01 / 核心原则

把高能力，用在关键判断上。

GPT × Codex × 工作室工具

任务分级、模型路由、资产复用与生产复盘

> 不是“每件事都用最强模型”，
> 而是“每件事都找到可靠、够用的路线”。

先选方法

检索 · 脚本 · 软件 · 生成

再选模型

能力 · 推理 · 验收 · 成本

按需升级

证据 · 风险 · 边界 · 交接

团队起点

常规多步骤制作，先用已验证可用的 **GPT-5.6 Sol / Medium**。轻任务再向下分流，关键难点才局部升级。确定性重复操作优先交给脚本和软件。

这份指南解决什么

把“任务该用什么模型、多少推理、哪条工具链、什么时候停止”变成可复用规则。配套 Skill、资源表、启动指令，以及不调用模型的路由建议与报告脚本。

以“小猪角色资产库 → 动作映射 → Blender 骨骼控制”流程为实践起点。案例来自本次工作记录与截图，不是独立模型性能或成本测试。

02 / 全员速查

不是一个开关，而是五条路线。

下面是工作室试行默认值，不是模型能力上限。高风险任务无需从低档逐级试错。

级别 / 判断	任务示例	建议起点	交付边界
R0 · 规则已固定	已有脚本的改名、汇总、固定参数渲染	**现成脚本 / 软件** 不逐项调用模型	输入、配方或测试变化时才重新审查
R1 · 清晰轻任务	短文改写、格式转换、固定字段提取	已批准 Luna / Terra 低档；GPT 用实际可用快速档	先抽查，错误可定位、可修复
R2 · 标准制作	brief 拆解、局部开发、资产映射、Blender 既定链路	**Sol / Medium**	先小样，达标后批量；不重做已确认部分
R3 · 关键难点	跨工具问题、架构边界、持续身份/动作漂移	Sol / High；必要时申请 Astra / Medium 或 High	只诊断一个难点，输出规则后降回执行档
R4 · 高风险审查	不可逆生产写入、重大交付判断、敏感数据流向	合适高能力模型 **+ 证据 + 人工审批**	模型通过不等于获准执行

两个容易弄反的判断

任务规模大，不等于推理难。一千张已有规范的文件整理，可能比一处骨骼映射错误更适合脚本。

便宜模型，不等于总成本低。当返工代价很大、问题高度不确定时，一次有边界的高能力诊断可能更划算。

公开目录包含 Sol / Terra / Luna 与 Astra；支持的推理档、客户端开放情况和团队权限需分别核验。界面 “Instant / Medium / High / Extra High / Pro” 不应直接当成所有 API 的统一参数。 [S01] [S02] [S03] [S13]

03 / 接任务的流程

先验收，再路由；先小样，再批量。

路由本身也有成本。简单任务直接处理；复杂任务只在开始、明显失败或风险改变时重新路由。

1. **把交付说清楚**
   输出是什么？依据哪份资料？哪些不能改？什么才算通过？已有答案不重复问。
2. **先找已完成的部分**
   检索官方参考、现有资产、历史通过版本和脚本；不要先从聊天中自由重建整个项目。
3. **选最小可靠路线**
   先决定检索 / 脚本 / 软件 / 生成，再选模型和推理。真正未知的部分才进入推理。
4. **验证代表样本**
   覆盖正常、困难和边界样本。图像低清看方向，最终细节仍需原分辨率验收。
5. **按证据修复**
   初次执行后，默认最多两次针对性修复。每次必须有新诊断；同一无效调用不循环。
6. **交付即收尾**
   保存结果、验收证据、版本、HANDOFF 与用量。通过就停止，不自动扩展成更大的项目。

复杂任务只要四行路由卡

```
推荐：Sol / medium；实际运行：未核验（等待运行记录）。
方法：动作库检索 → 骨骼参数 → 低清预览 → 验收后批量。
小样：3 个代表姿态；身份、动作语义、接触与可重复运行通过。
边界：只做批准范围；初次 + 最多 2 次有证据修复，超出申请诊断。
```

四行是路由结论，不要求展开模型的私有推理。这里的修复次数和小样数量是试行建议，不是平台硬限制。

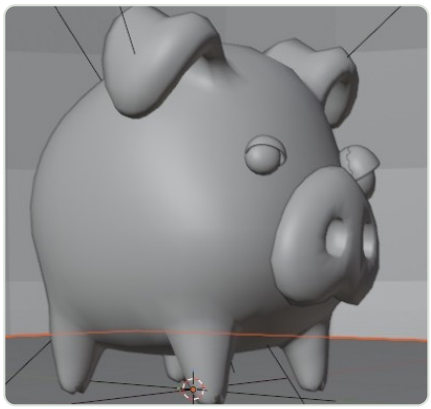
04 / 核心案例

小猪项目的价值，不是“换了一个模型”。

而是把反复猜测角色和动作，改造成有资产、有控制关系、有验收的制作流程。

已有角色 / 画面参考

动作与姿态样本



Blender 代理与结构控制

已知案例事实

实测反馈：使用“5.6 Sol”中推理度，完成项目设定与工作流打通；把既有小猪形象、动作建库映射，通过骨骼控制 Blender，提高一致性控制。没有使用所称的“6 模型”。

以上来自工作记录；截图能看到角色参考与 Blender 代理，不能据此推算真实用量、成本或自动化覆盖率。

官方参考 / 资产库 → 动作映射 → 骨骼 / 相机控制 → 预览与验收

应沉淀什么	建议最小字段 / 检查
角色资产索引	角色 ID、批准参考、版本、鼻耳眼与体型约束；引用原始来源。
动作映射表	动作 ID、来源帧、骨骼/姿态参数、关键接触和遮挡；语义不只写“开心”。
执行与回归	脚本版本、场景版本、相机/渲染设置、输出路径、通过样本与失败证据。
双层验收	“像同一只猪”与“动作表意正确”分开判断；还要检查穿插、落地、动作幅度和可重复运行。

可复用的经验：让模型建立规则和解决新难点，让数据库、脚本与软件负责稳定重复。这个案例支持一种工作方法，不证明所有复杂项目都能用同一档位完成。

05 / GPT 的日常使用

聊天负责判断，不要承担所有执行。

建议按工作性质分流，而不是强行给所有同事安排同一种工具。

优先在 GPT 中完成

brief 理解、沟通文案、报价说明、排期讨论、方案取舍与商业判断。

先给明确问题和相关资料；默认要求直接可用的答案，不要求每次输出长篇过程。

优先转给 Codex

软件开发、批量文件、表格规则校验、资产索引、软件控制、内部系统和重复流程。

把“每次聊着做”改为“写一次脚本，再按版本运行与检查”。

低效委托	更明确的委托
“帮我把这个项目全面优化一下。”	“只审查第 3 段动作连续性，其他镜头、影调、时长不改；输出需要修改的三处。”
“把所有资料读一遍再开始。”	“先看 brief 与资产索引，定位相关章节；缺证据时才读取原文。”
“再给我一版更好的。”	“身份和影调保持，只增加动作幅度；通过条件是侧向姿态仍能读出指定动作。”

可直接放进新对话的使用约定

请先判断这件事是否需要当前模型和推理档。简单任务直接给结果；复杂任务用一句话推荐模型与推理，并说明关键理由。优先复用我提供的资料，不无故扩大调研、不重复重写已确认内容。资料不足先定位缺项，不能把猜测写成事实。只处理我指定的修改范围。你无法核实或切换实际模型时，请明确说是建议，不要声称已切换。涉及批量文件、软件操作或可重复流程时，建议转为 Codex + 脚本，而不是逐条聊天执行。

重要区别：提示词可以让模型给建议，却不是设置控件。普通 ChatGPT 应在实际界面选择可用模型/推理档；已经发出的请求不会因为一句“改用低档”自动重新计费。界面可用档位依套餐和工作区而异。 [S03]

06 / Codex 启动指令

先告诉它怎么工作，再告诉它做什么。

这段可直接复制。完整文本也在 Starter Kit 的 START_HERE.md；无需每轮重复粘贴。

请按“生产路由与节约用量”方式协作，不要默认最高模型、最高推理。

接到新任务，先确认交付物、验收条件、已有资产和可用资源；能从现有资料确定的，不重复提问。简单任务直接处理，复杂任务先给不超过四行的路由卡：推荐模型/推理、工作方法、预算边界、升级条件。说明推荐值与实际运行值；看不到运行信息就写“未核验”，不要声称已切换。

优先顺序：复用已验证结果 → 检索已有资产和官方参考 → 运行现成脚本/软件 → 模型完成必要的新判断。常规多步骤制作以已验证可用的 GPT-5.6 Sol / medium 为起点；简单任务可用资源表中已批准的轻量模型；只有难点、失败证据或高风险才升级。不要让高价主模型持续监督低价模型。高风险、不可逆操作可直接提高审查等级，不必先用低档试错。

先读取 resource-registry.local.json；不存在时给出缺项，不编造接口、额度和工具能力。模型、推理档、生图/视频渠道、内网工具均按实际可用性、能力、数据权限、剩余额度和返工成本选择。支持真实子代理/调用接口时，按批准配置执行并核验实际模型；不支持时给出切换建议，等我切换，不假装完成。

先做小样验证，再批量；初次执行失败后最多两次针对性修复，每次必须有新的诊断证据。资料不足、权限不足、服务故障先处理原因，不靠更强推理硬顶。不得擅自开启付费备用渠道、突破预算、扩大权限或外发客户素材。

只读取必要文件、章节、日志和图片；阶段完成就保存资产索引、脚本、验收结果和 HANDOFF.md，不反复重读整个项目。并行仅用于真正独立的任务，默认至多一个额外子代理。

结束时交付：结果路径、验收证据、未解决事项、实际模型/工具及可获得的用量。没有账单或运行记录的数字写“未知”。日报/周报/月报先由脚本聚合，只让模型分析异常和最多三条改进。达到验收即停止扩展。

本段是工作方法与授权约定，不会自行配置账户、内网或硬预算。真实模型选择、子代理调度和运行核验见下一页。 [S04] [S07]

07 / 真实切换与配置

“推荐用了什么”，不等于“实际跑了什么”。

先把便宜且够用的配置设成入口，再考虑自动调度；不要先用最高档长篇讨论怎样省用量。

方式	能做什么	必须核验
普通 ChatGPT	模型先建议，用户在界面选择实际可用档位。	可用模型与控件；不能用一句自报“已切换”当证据。
Codex / 子代理	支持的本地版本可指定子代理 model 与 model_reasoning_effort。	实际派生型号、推理档、父会话继承权限与所有调用用量。
自建调用层 / 内网	根据资源、任务和预算选择已批准的真实接口。	认证、接口能力、硬预算、幂等、回执、失败重试、数据流向。

先建立一个可靠的常规入口

```
codex --model gpt-5.6-sol -c 'model_reasoning_effort="medium"'
```

仅在账户已开放、客户端支持且管理员允许时执行。这会启动真实会话并可能产生用量；本交付没有替你运行。

常规主线程

Sol / Medium

读取必要资料，执行已有方法，组织小样与验收。

优先单线程；不为了“显得自动化”额外派代理。

一次性难点复核

Sol / High 或批准的 Astra

只处理失败证据和阻断问题。

输出修复规则后结束，将执行交回标准档。

必须避免的“纸面节省”

主模型仍是最高档，一直重读全量上下文、监督低档子代理。低档子代理省下的，可能又被主线程的重复规划和汇总吃掉。阶段交接应保存 HANDOFF，必要时新开标准档会话。

官方文档支持 CLI 选择模型及配置参数、自定义子代理文件。它们不保证每个旧客户端或账户都支持同样字段；生效情况必须实测。 [S04] [S07]

08 / 资源表与权限

开放的是能力，不是全部账号与权限。

Codex 要知道“有什么、能做什么、允许处理什么、怎样计费”，而不是只得到一串平台名字。

资源类型	管理员应登记
文本模型	准确 model ID、客户端/接口、已测推理档、可用账户、数据许可、计费渠道、余量与记录时间。
生图 / 视频渠道	参考/局部编辑能力、分辨率、输出类型、单次费用或积分、并发、队列、素材外发范围。
Blender / 本地工具	实际命令或 MCP 动作、读写目录、允许的控制项、版本、预览/批量成本、失败与回滚方式。
资产库 / 内网服务	检索入口、索引字段、文件/资产 ID、认证方式、动作白名单、数据边界、连通测试和负责人。

01 可用性

02 能力匹配

03 数据许可

04 预算批准

过完四道检查，才比较价格与速度

推荐表中所有访问默认 **unverified / disabled**。公开模型存在，不等于团队账号已开放；资源清单中的别名也不是现成接口。管理员完成最小测试后，才能启用。

可以主动调度的范围

已获授权的检索、预览、固定规则批处理、批准数量内的生成。优先通过窄接口调用，参数可校验、结果有回执。

必须保持审批的动作

付费备用渠道、充值、超额生成、客户素材外发、生产系统修改、覆盖和删除。密钥放凭据系统，不进聊天与 Skill。

MCP 可以按工具配置允许/禁止列表和审批策略；但写一份资源表不等于接通 MCP。Skill/JSON 也是行为约定，不是预算硬限额，硬控制要放在执行器或服务端。 [S08]

09 / 工具与生成渠道路由

不是所有视觉问题，都该重新生成。

把“工作方法路由”放在“模型路由”之前。对错误步骤提高推理，并不能修复方法本身。

实际需求	优先方法	小样验收
要同一个角色、同一个动作	先找批准资产/动作，再复用姿态、骨骼或参考帧。	身份比例与动作语义分别检查。
要精确相机、空间或接触	优先用现有 3D、骨骼、相机与几何控制。	接触、遮挡、视角、投影关系成立。
底图不变，只改局部	选择已批准且具备局部/参考编辑能力的渠道。	指定区域改变，非目标区域没有漂移。
需要未知风格 / 新构图	先少量候选测试方向，确定母版后收敛。	负责人先确认方向，不直接大批量铺量。
静帧要运动化	使用已批准的视频入口与参考资产。	先验动作、时长和连续性，再做高规格。

不要只看单次价格

渠道的候选数、失败重试、等待、返工、身份稳定性，都要纳入比较。低价但反复重做，不一定更划算。

参数不能乱套

文本模型的推理档，不是所有生图/视频工具的通用参数。资源表须分别登记每个渠道实际支持的控制项。

建议的小样停损规则

图像首轮最多两个候选、视频首轮最多一个候选，均需事先批准渠道与数量。选中方向后固定资产与约束，只改一类变量。达到预算或连续失败边界就停，先诊断，不自动切到另一付费渠道。

以上是试行方法，不是固定的最佳候选数。重要镜头可批准更大的探索预算；不能用低清通过代替最终成片验收。

10 / 上下文与可复用资产

别让下一次对话，重新支付上一次的理解。

项目事实保存在文件中；聊天负责推进当前阶段，不承担唯一的项目记忆。

留在哪里	保留什么	如何使用
AGENTS.md	短小的默认规则、权限与指向 Skill 的入口。	合并现有规则，不塞完整手册。
SKILL.md	任务触发、分级、工具选择、验收与停止流程。	需要时加载，具体参考与模板分开。
资产索引 / 映射表	来源、ID、版本、通过样本、失败标记。	按需检索，不每次重看整个素材库。
脚本 / 测试 / manifest	已验证步骤、参数、依赖版本、输出位置。	用差异更新和回归测试替代整体重建。
HANDOFF.md	已确认项、现有文件、失败证据、下一步、预算。	换模型、换线程或交同事时读取。

日常习惯：只加新信息

同一任务优先继续当前有效上下文，不反复贴整份旧资料。不同项目分线程；阶段切换用交接摘要，但保留可回查的原始证据。错误日志先截取相关段，图像先看代表样本，再精读需要验收的细节。

缓存不是“免费记忆”

API 提示缓存依赖可复用的前缀；上下文、模型或配置变化可能影响命中。它是输入处理缓存，不是输出结果缓存。不要只为追求缓存而堆无关文本，也不要假设每次压缩都一定更便宜；看实际回执与总成本。

Skill 的本地发现路径、AGENTS 的加载规则、API 缓存行为均有官方说明。对本工作室来说，最优先的是保存已验证资产与脚本，而不是依赖聊天“应该记得”。 [S05] [S06] [S11]

11 / 用量与生产记录

每天记一条事实，周报才有真实结论。

先把日志留下，再做统计。不要要求模型凭聊天长度猜用量，更不能把未知费用记成零。

记录层	最小字段
任务与结果	task_id、项目、任务类型、结束时间、通过/失败/阻塞、验收证据、结果路径。
实际执行	推荐模型与推理、实际调用模型与推理、核验来源、主线程与所有子代理/工具回执。
尝试与返工	初次 + 修复的尝试次数、升级原因、人工返工分钟、复用资产或脚本。
账目与额度	渠道、单位、实测 amount、回执来源、完整性。API、订阅额度、平台积分分开。

三种“用量”不能混着算

ChatGPT / Codex 套餐：记录实际显示的额度或管理员统计；不要按聊天轮数硬算。
API：记录原始 usage 与账单；推理 token 若已含于输出，不能再加一次。
生图 / 视频 / 本地算力：分别保留积分、货币或算力单位，不强行合并。

```
python .agents/skills/production-router/scripts/production_report.py \
  --input logs/production.jsonl \
  --period week --date 2026-09-29 \
  --timezone Asia/Shanghai \
  --output reports/2026-W36.md
```

改为 `--period day` 生成日报，改为 `month` 生成月报。脚本离线运行，不调用模型；输入日志要由获准的数据源提供。示例记录默认不计入真实报告。

ChatGPT 与 API 平台账务分开；适用工作区的实际消耗还会受到任务规模、模型、自动化与并发影响。管理员分析能力也依套餐和权限开放。本包未接入你们的账户或账单。 [S09] [S10] [S12] [S14]

12 / 周报与月报

报告用来修路线，不是排名谁花得少。

程序计算数字，模型分析异常；最后由负责人决定下周期的默认配置。

每周：修正在发生的浪费

交付与质量：通过了什么、哪些返工。
失败分类：缺参考、工具故障、选型错误或过度扩张。
复用资产：哪些脚本和方法下次能直接用。
行动：最多三条，有负责人和验收。

每月：调整生产默认值

同类任务：比较模型、推理和工具链。
资源配置：哪些订阅/渠道需要保留或调整。
标准化：哪些成功案例升级为 Skill。
更新：资源表、配额和版本兼容性。

指标	应采用的口径
一次通过率	已验收且 attempts=1 的通过任务 / 已完成验收且 attempts 已知任务；披露缺项。
真实模型分布	只统计有运行/回执证据的实际调用；推荐配置不算。
账目完整性	完整核对任务比例 + 未覆盖渠道；已知账目合计不冒充总成本。
同类任务成本	比较相同任务、相同验收；算入失败调用及人工返工，不能只看成功那一次。
可复用成果	新增脚本、资产索引、动作映射、测试样本与批准母版。

给报告分析模型的一句话

只依据这份已脱敏的聚合报告分析，不改统计数字。
区分数据缺口、工具/资料问题和模型能力问题。
给出最多三条下周期可验证的改进，写明依据、负责人待填项和验收方式。
没有同类基线与完整账目，不宣称节省比例。

定时执行尚未启用。先由管理员手动运行成功，再接工作室已有调度器、cron / launchd / CI；固定时区、受控目录和失败告警。自动邮件或群发送需另行授权。

13 / 从今天开始落地

先跑通一个项目，再把默认值发给全员。

把“使用习惯 → 可复用 Skill → 真实调度 → 生产复盘”分开上线，避免一次做成过重的平台工程。

A 第一步：立刻可用的规则

发出团队通知与启动指令；常规制作从 Sol / Medium 开始。只合并短版 AGENTS，复杂规则放 Skill。

B 第二步：核验实际资源

管理员补齐模型、账户、客户端、生图渠道和内网工具。每种资源做一次最小测试，记录真实能力和许可范围。

C 第三步：做同类任务小样

先选小猪资产/动作流程或另一条可回滚制作链。使用相同样本与验收，对比路线、重试和人工返工。

D 第四步：再接自动路由

确定性任务交脚本；只有真独立子任务才派代理。加硬预算、幂等、失败回执与审批，不靠提示词模拟控制。

E 第五步：持续复盘

每任务一条记录，每周三项改进，每月更新资源和默认路线。没有实测不承诺固定节省比例。

先复制 Skill，再运行离线检查

```
.agents/skills/production-router/SKILL.md

python -m unittest discover \
  -s .agents/skills/production-router/scripts \
  -p 'test_tools.py' -v
```

Starter Kit 提供 10 项合成数据测试，已在本次交付环境通过。它们只验证离线路由/统计逻辑，不代表工作室接口已接通。

当前官方 Skill 文档使用项目 `.agents/skills` 目录；AGENTS 应合并而非覆盖已有安全约束。子代理示例放在 examples，需确认版本后才复制到正式配置目录。 [S05] [S06] [S04]

14 / 来源与使用边界

把建议和已验证事实，清楚地分开。

官方文档核验于 2026-09-29。所有链接可点击；产品更新后请重新验证账户、客户端与配置。

S01 **OpenAI｜GPT-5.6 Sol 模型卡**
核验模型 ID 与支持的推理参数；不是团队账户权限证明。

S02 **OpenAI｜GPT-6 Astra 模型卡**
核验模型 ID 与支持的推理参数。

S03 **OpenAI Help｜GPT-5.6 and GPT-6 Pro in ChatGPT**
界面档位与可用性依套餐、工作区和逐步开放而变。

S04 **OpenAI｜Subagents**
子代理的模型、推理、配置文件与运行权限继承。

S05 **OpenAI｜Build skills**
SKILL.md 元数据、按需使用和 .agents/skills 本地发现路径。

S06 **OpenAI｜Custom instructions with AGENTS.md**
项目指令加载与局部覆盖机制。

S07 **OpenAI｜Codex CLI reference**
--model、--config、--sandbox 等命令行参数。

S08 **OpenAI｜Model Context Protocol**
MCP 接入、工具允许列表、凭据及动作审批配置。

S09 **OpenAI Help｜ChatGPT Rate Card**
适用工作区的用量影响因素；不能推广成所有订阅统一换算。

S10 **OpenAI Help｜Managing billing for ChatGPT and the API platform**
ChatGPT 与 API 平台账务区分。

S11 **OpenAI｜Prompt caching**
稳定前缀、缓存读写、实际回执口径；不是结果缓存。

S12 **OpenAI Help｜Using Codex with your ChatGPT plan**
账户/工作区使用边界与符合条件的管理员分析权限。

S13 **OpenAI｜Model catalog**
Sol/Terra/Luna 与 Astra 公开模型目录；访问须另验。

S14 **OpenAI｜Reasoning models**
推理与输出 token 的原始统计口径；避免重复计入费用。

U1 · 工作室案例：来自工作记录与所附截图；图片仅裁取案例区域，未包含桌面路径信息。并非独立测试、成本审计或已验证自动系统。

本指南提出：R0–R4 分级、默认起点、两次针对性修复、小样数量与上线阶段，均为试行政策，不是 OpenAI 的硬性限制。

本次实际交付：手册、Skill、模板、可选配置与离线脚本。未接入工作室内网、未改变账号配置、未创建周期任务、未取得全员聊天或账单。