

RagFlow完整配置与使用指南

概述

RagFlow是星核电脑预装的强大地知识库软件，支持多种AI模型类型和文档处理功能。通过本指南，您可以完整掌握从模型配置到知识库使用的全流程操作。

1. RagFlow初始设置

注册登录流程

首次进入RagFlow：

- 访问：`http://127.0.0.1:3368`
- 在登录页面点击"注册"
- 输入任意邮箱和密码（本地注册，无需真实邮箱）
- 直接点击"注册"按钮进入系统

用户信息配置

个人设置页面：

- 用户名：xh（默认）
- 语言：简体中文
- 时区：UTC+8 Asia/Shanghai
- 邮箱地址：localhost.com（示例）

2. 模型配置详解

星核标准端口配置

本地模型服务端口：

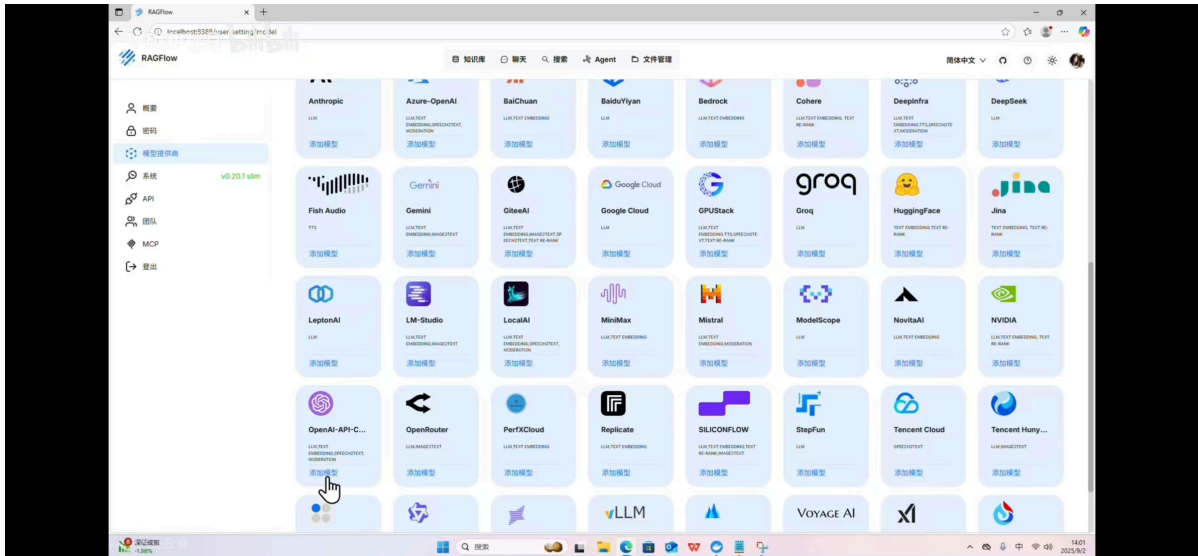
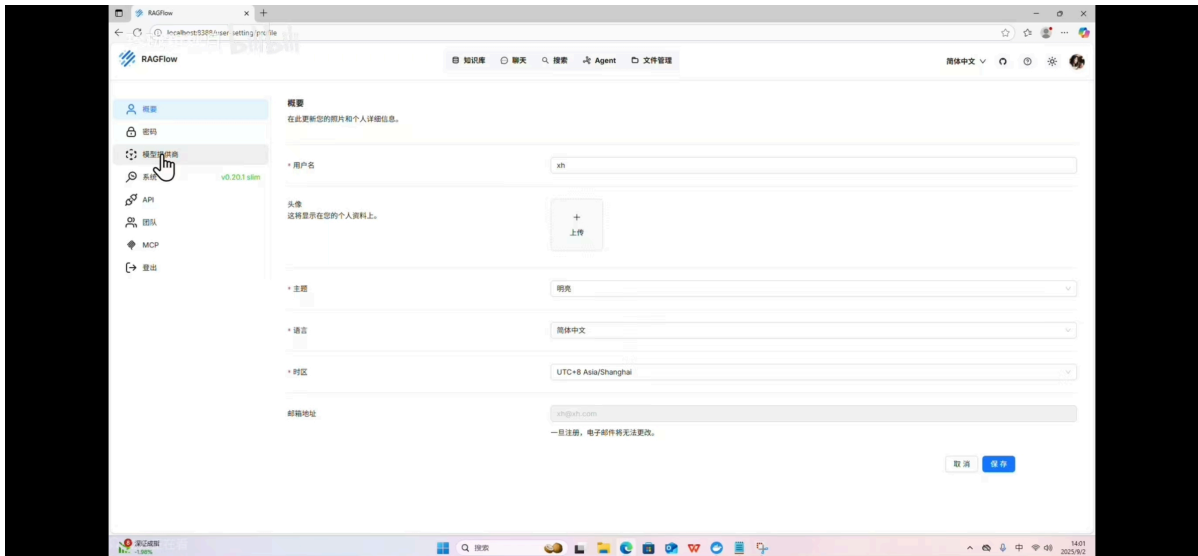
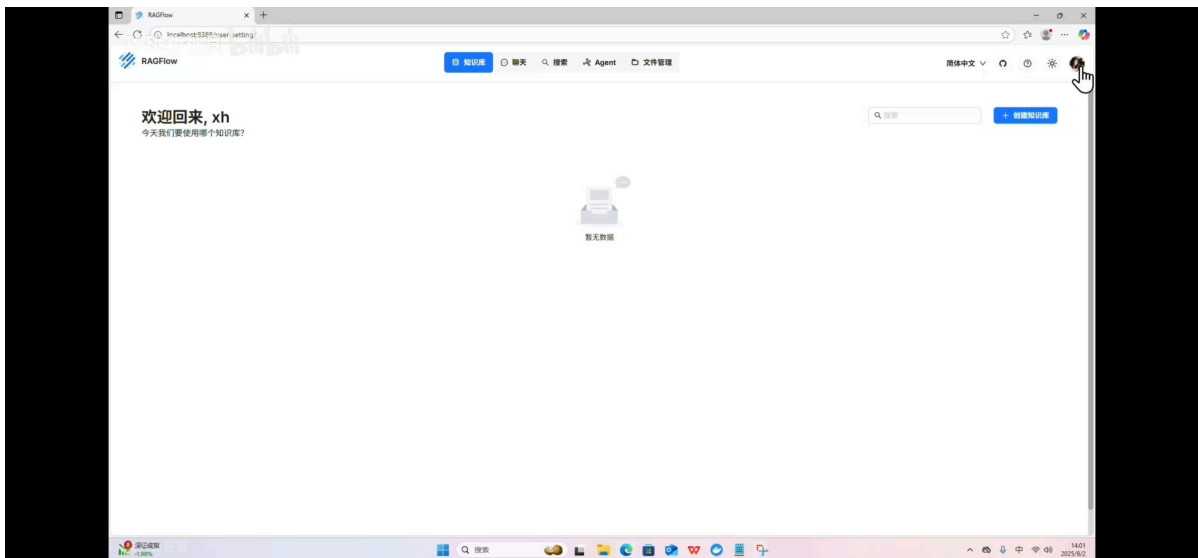
- Chat模型：`http://host.docker.internal:1234/v1`
- Embedding模型：`http://host.docker.internal:1278/v1`
- Image2Text模型：`http://host.docker.internal:1256/v1`

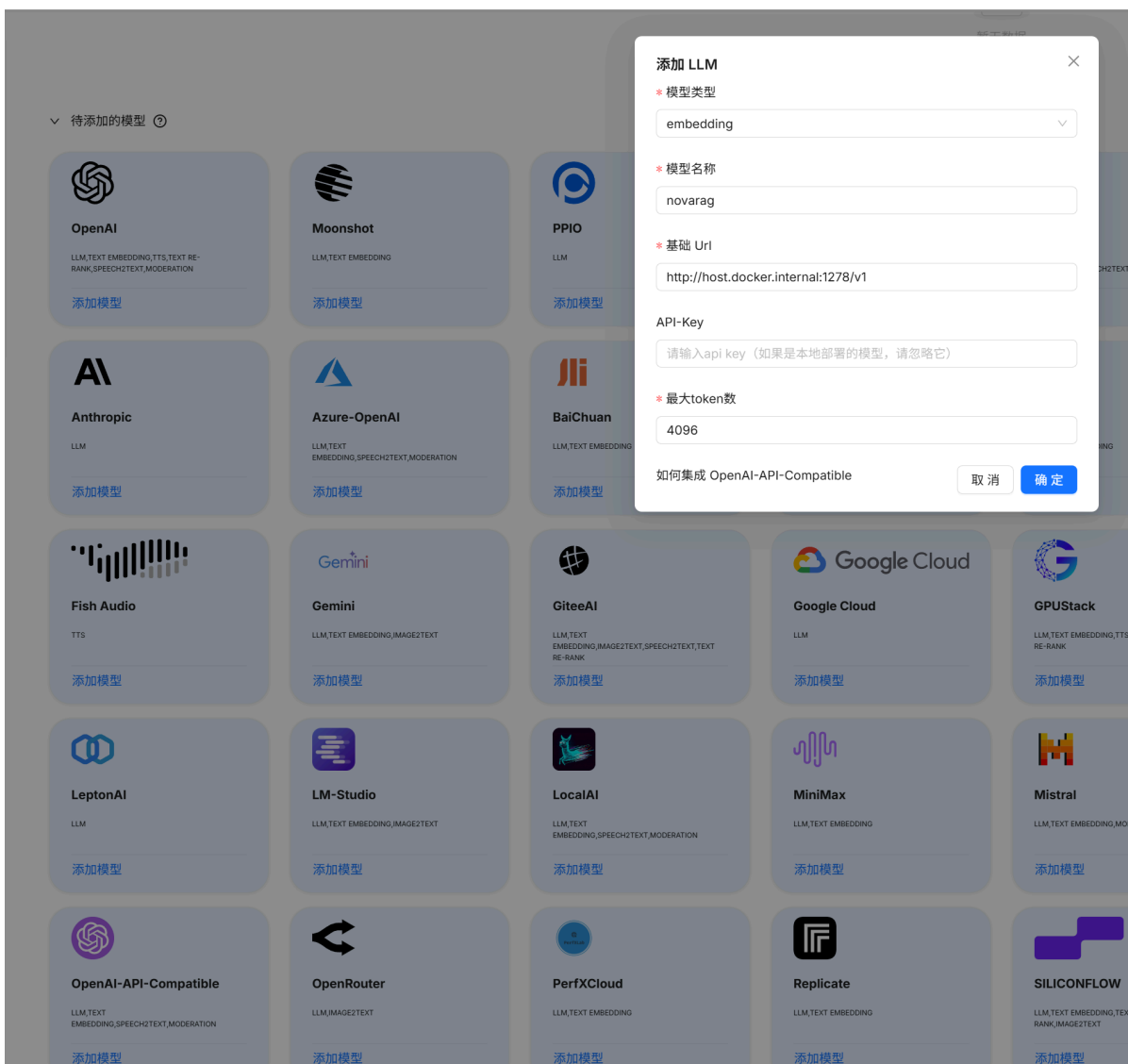
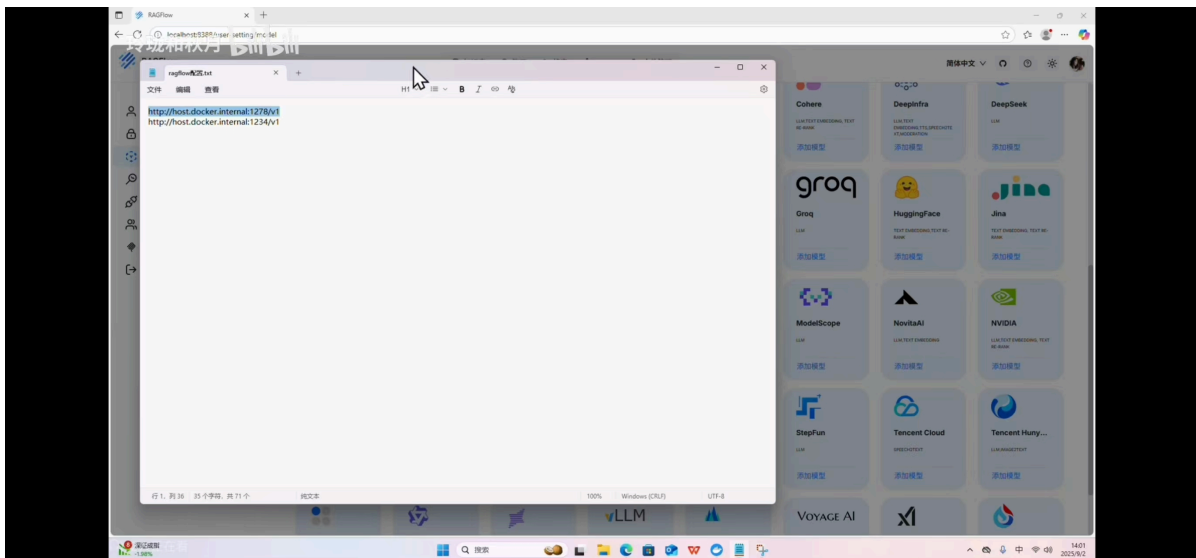
2.1 Embedding模型配置（向量化模型）

配置步骤：

- 选择模型类型：embedding
- 填写配置信息：

模型类型：embedding
模型名称：novarag
基础URL：`http://host.docker.internal:1278/v1`
API-Key：（留空）
最大token数：4096





2.2Chat模型配置（对话模型）

配置步骤：

1. 点击"模型管理" → "添加LLM"
2. 填写配置信息：

模型类型： chat

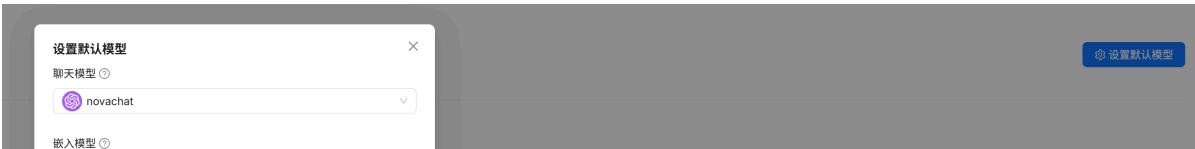
模型名称： novachat

基础URL: http://host.docker.internal:1234/v1

API-Key: （留空）

最大token数： 30000

是否支持Vision： 关闭（根据模型能力调整）



添加 LLM

* 模型类型

chat

* 模型名称

novachat

* 基础 Url

http://host.docker.internal:1234/v1

API-Key

请输入api key（如果是本地部署的模型，请忽略它）

* 最大token数

30000

是否支持 Vision

☐

如何集成 OpenAI-API-Compatible

取消

确定

2.3 Image2Text模型配置（图像识别模型）

2.3.1首先：到智玲同学首页——模型管理里面MimoVL-7B 进行参数设计，端口号#1234#改为#1256#

MimoVL-7B

Qwen

已下载

具备推理能力的多模态模型，体积小巧，性能强大，在图片识别和推理方面有不错的表现

大小

6GB

下载量

10万+

授权

开源

启动

参数

模型参数

MimoVL-7B

自动启动: ☐ 禁用

模型参数:

-c 30000 -ctk q8_0 -ngl 100 --host 0.0.0.0 --mmproj %ModelDir%/MimoVL-7B/mmproj-MiMo-VL-7B-RL_BF16.gguf -b 8192 -ub 4096

端口号:

1234

保存配置

模型参数

MimoVL-7B

自动启动: ☐ 禁用

模型参数:

恢复默认参数

```
-c 30000 -ctk q8_0 -ngl 100 --host 0.0.0.0 --mmproj %ModelDir%/MimoVL-7B/mmproj-MiMo-VL-7B-RL_BF16.gguf -b 8192 -ub 4096
```

端口号:

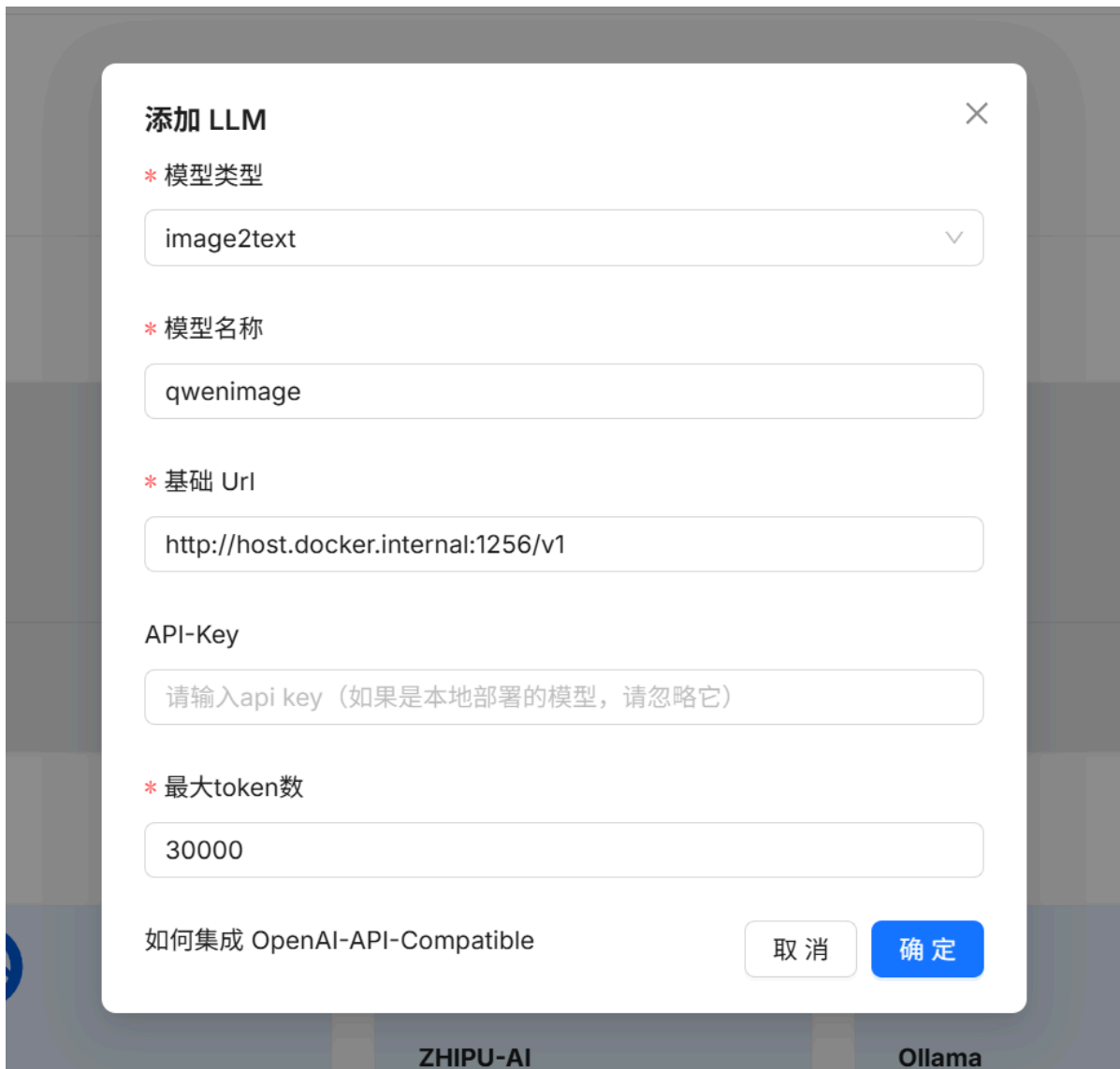
1256

保存配置

2.3.2配置步骤:

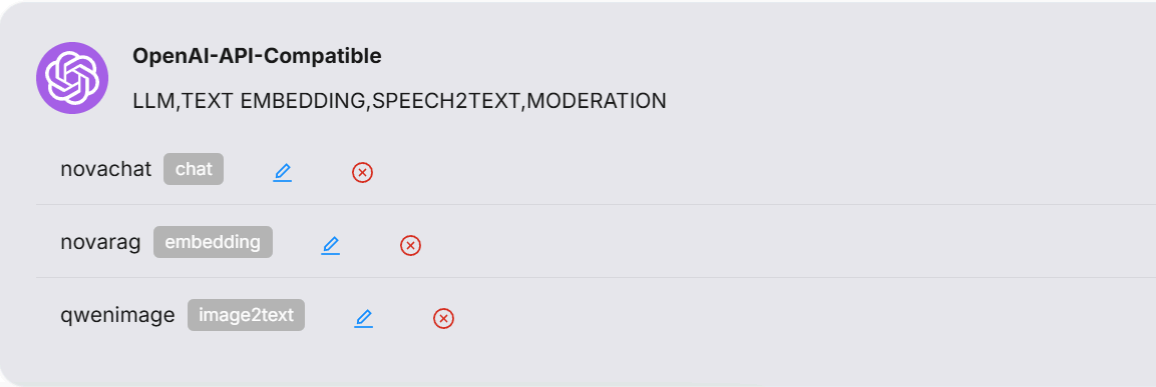
1. 选择模型类型: image2text
2. 填写配置信息:

模型类型: image2text
模型名称: qwenimage
基础URL: http://host.docker.internal:1256/v1
API-Key: (留空)
最大token数: 30000

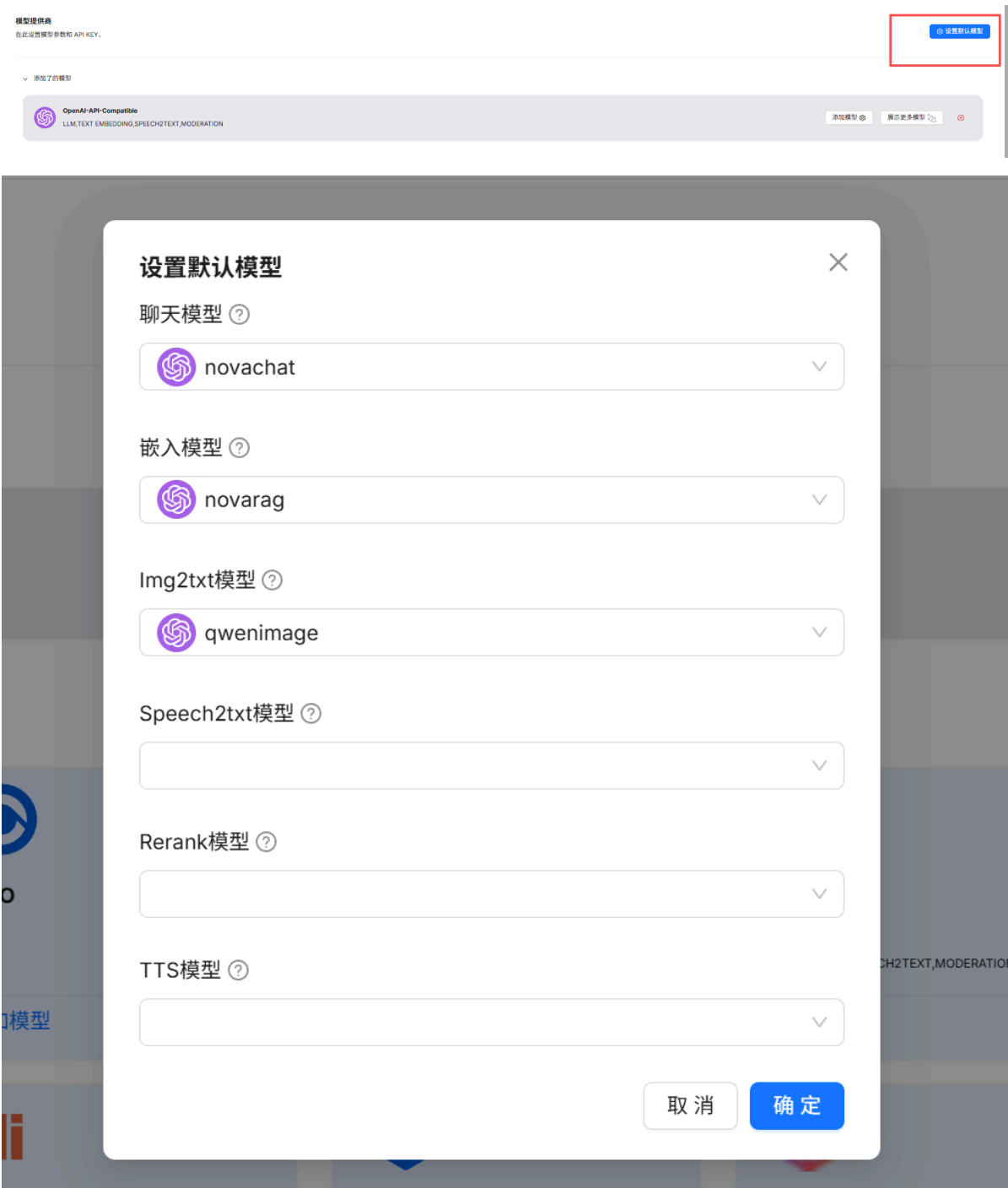


3. 模型配置成功

✓ 添加了的模型



4. 设置默认模型（下小心错过，谢谢 L。全程辅导提醒）



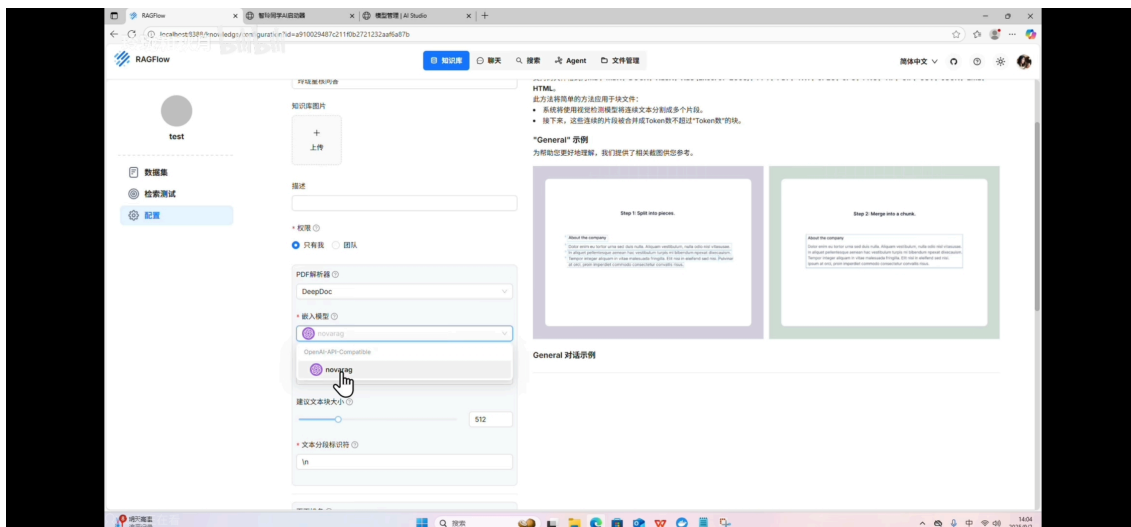
5. 知识库创建与配置

创建知识库

操作流程：

1. 点击"知识库" → "创建知识库"
2. 输入知识库名称（如：test）
3. 选择配置选项：
 - 权限：只有我
 - PDF解析器：DeepDoc
 - 嵌入模型：选择已配置的embedding模型

- 切分方法：General
- 语义文本最大块：512
- 文本分片段识别：\n



文档上传与处理

支持的文档格式：

- PDF文件
- Word文档 (.doc/.docx)
- TXT文本文件
- Markdown文件

上传步骤：

1. 进入已创建的知识库
2. 点击"上传"按钮
3. 选择本地文档文件
4. 等待文档解析和向量化处理
5. 查看处理状态（绿色表示成功）

解析模板选择

General模板配置：

- 用途：适用于大多数文档类型
- 特点：通用性强，处理稳定
- 推荐：日常文档处理首选

6. 聊天助手配置

创建聊天助手

基本配置：

1. 点击"聊天" → "聊天助手"
2. 点击"创建助手"按钮

3. 填写助手信息：

- **配置模式**：显示引用
- **聊天模型**：选择已配置的chat模型
- **知识库**：选择要关联的知识库

高级参数设置

自由度参数：

- **温度**：0.7（推荐值，平衡创造性与准确性）
- **Top P**：0.8（推荐值）
- **存在处罚**：0.2
- **频率处罚**：0.7
- **最大token数**：2048

系统提示词配置

星核专用提示词：

你就是星核智能电脑，请总结知识库的内容来回答问题，请列举知识库中的数据详细回答，知识库中有关星核的介绍，其实就是对你自己的介绍，请再次牢记——你就是星核。如果用户的问题在知识库没有直接答案，但是与知识库中的内容有关，你需要基于知识库中的事实依据，并且适当结合你自己的知识，进行推理性的扩展回答，但需要告诉用户哪些是你推理的。当所有知识库内容都与问题无关时，你的回答必须包括"知识库中未找到您要的答案！"这句话。回答需要考虑聊天历史。

以下是知识库：

{knowledge}

以上是知识库。

功能开关配置

推荐设置：

- **显示引用**：开启（显示信息来源）
- **关键词分析**：关闭
- **文档精确度**：开启

7. 第三方模型服务商支持

云端模型提供商

星核电脑支持接入多种第三方AI服务：

国际主流服务商：

- **OpenAI** - GPT系列模型
- **Anthropic** - Claude系列模型
- **Google Cloud** - Gemini、PaLM模型
- **Azure-OpenAI** - 微软Azure上的OpenAI服务
- **AWS Bedrock** - 亚马逊AI模型服务

国内服务商：

- ZHIPU-AI - 智谱AI (GLM系列)
- BaiChuan - 百川智能
- BaiduYiyan - 百度文心一言
- Tencent Cloud - 腾讯云AI
- Moonshot - 月之暗面 (Kimi)
- DeepSeek - 深度求索

第三方服务配置示例

OpenAI配置：

服务商：OpenAI
模型类型：chat
模型名称：gpt-4
基础URL：https://api.openai.com/v1
API-Key：sk-your-api-key-here
最大token数：8192

8. 实际使用流程

知识库问答测试

测试步骤：

1. 进入聊天界面
2. 选择配置好的聊天助手
3. 输入测试问题
4. 查看回答质量和引用信息
5. 根据需要调整参数

多文档管理

批量处理建议：

- 按主题分类上传文档
- 使用统一的命名规范
- 定期检查处理状态
- 及时处理错误文档

9. 性能优化建议

硬件资源管理

显存分配建议：

- 基础配置：64G显存 (Chat + Embedding)
- 完整配置：96G显存 (支持所有模型类型)

- **内存要求：**建议保留16G以上系统内存

模型选择策略

按使用场景优化：

使用场景	Chat模型	Embedding模型	备注
日常问答	本地novachat	本地novarag	成本最低
专业分析	云端GPT-4	本地novarag	平衡性能和成本
大规模处理	本地235B	本地novarag	性能最佳

10. 故障排除

常见问题及解决方案

问题1：模型连接失败

症状：模型显示红色状态

原因：模型服务未启动或端口错误

解决：

1. 检查星核模型管理中对应模型状态
2. 确认端口配置正确
3. 重启模型服务

问题2：文档处理失败

症状：文档状态显示错误

原因：文档格式不支持或内容过复杂

解决：

1. 转换文档格式为PDF或TXT
2. 清理文档中的图片和表格
3. 分割大文档为小块处理

问题3：回答质量不佳

症状：回答不准确或无法找到相关信息

原因：参数设置不当或知识库内容不足

解决：

1. 调整温度和Top P参数
2. 增加相关文档到知识库
3. 优化系统提示词

问题4：系统响应慢

症状：查询响应时间过长

原因：系统资源不足或并发请求过多

解决：

1. 关闭不必要的模型服务
2. 减少并发查询数量
3. 优化文档切片大小

网络连接问题

本地服务访问：

- 确保Docker服务正常运行
- 检查端口占用情况
- 验证防火墙设置

第三方API调用：

- 验证API密钥有效性
- 检查网络连接稳定性
- 确认服务商API状态

11. 高级功能使用

API集成

RESTful API使用：

- 支持OpenAI兼容格式
- 可集成到其他应用中
- 提供完整的API文档

数据导入导出

知识库备份：

- 支持导出知识库配置
- 可备份处理后的向量数据
- 提供批量恢复功能

多用户管理

权限控制：

- 支持多用户注册
- 可设置知识库访问权限
- 提供使用统计功能

12. 最佳实践建议

知识库构建策略

文档质量要求：

- 使用结构清晰的文档
- 避免过多图表和格式
- 保持信息的时效性
- 建立标准化命名规范

性能监控

定期检查项目：

- 模型服务运行状态
- 系统资源使用情况
- 查询响应时间
- 错误日志信息

扩展应用

集成建议：

- 与星核其他AI功能联动
- 开发自定义应用接口
- 建立企业级知识管理体系
- 实现智能客服系统

通过本指南的详细配置，您可以充分发挥RagFlow在星核电脑上的强大功能，构建专业级的智能知识库系统。