# （一）大模型驱动的协同标注系统

## 系统背景:

在第二届全球数字贸易博览会上，国家数据局局长刘烈宏指出，数据基础设施将对扩大数据产业规模、繁荣数据产业生态、促进数字经济高质量发展产生重要意义。将AI算法比作“火箭”，那么数据标注就是推动AI的“燃料”，对训练模型、提高准确率至关重要。在当前竞争激烈的AI市场，数据标注的效率和质量将直接影响企业和组织的AI应用效果和竞争力。

另一方面，OpenAI发布的ChatGPT等大模型掀起了一轮研究热潮。结合了超大规模的模型、数据和算力，大模型初步涌现了通用智能，对众多行业形成广泛的应用潜力。随着大模型的发展，数据标注也有了新机遇。首先，大模型在垂直领域的应用更需要大量的标注数据，以实现大模型的领域微调。其次，大模型内蕴的通用知识，能够以低成本构建高质量数据，在数据标注领域将会带来新的可能性和突破口。

在国内外，以网易伏羲有灵、Snorkel Flow、火山引擎等为代表的数据标注产品已将大模型融入标注的过程中。然而，面对垂直领域的产业需求时，通用大模型往往难以直接解决标注问题，这些标注产品仍需大量借助人类知识进行数据的标记、校验和修复。因此，亟需开发基于

## 主要功能

文档管理系统开发的主要目的是结合RFID技术对文档进行管理，主要功能模块包括用户管理、文档管理和业务管理三大模块。

（1）大小模型协同的弱标注模块

利用大模型（LLM或VLM）作弱标注者+小模型（SML）做蒸馏（样例选择）建设无监督/弱监督主动学习标注系统。

（2）人工标注调度模块

利用人工智能技术辅助人工标注工作，包括预标注数据推荐、人工指导标注推荐、标注过程可视化。

（3）业务管理

利用工程化技术，提供流程B/S架构应用系统，为用户提供操作流畅的一站式自动生成标注数据服务。包括前端、后端、算法相关技术栈。

## 开发环境

操作系统及版本：windows xp

支撑软件及版本列表：Python + Vue.js

数据库环境：MySql

## 基本的架构体系

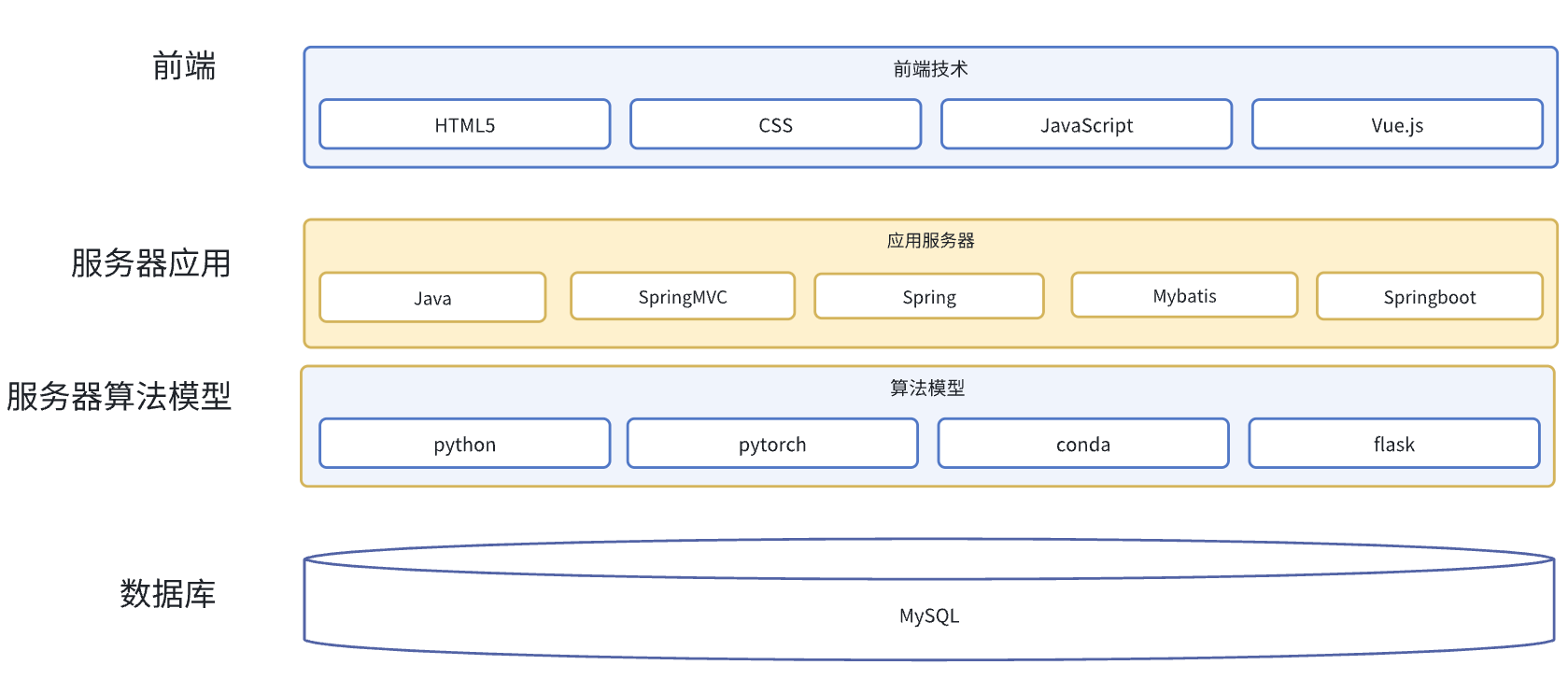


图 1 大模型驱动的标注系统功能结构图

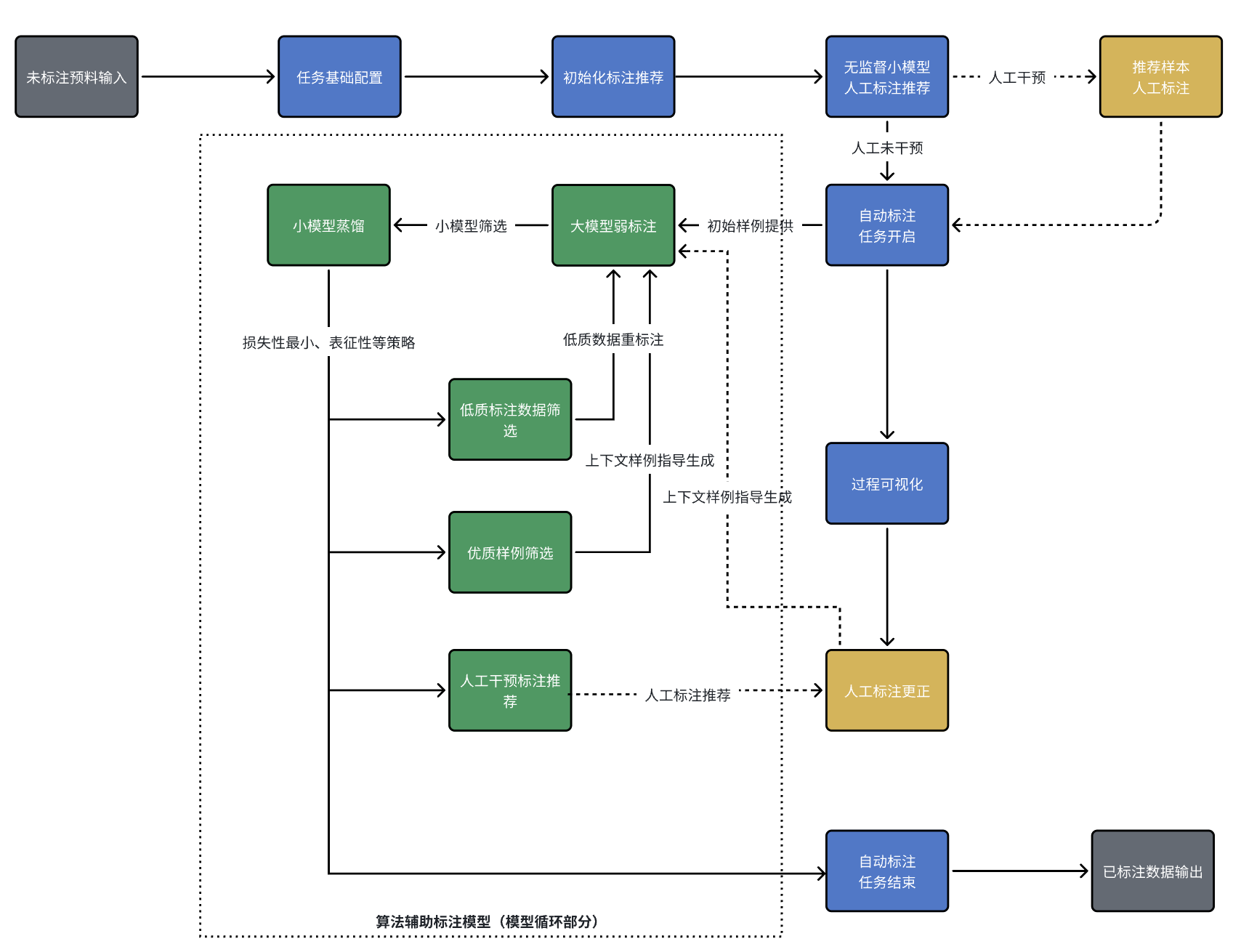


图2 核心功能流程图

用户管理：包括数据生产商，可发布数据标注任务，收集众包标注人员反馈的标注信息。以及众包标注用户，管理标注结果。

标注模块：由众包标注用户操作，用户可登陆和退出系统，配置大小模型进行自动化标注，再进行人工修正。

数据管理模块：包含数据的存储、下载、保存等操作。