2025

开源操作系统助力 RISC-V架构上的AI计算能力

先进计算与关键软件海河实验室基础软件部　部长
OpenAtom openKylin社区技术委员会　委员
王文竹

开放原子开源基金会 OPENATOM FOUNDATION | openKylin

2025年7月18日

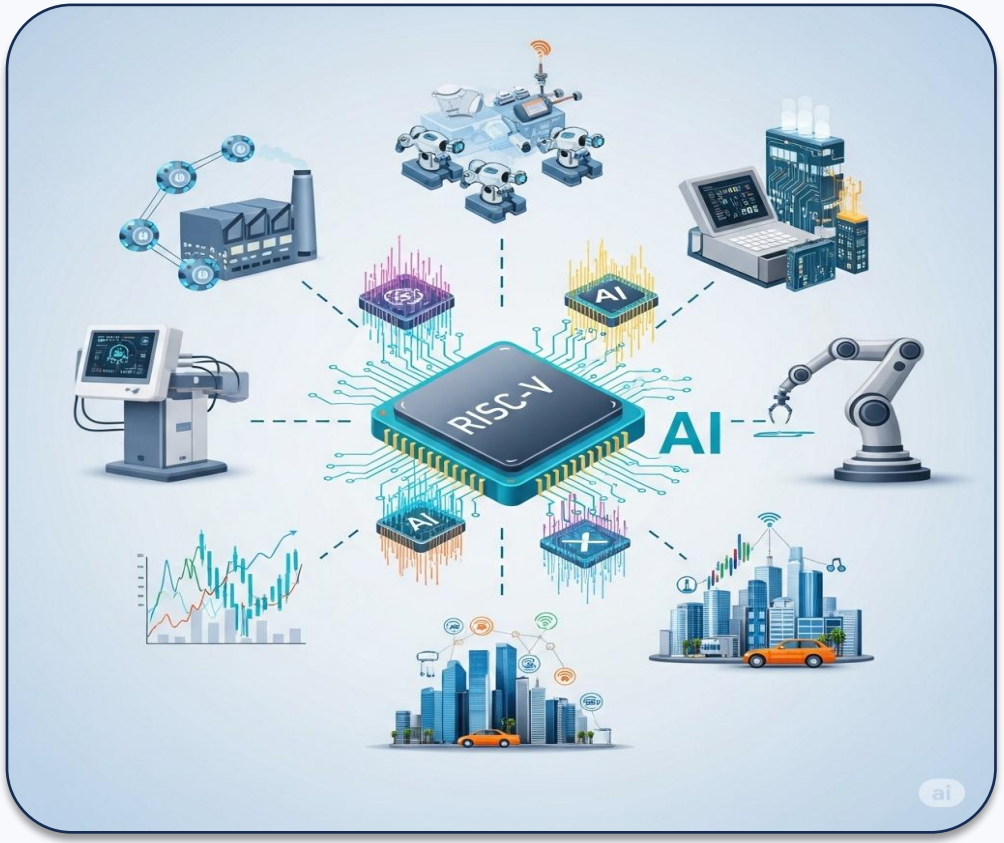
— OPENATOM OPENKYLIN —

人工智能技术的发展

开放原子开源基金会 OPENATOM FOUNDATION | openKYLin

人工智能正重塑数字世界底层逻辑，改变各行业的运作模式

逻辑	从基于规则到基于学习
数据	从数据存储到数据智能
行为	从人工操作到自主化



RISC-V在AI时代的机遇

开放原子开源基金会 OPENATOM FOUNDATION | openKYLin

多元AI算力

指 令 集：向量、矩阵、张量

体系架构：多核、 异构计算

AI加速器：GPU、DSA

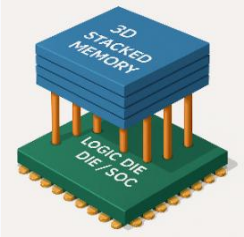
先进封装：Chiplet

RISC-V®

RVV

Matrix

Tensor



开源AI生态

AI 框架：PyTorch、ONNX

算子库：矩阵乘、激活函数

扩展性：通讯协议、通讯库

加速卡：硬件驱动、运行时库

PyTorch

Triton

RISC-V目前在AI时代的挑战

开放原子开源基金会 OPENATOM FOUNDATION | openKYLin

AI软件生态成熟度不足	AI框架移植、编译器支持等
AI应用开发环境不完善	算力和模型管理、开发接口等
AI计算性能优化刚起步	算子库、存储、分布式优化等

操作系统的发展历程

经过半个多世纪的发展与演化，操作系统的主要任务一直没有变化

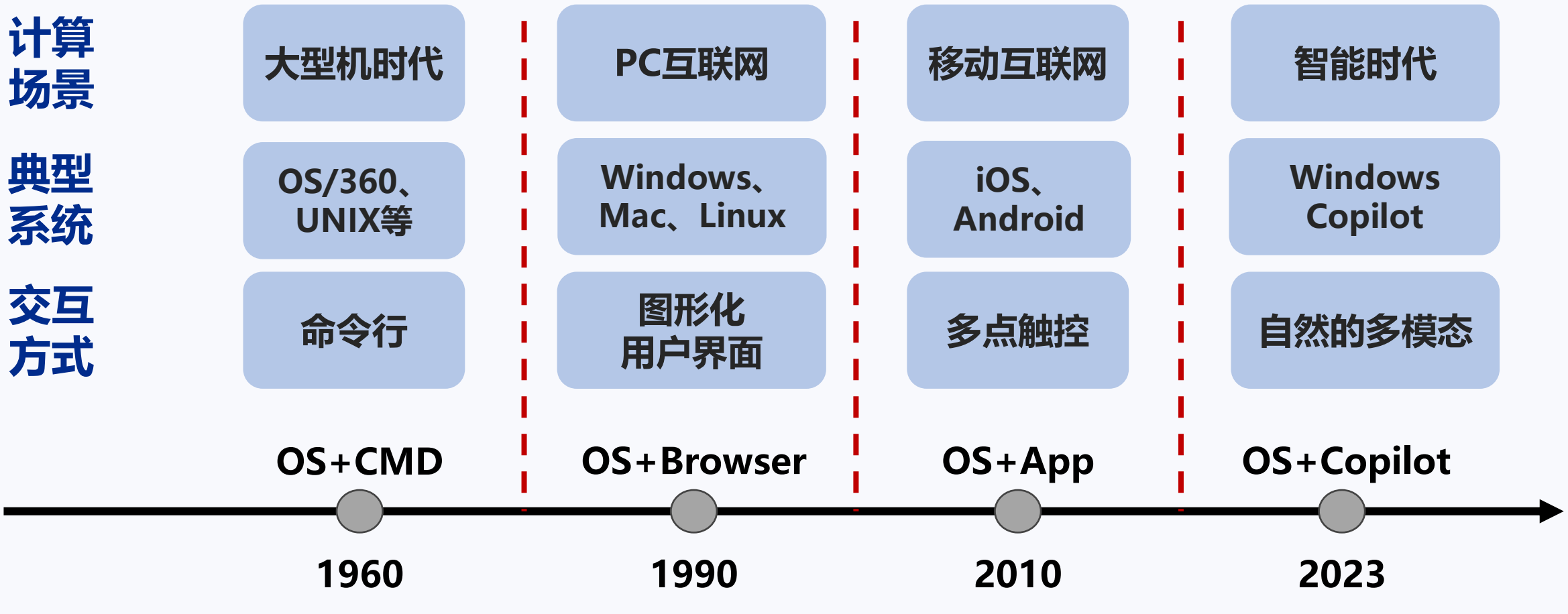
- 高效管理计算机系统中的各种软硬件资源
- 为上层软件和用户提供运行环境（命令）和开发环境（API）

操作系统发展的两大源动力

- 计算机系统的不断发展
- 计算机应用的不断发展

操作系统的发展历程

开放原子开源基金会 OPENATOM FOUNDATION | openKylin



操作系统的演化

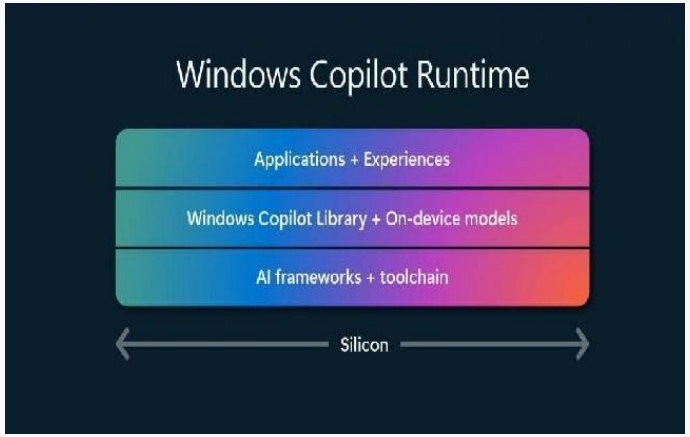
国外操作系统在AI上的发展

开放原子开源基金会 OPENATOM FOUNDATION | openKylin

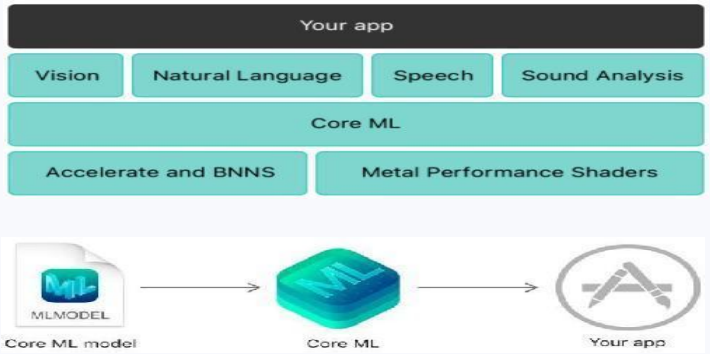
5月20号发布 Copilot+PC

6月11号发布 Apple Intelligence

8月13号在“Made by Google”大会宣布将Gemini和Android深度融合

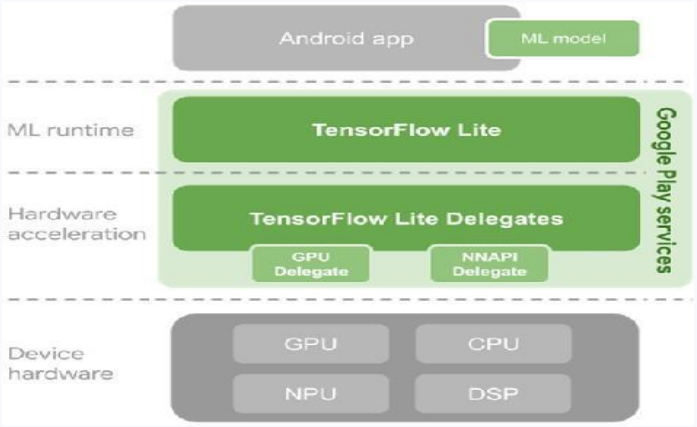


1. 通过Copilot Runtime给提供AI SDK
2. 通过onnxruntime支持本地模型推理
3. 通过DirectML支持各种硬件加速
4. 集成40+本地模型以及云端模型



Core Machine Learning

1. 提供自然语言处理、视觉处理、语音处理能力的多种API
2. 支持硬件加速和本地模型推理
3. 集成本地模型、云端模型



1. 通过AI Edge SDK和AICore提供APP调用本地AI能力
2. 支持本地模型推理和硬件加速
3. 集成Gemini以及云端模型

AI时代OS的两大推动力

开放原子开源基金会 OPENATOM FOUNDATION | openKylin

强大的新推理引擎

- 管理更复杂的本地资源
- 提供各种数据源和服务

更自然的交互方式（自然的多模态）

- 注重人机协作和智能化的用户界面
- 语言理解和机器学习技术理解用户需求，提供个性化的建议和辅助功能

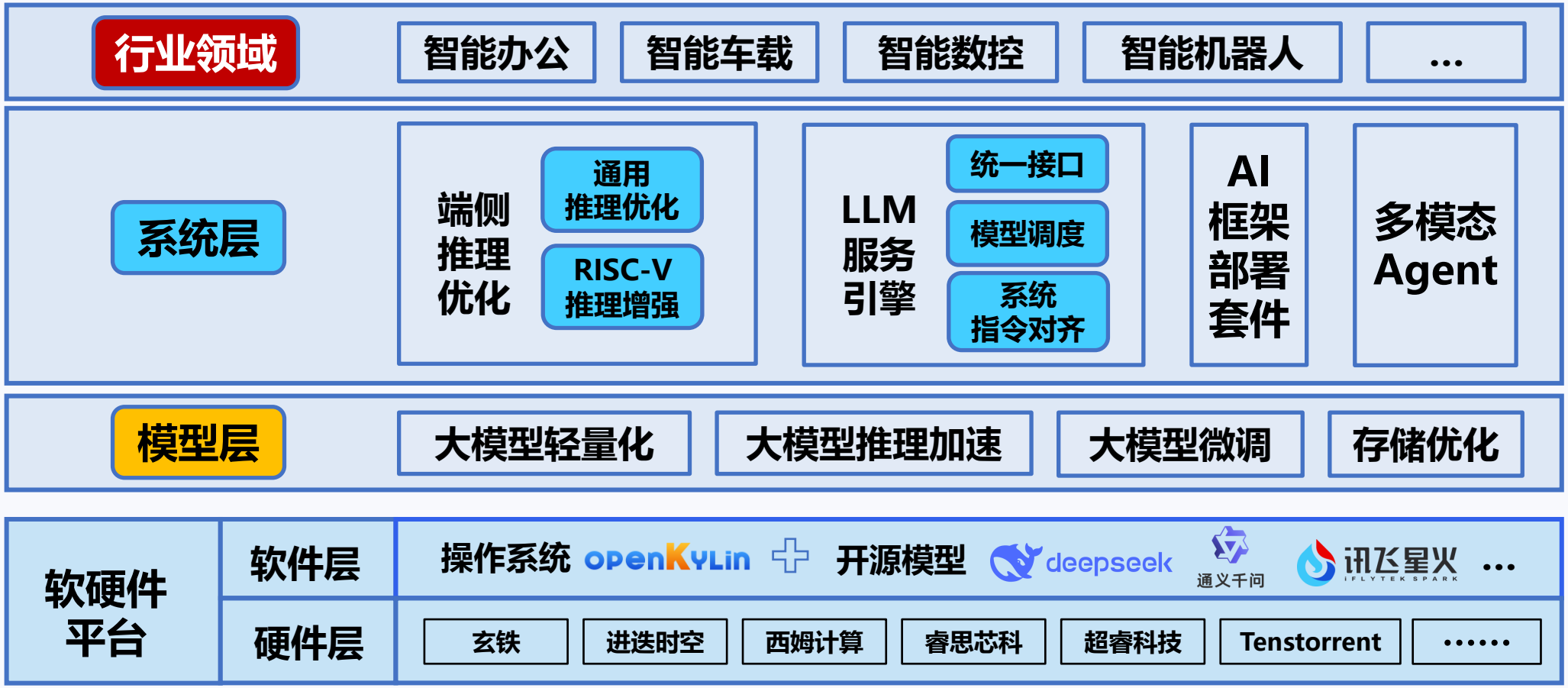
AI操作系统的设计目标

开放原子开源基金会 OPENATOM FOUNDATION | openKYLin

应用交互层	OS with AI	操作系统面向用户交互的基础功能和应用全面融入AI特性，以AI赋能提升和突破OS能力
系统架构层	OS + AI	针对交互层需求构建AI子系统框架，统一封装多家厂商的云端各类AI接口，实现本地推理和向量数据库等AI技术
硬件适配层	OS for AI	构建AI HAL，兼容各RISC-V CPU和加速卡等AI算力芯片，构建RISC-V AI硬件生态体系。

RISC-V架构AI OS设计

开放原子开源基金会 OPENATOM FOUNDATION | openKylin



以RISC-V AI开源软件栈为**核心**，高效管理和利用RISC-V智算硬件，支持**多场景**AI应用。

一、云端融合的RISC-V AI子系统

开放原子开源基金会 OPENATOM FOUNDATION | openKYLin

- 硬件资源统一管理
- AI模型协同调度
- AI应用API和SDK
- AI模型的云端融合
- 支持各类AI应用的开发

一、云端融合的RISC-V AI子系统

开放原子开源基金会 OPENATOM FOUNDATION | openKYLin

完成AI SDK V1.0版本提供40+个API

API接口：提供对外的接口，供开发者调用SDK的功能，是SDK与开发者交互的核心

通 信 层：提供与AI-Runtime的会话管理和协议通信

- 自然语言处理
  - 本地大语言模型
  - 主流云端大语言模型
  - 主流云端文生图模型
- 视觉处理
  - 本地OCR模型
  - 本地图片美化
- 语音处理
  - 本地语音识别
  - 本地语言说话人识别
  - 本地语音合成
  - 主流云端语音模型
- 模型配置
  - 本地模型配置
  - 云端模型配置

一、云端融合的RISC-V AI子系统

- **AI SDK接口的具体实现：**针对每种能力的AI能力都有对应的AI服务进行处理
- **集中管理AI任务：**通过AI引擎对接云、端模型，实现AI能力
- **AI内置应用接口：**服务配置及内置服务
- **三个核心组件：**向量数据引擎、意图识别模块和AI Runtime Core

二、AI计算性能优化

基于RISC-V的核心算子优化

- 针对矩阵乘法类、基础运算类和激活函数类等AI算子进行优化，性能提升1.6~7.5倍
- 使用向量寄存器和向量指令，将循环和数据操作转换为向量操作
- 处理多个计算序列，减少访存次数
- 结合指令融合、掩码操作、量化、算子融合等技术，实现高效的并行计算，加速卷积、全连接等AI算子运行
- 基于XNNPACK构建高性能AI算子库

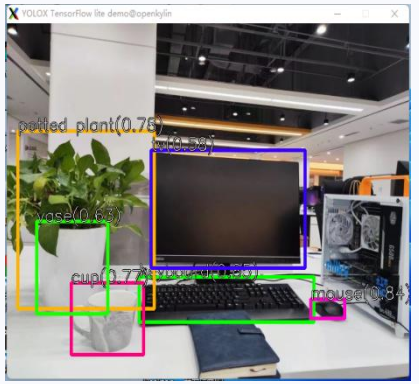
二、AI计算性能优化

开放原子开源基金会 OPENATOM FOUNDATION | openKYLin

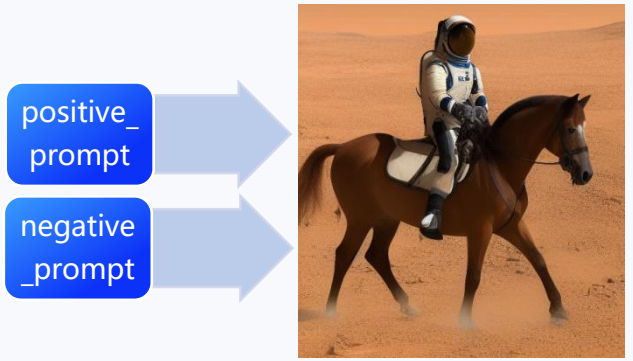
RISC-V端侧推理增强

- 目标检测耗时：性能平均提升60%
- 图像分割：性能平均提升50%
- 通用语言模型：性能平均提升16%
- 文生图：性能平均提升10%

实现功能	推理框架	推理模型	性能提升均值
目标检测	TensorFlow Lite	YOLOX	60%（FPS）
图像分割	TensorFlow Lite	YOLOv8	50%（FPS）
通用语言模型	InferLLM	ChatGLM-6B	16%（推理速度）
文生图	OnnxStream	Stable Diffusion	10%（推理速度）



目标检测效果图



文生图效果图

三、软件生态支持

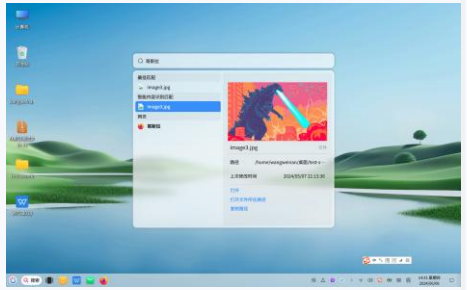
开放原子开源基金会 OPENATOM FOUNDATION | openKYLin

RISC-V推理框架支持

- 支持高效运行PyTorch、TensorFlow Lite、TensorFlow、ONNX Runtime等AI推理框架
- 高效运行大语言模型框架、适用于移动边缘设备框架、通用跨平台框架
- 完成RISC-V多硬件平台的AI软件栈适配，兼容各类RISC-V芯片和硬件实现

三、软件生态支持

开放原子开源基金会 OPENATOM FOUNDATION | openKylin



智能模糊搜索

支持语义搜图

支持语义搜文

支持索引动态更新

01

02

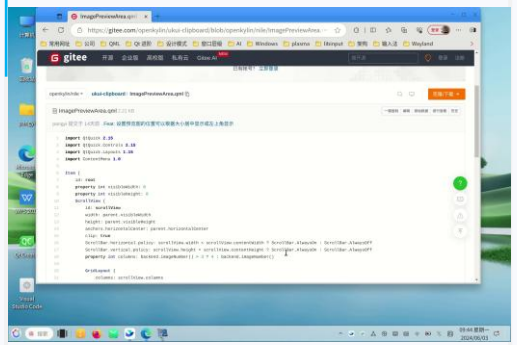
智能剪切板

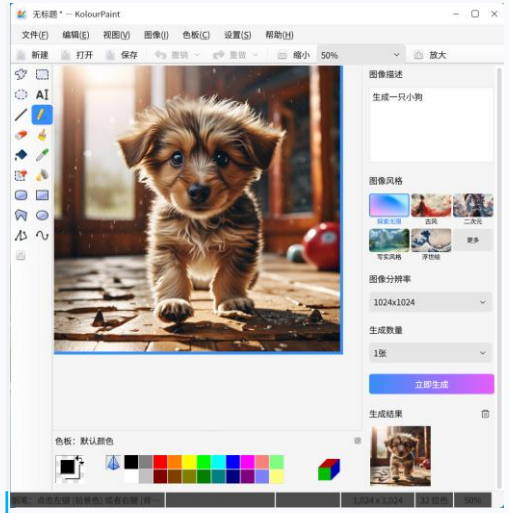
Win+V唤起

富文本转图片

纯文本转图片

图片转文本





AI画图

支持多种画图风格

支持实时图片绘画和编辑

支持不同模型图像

生成尺寸自适应布局

03

04

AI看图

智能图片处理

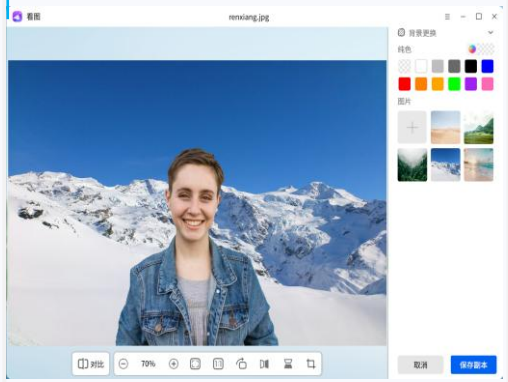
支持人像抠图功能

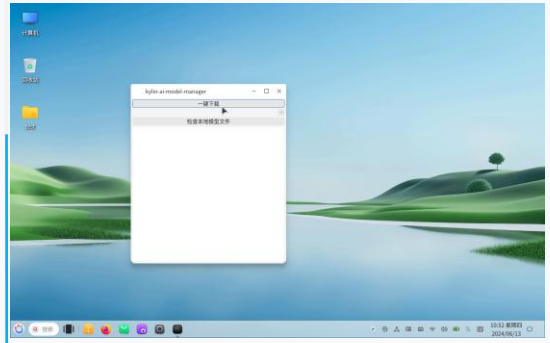
支持人像纯色背景

自定义背景替换功能

支持抠图证件照

多样式尺寸裁剪功能





05

本地模型推理

本地模型与云端模型的选择

开发模型管理工具

从模型仓库下载更新模型

本地大语言模型推理

本地语音识别处理

本地图像处理

本地文本向量化模型

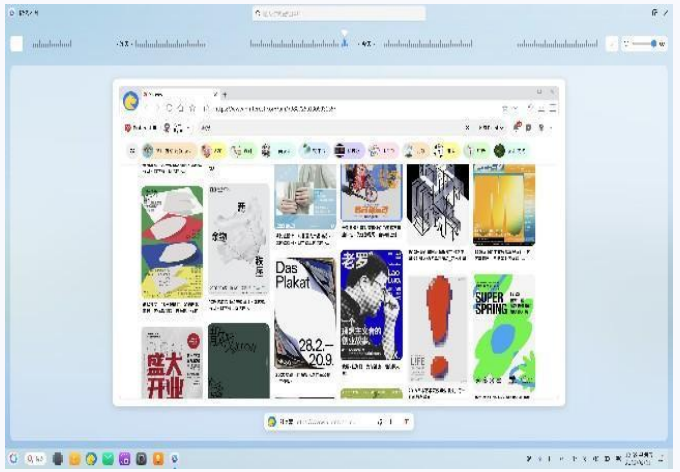
AI系统组件和应用示例

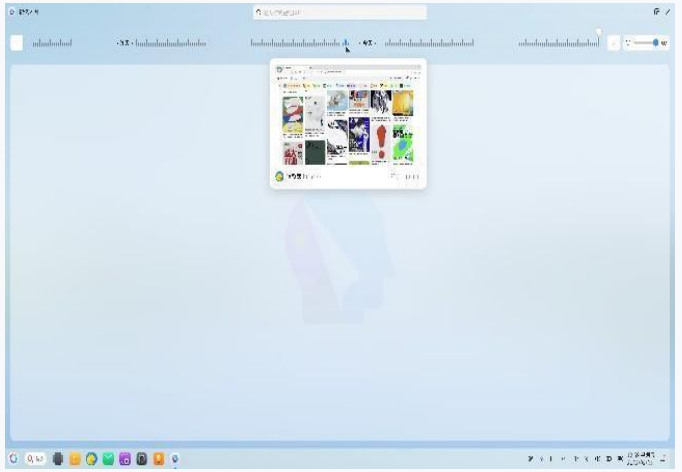
开放原子开源基金会 OPENATOM FOUNDATION | openKylin

数据记录提效-记忆地图

1 功能简介

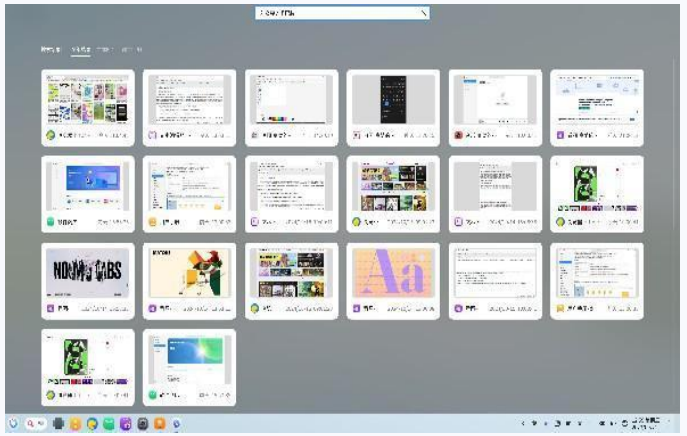
- 记录所有浏览内容
- 记录应用活动记录
- 支持基于时间轴记忆浏览
- 支持语义内容理解的以文搜文和以文搜图

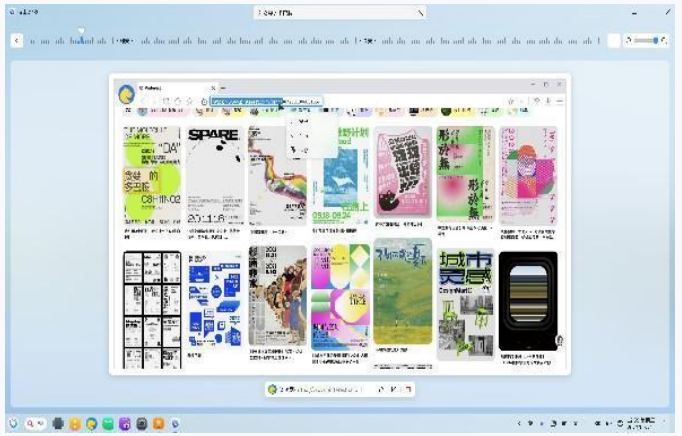




2 实现原理

- 通过显示服务高效获取应用显示内容
- 通过OCR识别出文字内容和区域
- 通过向量化模型提取文本和图像特征
- 通过向量数据库存储数据和检索





AI系统组件和应用示例

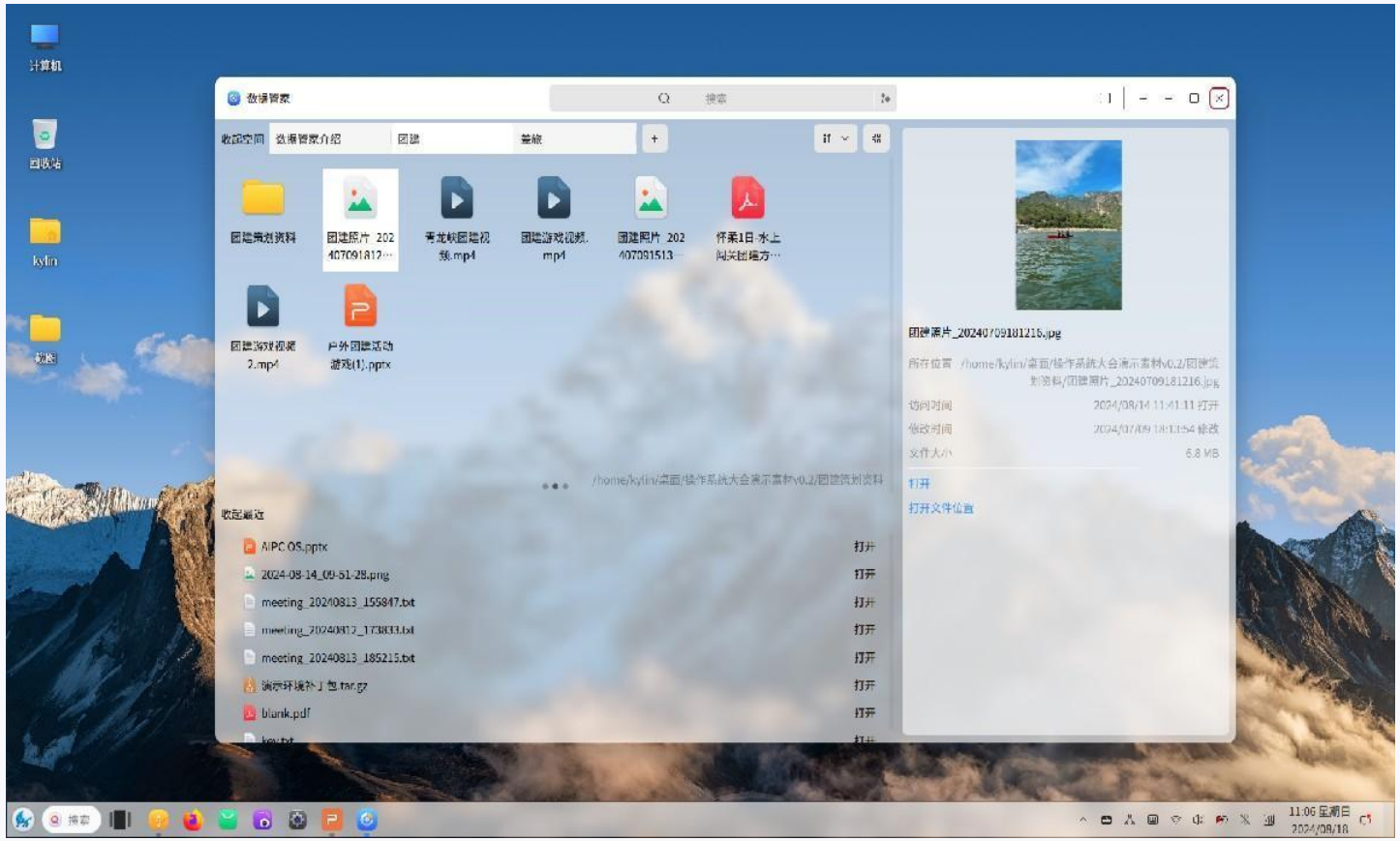
数据管理提效-数据管家

1 功能简介

- 基于文件内容的逻辑空间管理
- 支持对文件内容摘要提取和预览
- 支持图片和文档标签提取
- 支持文件聚类

2 实现原理

- 基于ebpf实现高效的文件管理
- 基于向量化模型实现文件特征提取
- 基于多模态模型实现图片摘要提取
- 基于大语言模型实现文件摘要提取



AI系统组件和应用示例

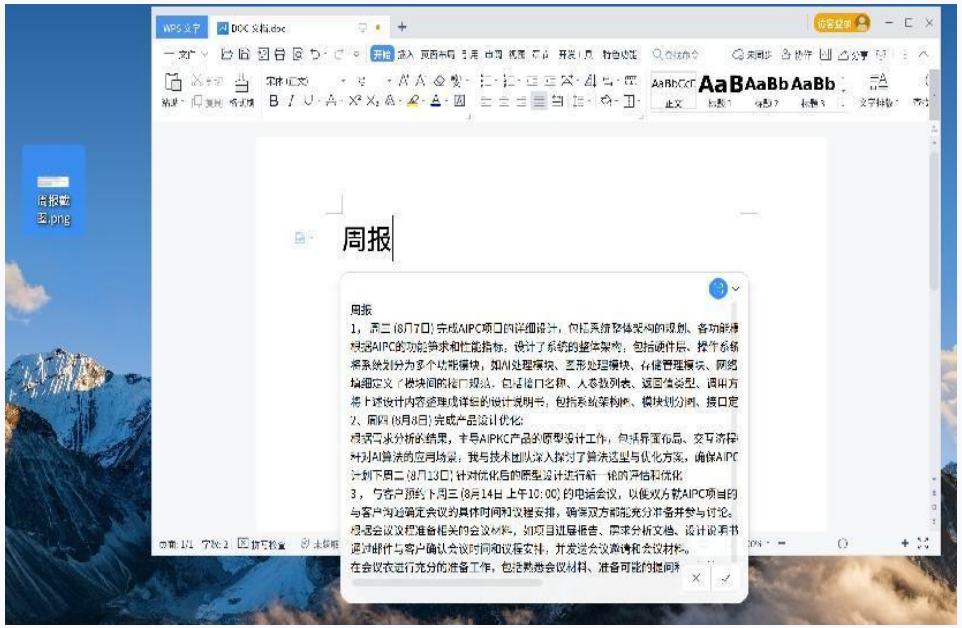
开放原子开源基金会 OPENATOM FOUNDATION | openKylin

数据查找提效-智能模糊搜索

功能简介

- 支持语义搜图
- 支持语义搜文
- 支持索引动态更新

数据流转提效-智能剪切板



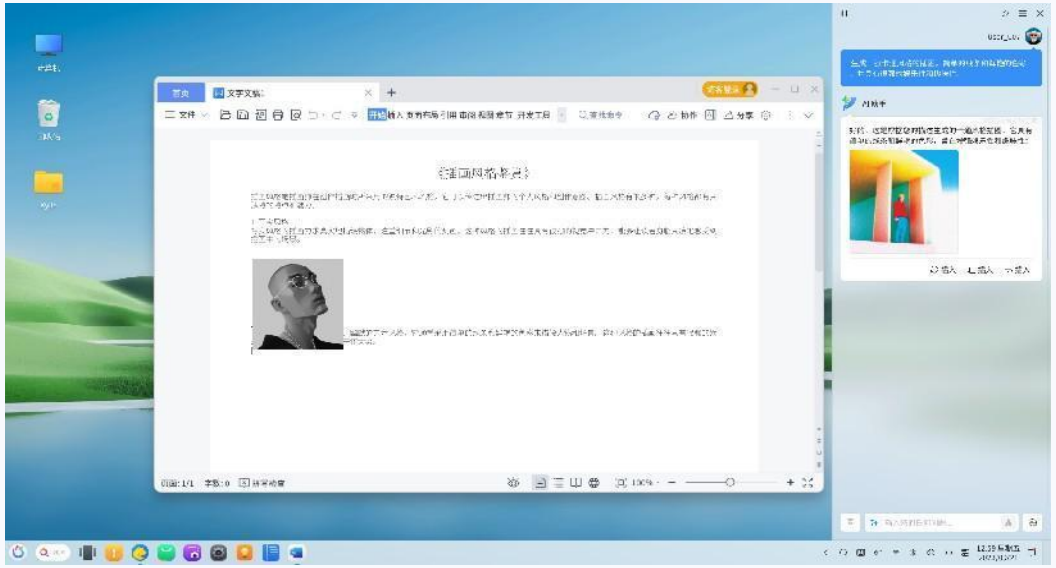
功能简介

- 支持富文本转图片
- 支持文本转图片
- 支持图片转文本

AI系统组件和应用示例

开放原子开源基金会 OPENATOM FOUNDATION | openKylin

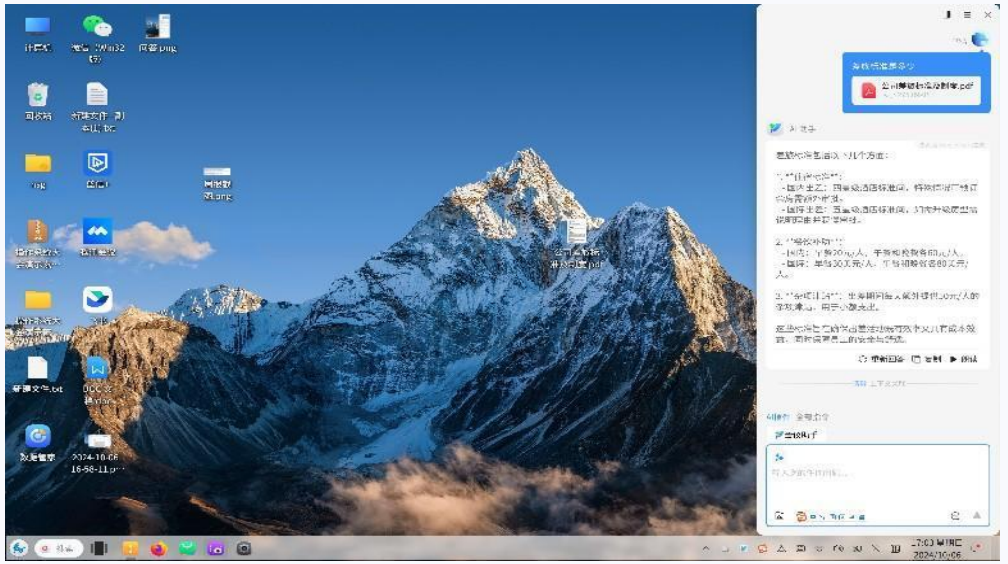
系统控制



功能简介

- 支持200+条系统控制指令，如打开应用、音量调节等
- 支持10+条快捷指令，如翻译、总结、整理周报、整理代办等
- 支持文生图

知识问答



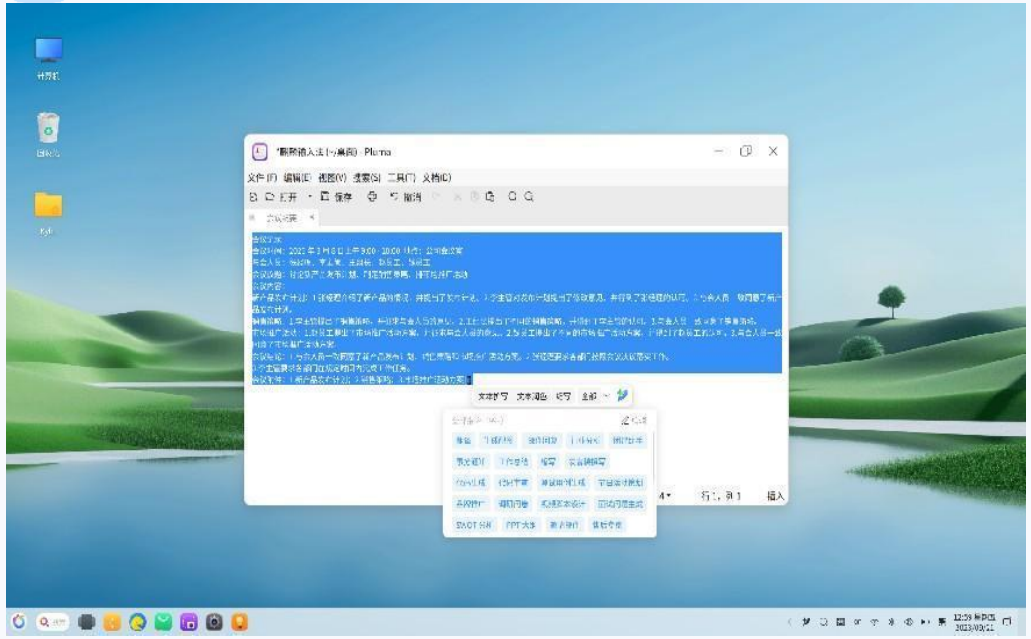
功能简介

- 支持本地知识库问答
- 支持本地文档的问答
- 支持联网实时知识问答

AI系统组件和应用示例

开放原子开源基金会 OPENATOM FOUNDATION | openKylin

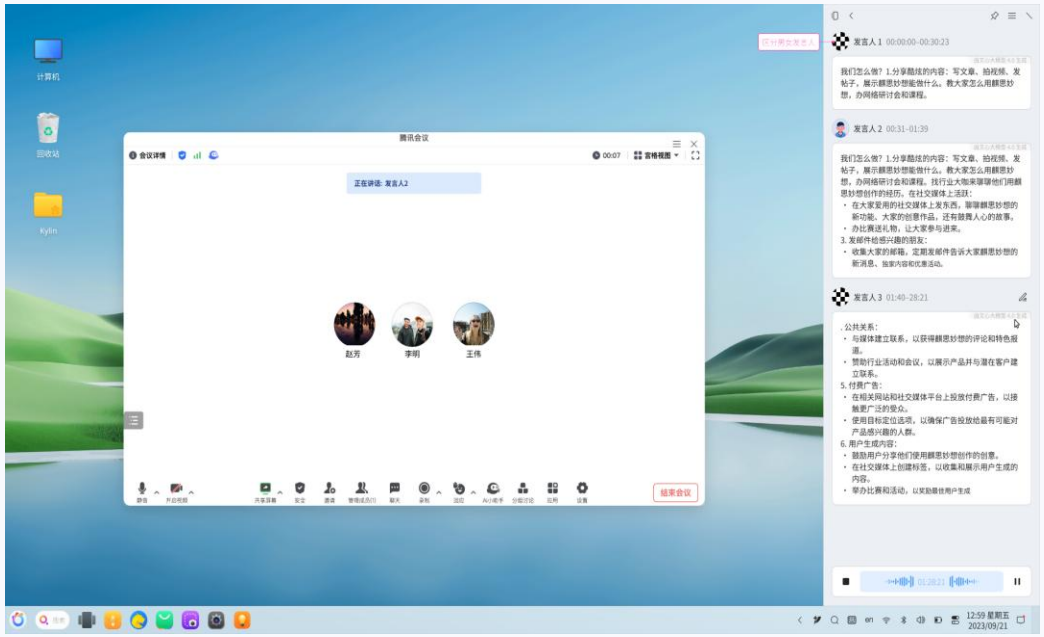
应用联动



功能简介

- 支持全局的划词文本提取
- 支持划词指令与AI助手联动

助手插件



功能简介

- 支持会议助手
- 支持扩展各种助手插件

端侧大模型+openKylin智能桌面

开放原子开源基金会 OPENATOM FOUNDATION | openKylin

RISC-V AI硬件平台

- 完成RISC-V AI硬件平台软件栈适配
- 支持单卡或多卡集群等RISC-V异构AI平台

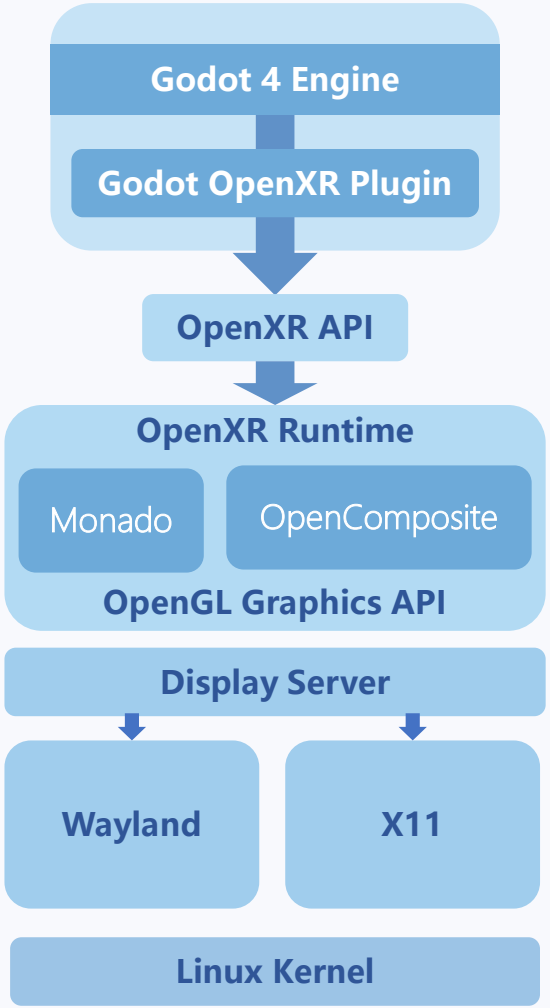
openKylin智能桌面应用

- 智能模糊搜索、智能剪切板、AI画图看图、AI助手等

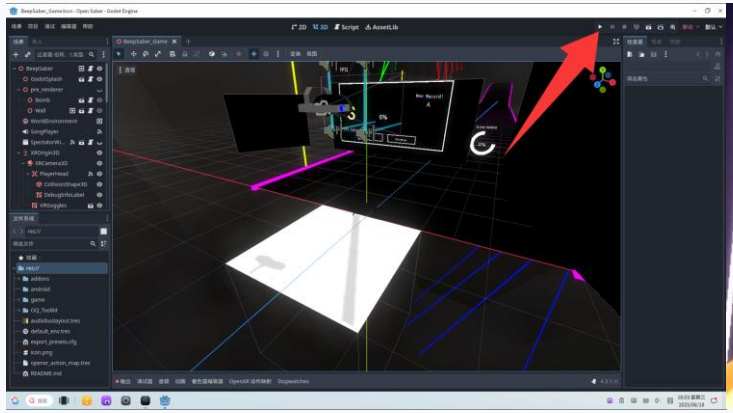
多种类模型支持

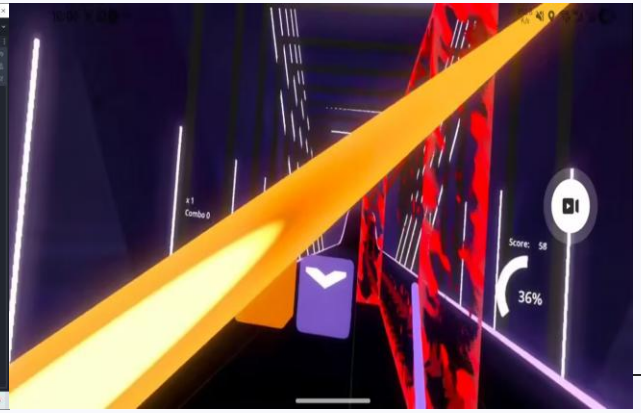
- 实现各类主流AI模型（大语言模型、图像识别、文生图、语音识别）的本地化部署
- 将openKylin AI子系统接入本地RISC-V AI硬件平台，为RISC-V端侧智能应用开发提供支持

openKylin RISC-V XR支持



- 将OpenXR 软件栈移植至 RISC-V 平台，并实现了与 XR 头显的协同运行，构建了OpenXR 远程执行环境
- 支持运行空间类 OpenXR 应用与游戏，特别针对 Godot 引擎进行了适配与优化
- 对OpenXRloader、OpenXRruntime、Godot OpenXRplugin等软件的兼容性进行验证，支撑多款开源 XR 应用运行





总结与展望

开放原子开源基金会 OPENATOM FOUNDATION | openKylin

技术合作交流

开源社区、openKylin
RISC-V AI SIG组等平台

基础软件环境支持

玄铁、Tenstorrent、进迭时空等厂商AI硬件平台支持

操作系统自身优化

高性能优化、RISC-V Conda环境开发、RISC-V Python AI库等

deepseek

通义千问

讯飞星火 IFLYTEK SPARK

RISC-V®

基于RISC-V软硬件开源体系和协同优化，共筑RISC-V AI 生态

2025

开放原子开源基金会 OPENATOM FOUNDATION | openKylin

THANKS

谢谢！

www.openKylin.top

contact@openKylin.top