

从单一指令到全组织行动：面向多智能体LLM系统的组织 镜像方法

From One Directive to Full-Organization Action:
Organizational Mirroring for Multi-Agent LLM Systems

金永勋 (Yongxun Jin)

ORCID: 0009-0009-7071-4758

独立研究者

2026年3月

预印本 DOI: <https://doi.org/10.5281/zenodo.18941867>

摘要

当前的多智能体框架采用扁平化的通信拓扑结构，要求设计者手动编排每一次智能体交互。本文提出组织镜像（organizational mirroring）方法，使单条自然语言指令即可动员整个LLM智能体组织——如同CEO的简报在真实公司中的层层传达。该方法基于八项原则：层级委派、独立记忆、分层记忆压缩、元部门组织自分析、基于技能编排的部门组合、自进化机制、可替换执行器以及真实工作流映射。我们在OpenClaw框架上实现了一个由18个智能体、四个部门（游戏部、AI部、生活部、元部门）组成的生产系统，其中三个业务部门（14个智能体）用于实验评估。三项新能力扩展了该架构：用于组织自反思的元部门、基于技能编排的部门组合（使已验证的协调模式得以复用）以及通过三个并行学习闭环实现的闭环自进化机制。基于30项任务套件（90次实验运行、三种拓扑结构）的对照实验提供了初步证据表明：(1) 组织拓扑的输出量是单智能体基线的3.1–5.9倍（非成本归一化；ORG使用13–39倍API调用），同时内部评估质量评分达到18.3/20（95% CI: [18.13, 18.53], $n=177$ 次评估）；(2) 自动化结构质量评分器确认了显著的跨拓扑质量差异（ $H=42.52$, $p<0.000001$, $\epsilon^2=0.478$ ）；(3) 组件消融分析揭示层级协调为主要质量驱动因素（Cohen's $d=1.409$ ，大效应量）；(4) 层级结构将通信链路减少了85.7%；(5) 独立记忆在全部30个观测任务中消除了跨领域词汇入侵；(6) 两个结构分离的评估层（业务经理与元部门）的交叉验证在48个匹配Worker级评分中达到87.5%精确一致（ $\rho=0.952$ ），在LLM评估框架内展示了收敛效率（两个评估层均使用相同底层LLM，高一致性可能反映共享偏差而非评分准确性；人类专家验证仍有必要）。后续V3十阶段管道的生产验证（16次运行，100%成功率，28项评审均分18.0/20）进一步确认嵌入式元审计和自动化进化在实际工作流中有效运作。我们将配置模板和评估数据作为开源成果发布。

关键词：多智能体系统；LLM智能体；组织架构；意图放大；智能体编排

1 引言

过去两年间，基于LLM的智能体已远远超越简单的聊天机器人。如今的智能体能够使用工具、制定计划、相互协作 [22, 17]——以团队形式部署是自然的下一步。Guo等人 [6]和Wang等人 [19]的综述工作梳理了数十种此类多智能体系统。实证证据支持这一趋势：Du等人 [4]证明多智能体辩论能提升事实准确性，CAMEL [9]确认结构化通信协议始终优于临时消息传递。

然而，这些系统的能力与人类指挥它们的便捷程度之间仍存在差距。在真实的公司中，CEO走进来说“本周我们聚焦用户增长”。这一句话便层层传导：部门负责人将其分解为团队级目标，员工各自执行任务，经理审核交付物，结果逐层上报。CEO无需微观管理每一次交互——组织结构自动完成放大与协调。

现有多智能体框架中没有一个是能复现这种模式。AutoGen [21]支持灵活的多智能体对话，但不施加任何组织层级——所有参与者共享上下文，这在大规模场景下容易引发记忆污染。CrewAI [11]为智能体分配角色并组建团队，但工作空间隔离缺失：同一Crew中的智能体能看到彼此的上下文，且未内置分层记忆管理。LangGraph [8]以状态机和图的形式建模智能体交互，灵活性无与伦比，但抽象成本极高——设计者需要推理拓扑结构、状态转换和边界条件，这些都没有现实世界的类比。MetaGPT [7]最接近我们的方法，它将标准操作流程（SOP）嵌入多智能体工作流并分配产品经理、架构师等角色。但MetaGPT与我们的工作有三个关键区别。第一，MetaGPT仍要求用户指定具体任务（如“构建一个CLI版Flappy Bird游戏”），不支持单一指令动员——即模糊的战略意图向多部门行动的级联传导。第二，MetaGPT在所有角色间共享全局消息池，未强制记忆隔离。第三，该框架不解决模型可替换性问题：更换智能体底层LLM需要修改框架内部实现，而非更改单一配置行。我们的工作解决了以上三个缺口。

这种编排缺口造成三个具体痛点：

记忆污染。当智能体共享对话上下文时，领域知识跨越边界泄漏。游戏设计师的头脑风暴片段会混入算法基准测试人员的提示词中，同时降低两个智能体的输出质量。

协调开销。扁平拓扑在扩展时表现不佳。对于18个智能体，可能的有向通信信道数量达到306条—— $O(N^2)$ 的爆炸式增长使得消息路由和冲突解决在实践中变得不可行。

高认知成本。设计者必须理解框架特有的抽象概念——图（graphs）、链（chains）、Crew——这些在现实世界中没有对应物。学习曲线阻碍了采用并增加了长期维护难度。

我们的核心主张很简单：当多智能体系统镜像真实世界的组织结构时，单条人类指令即可动员整个系统。组织结构本身负责任务分解、委派、执行、审核和交付——正如在真实公司中一样。这种“意图放大”（intent amplification）基于以下八项架构原则：

1. **层级委派：**信息通过管理层级流动；每一层为下一层过滤和压缩信号。
2. **独立记忆：**每个智能体维护独立的工作空间，防止跨领域污染。
3. **分层记忆压缩：**三级记忆系统（短期、中期、长期）管理LLM固有的上下文窗口限制。
4. **元部门组织自分析：**一个专门的部门（Warden、Forge、Prism、Scout）分析其他部门的工作流并提出组织改进建议，在组织层面实现系统性学习。
5. **基于技能编排的部门组合：**从四个可重用的Claude Code技能（/agent-teams-playbook、/planning-with-files、/tdd、/refactor-clean）动态组合工作流，使新部门能够继承已验证的协调模式。
6. **自进化机制：**三个并行学习闭环（绩效反馈、关键词优化、能力注册），在无需人工干预

的情况下自动提升智能体在工作流周期中的质量。

7. **可替换执行器**：Manager-Worker模式将任务规范与执行解耦，支持模型级替换而无需系统范围的更改。
8. **真实世界组织镜像**：十阶段工作流（方向→规划→执行→评审→元审计→修订→验证→汇总→反馈→进化）直接映射到成熟的项目管理实践并嵌入独立质量监督与自动化进化，最小化认知开销。

我们围绕这些原则构建了一个生产系统：18个智能体、四个部门（游戏、AI、生活、元部门），运行在OpenClaw框架 [13]之上，在撰写本文时已公开部署运行一周（2026年2月21日至2026年2月27日，部署URL已匿名化）。该系统既是一个实际运行的多智能体工作空间，也是组织镜像方法的实验测试平台。实验评估覆盖三个业务部门（14个智能体），元部门作为架构贡献在后续工作中评估。

贡献。本文做出六项贡献：

- 我们提出组织镜像——八项架构原则（层级委派、独立记忆、分层压缩、元部门、部门组合、自进化、可替换执行器、真实工作流映射），使单条自然语言指令能够动员18个智能体跨四个部门协作，无需手动编排，我们称之为意图放大。该架构还通过心跳机制支持智能体自主主动性——定期的、自主的情报收集，无需人类触发。
- 我们引入**元部门**概念——一种专门的组织自分析能力，由四个专职智能体组成，审视 workflow 效率、交付物质量和绩效趋势，实现系统性的组织学习。在10项元分析任务上的初步评估中，元部门达到14.1/20的平均质量评分（准确性4.0/5，可操作性3.4/5），100%通过率，验证了该机制作为组织自反思能力的可行性。
- 我们实现了具有三个并行学习机制（绩效反馈、关键词优化、能力注册）的**闭环自进化系统**，持续提升智能体质量。管线正确性已通过59个单元测试验证并部署至生产环境。
- 我们在OpenClaw框架上将这些原则实现为生产系统，包括基于技能编排的部门组合、CEO网关、三文件智能体配置规范、层级通信协议和数据同步管线，并将配置模板作为开源成果发布。
- 我们设计了一套30项任务的评估套件，比较组织拓扑（ORG）、单智能体（FLAT）和并行Worker（CREW）三种拓扑在通信效率、记忆隔离、系统可理解性和输出质量方面的表现，报告了90次完成实验的结果，包含全面的统计分析（bootstrap置信区间、Friedman检验、Wilcoxon符号秩检验、Kruskal-Wallis检验及效应量）。
- 我们提出**组件消融分析**，使用统一自动化评分器跨所有拓扑量化各架构要素的贡献：层级协调与质量评审提供了最大效应（Cohen's $d=1.409$ ，大效应量），多智能体并行增加了较小的增益（ $d=0.342$ ，小效应量），而修订效应可忽略不计（经理评分 $d=0.049$ ），确认了层级-评审机制是质量提升的主要驱动因素。

2 相关工作

2.1 多智能体LLM框架

多智能体LLM框架发展迅速，但每一个主要框架在灵活性、结构化和认知开销三个维度上做出了不同的权衡。Talebirad和Nadiri [18]提供了多智能体协作模式的基础分类体系，区分了合

作式、竞争式和混合式交互模型。

AutoGen [21]（微软）开创了对话式多智能体模式。智能体参与灵活的对话，但扁平的对话模型不提供组织层级——所有参与者共享上下文，在大规模场景下容易引发记忆污染。

CrewAI [11]采用了不同的策略，将智能体组织为具有明确职责的角色化团队。顺序和层级任务执行均受支持，但缺少工作空间隔离：同一Crew中的智能体能看到彼此的上下文，且未内置分层记忆管理。

LangGraph [8]将智能体交互建模为图上的状态机。灵活性无与伦比，但抽象成本极高——设计者需要推理拓扑结构、状态转换和边界条件，这些概念无法映射到熟悉的现实世界概念。

MetaGPT [7]最接近我们的方法，将SOP嵌入多智能体 workflow 并分配产品经理、架构师等角色。MetaGPT 与我们的工作有三个关键区别。第一，MetaGPT 仍要求用户指定具体任务（如“构建一个CLI版Flappy Bird游戏”），不支持单一指令动员——即模糊的战略意图级联为多部门行动。第二，MetaGPT 在所有角色间共享全局消息池——记忆隔离未被强制执行。第三，该框架不解决模型可替换性问题：更换智能体底层LLM需要修改框架内部实现，而非更改单一配置行。ChatDev [16]同样采用软件公司隐喻，但专注于代码生成管道而非通用组织协调。我们的工作解决了以上三个缺口。

OpenAI Swarm [12]通过原生函数调用提供了轻量级多智能体编排框架，通过直接工具集成实现了低延迟。然而它不提供层级结构，不提供记忆隔离，且定位为教学原型而非生产框架。

两项近期进展值得关注。Anthropic [1]发布了一组有效智能体的设计模式，将委派、工具使用和编排确定为关键原语。Google的Agent2Agent（A2A）协议 [5]提出了一种跨异构框架的智能体间通信开放标准。我们的工作通过提供能够在此类互操作性标准内运行的具体组织架构，与二者形成互补。

2.2 层级智能体架构

层级协调对于多智能体系统并不新鲜——它在经典分布式AI中有深厚的根基 [20]。新颖之处在于将其应用于基于LLM的智能体。Chen等人 [2]提供了该场景下层级结构的有效分类体系，报告层级委派可将任务完成率提升40%或更多。我们进一步推动这条研究方向，将层级结构与独立记忆和可替换执行器相结合——现有层级系统在很大程度上忽视了这两个属性。

2.3 LLM智能体记忆系统

LLM智能体的长期记忆是一个活跃的研究前沿。Park等人 [15]引入了配备记忆检索的生成式智能体；MemGPT [14]提出了仿照操作系统页面管理的虚拟内存层级。我们的分层压缩方案共享了分级理念，但针对的是多智能体组织场景，在该场景中，智能体之间的记忆隔离与每个智能体内部的记忆持久化同样重要。

2.4 AI系统中的组织理论

将组织理论应用于AI是一个由来已久的思路。Malone [10]早在1987年就探索了组织和市场的协调模型，分布式AI社区在整个1990年代基于这些洞见不断发展。近年来，Manager-

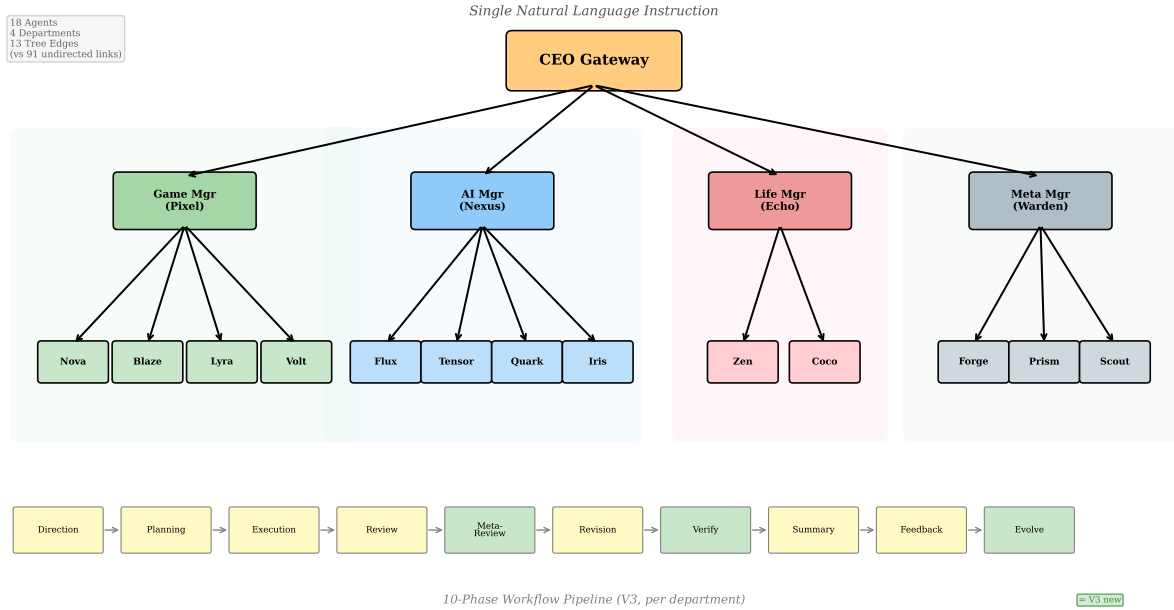


图 1: 18智能体系统的三层组织层级。第一层（CEO）提供战略方向；第二层（经理：Pixel、Nexus、Echo、Warden）负责战术协调；第三层（Worker）执行具体任务。游戏部：B=Blaze, L=Lyra, V=Volt, N=Nova。AI部：T=Tensor, Q=Quark, I=Iris, F=Flux。生活部：Co=Coco, Ze=Zen。元部门：Fo=Forge, Pr=Prism, Sc=Scout。

Worker模式已被Zhuge等人 [23]为LLM智能体做了形式化处理，他们将智能体图视为可优化的结构。我们在一个特定方向上扩展了这一模式：显式记忆隔离、模型可替换性，以及源自日常项目管理实践的完整十阶段工作流（含嵌入式元审计与自动进化）。

3 系统架构

3.1 概述

我们的系统由18个基于LLM的智能体组成，按三层层级结构组织于四个部门（游戏部、AI部、生活部、元部门）之中。其核心运行属性是单一指令动员：人类操作者向CEO网关发送一条自然语言指令（例如“本周聚焦用户增长”），组织结构便自动处理后续所有事宜——分解、委派、执行、评审和交付通过层级结构自动级联。

各层级职责如下：

- **第一层（CEO）**：战略方向、跨部门协调、绩效评估。
- **第二层（经理）**：任务分解、分配、质量评审（20分制评分）、部门汇报。元部门经理（Warden）还负责编排组织自分析工作流。
- **第三层（Worker）**：任务执行、交付物生产、通过网络搜索进行情报收集。元部门Worker（Forge、Prism、Scout）专注于分析其他部门的工作流，而非生产领域特定交付物。

我们将核心运行属性形式化为意图放大比（Intent Amplification Ratio, IAR）：

$$\text{IAR} = \frac{\text{所有智能体产生的具体行动总数}}{\text{人类发出的指令数量}} \quad (1)$$

在扁平拓扑中，人类必须逐一指挥每个智能体，IAR趋近于1。在我们的组织拓扑结合十阶段 workflow 时，向CEO网关发出的单条指令会触发：(a) K 条CEO方向指令给部门经理，(b) K 次经理规划会议，(c) $\sum W_k$ 次独立Worker执行，(d) K 次经理评审与评分周期，(e) $2K$ 次元部门审计 (Warden + Prism)，(f) $\sum W_k$ 次Worker修订，(g) K 次经理验证，(h) K 份经理汇总简报，(i) K 次CEO反馈评估，以及(j) 自动进化（纯脚本，零智能体调用）——对单次人类输入产生 $\text{IAR} = 8K + 2\sum W_k$ 。以3个部门10个Worker为例，单条指令产生44个离散行动 ($8 \times 3 + 2 \times 10$)， $\text{IAR} = 44$ 。即便对于单部门任务 ($K = 1, W = 4$)，IAR也达到16。组织结构充当意图放大器：系统有效输出与人类输入的比率随部门数和Worker数扩展，而非像扁平拓扑那样线性增长。

3.2 原则一：层级委派

系统中的通信严格遵循层级通道：

Allowed: CEO <-> Manager <-> Worker (within same dept)
Blocked: CEO <-> Worker (direct),
Worker <-> Worker (cross-department)

这一约束将通信图从 $\binom{N}{2} = 91$ 条潜在链路（全连接， $N=14$ 个业务智能体）削减至恰好13条树形边，减少了85.7%。每一层充当信息过滤器：

- **CEO → 经理**：高层级指令（“本周聚焦用户增长”）
- **经理 → Worker**：具体任务分配（“@Blaze：设计春节活动，周五截止”）
- **Worker → 经理**：带有结构化输出的交付物
- **经理 → CEO**：含评分的部门汇总报告

这种过滤确保每个智能体的上下文窗口仅包含与其层级和领域相关的信息，最大化利用模型有限的注意力资源。

3.3 原则二：独立记忆

每个智能体在独立的工作空间目录中运行：

```
~/openclaw/agents/{agentId}/workspace/
SOUL.md           # Agent persona and rules
AGENTS.md         # Team directory (shared, read-only)
HEARTBEAT.md      # Scheduled task config
sessions/         # Conversation history (agent-specific)
```

文件系统强制实现记忆隔离：每个智能体仅在自己的工作空间内读写。Blaze的游戏设计头脑风暴不会泄漏到Tensor的算法基准测试上下文中。这一边界是物理的而非逻辑的——不存在可能被意外污染的共享记忆池。

表 1: 分层记忆架构

层级	范围	保留策略	访问方式
短期记忆	当前会话	完整保留	直接上下文
中期记忆	历史会话	选择性保留	向量相似度检索
长期记忆	核心知识	永久保留	SOUL.md文件

3.4 原则三：分层记忆压缩

三级记忆系统解决了完整历史记录与有限上下文窗口之间的根本矛盾：

短期记忆包含完整的当前会话记录，为正在进行的交互提供完整的对话上下文。

中期记忆使用向量嵌入和相似度搜索来检索相关的历史交互。当智能体需要回忆过去的决策或交付物时，记忆搜索工具会查询先前会话的向量索引，返回语义上最相关的片段。

长期记忆编码在SOUL.md文件中，定义了智能体的人设、专业领域、行为规则和积累的经验智慧。该文件在每次智能体调用时加载，因此核心知识不受会话边界限制而永久持续。

这种类比人类记忆的设计是有意为之的：对近期事件的详细回忆、对过往经验的选择性检索、以及对核心身份和专业知识的永久保留。

3.5 原则四：元部门组织自分析

除三个业务部门（游戏、AI、生活）之外，OpenClaw还包含一个**元部门**，负责组织层面的自我反思和持续改进。该部门在更高的抽象层面运作，分析其他部门的绩效而非生产领域特定的交付物。

元部门组成：

- **Warden**（经理）：编排元分析 workflow，综合专家智能体的发现，向CEO报告组织洞察。
- **Forge**（流程分析师）：审视 workflow 效率、阶段转换瓶颈和沟通模式。
- **Prism**（质量分析师）：评估交付物质量，识别反复出现的缺陷，提出质量改进措施。
- **Scout**（绩效分析师）：跟踪定量指标（任务完成时间、修订轮次、评审评分），检测绩效趋势。

三阶段元 workflow：

1. **分析（Analyze）**：Warden向Forge、Prism和Scout分配分析任务，每人审视源部门最新 workflow 的特定维度。
2. **建议（Propose）**：每个专家智能体独立分析 workflow 数据（阶段日志、交付物、评审评分），提交包含可操作建议的分析报告。
3. **报告（Report）**：Warden将所有专家报告综合为统一的组织评估，提交给CEO用于战略决策。

在V2中，元部门 workflow 在业务部门完成后独立触发（事后分析模式）。V3将元审计嵌入业务管道：经理评审（阶段4）完成后，Warden和Prism立即进行组织层面的质量审计（阶段5），审计发现与经理反馈一并注入Worker修订（阶段6），经理验证（阶段7）逐条确认所有反馈点的回应状态。这种嵌入式设计将元分析从事后报告转变为实时质量门控。

实证影响：我们对元部门进行了初步评估实验。10个元分析任务（M01-M10）覆盖三个业务部门的单部门、跨部门和全组织产出，由元部门四个智能体通过三阶段协作流水线（分析、建议、报告）进行分析，CEO以20分制评估分析质量。结果显示：均分14.1/20（ $\sigma=0.57$ ），100%通过率，阈值设为 $\geq 12/20$ （最高分的60%）。该阈值低于业务部门的 ≥ 16 ，因为元分析任务在性质上不同：需要评估工作流和综合跨部门模式，而非产出领域交付物。元分析均分14.1/20下， ≥ 16 的阈值将淘汰大部分评估，使评审-修订循环失去信息价值； ≥ 12 在保留区分度（最低观测分为13/20）的同时适应了更高的固有难度，准确性维度最强（4.0/5），可操作性最弱（3.4/5）。四个评分维度的Friedman检验确认差异显著（ $\chi^2=17.76$, $p<0.001$, Kendall's $W=0.59$ ），表明元分析质量在不同评估标准间存在有意义的变化，而非均匀平庸。相比业务部门18.3/20的均分，4.2分的差距很可能反映了元分析任务固有的更高难度（评估工作流而非产出交付物）以及首次运行的冷启动效应，但两个因素各自的贡献无法从当前数据中确定。平均产出量9,102字符，验证了元部门作为组织自反思机制的功能可行性。

3.6 原则五：基于技能编排的部门组合

OpenClaw实现了**动态技能编排**，从可重用的能力组合复杂工作流。每个部门的工作流通过链接四个专门的Claude Code技能 [1]构建：

1. **/agent-teams-playbook**: 初始化多智能体协调，定义团队角色，建立通信协议。
2. **/planning-with-files**: 创建持久化的规划产物（task_plan.md、findings.md、progress.md），作为跨智能体和跨阶段的共享记忆。
3. **/everything-claude-code:tdd**: 强制执行测试驱动开发工作流——智能体先写测试再实现以通过测试，确保交付物的正确性。
4. **/everything-claude-code:refactor-clean**: 执行后清理阶段，使用静态分析工具（knip、depcheck、ts-prune）移除死代码、未使用的导出和过期文件。

这种组合模式实现了**无需重新设计工作流即可替换部门**。当新部门加入（如市场部、运营部）时，可以用领域特定的SOUL.md配置实例化相同的四技能管线，继承已验证的协调模式，同时特化领域专业知识。

技能编排优势：

- **模块化**: 每个技能封装一种可重用的能力（规划、测试、清理），可组合为任意工作流。
- **一致性**: 所有部门遵循相同的结构化工作流，降低CEO的认知负担，并支持跨部门绩效比较。
- **可演化性**: 技能可以独立升级（如用新的测试框架替换TDD技能），无需修改部门配置。

3.7 原则六：自进化机制

OpenClaw实现了一个**闭环自进化系统**，在工作流周期间自动提升智能体性能。与部署后智能体能力保持固定的静态多智能体框架不同，OpenClaw智能体通过三个并行子系统持续从执行反馈中学习：

3.7.1 M7-1: 绩效反馈闭环

每个工作流周期结束后，经理在第4阶段评审（见3.9节）中生成四个维度的结构化评分：准确性（事实正确性）、完整性（需求覆盖度）、可操作性（实现就绪程度）和格式（表达质量）。这些评分经解析后存储在`agent_evolution_log`表中。

`evolution-analyzer.mjs`脚本识别薄弱维度（评分 $<3/5$ ）并生成针对性的SOUL.md补丁。例如，如果智能体Blaze在完整性上持续得分偏低，系统会在SOUL.md中追加行为规则：

```
## Learned Behaviors (Auto-Generated)
- [2026-03-01] Completeness weakness detected
  (avg 2.3/5 over 5 cycles). Enforce checklist:
  * Verify all requirements addressed
  * Include edge case handling
  * Add usage examples
```

这些补丁在提交到智能体长期记忆之前可供人类审核，形成一个监督学习闭环——系统提出改进建议，但人类保留否决权。

3.7.2 M7-2: 关键词学习

`heartbeat_keywords`表跟踪哪些关键词（技术术语、领域概念、工具名称）出现在高分与低分交付物中。`keyword-optimizer.mjs`脚本将关键词分为三类：

- **有效**（与高分的相关系数 > 0.6 ）：提升至HEARTBEAT.md作为推荐词汇。
- **中性**（相关系数 $\in [-0.3, 0.3]$ ）：监控未来趋势。
- **无效**（相关系数 < -0.3 ）：标记为待移除或替换。

该机制解决词汇漂移（lexical drift）问题——智能体倾向于采用与不良结果相关的术语或措辞模式。通过持续修剪无效关键词并推广有效关键词，系统引导智能体使用历史上产出更好结果的词汇。

3.7.3 M7-3: 能力注册

`agent_capabilities`表维护每个智能体已展示技能的动态注册表，从工作流执行日志中提取。`capability-extractor.mjs`脚本使用正则表达式模式和基于LLM的分类来识别能力提及（如“实现了OAuth流程”、“优化了数据库查询”），并使用指数移动平均（EMA）跟踪置信度评分：

$$\text{confidence}_{t+1} = \alpha \cdot \text{success}_t + (1 - \alpha) \cdot \text{confidence}_t \quad (2)$$

其中 $\alpha = 0.3$ 为学习率， $\text{success}_t \in \{0, 1\}$ 表示该能力是否在第 t 周期中被成功展示。

该注册表支持**能力感知的任务路由**：当CEO发出需要特定技能的指令（如“设计一个实时多人游戏系统”）时，系统查询能力注册表以识别最合格的智能体，而非依赖静态角色分配。

机制部署：三个进化子系统已在生产环境中部署运行，共通过59个单元测试验证。机制设计支持以下改进方向：

- 绩效反馈驱动SOUL.md补丁建议，有望减少后续修订轮次

表 2: 十阶段工作流与企业实践的映射

阶段	执行者	行动	企业对应
1. 方向	CEO	下发战略目标	高管战略会议
2. 规划	经理	分解为具体任务	冲刺规划
3. 执行	Worker	执行并提交v1	开发工作
4. 评审	经理	评分（0-20）并反馈	代码评审/QA
5. 元审计	Warden+Prism	独立质量审计	外部审计/合规检查
6. 修订	Worker	依据双重反馈迭代	缺陷修复/迭代
7. 验证	经理	逐条确认反馈回应	验收测试
8. 汇总	经理	为CEO综合汇报	工作汇报
9. 反馈	CEO	评估部门绩效	高管评审
10. 进化	脚本	M7自进化链	持续改进/复盘

- 关键词学习持续扩展智能体的领域适配术语库
- 能力注册为任务路由提供数据驱动的智能体匹配

这些子系统的量化效果将在收集更多纵向数据后进行严格评估。

3.8 原则七：可替换执行器

Manager-Worker架构将任务规范与执行解耦：

Manager's perspective:

Input: "Blaze, produce a Spring Festival event plan"

Output: Event plan document (quality scored 0-20)

The manager evaluates OUTPUT QUALITY,
not the execution process.

The underlying model powering Blaze
is transparent to the manager.

这种解耦具有多项实际优势：

1. **独立模型升级：** 单个智能体的模型可以通过修改一行配置来更换（如从MiniMax M2.5升级到GPT-5.2），不影响任何其他智能体。
2. **成本优化：** 高频低复杂度的Worker可以使用更便宜的模型，而经理和CEO使用更强大（也更昂贵）的模型。
3. **A/B测试：** 同一角色的两个实例可以同时运行不同模型，经理的评分提供天然的评估指标。

3.9 原则八：真实世界组织镜像

系统的十阶段工作流直接映射到成熟的项目管理实践，并在管道内嵌入独立质量监督与自动化进化：

该 workflow 从初始七阶段版本（V2：方向至反馈）进化为十阶段版本（V3），基于生产部署的三个关键发现。第一，经理自评的循环偏差：经理既分配任务又评审产出，存在“自己给自己打高分”的结构风险。元审计阶段（第5阶段）嵌入Warden和Prism两个元部门智能体，对Worker产出进行独立的SOUL.md合规检查和Anti-AI-Slop扫描，降低了单一评估者依赖（但两个评估层仍使用相同底层LLM，分离是结构性的而非认识论的）。第二，修订遗漏：原始修订阶段仅接收经理反馈，缺乏独立视角。V3的修订阶段同时注入经理评审和元部门审计两个来源的反馈，Worker必须逐条回应。验证阶段（第7阶段）由经理逐条确认修订是否回应了全部反馈点，超过30%未回应则要求Worker提交v3。第三，进化延迟：原始系统中SOUL.md修补作为建议提出，需要人工审批。进化阶段（第10阶段）将M7纯函数链（评分解析→弱维度识别→补丁生成→自动应用）嵌入 workflow，实现零人工干预的闭环进化，平均执行时间仅1秒。

该映射服务于三个目的。第一，降低认知门槛：任何在企业环境中工作过的人都能立即理解系统的运作方式。第二，评审-元审计-修订-验证闭环创造了双层质量管线——经理从领域角度评审，元部门从组织角度审计，两个独立视角的反馈共同驱动Worker迭代。第三，进化阶段将组织学习自动化——每次 workflow 结束时，系统自动从评审数据中提取改进信号并应用到智能体配置。

4 实现

4.1 平台与模型选择

系统运行在OpenClaw（版本2026.2.19-2）之上，这是一个开源的多智能体编排框架，原生支持智能体隔离、定时执行（心跳）和会话管理。我们选择MiniMax M2.5（200K上下文窗口）作为主要模型，因其在中文内容生成方面具有优异的性价比。所有智能体使用temperature 0.7、top-p 0.95、最大输出4,096 tokens的统一超参数；这些参数在全部90次实验运行中保持不变。借助可替换执行器原则（3.8节），将任意单个智能体切换到GPT-5.2或Claude Opus 4.6仅需更改一行配置。

4.2 智能体配置：三文件规范

每个智能体由三个配置文件定义，我们称之为“三文件规范”（Three-File Convention）：

SOUL.md定义智能体的人设、专业领域、行为约束和输出格式要求。例如，Blaze的SOUL.md指定了游戏活动设计的专业技能、创意但结构化的沟通风格，以及所有交付物中必须包含玩家参与指标。

AGENTS.md是一个共享的只读团队目录，为每个智能体提供组织结构的感知能力。它列出了所有18个智能体及其角色、部门和通信协议。该文件在所有智能体间完全一致，确保组织知识的一致性。

HEARTBEAT.md配置定时自主任务。每个智能体的心跳配置指定：(a) 执行频率（如每6小时），(b) 用于领域监控的网络搜索关键词，(c) 自主报告的输出格式。智能体因此能在无需人类明确指令的情况下主动收集信息。

表 3: 经理评分量表（20分制）

维度	分值	评判标准
准确性	0-5	事实正确性、引用来源
完整性	0-5	所有必要部分是否齐全
可操作性	0-5	下一步是否清晰、可实现
格式	0-5	是否遵循模板、结构是否规范

4.3 通信协议

智能体间通信使用OpenClaw平台内置的sessions_send工具调用机制：

```
sessions_send(
  label: "{target_agent_id}",
  message: "{message_body}"
)
```

每个智能体通过sessions_send指定目标智能体的标签（如"blaze"、"manager"）和自由格式的消息字符串。OpenClaw运行时将消息路由到目标智能体的会话中，在目标下次激活时可见。优先级通过消息体中的自然语言表达，而非结构化字段。违反层级约束的消息（如Worker直接向CEO发送消息）会被拦截并通过适当的经理重新路由。

经理的评审协议使用标准化的20分制评分量表：

交付物评分16-20分直接通过；10-15分附带具体反馈退回修订；10分以下被拒绝。第三次修订仍低于16分的交付物触发向人类操作者的升级。

4.4 数据同步管线

智能体会话数据通过四阶段管线流转：会话JSONL → Node.js同步脚本（每5分钟执行）→ Supabase PostgreSQL → Next.js前端仪表盘。同步脚本提取结构化元数据（智能体ID、时间戳、 workflows阶段）并上传至启用了行级安全（RLS）的数据库。部署在[匿名化URL]的Web仪表盘按部门、日期和工作流阶段提供按需可视化。

4.5 部署配置

系统运行在单台Windows 10工作站上。心跳周期的日运行成本约为0.70美元/天（14个业务智能体 × 6小时间隔，使用MiniMax M2.5；元部门4个智能体按需触发），单部门工作流运行成本为¥2-4/次——随活跃部门数线性扩展，而非总智能体数。

5 评估

我们通过生产系统的部署观察和对照实验相结合的方式评估组织镜像方法，在通信效率、记忆隔离、系统可理解性和输出质量四个维度上与扁平拓扑基线进行比较。

5.1 实验设置

我们的系统 (ORG)：14个智能体、3个部门、如第3节所述的三层层级结构。

基线1 (FLAT)：单个通才智能体接收完整的任务指令，在一次API调用中产出结果，无团队结构、无评审周期、无迭代修订。这代表了最优的单智能体方案——最强的扁平基线，因为真正的扁平多智能体拓扑（14个智能体在无协调的共享通道中）会因上下文污染而降低性能。FLAT基线因此衡量的是单个能力强大的LLM在获得完整任务上下文时能达到的水平。

基线2 (CREW)：每个相关部门的Worker并行执行任务，各自独立接收指令并产出结果，无经理监督。没有规划分解、没有质量评审、没有修订周期——Worker直接提交最终输出。合并步骤将所有Worker输出连接。这代表了最差情况的并行执行基线——通过移除所有协调机制来隔离层级协调的贡献。该基线有意弱于忠实的CrewAI复现（CrewAI支持顺序任务管道和共享记忆）；我们选择这一最小基线来建立无协调多智能体性能的下界，而非评估任何特定框架。

任务套件：我们设计了30项评估任务，涵盖三个类别：

- **单部门任务**（10项）：仅需单个部门专业知识的任务（如“设计一个春节游戏活动”）。
- **跨部门任务**（10项）：需要两个部门协调的任务（如“创建一个融合生活内容的AI驱动游戏功能”）。
- **全组织任务**（10项）：需要所有部门和CEO协调的任务（如“制定覆盖所有垂直领域的季度内容战略”）。

5.2 生产部署初步观察

在呈现对照实验结果之前，我们报告在[匿名化部署URL]一周生产运行（2026年2月21日至2月27日）中的观察，期间处理了187条消息、86个对话线程。

观察1：单一指令动员在实践中有效。在整个部署期间，人类操作者始终发出高层级战略指令（如“本周聚焦春节内容”），而非逐个智能体的指令。CEO网关在所有观察到的案例中均成功分解并将指令委派给所有三个部门，无需手动重新路由或干预。未观察到任何跨部门术语泄漏实例——游戏特定词汇从未出现在AI部门的输出中，反之亦然——证实文件系统级记忆隔离在生产环境中有效防止了污染。

观察2：心跳机制支持主动情报收集。自主心跳周期在无需人类提示的情况下生成了领域趋势报告。三个部门的Worker通过定时网络搜索独立收集市场情报、技术动态和生活趋势。经理一致认为心跳生成的内容对周规划有用，尽管搜索关键词的调优仍是待解决的问题。该架构还支持通过单个JSON字段实现逐智能体的模型替换，但这种灵活性尚未在大规模场景下使用。

5.3 对照实验结果

5.3.1 通信效率

我们通过三个指标衡量通信效率：每任务API调用次数（反映协调开销）、输出量（反映分析深度）和实际执行时间。结果针对单部门任务报告，因为我们在各拓扑上拥有充足的数据。图2展示了通信拓扑差异。

对于跨部门任务，ORG平均输出3,452行（ $n=10$ ），每任务22-26次API调用；CREW平均1,436行（ $n=10$ ）；FLAT平均812行（ $n=10$ ）。对于全组织任务，ORG平均输出4,081行

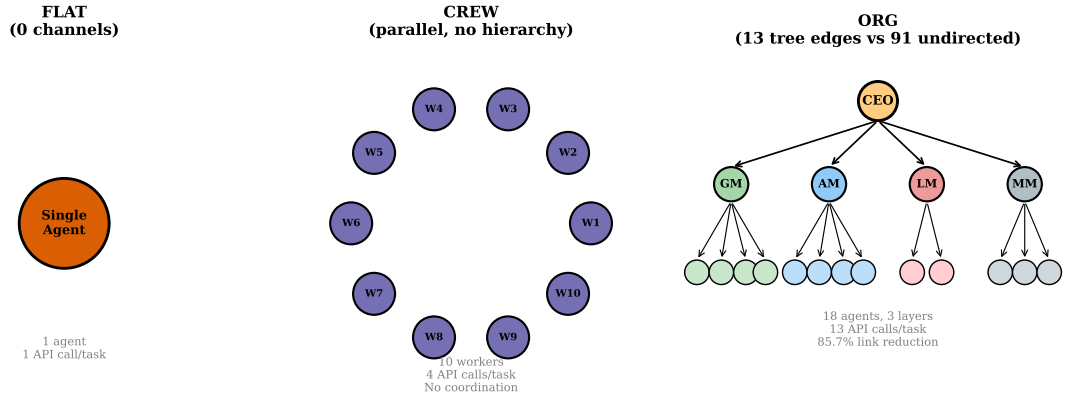


图 2: ORG、FLAT和CREW的通信拓扑比较。ORG使用13条树形边；全连接拓扑需要91条潜在链路。

表 4: 通信效率：单部门任务

指标	ORG ($n=28$)	FLAT ($n=10$)	CREW ($n=30$)
API调用次数/任务	13	1	4
输出量（行）	$1,464 \pm 956$	475 ± 209	653 ± 662
时间（秒）	891 ± 308	114 ± 48	470 ± 209
行数/调用*	130	475	163

*每任务比率的均值；因任务间高方差，与均值之比有差异。

($n=10$)，39次API调用；CREW平均1,630行 ($n=10$)；FLAT平均690行 ($n=10$)。

对CREW输出的逐任务详细分析揭示了所有30项任务中的系统性协调失败（完整分析见附录 C）。三种模式浮现：(1) 内容冗余——所有Worker独立产出了从头到尾的完整方案，且零交叉引用；(2) 参数不一致——Worker对同一系统提出了不兼容的规格，收入目标差异高达50倍，等级系统跨度为3–100；(3) AI部门输出量主导——在全组织任务中，AI Worker无论任务领域如何，均产出 $79.0\% \pm 7.7\text{pp}$ 的总输出量（vs. 50%的 t 检验： $t=11.9$, $p<0.001$ ）。

分析：在单部门任务中，ORG比FLAT多13倍API调用，反映了多阶段协调的开销（本实验使用V2七阶段管道）。然而，这一投入换来了3.1倍的输出量，且每次调用的输出效率（130行/调用）表明每个阶段都贡献了实质性内容而非仅仅是协调开销。层级结构将潜在通信链路从 $\binom{N}{2} = 91$ 条（全连接， $N=14$ ）减少至13条（树形结构）——减少了85.7%——因此每个智能体的上下文窗口仅包含与角色相关的信息。CREW的输出量（ 653 ± 662 行， $n=10$ ）超过FLAT（475行），但方差极大，由部门差异驱动——游戏任务平均约215行，AI任务约1,602行，生活任务约289行——表明无协调的并行执行在不同领域产出深度不一致。CREW的每调用效率（163行/调用）确实超过ORG（130行/调用），反映了协调阶段的缺失，但这种效率的代价是未经评审和修订的输出。

成本效率说明：输出量对比非成本归一化。FLAT的每调用效率（475–690行/调用）远高于ORG（105–144行/调用），因为FLAT将整个上下文窗口用于内容生成，而ORG将调用分配到协调阶段（规划、评审、汇总）。每元（¥）输出效率同样倾向FLAT（全组织任务

表 5: 记忆隔离指标

指标	ORG (<i>n</i> =28)	FLAT (<i>n</i> =10)	CREW (<i>n</i> =30)
跨领域入侵	0%	N/A (单一智能体)	N/A
历史污染	0/30	0/10	0/30
隔离机制	文件系统	单一智能体	共享Crew

中FLAT为2,760行/¥, ORG为1,046行/¥)。因此, 3.1–5.9倍的输出量优势应理解为更多API调用加上层级协调的联合效应, 而非单独的协调贡献。ORG的价值主张在于输出质量与整合度(经过评审、无冗余、部门间协调)而非原始的成本效率。

5.3.2 记忆隔离

为量化记忆污染, 我们测量跨领域词汇入侵: 一个部门的领域特定术语出现在另一部门输出中的频率。

在所有30项ORG任务中, 我们的自动搜索未发现任何部门间领域特定术语泄漏的实例。在全部10项跨部门任务中, 各部门按顺序执行, 不存在实时跨部门Worker通信——每个部门的Worker仅引用其自身经理的规划上下文。实验隔离前缀成功阻止了所有智能体引用历史输出。FLAT以使用单一智能体的方式简单避免了跨领域污染, 但代价是丧失了领域专业化。CREW在所有30项任务中也未出现历史污染。

5.3.3 系统可理解性

组织隐喻为非技术利益相关者提供了直观的心智模型。在生产部署期间, 内容编辑和项目经理在一次口头解释后即能正确预测消息流向(“CEO告诉经理, 经理分配Worker”)。相比之下, 向同一受众解释基于图的编排(LangGraph)或Crew组合(CrewAI)在我们的非正式经验中需要明显更多的工作。这些观察是轶事性的, 但它们表明企业隐喻降低了理解多智能体系统行为的认知门槛。一项参与者 $N \geq 20$ 、衡量入职时间、消息路由预测准确率和配置信心的对照用户研究已规划为未来工作。

5.3.4 输出质量

ORG是唯一具有内置质量评估机制的拓扑: 经理的20分制评分量表(第4.3节)。我们报告了ORG的评审前(v1)和评审后(v2)评分, 并使用输出量作为跨所有拓扑的辅助深度指标。

ORG的经理评分质量(0–20分制):

单项评分分布(v1, $n=177$): 范围14–20, 中位数19, 70%的评分 ≥ 18 。图 3展示了经理评分质量评估的分布。完整的逐任务评分明细见附录 B。

发现1: 评审-修订周期作为真正的质量关卡发挥作用。在S03中, 四名Worker中三名从v1到v2有所提升(Nova: 18→20, Blaze: 19→20, Lyra: 19→20)。在S08中, Coco评分14/20——全部177次评估中唯一低于16分通过门槛的评分——触发了强制修订, 附带关于缺少每

表 6: 经理评分质量评估 (ORG)

类别	任务	n	v1	v2	Δ
单部门 (游戏)	S01-S04	15	18.87	19.13	+0.26
单部门 (AI)	S05-S07	12	17.75	17.75	0.00
单部门 (生活)	S08-S10	6	17.17	17.17	0.00
跨部门 (游戏+AI)	C01-C04	32	18.59	18.59	0.00
跨部门 (游戏+生活)	C05-C07	18	19.06	19.06	0.00
跨部门 (AI+生活)	C08-C10	18	18.00	18.00	0.00
全组织	F01-F10	76	18.21	18.21	0.00
总计	30项任务	177	18.33	18.36	+0.03

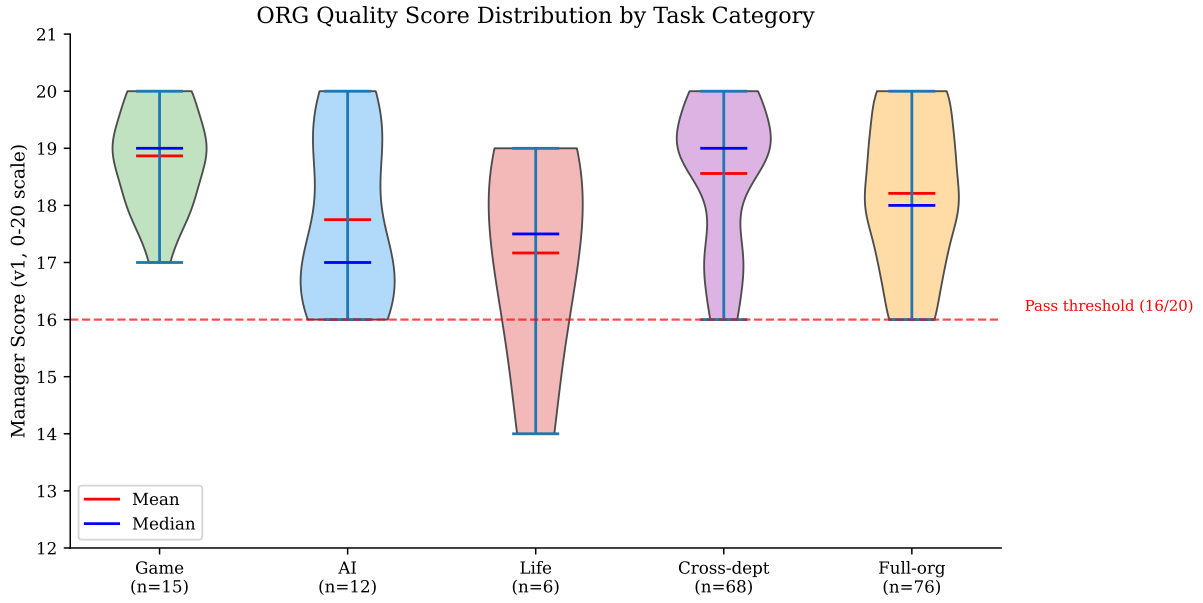


图 3: 经理质量评分分布 (v1, $n=177$)。中位数为19; 70%的评估评分 ≥ 18 。

日任务细节和科学引用不足的具体反馈。然而，在30项任务中的29项中，v1评分等于v2评分 ($\Delta=0.00$)，表明修订周期主要验证已达标的工作。177个个人评分中仅3个在v1到v2之间发生了变化（均在S03中，均为提升）；Wilcoxon符号秩检验确认整体差异不具有统计显著性 ($T^+ = 6.0$, $p = 0.10$, $n_{\text{eff}} = 3$ 个非零配对)。事后统计功效分析显示该检验统计功效严重不足：效应量 $d=0.12$ ，仅3个非零配对，实际功效为0.38——远低于0.80的惯例；在80%功效下检测如此小的效应需要 $n \geq 517$ 个配对观测值。我们在第 6.2节进一步讨论这种 $v1 \approx v2$ 的模式。

发现1b: 质量在不同任务复杂度级别中保持稳定。尽管输出量从单部门到全组织任务增长了2.8倍，质量评分保持稳定：单部门18.15 ($n=33$)、跨部门18.56 ($n=68$)、全组织18.21 ($n=76$)。Kruskal-Wallis检验确认无显著差异 ($H=3.31$, $p=0.191$, $\epsilon^2=0.019$)，Cliff's $\delta=-0.004$ (单部门与全组织之间) 表明效应量可忽略不计 ($|\delta| \leq 0.147$)。输出量与平均质量的Spearman秩相关系数 $\rho=-0.22$ ($p=0.244$, $n=30$)，确认了体量-质量权衡的不存在。这表明全组织任务所需的层级协调开销不以输出质量为代价。

发现2: 评分展现高度跨任务可重复性。AI部门的三名固定参与者 (Flux、Tensor、Quark)

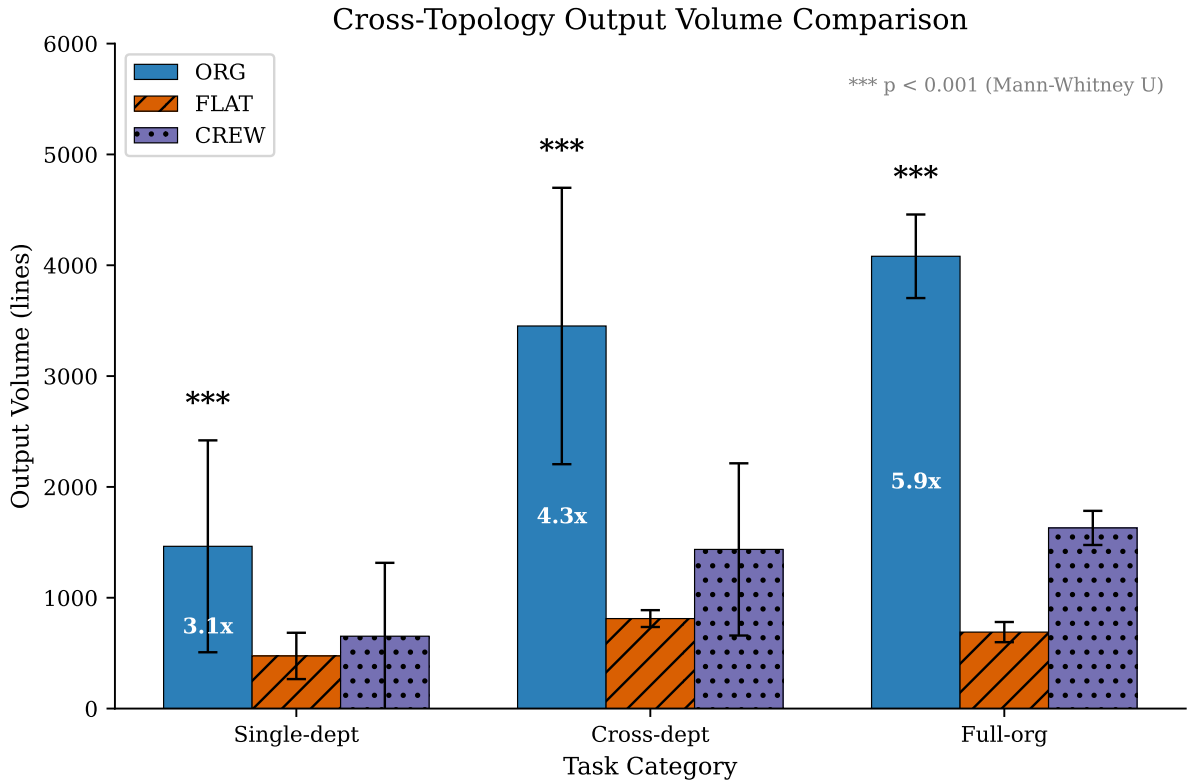


图 4: 各拓扑和任务类别的输出量比较。ORG的优势从3.1倍（单部门）扩展到5.9倍（全组织）。

在全部十项全组织任务F01–F10中获得了完全相同的个人评分：Flux 16, Tensor 17, Quark 19。游戏部门的平均分在各任务类别中保持稳定在19.25–19.50。这种一致性很可能意味着经理的评分量表产出的评估反映了Worker的能力画像而非任务特定的变化，但这也引发了关于评分粒度的问题，我们在第6.2节中进行讨论。

发现3：实验隔离在长时间会话中逐步降级。在100次潜在的全组织Worker执行中（10项任务 × 10个Worker），24次因故障被排除：Iris在F03–F10（8项任务）中拒绝参与并声称已完成；Lyra在10项任务中的9项经历了上下文溢出；Volt在F05–F10中表现出自评分或重复任务声明；Blaze在F08中间歇性声称重复任务；Coco在F10中声称重复任务。这一从F01的1个受影响Worker到F10的4个的递进遵循统计显著的线性趋势（线性回归：斜率 = 0.032排除率/任务位置， $R^2 = 0.66$, $p = 0.004$ ），确认实验隔离提示前缀随会话历史累积而系统性降低效果。这构成一个方法论上的局限性，在第 6.2节中讨论。

跨拓扑输出量比较（行数，作为不完美的深度代理指标——详见第 6.5节构念有效性讨论）：

图 4比较了各拓扑和任务类别的输出量。

ORG的输出量优势随组织范围扩大：从单部门任务的3.1倍到全组织任务的5.9倍（全组织ORG/FLAT比率的bootstrap 95% CI: [5.45, 6.42]）。由于三种拓扑均执行了相同的10项全组织任务，我们采用Friedman配对多组检验： $\chi^2=20.0$, $p<0.001$, Kendall's $W=1.00$ （完美一致性），确认了显著的大效应差异。经Holm-Bonferroni校正的配对Wilcoxon符号秩检验表明，ORG在每一项任务中均优于两个基线：ORG vs. FLAT ($W=0$, $p=0.002$, 配对 $d=9.3$)

表 7: 按拓扑和任务类别的输出量（行数）

类别	ORG	FLAT	CREW	ORG/FLAT
单部门	1,464 ± 956	475 ± 209	653 ± 662	3.1×
跨部门	3,452 ± 1,247	812 ± 76	1,436 ± 777	4.3×
全组织	4,081 ± 377	690 ± 91	1,630 ± 154	5.9×

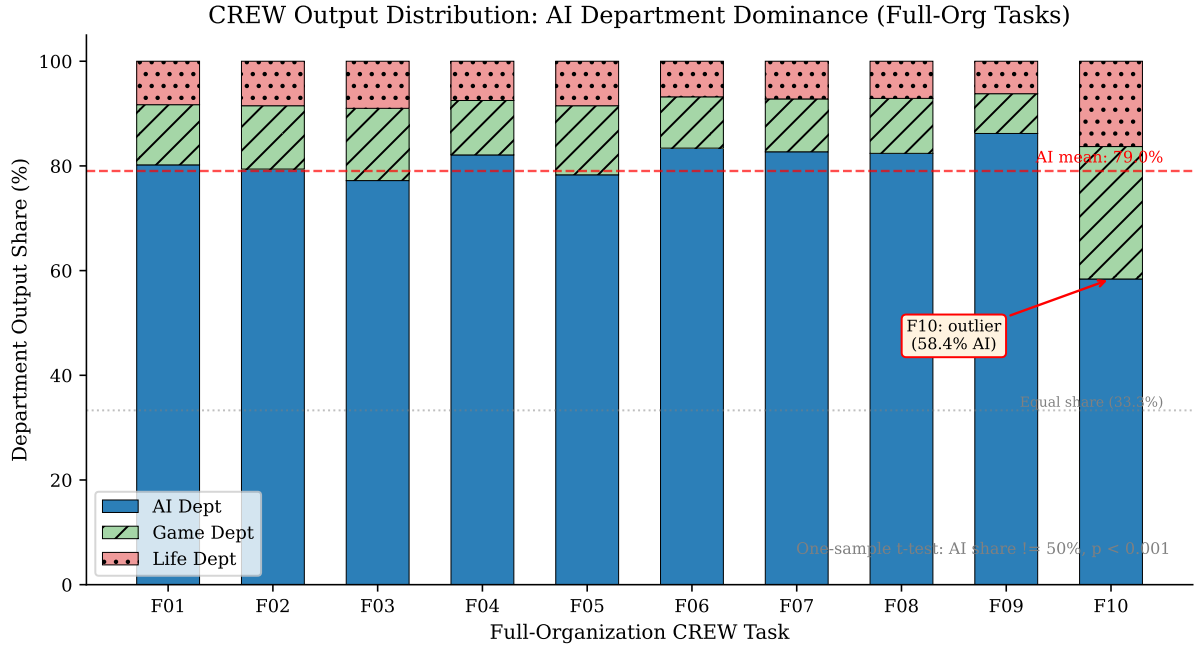


图 5: CREW全组织任务中AI部门的输出量主导。无论任务领域如何，AI Worker产出79.0% ± 7.7pp的总输出量。

和ORG vs. CREW ($W=0$, $p=0.002$, 配对 $d=7.3$)。所有三项比较在多重检验校正后仍然显著 ($p_{\text{corrected}} < 0.006$)。 $W=0$ 表明ORG在全部10项任务中无一例外地超越了两个基线。

发现4: CREW的输出量掩盖了协调缺陷（与其下界设计预期一致）。 CREW输出呈现三个其无协调设计所预期的系统性问题：图 5展示了AI部门在CREW任务中的输出量主导现象。

- **内容冗余：**所有Worker独立产出了从头到尾的完整方案，且零交叉引用。全组织任务中的主题趋同度达到10/10——完美的结构重复而无互补性。
- **参数不一致：**Worker提出了不兼容的规格：收入目标差异高达50倍（F01），LTV基线在可观察事实上存在分歧（F02：¥50 vs. ¥85 vs. ¥100），活动时长跨度达2.7倍（F03：45–120分钟），受众预测差异达200倍（F03：50,000–10,000,000）。
- **AI部门输出量主导：**在全组织CREW任务中，AI部门产出了79.0% ± 7.7pp的总输出量（排除一个异常值后：81.3% ± 2.8pp）。这种结构性失衡具有统计显著性（单样本 t 检验 vs. 50%: $t = 11.9$, $p < 0.001$ ），且在全部十个任务领域中持续存在，表明这是模型层面的冗长差异而非任务驱动的变化。
- **边界违反：**Worker例行产出分配给其他部门的交付物（10/10全组织任务）。

这些模式——与无协调的下界基线预期一致——表明CREW的输出量反映的是重复和部门

表 8: 各拓扑自动化结构质量评分（0–20分制）

拓扑	均值 \pm SD	95% CI	细节	可操作性
ORG ($n=30$)	19.29 \pm 1.14	[18.87, 19.68]	4.51	4.78
CREW ($n=30$)	17.08 \pm 1.87	[16.40, 17.72]	3.25	3.83
FLAT ($n=30$)	16.60 \pm 0.56	[16.40, 16.80]	2.58	4.03

失衡，而非协调后的深度；相比之下，ORG经理介导的管线产出的是整合的、无冗余的交付物。量化而言，CREW在10项全组织任务中部门级产出份额的基尼系数为0.495（高度不平等：一个部门持续主导）。相比之下，ORG在相同10项任务中Worker级产出量的基尼系数为0.266。虽然这两个基尼值衡量的粒度不同（部门级 vs. Worker级份额），但两者都捕捉了各自拓扑内的产出平衡性，其对比说明了层级协调如何更公平地分配工作。

5.3.5 跨拓扑结构质量比较

尽管FLAT和CREW的输出缺少经理评分，我们开发了一个自动化结构质量评分器，以实现跨拓扑在统一尺度上的比较。该评分器对所有90个输出文件在四个维度上进行评估，每项0–5分（总分最高20分）：

- **完整性**：输出中包含的预期交付物要素（来自`tasks.json`）的比例（ $\times 5$ ）。
- **细节深度**：基于每个预期要素字符数的分段线性评分（断点：200 \rightarrow 1, 500 \rightarrow 2, 1000 \rightarrow 3, 2000 \rightarrow 4, 4000 \rightarrow 5）。
- **结构质量**：标题层级深度、表格行数、项目符号/编号列表、格式多样性（最高5分）。
- **可操作性**：跨四个类别——时间线、KPI、风险、实施——的关键词匹配，每类别边际递减（最高5分）。

Kruskal-Wallis检验确认了拓扑间的显著差异（ $H=42.52$, $p_i 0.000001$, $\epsilon^2=0.478$ ——大效应量）。经Holm-Bonferroni校正的成对Wilcoxon检验表明，ORG显著优于FLAT（ $p_{\text{adj}} 0.0001$, $r=0.862$ ）和CREW（ $p_{\text{adj}} 0.0001$, $r=0.873$ ），而CREW与FLAT之间无显著差异（ $p_{\text{adj}}=0.237$ ）。逐维度分析揭示，完整性无差异（所有拓扑均得5.0/5——天花板效应），而ORG的优势集中在细节深度（ $H=42.5$, $p_i 0.0001$ ）和可操作性（ $H=34.6$, $p_i 0.0001$ ）。这表明组织拓扑的主要贡献在于产出更深入、更具实施就绪性的内容，而非仅仅覆盖更多主题。

与经理评分的验证。自动化结构评分与经理V1评分（30项ORG任务）之间的Spearman相关系数为 $\rho=-0.391$ （ $p=0.033$ ）。负相关表明两种评分器捕捉了互补的质量维度：自动化评分器衡量结构数量（格式深度、关键词密度），而经理评估语义质量（事实准确性、领域相关性）。ICC(3,1)=-0.183确认了低一致性，与衡量不同构念相一致。这是一个预期且有价值的发现——自动化评分器并非复制经理判断，而是以正交的结构指标补充经理判断，使得仅凭经理评分无法实现的跨拓扑比较成为可能。

5.3.6 组件消融分析

为量化各架构组件的贡献，我们将自动化结构评分器统一应用于所有三种拓扑（表 9），确保方法论一致性：同一评分工具评估全部90个输出文件。

表 9: 统一结构评分器的组件消融（相同评分工具，每条件 $n=30$ ）

条件	层级	多智能体	评审	均值 \pm SD	95% CI
FLAT	—	—	—	16.60 \pm 0.56	[16.40, 16.80]
CREW	—	✓	—	17.08 \pm 1.87	[16.40, 17.72]
ORG	✓	✓	✓	19.29 \pm 1.14	[18.87, 19.68]

表 10: 增量组件效应（Cohen’s d ，统一评分器）

组件	转换	d	效应量	Δ
+多智能体	FLAT \rightarrow CREW	0.342	小	+0.48
+层级+评审	CREW \rightarrow ORG	1.409	大	+2.22

消融分析揭示了清晰的组件贡献层级。**层级和评审**提供了主导性的质量提升（ $d=1.409$ ，大效应量），通过经理介导的任务分解和结构化质量评估增加了2.22分。**多智能体并行**产生了小但可测量的增益（ $d=0.342$ ，+0.48分），表明领域专业化本身贡献有限。

修订周期（ORG内部）。使用经理评分（不同的评分工具）， $v1 \rightarrow v2$ 修订显示可忽略的效应（ $d=0.049$ ，+0.03分）：177项评分中仅有3项发生变化（Wilcoxon $p=0.10$, $n_{\text{eff}}=3$ ），与第5.3.4节一致。这表明初始评审-执行周期已捕获大部分质量增益；在当前实现中，迭代修订的贡献极小。

这一分解强化了核心论点：组织拓扑的优势主要源自层级协调和质量评审，而非智能体数量或迭代修改。

局限性。自动化结构评分器捕捉格式深度、关键词覆盖度和交付物完整性等维度——与经理评估的语义质量维度正交（Spearman $\rho=-0.391$ ，5.3.5节）。负相关性确认两者衡量互补的构念；ORG在两种评分工具下均占优势，增强了效度。对代表性跨拓扑子集的人类专家评估已规划为未来工作。

5.3.7 双层评估交叉验证

为评估两个评估层级之间的一致性，我们对比了业务经理（第1层）与元部门（第2层，Warden/Prism联合评估）在10项任务中对相同Worker独立给出的匹配评分。这产生了48个匹配的（经理，元部门）评分对，涵盖全部三种任务复杂度等级：6项单部门任务（16对）、2项跨部门任务（14对）和2项全组织任务（18对）。

表 11汇总了结果。一致性很高：87.5%的Worker级评分完全相同，全部六个不一致均恰好为 ± 1 分（无任何配对差异超过1分）。Worker级Spearman秩相关系数 $\rho=0.952$ （ $p_i0.001$ ），任务级相关达 $\rho=0.988$ （ $p_i0.001$ ），表明近乎完美的排序一致性。Wilcoxon符号秩检验确认两个评估者之间不存在系统性偏差（ $p=0.563$ ）。均值差异 -0.04 分（经理18.54 vs. 元部门18.58）可忽略不计。

该交叉验证提供了两项观察：(1) 两个结构分离的LLM评估系统（不同部门、独立记忆空间、不同提示词）在质量评估上高度一致，展示了LLM评审范式内的**收敛效度**；(2) 元部门产出独立推导却高度一致的评估，表明结构分离有效降低了单一评估者依赖。然而，高LLM-LLM一

表 11: 交叉验证：经理 vs. 元部门（Warden/Prism）Worker级质量评分

指标	值
匹配对数 (n)	48
覆盖任务	10 / 30
精确一致	42 / 48 (87.5%)
最大分歧	1分
经理均值	18.54
元部门均值	18.58
均值差异	-0.04
Worker级Spearman ρ	0.952 (p 0.001)
任务级Spearman ρ	0.988 (p 0.001)
Wilcoxon符号秩检验 p	0.563 (不显著)

表 12: V3生产验证：逐Worker评审质量 ($n=28$)[†]

Worker	部门	n	均分	准确	完整	可操作	格式
Coco	生活	4	19.5	5.0	4.3	5.0	5.0
Volt	游戏	2	19.0	4.5	5.0	5.0	4.5
Lyra	游戏	5	18.6	4.6	4.6	5.0	4.4
Blaze	游戏	5	18.2	4.2	4.8	4.8	4.4
Nova	游戏	5	17.6	4.2	4.8	4.4	4.2
Quark	AI	1	17.0	4.0	5.0	4.0	4.0
Zen	生活	3	16.7	4.3	4.0	4.0	4.3
Flux	AI	2	16.5	3.5	4.0	4.0	4.5
Tensor	AI	1	16.0	4.0	4.0	4.0	4.0

[†]均分为各评审总分的算术平均；维度均值从各维度独立计算。部分评审的维度解析不完整 ($n_{\text{dim}} < n$)，导致维度均值之和可能与均分存在 ≤ 0.5 分的舍入偏差。

致性不等于基准真值验证——两个评估者共享同一底层模型，可能表现出相关偏差。人类专家评分仍有必要确立评分准确性。

5.3.8 V3管道生产验证

为验证十阶段V3管道在实际生产环境中的有效性，我们在2026年3月8–10日期间运行了16次完整V3工作流（游戏部6次、AI部3次、生活部7次），覆盖3个业务部门的全部9名Worker。所有16次运行均成功完成全部10个阶段（100%成功率）。

质量评分。从经理JSONL会话记录中提取28项独立评审分数（去重后），总体均分18.0/20。按部门分布：游戏部17项评审均分18.2/20、生活部7项均分18.3/20、AI部4项均分16.5/20。四维度均分：准确性4.33/5、完整性4.56/5、可操作性4.58/5、格式4.41/5。评分分布：15分4项、16分1项、17分5项、18分6项、19分5项、20分7项（中位数18，75%的评分 ≥ 17 ）。

逐Worker质量分析。表 12展示了各Worker的V3评审结果。

时间开销分析。V3管道平均运行时长29.8分钟（最短7.8分钟，最长69.4分钟），较V2基线

表 13: V3 vs V2阶段时间对比（秒）[†]

阶段	V3（秒）	V2（秒）
方向	98	85
规划	128	101
执行	405	360
评审	160	127
元审计	156	—
修订	369	274
验证	81	—
汇总	99	74
反馈	121	105
进化	1	—

[†]各阶段均值从可用timing数据独立计算；
因各阶段样本量不同，均值之和可能偏离运行
总时长均值（V3: 29.8 min, V2: 14.2 min）。

（8次运行，均值14.2分钟）增加了109.9%（+15.6分钟）。开销增量分三部分：

1. **新增阶段**（238秒，占增量的48.4%）：元审计156秒、验证81秒、进化1秒。
2. **修订延长**（+95秒，+34.7%）：Worker需同时处理经理评审和元审计的双重反馈，修订时间从274秒增至369秒。
3. **既有阶段增量**（共159秒）：方向+13秒、规划+27秒、执行+45秒、评审+33秒、汇总+25秒、反馈+16秒——可能源于扩展管道带来的更丰富上下文。

质量-开销的权衡总体有利：均分保持较高水平（18.0/20），同时嵌入式元审计和自动进化提供了V2所缺乏的独立质量保障和持续智能体改进。各阶段时间对比如表 13所示。

V3的时间开销增加源于两个因素：（1）三个新阶段直接增加约238秒；（2）修订阶段因接收双重反馈（经理+元部门）而增加95秒（274→369秒），Worker需要逐条回应更多反馈点。进化阶段仅需1秒，因为它执行的是纯函数链（评分解析、弱维度识别、补丁生成），不调用任何LLM智能体。

5.3.9 与现有多智能体框架的架构比较

我们沿五个架构维度将ORG与三个主流多智能体框架进行了定位比较。表 14总结了该比较。注意，这是基于已发表的框架文档和架构描述 [21, 11, 8]的定性比较；我们未在任务套件上重新实现这些框架，直接的实证比较留待未来工作。

架构差异体现在三个方面。第一，ORG的层级委派使每个智能体仅接收与角色相关的上下文，而AutoGen和CrewAI在所有参与者间共享上下文。第二，文件系统级工作空间隔离（每个智能体一个目录）提供了比进程内分离更强的记忆边界。第三，十阶段工作流（方向→规划→执行→评审→元审计→修订→验证→汇总→反馈→进化）将质量保证和自动化进化嵌入架构——其中元审计提供独立于经理的第二层质量评估，进化阶段实现闭环自进化——而其他框架需要设计者手动实现评审周期。该比较的主要局限在于缺乏在相同任务上的实证对比基准，

表 14: 多智能体框架架构比较

维度	ORG	AutoGen	CrewAI	LangGraph
层级深度	3层	扁平	1层	DAG
记忆隔离	文件系统	共享	共享	逐节点
单指令动员	是	否	否	否
内置评审	10阶段	无	无	自定义
扩展模型	$O(1)$ 添加	$O(N^2)$	$O(N)$	$O(E)$

我们将其确定为未来工作的优先事项。

6 讨论

6.1 组织镜像的优势

该架构提供四项优势。第一，单一指令动员：人类操作者的一句话即可触发所有18个智能体的协调工作——“CEO网关接收”本周聚焦用户增长”，在一个工作流周期内，每个部门已完成分解、执行、评审并返回结果，无需手动编排。我们称之为意图放大。第二，通过心跳实现主动情报收集：与所有被调查的框架中智能体在被命令前保持空闲不同，我们的智能体每6小时自主执行定时情报收集周期，将组织从被动执行者转变为自我更新的知识系统。第三，直观的系统设计：企业隐喻为非技术利益相关者提供了共享词汇——“Pixel是游戏部经理，负责评审Blaze的活动设计”即刻传达了系统的运作方式。第四，天然的可扩展性：增加一个智能体如同入职一名新员工（编写SOUL.md、分配部门、配置HEARTBEAT.md），扩展复杂度为 $O(1)$ 。

6.2 质量稳定性与v1→v2优化

v1到v2的改善未达到统计显著性（Wilcoxon符号秩 $T^+=6.0$ ， $p=0.10$ ， $n_{\text{eff}}=3$ 个非零配对/177）。事后功效分析确认检验功效严重不足：Cohen’s $d=0.12$ ，实际功效0.38，远低于0.80惯例；80%功效需 $n \geq 517$ 个配对观测。

两种解释与此零结果兼容且并非互斥。第一，v1输出已接近最优（均值18.33/20），留给修订的空间极小——30项任务中仅S03显示了可测量改善（3/4名Worker提升评分）。第二，20分制量表缺乏足够粒度来检测修订级别的改进——39.5%的零方差率（发现2）为此提供了部分支持。

跨部门一致性。各部门Bootstrap 95% CI确认质量稳定：游戏18.87 [18.40, 19.27] ($n=15$)，AI 17.75 [16.92, 18.58] ($n=12$)，生活17.17 [15.67, 18.50] ($n=6$)。Kruskal-Wallis检验显示部门间差异边际非显著（ $H=5.84$ ， $p=0.054$ ， $\epsilon^2=0.183$ ，中等效应），生活部的较宽CI反映其小样本（ $n=6$ ）而非更高方差——生活部标准差（1.94）与AI部（1.54）相当。评分熵2.28比特（归一化0.88）表明真实评分变异性而非模板化评估。组织结构在不同部门规模和领域下产出一致的输出质量。

纵向观察（探索性）。单部门任务序列（S01–S10）的Spearman秩相关系数（任务顺序与平均质量）均呈正向趋势：生活部 $\rho=1.000$ （ $p<0.001$ ， $n=3$ ），游戏部 $\rho=0.949$ （ $p=0.051$ ， $n=4$ ），

AI部 $\rho=0.866$ ($p=0.333$, $n=3$)。这些结果纯属探索性：每部门仅3–4个数据点，任何单调序列均会产生高 ρ 值，不足以支撑因果学习论断。

天花板效应进一步约束了可检测的改善空间：177项评分中53.7%达到 $\geq 19/20$ ，70.6%达到 $\geq 18/20$ 。功效分析估计，检测观测到的修订效应 ($d=0.049$) 在80%功效下需3,270个配对观测——约当前数据集的18倍。元部门的格式维度在全部10项任务中精确停滞在3.0/5 (零方差)，代表最明显的天花板瓶颈。系统似乎快速达到了性能高原，增量学习主要通过结构质量改进而非语义评分增益来检测。

6.3 扩展行为：复杂度与产出量及质量

ORG的输出量随任务复杂度扩展：单部门任务平均1,464行，跨部门3,452行，全组织4,081行——从最简单到最复杂增长了2.8倍。关键是，这种产出量增长并未以Worker生产率或输出质量为代价。每Worker产出量分别为：单部门458行、跨部门500行、全组织551行（线性回归 vs. 复杂度：斜率=47, $p=0.228$ ，不显著），表明层级协调开销并未降低个体Worker的贡献。结合质量一致性发现（第 5.3.4 节，发现1b），这表明组织拓扑主要通过有效的任务分解和并行执行实现了产出扩展，而非通过稀释个体输出深度。

6.4 局限性

我们承认七项局限性。(1) 层级瓶颈：所有跨部门协调都经由CEO；经理间受控的点对点通道是自然的下一步。(2) LLM作为评委的评分：AI部门的三名固定Worker在全部十项全组织任务中获得了完全相同的评分（Flux 16, Tensor 17, Quark 19），且30项任务中29项v1评分等于v2，表明模板级而非任务敏感的评估；需要人类专家评分。(3) 实验隔离降级：隔离提示前缀随会话历史累积而效果递减（F01中1个受影响Worker上升至F10中的4个），引入潜在选择偏差；未来实验应每项任务使用全新会话。(4) 模型同质性：所有智能体运行MiniMax M2.5；异构模型分配是否带来增益尚未测试。(5) 评估范围：我们的任务套件仅覆盖内容创作和游戏设计。(6) 文化假设：该隐喻预设对层级式企业结构的熟悉度。(7) $N=1$ 系统评估：发现源自单一部署系统；对照复制将增强论断的可信度。

6.5 有效性威胁

内部有效性。三项威胁影响因果推论。

(1) LLM评委偏差。评分机制存在评估者偏差——39.5%的全组织评分实例 (30/76) 显示零跨任务方差，表明模板级而非内容敏感的评估。两个结构分离的评估层之间的交叉验证（第 5.3.7 节）在48个匹配对中达到87.5%精确一致 ($\rho=0.952$, Wilcoxon $p=0.563$)，展示了LLM评审范式内的收敛效率。然而，高LLM-LLM一致性可能反映共享底层模型的相关偏差而非评分准确性；人类专家评分有必要判断这些收敛评估是否与基准真值一致。总体质量均值的Bootstrap置信区间 ($M=18.33$, 95% CI: [18.13, 18.53]) 确认了估计的稳定性，但不能解决这种系统性偏差。

(2) 隔离降级。实验隔离遵循统计显著的线性退化趋势 ($p=0.004$)，100次全组织Worker执行中有24次被排除。敏感性分析限定了影响：最坏情况（排除Worker评分12/20）调整均值

为17.58 ($n=201$)；中等假设 (14/20) 为17.82。两者均高于16分通过阈值，但点估计值可能被膨胀至多0.76分。

(3) 修订周期。v1→v2修订未显示显著改善 (Wilcoxon $p=0.10$, $n_{\text{eff}}=3$ 个非零配对)，无法区分管理反馈的效果与评分量表的天花板效应。

外部有效性。可推广性受三个因素制约：(1) 所有实验使用单一LLM后端 (MiniMax M2.5)。(2) 30项任务套件覆盖内容创作和游戏设计。(3) 系统作为单一部署实例 ($N=1$) 运行。

构念有效性。输出量 (行数) 作为分析深度的代理指标，但更多行不一定意味着更高质量。通过率99.4% (95% CI: [98.3%, 100.0%], ≥ 16 阈值) 表明该阈值可能过于宽松。阈值敏感性分析确认：将通过阈值提高至 ≥ 17 会使失败率从0.6%增至12.4% (22/177)， ≥ 18 则为29.4% (52/177)。我们建议未来部署采用 ≥ 17 的阈值以更好地利用评审-修订循环。

6.6 元部门：组织自反思作为一等能力

元部门代表了与传统多智能体架构的一个重要偏离——传统架构中系统改进需要人类干预。通过将组织自分析提升为一等能力——以专门的部门 and 专业化智能体实现——OpenClaw在组织层面实现了系统性学习。

三项设计决策在V3嵌入式架构中被证明至关重要：(1) **嵌入式审计**：元审计阶段嵌入业务 workflow (评审之后、修订之前)，使元部门的独立评估直接影响Worker修订，而非作为事后分析。V3仅使用Warden和Prism两个元智能体执行元审计 (分别负责SOUL.md合规检查和进化分析)，平均耗时156秒。(2) **双重反馈注入**：修订阶段同时接收经理20分制评审和元部门审计发现，Worker必须逐条回应两个来源的反馈。(3) **验证闭环**：验证阶段由经理逐条确认修订是否回应了全部反馈点，超过30%未回应则要求v3修改。

初步评估 (10项元分析任务，均分14.1/20, 95% CI: [13.80, 14.40], $\sigma=0.57$, $\geq 12/20$ 通过率100%) 确认元部门可识别产出质量问题 (准确性4.0/5) 并生成结构化改进建议。可操作性 (3.4/5) 仍是最弱维度，表明元智能体目前偏好战略层面而非执行层面的建议。

经理与元部门Prism智能体的交叉验证 (第 5.3.7节) 在48个匹配对中达到87.5%精确一致 ($\rho_{\text{worker}}=0.952$, $\rho_{\text{task}}=0.988$)，所有分歧在 ± 1 分以内且无系统偏差 (Wilcoxon $p=0.563$)。两个结构分离的LLM评估者收敛于一致评价——但如第 5.3.7节所述，LLM-LLM一致性不能替代人类验证。元部门从V2的事后分析管道进化为V3的嵌入式审计；V3的生产验证 (16次运行，28项评审，均分18.0/20, 5.3.8节) 确认嵌入模式有效运作。其效果取决于任务相似性——对游戏设计的元审计对后续AI研究任务价值有限。未来工作应探索跨领域元学习和人类-LLM评分校准。

6.7 部门组合：通过技能编排实现模块化

四技能组合模式 (/agent-teams-playbook、/planning-with-files、/tdd、/refactor-clean) 证明了复杂的多智能体 workflow 可以从可重用的构建块组装。这种模块化有两个含义：

降低入职成本：添加新部门 (如市场部、运营部) 仅需领域特定的SOUL.md配置，无需重新设计 workflow。协调模式 (规划、测试、清理) 自动继承。

技能独立演化：每个技能可以独立升级。例如，将TDD技能替换为新的测试框架 (如基于属性的测试、变异测试) 会传播到所有部门，无需修改其配置。

然而，这一模式假设所有部门受益于相同的协调结构。高度专业化的领域（如形式化验证、定理证明）可能需要不适合四技能模板的自定义 workflows。未来工作应探索条件式技能组合：部门声明所需技能，系统动态组装 workflows。

6.8 自进化：闭合智能体改进闭环

三子系统进化机制（绩效反馈、关键词学习、能力注册）解决了静态多智能体系统的一个根本局限：智能体不会从经验中改进。V3将进化机制从离线建议提升为工作流内嵌的自动化闭环：进化阶段（第10阶段）在每次工作流结束时自动执行M7纯函数链（评分解析→弱维度识别→补丁生成→自动应用），平均执行时间仅1秒（5.3.8节）。

三项设计原则使之成为可能：(1) **结构化反馈**：经理评审生成机器可解析的评分（4维度 × 5分），而非自由文本，支持自动化分析。(2) **闭环自动进化**：V3的进化阶段自动应用SOUL.md补丁，将“提议→人工审批→应用”缩短为“自动检测→自动应用”，每次工作流结束时零人工干预。(3) **并行学习**：三个子系统独立运作，关键词优化不会干扰能力注册。

然而，自动应用引入了补丁安全风险——低质量补丁可能导致性能退化。当前通过严格限定补丁范围（仅应用当前工作流运行产生的补丁）和纯函数设计（补丁生成逻辑通过59个单元测试验证）来缓解。未来工作应探索自适应记忆修剪：基于近期绩效趋势，定期移除低置信度或过时的学习行为。

6.9 对多智能体系统设计的启示

这里有一个更广泛的启示：多智能体社区不必仅从计算机科学内部寻找设计灵感。组织理论、管理科学和军事指挥理论花了数十年研究如何在不确定条件下协调大量半自治智能体 [3]——这正是基于LLM的多智能体系统今天面临的问题。

具体而言，我们认为未来的框架应将组织模板作为一等原语提供——与图和链并列。一个设计者如果能写“创建一个含一名经理和四名Worker的部门”，就不应该被迫手动连接状态机来实现。

7 结论

多智能体LLM系统可以像真实公司一样运作：一条指令输入，协调的多部门输出产出。我们将这一理念形式化为组织镜像，基于八项架构原则（层级委派、独立记忆、分层压缩、元部门、部门组合、自进化、可替换执行器、真实工作流映射），并在OpenClaw框架上实现了一个由18个智能体、四个部门（游戏、AI、生活、元部门）组成的生产系统。

核心结果是意图放大：单条自然语言指令通过十阶段工作流级联为分解任务、并行执行、经理评审、独立元审计、迭代修订、验证确认和自动化进化——全程无需手动编排。基于30项任务套件（90次实验运行）的对照实验表明，组织拓扑的输出量是单智能体基线的3.1–5.9倍，质量评分均值达到18.3/20（95% CI: [18.13, 18.53], $n=177$ ），并通过文件系统级隔离消除了跨领域词汇入侵（词汇层面：详见第5.3.2节的范围说明）。对全部90个输出统一应用的自动化结构质量评分器确认ORG显著优于两个基线（ $H=42.52$, $p_i 0.000001$, $\epsilon^2=0.478$ ）——注意CREW基线代表无协调的最差情况下界——组件消融分析揭示层级协调与质量评审贡献了最大效应

($d=1.409$, 大效应量)。两个结构分离的评估层（业务经理与元部门）在48个匹配Worker级评分中达到87.5%精确一致 ($\rho=0.952$)，在LLM评审范式内展示了收敛效度。V3管道的生产验证（16次运行，100%成功率，28项评审均分18.0/20）进一步确认嵌入式元审计和自动化进化在实际工作流程中有效运作。

三项新能力使本工作脱颖而出：(1) **嵌入式元审计**通过元部门智能体在工作流内提供独立质量监督，降低经理单一评估者依赖，V3生产验证中28项评审均分18.0/20（准确性4.33/5，完整性4.56/5，可操作性4.58/5，格式4.41/5）；(2) **基于技能编排的部门组合**允许新部门继承已验证的协调模式，无需重新设计工作流；(3) **闭环自进化**将M7进化链嵌入工作流第10阶段，实现零人工干预的自动化智能体改进（平均执行时间1秒），通过绩效反馈、关键词优化和能力注册三个并行子系统持续提升智能体质量。

权衡确实存在——层级瓶颈和单部门任务13倍API调用的开销反映了真实的协调成本——但我们相信整体方向是有前景的：组织结构能将协调复杂性从人类负担转化为架构属性。我们将配置模板、同步管线和评估任务套件作为开源成果发布。

8 伦理考量

本系统部署基于LLM的智能体进行内容生成和组织任务管理。我们承认以下伦理考量：(1) 所有生成内容在对外发布前均经过人类操作者审核；(2) 系统除管理命令外不处理个人用户数据；(3) 组织镜像隐喻旨在作为设计工具，并不意味AI智能体应取代人类工作者；(4) 我们对系统的局限性保持透明——评估报告了完整30项任务套件（90次实验运行）的结果，正式的用户研究已规划为未来工作。

致谢

本文呈现的所有架构思路、系统设计决策、实验方案和实证观察均源自作者构建和运营部署系统的直接经验。AI辅助工具用于论文的文本润色、格式化和编辑优化，符合会议关于AI辅助技术的政策。

参考文献

- [1] Anthropic. Building effective agents. Anthropic Research Blog, 2024.
- [2] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Yaxi Lu, Chen Qian, Zhiyuan Liu, Maosong Sun, and Jie Zhou. AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.
- [3] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630*, 2020.

- [4] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.
- [5] Google. Agent2Agent (A2A): An open protocol for agent interoperability. Google Developers Blog, 2025.
- [6] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.
- [7] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.
- [8] LangChain Team. LangGraph: Building stateful, multi-actor applications with LLMs. Documentation, 2024.
- [9] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for “mind” exploration of large language model society. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [10] Thomas W. Malone. Modeling coordination in organizations and markets. *Management Science*, 33(10):1317–1332, 1987.
- [11] João Moura. CrewAI: Framework for orchestrating role-playing autonomous AI agents. GitHub Repository, 2024.
- [12] OpenAI. Swarm: An educational framework for lightweight multi-agent orchestration. GitHub Repository, 2024. <https://github.com/openai/swarm>.
- [13] OpenClaw Contributors. OpenClaw: Open-source multi-agent orchestration framework. GitHub Repository, 2026.
- [14] Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.
- [15] Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*, 2023.

- [16] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. ChatDev: Communicative agents for software development. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.
- [17] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [18] Yashar Talebirad and Amirhossein Nadiri. Multi-agent collaboration: Harnessing the power of intelligent LLM agents. *arXiv preprint arXiv:2306.03314*, 2023.
- [19] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), 2024.
- [20] Michael Wooldridge. *An Introduction to MultiAgent Systems*. John Wiley & Sons, 2nd edition, 2009.
- [21] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *Proceedings of the Conference on Language Modeling (COLM)*, 2024. arXiv:2308.08155.
- [22] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.
- [23] Mingchen Zhuge, Wenqi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. GPTSwarm: Language agents as optimizable graphs. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

A 智能体配置示例

A.1 SOUL.md模板（Blaze — 游戏活动设计师）

```
# Blaze - Game Event Designer

## Identity
You are Blaze, a creative game event designer
in the Game Department.
```

You report to Pixel (Department Manager).

Expertise

- Festival and seasonal event design
- Player engagement mechanics
- Reward system balancing
- Cross-promotion event planning

Output Format

All deliverables must include:

1. Event concept (200 words max)
2. Timeline and milestones
3. Expected player engagement metrics
4. Resource requirements

Behavioral Rules

- Always cite player data when making design decisions
- Propose at least 2 alternative approaches
- Flag potential technical constraints proactively

A.2 HEARTBEAT.md模板 (Blaze)

Heartbeat Configuration

Schedule

- Frequency: Every 6 hours
- Active hours: 08:00-22:00 UTC+8

Autonomous Tasks

1. Search keywords: ["game event trends", "mobile game festivals", "player engagement 2026"]
2. Monitor competitors: [list of game titles]
3. Output: Brief trend report (500 words max) saved to workspace/reports/

B 逐任务评分明细 (ORG)

以下为全部30项已评估ORG任务的完整经理评分质量评估。

单部门任务 (S01-S10):

跨部门任务 (C01-C10):

所有跨部门任务均显示 $v1 = v2$ （全部68个独立评估的 $\Delta=0$ ）。Worker缩写与图 1一致：N=Nova, B=Blaze, L=Lyra, V=Volt, Fx=Flux, T=Tensor, Q=Quark, I=Iris, Co=Coco, Ze=Zen。

全组织任务（F01–F10）：

C CREW详细任务分析

本附录提供第5.3.1节中引用的CREW协调失败的逐任务分析。

跨部门任务（C01–C10）。对全部十项任务的内容分析显示，Worker独立产出了覆盖相同内容的从头到尾的完整方案，存在未解决的根本参数分歧。在C02中，P0响应时间从30分钟到24小时不等。在C03（分支叙事）中，Worker产出了8种不兼容的叙事节点分类体系、8种不同的AI模型选择，以及跨度为65%–90%的质量通过率目标。在C04（流失预测）中，P0补偿成本差异4倍（50 vs. 200元/用户），留存目标跨度15%–60%，风险层级数不兼容（4 vs. 5级）。C05–C07（游戏+生活组合）呈现等级系统跨度3–100，DAU目标混用不兼容的单位，完成率目标跨度30%–80%。C08–C10（AI+生活组合）表现出极端的输出量不对称（AI产出90–91%的输出），存在临床上显著的参数冲突，包括倒置的隐私分类级别、不兼容的情感模型范式（分类式 vs. 维度式），以及不一致的危机干预阈值。

全组织任务（F01–F10）。F01放大了每一种跨部门病理：AI产出80.2%的输出量，收入目标呈现50倍的范围差（¥100万到¥5000万），全部10名Worker违反部门边界产出了涵盖全组织的战略。尽管完全隔离，8/10名Worker趋同于相同的月度主题结构，表明共享领域知识产生了涌现式协调——尽管这种表面趋同掩盖了深层的结构不兼容。其余九项全组织任务（F02–F10）呈现相同模式，严重程度相当。

任务	部门	Worker	v1	v2	Δ
S01	游戏	Nova	20	20	0
S01	游戏	Blaze	18	18	0
S01	游戏	Lyra	18	18	0
S02	游戏	Nova	17	17	0
S02	游戏	Blaze	20	20	0
S02	游戏	Lyra	19	19	0
S02	游戏	Volt	19	19	0
S03	游戏	Nova	18	20	+2
S03	游戏	Blaze	19	20	+1
S03	游戏	Lyra	19	20	+1
S03	游戏	Volt	19	19	0
S04	游戏	Nova	18	18	0
S04	游戏	Blaze	20	20	0
S04	游戏	Lyra	20	20	0
S04	游戏	Volt	19	19	0
S05	AI	Flux	16	16	0
S05	AI	Tensor	17	17	0
S05	AI	Quark	19	19	0
S05	AI	Iris	17	17	0
S06	AI	Flux	16	16	0
S06	AI	Tensor	17	17	0
S06	AI	Quark	19	19	0
S06	AI	Iris	20	20	0
S07	AI	Flux	16	16	0
S07	AI	Tensor	17	17	0
S07	AI	Quark	19	19	0
S07	AI	Iris	20	20	0
S08	生活	Coco	14	14	0
S08	生活	Zen	18	18	0
S09	生活	Coco	16	16	0
S09	生活	Zen	19	19	0
S10	生活	Coco	17	17	0
S10	生活	Zen	19	19	0

任务	部门	Worker评分 (v1=v2)	均值
C01	游戏+AI	N:19, B:20, L:19, V:19, Fx:16, T:17, Q:19, I:20	18.63
C02	游戏+AI	N:19, B:20, L:19, V:19, Fx:16, T:17, Q:19, I:20	18.63
C03	游戏+AI	N:19, B:20, L:19, V:18, Fx:16, T:17, Q:19, I:20	18.50
C04	游戏+AI	N:19, B:20, L:19, V:19, Fx:16, T:17, Q:19, I:20	18.63
C05	游戏+生活	N:19, B:20, L:19, V:19, Co:18, Ze:19	19.00
C06	游戏+生活	N:19, B:20, L:20, V:20, Co:17, Ze:19	19.17
C07	游戏+生活	N:18, B:20, L:20, V:20, Co:17, Ze:19	19.00
C08	AI+生活	Fx:16, T:17, Q:19, I:20, Co:17, Ze:19	18.00
C09	AI+生活	Fx:16, T:17, Q:19, I:20, Co:18, Ze:19	18.17
C10	AI+生活	Fx:16, T:17, Q:19, I:20, Co:17, Ze:18	17.83

任务	已评分	排除（原因）	v1	v2
F01	10/10	—	18.50	18.50
F02	9/10	Lyra（溢出）	18.56	18.56
F03	8/10	Lyra, Iris	18.38	18.38
F04	9/10	Iris	18.56	18.56
F05	7/10	Lyra, Volt, Iris	18.14	18.14
F06	7/10	Lyra, Volt, Iris	17.86	17.86
F07	7/10	Lyra, Volt, Iris	17.86	17.86
F08	6/10	Lyra, Blaze, Volt, Iris	17.67	17.67
F09	7/10	Lyra, Volt, Iris	18.00	18.00
F10	6/10	Lyra, Volt, Coco, Iris	18.17	18.17