）：决议、决定、命令（令）、公报、公告、通告、意见、通知、通报、报告、请示、批复、议案、函、纪要。“文种”字段按文件自称填写，“政策形式”字段按标题末尾的形式词填写，两者分开记。

3.4 算一份还是几份

情形	计数	做法
印发通知＋附件	一份	标题、政策形式按附件记，通知标题记入 title_notice；原件和附件挂在同一 doc_id 下
同一文件多个来源（官网、公报、法宝）	一份	同一 dup_group，source 表记多条来源，主原件取优先级最高的
转发上级文件	单独一份（主库–转发）	与被转发文件建“转发”关系；转发说明本级做出了响应，是扩散证据
修订版、新年度版	各算一份	建“修订”关系；年度工作要点逐年各算一份
解读、图解、问答	不计数	进线索库，挂到原文件
征集通知与入选名单	各算一份（实施类）	建“名单”关系
联合发文	一份	归牵头机关所在单元；全部机关记入 issuers；跨层级联合发文（如省与市）归高一层
配套细则、任务分工方案	各算一份	建“配套”关系

3.5 时间边界

- **不设起始年份。**所有检索都选“全部时间”。2015 年以前的命中照常登记，另加“前史”标记，研究时可单独处理。我国较早的人工智能相关政策表述包括 2015 年“互联网＋”行动中的“互联网＋人工智能”、2016 年的“互联网＋”人工智能三年行动实施方案；更早的科技计划里有“智能计算机”“模式识别”等提法，用 K3 历史词去找。
- **冻结日。**v1.0 按发布日期截止到冻结日（默认 2026–12–31）。冻结日之后发布的文件进入下一版。
- **增量。**冻结后每月对每个单元跑一次增量检索（只查上次检索日以后的发布），新文件走同样流程，每季度发布一个小版本。
- **时效。**已废止、已修订的文件照样收录，用 validity 字段标明；研究政策存量时按时效筛选。

3.6 判例

以下 20 条判例覆盖了试点前能预见的边界情形，A5 研判智能体把它们作为提示词的一部分。试点中出现的新情形由负责人裁决后追加（表单“13 判例”）。

编号	情形	判定	理由
J01	某市政府工作报告写“推进人工智能+政务服务”	辅库·政府工作报告	不是政策文件，只进辅库
J02	某省“人工智能+”行动实施方案	A，主库	以人工智能为对象
J03	某市数字政府“十四五”规划，设“推动人工智能等新技术应用”一节，下有 3 条具体任务	B，主库	独立一节且有具体任务
J04	某市优化营商环境若干措施，只有一句“运用人工智能等技术提升审批效率”	C，登记	仅修饰，无独立任务
J05	某市政务云管理办法，全文无人工智能词	X，排除	无核心词；旧库中此类有 84 份（机器预判）
J06	某市“智慧城市”建设方案，正文有“城市大脑”但无人工智能词	X，排除；邻近领域=智慧城市	K2 命中但无 K1 词
J07	某市智能网联汽车道路测试管理办法，正文多处写“自动驾驶系统”但无人工智能词	待定→人工	自动驾驶是人工智能应用，但是否把邻近领域整体纳入需负责人决定；裁决后写入判例
J08	某省算力基础设施行动计划，有一节“建设智算中心支撑大模型训练”	B，主库	有以人工智能为对象的独立任务
J09	某市关于征集人工智能应用场景的通知	A，主库–实施类	征集通知是实施类文件
J10	某市人工智能应用场景入选名单公示	A，主库–实施类；与征集通知建立“名单”关系	
J11	某市转发省人工智能行动方案的通知（无本地内容）	A，主库–转发；建立“转发”关系	转发本身是扩散记录
J12	某市人工智能行动计划的政策解读	线索库，挂到原文件	解读不独立计数
J13	新闻报道：某区上线“AI 政务助手”	线索库	据此反查有无对应政策文件
J14	某区法治政府建设年度报告提到人工智能	排除	报告类
J15	北大法宝下载件只有印发通知，附件显示“无相关内容”	完整性=缺附件，回官网补件	旧库中此类 16 份
J16	某市政府与某企业签订人工智能战略合作协议的新闻	线索库	协议不是政策文件；如官网公开了协议文本，登记为“其他”交人工
J17	某省人工智能地方标准	辅库·标准	
J18	某市“十四五”科技创新规划，有“人工智能”一节	B，主库	
J19	2012 年某市“智能交通”规划，有“模式识别”技术应用任务	B，主库；前史	2015 年以前命中进前史标记
J20	某直辖市区人民政府人工智能产业扶持办法	A，主库；L3 第三层，单元类型=直辖市区县	旧库把此类误记为地级

第 4 章

谁的政策：层级、单元与发文主体

“算谁的”决定了央地比较能不能做。旧库把直辖市区县当地级、按文件夹而不是按发文机关定省份，都出在这里。本章给出层级定义、检索单元总表、每个单元要查的机关，以及特殊情形的处理。

4.1 四个层级

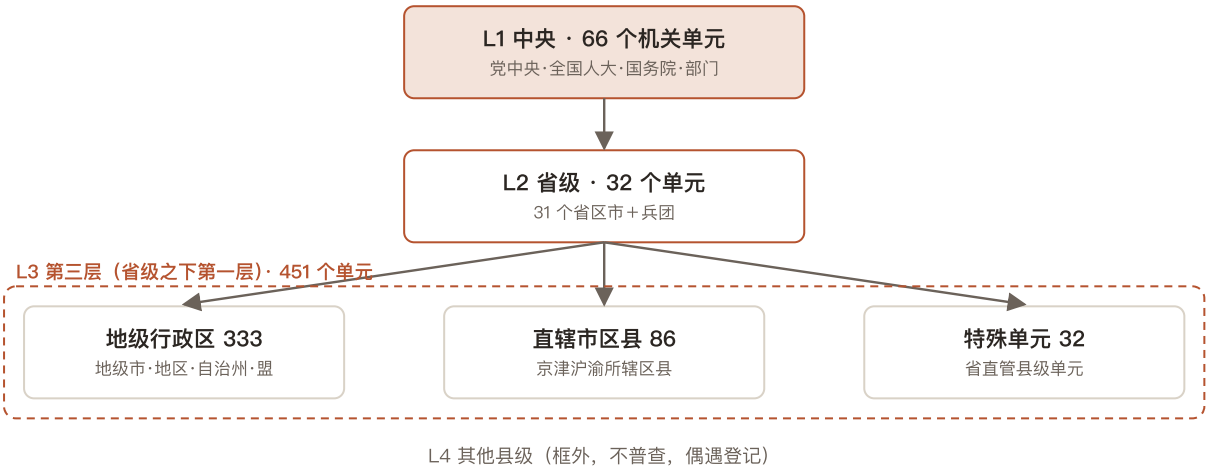


图 3　层级与检索单元

代码	层级	包括	单元数
L1	中央	党中央、全国人大及其常委会、国务院、国务院办公厅、国务院部门及部委管理的国家局、中央议事协调机构	中央检索单元 66 个（以机关为单元）
L2	省级	31 个省、自治区、直辖市的党委、人大、政府及其部门；新疆生产建设兵团作为特殊单元	31＋1
L3	第三层（地级）	333 个地级行政区（地级市、地区、自治州、盟）；4 个直辖市所辖 86 个区县；32 个省直管县级单元（单独标记，分析时可剔除）	333＋86＋32＝451
L4	县级（框外）	其余县级政府及部门	不普查，偶遇登记

“层级”按牵头发文机关所属的政府确定，不按文件存放位置，也不按文件适用范围。另设“单元类型”字段区分地级市、地区、自治州、盟、直辖市区县、省直管单元，这样分析时既能把第三层合并，也能把直辖市区县、特殊单元剔除。

4.2 检索单元总表

层级	单元数	名录来源与说明	表单
L1 中央	66	核心 15 个、扩展 49 个、司法辅库 2 个。以中国政府网“国务院组织机构”页及党中央机构设置为准，G0 前核对一遍	02 中央检索单元
L2 省级	32	31 个省、自治区、直辖市，加新疆生产建设兵团	03 省级检索单元

层级	单元数	名录来源与说明	表单
L3 地级	333	地级市 293、地区 7、自治州 30、盟 3（以名录数据计）；其中副省级 15 个另行标记	04 第三层–地级单元
L3 直辖市区县	86	北京 16、天津 16、上海 16、重庆 38	05 第三层–直辖市区县
L3 特殊单元	32	河南、湖北、海南的省直管县级单元，新疆自治区直辖县级市	06 第三层–特殊单元
合计	**549**	区划名录来自公开数据集 modood/Administrative-divisions-of-China，G0 前与民政部最新区划代码逐条核对；有出入以民政部为准	

检索单元编号统一为 `U-` 加 6 位区划码（省级为 `U-330000`，地级为 `U-330100`，区县为 `U-310115`），中央为 `U-C01` 至 `U-C66`。上一轮“方案乙”197 个单元的编号规则相同，可以直接对接。

4.3 每个单元要查哪些机关

一个政府单元的人工智能政策，可能由党委、人大、政府、政府办公厅、十几个部门和议事协调机构发出。A1 建档智能体为每个单元列出实际存在的机关和网站，A2 按机关逐个检索。核心机关必查，扩展机关在核心完成后用站内全文检索过一遍。

层级	类别	机关	优先级	说明
省级	党委	省（区、市）委、省委办公厅	核心	含与政府联合发文
省级	党委	省委网信办	核心	算法、生成式人工智能属地管理
省级	人大	省人大常委会	核心	地方性法规、决定
省级	政府	省人民政府	核心	省政府令、省政府文件
省级	政府	省人民政府办公厅	核心	
省级	部门	发展和改革委员会	核心	
省级	部门	工业和信息化厅（经济和信息化厅/委）	核心	名称各省不同
省级	部门	科学技术厅	核心	
省级	部门	数据局（大数据局/政务服务和数据管理局）	核心	名称和职能各省不同，按实际机构
省级	部门	政务服务管理局（单设的省份）	核心	
省级	部门	教育厅	核心	
省级	部门	卫生健康委员会	核心	
省级	部门	公安厅	核心	
省级	部门	人力资源和社会保障厅	核心	
省级	部门	交通运输厅	核心	
省级	部门	市场监督管理局	核心	
省级	议事协调	人工智能/数字经济/数字政府领导小组（办公室）、产业链链长办公室	核心	常以“办公室”名义发文

层级	类别	机关	优先级	说明
省级	部门	农业农村厅、民政厅、司法厅、财政厅、商务厅、文化和旅游厅、住建厅、自然资源厅、生态环境厅、应急管理厅、医保局、税务局、地方金融管理局等	扩展	在核心机关检索完成后逐个过一遍
第三层	党委	市委、市委办公室、市委网信办	核心	
第三层	人大	市人大常委会（有立法权的设区的市）	核心	地方性法规
第三层	政府	市人民政府、市政府办公室（厅）	核心	
第三层	部门	发展和改革委员会	核心	
第三层	部门	工业和信息化局（经信局）	核心	
第三层	部门	科学技术局	核心	
第三层	部门	数据局（大数据局/政务服务和数据管理局/行政审批服务局）	核心	
第三层	部门	教育局、卫生健康委、公安局、人社局、交通运输局、市场监管局	核心	
第三层	议事协调	人工智能/数字经济/数字政府领导小组办公室	核心	
第三层	功能区	国家级新区、高新区、经开区管委会	扩展	归到所在单元，机关类型记为功能区
第三层	部门	其余市直部门	扩展	以站内全文检索为主

4.4 特殊情形

情形	处理
副省级城市、计划单列市	仍归 L3，加“副省级”标记。这些城市发文量大、部门多，建档时机关清单要列全。
有立法权的设区的市、经济特区	人大常委会的地方性法规必查（R5）。深圳等经济特区有特区立法权，单独留意“经济特区条例”。
国家级新区、高新区、经开区、自贸片区	文件归所在单元（如浦东新区归 U-310115），机关类型记“功能区”。跨行政区的新区（如雄安）归所在省份的对应单元，在备注中写明。
议事协调机构（领导小组办公室、链长办公室）	按所属政府定层级，机关类型记“议事协调机构”。这类机构常是人工智能政策的实际发文者，建档时要找到它们挂靠在哪个部门网站上发布。
区划调整、撤并、更名	建档时写清 2015 年以来的沿革。被撤并单元的旧文件归继承单元，备注写原单元名；更名前后视为同一单元。
省市联合、部省联合发文	归高一层的牵头机关；低一层单元的记录中通过 issuers 字段仍可检索到。
党委与政府联合发文	机关类型记“党政联合”，归该级政府单元。
县级文件	直辖市区县以外的县级文件，检索中偶遇时登记到“框外表”（level=L4），不纳入主库统计。

第 5 章

检索：怎么找全

找全靠三件事：来源要多（不同来源漏的东西不一样）、词要全（人工智能政策的提法每隔两三年变一次）、记录要细（没有记录就无法证明找全）。本章给出来源矩阵、词表、路由、单元检索步骤和停止规则。

5.1 来源矩阵

来源	中央	省级	第三层	长处与短板
政府门户政策文件库 / 信息公开	●	●	●	官方、最新；但站内检索质量参差，旧文件常下线，部分站点有反爬和验证码
部门网站	●	●	●	部门文件很多不进门户文件库；工信、科技、数据部门最重要
政府公报	●	●	○	覆盖早年文件，目录稳定；只收部分文件，部门文件少
法律法规数据库、规章库	●	●	○	法规、规章权威；不含一般规范性文件
北大法宝等商业库	●	●	●	全文检索好、跨地区统一；有附件缺失、收录滞后，需机构账号
上级文件反查	—	●	●	能找到“贯彻落实”“转发”类文件；本身就是扩散数据
通用搜索引擎	○	○	●	兜底查漏；结果只当线索
政务新媒体、新闻	○	○	○	能发现未在网站公开的文件名；只当线索

● 必做　○ 选做或兜底。全国性来源的入口见表单“10 来源站点清单”，各单元的站点由 A1 建档时登记。

5.2 检索词表

层	检索词	用法	备注
K1 核心词	人工智能	单独检索	最主要的词；单独命中即进入候选
K1 核心词	AI	单独检索，结果需人工过目	英文缩写，噪声稍多
K1 核心词	“人工智能+”/“AI+”/“智能+”	单独检索	2019 年起的提法，注意引号与全角加号
K1 核心词	大模型、大语言模型、行业大模型	单独检索	2023 年起高频
K1 核心词	生成式人工智能、AIGC、生成合成	单独检索	
K1 核心词	智能体、AI Agent	单独检索	2025 年起高频
K1 核心词	机器学习、深度学习、神经网络	单独检索	
K1 核心词	计算机视觉、自然语言处理、语音识别、知识图谱	单独检索	
K1 核心词	具身智能、人形机器人、通用人工智能	单独检索	
K1 核心词	算法推荐、深度合成	单独检索	治理类规章用语

层	检索词	用法	备注
K2 邻近词	智能算力、智算中心、算力	与 K1 组合；单独检索时需核对正文	算力文件常含人工智能专节
K2 邻近词	机器人、智能机器人、服务机器人	同上	
K2 邻近词	智能制造、智能工厂	同上	多数不涉及人工智能技术本身
K2 邻近词	自动驾驶、智能网联汽车、无人驾驶、无人机、低空经济	同上	
K2 邻近词	智慧城市、城市大脑、数字孪生	同上	
K2 邻近词	数字政府、智慧政务、智能审批、无人干预自动审批、智能客服、数字人	同上	政务服务子集的主要入口
K2 邻近词	人脸识别、智能语音、智能问答、机器翻译	同上	
K3 历史词	智能计算机、专家系统、模式识别、机器智能	不限年份检索	用于 2015 年以前的前史
K3 历史词	“互联网+”人工智能、新一代人工智能	不限年份检索	2015—2017 年的典型提法

- 门户网站站内检索通常不支持布尔式，逐词检索；每个 K1 词都要检索一遍并登记命中数。
- 北大法宝、搜索引擎支持组合式：(K1 任一词) 限定地区与机关，不加主题限定词。旧方案把检索限定在“政务、政务服务”上，会漏掉产业、治理类文件，新方案取消这一限定。
- K2 词只在必做路由完成后、为查漏而用；K2 命中但正文没有 K1 词的文件直接按 X 处理并登记。
- 检索不设起始年份，2015 年以前的命中加“前史”标记。门户站内检索若默认只查近几年，要手动改为“全部时间”。
- 词表是活的：试点阶段每发现一个新提法（如某地自造的“AI+政务”品牌名），登记到问题单，由负责人决定是否加入，加入后已完成的单元只补检该词。

5.3 检索路由

路由	来源	要求	做法	注意
R1	本级政府门户的政策文件库/政府信息公开栏目	必做	逐个 K1 词站内检索；再按“政策文件”栏目按年份浏览标题（2015 年至今）	门户检索常有年份默认值和结果条数上限，需翻到最后一页
R2	核心部门网站（见机关模板）	必做	每个核心部门的“政策文件/规范性文件/通知公告”栏目站内检索 K1 词	数据局、工信、科技三个部门最容易出文件
R3	政府公报（本级）	必做（2015—2020 年为重点）	逐期目录浏览，标题含 K1/K2 词的打开核对	旧文件常从网站下线，但公报目录还在
R4	北大法宝（中央法规库/地方法规库）	必做	全文检索 K1 词，发布地区或发布部门限定为本单元，不限日期；导出结果清单	由团队成员用机构账号人工登录，智能体不输入账号密码；下载件必须核对附件是否完整
R5	法律法规与规章数据库	中央、省级、有立法权的市必做	国家法律法规数据库、国家规章库检索 K1 词	法规、规章以此为准

路由	来源	要求	做法	注意
R6	上级文件反查	必做	对本单元上级的每份 A 级文件，检索“贯彻落实/转发 ×× ”+本单元名	直接服务纵向扩散研究，也是查漏最有效的一步
R7	通用搜索引擎	必做（兜底）	“单元名”(K1 词) site:gov.cn；再用“印发/通知/方案”组合	只当线索：命中后必须回到官网找原件
R8	政务新媒体与新闻	选做	微信公众号、新闻发布会、媒体报道中提到的文件名	只当线索，进线索库

为了控制检索量，各路由用词不同：R1 门户逐个检索全部 K1 词；R2 部门网站检索“人工智能”“大模型”“智能体”三个词，站内检索失效时改为按年份浏览政策栏目标题；R4 北大法宝用一个组合式（全部 K1 词 OR）；R7 用两条组合式。这样一个地级单元大约 50—60 次检索。

5.4 单元检索步骤（A2 执行）

1. **读档案。**读取 A1 生成的单元档案：现行名称、沿革、机关清单、站点入口、上级单元的 A 级文件清单。
   检查点 档案中每个核心机关都有站点入口或“未找到”说明
2. **R1 门户。**在本级门户政策文件库逐个检索 K1 词，时间设为全部；结果超过 200 条时按年份切分。每次检索登记一行日志。再按年份浏览“政策文件”栏目标题，补登标题含 K1/K2 词的文件。
   检查点 每个 K1 词都有一行日志
3. **R2 部门。**按机关清单逐个进入核心部门网站检索。数据、工信、科技三个部门同时浏览其“通知公告”栏目。
   检查点 每个核心机关至少一行日志
4. **R3 公报。**浏览本级公报目录，2015—2020 年为重点。
   检查点 日志写明浏览的年份区间
5. **R4 北大法宝。**在人工登录的会话中，按“发布地区＝本单元”全文检索组合式，不限日期，导出结果清单，与已登记候选比对后补登。
   检查点 日志记录检索式和结果总数
6. **R5 法规。**中央、省级和有立法权的市检索法律法规数据库和规章库。
7. **R6 上级反查。**对上级单元（第三层的上级是省级和中央，省级的上级是中央）的每份 A 级文件，检索“贯彻落实 ××”“转发 ××”加本单元名。
   检查点 上级 A 级文件逐一有日志
8. **R7 搜索引擎。**“单元名”(K1 任一词) site:gov.cn，再加“印发 OR 通知 OR 方案”各一次。命中只当线索，回官网找原件。
9. **补查。**对 R4—R7 新发现文件的发文机关，如果它不在核心机关清单里，补查该机关网站（R2 扩展）。
10. **停止判断。**按 5.5 的规则判断能否停止；能停止则写单元小结，状态改为“检索完成”。

5.5 停止规则与“0 份”结论

1. R1—R7 中适用的必做路由全部完成，每条都有检索日志（含 0 命中）。
2. 最后完成的两条路由没有新增 A、B 级文件。若有新增，按新文件的发文机关补查该机关网站，再判断一次。

3. 所有“待定”已写明理由并提交人工。

4. 门户或核心部门网站不可访问的，登记“站点异常”，不能据此得出 0 份结论，转人工处理。

> **“0 份”结论必须经得起复查**
>
> 某个单元一份 A、B 级文件都没找到，是重要的研究信息（这个地方没有出台人工智能政策），但也是漏检最容易藏身的地方。所以 0 份单元有三道额外关口：自检规则 V16 检查日志完整；A7 用不同来源做一次独立检索（提示词 P8）；核验员人工复核其中 30%。人工复核发现漏检时，同批全部 0 份单元重做。

5.6 检索日志要记什么

日志是“找全”的证据。每执行一次检索（一个站点×一个词或一个栏目）记一行，字段见表单“16 检索日志”：路由、站点、检索式或栏目路径、时间范围设置、结果总数、实际浏览数、新增候选、站点异常、执行者、执行时间、截图或存档路径。“实际浏览数”小于“结果总数”时必须在备注中说明原因（例如按年份切分后分别浏览）。

5.7 站点访问规范与人机交接

情况	智能体怎么做	人工怎么做
正常访问	同一站点请求间隔不少于 3 秒，并发不超过 2；遵守网站声明的访问限制；只取公开页面	—
验证码、人机验证	立即停止该站点，日志记“有验证码”，问题单登记“需人工”，转做其他路由	人工在浏览器中完成验证或手动检索，把结果交回
需要登录（北大法宝等）	不输入任何账号密码；请人工在同一浏览器会话中登录，登录后继续	用机构账号登录；批量导出遵守授权协议
访问被拒、超时、检索功能失效	重试不超过 3 次，间隔递增；仍失败记“站点异常”，改用公报、法宝、搜索引擎路由，不得据此下 0 份结论	判断是否需要换网络环境或人工检索
网站改版、链接失效	用站内检索找新地址；找不到时查公报和互联网档案馆	—

政府网站常用的防护措施会把高频访问判为攻击。宁可慢，不要被封；被封后整个单元的检索都要转人工。

第 6 章

下载与存证

一份文件进库时，要能回答四个问题：从哪里拿的、什么时候拿的、拿到的是不是全文、以后能不能证明没被改过。本章的规定都是为了回答这四个问题。

6.1 原件优先级

1. 官网发布的原件附件（PDF、OFD、DOC、DOCX、WPS）。
2. 官网网页正文（文件以网页形式发布、无附件时）。
3. 政府公报刊登版。
4. 国家法律法规数据库、规章库。
5. 北大法宝等商业数据库。
6. 其他网站转载（只在前五种都找不到时使用，completeness 记“待核”，进高风险抽检）。

同一文件找到多个来源时全部登记到 source 表，主原件取优先级最高的。官网有原件时不得用北大法宝件代替。

6.2 一份文件要存什么

内容	要求
原件	正文文件和全部附件，保持原格式、原文件名另记
网页快照	文件所在网页存两份：HTML 单文件（含图片）和打印 PDF；网页正文型文件的快照就是原件
来源信息	最终链接（跳转后的地址）、来源类型、获取时间（精确到分钟）、获取者
哈希值	每个文件的 SHA-256，写入 file 表
文本	原始提取文本 raw、清洗后文本 clean、条款块 chunks（第 7 章）
元数据	meta.json，内容与数据库记录一致

6.3 命名与目录

路径（工作目录下）	内容
`00_admin/`	手册、执行包、版本说明、问题单导出、交接记录 handoff/
`01_roster/`	units.csv、agencies.csv、sites.csv
`02_search_logs/{unit_id}/`	search_log.csv；shots/ 截图与存档
`03_originals/{level}/{prov_code}/{unit_id}/{doc_id}/`	main.pdf 主原件（统一命名）；att01.docx… 附件按正文顺序编号；page.html、page.pdf 网页快照；meta.json

路径（工作目录下）	内容
`04_text/`	raw/{doc_id}.txt；clean/{doc_id}.txt；chunks/{doc_id}.jsonl
`05_db/`	ai_policy.sqlite；exports/ 按版本导出的 csv、xlsx
`06_qc/`	batches/{batch_id}/ 抽检清单与结果
`07_legacy/`	旧库原始备份、对照表与迁移日志
`08_clues/`	clues.csv 及快照
`09_aux/`	gov_reports/ 政府工作报告；standards/ 标准；judicial/ 司法文件

工作目录和其下所有文件夹、文件名一律用 ASCII 字符（建议工作目录为 `~/ai_policy_db/`），中文名称只出现在数据库和 meta.json 里。这样做是因为执行主体换成 Codex 后，大部分工作由脚本完成，中文路径在命令行工具、无头浏览器和压缩备份中都容易出错（本机已遇到过 Chrome 读中文路径失败的情况）。文件一律用 doc_id 定位，原文件名记在 file 表的 orig_name 字段，也就不会再有旧库那种“关于印发《东莞市数字…”的截断名。

6.4 完整性检查

问题	判据	处理
通知壳	正文出现“现印发给你们”“附件：”，但附件正文缺失	回官网下载附件；北大法宝件出现“附件预览 无相关内容”必查。旧库有 21 份属此类
截断	末尾没有落款、日期，或条款编号中断，或页数与官网标注不符	换来源重下
仅摘要	只有摘要、目录或第一页	换来源重下
无文字层	每页可提取字符少于 100	OCR，并标 ocr=1；OCR 后人工抽看首页和末页
乱码	提取文本中非中文、非标点字符占比异常	换提取工具或 OCR
附件多份	正文列出附件 1、附件 2……	逐个下载，全部挂到同一 doc_id
图片型附件	清单、图表以图片形式附在网页上	保存原图并 OCR

完整性不是“完整”的记录不能进入研判。A3 自动跑前三条检测规则；判不准的标“待核”，进高风险抽检。

6.5 格式转换与 OCR

格式	处理
PDF（有文字层）	`pdftotext -layout` 提取；每页可提取字符少于 100 视为无文字层
PDF（扫描件）	Tesseract 中文简体模型 OCR，或调用多模态模型识别；ocr=1，人工抽看首末页
DOC、DOCX、WPS	macOS 自带 `textutil -convert txt`，DOCX 也可用 python-docx（含表格）；WPS 文件先尝试按 DOC 处理
OFD	需另装开源 OFD 转换工具转 PDF 或文本（本机目前未装，试点前装好并测试）

格式	处理
HTML	提取正文区域，去掉导航、页脚、“浏览次数”等网站元素
图片附件	保存原图，OCR 后作为附件文本

清洗规则：统一 UTF-8；去掉页眉页脚、页码、北大法宝版权声明和引证码；保留原段落和条款编号；全角半角不做改动（文号中的六角括号除外，统一为〔〕）。raw 和 clean 都保留，抽取和研判用 clean。

第 7 章

梳理：元数据、研判、去重与关联

梳理把一堆文件变成一张能查询、能统计、能追溯的表。顺序是：抽元数据（A4）→ 研判相关性和去向（A5）→ 去重归并、识别关系、切分条款（A6）→ 自检（A0）。

7.1 元数据从哪里取

同一字段可能有多个来源：原件正文（版头、文号、落款）、网页元数据（发布日期、索引号、有效性）、北大法宝元数据（制定机关、效力位阶、时效性）。冲突时按下列顺序取值，并把冲突写入记录的 conflicts 备注：

1. 官方原件正文；
2. 官网网页上的元数据（政府信息公开目录的“索引号、发布机构、成文日期、发布日期、有效性”）；
3. 政府公报；
4. 北大法宝等数据库；
5. 推断（只用于 date_published 只知年月等情形，必须在备注说明）。

7.2 关键字段的规则

字段	规则
标题	取原件首页标题全称；印发型取附件标题，通知标题另记；书名号、引号按原文；不截断、不加注释
发文字号	统一为“机关代字〔年份〕序号”；无文号填“无”；部门令写“令第 × 号”
发文机关	规范全称，联合发文全部列出，按版头或落款顺序；简称、代字不能代替全称
日期	成文日期取落款，发布日期取网站发布或公报刊登日期；两者都尽量填；只知年月写 YYYY-MM-00
层级与单元	由牵头机关决定（第 4 章），不按文件夹、不按适用地域
效力类别	法律法规规章按制定主体和程序判定；其余政府文件中标明“行政规范性文件”或有规范性文件登记号的记“行政规范性文件”，其他记“其他政策文件”
时效	官网标注优先；发现后续废止或修订文件时更新，并建关系
规划/实施期	标题或正文写明“2025—2027 年”的填起止年；“十四五”填 2021—2025

全部字段的定义、类型和取值见附录 C 与表单“14 字段字典”。

7.3 相关性研判

1. **机器初筛。**统计 K1 词在正文中的出现次数和位置（ai_hits），标题是否含 K1 词。初筛只用来排序和发现异常：命中为 0 的直接预标 X；标题含 K1 词的预标 A。
2. **模型研判。**A5 按提示词 P5 读全文，给出等级、去向、技术形态、领域和 1—3 处原文证据。长文件按条款块逐块判断“是否有以人工智能为对象的独立任务”，再汇总。

3. **规则校验。**自检 V10、V11 检查证据能否在正文中逐字找到、等级与命中数是否矛盾。
4. **多模型复核（试点和高风险记录）。**试点阶段由 Codex 和 Claude 各判一遍，不一致的交人工；正式阶段由 Claude 对“待定”“低置信度”和抽中的记录做复核，保证检查者和执行者不是同一个模型。
5. **人工裁决。**待定和分歧由核验员裁决，新情形由负责人写入判例。

> **技术形态与领域标签**
>
> A、B 级记录必须填 ai_forms（通用人工智能、大模型、生成式、智能体、机器学习与视觉语音、具身智能与机器人、算力、算法治理、数据）和 domains（产业发展、科技研发、政务服务、政府治理与监管、公共安全、医疗卫生、教育、交通等 19 类）。标签按“独立任务所在领域”打，可多选。政务服务子集＝domains 含“政务服务”或“政府治理与监管”的 A、B 级记录。

7.4 去重与归并

层级	判据	处理
一级：完全相同	SHA-256 相同	自动合并为同一文件，保留全部来源
二级：同一文件	规范化标题相同且文号相同；或规范化标题相同、发文机关相同、日期相差 30 天内	自动归入同一 dup_group，主原件按优先级选取
三级：疑似	标题相似度≥0.85，或正文前 1500 字相似度≥0.9	A6 按 P6 判断是同一文件、修订版、年度版、转发件还是配套文件；判不准交人工

“规范化标题”指去掉书名号、引号、空格、括号内的文号和“（试行）”以外的修饰后的标题。标题相同但文号不同的，绝不自动合并。

7.5 关系

关系	方向	说明
印发	通知—附件	同一文件的两部分，合并为一条记录；此关系只在拆分存储时使用
转发	转发文件—被转发文件	下级转发上级或同级文件
依据	本文件—上级依据文件	正文写明“根据/为贯彻落实/按照”
配套	细则/办法—主文件	实施细则、配套措施、任务分工方案
修订	新版本—旧版本	修订、修正、替代
废止	废止文件—被废止文件	
解读	解读—原文件	解读材料挂靠，不独立计数
名单	名单公示—征集通知	征集与结果的对应

关系表是纵向扩散分析的直接材料：转发和依据关系记录了下级对上级文件的响应，名单关系记录了场景征集的结果。A6 从正文中“根据”“为贯彻落实”“按照”“现将……转发给你们”等句式识别关系，证据句逐字写入 evidence。

被引用的上级文件不在库里时，先建一条“候选”记录，交 A2 补检。

7.6 条款切分（为工具挖掘做准备）

A、B 级文件按公文结构切成条款块，供后续政策工具抽取使用。这一步对应朱旭峰工作流的“结构与语义约束分块”。

- 识别层级标记：编、章、节、条，以及“一、（一）1.（1）”四级序号；按最细一级有完整语义的单位切分。
- 每块不超过约 800 字；超长的按句号切分并保留上级标题路径。
- 每块记录：chunk_id（doc_id＋序号）、标题路径（如“三、重点任务 / （二）推进政务应用”）、正文、字数、是否含 K1 词。
- 附件中的表格（任务分工表、场景清单）按行切分，每行一块，保留表头。

```
{"chunk_id":"M330100-2025-0007-023","doc_id":"M330100-2025-0007",
 "path":["三、重点任务","（二）推进人工智能政务应用"],"seq":23,
 "text":"……","chars":312,"k1_hit":true}
```

第 8 章

数据库

数据库用 SQLite：单文件、免安装、能被 Python、R 和 Excel 插件直接读，便于在团队之间传递和版本冻结。完整建表语句见附录 B。

8.1 表结构

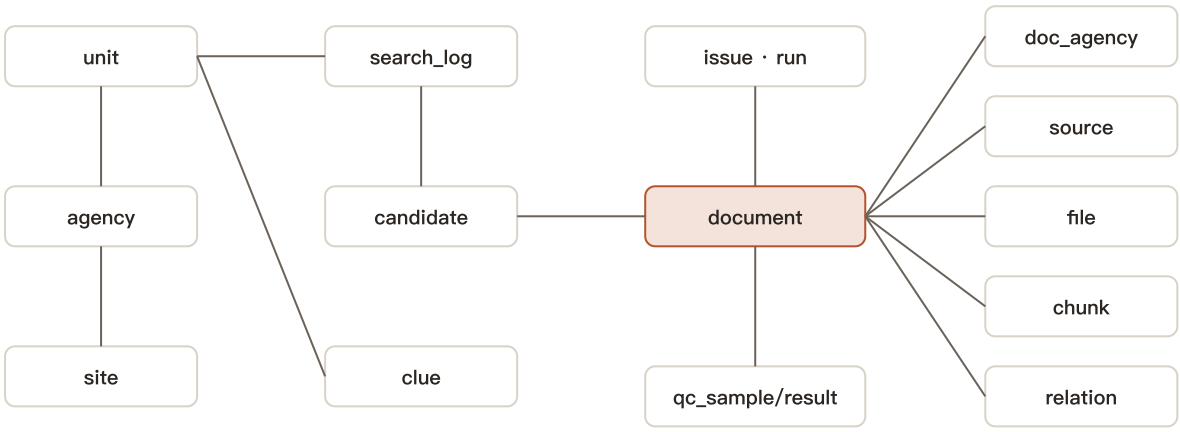


图 4　数据库表关系（document 为主表）

表	一行是什么	主要字段
unit	一个检索单元	unit_id、层级、单元类型、名称、区划码、上级、状态、批次
agency	单元内的一个机关	agency_id、unit_id、名称、类别、核心/扩展、网站
site	一个检索入口	site_id、agency_id、类型、网址、验证状态
search_log	一次检索	log_id、unit_id、路由、站点、检索式、时间范围、结果数、浏览数、新增候选、异常
candidate	一条候选	cand_id、unit_id、标题、链接、发现路由、处理结果（转为 doc_id 或排除原因）
document	一份文件（主表）	字段字典全部字段
doc_agency	文件与发文机关的对应	doc_id、机关名称、顺序、是否牵头
source	文件的一个来源	doc_id、来源类型、链接、获取时间、是否主来源
file	磁盘上的一个文件	file_id、doc_id、角色（正文/附件/快照）、路径、格式、哈希、页数、原文件名
relation	两份文件的关系	from、to、类型、证据
chunk	一个条款块	chunk_id、doc_id、路径、序号、正文、K1 命中
clue	一条线索	链接、类型、提到的文件名、是否找到原文

表	一行是什么	主要字段
qc_sample / qc_result	一条抽检样本 / 一个字段的核验结果	批次、抽样类型、字段、库内值、核验值、结果
issue	一张问题单	类型、等级、描述、处理、状态
run	一次智能体运行	智能体、模型名称与版本、提示词版本、开始结束时间、处理对象

8.2 编号规则

对象	格式	示例
检索单元	U–区划码6位；中央 U–C两位	U–330100；U–C11
文件	层级码＋区划码6位–发布年–流水4位	C000000–2017–0001；P330000–2024–0012；M330100–2025–0007
检索日志	L–unit_id–流水4位	L–U–330100–0031
候选	K–unit_id–流水4位	K–U–330100–0105
条款块	doc_id–流水3位	M330100–2025–0007–023
抽检批次	B–范围–序号	B–33–01（浙江第一批）；B–C–01（中央）

层级码：C 中央、P 省级、M 第三层、X 框外。流水号在“层级码＋区划码＋年份”内递增，由 A0 统一分配，智能体不得自行编号，避免并发冲突。

8.3 版本、冻结与增量

- **口径版本。**手册、字段字典、词表、判例、提示词各有版本号；任何修改都登记问题单并升版本。每条记录在 run 表里能查到处理它的口径版本和模型版本。
- **数据冻结。**过验收门 G4 后，把数据库、原件、文本打包为 v1.0，计算整包哈希，存两处（项目文件夹和另一块介质或云盘）。冻结版只读，后续修改在新版本中进行。
- **增量。**每月对每个单元跑一次增量检索，只查上次检索日以后的发布；新文件走完整流程，并入下一个小版本（v1.1、v1.2……），变更记录写明新增、修改、删除的 doc_id。
- **导出。**每个版本导出 CSV（UTF–8 带 BOM，方便 Excel 打开）和 xlsx，字段名用英文、第二行附中文名。

第 9 章

智能体作业规程

本章写给运行智能体的人，也直接作为智能体的工作说明。每个智能体只做一件事，输入输出都是文件或数据库记录，便于检查和重跑。

9.1 角色分工

编号	角色	职责	输入	输出
A0	调度（脚本＋Codex）	分配工作包、维护状态、汇总进度、跑自检脚本、生成抽检样本	名录、状态表	工作包、进度看板、抽检清单
A1	建档智能体	为每个单元核对区划、列机关名录、找齐站点入口	单元名录、机关模板	单元档案（机关表、站点表）
A2	检索智能体	按路由逐条检索、登记每次检索、产出候选清单	单元档案、词表、路由表	检索日志、候选清单
A3	下载存证智能体	下载原件与附件、保存网页快照、计算哈希、做完整性检查	候选清单	原件包、文件表
A4	解析抽取智能体	提取正文、OCR、清洗、抽取元数据	原件包	正文文本、元数据 JSON
A5	研判智能体	判定相关性等级、文件类别、领域与技术形态，给出原文证据	正文、元数据	研判结果 JSON
A6	归并关联智能体	跨来源去重、识别转发/依据/修订关系、切分条款	全部已研判记录	去重组、关系表、条款块
A7	质检智能体（Claude）	对抽中的单元做二次独立检索；对抽中的记录做独立复核；由与执行主体不同的模型担任	抽检清单	复核结果、问题单
H1	核验员（人工）	抽检、裁决待定、复核 0 份单元、处理站点异常与登录	抽检清单、问题单	抽检记录、裁决
H2	负责人（人工）	拍板口径、处理升级问题、过验收门、冻结发布	进度看板、问题单	裁决记录、版本

9.2 工作包与并发

- **工作包＝一个检索单元。**A0 从名录中按批次取单元，生成工作包（单元档案、词表、路由、上级 A 级文件清单），分派给 A1—A5 依次处理。A6 在一批全部完成后统一运行（去重和关系需要看全批）。
- **并发。**同时处理 4—6 个单元；同一省份的单元可以并行，因为它们访问的是不同网站。同一网站（例如省政府门户）只允许一个智能体访问。
- **顺序。**先中央、再省级、再第三层。第三层的 R6 上级反查依赖中央和省级的 A 级文件清单，所以省级必须先于本省第三层完成。
- **运行环境。**执行主体为 Codex：由调度脚本为每个工作包启动一次运行，网页用脚本和无头浏览器访问，下载、哈希、文本提取、自检用 Python 脚本完成（详见 9.6—9.9）。每次运行在 run 表记录执行主体、模型名称

与版本、执行包版本和起止时间。

- **中断恢复。**所有中间产物写盘；任何一次运行都不依赖上一次运行的记忆，重启后从状态表和交接记录读取进度继续，不重复已完成的检索（按 search_log 去重）。

9.3 状态机

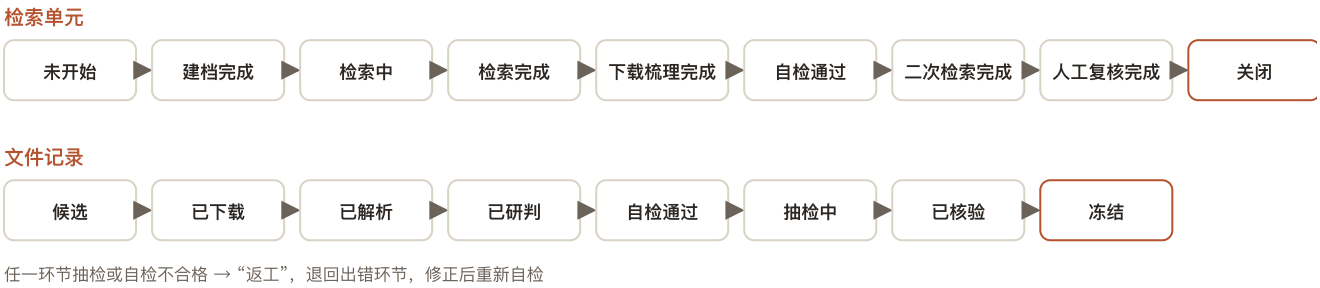


图 5　状态机

状态只能由 A0 根据自检结果推进，智能体自己不能把记录标为“已核验”。“返工”状态的记录退回到出错的环节，修正后重新走自检。

9.4 各智能体作业卡

A1 建档智能体

输入	单元名录一行；机关模板；全国性来源清单
步骤	1. 核对现行名称与区划码，查 2015 年以来的撤并、更名，写明依据。 2. 按机关模板逐个确认实际机关名称及网站首页、政策文件栏目、站内检索入口。 3. 找本级政府公报目录、人大常委会法规栏目（有立法权的）。 4. 试一次站内检索（“人工智能”），确认检索功能可用、能否设全部时间。
检查点	核心机关都有入口或“未找到”说明；所有网址都实际打开过
输出	单元档案 JSON；site 表记录
异常	站点打不开、需登录：记问题单“需人工”，档案标注，不阻塞其他单元

A2 检索智能体

输入	单元档案；K1 词表；路由表；上级 A 级文件清单
步骤	1. 按 5.4 的十步执行 R1—R7，每次检索登记日志。 2. 标题或摘要含 K1 词、或疑似相关的，登记为候选；新闻解读中的文件名登记为线索。 3. 按停止规则判断能否结束，写单元小结。
检查点	每个 K1 词在 R1 有日志；每个核心机关在 R2 有日志；上级 A 级文件在 R6 逐一有日志
输出	search_log、candidate、clue
异常	验证码、登录、封禁：停止该站点，记“站点异常”，转其他路由

A3 下载存证智能体

输入 候选清单

步骤
1. 按 6.1 优先级取原件和全部附件，存网页快照。
2. 计算 SHA-256，记录最终链接、来源类型、获取时间。
3. 跑完整性检测（通知壳、截断、仅摘要、无文字层）。

检查点 completeness 不是“完整”的都写了问题；北大法宝件已核对附件

输出 原件包、file 表、source 表

异常 三次下载失败记“下载失败”，候选保留，交人工

A4 解析抽取智能体

输入 原件包；网页元数据；法宝元数据

步骤
1. 提取文本，必要时 OCR；清洗出 clean 文本。
2. 按 P4 抽取元数据，冲突按 7.1 顺序取值并记录。
3. 统计 ai_hits，写入 document 表（状态：已解析）。

检查点 V01—V09、V12 通过

输出 raw/clean 文本；document 记录

异常 抽取字段与原文对不上：标低置信度，进复核

A5 研判智能体

输入 clean 文本或条款块；元数据；分级规则与判例

步骤
1. 按 P5 判定相关性、去向、技术形态、领域。
2. 摘录 1—3 处逐字证据并标位置。
3. 拿不准的标“待定”，写清卡在哪条规则。

检查点 V10、V11 通过；A/B 级都有证据

输出 研判结果写入 document（状态：已研判）

异常 证据在正文中找不到：判定作废，重跑一次；再失败交人工

A6 归并关联智能体

输入 一批全部已研判记录

步骤
1. 一级、二级去重自动执行，三级按 P6 判断。
2. 识别转发、依据、配套、修订、废止、解读、名单关系。
3. 对 A/B 级切分条款块。

检查点 V13、V14 通过；被引用但不在库的上级文件已建候选

输出 dup_group、relation、chunk

异常 无法判断是否同一文件：两条都保留，问题单交人工

A7 质检智能体（Claude）

输入	抽检清单（A0 生成）
步骤	1. 对双路样本单元按 P2 做独立检索（与原路由来源不同）。 2. 对 0 份单元按 P8 审查与再检。 3. 对抽中记录按 P7 盲填关键字段后比对。
检查点	复核结果逐条有依据；发现漏检立即登记 P1 问题
输出	qc_result、issue
异常	与原结果分歧：不自行改库，交核验员裁决

9.5 自检规则

A0 在每个单元完成后、每批抽检前运行以下规则。“阻断”项不通过的记录不能进入下一状态，“报警”项写入问题单由人工判断。

编号	规则	检查内容	级别
V01	必填字段齐全	按字段字典“必填”列逐条检查	阻断
V02	枚举值合法	枚举字段只能取字典列出的值	阻断
V03	日期合法	格式 YYYY-MM-DD；发布日期不晚于获取日期；成文日期不晚于发布日期 30 天以上时报警	阻断/报警
V04	文号格式	统一为〔〕；文号中的年份与成文年份不一致时报警	报警
V05	层级一致	level、unit_kind、admin_code 与名录一致；牵头机关名称中含的地名与单元一致	阻断
V06	链接可达	source_url 返回 200 或已有快照	报警
V07	哈希与文件	sha256 与磁盘文件一致；同一哈希只能对应一个 doc_id	阻断
V08	完整性	completeness 为“完整”才可进入已研判；通知壳检测规则自动跑一遍	阻断
V09	正文长度	主库文件正文少于 500 字报警（名单、短通知除外）	报警
V10	相关性证据	A/B 级必须有 1 条以上引文，且引文能在正文中逐字找到	阻断
V11	核心词一致	relevance 为 A/B 但 ai_hits 为 0 时阻断；relevance 为 X 但 ai_hits≥5 时报警	阻断/报警
V12	标题一致	title 与原件首页标题的相似度低于 0.9 报警	报警
V13	去重	标题规范化后相同且文号相同的记录必须同组	阻断
V14	关系闭合	relation 中引用的 doc_id 必须存在；转发文件必须有被转发对象	报警
V15	日志对账	每个单元：检索日志中纳入数之和＝该单元主库记录数	阻断
V16	0 份单元	0 份单元必须有全部必做路由日志，且已安排二次独立检索	阻断
V17	文件命名	文件名符合命名规则，路径与 unit_id 一致	阻断
V18	OCR 标记	ocr=1 的记录需人工抽看首末页	报警

9.6 执行主体换成 Codex：哪些要跟着变

多智能体工作改由 Codex 执行，Claude 只承担少量独立复核。口径、流程、验收标准都不变；变的是工具、指令和状态的传递方式、人机交接的方式。下表逐项说明。

方面	原来的设想	交给 Codex 后的做法
指令依据	对话＋手册	Codex 能读到我们的对话，但对话里有不少已经放弃的方案（效率论与风险论、方案乙的 197 单元、2017 年起、检索限定“政务”）。规定：**执行包是唯一依据**，对话只作背景；两者冲突时以执行包为准，并登记问题单。
指令文件	会话提示	工作目录根放 AGENTS.md（Codex 启动时会读取）。本机主目录是一个 git 仓库，里面已有一份关于课程网站设计的 AGENTS.md，可能被一并读入；工作目录要单独 `git init`，并在项目 AGENTS.md 开头声明以本文件为准。是否生效在执行环境检查中实测。
多智能体与并发	一个调度会话调用子智能体	调度脚本为每个工作包启动一次非交互运行（如 `codex exec`，以本机版本支持的方式为准），同时 4—6 个。每次运行只处理一个工作包，只读写约定的文件；doc_id 由调度脚本统一分配。
联网	内置网页搜索、抓取和浏览器	Codex 的沙箱默认可能不允许联网，需在配置中为工作目录开启网络访问。网页访问改为 Python 脚本（requests）＋无头浏览器（Playwright）处理脚本化页面；通用搜索引擎检索用 Codex 自带的网页搜索（如已开启），否则由人工补做 R7。
登录与验证码	人在应用内浏览器里登录	用有界面的 Playwright 持久化浏览器配置：由团队成员打开这个浏览器、用机构账号登录一次，脚本复用该会话。Codex 不接触账号密码；遇到验证码一律停下交人。
跨模型复核	两个厂商的模型	Codex 承担量大的 A0—A6；A7 独立复核、0 份审查、试点期双模型比对用 Claude，量约占记录的 10%—15%。
记忆与续跑	会话上下文	Codex 每次运行之间不共享记忆。状态只放在磁盘：SQLite、工作包目录、交接记录（9.9）。
路径与编码	中文文件夹名	工作目录及其下全部路径改用 ASCII（6.3），中文只出现在数据库和 meta.json 中。
运行记录	记录 Claude 模型	run 表的 agent 字段写“codex”或“claude”，model 字段写实际模型名与版本，prompt_version 写执行包版本号。
验收	—	不论之前是否有人用 Claude 试过部分单元，Codex 的结果都要单独通过试点门 G1。

9.7 执行包

执行包把手册中执行者需要的内容转成机器可读的文件，与本手册由同一份口径数据生成，内容一致。它放在项目文件夹的“政策库执行包”中，开工前整体复制到工作目录根。

文件	内容
`AGENTS.md`	任务说明、依据优先级、十条硬性规则、目录约定、每次运行的开始和结束动作
`rules.json`	字段字典与枚举、相关性分级与判定规则、文种去向、判例、词表、路由、停止规则、完整性检查、自检规则、抽检方案、关系类型、状态机、验收门

文件	内容
`prompts/P1.md` — `P8.md`	八个提示词全文（附录 A）
`roster/units.csv`	549 个检索单元（与表单 02—06 一致）
`roster/agency_templates.csv`	省级、第三层机关模板
`roster/known_items.csv`	已知项（只给调度脚本和 A7 用，不得出现在检索运行的上下文里）
`legacy/oldlib_diagnosis.csv`	旧库 673 个文件的诊断与建议去向
`schema.sql`	附录 B 建表语句
`VERSION`	执行包版本号；口径修改后升版本，重新生成

口径只在一处修改：负责人改手册口径，由同一份口径数据重新生成手册、表单和执行包。Codex 不得自行修改 rules.json 或提示词；发现规则不够用时登记问题单。

9.8 脚本做什么，模型做什么

Codex 长于写代码和跑代码。凡是规则能写清楚的工作，先让 Codex 写成脚本、用试点单元测过，再批量运行；只有需要读懂文本的判断才交给模型逐份处理。这样结果可复现，也能省算力。

环节	脚本	模型
名录	区划核对、站点可达性检测	找机关的实际名称和栏目入口（P1）
检索	门户检索请求、翻页、日志写入、候选去重	判断标题是否列为候选；应对网站改版（P2）
下载存证	下载、快照、哈希、完整性规则	判断附件链接在页面哪里（个别站点）
文本	提取、OCR、清洗、条款切分	—
元数据	正则抽取文号、日期、规划期	发文机关、政策形式、效力类别等需要理解的字段（P4）
研判	核心词统计；证据逐字校验（V10）	相关性、去向、标签（P5）
去重关系	一级、二级去重；相似度计算	三级去重、关系识别（P6）
质检	抽样（固定随机种子）、V01—V18、统计与召回估计	独立复核与 0 份审查（P7、P8，由 Claude 执行）

模型研判可以由 Codex 在运行中逐份阅读，也可以由脚本调用模型接口、按 JSON Schema 批量处理。后者更容易复现，但需要接口额度。两种做法都要把模型名、版本和提示词版本写进 run 表。

9.9 交接与续跑

1. **每次运行的开始动作。**依次读 AGENTS.md、rules.json、本工作包、最近一份交接记录、与本批有关的未关闭问题单。

   检查点　运行日志第一段写明读了哪些文件、执行包版本号

2. **每次运行的结束动作。**写交接记录 `00_admin/handoff/{日期}_{工作包}.md`：完成了哪些单元和记录、哪些没完成、卡在哪里、下次从哪一步开始、登记了哪些问题单。

3. **首次交接的理解测试。**开工前让 Codex 用自己的话回答五个问题：A、B、C、X 怎么区分；什么情况下可以下“0 份”结论；遇到验证码怎么办；官网原件和北大法宝件冲突听谁的；直辖市的区算哪一层。负责人看过答

案、纠正误解后才算通过 G0。

4. **人在环中的节点。**需要登录、验证码、站点异常、待定、P1 问题时，Codex 停下该项、登记问题单、继续做其他能做的事，不等待、不绕过。操作者每天集中处理一次。

5. **Claude 的介入点。**试点期的双模型研判、每批抽检中的独立复核（P7）、0 份单元审查（P8）、口径疑难的初步分析。调度脚本按抽检清单把对应记录和原件打包交给 Claude，结果写回 qc_result。

第 10 章

人工核验与抽检

智能体能把量做上去，但“有没有漏”“判得对不对”最终要由人确认，并且要能用数字说清楚。本章规定人在哪里介入、抽多少、怎么判合格、怎么估计召回率。

10.1 三道关

关口	谁做	做什么
第一道：自检	A0（程序）	V01—V18 规则全量检查；阻断项不通过不放行
第二道：独立复核	A7（另一模型或另一次独立运行）	对抽中的单元二次独立检索；对抽中的记录盲填关键字段后比对
第三道：人工抽检	核验员	按 10.2 的方案抽检；裁决待定和分歧；复核 0 份单元；处理站点异常

10.2 抽样方案

对象	比例	核什么	合格线
A/B 级纳入记录	每批随机 10%，不少于 20 条；不足 20 条全检	全字段对照官方原件；相关性等级与证据	关键字段错误率≤2%；相关性一致率≥90%
C 级登记记录	每批 5%，不少于 10 条	是否应升为 B 级	升级比例≤5%
排除记录（X 级）	每批 10%，不少于 20 条	是否误排	误排 A/B 为 0；误排 C≤5%
待定记录	100%	人工裁决并写理由	—
高风险记录	100%	OCR、通知壳补件后、无文号、日期冲突、层级存疑、来源非官网	—
0 份单元	智能体二次独立检索 100%；人工复核 30%	检索日志是否完整；人工换路由再查	人工复核发现 A/B 文件即该批 0 份结论全部重做
双路检索样本单元	每批 10% 的单元，不少于 3 个	两个智能体分别走不同来源，计算重叠，用于召回估计	估计召回率≥95%（A/B 级）
已知项回收	全部已知项	已知存在的文件是否被找回	A 级回收率≥95%，漏掉的逐个查原因

关键字段：标题、发文字号、发文机关、发布日期、层级、检索单元、效力类别、文件类别、相关性等级。抽样用随机数种子固定（种子写入批次记录），保证别人能复现抽中了哪些。“高风险记录”包括 OCR 件、补件后的通知壳、无文号、日期冲突、层级存疑、来源不是官网。

10.3 批次判定

判定	条件	处理
通过	关键字段错误率≤2%，相关性一致率≥90%，无 A/B 级漏检或误排	该批记录状态改为“已核验”，进入冻结候选
定向返工	关键字段错误率 2%—5%，或错误集中于一种类型	智能体按错误类型批量修正该批全部记录，再抽一次（新样本）
整批返工	关键字段错误率＞5%，或发现 A 级漏检、误排	退回检索或研判环节重做该批，修正提示词或规则后重跑

10.4 召回评估：怎么知道漏了多少

只抽检“已经收进来的”记录，只能说明精确度，说明不了漏检。本手册用两种办法估计召回率。

办法一：已知项回收

事先准备一份“已知存在”的文件清单（表单“22 已知项回收”），不告诉检索智能体，检索完成后看找回了多少。清单来源：中央层由负责人或核验员从官方文件库挑选核心文件（表单已列 23 条起步，需逐条核实）；省级和第三层用旧库中经人工确认的 A、B 级文件。

回收率＝流程找回的已知项数 ÷ 已知项总数。A 级已知项回收率要求≥95%；每个没找回的已知项都要查原因（词表缺词、路由没走、站点异常、判错等级），并据此修正流程。

办法二：双路独立检索与捕获—再捕获估计

每批随机抽 10% 的单元（不少于 3 个），由两个智能体各自独立检索：甲以门户和部门网站为主（R1—R3），乙以北大法宝、法规库、搜索引擎和上级反查为主（R4—R7）。两路都完成研判后，统计 A、B 级文件：

> n_1＝甲路找到的数量；n_2＝乙路找到的数量；m＝两路都找到的数量。
> 总量估计（Chapman 公式）：$\hat{N} = (n_1+1)(n_2+1) \div (m+1) - 1$
> 两路合并找到的数量：$U = n_1+n_2-m$；估计召回率 $= U \div \hat{N}$
>
> **算例**（示意数字，非实测）：甲路 40 份，乙路 36 份，重叠 30 份。$\hat{N} = 41\times37\div31-1\approx47.9$，合并找到 46 份，估计召回率≈96%。

这个估计假定两路“抓到”某份文件的概率互不相关。实际上容易被找到的文件两路都容易找到，会让 $\hat{N}$偏低、召回率偏高，所以两路必须用不同类型的来源，并把估计值当作上限来解读。召回率低于 95% 时，看乙路独有、甲路漏掉的文件属于哪类，针对性补查全批。

10.5 一致性与仲裁

- **试点阶段**：两名核验员各自独立对同一批 100 条记录（含 A/B/C/X 各类）判定相关性等级，计算 Cohen’s κ，要求≥0.80；不足时讨论分歧、修改规则和判例后再测一批。
- **模型与人**：同样计算 A5 判定与人工判定的 κ，作为是否可以放量的依据（试点门 G1）。

- **仲裁**：核验员之间、核验员与模型之间的分歧，由负责人裁决；裁决理由写入问题单，属于新情形的追加判例。

10.6 返工与问题单

所有问题都登记到问题单（表单“21 问题单”），分三级：P1 影响口径或整批（如发现词表漏了一个常用提法），P2 影响若干记录，P3 个别记录。P1 问题处理完之前，相关批次不得过批次门。返工时，智能体按错误类型批量修正该批全部记录，而不只是改被抽中的那几条。

10.7 核验员操作要点

1. **打开抽检清单**（数据库导出或表单“20 抽检记录”），每条记录同时打开库内原件和 source_url 指向的官网页面。
2. **先核“是不是这份文件”**：标题、文号、发文机关、日期与官网一致。
   检查点 官网打不开时用快照核，结果记“无法核”并写原因
3. **再核“全不全”**：附件是否齐全，末页是否有落款。
4. **再核“算谁的”**：层级、单元、机关类型。
5. **最后核“判得对不对”**：读相关性证据所在段落，按第 3 章规则判断等级和去向是否正确。
6. **记录**：逐字段填“对/错/无法核”；错的写正确值和依据。熟练后一条 A/B 级记录约需 3—5 分钟。
7. **0 份单元复核**：先看检索日志是否完整，再自己用两条最可能出结果的路由（通常是门户站内检索“人工智能”和北大法宝）查一次。

第 11 章

旧库迁移

旧库 673 个文件不直接搬进新库，而是作为种子和已知项，逐个按新流程核对后进库。表单“23 旧库诊断与迁移”已给出每个文件的机器预诊断和建议去向。

11.1 建议去向汇总

建议去向	份数	下一步
按新流程重新登记	502	找官网原件和链接，补元数据，按新口径研判
人工确认后多半排除	76	正文无人工智能核心词；人工确认后排除，保留记录
移入线索库	27	新闻、解读、动态、答复；据此反查原文
排除（报告类）	27	年度报告、法治政府建设报告等
OCR 后重新研判	19	无文字层
移入辅库·政府工作报告	14	v1.0 只收提到人工智能的报告
补附件后重新登记	6	通知壳（另有部分通知壳已含在其他类别中，迁移时一并补件）
删除重复副本	2	完全重复
合计	673	

11.2 迁移步骤

1. **冻结旧库。**把现有文件夹整体复制到 `07_旧库迁移/原始备份/`，以后不再改动原文件夹。
2. **人工确认去向。**核验员按表单 23 逐条确认“建议去向”（重点看“人工确认后多半排除”“移入线索库”两类），在“人工确认”列填写。
3. **挂到单元。**按新口径确定每个文件的层级和检索单元（直辖市区县 129 个改为 L3 直辖市区县；白银市文件改归甘肃）。
4. **找官方来源。**对确认保留的文件，A2 在对应单元检索时把它当作候选之一，找到官网原件和链接；找不到官网版本的，source_type 记实际来源，进高风险抽检。
5. **补件与重提取。**通知壳回官网补附件，无文字层的 OCR，截断文件名的按原件补全标题。
6. **按新流程研判。**旧文件和新检索到的文件一起走 A4—A6，不享受“免检”。
7. **回填对照。**在表单 23 的“新 doc_id”列填对应编号，形成旧—新对照表；旧库里有、新流程没找回的文件逐个说明原因（这一步同时是已知项回收测试）。

上一轮“方案乙”已做的工作

《方案乙–地级检索执行清单_197 单元》里已登记的检索记录和文件映射可以直接导入新表单的检索日志和候选表。但当时的检索式带有“政务”限定，这 197 个单元仍需按新词表补检 R1、R2、R4，不能视为已完成。

第 12 章

执行计划

先小范围试点把口径和工具调准，再按“中央—省级—第三层”顺序分批铺开，每批过门再进下一批。下面的时长是开工前的估计，试点结束后用实测数据重排。

12.1 阶段与验收门

阶段	建议时长	内容	出口
P0 准备	1 周	完成执行环境检查（表单 28）和执行主体理解测试；生成执行包；核对区划名录和中央机关名录；验证全国性来源网址；建库脚本、自检脚本、下载与提取脚本跑通；安装 OFD 转换和中文 OCR；旧库备份	G0 开工门
P1 试点	2 周	11 个试点单元（见 12.2）全流程；两名核验员测 κ；模型与人工 κ；已知项回收；修订词表、判例、提示词，定稿 v1	G1 试点门
P2 中央	1—2 周	66 个中央单元；中央 A 级文件清单是省级 R6 的输入	G2（中央批）
P3 省级	2—3 周	32 个省级单元，按东部、中部、西部、东北四批	G2（各批）
P4 第三层	8—10 周	451 个单元，按省分批；每省完成省级后才开始其第三层	G2（各省批）
P5 全库质检	2 周	全库去重和关系检查；分层级召回估计；旧库迁移收尾；数据说明书	G3、G4
P6 增量	持续	每月增量检索，每季度小版本	—

门	名称	通过条件	签字
G0	开工门	执行环境检查表全部通过；执行主体完成理解测试；名录核对完毕；站点清单覆盖全部核心机关；词表、字段字典、提示词冻结为 v1；SQLite 建库脚本跑通	负责人
G1	试点门	试点单元全部关闭；已知项回收率≥95%；κ≥0.80；关键字段错误率≤2%；试点问题单全部处理	负责人＋核验员
G2	批次门（每省/中央一批）	该批全部单元关闭；自检规则 V01—V18 无阻断项；抽检通过	核验员
G3	全库门	各层级召回估计≥95%；全库去重与关系检查通过；旧库迁移完成	负责人
G4	冻结发布	生成版本号、数据说明、变更记录；数据库与原件打包备份两处	负责人

12.2 试点设计

试点单元要能暴露各种难处：不同层级、发达与欠发达地区、旧库文件多与少、直辖市区县。

层级	试点单元	选择理由
中央	中央网信办、科技部、工业和信息化部	人工智能文件最多、类型最全（规章、规划、通知、标准、名单）
省级	浙江省、甘肃省	东西部对照；旧库各有 5 份和 1 份

层级	试点单元	选择理由
第三层	杭州市、衢州市、兰州市、定西市	副省级与一般地级、东部与西部；衢州、定西在旧库中为 0 份，用来检验“0 份”结论
直辖市区县	上海市浦东新区、天津市和平区	新区与老城区；天津区县在旧库中文件最多，用来检验迁移

试点期间每个单元都做双路检索，所有记录都做人工核验，作为后续抽检比例的依据。试点要交出：每单元耗时（智能体与人工）、各路由的新增贡献、常见错误类型清单、修订后的词表和判例。

12.3 人员与工作量

- **负责人 1 名**：口径裁决、验收签字，每周固定时间处理 P1 问题单。
- **核验员 2—3 名**：试点期全程投入；铺开后按批抽检。按 10.2 的比例，一个约 15 个地级单元的省份批次，预计需要人工 1.5—3 个工作日（抽检、0 份复核、站点异常处理），试点后校准。
- **Codex 操作者 1 名**（可由负责人或核验员兼任）：启动和监控运行、每天集中处理登录、验证码和站点异常、审看交接记录、维护脚本；同时负责把抽检样本交给 Claude 复核。
- **智能体时长**：一个地级单元约 50—60 次检索，加下载、抽取、研判，估计 1—3 小时（视站点状况）；中央核心机关单元更长。并发 4—6 个单元时，第三层 451 个单元约需 8—10 周。以上均为估计，以试点实测为准。

12.4 主要风险与应对

风险	应对
政府网站反爬、验证码导致大面积“站点异常”	降低访问频率；改走公报、北大法宝、搜索引擎；人工补检；站点异常单元不得下 0 份结论
北大法宝账号与授权限制	开工前确认学校授权范围；以检索和查漏为主，原件以官网为准
口径在执行中漂移	判例和问题单制度；口径改动升版本，已完成单元是否补做由负责人决定并记录
模型判定偏松或偏严	试点测 κ；正式阶段按批抽检；证据必须逐字可查
工作量超出预期	先保证核心机关、A 级文件；扩展机关和 C 级登记可以后补；分批冻结，先出中央＋省级的 v0.9
执行主体换成 Codex 后口径理解出现偏差	执行包是唯一依据；首次交接做理解测试；试点期双模型比对并逐条人工核验；Codex 结果单独过 G1
Codex 或 Claude 额度不足、运行中断	所有状态写盘，任何运行都可续跑；Claude 只用于抽检复核，额度按每批抽检量预留；优先保证核心机关和 A 级文件
早年文件下线找不到	公报目录、法规数据库、互联网档案馆；仍找不到的登记为“已知存在、原文缺失”

12.5 交付物清单

- SQLite 数据库与 CSV、xlsx 导出（每个版本一套）；
- 原件库、文本库、条款块库；

- 名录、站点、检索日志、线索、关系、抽检、问题单全部记录；
- 数据说明书：口径、来源、各层级覆盖与召回估计、已知局限、版本变更记录；
- 旧—新对照表。

附录 A

提示词模板

花括号为变量。所有提示词都要求输出结构化 JSON，枚举字段只能取字段字典中的值。提示词每次修改升版本号，并登记问题单。

P1　单元建档（A1）

【角色】你是政府信息检索员，负责为一个检索单元建立档案。
【输入】unit_id={unit_id}；单元名称={name}；上级={parent}；单元类型={kind}；机关模板=见附表。
【任务】
1. 核对该单元的现行名称和区划代码，查明 2015 年以来是否有撤并、更名、升格（写明时间和依据）。
2. 按机关模板逐一确认本单元实际存在的机关名称（如“数据局”在本地叫什么），列出其官网首页、政策文件栏目、站内检索入口；找不到的写“未找到”和你试过的入口。
3. 找到本级政府公报的目录页（如有）。
【规则】只记录你实际打开过的页面；不推测网址；遇到验证码或登录页立即停止该站点，记“需人工”。
【输出】JSON：{"unit_id","name_current","history":[...],"agencies":[{"name","role":"核心/扩展","home","policy_page","search_page","status"}],"gazette":"","issues":[...]}

P2　检索执行与登记（A2）

【角色】你是政策文件检索员。
【输入】单元档案 JSON；词表 K1（逐词）；路由表 R1—R7；本单元上级 A 级文件清单（用于 R6）。
【任务】按 R1→R7 的顺序执行。每执行一次检索（一个站点 × 一个词或一个栏目），在检索日志追加一行：路由、站点、检索式或栏目路径、时间范围设置、结果总数、实际浏览数、新增候选数。
【候选登记】凡标题或摘要含 K1 词、或标题像政策文件且可能含人工智能内容的，登记为候选：标题、发文机关、日期、链接、发现路由。只读标题无法判断时也登记，交后续环节判断。
【规则】
– 不设起始年份；站点默认只查近年的，改成全部时间并在日志注明。
– 结果超过 200 条时，按年份或栏目切分后分别浏览，不得只看前几页。
– 新闻、解读、动态页里提到的文件名，登记到线索清单，并尝试找到原文。
– 遇验证码、登录、访问拒绝：停止该站点，日志记“站点异常”，不要绕过。
– 不要因为“看起来和政务无关”而不登记；产业、教育、医疗等领域同样收。
【停止】适用的必做路由全部完成，且最后两条路由没有新增 A/B 候选。
【输出】search_log.csv（追加）；candidates.csv；clues.csv；一段 100 字以内的单元小结（含站点异常）。

P3　下载与完整性检查（A3）

【角色】你是文件存证员。
【输入】candidates.csv 中的一行。
【任务】
1. 按优先级获取原件：官网附件原件 > 官网网页正文 > 政府公报版 > 法律法规数据库 > 北大法宝 > 转载。
2. 保存：原件（含全部附件）、网页快照（HTML 与打印 PDF 各一份）、获取时间、最终链接。
3. 完整性检查：通知壳、截断、仅摘要、无文字层、附件多份、图片型附件（规则见完整性表）。
4. 计算每个文件的 SHA-256。
【规则】官网有原件时，不得用北大法宝件代替；北大法宝件出现“附件预览 无相关内容”必须回官网补附件。下载失败三次记“下载失败”并写原因。
【输出】meta_download.json：{"cand_id","files":[{"path","role":"正文/附件n/快照","format","sha256","pages"}],"source_url","source_type","completeness","problems":[...]}

P4　元数据抽取（A4）

【角色】你是公文元数据编目员，熟悉《党政机关公文处理工作条例》。
【输入】清洗后的正文全文；网页上的元数据（如有）；北大法宝元数据（如有）。
【任务】按字段字典抽取：title、title_notice、doc_number、issuers、issuer_type、legal_force、official_genre、policy_form、date_written、date_published、date_effective、period_start、period_end、validity、parent_doc_id 候选（上级依据文件的标题与文号）。
【规则】
– 来源冲突时以官方原件正文为准，其次官网网页元数据，再次北大法宝；把冲突写入 conflicts。
– 原文没有的字段填 null，不推断。日期只知年月时填 YYYY-MM-00。
– 印发型文件：title 取附件标题，title_notice 取通知标题，policy_form 按附件判断。
– 发文机关用原件上的规范全称，不用简称；联合发文全部列出，牵头机关在前。
– 枚举字段只能取给定值。
【输出】严格符合 JSON Schema 的对象；附 evidence：每个字段取自原文哪一句（逐字）。

P5　相关性研判（A5）

【角色】你是人工智能政策研究的编码员。
【输入】正文全文（或按条款切分后的全文）；元数据；相关性分级定义与 6 条判定规则；判例 20 条。
【任务】
1. 判定 relevance：A/B/C/X/待定。
2. 给出证据：逐字摘录 1—3 处最能支持判定的原文，并标明位置（章、节、条或段落序号）。
3. 判定 doc_class（去向）、ai_forms（技术形态）、domains（领域）、adjacent_field。
【规则】
– A：以人工智能为对象的文件；B：有以人工智能为对象的独立章、节、条、段或具体任务；C：仅列举或修饰；X：无核心词或“智能”仅指一般信息化。
– 判断依据是“是否有以人工智能为对象的独立任务”，不是出现次数。
– 依据只能来自本文件正文，不得引用你对该地政策的背景知识。
– 介于两级之间时选“待定”，写清楚卡在哪条规则。
【输出】{"relevance","evidence":[{"quote","loc"}],"doc_class","ai_forms":[],"domains":[],"adjacent_field","reason":"50字以内","confidence":"高/中/低"}

P6　去重与关系识别（A6）

【角色】你是政策文件库的编目审校员。
【输入】同一单元（或同一标题簇）的多条记录及其正文开头 1500 字、结尾 500 字。
【任务】
1. 判断哪些记录是同一文件的不同来源（同组），哪些是不同文件（如修订版、转发件、配套细则）。
2. 识别关系：转发、依据、配套、修订、废止、解读、名单（定义见关系表），写明证据句。
【规则】标题相同但文号或日期不同的，不得直接合并，判断是否为修订或年度文件；转发件单独成条。
【输出】{"groups":[[doc_id,...]],"relations":[{"from","to","type","evidence"}],"doubts":[...]}

P7　独立复核（A7）

【角色】你是独立质检员。你看不到原研判结果。
【输入】抽中的记录：原件文件与正文；字段字典；相关性分级规则。
【任务】从头独立填写关键字段（标题、文号、发文机关、发布日期、层级、效力类别、文件类别、相关性等级与证据）。
【规则】独立完成后才允许读取原记录并逐字段比对；不一致处写明谁对、依据哪句原文；无法判断的交人工。
【输出】{"doc_id","my_fields":{...},"diff":[{"field","orig","mine","who_is_right","evidence"}]}

P8　0 份单元审查（A7）

【角色】你是检索审计员。
【输入】一个结论为“0 份”的单元：单元档案、全部检索日志。
【任务】
1. 审查日志：必做路由是否全部完成、是否有站点异常未处理、时间范围是否为全部时间、结果多的检索是否翻完。
2. 换一种做法独立再查：用与原检索不同的来源（例如原以门户为主，则以北大法宝和公报为主），并做 R6 上级文件反查。
【输出】{"unit_id","log_ok":true/false,"log_problems":[...],"found":[候选...],"conclusion":"维持0份/推翻"}

附录 B

SQLite 建表语句

以下语句可直接执行，建出第 8 章的全部表。字段含义见附录 C。

```
PRAGMA foreign_keys = ON;
CREATE TABLE unit (unit_id TEXT PRIMARY KEY, level TEXT NOT NULL, unit_kind TEXT NOT NULL,
  name TEXT NOT NULL, admin_code TEXT, prov_code TEXT, parent_unit TEXT, region TEXT,
  sub_provincial INTEGER DEFAULT 0, batch_id TEXT, status TEXT DEFAULT '未开始',
  history TEXT, note TEXT);
CREATE TABLE agency (agency_id TEXT PRIMARY KEY, unit_id TEXT REFERENCES unit, name TEXT NOT NULL,
  category TEXT, priority TEXT, home_url TEXT, note TEXT);
CREATE TABLE site (site_id TEXT PRIMARY KEY, agency_id TEXT REFERENCES agency, site_type TEXT,
  url TEXT, verify_status TEXT, verified_at TEXT);
CREATE TABLE run (run_id TEXT PRIMARY KEY, agent TEXT, model TEXT, model_version TEXT,
  prompt_version TEXT, started_at TEXT, ended_at TEXT, target TEXT, note TEXT);
CREATE TABLE search_log (log_id TEXT PRIMARY KEY, unit_id TEXT REFERENCES unit, route TEXT,
  site TEXT, query TEXT, time_range TEXT, n_results INTEGER, n_browsed INTEGER,
  n_new_cand INTEGER, n_ab INTEGER, anomaly TEXT, run_id TEXT REFERENCES run,
  executed_at TEXT, evidence_path TEXT, note TEXT);
CREATE TABLE candidate (cand_id TEXT PRIMARY KEY, unit_id TEXT REFERENCES unit, title TEXT,
  issuer TEXT, date_hint TEXT, url TEXT, route TEXT, log_id TEXT REFERENCES search_log,
  outcome TEXT, doc_id TEXT, reason TEXT);
CREATE TABLE document (doc_id TEXT PRIMARY KEY, unit_id TEXT REFERENCES unit,
  title TEXT NOT NULL, title_notice TEXT, doc_number TEXT, issuers TEXT NOT NULL,
  issuer_n INTEGER, issuer_type TEXT, level TEXT NOT NULL, unit_kind TEXT, prov_code TEXT,
  admin_code TEXT, legal_force TEXT, official_genre TEXT, policy_form TEXT, doc_class TEXT,
  date_written TEXT, date_published TEXT, date_effective TEXT, period_start INTEGER,
  period_end INTEGER, validity TEXT, relevance TEXT, relevance_evidence TEXT,
  ai_hits INTEGER, ai_forms TEXT, domains TEXT, adjacent_field TEXT, parent_doc_id TEXT,
  source_url TEXT, source_type TEXT, retrieved_at TEXT, route TEXT, file_format TEXT,
  sha256 TEXT, pages INTEGER, chars INTEGER, ocr INTEGER DEFAULT 0, completeness TEXT,
  dup_group TEXT, version_of TEXT, prehistory INTEGER DEFAULT 0, status TEXT,
  agent_id TEXT, checker TEXT, conflicts TEXT, note TEXT);
CREATE TABLE doc_agency (doc_id TEXT REFERENCES document, agency_name TEXT, seq INTEGER,
  is_lead INTEGER, PRIMARY KEY (doc_id, seq));
CREATE TABLE source (source_id TEXT PRIMARY KEY, doc_id TEXT REFERENCES document,
  source_type TEXT, url TEXT, retrieved_at TEXT, is_primary INTEGER, note TEXT);
CREATE TABLE file (file_id TEXT PRIMARY KEY, doc_id TEXT REFERENCES document, role TEXT,
  path TEXT, orig_name TEXT, format TEXT, sha256 TEXT UNIQUE, pages INTEGER, bytes INTEGER);
CREATE TABLE relation (rel_id TEXT PRIMARY KEY, from_doc TEXT REFERENCES document,
  to_doc TEXT, rel_type TEXT, evidence TEXT, found_by TEXT, checked INTEGER DEFAULT 0);
CREATE TABLE chunk (chunk_id TEXT PRIMARY KEY, doc_id TEXT REFERENCES document,
  seq INTEGER, path TEXT, text TEXT, chars INTEGER, k1_hit INTEGER);
CREATE TABLE clue (clue_id TEXT PRIMARY KEY, unit_id TEXT, clue_type TEXT, title TEXT,
  url TEXT, mentioned_title TEXT, found INTEGER, doc_id TEXT, note TEXT);
CREATE TABLE qc_sample (sample_id TEXT PRIMARY KEY, batch_id TEXT, sample_type TEXT,
  target_id TEXT, seed INTEGER, assigned_to TEXT, status TEXT);
CREATE TABLE qc_result (sample_id TEXT REFERENCES qc_sample, field TEXT, db_value TEXT,
  checked_value TEXT, result TEXT, error_type TEXT, basis TEXT, checker TEXT,
  checked_at TEXT, written_back INTEGER DEFAULT 0);
CREATE TABLE issue (issue_id TEXT PRIMARY KEY, raised_at TEXT, raised_by TEXT, target TEXT,
  issue_type TEXT, severity TEXT, description TEXT, resolution TEXT, owner TEXT,
  status TEXT, closed_at TEXT);
CREATE INDEX ix_doc_unit ON document(unit_id);
CREATE INDEX ix_doc_rel ON document(relevance, doc_class);
CREATE INDEX ix_log_unit ON search_log(unit_id);
```

附录 C

字段字典

document 主表的全部字段。效力类别取值：法律、行政法规、部门规章、地方性法规、地方政府规章、党内法规、党委规范性文件、行政规范性文件、其他政策文件、标准、司法文件。政策形式取值：条例、规定/规则、办法、细则、规划/纲要、意见/指导意见/实施意见、方案/实施方案/行动方案、计划/行动计划、工作要点、若干措施/若干政策、指南/指引、清单/目录、征集/申报通知、名单/公示、决定、通知（其他）、标准/规范、其他。

字段	中文名	类型	必填	取值/格式	规则
doc_id	文件编号	文本	是	层级码+区划码6位–年份–流水4位	层级码：C 中央，P 省级，M 第三层，X 框外；年份取发布日期，未知填 0000
unit_id	检索单元编号	文本	是	见名录表	文件归属的检索单元；联合发文归牵头机关所属单元
title	标题（规范全称）	文本	是		取官方原件首页标题；印发通知型记附件标题，通知标题记入 title_notice；去掉书名号外的修饰
title_notice	印发通知标题	文本	否		印发型文件的通知标题
doc_number	发文字号	文本	否	机关代字〔年份〕序号	统一用六角括号〔〕；无文号填“无”而不是留空
issuers	发文机关（全部）	文本	是	规范全称，分号分隔	按原件落款或版头顺序；第一个为牵头机关
issuer_n	联合发文机关数	整数	是	≥1	
issuer_type	牵头机关类型	枚举	是	党委/人大/政府/政府办公厅/部门/议事协调机构/功能区/党政联合/其他	
level	层级	枚举	是	L1 中央/L2 省级/L3 第三层/L4 框外	按牵头机关所属政府
unit_kind	单元类型	枚举	是	中央机关/省级/地级市/地区/自治州/盟/直辖市区县/省直管单元/兵团	
prov_code	省级区划码	文本	是	2 位	中央填 00
admin_code	单元区划码	文本	是	6 位	中央填 000000
legal_force	效力类别	枚举	是	见“效力类别”列表	
official_genre	公文文种	枚举	否	15 种法定文种	按文件自称；印发型记“通知”
policy_form	政策形式	枚举	是	见“政策形式”列表	按附件标题末尾的形式词
doc_class	文件类别（去向）	枚举	是	主库/主库–实施类/主库–转发/辅库–标准/辅库–政府工作报告/辅库–司法/线索库/排除	见文种与去向表

字段	中文名	类型	必填	取值/格式	规则
date_written	成文日期	日期	否	YYYY-MM-DD	落款日期
date_published	发布日期	日期	是	YYYY-MM-DD	网站发布日期；与成文日期不同时两者都填；只知年月填 YYYY-MM-00
date_effective	施行日期	日期	否	YYYY-MM-DD	
period_start	规划/实施期起	整数	否	年份	如 2025—2027 年行动计划填 2025
period_end	规划/实施期止	整数	否	年份	
validity	时效状态	枚举	是	现行有效/已修订/已废止/失效/未知	依据官网“有效性”标注或后续废止文件
relevance	相关性等级	枚举	是	A/B/C/X/待定	见相关性分级
relevance_evidence	相关性证据	文本	A/B 必填	逐字引文 | 位置	1—3 处原文，标明章节或条款编号
ai_hits	核心词命中数	整数	是		K1 词在正文中出现次数（自动统计）
ai_forms	技术形态标签	多选	A/B 必填	通用人工智能/大模型/生成式/智能体/机器学习与视觉语音/具身智能与机器人/算力/算法治理/数据	
domains	领域标签	多选	A/B 必填	产业发展/科技研发/政务服务/政府治理与监管/公共安全/医疗卫生/教育/交通/农业/制造/金融/文化传媒/司法/社会民生/伦理与安全治理/人才就业/基础设施与算力/数据要素/其他	可多选，按独立任务所在领域
adjacent_field	邻近领域	枚举	否	智能制造/智能网联汽车/机器人/算力/数据要素/智慧城市/数字政府/无	主题属于邻近领域时填写
parent_doc_id	上级依据文件	文本	否	doc_id，分号分隔	正文明确写“根据/为贯彻落实××”的上级文件
source_url	官方原件链接	文本	是	http(s)://	优先官网原件页；无官网版本时填实际来源并在 source_type 标明
source_type	来源类型	枚举	是	官网原件/官网网页正文/政府公报/法律法规数据库/北大法宝/其他数据库/转载	
retrieved_at	获取时间	日期时间	是	YYYY-MM-DD HH:MM	
route	发现路由	枚举	是	R1—R8	首次发现该文件的路由
file_format	原件格式	枚举	是	pdf/ofd/doc/docx/wps/html/图片	
sha256	原件哈希	文本	是	64 位	主原件文件的 SHA-256

字段	中文名	类型	必填	取值/格式	规则
pages	页数	整数	否		
chars	正文字数	整数	是		清洗后正文字符数
ocr	是否 OCR	布尔	是	0/1	
completeness	全文完整性	枚举	是	完整/缺附件/截断/仅摘要/待核	见完整性检查
dup_group	重复组	文本	否		同一文件的不同来源共享组号
version_of	修订自	文本	否	doc_id	新版本指向旧版本
status	记录状态	枚举	是	候选/已下载/已解析/已研判/自检通过/抽检中/已核验/返工/冻结	见状态机
agent_id	处理智能体	文本	是		智能体名称、模型与版本
checker	核验人	文本	否		抽中时填写
note	备注	文本	否		

附录 D

配套表单说明

《中国政府人工智能政策库建设–工作表单.xlsx》共 29 张工作表。正式数据以 SQLite 为准，表单用于名录管理、进度跟踪、人工核验和小批量登记。

工作表	内容
00 封面与导航	使用约定与目录
01 层级与单元总数	层级定义与单元数
02 中央检索单元	66 个中央机关单元及进度列
03 省级检索单元	32 个省级单元
04 第三层–地级单元	333 个地级单元，含旧库文件数
05 第三层–直辖市区县	86 个区县
06 第三层–特殊单元	32 个省直管县级单元
07 机关名录模板	省级、第三层要查的机关
08 检索词表	K1/K2/K3
09 检索路由	R1—R8 与停止规则
10 来源站点清单	全国性来源与各省站点登记
11 相关性分级	A/B/C/X 与判定规则
12 文种与去向	八个去向
13 判例	J01—J20，可追加
14 字段字典	主表字段
15 文件登记表(模板)	主表 Excel 版，带下拉
16 检索日志(模板)	每次检索一行
17 线索表(模板)	新闻、解读等
18 关系表(模板)	八类关系
19 抽检方案	比例、合格线、批次判定
20 抽检记录(模板)	逐字段核验
21 问题单	P1—P3
22 已知项回收	召回测试清单（待核）
23 旧库诊断与迁移	673 个文件逐个诊断
24 自检规则	V01—V18
25 状态机与验收门	G0—G4
26 智能体与提示词	角色与 P1—P8 全文
27 进度看板	自动统计
28 执行环境检查	Codex 开工前的 20 项检查

术语

术语	含义
检索单元	检索、记录和验收的最小单位；中央为机关，省级和第三层为政府
第三层	省级之下第一层：地级行政区、直辖市区县、省直管县级单元
主库	A、B 级政策文件（含实施类、转发），研究分析的分母
路由	一种检索来源与做法的组合，R1—R8
候选	检索中登记、尚未下载和研判的文件
通知壳	只有印发通知、缺少附件正文的文件
已知项	事先确认存在、用于测试召回的文件
双路检索	两个智能体用不同来源独立检索同一单元，用于召回估计
冻结	数据版本定稿只读，后续修改进入新版本