基于检索增强生成（RAG）技术的上海云宇星空大模型知识库建设探索*

Exploration of Knowledge Base Construction for Shanghai Cloud Cosmos Large Model Based on Retrieval-Augmented Generation (RAG) Technology

徐 玮 Xu Wei

摘 要 为提升超大城市空间治理精细度，上海启动量子城市工作并积极探索研发规划资源行业基座大模型——云宇星空。以云宇星空大模型为研究对象，从检索增强生成（RAG）技术应用的全流程总结规划资源行业模型的创新要点。面对超大城市空间治理的多元需求，云宇星空大模型建立五步法知识工程建设流程；通过专题领域知识图谱建设和材料类型分类、规范标签体系等开展高质量数据集建设；语料生产和模型检索增强生成环节，以知识图谱叠加多路召回的方式，提升模型召回率和精准率；并依托语料动态管理平台实现语料从收集到生产入库及动态扩容的闭环管理。后续拟在模型空间智能能力深度拓展、知识图谱与RAG融合、多智能体协同的RAG架构等方面进一步提升完善。

Abstract To improve the precision of spatial governance in megacities, Shanghai has launched the Quantum City initiative and is actively researching and developing a foundational industry model for planning and resources—the Cloud Cosmos. Taking the Cloud Cosmos model as the research object, this paper summarizes the innovative features of the planning and resources industry model from the perspective of the full-process application of RAG technology. In response to the diverse needs of megacity spatial governance, the Cloud Cosmos model has established a five-step knowledge engineering process. It has developed high-quality datasets through the construction of domain-specific knowledge graphs, material type classification, and standardized tagging systems. In the stages of corpus production and model RAG, a combination of knowledge graphs and multi-path recall is employed to improve the model's recall rate and precision. Additionally, a dynamic corpus management system enables closed-loop management of the corpus from collection and production to database entry and dynamic expansion. Future work is planned to further enhance and refine the model in areas such as deep expansion of spatial intelligence capabilities, deep integration of knowledge graphs and RAG, and multi-agent collaborative RAG architectures.

关 键 词 检索增强生成；云宇星空大模型；知识库建设；空间智能；规划资源行业

Key words Retrieval-Augmented Generation (RAG); the Cloud Cosmos Large Model; knowledge base construction; spatial intelligence; planning and resources industry

文章编号 1673-8985（2026）04-0016-07 中图分类号 TU984 文献标志码 A DOI 10.11982/j.supr.20260403

作者简介

徐 玮

上海市数字城市规划研究中心

副主任，高级工程师，硕士

上海市量子城市空间智能创新重点实验室（筹）

副主任，大模型方向（语料建设）带头人

85367093@qq.com

0 引言

人工智能在算力、数据与技术的推动下飞速发展，基础大模型在通用领域表现优异，但在垂直领域，仍面临专业知识欠缺、数据隐私欠安全、模型幻觉等难题。当前解决方案主要分为微调预训练模型与检索增强生成（Retrieval-Augmented Generation，以下简称“RAG”）两类[1]2。其中，RAG依托外挂知识库，可以规避隐私风险、扩充知识范围，并能够访问实时更新的数据库或互联网资源，解决纯生成模型在准确性和时效性方面的不足[2]，是行业模型研发中广泛使用的技术手段。

云宇星空大模型由上海市规划和自然资源局发起，是针对规划资源领域业务数据繁杂、专业门槛高、决策支撑繁复等行业痛点而打造的行业基座大模型，于2025年12月正式

*基金项目：上海市人民政府决策咨询研究重大课题“上海深化量子城市空间治理大模型建设思路研究”（编号2026-A-19）资助。

《上海城市规划》https://www.shplanning.com.cn/. All rights reserved.

发布专业版。模型基于海量高质量“文本—图像—空间”多模态语料，构建城市空间治理领域全链路认知与决策能力，形成集“时空理解生成、知识整合检索、模型智能调度”于一体的技术底座，具备“问不倒、能调图、会统计、能识图、会报告”5大核心专业能力[3]。现阶段模型主要采用RAG技术，通过大规模、高质量可信数据的采集，打造专业知识库，以此解决行业模型知识来源有限的问题。

1 研究背景

为适应超大城市人口密度激增、空间结构复杂、空间资源不足、治理需求多元等发展挑战，上海推进量子城市空间治理行动。2026年2月，《上海市国民经济和社会发展第十五个五年规划纲要》将“深化量子城市空间智能建设、打造城市治理基础大模型”正式写入全市战略任务[4]。

量子城市空间治理行动是上海借助当前量子科技快速发展所带来的巨大引擎作用和颠覆式外溢效应，集中探索超大城市智能治理的实践行动[5]，依托精细化智能感知和人工智能大模型体系的强大模型能力，构建城市数据采集、分析、推演、决策的循环逻辑，改变传统城市规划的大尺度空间治理模式，将城市分析的精度提升到“米料”①级别，重塑城市空间治理范式。其中，云宇星空大模型是整个空间治理框架的核心底座。该模型以国产多模态大模型为基础，通过专业领域知识库构建和特定领域微调，构建城市运行和治理智能中枢。模型充分发挥RAG技术的优势，将知识外置并支持版本管理，能够在政策更新频繁的业务环境中快速迭代知识[6-7]，具有支持模型在复杂场景中的实用性[8]等特点，从而进行创新探索。

2 RAG技术的演进

RAG技术自2020年由Lewis等人正式提出[9]，相关研究可以分为两类。一是优化检索机制，致力于改进模型的信息检索性能，实现从大量信息中快速且高效的检索[10-12]；二是增强生成模型，试图减少模型幻觉[13]、优化检索与生成过程之间的协同作用[1]2。从行业应用看，RAG技术已在金融、法律、医疗等专业领域广泛落地，但在城市规划、自然资源管理等空间智能领域，尚缺乏系统性探索。

学科发展层面，云宇星空大模型首次将RAG技术体系化应用于规划资源这一高度专业化的空间治理领域，拓展了技术的应用边界。行业发展层面，云宇星空大模型的RAG实践为规划资源行业智能化转型提供了可落地的解决方案，模型通过构建高质量行业语料基座，系统整合全领域专业知识；同时建立随业务动态迭代的更新机制，确保了知识时效性，这些探索对于推动规划资源行业从“经验驱动”向“数据驱动”、从“人找数据”向“数据智能响应”转变具有重要示范价值。

3 基于RAG技术的大模型知识库建设

3.1 构建知识库建设顶层设计

作为行业基座模型，云宇星空大模型旨在构建全面、精准的行业知识体系，以支撑城市治理中的复杂决策，并提升公共服务响应效率与精确性[14]。为明确知识库的建设主体与实施路径，从工作机制与工作流程两个维度开展顶层设计。

建立政府牵头、技术引领、市场参与的工作机制。由上海市规划和自然资源局牵头，联合城市建设和治理领域相关业务主管部门，共同构建以规划资源知识为主体、其他城市治理领域知识相融合的多源知识体系，提升知识覆盖范围、可信度与专业水平。工作由业务团队、语料厂商、模型厂商构成的核心专班统筹推进，广泛联合高校、规划设计团队、市政工程设计团队、测绘团队等市场主体。

建立标准化工作流程。将知识库建设流程规范为5个标准环节：知识体系构建、材料收集、语料生产、入库调优、动态扩容。知识体系构建环节以业务需求为驱动，全面梳理行业知识架构，确保业务领域全覆盖；材料收集环节明确材料类型、格式与质量要求，通过质量检验机制保障材料质量；语料生产环节以提升模型召回率为目标，针对不同模态材料开展分类语料生产，确保语料质量；入库调优环节以实际应用为导向，引入专业人员进行内部测试，针对性开展模型调优；动态扩容环节实现语料的常态化自动更新。各环节均明确工作内容、质量要求及标注规范，确保语料的一致性和高质量，支撑模型稳定输出。

3.2 构建以业务为驱动的知识体系

系统梳理各领域知识体系，确保知识内容在宏观与微观、抽象与具体等维度上实现全覆盖与精细划分。全面梳理各自业务板块共同构成“两级两类”知识体系。“两级”包括“大领域”与“细分专题领域”，目前形成约80个细分专题领域（见表1）。“两类”包括业务领域与综合管理领域。业务领域涵盖规划、土地、项目、登记、执法、测调等规划资源业务管理全流程；综合管理领域涵盖办公、人事、党建、财务、法制、行政服务等专题。

以知识体系为框架组织材料收集。充分融合管理部门、高等院校及业务单位的多方视角，将材料分为管理、学术与实务三元体系。其中，管理类包括政策法规、技术标准、管理规

表1 规划资源“两级”领域示例表（节选）
Tab.1 List of two-level domains in planning and resources (excerpt)

大领域	细分领域
区域规划	跨省区域规划
总体规划	总体规划
	总规评估
专项规划	重点地区规划
	专项规划
控制性详细规划与城市设计	控制性详细规划
	城市设计
城市更新与社区规划	城市更新
	15分钟社区生活圈
建筑工程建设项目	优化营商环境
	建筑品质
	规划实施平台
	项目审批
公共空间艺术	上海城市空间艺术季
	城市雕塑管理
风貌保护与地名管理	风貌保护管理
	地名管理

资料来源：笔者自制。

注释：①“米料”源于人工智能大语言模型“语料”的类比迁移，被定义为“城市空间每个立方米的语料”，这里的每个立方米为虚指，以此代表空间语料的概念。“米料”以房屋建筑的“间”、城市空间的“盒”、市政管网的“段”为载体，将城市空间治理尺度从传统功能区块向下细化2—3个层级，实现从“宏观分区”到“微观单元”的范式转换。依托“间”“盒”“段”等空间载体，叠加“空（时空位置）、形（几何形态）、量（规模体量）、质（属性特征）、联（关联组合）、智（动态感知）”6大维度全息信息，使“米料”成为训练空间智能大模型的基础语料，也是城市空间治理的最小单元。

《上海城市规划》https://www.shplanning.com.cn/. All rights reserved.

则、管理材料及审批成果；学术类包括书籍与期刊论文；实务类包括实务案例、技术手册等（见表2）。与预训练模型所需的海量语料不同，知识库建设更强调知识的有效性与精准性，重质不重量，每个专题知识库归集约150份精选材料。目前80个细分专题领域共归集逾两万份材料，累计形成约15亿词元语料。

3.3 材料收集强调标准化和高质量

通过标准化标签设置和多维度质检开展主动高效的材料收集，为知识复用与创新构建可信的知识资产。

针对材料设置标准化元数据通用标签和专项标签。通用标签包括基础信息、时空信息、类型信息、管理信息。基础信息规范材料基础属性，包括唯一序号、文件名；时空信息记载材料的生产/生效时间和适用的空间范围，助力模型时空理解能力的提升；类型信息从专题领域和材料类型两方面规范材料构成；管理信息明确材料的采集部门、审核部门和向公众开放的属性，便于材料溯源，并可针对模型公众版和专业版匹配提供不同版本的语料（见表3）。专项标签主要针对政策法规和技术标准，增加文号/标准号、发布机构/批准发布机构、发布日期、施行日期/生效日期、失效日期等元数据，便于开展版本管理，帮助模型准确按时间调用历史版本材料，定位权责主体（见表4）。

文件质检是以自动质检加人工校核的方式，对材料元数据信息和材料内容进行质量检查，确保原始材料的准确性、安全性、一致性。由规划资源专家进行质检规则梳理，并转化成质检工具后开展自动质检，对于报错问题经人工判断后予以修正完善。

针对元数据信息，质检内容包括标签的完整性、规范性和准确性。完整性关注元数据字段是否存在填写遗漏，确保大模型使用中相关信息全面、可溯源。规范性通过形成填写规范和数据字典统一表达（见表5）。如文件名规范性有助于查找重名文件，减少算力消耗；时间规范性有助于梳理政策法规版本关系，提升模型按时间调用的精准度。准确性有助于模型按需调用；材料类型归类准确，有助于模型针对类型差异选择不同的生成方式，云宇星空大模型针对法规和标准类材料均采用了原文作答方式，而其他类型材料则允许充分发挥模型的生成能力。

针对材料内容，质检项包括涉密、涉敏、内容质量检查。涉密检查避免材料中有秘密、机密、绝密、密级、机要的内容；涉敏检查对于存在个人身份、联系电话、住址等隐私信息进行提醒；内容质量检查主要对小于20 k的材料、文件大小完全一样的材料及存在解码问

表2 材料类型一览表
Tab.2 List of material types

大类	中类	小类
管理类	政策法规	法律、行政法规、部门规章、地方性法规、地方政府规章、行政规范性文件
	党内法规	党内法规
	技术标准	国家标准、地方标准、地方规程、行业标准、团体标准
	管理规则	一般管理文件、操作口径、办事指南、重要概念释义
	管理材料	管理案例、其他
	审批成果	大机成果、档案成果
学术类	书籍	书籍
	期刊论文	期刊论文
实务类	实务案例	科研课题、研究报告、获奖项目、方案征集、其他
	技术手册	技术手册
	其他	其他

资料来源：笔者自制。

表3 通用标签一览表
Tab.3 List of general tags

标签类型	标签内容	
基础信息	序号	
	文件名	
时空信息	时间	
	适用范围	
类型信息	专题领域	大领域
		细分领域
	材料类型	大类
		中类
		小类
管理信息	采集部门	
	审核部门	
	是否向公众开放	

资料来源：笔者自制。

表4 专项标签一览表
Tab.4 List of dedicated tags

标签类型	标签内容
政策法规	文号
	发布机构
	发布日期
	施行日期
	失效日期
技术标准	标准号
	技术归口单位
	批准发布机构
	发布日期
	生效日期
	失效日期

资料来源：笔者自制。

表5 标签规范性质检规则（节选）
Tab.5 Quality inspection rules for tag standardization (excerpt)

字段名称	填写要求
文件名	1.符号：首尾不应加《》；《》、〈〉需各自成对出现；〈〉不可单独出现，必须搭配《》同时出现 2.文件名里不应出现文号 3.文件名里不应出现日期前缀或后缀，如20260523、202605、2026.05.23、2026.05 4.文件名里不应出现回车符 5.文件名里不应出现格式后缀，如.pdf、.doc、.xlsx等 6.文件名里不应出现“征求意见稿”“过程稿” 7.文件名未使用全称：当文件名中的数字不纳入统计时，文件名长度应大于4个字。如“东京2040”，文件名长度2个字，判为不合格
时间	x年x月x日；x年x月
适用范围	全国；x省；x市；x市x区；特定范围（需提供范围描述的专有名词）
专题领域	对应关系符合《规划资源两级专题领域一览表》
材料类型	对应关系符合《材料类型一览表》

资料来源：笔者自制。

《上海城市规划》https://www.shplanning.com.cn/. All rights reserved.

题的材料做出提醒，前者可能存在知识密度太低，中间者可能存在内容重复，后者可能存在文件错误导致无法加工为语料。

3.4 以提高召回率为目标的语料生产

文本、图纸和表格等多模态信息构成了规划资源领域的核心知识承载，以提高多模态数据召回率②为主要目的，探索适合云宇星空大模型的语料生产方式。

3.4.1 文档材料的多路召回

RAG技术的核心是根据问题向量相似性在知识库中检索出最相似的若干个文档片段，并将这些片段注入大模型的输入中，大模型基于注入的上下文与原始问题进行语义理解、信息整合与推理，最终生成自然语言答案[15]。为提高召回率，云宇星空大模型尝试在一般的稠密向量检索（向量相似度）和稀疏向量检索（关键词）的基础上，进行多路召回，配合这一调用方式，开展了多层次的语料加工。

以图1的政策文件语料切片为例，成果包括正文、目录层级、关键词、摘要、问题5部分。正文（content）即原文切片，尽可能保持原文基本特征；目录层级（level）明确切片所在材料的目录层级，支持多级目录标注；关键词（keywords）针对切片提取3个涉及业务的高频词作为关键词；摘要（summary）是对文本切片的总结，尽可能准确精炼地描述文段。问题（question）是针对当前切片做提问，列举若干可能的问题，可用于模型针对相似问题召回文档片段；上述内容均可用于多路召回。经80个专题约9 300个问题测试，模型在单一稠密向量检索时，召回率仅为46%；稠密向量检索和稀疏向量混合检索时，召回率提高至53%；进一步融合其他项召回后，召回率提升至90%以上。

此外，采用“机器抽取+人工校验”的半自动化方式构建知识图谱，包括基于大领域和细分专题领域生成专题领域知识图谱；面向部分场景应用，建立业务流程和规划管理逻辑类知识图谱；针对政策法规和重大政策文件，建设内容关联性和版本关系图谱。通过知识图谱叠加多路召回，提升模型调用的效率和准确性。如图2中，针对约2万篇中央、本市重大政策文件，构建分节点关系模式。政策文件的节点包括话语论述、中央政策、部委政策、本市政策4类，节点大小映射关键词出现频次与网络中心度；关系包含理论指导（红线）、层级隶属（粗黄线）、任务承接（绿线）、政策协同（蓝线）4类语义关系。话语论述作为理论顶层作用于中央政策；中央政策一方面通过大量协同引用关系形成国家政策网络，另一方面通过层级隶属关系向下映射至部委政策；部委政策进一步向下传导至本市政策。知识图谱可清晰呈现各类重大政策和治理要求从顶层理论、国家政策到地方落地实施的完整语义链路。

3.4.2 表格公式的定制算法开发

面向使用需求对表格进行定制算法开发。以政务服务办事指南为例，千余项行政服务事项的办理要求由若干巨型表格组成，业务人员对表格逻辑进行关联、拆解，明确表格查询的思维链，开发人员根据思维链进行生产算子定制化开发。经对80个专题中表格材料的10%抽样测试，定制化开发后产出的语料对比原表格还原度达96%。

针对复杂的数学公式，采用文本化处理方式，经算法定制支持通过Html准确还原公式，帮助模型理解复杂公式。

3.4.3 专业图纸的多路召回

图纸方面，在图名、图面文字信息的基础上，通过关联文字生产图文语料使图纸附带更多的规划信息，增加向量匹配度，提高召回率。如表6中，将南汇新城绿环空间结构图关联规划说明书，自动生产图文语料，补充时间、关键词、规划内容等信息。模型借用以上信息对图纸进行多维向量匹配，帮助模型在输出文字时同步输出配套图纸，提升模型专业度。

3.5 业务场景驱动的测试与调优

以业务场景为驱动，构建自动评测与人工评测相结合的测试体系。自动评测主要使用基于文档切片自动生成的问答对进行测试，侧重对于单一语料片段准确召回的评测。人工评测主要使用专家人工编写的问答对进行测试，

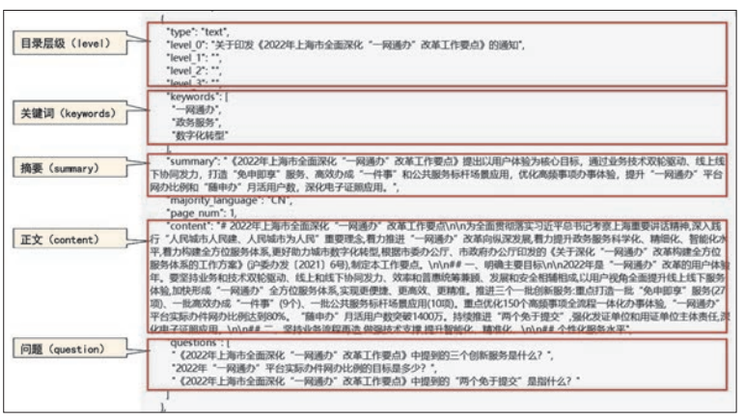


图1 语料切片生产成果示例
Fig.1 Example of corpus chunking production results
资料来源：笔者自绘。



图2 中央、本市重大政策文件知识图谱示例
Fig.2 Knowledge graph example of major central-municipal policy documents
资料来源：笔者自绘。

注释：② 召回率（ID recall）：基于文档名的检索召回率，判断是否准确召回希望召回的材料。

《上海城市规划》https://www.shplanning.com.cn/. All rights reserved.

侧重对于模拟现实提问的综合检索生成能力的评测。

3.5.1 自动评测

在自动生产问答对中预留10%作为自动评测集，评测语料片段的召回率、精准率[③]。当召回率大于95%、精准率大于75%时，专题知识库可进入人工评测流程（见表7）。

3.5.2 人工评测

业务专家针对每个专题人工撰写约100个测试问答对作为人工评测集。人工测试集重点聚焦业务应用情境，从覆盖业务需求、融入多元主体视角、侧重系统综合能力等角度构建评测框架，并编制评测问答集（见表8），评测模型在面对真实业务问答时的检索与生成两方面综合能力。

业务人员用测试问题集在模型测试版中逐一提问，对答案进行赞踩评价，并进一步补充问题“应召回的文档名称、文档片段、答题关键点”等。模型厂商通过大模型过程追踪平台（Langfuse）中的追踪（tracing）记录，对每个专题知识库的召回率、精准率、事实性[④]、正确性[⑤]进行评价，并综合给出大模型评分[⑥]。经80个专题约9 300个人工测试问题测试，经多路召回后，业务人员给出的点赞率和模型相关评测指标都达到较高水平（见表9）。

考虑到80个专题分3批入库，为评测新入库专题对已入库专题的影响，在每个已入库专题的人工测试问答对中预留10对回答正确的问答对（共计500个）作为回归测试集，新专题入库时，经模型测试和人工审核，回归测试需达到95%准确率，方可上线。

4 以联动平台实现知识库的系统管理和动态更新

批量梳理解决的是存量知识接入大模型的问题，随业务管理实现新增知识和数据的动态接入是知识库建设的重要议题。搭建语料动态管理平台（见图3），通过构建材料管理和语料中枢两大功能，实现业务材料自动进入材料归集、语料加工、知识入库的全流程闭环。

材料管理功能支持知识采集质检、存储更新、元数据管理、检索统计等需求。通过一键导入实现数据高效汇聚，满足各专题的首次导入与动态更新，大幅提升材料交付效率。通过实时互通实现协同联动强化，采用消息推送机制确保材料变更消息自动同步至材料审核和使用部门，为多单位共建共享提供坚实支撑。通过全程追踪实现流程闭环透明，材料上传后自动进入语料加工与入库流程，全生命周期状态实时可见，确保各单位随时掌握建设进度。

表6 图文语料生产示例

Tab.6 Example of text-image corpus production

文档	对应图纸
"type":"meta", "origin_file_name":"6-图1-南汇新城绿环空间结构图.jpg", origin_file_path":"6-南汇新城绿环专项规划/6-图1-南汇新城绿环空间结构图.jpg", "时间":2023年2月 "适用范围":上海市浦东新区 "优先级":1, 是否共享实训场":"是", "type:imagelevel_0":"南汇新城绿环专项规划:6-图1-南汇新城绿环空间结构图", "level_1":" 南汇新城绿环专项规划", "level_2":"6-图1-南汇新城绿环空间结构图", "keywords":["南汇新城" "空间结构" "空间布局" "绿环"], "content":"南汇新城绿环专项规划6-图1-南汇新城绿环空间结构图\n南汇新城绿环构建了一环五段十单元十二寻"的空间结构，将全环划分为西部森林漫游段、西南科教水廊段、南部滨海文化公园段、东部滨海生态公园段和北部运河绿廊段等五大功能段，环绕主脉形成凤舞林菲、水镜杉影、申港彩林、书香涵逸、秋水观澜、绿苇野径、海角浣尘、日升霞冠、白鹤仙踪、露泽栖鹃、静练水幕、花语如诉等十二个空间节点。" "summary":"南汇新城绿环构建了"一环五段十单元十二寻"的空间结构，将全环划分为西部森林漫游段、西南科教水廊段、南部滨海文化公园段、东部滨海生态公园段和北部运河绿廊段等五大功能段，环绕主脉形成凤舞林菲、水镜杉影、申港彩林、书香涵逸、秋水观澜、绿苇野径、海角浣尘、日升霞冠、白鹤仙踪、露泽栖鹃、静练水幕、花语如诉等十二个空间节点。" "questions":["南汇新城绿环的空间结构是什么样的？"]	

资料来源：笔者自制（表中图纸来自《南汇新城绿环专项规划》，上海市规划和自然资源局，2023年1月）。

表7 自动评测阶段模型测试数据（以第一批10个专题知识库为例）

Tab.7 Test data of model automatic evaluation (taking the first batch of 10 thematic knowledge bases as an example)

专题名称	模型自动评测指标	
	召回率/%	精准率/%
15分钟社区生活圈	96.60	77.32
沪派江南	96.40	80.51
城市总体规划	97.80	75.23[a]
品质市政	95.41[b]	78.58
城市设计	98.67	74.13
风貌保护管理	94.33	77.93
存量产业用地盘活	98.00	76.54
土地供给	97.00	79.18
生态复合单元和湿地空间	99.50	79.01
优化营商环境	96.30	77.97

注：a表示表中所列数据与其他专题非同期数据，同期数据为74.89%，后经模型调优后精准率提升至75.23%，本专题方进入人工评测。b表示表中所列数据与其他专题非同期数据，同期数据为92.6%，后经模型调优后召回率提升至95.41%，本专题方进入人工评测。

资料来源：笔者根据第一批专题知识库模型自动测试结果整理。

表8 人工测试问题类型及示例（以15分钟生活圈专题为例）

Tab.8 Test question types and examples (taking the thematic topic of 15-minute living circles as an example)

提问方向	序号	具体问题示例
覆盖业务需求	1	编制社区规划的一般流程是怎样的？
	2	上海市不同区的社区规划师制度有什么差异？
融入多元主体视角	3	上海有哪些公共服务示范性点位？
	4	市民可以通过哪些途径参与到社区生活圈的营造工作中？
侧重系统综合能力	5	社区规划师制度的演进与发展是怎样的？
	6	2025年上海市“15分钟社区生活圈”行动方案中提出的“多格合一”是什么意思？其作用意义是什么？

资料来源：笔者根据测试问题集整理。

注释：③ 精准率（ID precision）：基于文档名的检索精准率，判断是否召回了无需召回的材料。

④ 事实性（faithfulness）：基于召回的文档内容块来回答，评判大模型回答是否依据事实来回答，而不是随意发挥。

⑤ 正确性（keypoint answer correctness）：回答内容是不是已经包含了期望内容里的关键点。

⑥ 大模型评分（correctness）：回答内容与期望内容的匹配程度，交由大模型进行评判打分。

《上海城市规划》https://www.shplanning.com.cn/. All rights reserved.

语料中枢功能支持平台从后端管理转变为向采集端和生产端延伸的管理中枢。在采集端，平台与规划资源“一厅八室”信息管理平台⑦（以下简称“一厅八室”）深度联动，破除业务办公与知识储备体系间的数据隔阂，构建闭环式知识更新体系，解决知识库内容滞后、脱离业务实际的问题。通过标准化数据接口，在“一厅八室”业务流程办结后，自动触发政策法规发文成果、规划项目文件、不动产登记成果等资料向语料管理平台推送。在生产端，管理平台将存储的知识材料推送至语料生产系统，自动匹配对应加工方案，依次完成解析、切片后回传管理平台；管理平台将语料生产系统回传的语料推送至模型入库平台完成向量转换与入库存储。整套联动机制构建出业务催生知识、知识反哺业务的良性循环体系，持续推动规划资源治理模式智能化转型。

通过强化语料时间标签和模型时间检索实现对自动入库语料的版本管理。语料方面，针对政策法规和技术标准，按表4要求完善生效时间和失效时间标签，同一政策文件新版本入库触发旧版本自动添加失效时间。模型方面，建立不同时间触发机制，提问不涉及时间时，模型自动返回最新的政策文件；提问涉及确定时间段时，时间作为前置检索项，模型只返回该时间段内匹配的文档材料。

5 模型安全与伦理

规划资源领域存在政策迭代快、涉及自然资源权益和公共利益等敏感问题，虽然在材料采集阶段已开展了涉密、涉敏的内容筛查，但在知识库的不断扩容迭代中，仍会面临存量知识过时、新旧条文冲突、不同领域概念表述模糊等情况，可能导致模型输出错误，造成行业隐患。基于此，通过材料分级质检、材料溯源和时间检索、政策原文检索等手段，提升模型输出的安全性和准确性。

材料分级质检是指相关知识材料经采集单位人工质检、机器质检，责任单位审核，语料管理单位校核等多轮质检，充分确保材料准确、可靠。基于表9所列专题在模型人工测试阶段的事实性和正确性得分，可明显看到知识库材料本身真实性和可信度高，幻觉控制良好。

材料溯源和时间检索依托语料标签的规范化，实现按问题设定的时间检索；在模型输出内容中自动标注引用的语料片段，并提供对政策法规、技术标准等权威文件的下载；帮助用户对生成内容溯源、校核，有助于用户精准研判模型输出内容。

针对政务模型要求输出内容精准的需求，特设“政策ABC”技能，实现对政策文件的原文引用和原汁原味输出。框定检索对象，仅允许在政策法规、党内法规和技术标准3类材料中进行检索，同时抑制模型生成，要求其原文输出，并按需展示所引用政策文件的版本信息、条文演变等内容，使模型成为可靠的政策查询工具。

6 结论与探讨

本文系统总结了RAG技术在上海云宇星

表9 人工测评阶段模型测试数据（以第一批10个专题知识库为例）
Tab.9 Test data of model manual evaluation (taking the first batch of 10 thematic knowledge bases as an example)

专题名称	知识库材料数量/项	测试问答对数量/对	人工测试					
			人工评价指标	模型自动测算指标				
			点赞率/%	召回率/%	精准率/%	事实性	正确性	大模型评分
15分钟社区生活圈	85	134	90.4	95.15	79.48	95.4	91.1	81.9
沪派江南	135	100	93.3	93.43	78.52	92.3	89.2	80.4
城市总体规划	90	100	93.7	91.25	76.55	95.8	89.0	80.8
品质市政	68	112	96.1	91.00	78.17	96.1	93.1	82.9
城市设计	26	110	98.1	90.00	76.35	95.6	90.1	81.6
风貌保护管理	60	100	98.2	95.33	75.62	95.7	89.2	80.9
存量产业用地盘活	98	102	95.6	91.96	75.65	95.3	89.1	80.8
土地供给	151	101	97.3	93.43	81.30	98.4	91.1	81.5
生态复合单元和湿地空间	150	100	96.4	93.50	77.42	95.1	91.8	80.9
优化营商环境	58	100	98.2	97.33	78.82	96.6	92.8	81.8

资料来源：笔者根据第一批专题知识库模型人工测试结果整理。

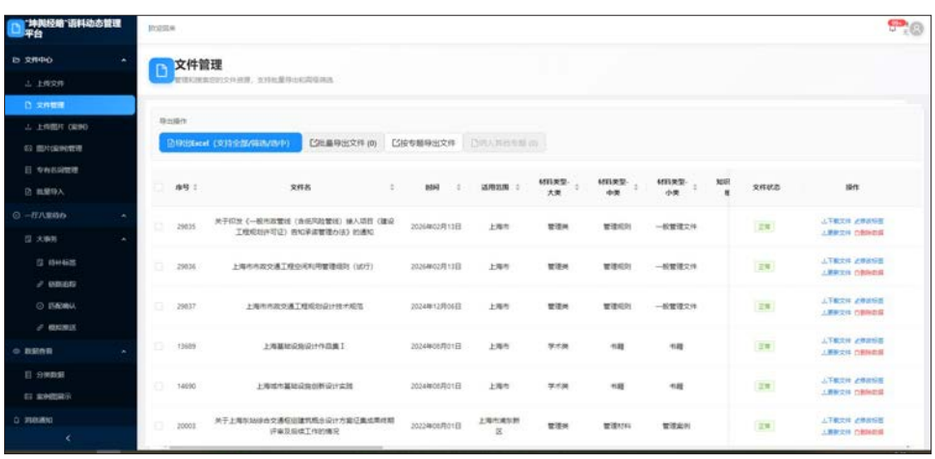


图3 语料动态管理平台
Fig.3 Dynamic corpus management system

资料来源：笔者截取自语料动态管理平台。

注释：⑦“一厅八室”是上海市规划和自然资源局在信息化建设领域提出的系统架构模式，旨在整合规划资源业务信息系统，实现全系统各层级、各板块的信息融合与统一管理。“一厅”即统一门户，为全市规划资源部门提供统一入口，为管理服务对象提供统一界面。“八室”是根据机构改革确立的职能配置架构，定义整体系统中有机统一的8个功能模块。具体包括：大规划（国土空间规划管理）、大土地（土地资源管理）、大项目（建设项目审批管理）、大事务（综合事务管理）、大地质（地质矿产资源管理）、大登记（自然资源确权登记）、大测调（测绘调查管理）、大监督（监督管理）。这一架构覆盖规划资源系统全层级、全板块，是上海市落实自然资源部国土空间规划“一张图”建设要求的重要支撑。

《上海城市规划》https://www.shplanning.com.cn/. All rights reserved.

空大模型建设中的探索应用，在以下方面实现了技术落地。第一，构建了垂直行业RAG知识工程的方法论体系，云宇星空大模型建立了“体系构建—材料采集—质检入库—语料生产—测试调优—动态更新”的完整知识库建设流程，形成以80个细分专题领域为框架的知识架构设计和“精准归集、质量可控、安全合规”的材料采集原则，为专业领域RAG知识工程建设提供了可复用的方法论框架。第二，云宇星空大模型针对规划资源垂直领域，实现了基于图像文本元数据的辅助检索。第三，研发了知识生产与业务联动的平台化工具，知识管理平台实现了材料收集、语料生产、知识入库的全流程数字化，并与业务审批平台建立自动化联动机制，构建了知识从业务中产生又赋能业务发展的良性循环。

未来有待改进之处包括以下3个方面。一是空间智能能力的深度拓展。当前云宇星空大模型的RAG能力主要集中于文本和二维图像知识，三维空间数据的RAG技术尚处于探索阶段，未来将进一步依托“米料”，实现模型对于三维数据的空间测量、空间重构、空间关系理解等核心能力的检索增强。此外，在二维图像的视觉理解方面，模型目前正在探索识图能力的提升，后续需进一步研发视觉—语言联合嵌入模型，完成图像内容向量化检索，增强图纸召回的语义关联度。二是知识图谱与RAG的深度融合。当前知识库以向量检索为主，仅针对专题领域、部分管理流程和政策法规版本建立了知识图谱，知识图谱的结构化推理能力尚未充分释放。考虑到规划资源领域存在大量业务关系以图谱形式表达更为直观清晰，包括如地块与规划指标的关联，不同管理环节在流程关系、业务内容关系、专业名词相关性等方面的关系等，未来需进一步加强各类专业知识的图谱构建，实现基于图结构的复杂推理，提升模型对规划资源领域复杂关联关系的理解能力。三是多智能体协同的RAG架构。随着量子城市应用场景拓展，单一RAG系统难以满足复杂任务需求，未来需探索多智能体架构下的分布式RAG，实现跨领域知识的协同检索和综合生成。

云宇星空大模型的RAG技术探索，为规划资源行业智能化转型提供了实践样本，也为其他专业领域行业大模型建设提供了经验借鉴。随着空间智能技术的持续发展，RAG技术将在超大城市治理中发挥更加重要的作用，助力谱写上海“量子城市”与人民城市建设的崭新篇章。

（上海市数字城市规划研究中心奚文沁、欧阳琳欣、徐望悦、王亚楠，库帕思科技杜展展、王圣峰，商汤科技邱雷、成修对本文有贡献，在此深表感谢。）

参考文献 References

[1] 刘雪颖，云静，李博，等. 基于大型语言模型的检索增强生成综述[J]. 计算机工程与应用，2025，61（13）：1-25.
Liu Xueying, Yun Jing, Li Bo, et al. Survey on retrieval-augmented generation based on large language models[J]. Computer Engineering and Applications, 2025, 61(13): 1-25.

[2] 任海玉，刘建平，王健，等. 基于大语言模型的智能问答系统研究综述[J]. 计算机工程与应用，2025，61（7）：1-24.
Ren Haiyu, Liu Jianping, Wang Jian, et al. Review of intelligent question answering systems based on large language models[J]. Computer Engineering and Applications, 2025, 61(7): 1-24.

[3] 上海市规划和自然资源局.“云宇星空大模型”——上海量子城市空间智能建设之规划资源行业大模型[EB/OL]（2025-12-21）[2026-06-07]. https://mp.weixin.qq.com/s/ZRpzwKrcMhT-EEBayiN11Q.
Shanghai Municipal Bureau of Planning and Natural Resources. “The Cloud Cosmos Large Model”: domain foundational large model for the planning and resources industry in Shanghai Quantum City spatial intelligence construction[EB/OL]. (2025-12-21) [2026-06-07]. https://mp.weixin.qq.com/s/ZRpzwKrcMhT-EEBayiN11Q.

[4] 上海市人民政府. 上海市国民经济和社会发展第十五个五年规划纲要[Z]. 2026.
Shanghai Municipal People’s Government. The 15th Five-Year Plan for national economic and social development of Shanghai[Z]. 2026.

[5] 上海规划资源. 量子无界、智治新篇，一文读懂上海量子城市空间智能建设行动[EB/OL].（2025-12-21）[2026-06-07]. https://mp.weixin.qq.com/s/08g9LBvUoUpWki_Lo9saFQ.
Shanghai Guihuaziyuan WeChat Account. Quantum boundless, new chapter of smart governance: a guide to Shanghai Quantum City spatial intelligence construction action[EB/OL]. (2025-12-21) [2026-06-07]. https://mp.weixin.qq.com/s/08g9LBvUoUpWki_Lo9saFQ.

[6] 卓梁雨，万振龙，马超凡. RAG及AI编码技术在海关知识库平台中的应用与优化[J]. 中国口岸科学技术，2026，8（1）：10-15.
Zhuo Liangyu, Wan Zhenlong, Ma Chaofan. Application and optimization of RAG and AI coding technology in customs knowledge base platform[J]. China Port Science and Technology, 2026, 8(1): 10-15.

[7] Gao Y F, Xiong Y, Gao X Y, et al. Retrieval-augmented generation for large language models: a survey[PP/OL]. V1. (2023-12-18) [2026-06-07]. https://doi.org/10.48550/arXiv.2312.10997.

[8] 籍欣萌，昝红英，崔婷婷，等. 大模型在垂直领域应用的现状与挑战[J]. 计算机工程与应用，2025，61（12）：1-11.
Ji Xinmeng, Zan Hongying, Cui Tingting, et al. Current status and challenges of large models in vertical domain applications[J]. Computer Engineering and Applications, 2025, 61(12): 1-11.

[9] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[PP/OL]. V4. (2021-04-12) [2026-06-07]. https://doi.org/10.48550/arXiv.2005.11401.

[10] Zhao T, Wallace E, Feng S, et al. Calibrate before use: improving few-shot performance of language models[C]//Proceedings of the International Conference on Machine Learning, 2021: 12697-12706.

[11] Cheng X, Luo D, Chen X, et al. Lift yourself up: retrieval augmented text generation with self-memory[C]//Advances in Neural Information Processing Systems, 2024.

[12] Gao L, Madaan A, Zhou S, et al. PAL: program-aided language models[C]//Proceedings of the International Conference on Machine Learning, 2023: 10764-10799.

[13] Huang L, Yu W J, Ma W T, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions[J]. ACM Transactions on Information Systems, 2025, 43(2): 1-55.

[14] 张玉鑫. 打造量子城市 构建数字化赋能超大城市治理新模式[J]. 中国测绘，2025（5）：7-10.
Zhang Yuxin. Building a Quantum City to construct a new digital empowerment model for mega-urban governance[J]. China Surveying and Mapping, 2025(5): 7-10.

[15] 程南炜，徐见微，王进. 面向城市规划设计的多模态RAG知识库系统构建方法研究[J]. 中国信息界，2026（3）：85-89.
Cheng Nanwei, Xu Jianwei, Wang Jin. Research on the construction method of multimodal RAG knowledge base system for urban planning and design[J]. China Information Times, 2026(3): 85-89.

《上海城市规划》https://www.shplanning.com.cn/. All rights reserved.