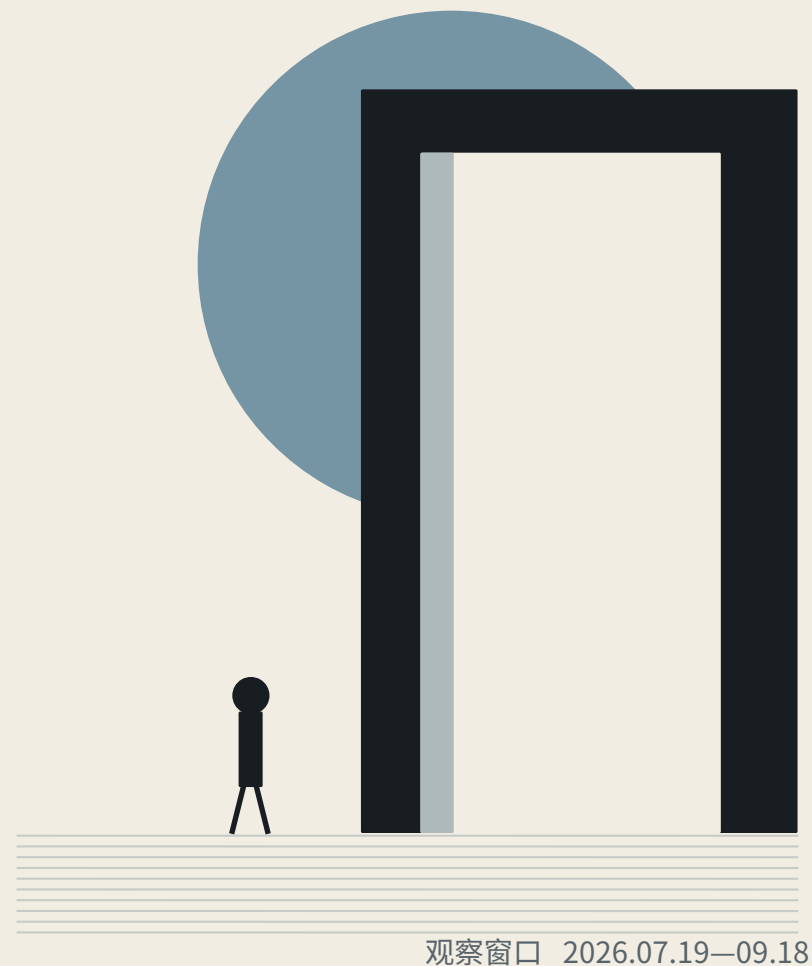


语言模型 和 世界模型 之争

操作专业软件 · 生成空间 · 结构化决策

郭宇 · 东京银座店 · 14:30-16:00



观察窗口 2026.07.19-09.18

这两个月，三种方向同时加速

07.28

World Labs

现实 → 模拟 → 现实

机器人训练与评估

09.03

OpenAI / Astra

更强的计算机操作

新一代通用模型

08.13

OpenAI / Sol

Ultrafast 预览

工作流的速度前史

09.12 JST

创作者实验

Astra × 服装型纸

Marvelous Designer

09.01

World Labs / Atlas

生成 + 重建 + 时空

新模型，早期接入

09.15

TypeSafe / Jev

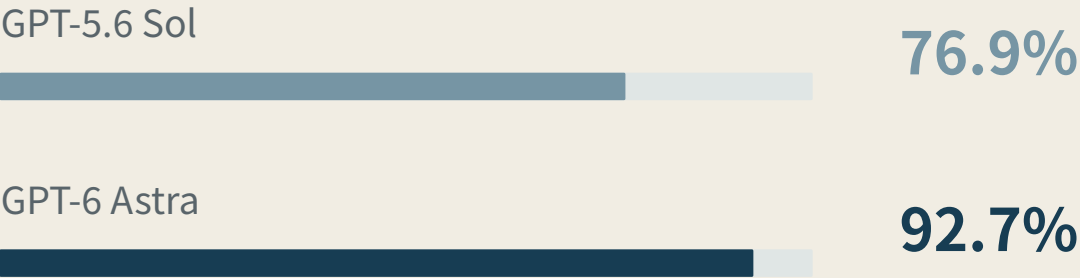
直接输出结构化决策

新模型，早期接入

进步在于：看屏幕，做动作，再纠错

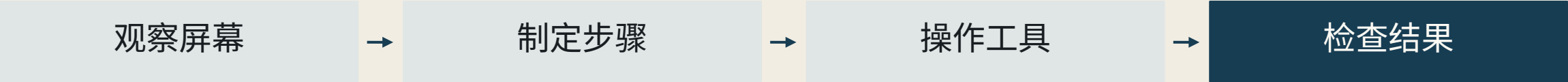
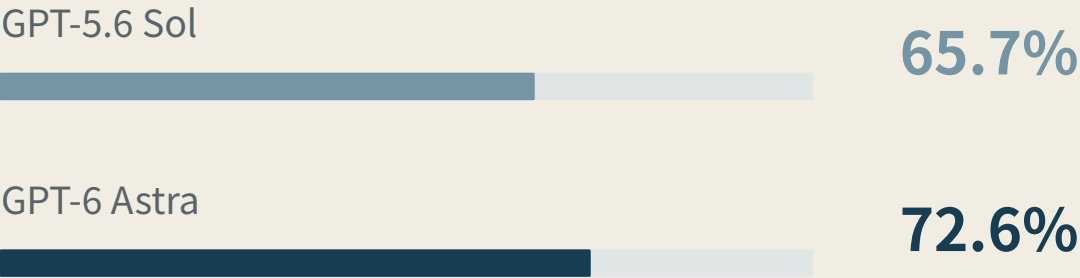
屏幕定位

ScreenSpot-Pro · 无工具



桌面任务

OSWorld 2.0 · 部分得分



厂商报告；OSWorld 2.0 为 v2026.08.08 离线集。两种指标分别衡量定位与任务完成，不能混算。

从参考图，到能编辑的三维资产



组合流程示意 · 经 MCP 操作 Blender，逐步检查。

09.18 · Stefan 的 Smart UV 展示（原帖视频封面）

UV：给三维表面一张二维“版型”

网格 → Smart UV 展开 → UV 布局

Image Generation 接着做什么？

布局 / 蒙版 + 参考图 → 生成或修改纹理
贴回 Blender → 检查接缝、拉伸与材质

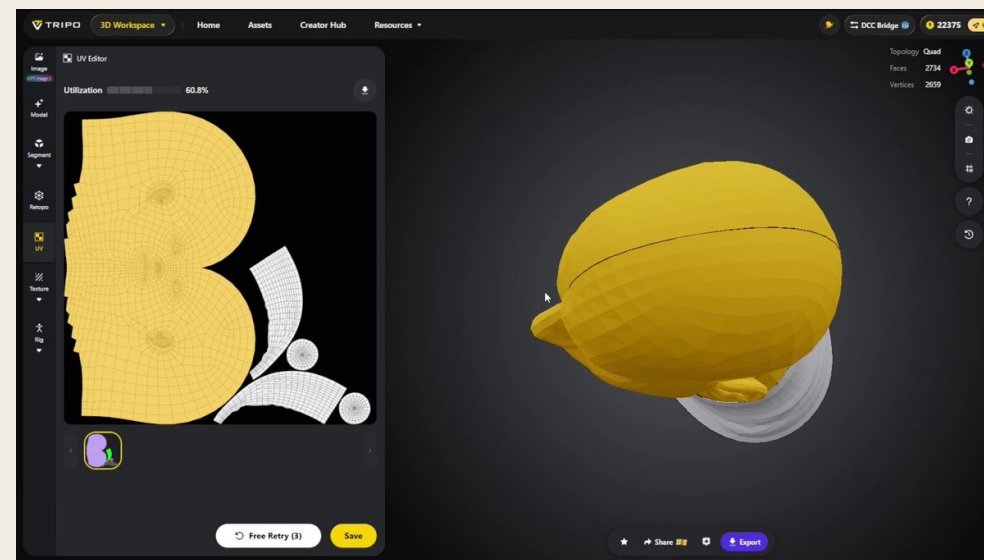
可组合的纹理流程示意；生成纹理不等于求解 UV 坐标。

[Tripo · Astra 角色流程](#)

[Stefan · Smart UV 原帖](#)

[Blender · UV 布局](#)

[Tripo · 图像与纹理](#)



服装已经能做，人的反馈仍然关键

2026.09.12 JST · 创作者自述

Astra → MCP → 型纸与服装

可继续编辑

输出进入专业软件的原有制作流程。

仍需多次修正

复杂褶边、丝带与细节，仍考验审美和判断。



图：Hirokazu Yokohara 原帖视频封面；公开实验观察，未复现。

[Yokohara · X 原帖](#)

[Yokohara · 后续说明](#)

[Marvelous Designer · API](#)

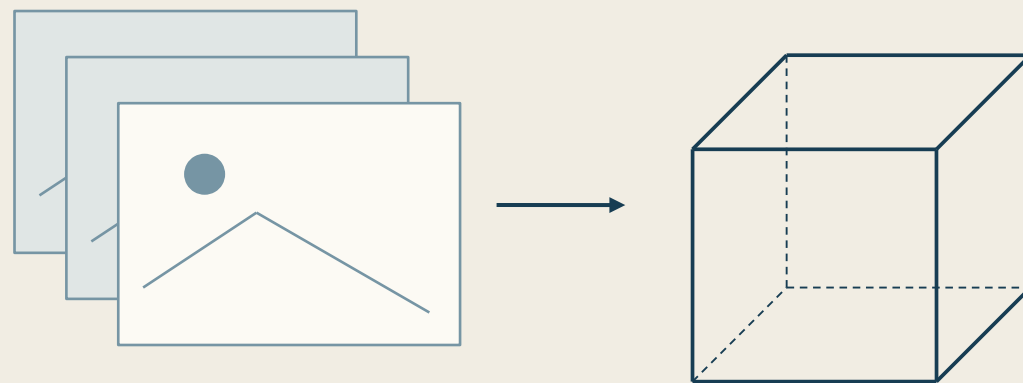
从几张照片，建立可探索的空间

09.01 Atlas

文本、图像、视频与 3D
进入共享的空间上下文

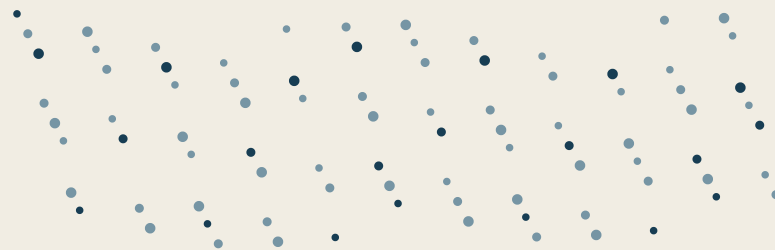
控制相机路径 / 补全未见区域
输出新视角、点云与 3D 高斯泼溅

Atlas：早期接入，非全量 Marble 更新



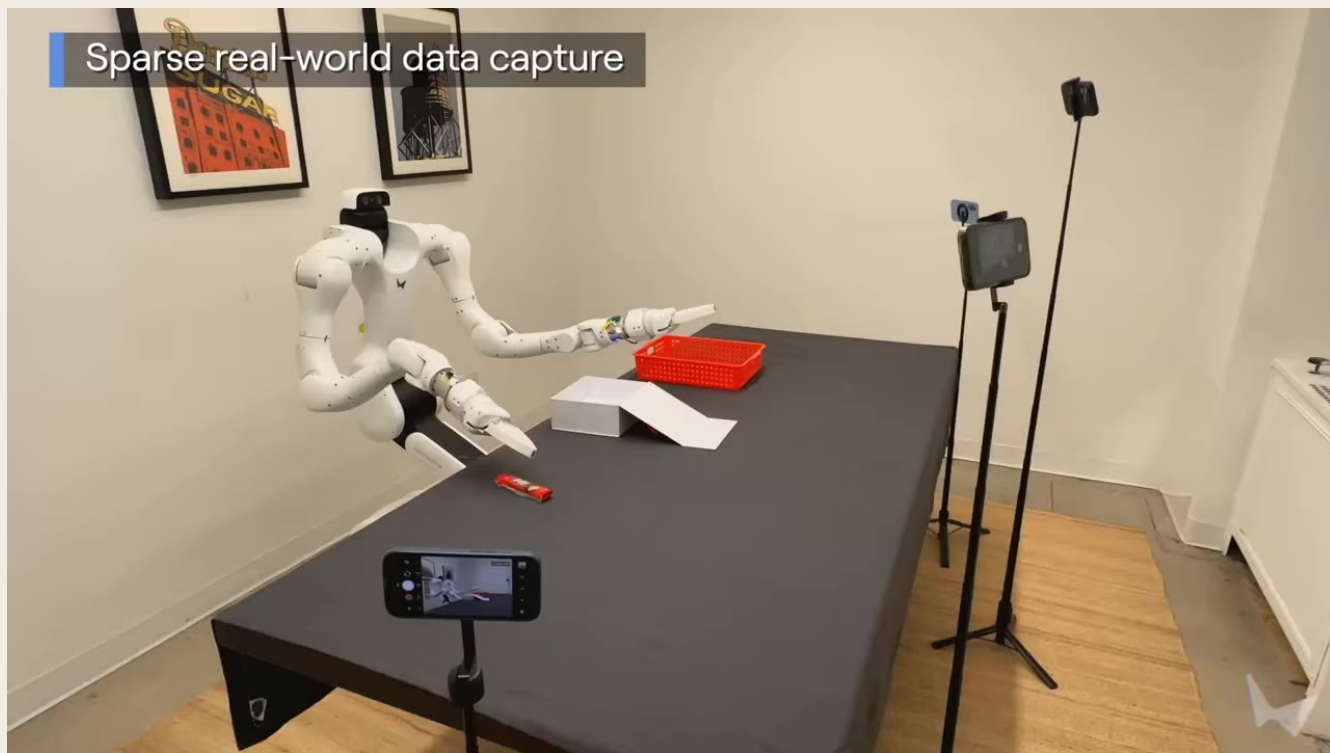
照片 → 空间表征 • 原创示意

点云示意



点云是点；splat 加入外观；mesh 是连接成面的几何。

分界正在移动：从看起来，到动起来



图：World Labs Atlas 官方视频封面 · 现实数据采集

已有能力 · 历史背景

Marble 已有碰撞网格

可以交给物理引擎使用；
几何仍需尺度、材质等设定。

近期进展 · 07.28 / 09.01

R2S2R → Atlas

从物理交互重建模拟，
再训练与评估机器人策略。

Jev：把判断压进毫秒级

Diogo Almeida • 前 OpenAI 研究员 • 09.15 发布 / 早期接入

70–500 ms

官方报告的端到端响应

美国西海岸测试；非延迟保证

\$0.042

每百万输入 tokens

Jev 1.13 • 官方 API 标价

\$0

输出费用

仅返回选择、评分和概率

为什么快：省去逐词写作，多个问题在一次请求中并行求值。

文本状态 + 预设问题



Jev 并行评估



选择 / 评分 / 概率

算例：每次总输入 1,000 tokens，10 万次约 \$4.20。

按 API 标价估算，未计重试、工具与基础设施；非本场实测。类型正确 ≠ 判断正确。

[TypeSafe • Jev 发布](#)

[TypeSafe • 模型与价格](#)

[TypeSafe • 决策接口](#)

速度、成本，要和判断质量一起看

TypeSafe 厂商评测 · 四类工作流等权平均 · 统一采用 workflow 配置

模型	平均耗时 / 例	平均费用 / 例	与参照答案一致率
Jev	0.4 秒	\$0.0004	67.8%
GPT-5.6 Terra	10.1 秒	\$0.0304	67.9%
Claude Opus 5	37.8 秒	\$0.1761	73.1%
DeepSeek V4 Flash	51.9 秒	\$0.0059	64.4%
Claude Sonnet 5	78.1 秒	\$0.1174	67.8%

对照 Terra：耗时约 1/25，费用约 1/76；参照答案一致率接近。

参照答案由 Astra + Fable 5.1 生成；一致率不等于人工验证准确率。

LLM 需输出概率、使用默认推理；倍数由公布的四舍五入值计算，非本场实测。

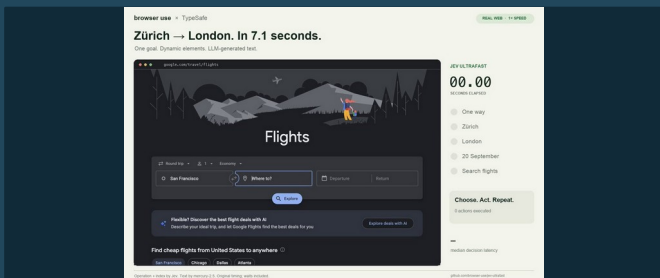
[TypeSafe · 工作流评测](#)

[TypeSafe · Jev 发布](#)

查航班，分邮件，玩游戏

浏览器查航班

@gregpr07 · 09.17



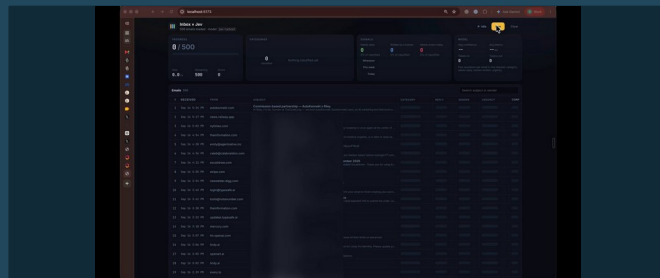
约 7 秒 · \$0.0039

Jev 选动作与页面元素
小型 LLM 生成输入文字

8,218 赞 · 原帖视频 ↗

批量分邮件

@rileybrown · 09.17



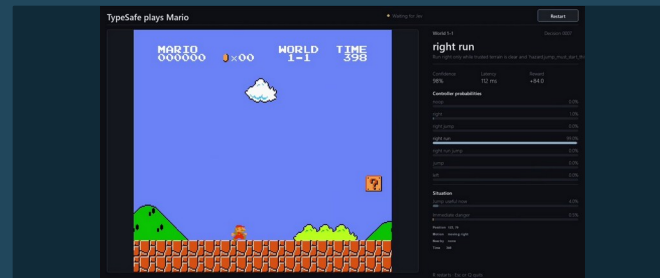
500 封 · \$0.035

批量分类邮件，并判断
是否回复、是否紧急

3,780 赞 · 原帖视频 ↗

实时玩马里奥

@faadilhshaik · 09.16



状态 → 动作

模拟器读出 JSON 状态
Jev 选择移动与跳跃

2,814 赞 · 原帖视频 ↗

讲者判断：高频判断交给 Jev，规划与生成交给通用模型。

耗时、费用为作者报告，未在本场复测。点赞：09.19 JST 快照；图片为原帖视频封面。

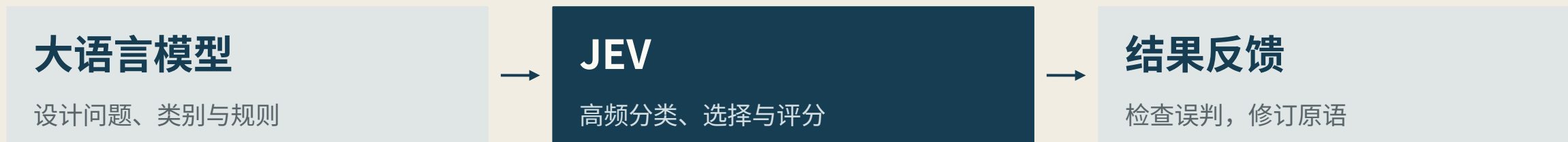
Browser Use · 代码与计时

Mario · 状态与动作代码

JEV：快速的超级通用分类器

我觉得 JEV 最大的用例是作为快速的超级通用分类器，它在大语言模型的时代，做了一件大语言模型时代之前、小型语言模型没有做好的事情，

**如果要用 JEV 来快速地组织信息并做出判断，
仍然需要大语言模型设计好足够丰富的设计原语**

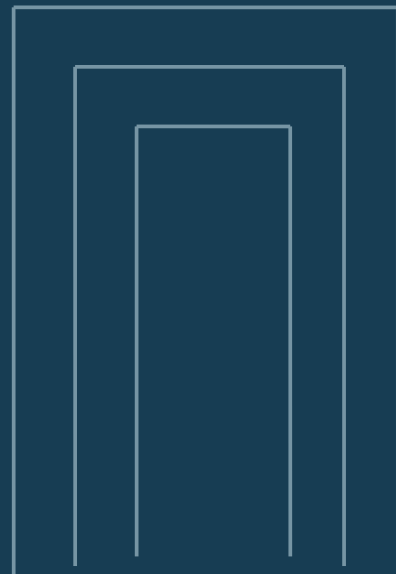


分工示意 · 以邮件为例：主题、紧急度、下一步动作。

使用智力获取超额收益的时代结束了，
但使用智能赚取复利的时代
却才刚刚开始。

我们每个月都处在一些事情改变的临界点。

今天在银座单向街书店的分享，我将对这个话题展开我的理解和解释。



让每次使用
留下下一次的起点