openKylin 系统智能体用户手册

V0.9.2

openKylin AgentOS SIG

国防科技大学、哈尔滨工业大学（深圳）、麒麟软件有限公司

2026 年 07 月 10 日

目　　录

1. 产品概述

openKylin 系统智能体（KylinAgent）是一款面向 openKylin 桌面环境的智能助手应用，支持连接本地或远程智能体运行时服务，为用户提供实时流式对话、会话管理、模型配置、技能扩展、记忆管理、过程追踪和运行日志查看等能力。用户可通过自然语言与系统智能体交互，完成信息查询、文件处理、系统操作辅助和跨应用协同等任务。

图 1 openKylin 系统智能体主界面

后端能力说明：系统智能体后端基于主流开源智能体运行时进行优化增强，包括"主干-分支"协同执行架构、轻量级记忆精炼模型、统一模型推理服务、系统操控 Skill 集成、CUA 智能体 Skill 封装等。通过统一模型推理服务有效降低大模型的切换次数与任务等待时间；通过"主干-分支"协同执行架构可有效压缩 token 消耗；通过轻量级记忆精炼模型过滤冗余信息，重点保留高价值任务经验，有效提升历史经验复用准确率；通过 API 调用方式实现对具备 API 接口的系统配置和应用功能的直接操作；对于缺少 API 接口的桌面应用，则通过 CUA 智能体 Skill 封装，采用界面感知和键鼠模拟方式完成操作，从而实现跨应用协同。这些优化增强用于提升复杂任务拆解、系统级操作协同、技能复用和过程可追踪能力。

运行环境维护说明：应用启动时会检查智能体运行时启动命令、API Server 配置、后端运行时版本、Kylin CUA 运行时和 CUA Skill 同步状态。发现缺失或版本不一致时，会提示用户执行安装或更新，并在进度窗口中展示执行日志。

本地模型能力说明：模型设置页内置 Qwen3.6-35b 与 Holo3.1-35b 两个默认本地模型入口。模型下载完成后会注册到本地推理服务，并写入后端模型配置，随后可在聊天模型选择器中使用。

实时流式对话：发送消息后可持续接收模型回复，并可在生成中停止。

会话管理：支持新建、切换、重命名、删除多个会话，最近会话显示在左侧列表顶部。

多模型支持：可配置 OpenRouter、Anthropic、OpenAI 或自定义模型服务，并在聊天界面选择使用。

附件输入：支持通过按钮选择文件，也支持拖放本地文件到输入区。

过程详情：可查看单条助手消息的生成过程，也可预览和导出整个会话的详细日志。

设置中心：当前界面提供通用、模型、技能、个性、记忆、日志和用量七类设置。

本地状态保存：窗口状态、侧边栏显示、模型选择、会话状态等会持久化保存。

模块	**主要用途**
主界面	聊天、选择模型、选择请求优先级、查看 Token 与耗时、发送附件、管理会话。
设置-通用	主题、服务地址、服务密钥、托盘、委派模式、详情按钮、语音相关选项。
设置-模型	添加、搜索、编辑、删除聊天可用模型。
设置-技能	查看、搜索、创建、导入、编辑、卸载技能，并控制技能自进化。
设置-个性	编辑 SOUL.md，用于定义智能体人格、语气和长期指令。
设置-记忆	查看会话/消息/记忆统计，编辑智能体记忆和用户画像。
设置-日志	查看运行日志。
设置-用量	查看服务端会话数、消息数、Token 和预估费用。

2. 快速开始

2.1 安装

安装方式包括以下两种：

（1）新装系统

下载并安装预包含 openKylin 系统智能体的镜像,下载链接 https://factory.openkylin.top/kif/download/build/SpPwtLXhoJH7fv8MnEdFx46ai0szljW2/openKylin-Desktop-V2.0-SP2-AgentOS-20260407-x86_64.iso。

（2）旧版本更新

如果已经安装 openKylin 2.0 SP2 版本，可通过以下两种方式安装系统智能体：

a. 图形界面安装方式，打开软件商店，搜索“系统智能体”或“kylin-agent”进行下载安装；

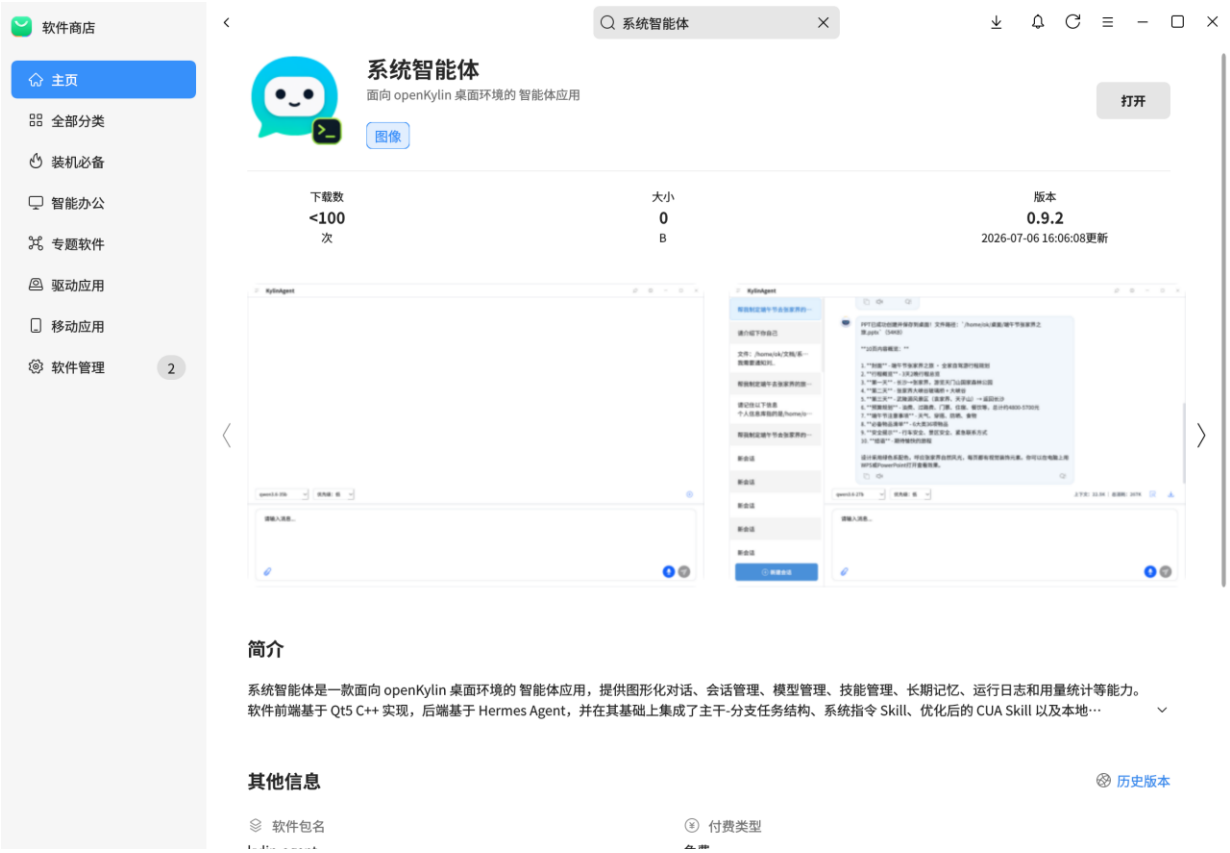

图 2 从软件商店下载安装

b. 终端安装方式，执行安装命令：sudo apt update && sudo apt install kylin-agent。（openKylin 2.0 SP2 版本引入磐石架构，默认不允许通过 apt install 方式安装软件包，可使用命令：sudo mm-cli -o&&reboot 进入维护模式再尝试执行上述安装命令。）

2.2 首次启动与运行环境检查

首次启动时，系统智能体会自动更新运行环境，弹出确认窗口，需点击“执行”后应用会自动完成应用运行环境更新操作。

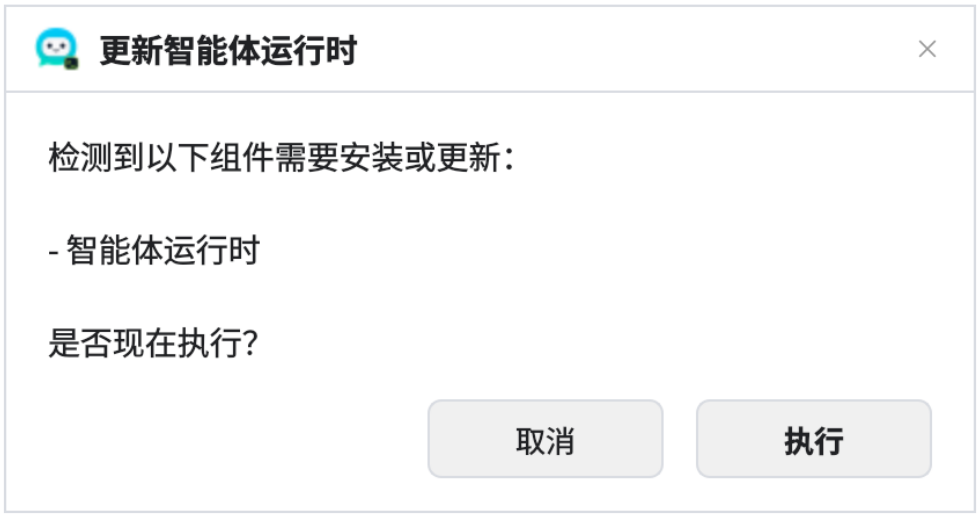

图 3 更新智能体运行时确认窗口

更新完成之后，会自动进入模型配置页面。

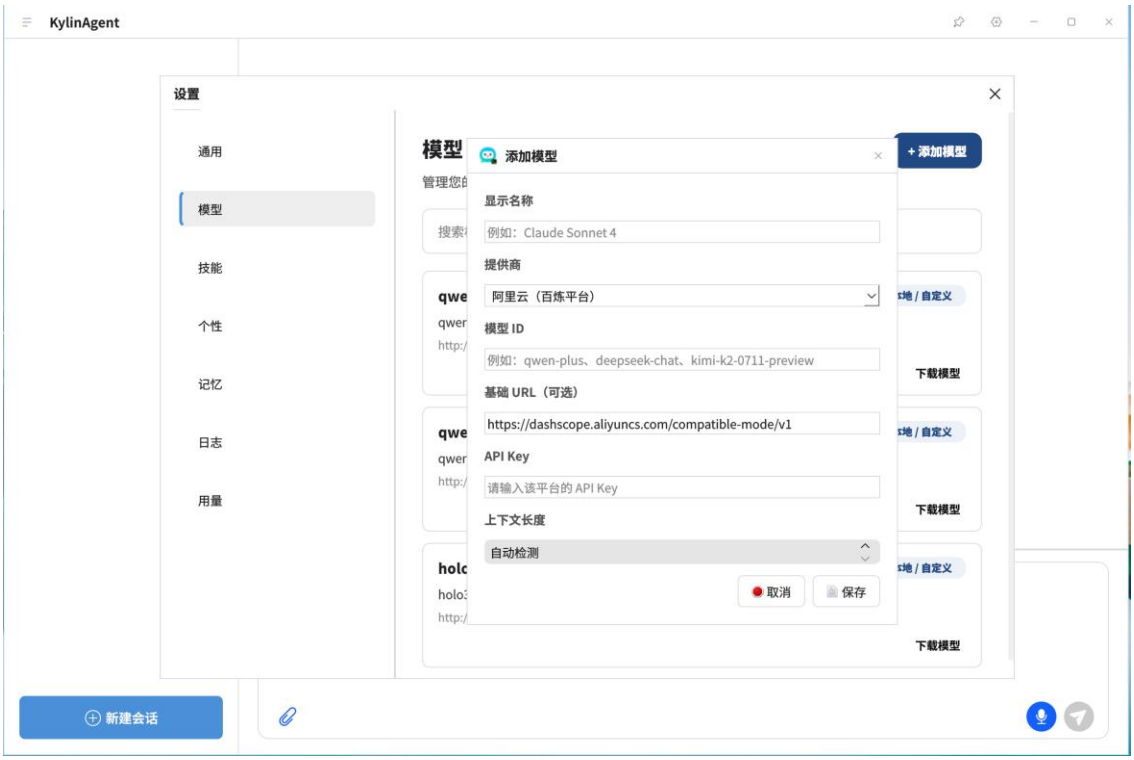

图 4 模型配置界面

2.3 模型配置

模型配置可采用云端推理和本地推理两种方式。

1. 云端推理方式:

(1) 添加大模型推理平台提供的模型服务，从添加模型窗口中的提供商中选择。目前支持阿里云（百炼平台）、百度智能云（千帆平台）、火山引擎（方舟平台）、科大讯飞（星火大模型）、Kimi（月之暗面）、MiniMax (M3)、深度求索（DeepSeek）、智谱 AI（GLM 系列）。根据提供商填入对应的显示名称、模型 ID 和 API Key，点击“保存”。

图 5 从选择的提供商中添加模型

(2) 添加其他的云端推理服务，填入对应的显示名称、模型 ID、基础 URL 和 API Key，点击“保存”。示例如下：

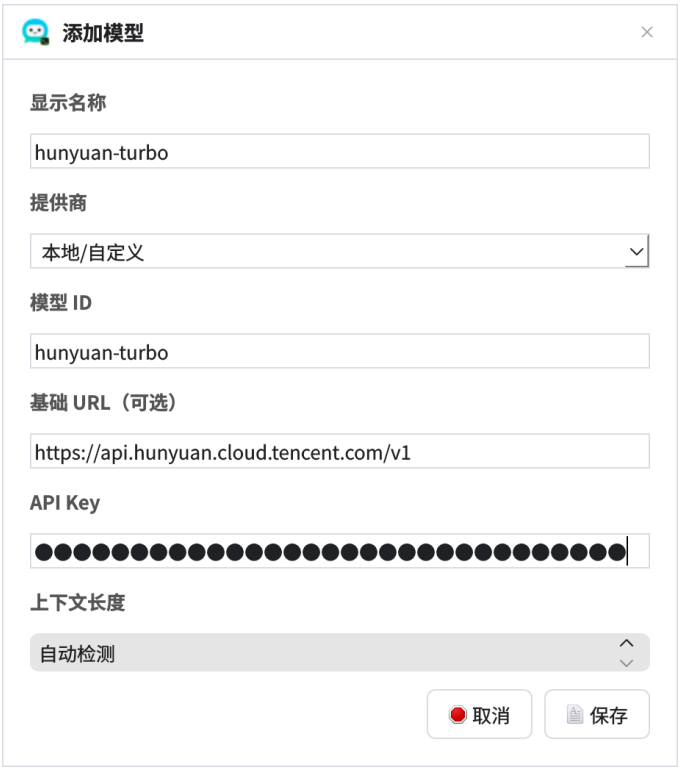

图 6 自定义方式添加云端模型

2. 本地推理方式

本地推理需使用 openKylin 系统智能体提供的模型统一推理服务（uni-model-infer）。该推理服务在主流推理框架的基础上，添加支持了最小切换调度、任务优先级调度，并提供友好的模型管理与服务监控界面。

（1） 安装统一推理服务

如果新装包含 openKylin 智能体的系统镜像，已默认集成统一推理服务，则略过该安装步骤。

如果基于 openKylin 2.0 SP2 版本更新，可通过以下两种方式安装统一推理服务：

a. 图形界面安装方式，打开软件商店，搜索“统一模型推理”或“uni-model-infer”进行下载安装；

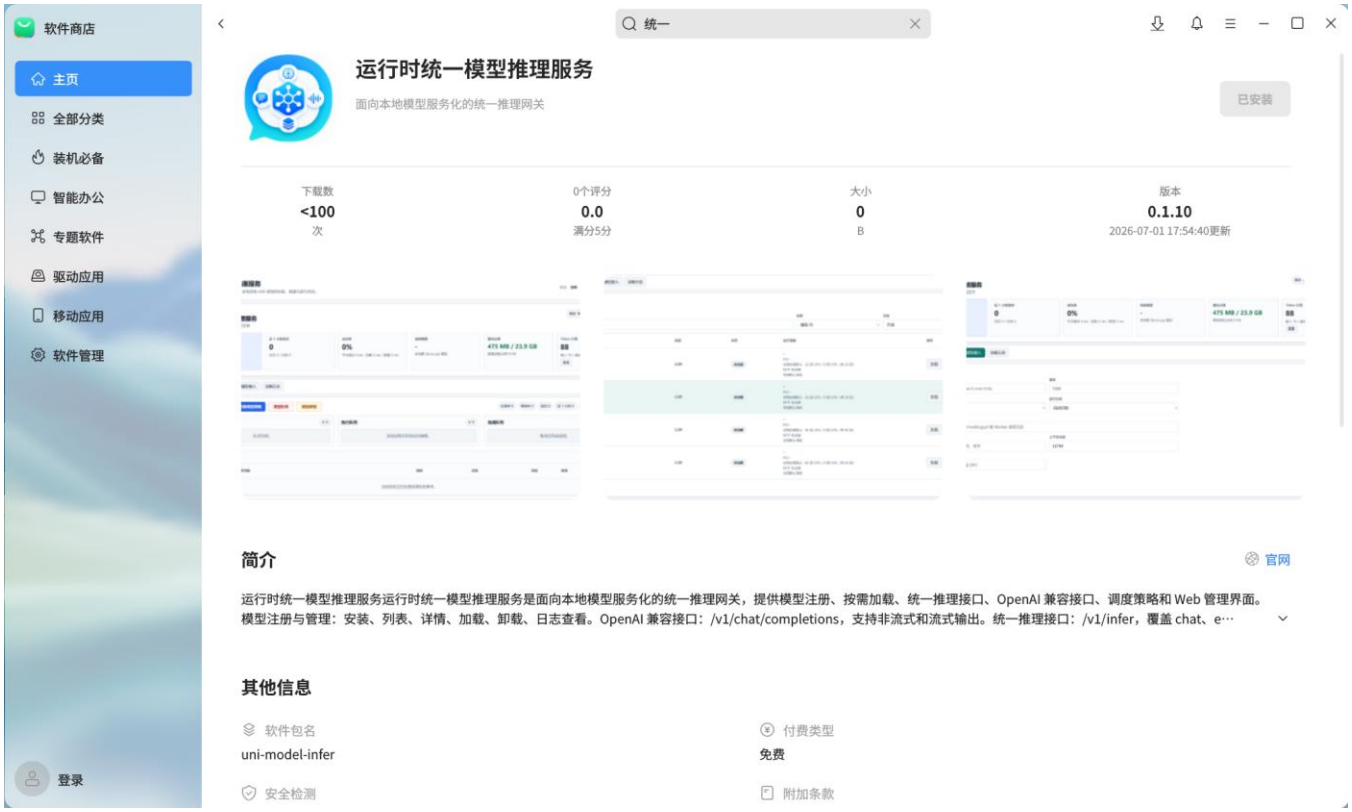

图 7 从软件商店搜索下载安装

b. 终端安装方式，执行安装命令：sudo apt update && sudo apt install uni-model-infer。（openKylin 2.0 SP2 版本引入磐石架构，默认不允许通过 apt install 方式安装软件包，可使用命令：sudo mm-cli -o&&reboot 进入维护模式再尝试执行上述安装命令。）

（2）环境检测

首先通过浏览器进入 WebUI 控制页面“127.0.0.1:18080”进入环境检测步骤，按要求检测显卡驱动、CUDA 环境以及 llama.cpp 环境。检测不通过则建议按照页面显示的安装步骤依次安装显卡驱动、CUDA 环境以及 llama.cpp 环境，点击“重新检测”按钮。验证通过后可以点击“进入控制台”进入运行时统一模型推理服务 WebUI 页面。

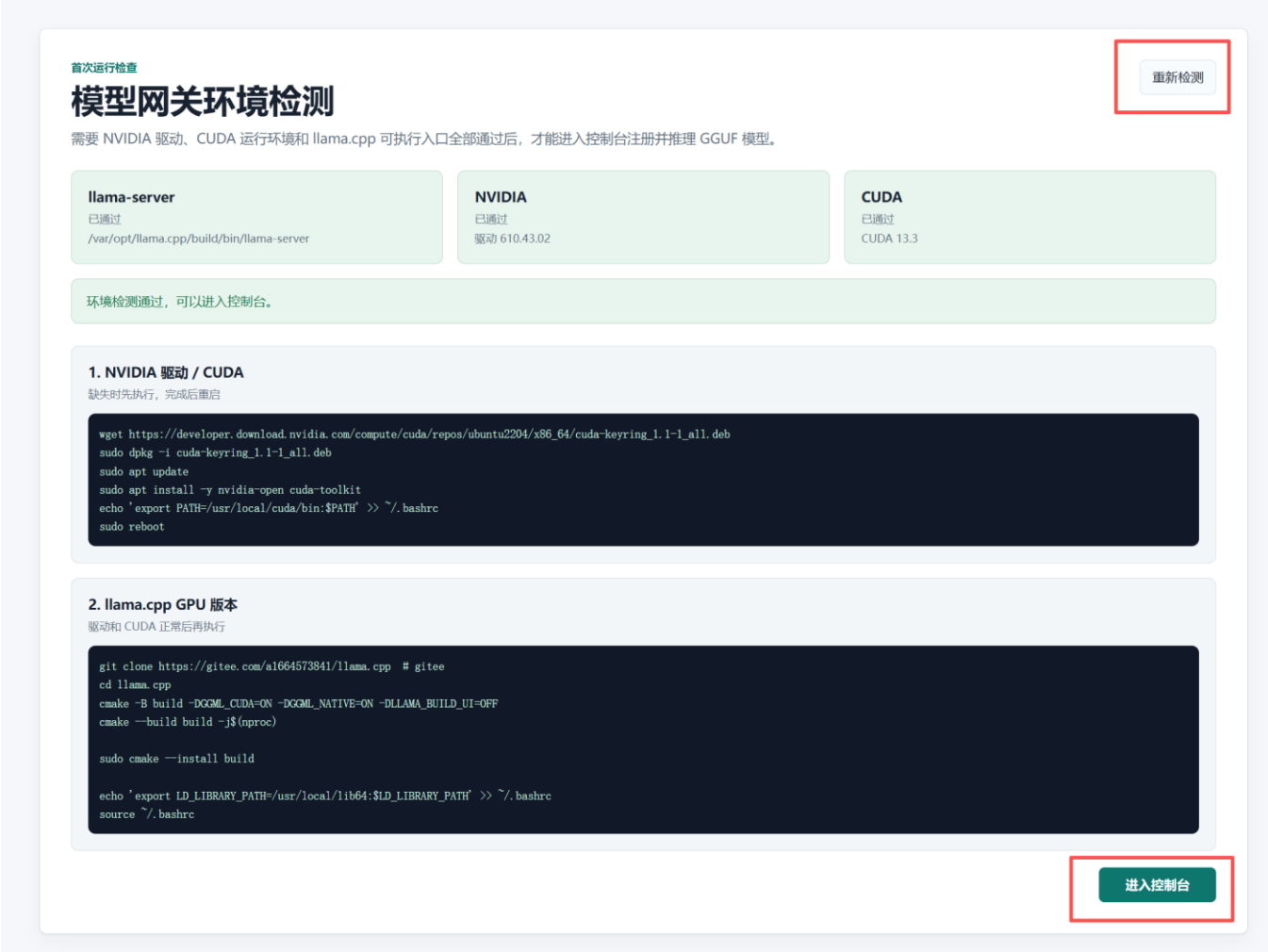

图 8 环境检测页面（按步骤完成推理环境配置）

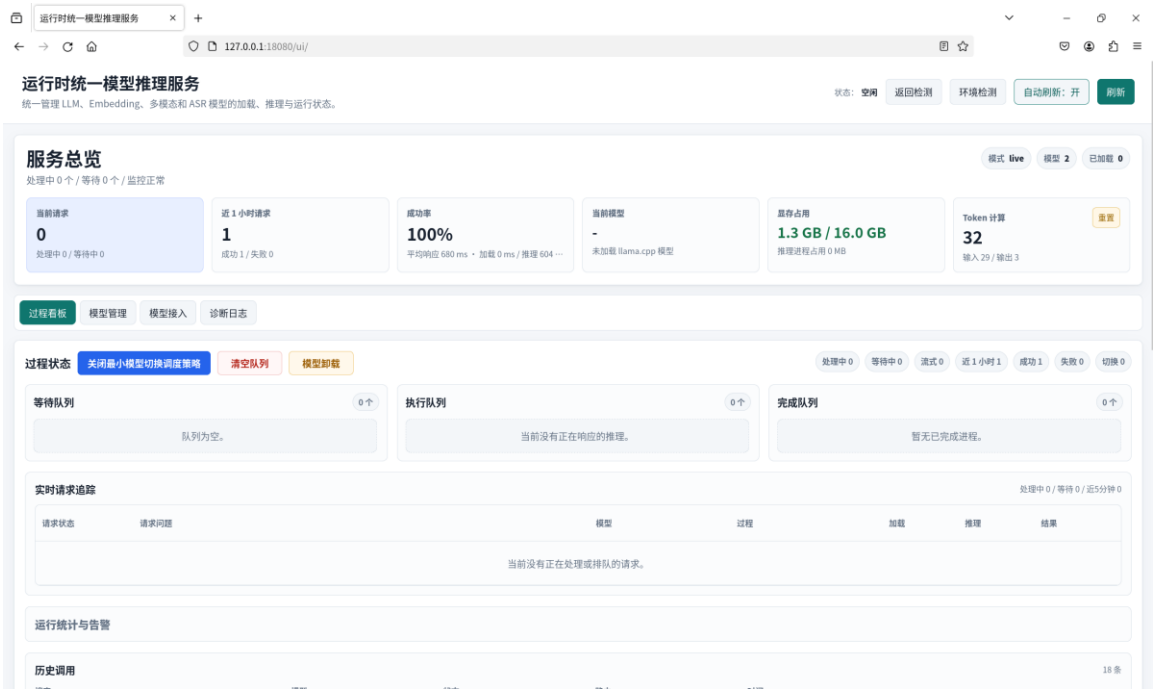

图 9 运行时统一模型推理服务 WebUI 页面

（3）模型接入

点击“模型接入”页面，填写模型 ID 以及对应模型（支持拖拽获取文件路径），多模态模型需要增加 mmproj GGUF 模型，添加完成以后，点击注册模型即可增加成功。

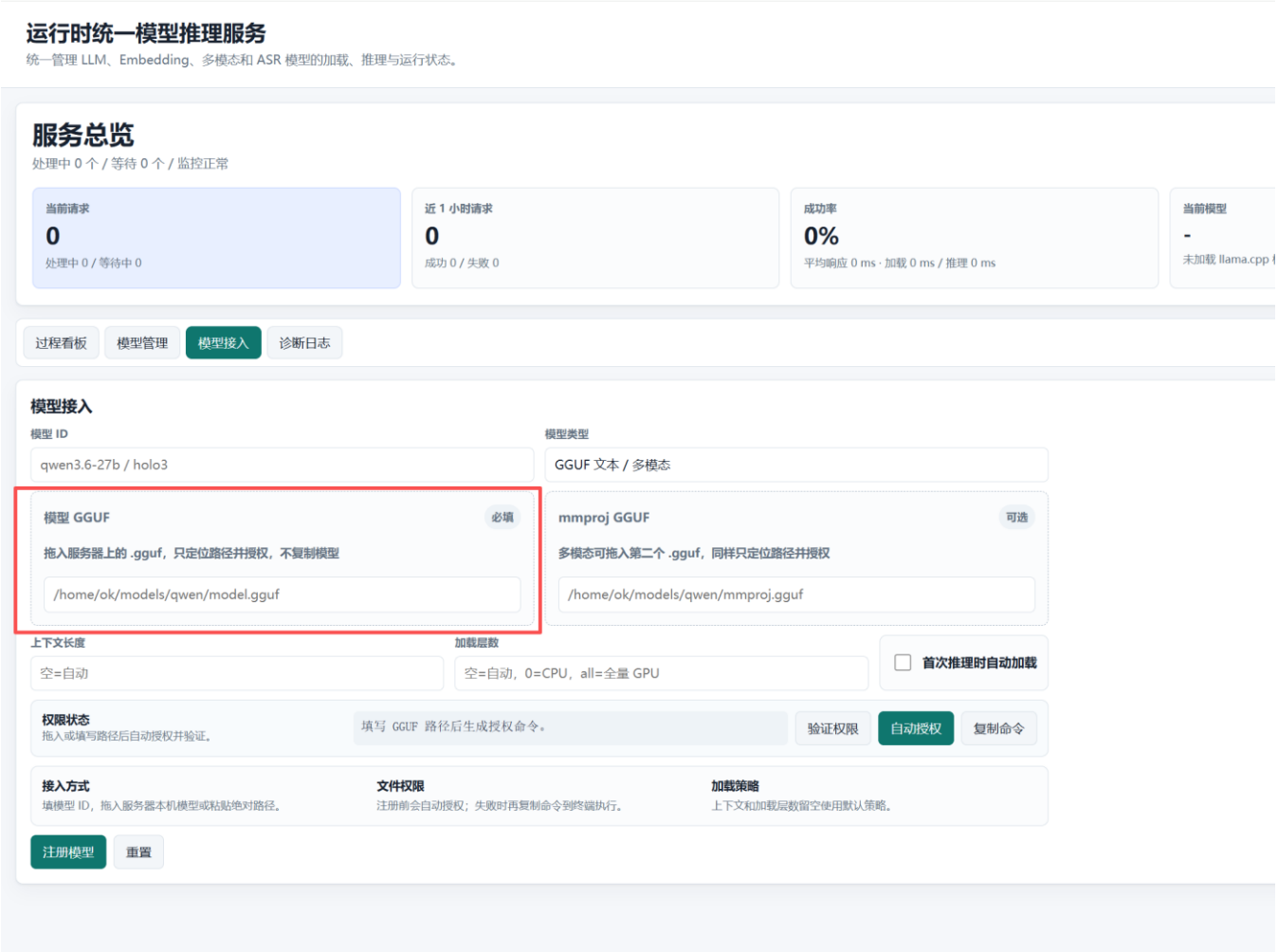

图 10 模型接入（可拖拽模型导入）

图 11 模型接入成功后可在模型管理显示

（4）配置模型

安装“统一推理服务”完成后，可以在“设置”中的“模型”页面，下载模型或手动添加模型。

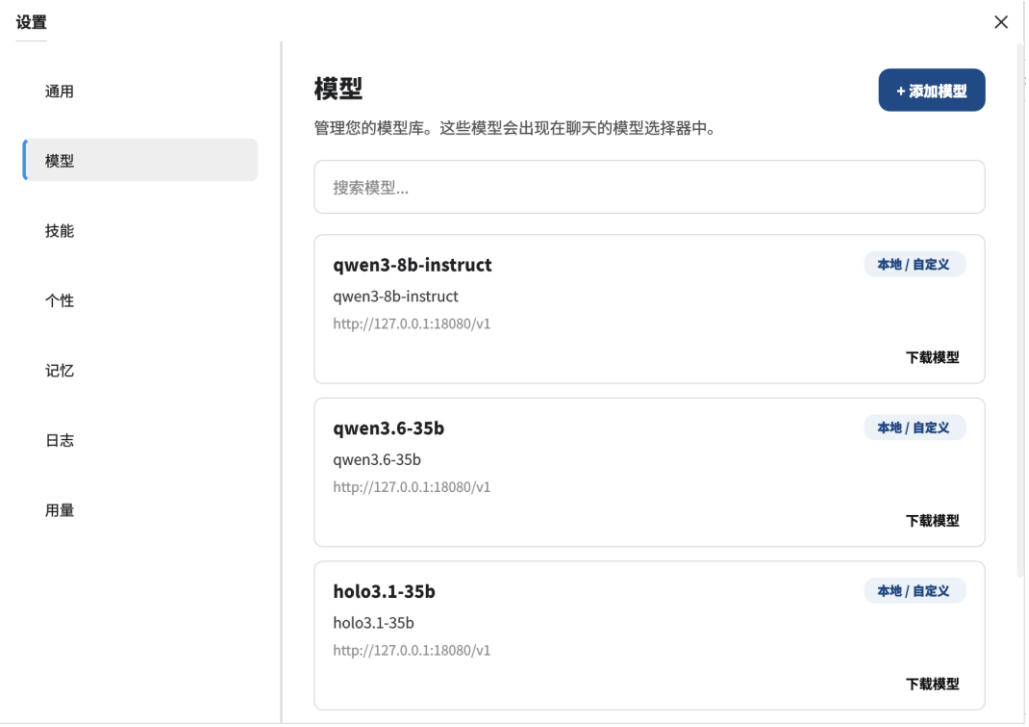

图 12 支持大模型统一推理服务的默认模型

图 13 添加大模型统一推理服务提供的模型

2.4 基本使用流程

1. 打开系统智能体，应用会自动加载最近会话；如果没有会话，可点击“新建会话”。
2. 在聊天区底部选择模型，并根据任务紧急程度选择“优先级：低/中/高”。
3. 在输入框中输入问题；如需提供文件，点击附件按钮选择文件，或将文件拖入输入区。
4. 点击发送按钮开始对话。生成过程中发送按钮会变为停止按钮，可用于中止当前回复。
5. 需要排查或回顾模型行为时，点击详情预览或单条助手消息上的详情按钮查看过程信息。

3. 主界面使用

3.1 窗口与标题栏

窗口采用无边框设计。标题栏左侧为侧边栏开关，中间显示 系统智能体，右侧依次提供靠边固定、设置、最小化、最大化/还原、关闭到托盘等按钮。

侧边栏开关：显示或隐藏左侧会话列表。隐藏后聊天区会显示独立的新建会话按钮。

靠边固定：用于将窗口固定到边缘或保持在合适位置，适合边工作边对话。

设置：打开设置中心。

关闭按钮：默认隐藏到系统托盘；可通过托盘菜单显示窗口或退出应用。

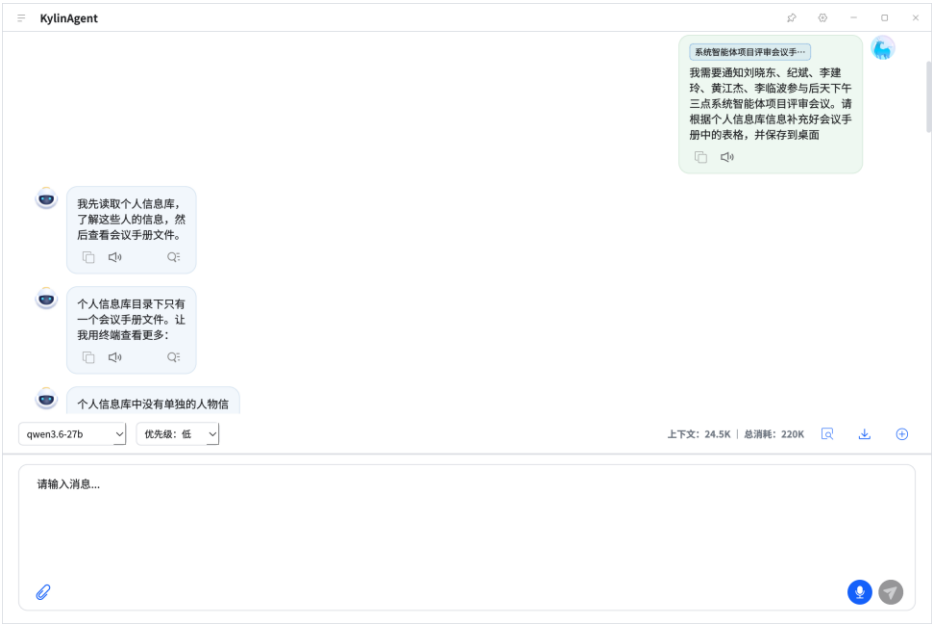

图 14 隐藏会话列表后的主界面

图 15 靠右贴边使用效果

3.2 会话列表

左侧会话列表展示最近会话，新建会话位于列表底部按钮。选择某个会话即可加载该会话的历史消息。

新建会话：点击“新建会话”创建空白会话，并自动切换到该会话。

重命名会话：在会话项上打开右键菜单，选择“重命名”，输入新名称后确认。

删除会话：在会话项右键菜单选择“删除”；多选时可删除多个会话。删除前会有确认。

运行状态：当某个会话正在流式生成时，列表会更新对应状态；其他会话有新内容时可标记未读。

图 16 隐藏侧边栏时的新建会话入口

3.3 聊天工具栏

聊天区下方工具栏包含模型选择器、请求优先级、会话统计、详情预览、导出和新建会话入口。部分按钮会根据通用设置和当前会话状态显示或隐藏。

控件	说明
模型选择	选择本轮对话使用的模型。列表来自“设置 > 模型”配置。
优先级	可选低、中、高。该值随会话保存，并随请求发送给智能体运行时服务。
会话统计	显示上下文、总消耗、耗时等信息；鼠标悬停可查看输入/输出/缓存/推理/API 调用等详细统计。

详情预览	打开会话详细信息预览，支持按关键信息、分组、时间查看。
导出	将当前会话详细信息导出为日志或文本文件。
新建会话	侧边栏隐藏时提供快捷新建入口。

3.4 输入、附件与发送

输入框提示为“请输入消息...”。输入内容或添加附件后，发送按钮会自动启用。

点击附件按钮可选择一个或多个本地文件。也可以将本地文件直接拖入输入区或消息区域。

附件会以标签形式显示在输入框上方；点击附件标签上的移除按钮可取消发送。

发送后，附件路径会以“文件：...”形式随用户消息发送给后端。

语音相关功能目前为预留或部分实现功能。界面中已提供语音按钮，通用设置中包含语音唤醒、识别后自动发送和自动播报等选项；其中部分能力需依赖后续版本或额外语音组件支持，具体以当前版本实际提示为准。

3.5 消息操作与过程详情

每条消息下方提供常用操作。用户和助手消息均可复制；助手消息还可展开生成过程。

复制：将当前消息内容复制到剪贴板。

展开/收起过程：助手消息可展开本条消息的思考、工具调用、返回结果摘要等详情。

会话详情预览：从工具栏打开，可按“关键信息”“分组排序”“时间排序”三种方式查看，并可筛选事件或导出当前预览列表。

图 17 会话详细信息预览

4. 设置中心

点击主窗口右上角的设置按钮进入设置中心。左侧为分类列表，当前实现并显示的分类包括：通用、模型、技能、个性、记忆、日志、用量。设置窗口支持拖动和调整大小。

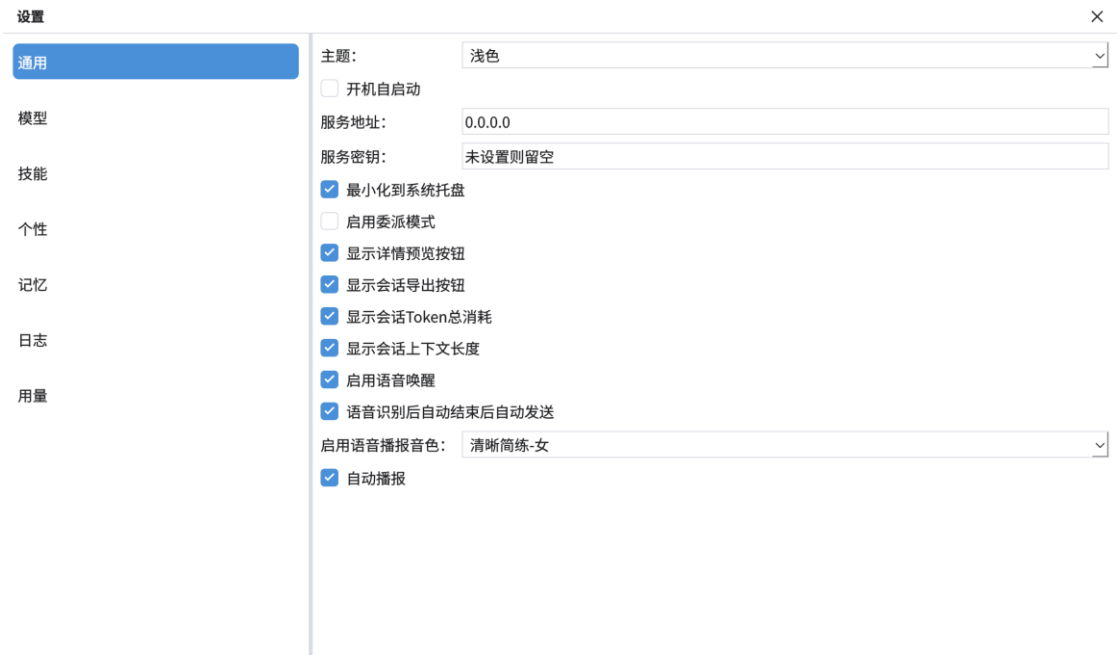

图 18 通用设置

4.1 通用设置

设置项	用途
主题	跟随系统、浅色、深色三种模式。切换后界面立即应用。
开机自启动	控制应用是否随系统启动。
服务地址	智能体运行时服务地址。编辑完成后自动规范化并保存。
服务密钥	服务鉴权密钥，未设置则留空。输入框以密码形式显示。
最小化到系统托盘	控制关闭/最小化行为是否进入托盘。
启用委派模式	控制是否启用委派模式。
显示详情预览按钮	控制聊天工具栏是否显示详情预览按钮。
显示会话导出按钮	控制聊天工具栏是否显示导出按钮。
显示会话 Token 总消耗	控制会话统计中是否显示总 Token 消耗。
显示会话上下文长度	控制会话统计中是否显示上下文长度。
语音相关选项	包含语音唤醒、识别后自动发送、音色选择、自动播报。

4.2 模型设置

模型页用于管理聊天界面可选模型。模型信息来自当前配置的智能体运行时服务，保存时通过服务配置 API 写回。

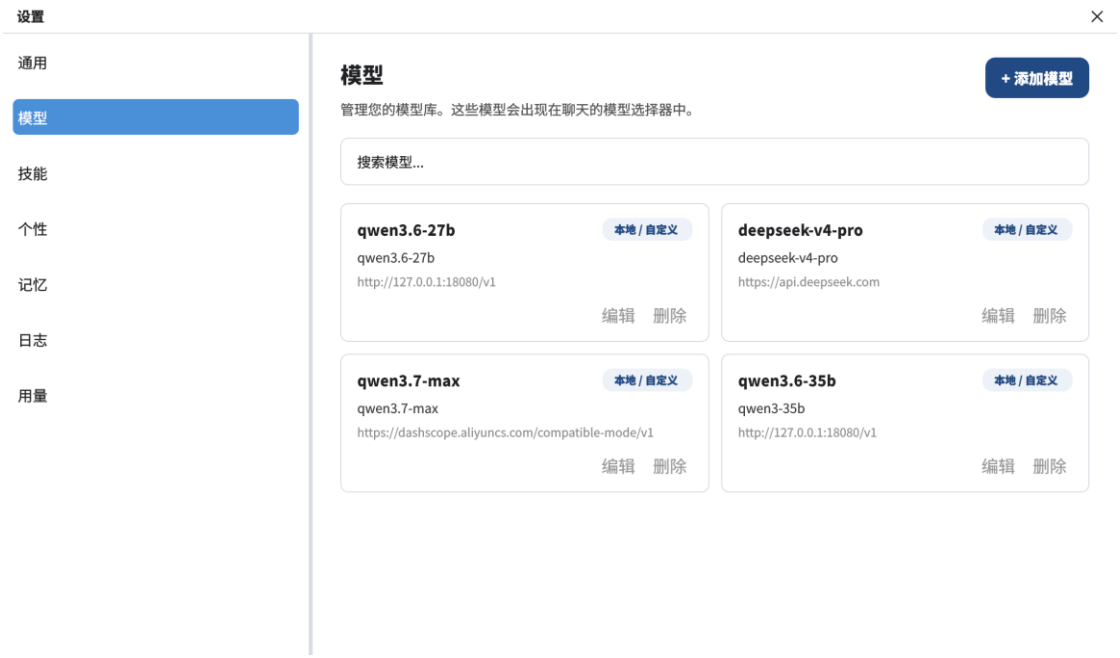

图 19 模型列表

搜索模型：在搜索框输入名称、模型 ID、提供商或基础 URL 的关键字。

添加模型：点击“+ 添加模型”，填写显示名称、提供商、模型 ID、基础 URL 和上下文长度。名称与模型 ID 为必填。

编辑模型：点击模型卡片上的“编辑”，修改字段后保存。

删除模型：点击“删除”并在确认框中确认。

上下文长度：可手动填写 token 数；设置为 0 时显示为“自动检测”。

字段	说明
显示名称	在聊天模型选择器和模型卡片中显示的名称，例如 Claude Sonnet 4。
提供商	支持 OpenRouter、Anthropic、OpenAI、本地/自定义；显示时也可识别 Ollama 等服务端返回项。
模型 ID	实际请求的模型标识，例如 anthropic/claude-sonnet-4-20250514 或 gpt-4.1。
基础 URL	自定义或本地模型服务端点，例如 http://localhost:1234/v1，可留空。
上下文长度	模型上下文 token 上限，用于统计与显示。

图 20 添加模型

图 21 编辑模型

默认本地模型下载与推理服务注册，模型页内置 qwen3.6-35b 与 holo3.1-35b 两个默认本地模型。未下载时卡片显示“下载模型”，下载中显示进度条和“取消下载”，下载完成后可编辑或删除。

默认模型：qwen3.6-35b、holo3.1-35b。

模型保存位置：~/.local/share/kylin-agent/models/<模型 ID>/；下载文件包含 model.gguf 与 mmproj.gguf。

下载完成后应用会生成 md5sum 标记，并自动调用 http://127.0.0.1:18080/v1/models:install 注册本地模型推理服务；注册成功后生成 install 标记。

模型写入智能体运行时后端配置后生成 config 标记，后端模型地址为 http://127.0.0.1:18080/v1，随后刷新聊天模型选择器。

注册推理服务依赖 openKylin apt 包 uni-model-infer；如需目录权限，应用会提示输入当前系统用户密码，并执行 sudo model-gateway-grant-access 授权模型目录访问。

4.3 技能设置

技能页用于扩展智能体能力。技能以卡片形式展示，支持搜索、预览、创建、导入、编辑和卸载。

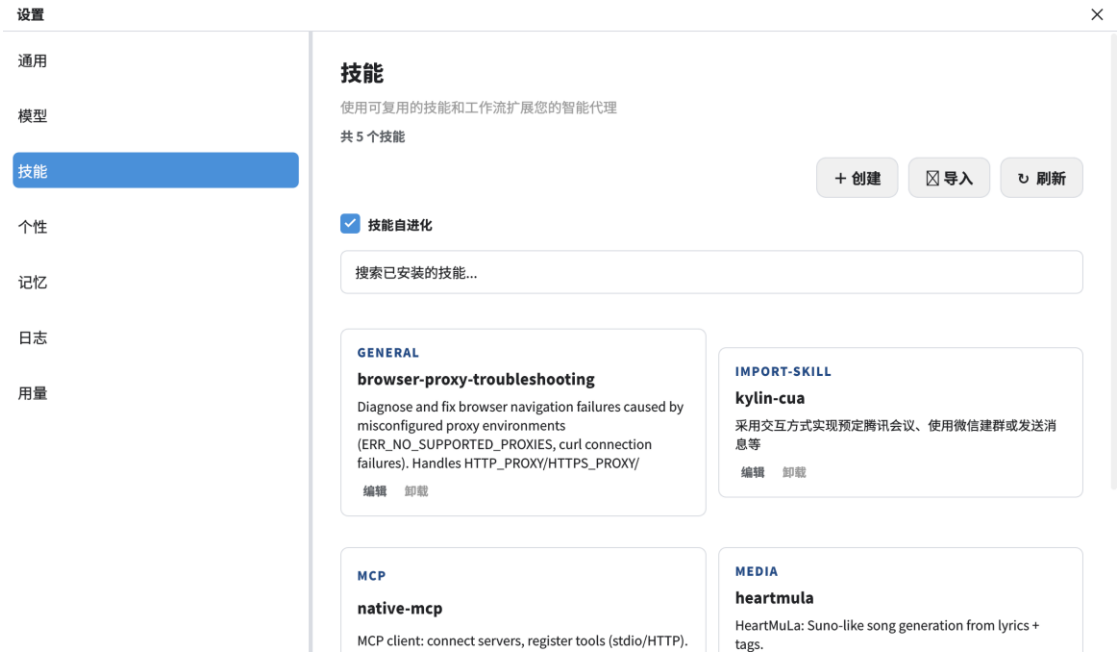

图 22 技能列表

创建：打开技能编辑器，从空白模板创建新的技能目录与 SKILL.md。

导入：支持导入 SKILL.md 文件、技能目录或压缩包。导入前会检查技能名称与 SKILL.md。

刷新：重新从当前智能体运行时服务读取技能列表。

技能自进化：允许代理在复杂任务后沉淀技能，并由后台维护代理整理技能库。

搜索：按技能名称、描述或分类过滤已安装技能。

预览：点击技能卡片查看 SKILL.md 内容。

编辑：在本地存储模式且技能目录存在时，可直接编辑技能文件树。

卸载：删除已安装技能，操作前会有确认。

图 23 技能搜索

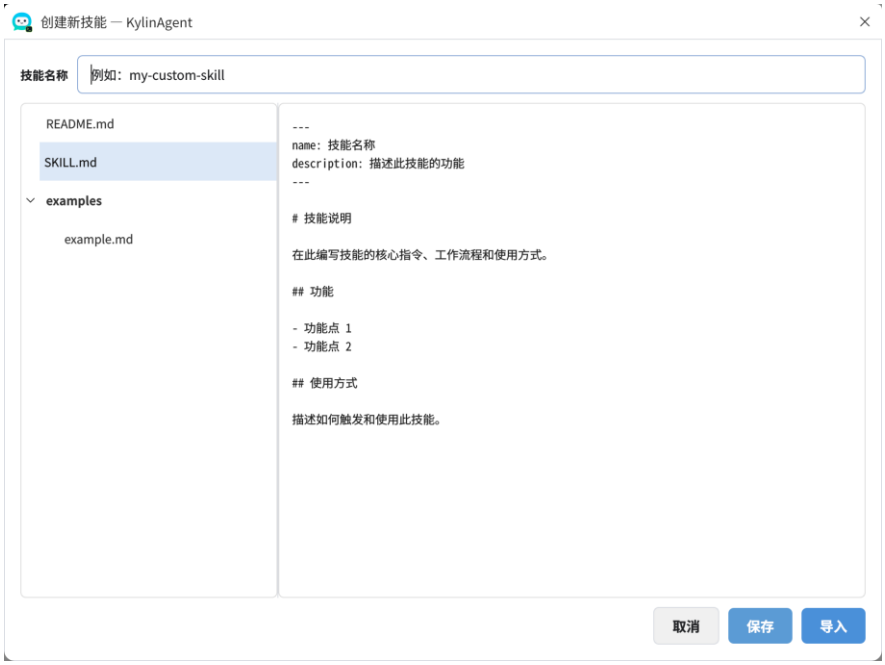

图 24 创建技能

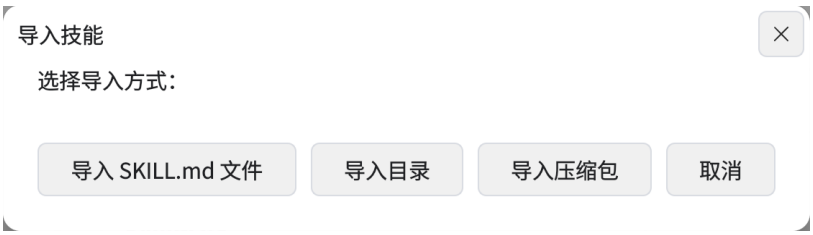

图 25 导入技能方式

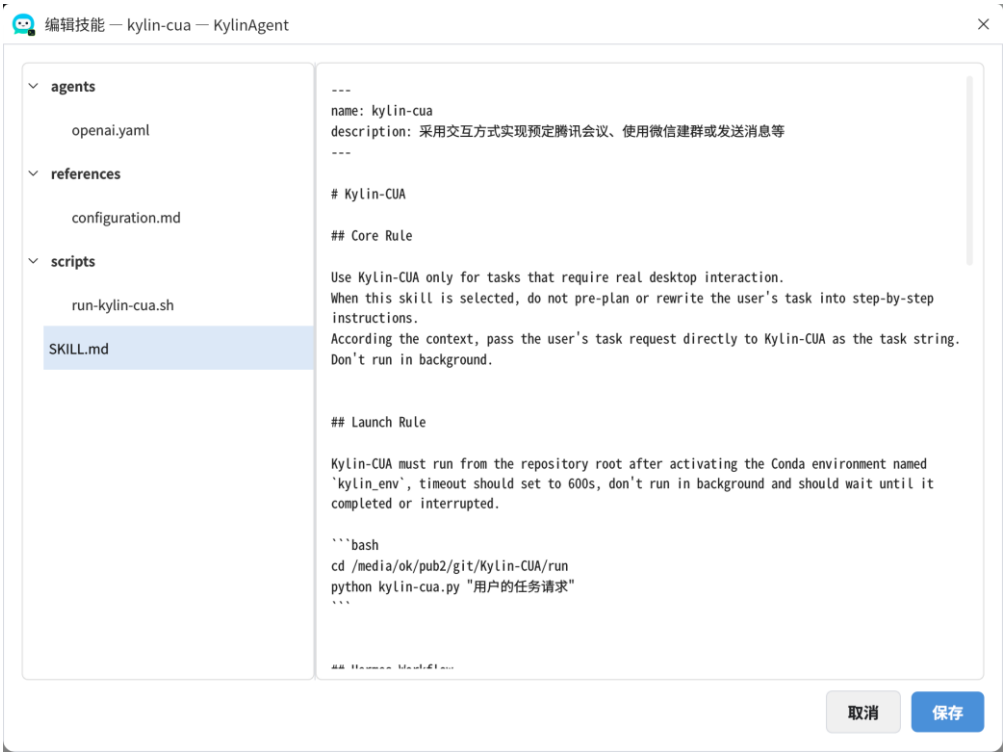

图 26 编辑技能

4.4 个性设置

个性页通过 SOUL.md 定义智能体人格、语气和长期指令。内容在每次会话时重新读取，修改后无需重启。

图 27 个性设置

在编辑器中输入人格、沟通风格、长期偏好或行为规则。

内容变更后会自动保存，并短暂显示“已保存”。

点击“重置”可恢复默认 SOUL.md 内容，重置前会弹出确认。

4.5 记忆设置

记忆页用于管理跨会话的智能体记忆和用户画像，并展示会话、消息、记忆数量以及容量占用。

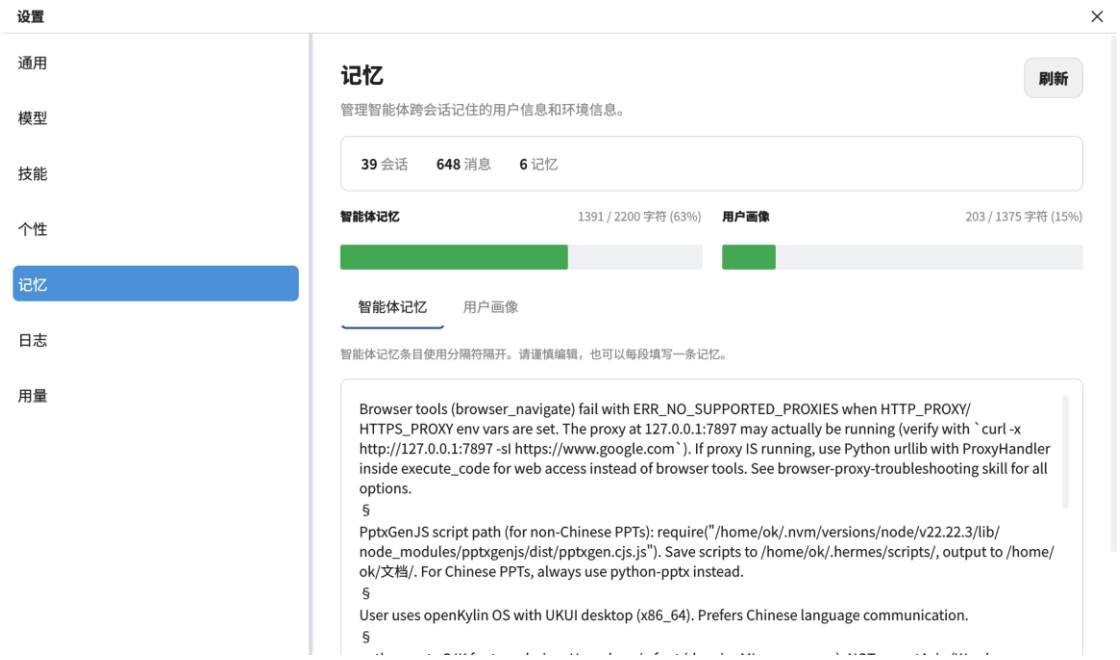

图 28 智能体记忆

图 29 用户画像

顶部统计：显示当前服务中的会话数、消息数和记忆条目数。

容量进度：智能体记忆和用户画像分别展示字符容量占用。

智能体记忆：用于记录代理跨会话保留的事实、偏好或环境信息。可按段落填写，条目由分隔符区分。

用户画像：填写姓名、角色、偏好、沟通风格、常用系统和时区等信息。

保存：编辑完成后点击“保存”写回当前内容。

4.6 日志信息

日志页面用于查看运行日志。左侧列出日志文件及大小，点击文件后右侧显示内容。大文件仅显示最后 200 KB，并自动滚动到底部。

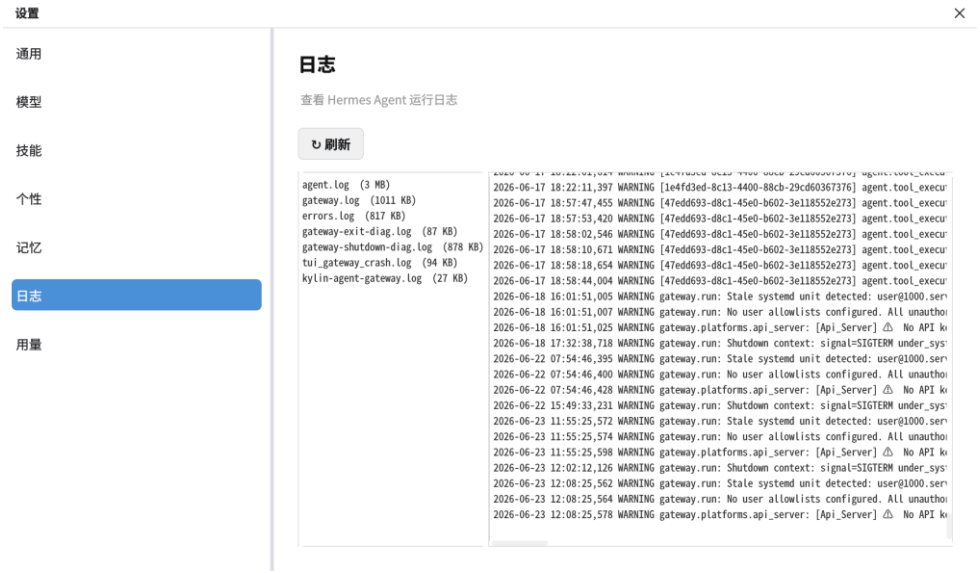

图 30 日志信息

4.7 用量统计

用量页展示当前智能体运行时服务返回的统计数据，包括总会话数、总消息数、输入 Token、输出 Token 和预估费用 USD。点击刷新可重新读取。

图 31 用量统计

5. 常见问题

问题	处理方法
聊天界面没有可选模型	进入“设置 > 模型”添加模型；确认智能体运行时服务的模型配置接口可用。
发送按钮不可用	确认输入框中已有文字或附件；空消息不会启用发送。
看不到详情预览或导出按钮	进入“设置 > 通用”，启用“显示详情预览按钮”或“显示会话导出按钮”。
技能无法直接编辑	只有本地存储模式且技能目录存在时才显示编辑入口；远程服务模式下可导入同名技能包覆盖。
日志为空	确认智能体运行时服务已运行并产生日志；点击刷新重新读取。
关闭后窗口不见了	默认会隐藏到系统托盘。可从托盘图标菜单选择“显示窗口”，或选择“退出”彻底关闭。
注册模型推理服务失败	确认已安装 uni-model-infer，且本地推理服务 127.0.0.1:18080 可用；重新启动应用可触发自动修复。
模型已下载但聊天中不可选	等待启动时“正在配置模型和注册模型推理服务...”完成；如仍失败，可在“设置 > 模型”删除后重新下载。
启动时提示更新智能体运行时	说明智能体运行时、CUA 运行时或 CUA Skill 缺失/版本不一致，点击“执行”即可自动安装或更新。

6. 附录：依赖安装

项目依赖 Qt 5.12+，以及 Qt5 的 Core、Gui、Widgets、Sql、Network、Xml、Svg 模块。SQLite 驱动为运行时必需，用于本地状态数据库。

系统	安装命令	说明
openKylin	sudo apt install build-essential cmake qtbase5-dev libqt5sql5-sqlite libqt5svg5-dev zlib1g-dev git curl rsync python3-pip libc6 libgcc-s1 | libgcc1 libstdc++6 zlib1g aria2 libqt5core5a libqt5gui5 libqt5widgets5 libqt5sql5 libsqlite3-dev libqt5network5 libqt5xml5 libqt5svg5 xclip xdotool wmctrl x11-utils x11-xserver-utils xdg-utils libglib2.0-bin libgtk-3-bin scrot python3-tk uni-model-infer kylin-agent-runtime-cache	提供 C++17 编译器、CMake、Qt5 开发组件、SQLite 驱动、QtSvg、ZLIB 开发库、Git/Curl/Rsync、Python pip，以及注册本地推理服务所需的 uni-model-infer。

如果运行时提示无法加载 SQLite、SVG 或本地模型推理服务相关组件，请优先确认 libqt5sql5-sqlite、libqt5svg5、zlib1g、uni-model-infer 等运行依赖已安装。其中 uni-model-infer 需要确保已经安装 NVIDIA 显卡驱动以及 CUDA13 及以上。