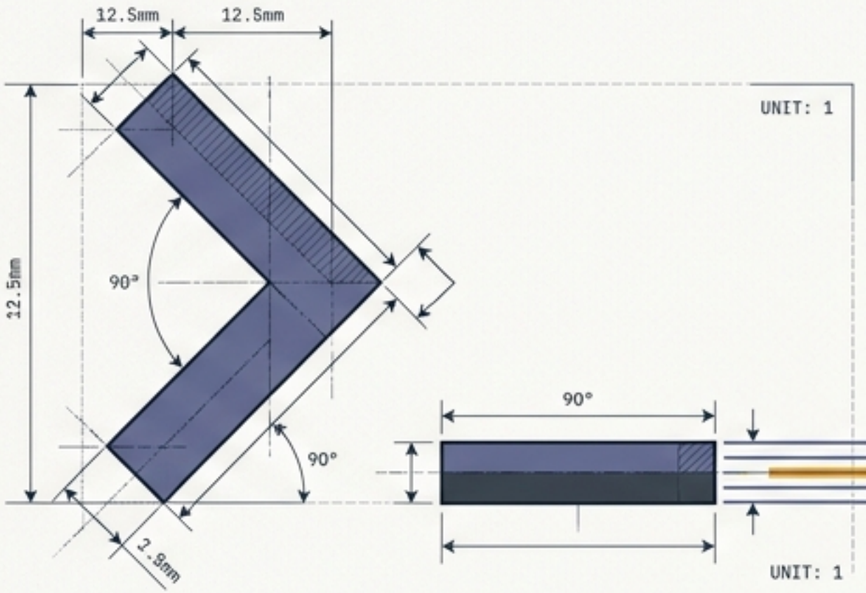
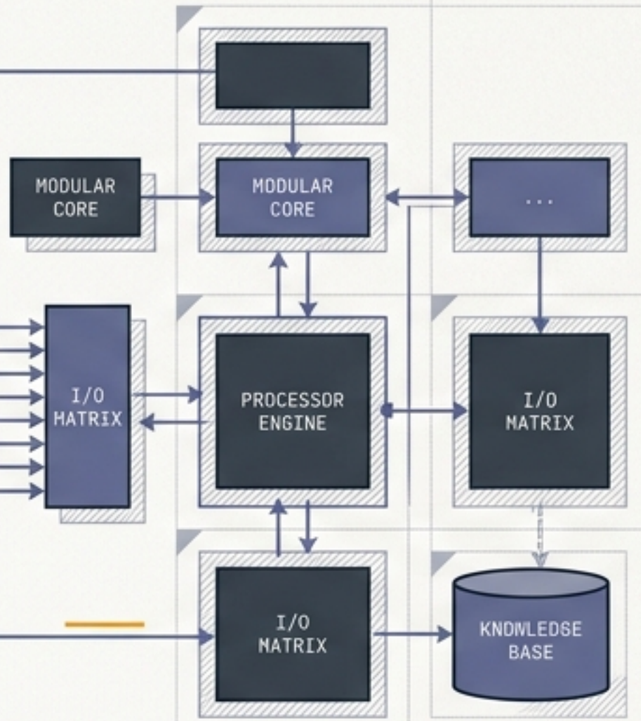
架构深度解析与能力跃迁指南

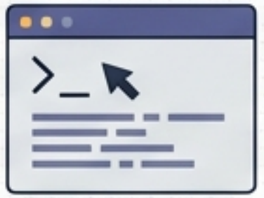




Hermes Agent：下一代终端原生 AI 智能体系统

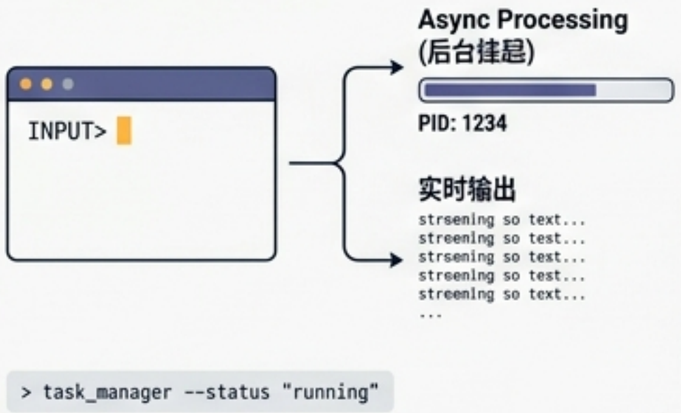
为极客、开发者与自动化团队打造的工程级中枢

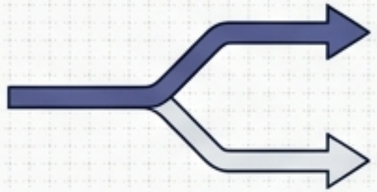
核心 DNA：不仅是对话，更是全能基础设施



交互层：原生 TUI / CLI

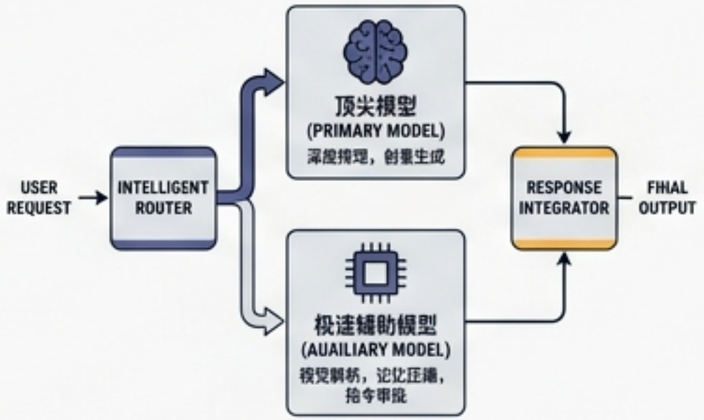
摒弃臃肿网页 UI，回归纯粹终端。支持无阻塞异步输入、多行代码粘贴、后台任务挂起与透明状态流。





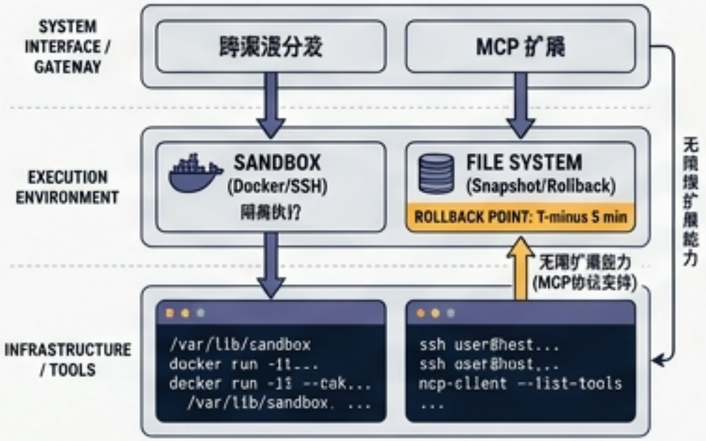
认知层：双轨模型路由

主辅模型彻底解耦。将深度推理交给顶尖模型，将视觉解析、记忆压缩、指令审批等后台任务下发给极速辅助模型。

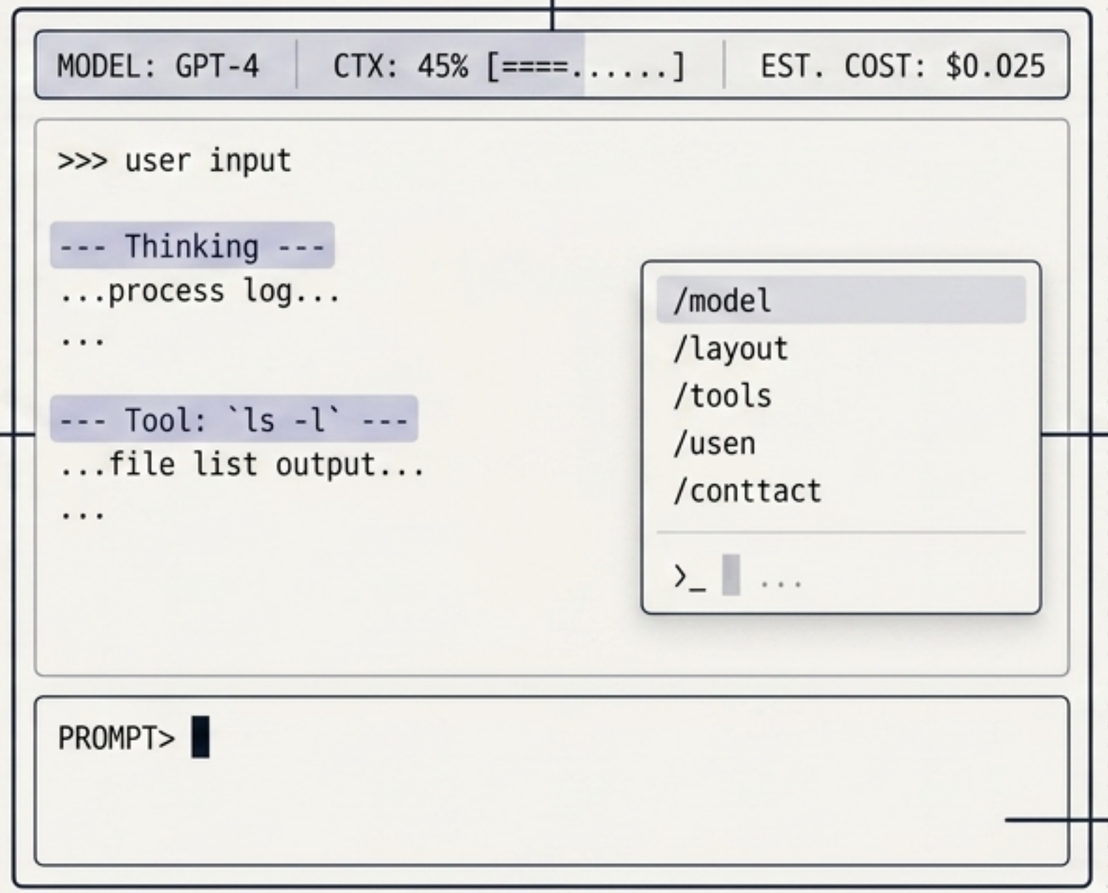


执行层：深度系统集成

内建 Docker/SSH 沙盒执行环境、跨渠道分发网关、文件系统快照回滚，以及基于 MCP 协议的无限工具扩展。



TUI 界面解剖：为终端居民打造的 HUD



全局 Token 雷达

实时展示模型状态、上下文压力条（带容量预警 `[==!]`）、Token 消耗与精确预估成本。

内联思维链与工具流

模型推理（Thinking）与工具调用（如终端命令 `cmd: ls`）分层渲染，过程绝对透明可控。

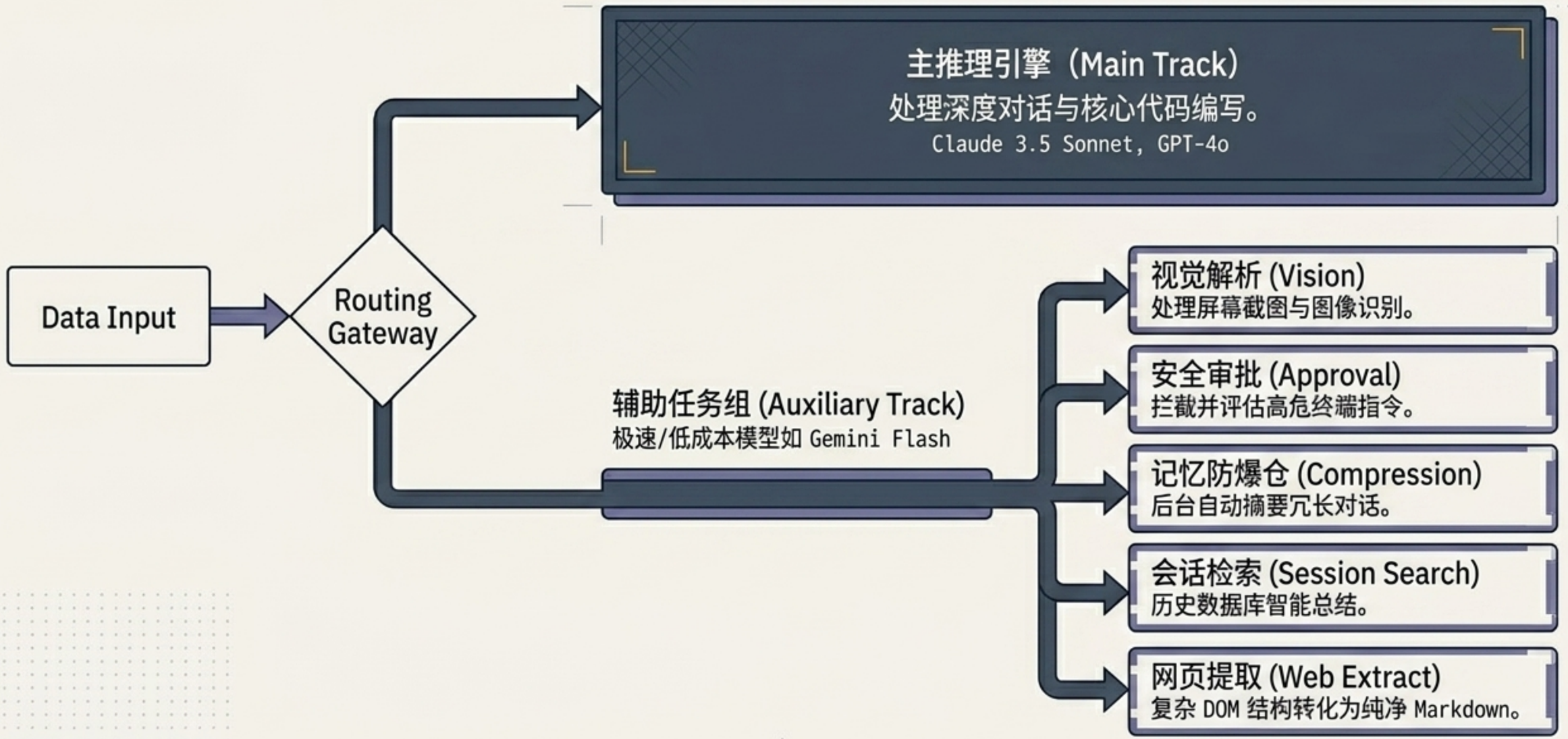
Slash Commands 矩阵

按键 `'/'` 即唤出，支持动态技能挂载、交互式模型切换与界面高定排版。

无阻塞多行编辑器

完美支持大段代码粘贴防抖，运行期间可随时安全中断或进行任务重定向（Steer）。

认知引擎：主辅解耦的双轨路由策略



行动引擎：终端后端选型矩阵

后端类型 (Backend)	隔离级别 (Isolation)	文件持久性 (Persistence)	最佳适用场景 (Best Scenario)
local	无隔离（本机直连）	宿主实时同步	本机日常开发与文件处理
docker	容器级隔离（剥夺Root）	跨会话持久挂载	代码沙盒、安全评测与 CI/CD 自动化
ssh	跨网络边界物理隔离	远程持久化 Shell	操控远程高算力服务器集群
modal / vercel	云端 Serverless 微机	快照级持久化恢复	云端临时高并发任务与销毁后同步
daytona	托管式云开发容器	工作区休眠与唤醒	云端标准化开发环境管理

安全护栏：将 AI 破坏力关进笼子

第一道：PII 数据脱敏

在网关层自动哈希化手机号、用户 ID 等隐私数据，防止敏感泄露。

第二道：Tirith 启发式扫描

执行前通过独立扫描器识别恶意终端特征，例如隐蔽后门或高危指令。

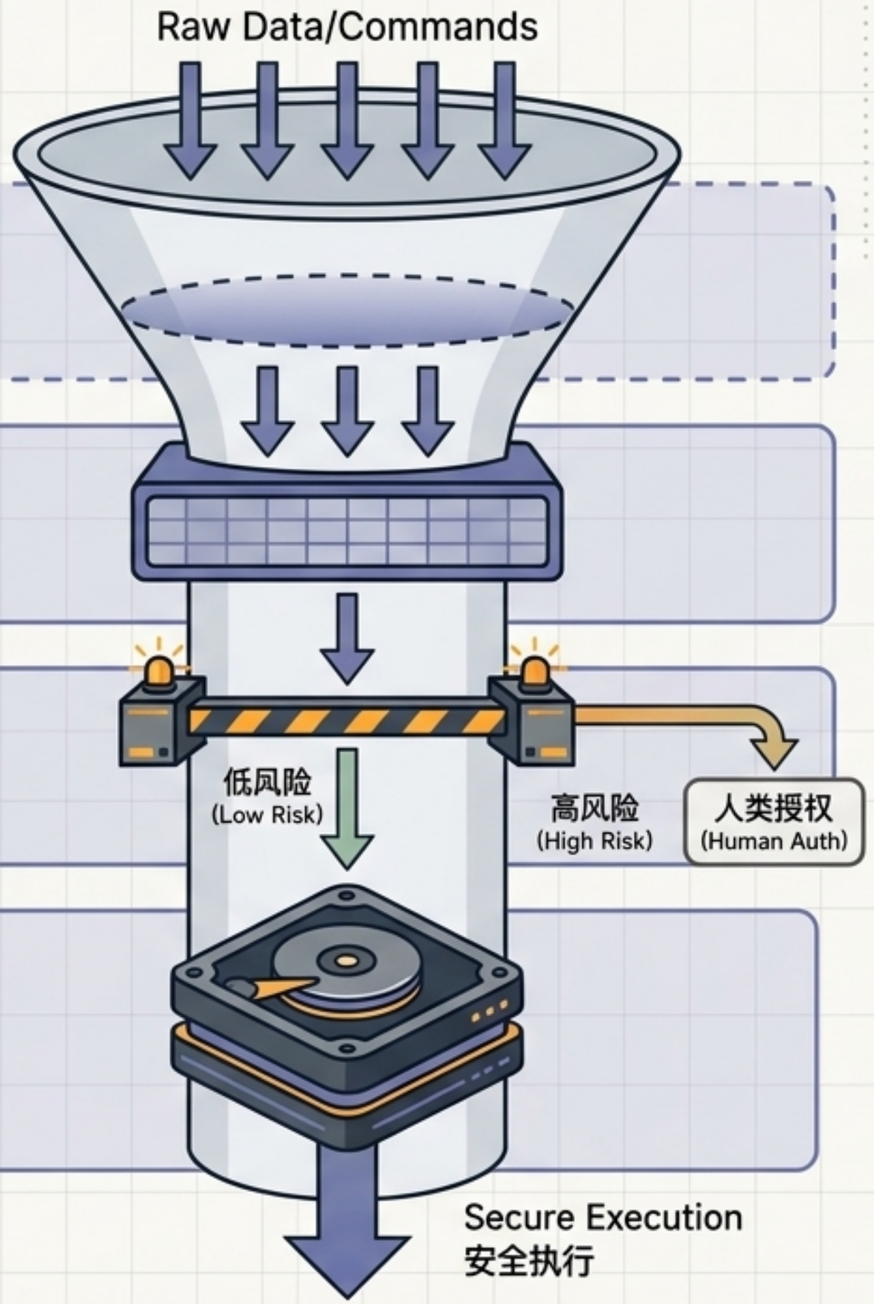
第三道：智能审批（Smart Approvals）

调用独立辅助模型判定风险等级。低风险静默放行，高风险触发拦截弹窗请求人类授权。

第四道：自动快照与回滚

执行文件修改前自动创建物理快照，支持一句命令瞬间无损恢复。

`/rollback`



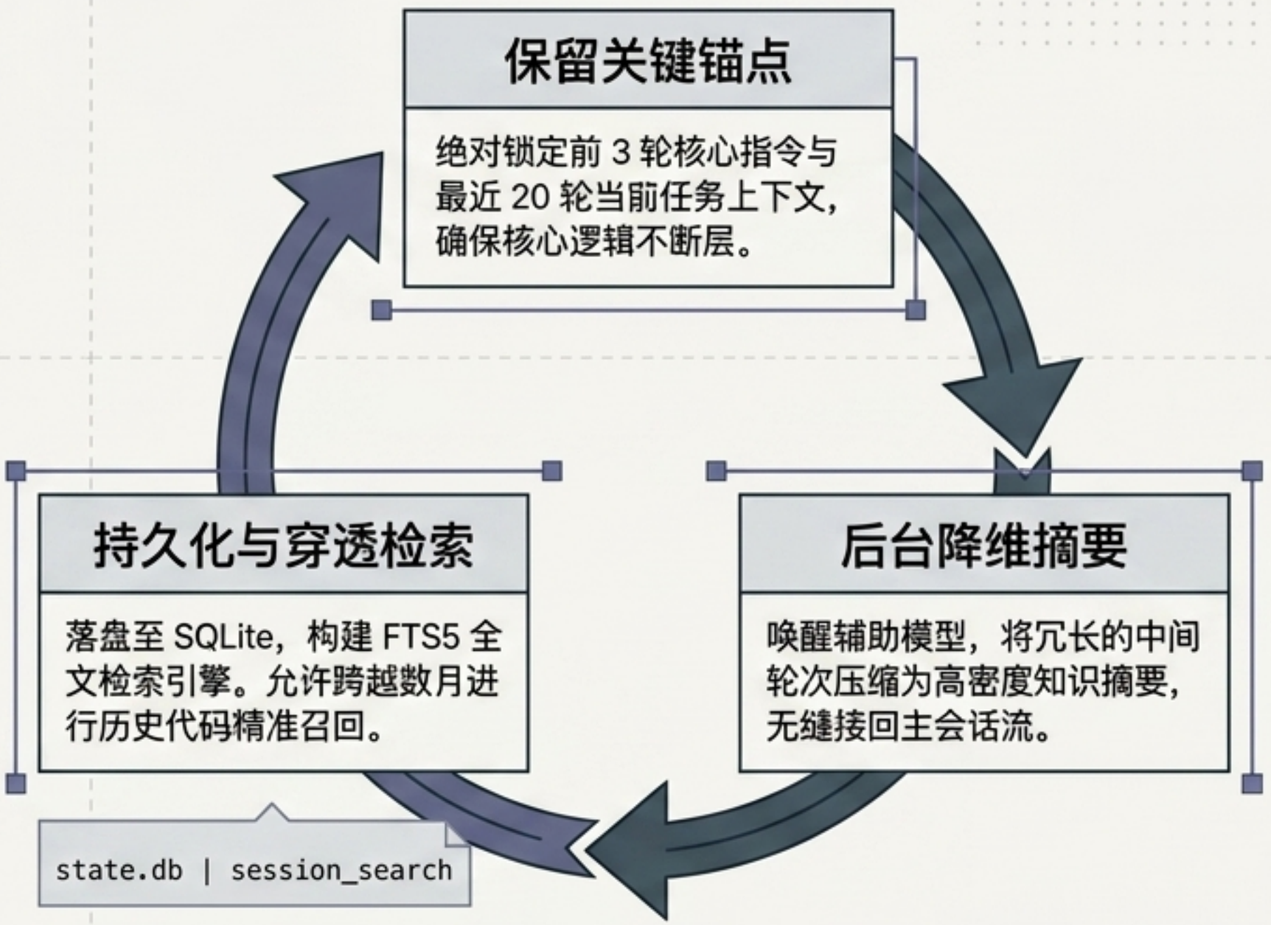
记忆生态：永不爆仓的上下文生命周期

Context Pressure Indicator

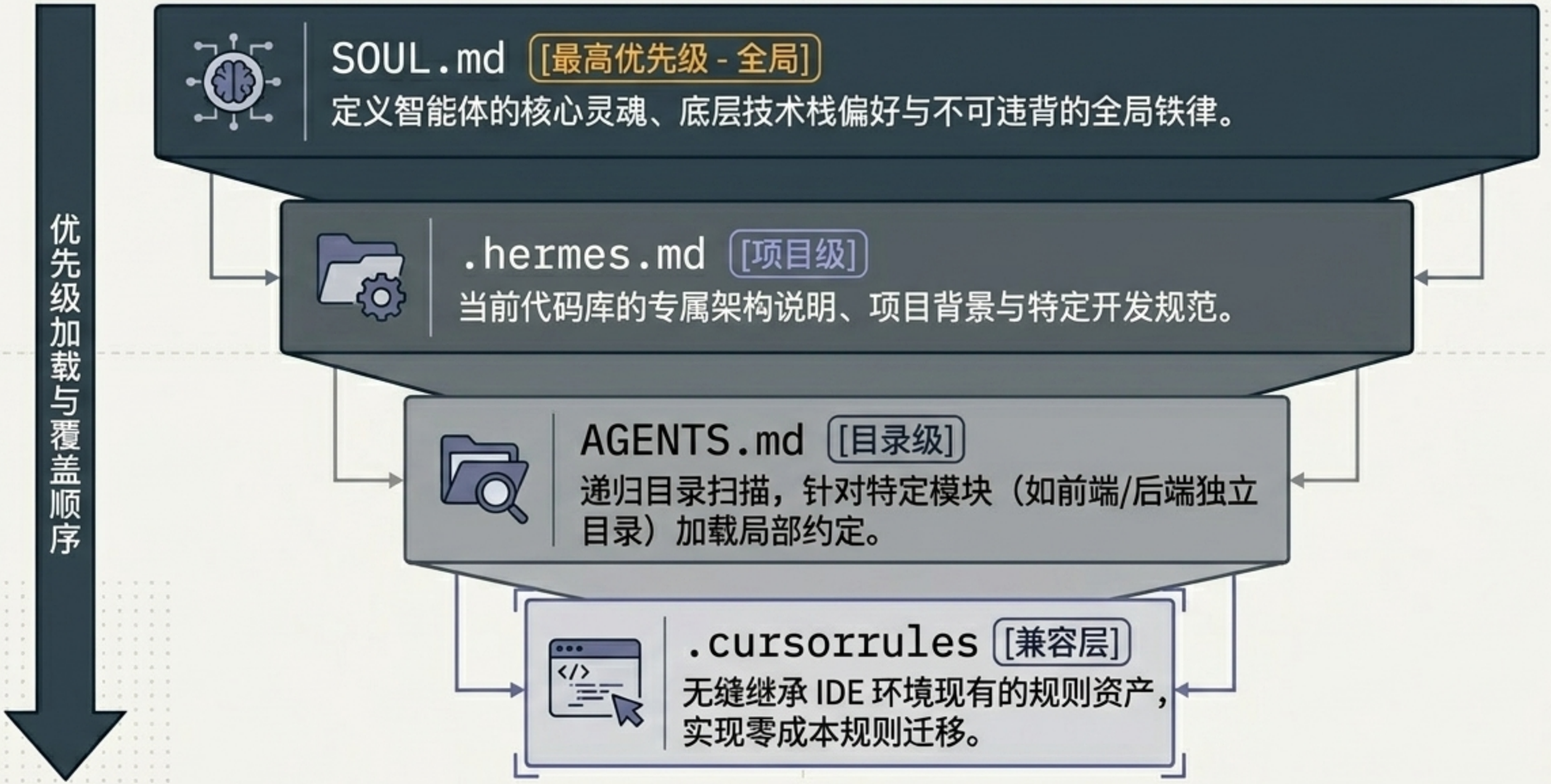
压力感知与自动折叠

系统毫秒级监控上下文容量。当对话代币触及警戒阈值，触发静默压缩机制。

Memory Lifecycle Flow

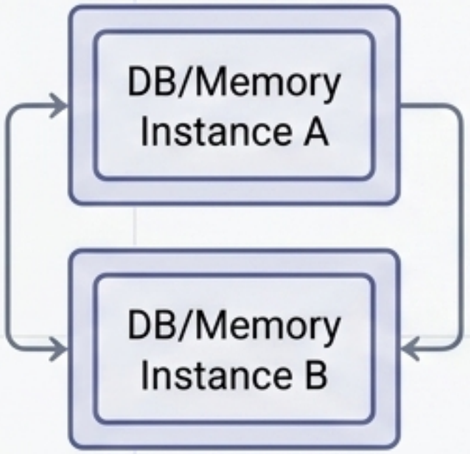


知识引擎：多级上下文注入瀑布



算力分身：Profiles 与 Worktrees 隔离

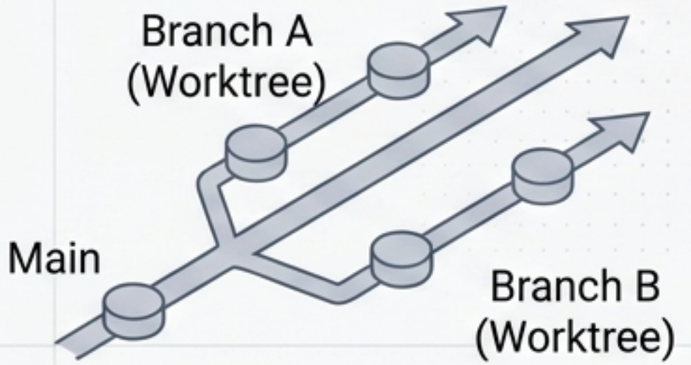
Profiles - 认知隔离



创建彼此独立的大脑实例。例如 “coder” 和 “bot” 拥有完全独立的环境变量、人格设定与记忆历史。

```
hermes profile create coder
```

Git Worktrees - 物理隔离



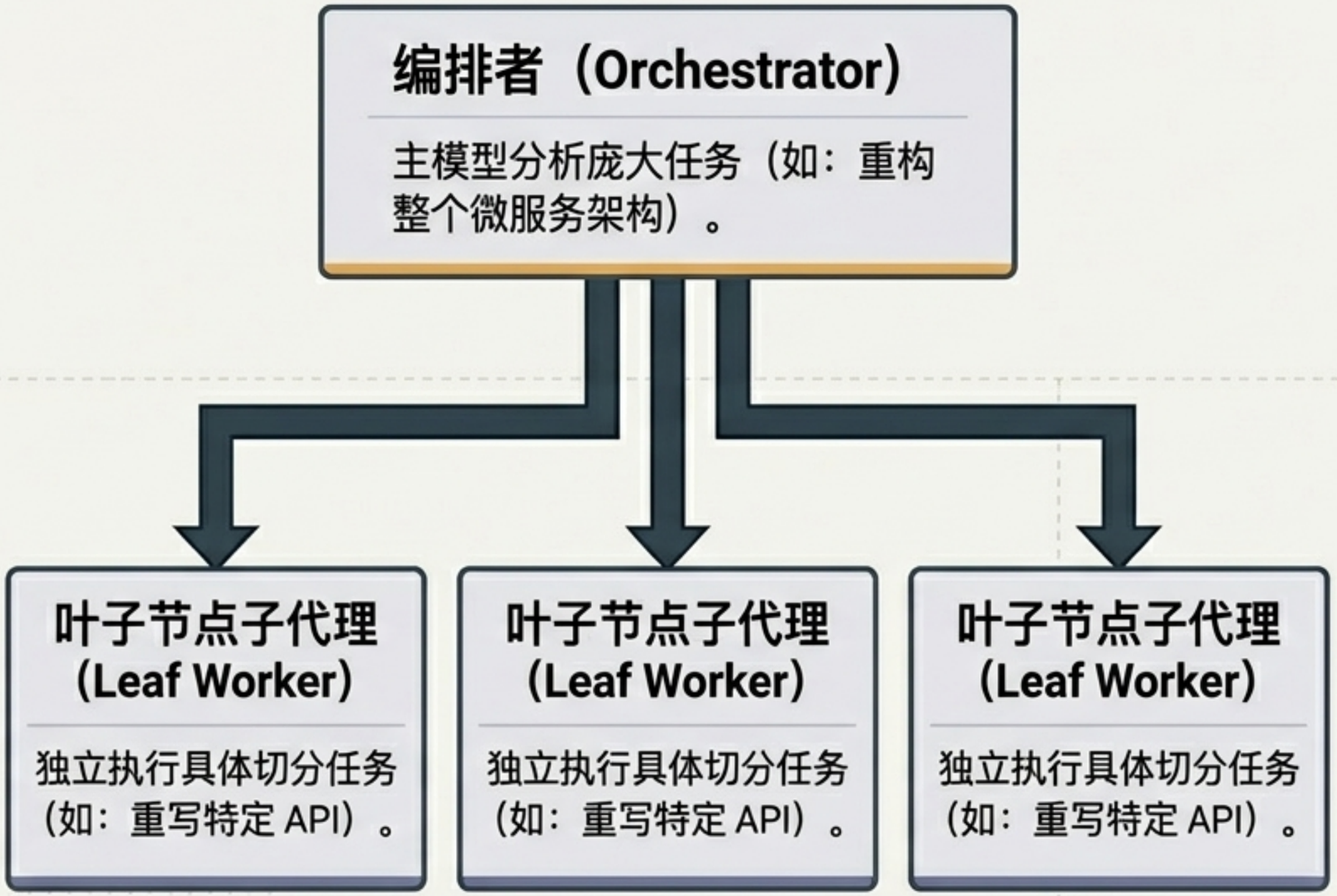
隐形的工作区魔法。在同一代码库并发拉起多个影子分支，Agents 平行修改代码，互不污染，最终提交流水线 PR。

```
hermes -w
```

极速并发分流

使用 /background 指令，在维持主工作流交互的同时，将耗时分析或高资源消耗任务抛入后台隔离沙盒运行。

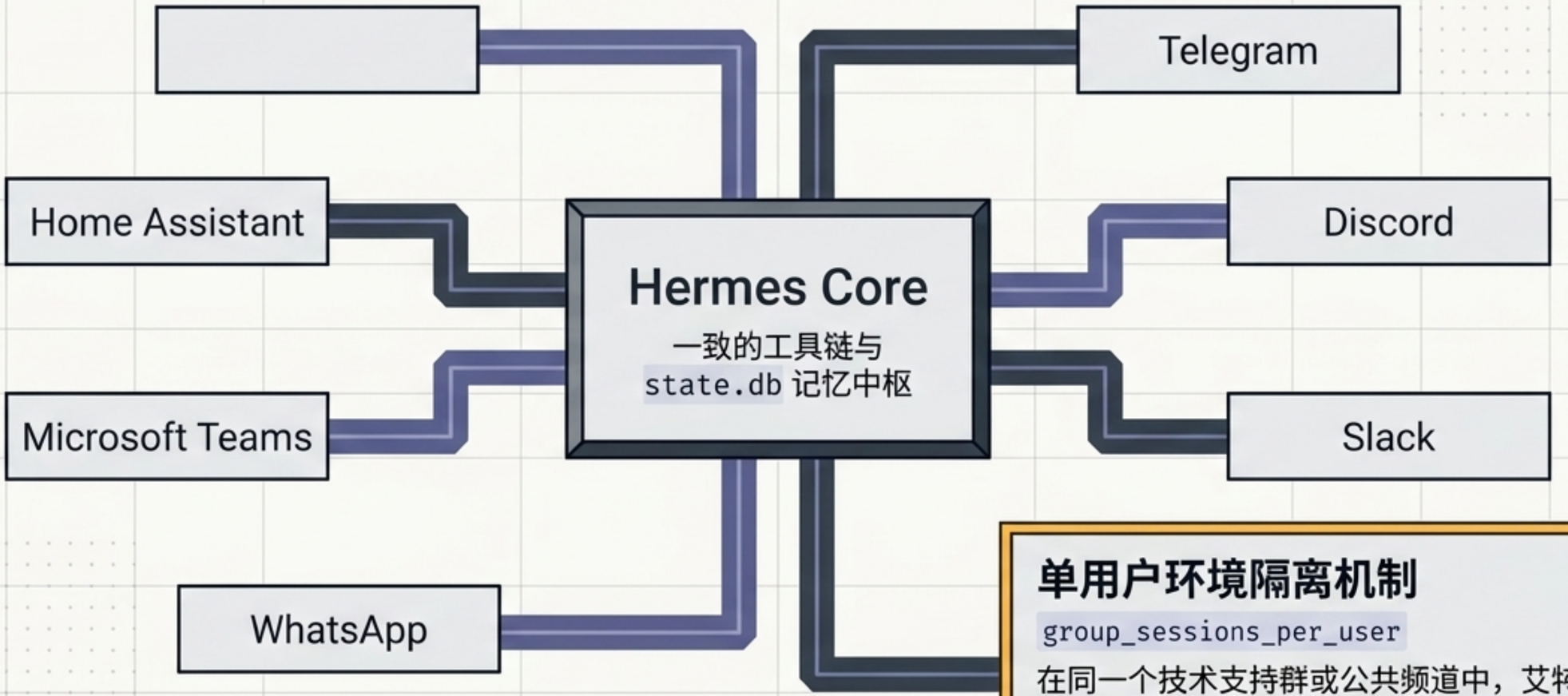
代理集群：深度子任务委派（The Swarm）



集群扩展参数

- `max_concurrent_children`
  并发宽度控制。同时拉起多个独立的子沙盒并行工作，算力横向扩展。
- `max_spawn_depth`
  衍生深度锁。支持最高三级深度树状裂变，严控爆炸半径。
- **独立运行环境**
  支持为每个子节点挂载专属隔离 Docker 镜像，满足异构工具链依赖。

全渠道分发：企业级消息网关（Gateway Hub）



单用户环境隔离机制

`group_sessions_per_user`

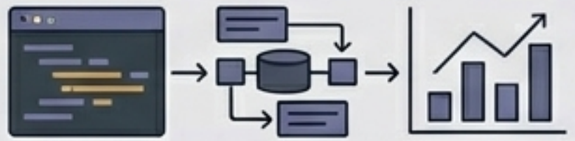
在同一个技术支持群或公共频道中，艾特机器人的每一位团队成员都拥有完全独立的上下文会话与成本核算账户，从根本上杜绝多用户对话串线与内存污染。

能力扩展：原生 Skills 与全局 MCP 协议

内置超级技能（Native Skills）

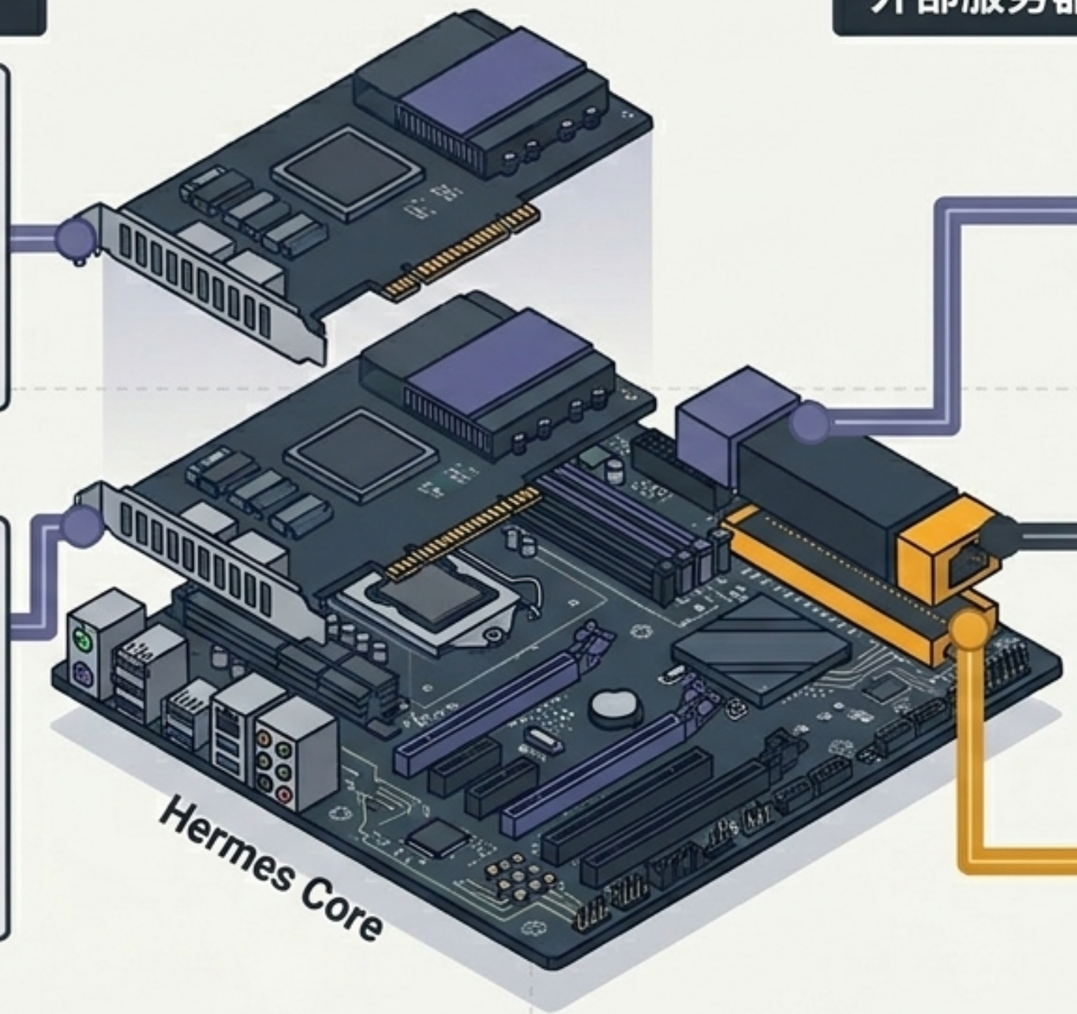
Python 自动化运行与数据分析

`RUN_PYTHON_AUTOMATION_ANALYTICS`



Firecrawl/SearXNG 深度网页检索

`FIRECRAWL_SEARXNG_SEARCH`



外部服务器无缝接入（MCP Protocols）

无缝桥接

接入 Github，本地文件系统，PostgreSQL 数据库等外部生态。

双向通信支持

支持本地命令行流（Stdio）与远程 HTTP/OAuth 2.1 鉴权连接。

反向触发

支持服务器发起的 LLM 采样请求（Server-Initiated Requests）。

极速部署：跨平台无缝接轨

一键极速安装

完美支持 Linux / macOS / WSL2。系统将自动处理 uv, Node.js, ffmpeg 等所有底层依赖环节。

```
curl -fsSL https://raw…../install.sh | bash
```

移动端全血运行

官方深度适配 Android Termux。构建完整的原生虚拟环境，将工业级大模型自动化引擎直接装进口袋。

```
pkg update && curl -fsSL ... | bash
```

声明式系统级服务

专为硬核运维团队设计。提供原生 Nix Flake 和 Module 支持。一键拉起附带 Sops-Nix 密钥加密、声明式 MCP 配置及持久化容器挂载的服务体系。

```
nixos-rebuild switch --flake .#server
```

高阶魔法：终端控制备忘录

核心系统指令（Core Slash Commands）

`"/model"`：
交互式唤出菜单，极速切换模型与 API 提供商。

`"/voice on"`：
唤醒本地 Whisper 语音听写与多厂商 TTS 引擎联动。

`"/reasoning"`：
动态调整模型深度思考算力分配（高/中/低）。

`"/title"`：
手动干预，快捷管理与命名当前记忆抽屉。

`"/tools"`：
展开模块化面板，开启或屏蔽特定执行能力。

自定义宏指令（Quick Commands）

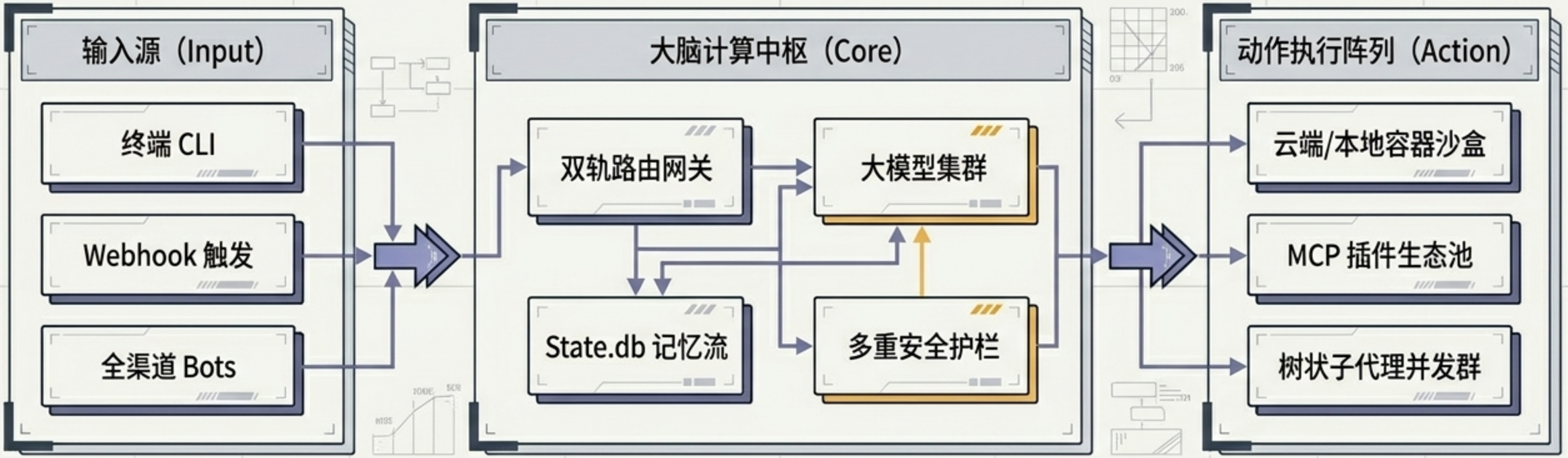
在 `config.yaml` 中配置 `type: exec`，实现零 Token 消耗瞬间调用宿主命令。将聊天框变为你的个人运维控制台。

输入 `"/gpu"`：
瞬间调用宿主 `nvidia-smi` 打印显卡状态。

输入 `"/restart"`：
一键重启挂载服务。

输入 `"/status"`：
实时监控磁盘与内存基准信息。

构筑属于你的赛博中枢



Hermes 拒绝成为生成代码片段的玩具。它是下一代自动化中枢核心：拥有持久深层记忆、无缝调度海量模型、在绝对安全的沙盒内执行真实世界的复杂指令，并随时听候全网多渠道的调遣。

```
curl -fsSL https://raw.githubusercontent.com/NousResearch/hermes-agent/main/scripts/install.sh | bash
```