ClawFlag 产品定义文档 v5.0 (完整版)

产品标语：掌控你的 AI，从信任开始。

AI Agent 治理平台与指挥中心

版本 5.0 | 机密 | 整合 38 位专家 5 轮评审及用户核心构想 | 2026年2月21日

clawflag.com ・ bitflag.com

---

0. 核心哲学 — 产品理念的演进

> 🎯 **我们交付的不是功能，是掌控感与确定性。**

自 v4.0 以来，我们对市场的认知、对用户的理解以及对技术趋势的判断都发生了深刻的演进。v5.0 的核心哲学，建立在对一个残酷现实的清醒认知之上：通用 AI Agent 的时代已经到来，但它仍是一片混乱、危险且不可信的“狂野西部”。我们的三大基石信念，在吸收了最新的市场反馈和专家共识后，得到了深化与重塑：

信念一：AI 的价值，始于可信。

一个无法被信任的工具，无论多么强大，其价值都将趋近于零。OpenClaw Agent 能够执行真实世界的操作，也就能造成真实世界的破坏。根据 SecurityScorecard STRIKE 的报告，超过 135,000 个 OpenClaw 实例暴露在公网，其中大量存在远程代码执行（RCE）漏洞[1]。用户最深层的恐惧是“失控”。因此，ClawFlag 的首要任务不是“赋能”，而是“约束”；不是“加速”，而是“刹车”。我们必须为用户提供一个在任何情况下都绝对可靠的“拉电闸”的权力。

信念二：信任的建立，源于激进的透明度。

用户对 AI Agent 的不信任，根植于其“黑箱”特性。当 Agent 显示“正在输入...”时，用户内心的疑问是：“它在干什么？要多久？它会不会又在胡说八道？”。v4.0 提出的“人类在环”理念是正确的，但仅仅“在环”是不够的，人类必须能“看透”这个环。ClawFlag 必须提供一个“驾驶舱”，将 Agent 的思考链、工具调用、记忆读写、成本消耗等内部状态，以一种人类可以直观理解的方式实时呈现出来。我们追求的不是简单的日志查看，而是“激进的透明度”（Radical Transparency）。

信念三：透明度的实现，依赖于零摩擦的连接。

如果查看“驾驶舱”需要复杂的网络配置，那么透明度就毫无意义。v4.0 设想的直连模式，将 99% 的非专业用户挡在了门外。我们必须承认，便利性是所有高级功能得以实现的基础。因此，v5.0 明确采纳了“云端盲中继”架构。这并非对 v4.0“零云数据主权”理念的背叛，而是一种哲学上的升华。

> **“盲管道”（Blind Pipe）原则：** ClawFlag 的中继服务器在物理上转发数据包，但在逻辑上对内容完全无知。通过端到端加密（E2EE），我们确保了中继服务器无法读取、存储或理解任何

> 通信内容。这比“我们承诺不看”的政策保证更强大，是“我们在技术上做不到看”的架构保证。它完美调和了连接便利性与数据主权之间的矛盾。

一言以蔽之： ClawFlag v5.0 的核心使命，是通过“零摩擦连接”实现“激进的透明度”，最终建立用户对 AI Agent 的“绝对掌控感”，从而构建“可信 AI”的基石。

---

1. 愿景与战略定位

愿景：

> **成为通用 AI Agent 时代事实上的“治理与安全”标准层。**

战略定位：从“工具”到“平台”的跃迁

十位跨领域顶级专家的共识是：ClawFlag 与 Happy Coder 等竞品的本质区别在于定位。Happy Coder 是一个极致的“赋能型工具”（Enabling Tool），而 ClawFlag 是一个雄心勃勃的“管控型平台”（Governing Platform）[2]。

对比维度	Happy Coder (工具)	ClawFlag (平台)
核心价值	便利性 (Convenience)	掌控感 (Control) & 信任 (Trust)
解决问题	“我不在电脑前时如何使用 Agent？”	“我如何确保我的 Agent 不会失控？”
产品形态	轻量级遥控器	高可靠指挥与控制中心
商业模式	社区贡献 (维生素)	风险管理与价值放大 (止痛药/保险)
护城河	先发优势，用户体验	生态标准，网络效应，信任壁垒

ClawFlag 的竞争对手不是 Happy Coder，而是 AI Agent 规模化应用后必然出现的信任、安全和治理危机本身。

---

2. 目标用户

目标用户画像保持 v4.0 的核心划分，但对其核心痛点的理解更加深刻。

层级	用户画像	核心痛点（v5.0 新认知）
P0 (70%)	**技术探索者**	**恐惧与无助：** 害怕 Agent 失控、产生高额费用、泄露隐私，但又不知道如何监控和阻止。

P1 (20%)	**多 Agent 运营者**	**混乱与低效：** 管理多个 Agent 的配置、记忆和成本如同噩梦，缺乏统一视图和批量操作能力。
P2 (10%)	**结果关注者**	**信息黑洞：** 完全不知道 Agent 的工作状态和成果，除非主动去问，无法评估其 ROI。

---

3. 最初五分钟 — 用户旅程 (v5.0 全新路径)

这是对产品核心逻辑的根本性重塑，是 MVP 必须验证通过的生命线。

✅ 理想路径 (零摩擦连接)

1. **下载 (15秒):** 用户在 App Store 或 Google Play 下载 ClawFlag App。
2. **创建连接 (15秒):** 打开 App，点击“+”创建新的“Bot 连接”，App 生成一个唯一的 Token (例如 `cf_tok_xxxxxxxx`)。
3. **安装 Agent (60秒):** 用户在自己的电脑（运行 OpenClaw 的地方）打开终端，执行 `npm install -g clawflag-agent`。
4. **配对 (10秒):** 在同一终端中执行 `clawflag-agent --token cf_tok_xxxxxxxx`。
5. **连接成功 (3秒):** 终端显示“✅ 已连接至 ClawFlag 云端中继”，同时手机 App 上的 Bot 状态变为绿色“在线”，并自动跳转到“驾驶舱”页面。

整个过程，用户无需了解 IP 地址、端口、防火墙或任何网络知识。

---

4. 最小可行治理产品 (MVGP) — 功能与优先级

根据专家委员会的强烈建议，v5.0 摒弃了 v4.0 的“大而全”功能列表，聚焦于一个“最小可行治理产品”（Minimum Viable Governance Product, MVGP）。这个 MVP 的唯一目标，就是解决用户最核心的恐惧，建立最基本的信任。

MVGP 核心三大支柱：

① 绝对可靠的“紧急停止按钮” (The Big Red Button)

这是产品的灵魂，是信任的基石。一个在 UI 上永远置顶、高亮、可一键触达的按钮。

- **功能：** 无论 App 或 Agent 处于何种状态，按下此按钮，必须在 1 秒内可靠地终止所有正在运行的 Agent 任务。
- **技术要求：** 该指令通过独立的、高优先级的信道（如 APNs/FCM 的紧急推送）发送，确保在网络拥堵甚至主应用无响应时依然能够触发。

- **价值主张：** 给予用户最终的、绝对的控制权。它是一个强大的信任符号。

② 基础版“驾驶舱” (Show Your Work Cockpit)

变“黑箱”为“玻璃盒”，提供最核心的“激进透明度”。

- **功能：** 在对话界面，当 Agent 回复时，旁边提供一个“探查”按钮。点击后，以可折叠的树状视图清晰展示该任务的思考链、工具调用和子任务分派。
- **技术实现：** `clawflag-agent` 实时监听 OpenClaw 的结构化日志，解析关键事件，加密后通过 WebSocket 实时推送到 App 进行渲染。
- **价值主张：** 解答用户“它在干什么？”的核心疑问，建立对 Agent 工作过程的基本认知。

③ 极简“成本熔断器” (Cost Fuse)

解决用户的财务焦虑。

- **功能：** 在 App 中提供一个简单的输入框：“当本日累计成本超过 [__] 美元时，暂停所有 Agent 活动并向我发送警报。”
- **技术实现：** `clawflag-agent` 持续监控成本数据，达到阈值后，立即暂停任务并向 App 推送高优先级通知。
- **价值主张：** 提供财务上的确定性和安全感，防止意外产生高额账单。

---

5. 完整功能版图 (Post-MVGP)

MVGP 是起点，不是终点。在核心治理能力得到市场验证后，我们将逐步解锁 v4.0 中规划的、并经过专家评审优化的深度功能，构建完整的产品护城河。这些功能将主要面向 Pro 和 Team 用户。

5.1 深度治理与安全

- **三层成本熔断器:** 从 MVGP 的单一阈值，升级为 v4.0 定义的“警告层”、“降级层”和“熔断层”的完整体系。
- **Gateway 安全审计:** 首次连接及之后定期运行，根据云端更新的安全清单，自动检查版本、公网暴露、认证配置、恶意技能等，并提供修复建议。
- **技能护盾 (Skill Shield):** 对 OpenClaw 技能进行安装前静态分析、安装后行为沙箱监控、更新时差异对比，实现完整的“软件供应链安全”。

5.2 智能运营与优化

- **ClawRouter 可视化模型路由:** 提供拖放式 UI，用于配置复杂的模型路由规则、备用链和决策预览。
- **成本顾问引擎:** 基于历史数据，主动提供可操作的成本优化建议，如“将此定时任务切换到更便宜的模型，每月可节省 XX 元”。

- **Agent 每周文摘:** 自动生成每周摘要报告，通过推送或邮件发送，总结任务、开销、安全事件和模型使用情况。

5.3 记忆与认知管理

- **永久记忆确定性编辑器:** 提供可视化界面，让用户能够安全、确定地编辑 Agent 的核心记忆（`SOUL.md`，`USER.md`），并提供版本控制。
- **分层零配置:** 通过 L0-L3 的分层界面，向不同水平的用户暴露不同复杂度的配置选项，兼顾易用性与专业性。
- **记忆智能:** 实现三阶段压缩、记忆保真度评分、记忆垃圾回收等高级功能，确保 Agent 记忆的长期健康。

5.4 增强型对话体验

- **工具调用内联卡片:** 在对话中以丰富的卡片形式展示工具调用的结果，如浏览器截图、文件差异、可折叠的代码块。
- **上下文管理:** 提供单条消息成本标签、上下文压缩警告及一键压缩按钮。

---

6. 信息架构

MVGP 阶段信息架构

标签	图标	功能
对话	💬	与 Agent 的基础对话。每个 Agent 回复旁都有“探查”按钮以打开“驾驶舱”视图。
控制台	🕹️	**紧急停止按钮**。成本熔断器设置。Bot 连接管理（添加/删除 Token）。
设置	⚙️	账户信息，通知设置，帮助与反馈。

完整版信息架构 (Post-MVGP)

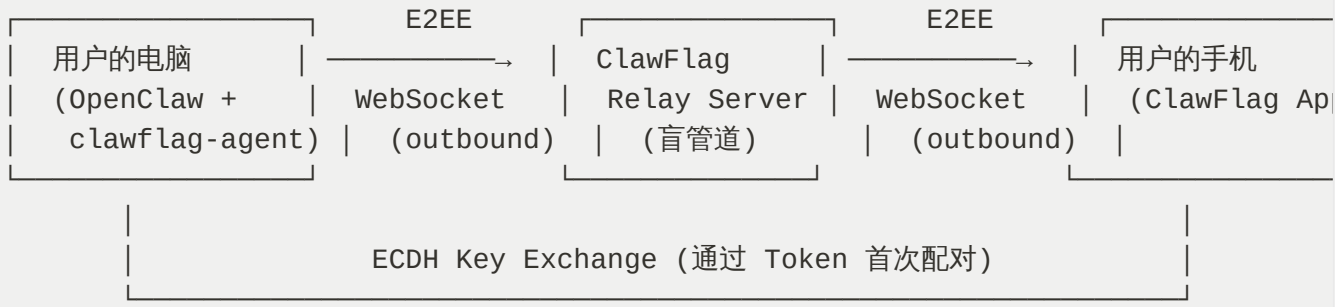
标签	图标	功能
对话	💬	增强型对话体验，包含内联卡片、成本标签、上下文管理等。
驾驶舱	📊	实时监控仪表盘，整合概览视图、会话列表、任务状态、成本

		图表、安全警报。
大脑	🧠	统一的记忆与配置中心，包含记忆浏览器、分层配置编辑器。
路由	⚡	ClawRouter 模型路由、成本顾问、预算与熔断器高级设置。
技能	🛡️	技能护盾，管理已安装技能的权限、安全状态和更新。

7. 技术架构

7.1 整体架构：三层盲中继模型

Plain Text



7.2 安全模型：盲管道原则的实现

- **密钥交换：** 用户在 App 生成 Token，`clawflag-agent` 使用此 Token 向 Relay Server 注册，服务器协助完成 App 与 Agent 之间的 ECDH 密钥交换。密钥本身永不经过服务器。
- **端到端加密：** 所有通信使用协商出的会话密钥，采用 AES-256-GCM 加密。Relay Server 只处理无法解密的二进制数据块。
- **开源可验证：** Relay Server 和 `clawflag-agent` 的代码将完全开源（MIT 许可），允许社区审计和自建。
- **前向保密：** 每个会话使用独立的临时密钥，确保历史通信的安全。
- **本地安全：** App 端使用 iOS Keychain / Android Keystore 存储敏感信息；关键操作需要生物识别确认。

7.3 `clawflag-agent` (Node.js)

作为 `npm` 全局包分发，其核心职责：

- 通过 WebSocket 维持与 Relay Server 的长连接。

- 实时监听 OpenClaw 的结构化 JSON 日志流。
- 监控成本和状态指标。
- 接收并执行来自 App 的加密指令（如“紧急停止”）。
- 将状态和日志数据加密后发送到 App。

7.4 技术栈

类别	技术
框架	React Native + Expo (PWA 备用)。CodePush 用于热更新。
通信	WebSocket + 高优先级推送 (APNs/FCM)
本地存储	WatermelonDB / SQLite (加密)
中继服务器	Go / Rust (高性能、低资源占用)
Agent	Node.js (TypeScript)

---

8. 商业化

Freemium (免费增值) 模型

- **免费版 (Free):** 包含完整的 MVGP 功能：紧急停止按钮、基础驾驶舱、成本熔断器。支持 1 个 Bot 连接。**核心逻辑是：用最核心的治理功能免费获取海量用户，建立事实标准。**
- **Pro 版 (¥68/月):** 在 MVGP 验证成功后推出。包含多 Bot 连接、高级驾驶舱分析、记忆编辑器、技能护盾、详细成本报告等深度治理功能。
- **团队版 (¥198/月/席):** 多 Gateway 管理、RBAC、统一审计日志、共享路由模板等团队协作功能。

核心逻辑转变： v4.0 的付费点是“解锁高级功能”，v5.0 的付费点是“从单一 Agent 的基础治理，升级到多 Agent 的高级运营”。

---

9. 市场进入策略 (GTM)

核心策略：成为 OpenClaw 社区的“首席安全官”。

1. **产品即营销：** 将 `clawflag-agent` 开源，并提供一个简单的“一键安全加固”脚本。该脚本能自动检查并警告常见的安全风险（如公网暴露、弱认证），并引导用户安装 ClawFlag App 以获得实时监控和“紧急停止”能力。这是最强大的增长钩子。

2. **社区深度绑定：** 在 OpenClaw 的 Discord 和 GitHub 中积极贡献，将 ClawFlag 定位为提升生态安全性的公益性项目，而非一个纯粹的商业工具。主动为社区头部用户提供支持，赢得他们的背书。
3. **升维竞争：** 如果 Happy Coder 等竞品进入 OpenClaw 市场，不进行功能对标。反复强调 ClawFlag 的“治理”和“安全”属性。宣传口径：“Happy Coder 让你‘看到’你的 Agent，而 ClawFlag 让你‘掌控’你的 Agent。”

---

10. 关键风险

风险	描述	缓解措施
执行风险	即使是 MVGP，技术复杂性依然很高，尤其是“紧急停止”的可靠性。	将 80% 的工程资源投入到“紧急停止”的可靠性上，将其作为产品的生命线。
信任鸿沟	用户可能不相信“盲中继”的承诺。	彻底开源，邀请知名安全研究员进行公开审计，提供自建 Relay Server 的详细文档。
生态依赖	OpenClaw 的 API 或日志格式发生破坏性变更。	构建灵活的适配器层，并与 OpenClaw 核心社区建立紧密沟通渠道，提前获取变更信息。
平台竞争	OpenClaw 官方或大型科技公司推出功能重叠的产品。	聚焦跨框架支持，定位为中立的第三方治理平台，并通过开源和社区建立防御性护城河。

---

11. 执行路线图

Phase 0: 核心验证 (2-4 周)

- 搭建 Relay Server，实现 `clawflag-agent` 和 App (PWA) 之间的加密消息传递。
- 实现最基础的“紧急停止”指令通路。

Phase 1: MVGP 发布 (4-8 周)

- 完善“紧急停止按钮”的 UI 和可靠性工程。
- 实现基础版“驾驶舱”（日志树状视图）。
- 实现极简“成本熔断器”。
- 上架 App Store 和 Google Play。

- 在 OpenClaw 社区正式发布，并同步开源 `clawflag-agent`。

Phase 2: 迭代与增长 (8-16 周)

- 根据 MVGP 的用户反馈，决定下一个要解决的核心痛点（可能是记忆管理或多 Bot 支持）。
- 推出 Pro 版订阅。

---

12. 长期愿景：BitFlag 网络

本节作为战略北极星，为架构决策提供方向性背景，其内容继承自 v4.0，并根据新架构进行了微调。BitFlag 网络旨在构建一个受管理、可审计、由人类控制的 Agent 社交网络，实现 Agent 之间的安全协作与交易。ClawFlag 作为工具层，是实现这一网络愿景的基石。

---

13. 参考文献

[1] SecurityScorecard STRIKE. (2026, February). The Rise of Exposed AI Agents: A New Attack Surface.

[2] Manus AI Expert Panel. (2026, February 21). ClawFlag vs. Happy Coder: 10-Expert Deep Dive & Strategic Opportunity Analysis.

[3] Happy Engineering. (2026). Happy Coder - Claude Code Anywhere.

[4] Manus AI Product Strategy Review. (2026, February 21). ClawFlag New Architecture: Top Product Expert Review.