

辰极智脑 - 认证与仪表盘需求文档

一、修订记录

修订日期	修订内容	修订人员	审核人员
2025-07-12	辰极智脑认证与仪表盘功能初稿	辰极智脑团队	辰极智脑团队

二、需求背景与目标

2.1 编写目的

本文档旨在明确“辰极智脑”系统中“用户登录”及“系统仪表板”模块的功能需求、交互逻辑及非功能性指标。该文档将作为UI设计、前端开发、后端开发及测试用例编写的依据。

2.2 产品背景

“辰极智脑”作为企业级AI服务中台，安全性与可观测性至关重要。

- 认证体系**：需要构建安全、高效的准入机制，确保仅授权用户可访问系统，并严格区分企业管理员与普通用户的身份边界。
- 系统仪表板**：作为用户登录后的首屏，需直观展示系统的运行健康度（如GPU/CPU状态）、服务可用性及核心业务指标（如文档处理量、查询量），使用户了解系统运行情况。

2.3 术语定义

- **Tenant (租户/企业)****：系统的顶层逻辑隔离单位，一个企业对应一个 Tenant。
- **JWT (JSON Web Token)****：用于在网络应用环境间传递声明而执行的一种基于JSON的开放标准，本系统用于身份验证。
- **QPS (Queries Per Second)****：每秒查询率，衡量系统吞吐量的重要指标。
- **GPU (Graphics Processing Unit)****：图形处理器，本系统中用于承载 LLM 推理和 Embedding 向量化计算的核心硬件。

2.4 系统边界与依赖

本系统（辰极智脑）专注于 AI 能力的供给与编排，**不包含**原生的用户注册与组织架构管理功能。所有身份认证与权限控制均依赖外部的 ****中央管理系统 (ts_manage)****。

- ts_manage 职责**：
 - 用户管理**：负责账号创建、密码重置、用户信息维护。
 - 租户管理**：负责企业（Tenant）的生命周期管理及组织架构树的维护。
 - 权限中心**：负责定义角色（Role）及分配菜单路由权限。
- 辰极智脑 职责**：
 - 认证代理**：提供登录界面，将认证请求透传至 ts_manage，并缓存返回的 Token。
 - 权限消费**：根据 ts_manage 返回的路由表，动态渲染前端菜单与页面。

三、用户角色与场景

3.1 用户角色

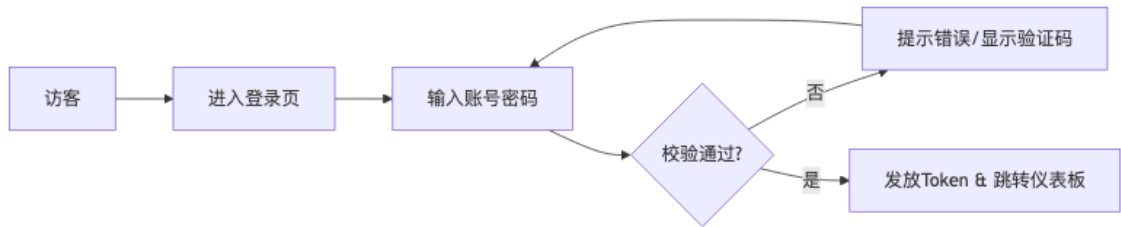
角色	职责描述	权限范围
访客	未登录用户	仅能访问登录页；无法访问系统内部功能。
普通用户	知识的使用者	登录系统；查看个人的业务统计数据（如个人文档数）；查看基础服务状态。
企业管理员	企业内系统的维护者	登录系统；查看本企业的全局业务统计数据；查看系统硬件资源监控（CPU/GPU/内存）及详细服务健康报告。
平台管理员	整个平台的超级管理者	管理所有租户（企业）的生命周期（开通/禁用）；管理全局系统配置（如模型接入）；查看全平台系统级监控与日志。

3.2 典型场景

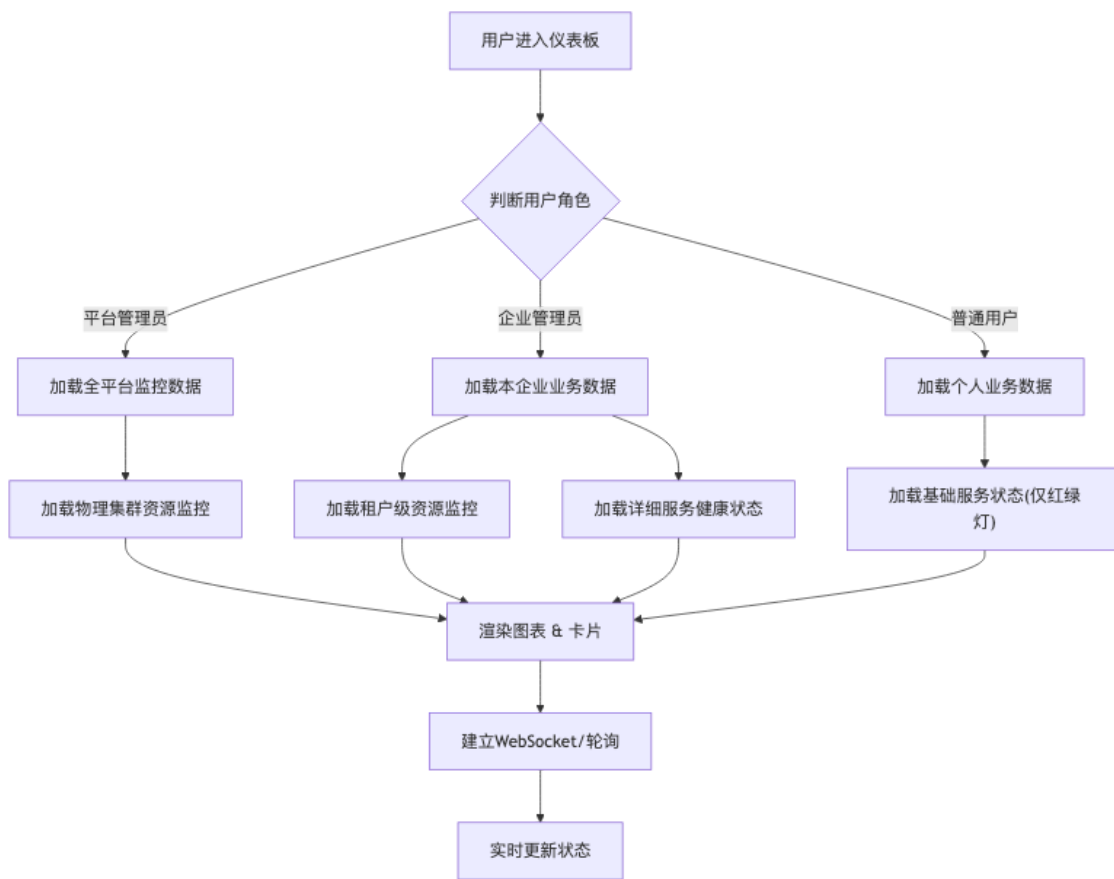
- 1. 账号分发：企业管理员在 **ts_manage** 后台为新员工创建账号并分配“普通用户”角色，员工获取账号后即可登录辰极智脑。
- 2. 日常登录：用户通过账号密码登录系统，系统自动向 **ts_manage** 发起认证；若**连续3次**输错密码，触发安全机制，**ts_manage** 将要求用户进行校验图形验证码。
- 3. 系统巡检：企业管理员登录后，在仪表板查看 GPU 显存使用率和 LLM 服务响应时间，评估是否需要扩容。
- 4. 业务概览：普通用户登录后，在仪表板查看自己上传的文档处理进度和今日查询次数。

四、业务流程

4.1 用户登录流程



4.2 仪表板数据加载流程



五、功能性需求

5.1 认证集成说明

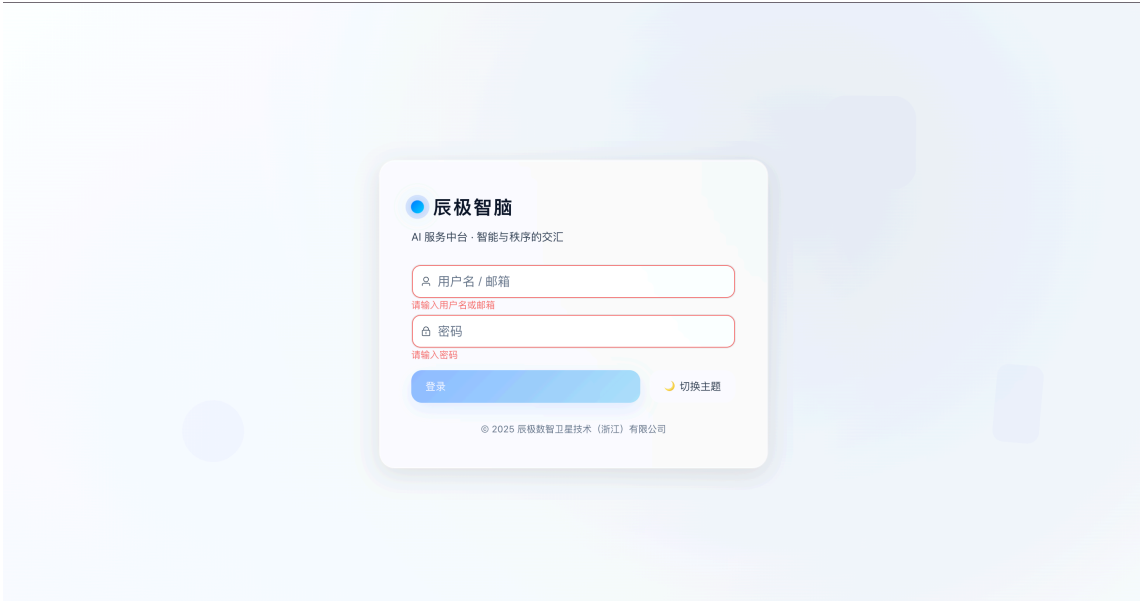
业务说明

本系统不提供用户注册入口。所有用户的生命周期管理（创建、编辑、禁用、删除）均在 **中央管理系统 (ts_manage)** 中完成。

1. 账号获取：
 - 企业开通：由平台管理员在 ts_manage 中创建“租户”及“企业管理员”账号。
 - 员工入职：由企业管理员在 ts_manage 中创建“员工账号”并分配权限。
2. 密码管理：
 - 初始密码由管理员在创建时指定。
 - 密码重置：由管理员在 ts_manage 后台操作。
3. 数据同步：
 - 用户登录成功后，本系统会缓存 ts_manage 返回的用户基本信息（如 UserID, TenantID, Role），用于后续的业务逻辑判断。

5.2 用户登录

原型图



业务说明

1. 登录方式：账号 + 密码。
- 交互流程：用户输入账号和密码后，点击“登录”按钮或按回车键。系统首先进行前端格式校验（如非空校验、长度限制），校验通过后自动提交至后台进行身份验证。
2. 认证代理逻辑：
- 安全传输：系统采用前后端分离的认证机制。前端界面收集用户凭证并加密传输，后端服务作为安全代理，将认证请求转发至中央管理系统 (ts_manage) 进行核验。
 - 会话保持：验证通过后，系统建立本地会话并缓存用户权限信息，保证用户会话连续有效。
3. 安全机制（依赖 ts_manage）：
- 验证码：当系统检测到异常登录尝试（如密码错误）时，会自动触发验证码机制，要求用户输入图形验证码以继续。
 - 账户锁定：为保障账户安全，系统执行严格的锁定策略：若用户连续输入错误密码达到 **10** 次，该账户将被临时锁定 **30分钟**。锁定期间无法登录，需等待自动解锁或联系管理员重置。
4. 登录反馈：
- 成功：跳转至系统仪表板，弹出通知“欢迎回来，{用户名}”。
 - 失败：直接展示 ts_manage 返回的错误信息（如“用户名或密码错误”）。
5. 主题切换：登录页提供“深色/浅色”主题切换按钮，偏好设置保存在本地浏览器。

界面字段说明

字段名	字段类型	是否必填	说明
用户名	文本输入	是	支持用户名或邮箱登录。
密码	密码输入	是	密文显示，支持“显示/隐藏”切换。

验证码	文本输入	条件显示	当后端检测到异常登录时触发。
-----	------	------	----------------

界面操作说明

操作名称	操作说明
登录	提交凭证，成功后获取 Token 并存储，跳转首页。
刷新验证码	点击验证码图片，重新获取一张新图片。(当出现验证码校验机制时)
切换主题	点击"切换主题"按钮，切换界面深色/浅色模式。

5.3 系统仪表板

原型图



模块化编排器服务状态

服务名称	状态	消息	响应时间
API服务	正常	running	-
任务管理器	正常	running	-
文件存储	正常	running	-
基础服务	正常	running	-
分块服务	正常	running	-
文档处理器	正常	running	-
LightRAG工厂	正常	running	-
知识库服务	正常	running	-
LLM服务	正常	running	-
rag_registry	正常	running	-
rag_provider	正常	lightrag	-
rag_provider_instance	正常	running	-
rag_lightrag_workspace	正常	running	-

rag_storage_check	正常	enabled	-
rag_working_dir_exists	正常	running	-
rag_provider_health	正常	healthy	-
rag_provider_response_tir	正常	0.0	0ms
rag_fallback	正常	running	-

© 2025 辰极数智卫星技术（浙江）有限公司

运行时间: 1天23小时30分钟 | 最后更新: 17:27:28

业务说明

1. 数据指标定义：

- **系统健康度**：系统整体运行状态的综合评分，采用"木桶原则"计算，即取所有核心微服务中最差的状态作为整体状态。

- **计算规则**：

子服务状态	综合状态
所有服务均为"正常"	正常（绿色）
至少一个服务为"警告"，且无"错误"	警告（橙色）
至少一个服务为"错误"	错误（红色）

- **服务存活率**：当前正常运行的微服务数量占总服务数量的比例，展示格式为" X / Y "（如 " $5 / 6$ "）。
- **租户总数**：统计从 `ts_manage` 同步的、当前状态为"已启用"的企业租户总数（仅平台管理员可见）。
- **文档总数**：已完成向量化索引且状态正常的文档总数。
- **查询总数**：系统上线以来，累计处理的检索（Retrieval）与对话（Chat）请求总次数。
- **存储总量**：所有文档占用的对象存储（OSS）空间总和，单位自动换算（MB/GB/TB）。
- **资源使用率**：底层物理计算集群的实时平均利用率，数据源于基础设施监控系统（如 Prometheus）。
 - **CPU 使用率**：所有 CPU 核心的平均占用百分比。
 - **内存使用率**：已用内存占总内存的百分比。
 - **GPU 显存使用率**：GPU 显存占用百分比。
 - **GPU 温度**：GPU 核心温度（摄氏度），超过 80°C 时标红预警。
- **服务详情指标**：
 - **状态**：通过心跳检测获取的实时状态。
 - **运行中（绿色）**：最近一次心跳检测成功。
 - **异常（红色）**：最近一次心跳检测失败或超时（超时阈值：5秒）。
 - **未知（灰色）**：服务刚注册，尚未完成首次心跳检测。
 - **响应时间**：服务接口的 P95 响应延迟（即 95% 的请求响应时间低于此值），单位毫秒（ms）。超过 1000ms 时标黄预警。

2. 角色视图权限：

指标项	平台管理员	企业管理员	普通用户
租户总数	全平台总数	-	-
文档总数	全平台总数	本企业总数	个人上传总数

查询总数	全平台总数	本企业总数	个人查询总数
今日上传/存储	-	本企业统计	个人统计
物理资源 (CPU/GPU)	全平台物理节点监控	本企业配额使用情况	-
服务健康状态	全局服务健康+异常日志	本企业相关微服务状态	仅核心服务红绿灯状态

3. 实时性要求：

- **数据刷新：** 页面所有数据每 30 秒自动刷新一次，用户也可点击刷新按钮手动刷新。

4. 状态可视化：

- **健康度：** 使用颜色区分（绿色-正常，橙色-警告，红色-错误）。
- **趋势图：** (可选) 展示近24小时的 QPS 或 Token 消耗趋势。

界面操作说明

操作名称	触发条件	操作说明	影响/结果
刷新数据	用户点击页面右上角的“刷新”按钮。	强制重新从后端 API 获取最新的监控数据与统计指标。	页面所有统计卡片与图表展示的数据将更新为最新状态；若请求失败，弹出错误提示。
查看详情	用户点击具体服务（如“知识库服务”）的状态标签。	查看该服务的详细运行日志或健康检查报告。	跳转至该服务的详情页面（或弹出侧边栏），展示更细粒度的性能指标与错误日志。

六、非功能性需求

6.1 安全性

- **传输加密：** 所有登录请求必须通过 HTTPS 传输。
- **敏感信息脱敏：** 日志中严禁记录明文密码。
- **防暴破：** 验证码机制必须在后端强制校验，防止绕过前端直接调接口。

6.2 性能要求

- **登录响应：** 登录接口响应时间 < 500ms。
- **仪表板加载：** 首屏数据加载时间 < 1秒。
- **监控延迟：** 资源监控数据的延迟不超过 5 秒。

6.3 浏览器兼容性

- 支持 Chrome 80+, Firefox 70+, Safari 13+, Edge 80+。
- 仪表板需支持响应式布局，适配 1920x1080 及 1366x768 分辨率，移动端仅展示核心统计卡片。