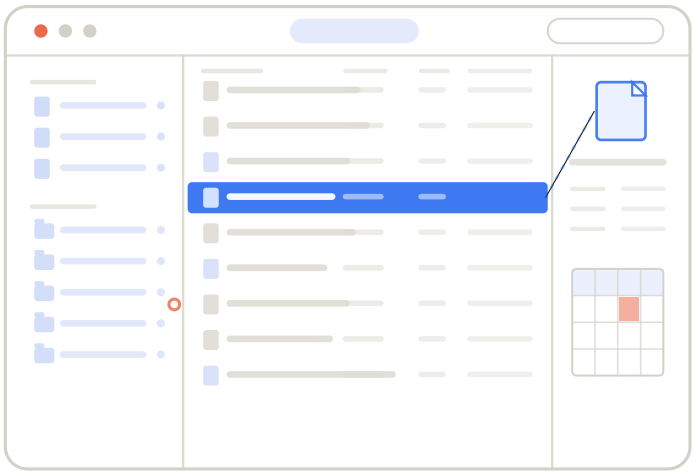
SoTALab
BY ZHIHU



MEASURABLE LIFT ON APEX-AGENTS

APEX-Agents 训练数据

APEX-Agents · Investment Banking

为真实、可交付的投行 / 金融 Agent 任务，
构造同分布、可训练、可评测的过程级数据。

变局 · FROM Q&A TO REAL WORK

答对问题，不等于能干成工作。

AI 的竞争正从「知识问答」转向「能不能干真实、**有经济价值的工作**」——这不只是我们的判断，Mercor 在推出 APEX-Agents 时也明确指出：现有大部分基准只在孤立的提示词或单一技能上测模型，**测不出 agent 能不能在真实、杂乱的工作场景里把一件事从头干到尾**。

衡量这件事，不再只是一道有标准答案的题，而是还需要有**一整个真实的工作环境**：一屋子文件、邮件与对话，一个必须自己翻找、多步推理才能干成的任务——Mercor APEX-Agents 正是衡量这种能力的公认标尺。

> **上百份文件、往来的邮件、一句模糊的题面——这才是真正的考验。**

APEX-AGENTS · REAL ECONOMIC-VALUE WORK

为什么重要 · 评测新趋势

一类更真实的 Agent 评测，正在成为共识。

Agent 评测正从**孤立问答**，走向**真实、长链路、贴近专业工作流**的任务——Mercor 的 APEX-Agents 正是这一趋势的代表：它衡量前沿 Agent 能否跨应用、长链条完成真实专业工作。

更真实
256 位从业者

由 **256 位资深从业者**（平均 12.9 年，来自 BCG / 麦肯锡 / 大摩 / 花旗）**按真实工作流编写**，覆盖投行 / 咨询 / 法律三类。

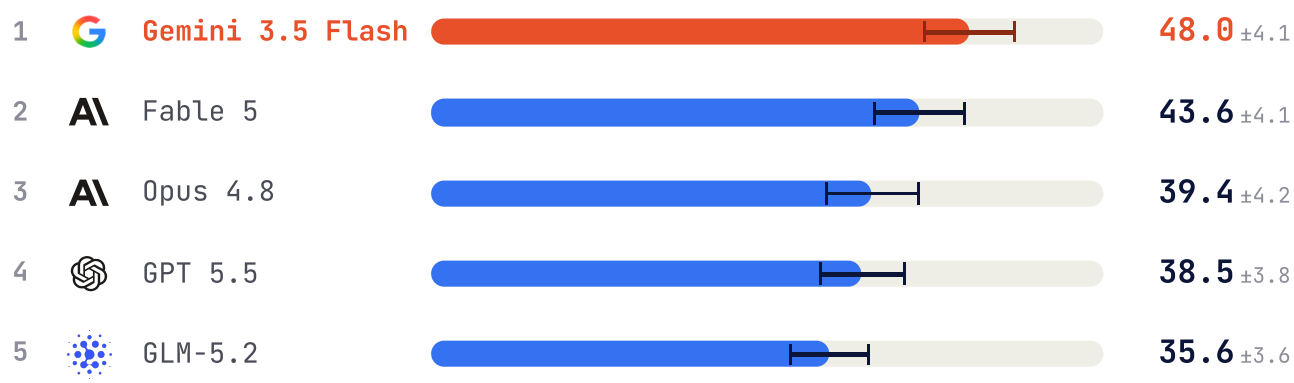
渐成共识
持续追踪

各大前沿实验室持续投入这类评测；第三方机构（Artificial Analysis）也**独立复现**了结果。

练通用能力
APPLIED COMPUTE

第三方（Applied Compute）：用这类数据训练的模型，**GDPVal +7.7pp、Toolathalon +8.0pp**——**迁移到未见任务**，而非拟合本榜。

MERCOR 官方榜 · APEX-AGENTS · PASS@1 · 来源 MERCOR.COM/APEX



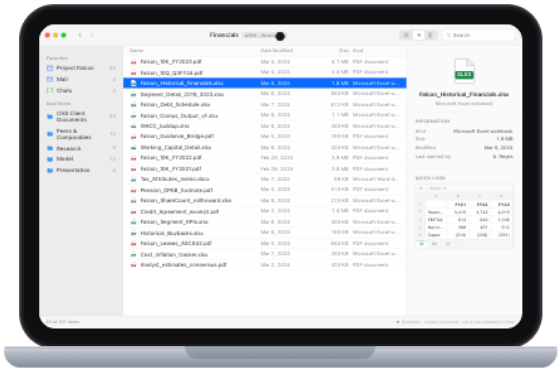
官方数据集 33 WORLDS · 480 TASKS · 覆盖投行 / 咨询 / 法律三类

官方榜**全员 Pass@1 不到一半**——又重要、又极难。我们造对口的同分布数据，**不是为了刷这个榜，而是训练这类真实工作背后的通用能力**——已验证能泛化到未见任务（GDPVal +7.7pp / Toolathalon +8.0pp）。

是什么 · DATA STRUCTURE

一个 World，一整套围绕它的 Task。

本期 SoTALab 数据专注**投行 / 金融**垂类。一个 **World** 模拟投行经理的真实工作电脑（文件、邮件、对话）；围绕它构建**一系列 Task**——每个 World 通常承载多道真实工作任务，每道配 Rubrics 与 Gold Response。



WORLD 219 · DEAL ROOM · FINANCIALS — 投行经理的工作电脑 · SYNTHETIC · EXPERT-REVIEWED

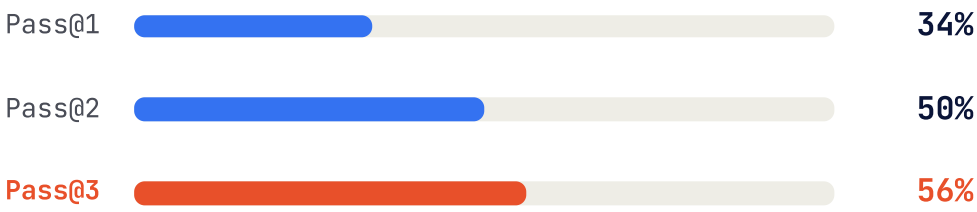
模块	元素	用途
World	File system	模拟投行经理的文件系统，平均含 **150+ 高质量附件**（PDF / 表格为主）。
	Email	模拟投行经理之间的邮件往来。
	Chat	模拟投行经理之间的 IM 即时消息往来。
Task	Task Prompt	一句简短题面，通常要求输出数字结论，或生成 / 修改文件。
	Rubrics	LLM-as-a-Judge 的判据，一条或多条精确检验 final answer 正确性的规则。
	Gold Response	给人类看的专家参考答案。
运行结果	Score	答对 / 答错 Rubrics 的得分。
	Trajectory	特定模型完成 Task 的执行轨迹。

抽样统计**与官方分布相似**（每 World 150–200 附件，PDF / 表格为主）；难度经三模型 90 次运行验证**可真实区分**。

为什么难 · 失败模式

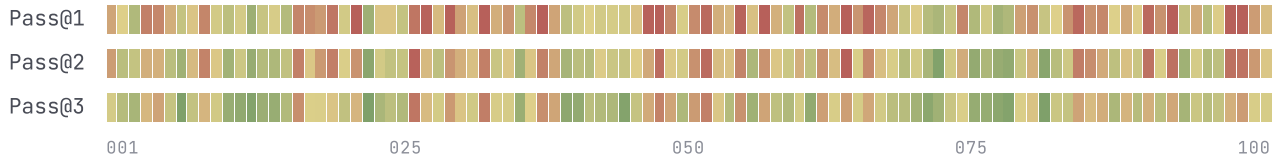
最强模型也只爬到一半——可观的 RL 空间就在这里。

即便最强的前沿模型，单次通过率 **Pass@1 也只有约 34%**；Pass@1 与 Pass@3 相差约 **20pp**，正是可观的 RL 提升空间。这不是问答，是**「文件之海」**——单题 ≈ **149 个文件**、~88 次工具调用才能干成一道。



PASS@K 爬升 · SOTALAB 自测 100 题 · 越难爬升越慢

运行 SCORE 热力图 · PASS@1/2/3 × 100 题 · 示意



从 PASS@1 到 PASS@3，绿色（做对）越来越多——「多次才对」的题，正是 RL 的空间

失败模式分布 · 抽样计数 · 高度集中

失败并非随机噪声，而是集中体现能力短板：**无法识别权威数据源、无法完成跨文件口径对齐、遗漏中间计算链路，或在多个候选答案间无法收敛到唯一最终答案**。

怎么造 + 谁造 · PRODUCTION LINE

造得出，还造得对。

公开榜禁止训练，能造出**同分布新题**的厂商极少——我们把「难造」拆成一条**可控、可质检、可准出**的生产管线，每一步都由系统合成 + 专家把关。

1 高质量附件采集

采集 **20 万+** 高质量财报、投资模型附件，Meta 打标供合成任务。

2 World 合成

系统合成 World 并**对齐官方统计分布**，由三层专家体系质检真实性。

3 Task 合成

准出线 = **Gemini 3.5 Flash 3 次至少失败 1 次**；**两位专家背靠背答题**核对 Gold Response。

4 分布准出

统计上剔除难度偏低、质量偏差题目，只**准出高质量题**。

谁造 · 三层专家体系

高校学术型
清北复交人 + 头部投行实习

资深实务型
3 年+ IPO / 并购 / 融资全职

专项研究型
CFA / FRM / CPA · 量化风控

系统合成保证**规模与分布**，专家把关保证**真实与正确**——两者缺一，这种题都造不出来、也造不对。

不断进化 · WORLD & TASK CO-EVOLUTION

Task 反哺 World，数据资产持续迭代。

数据不是一次性产出。每一轮：系统围绕 World 批量生成 Task → **三层专家体系对 Task 与 World 一起质检把关**，既筛掉不合格的 Task，也标出 World 里不够真实、不够完整的地方 → World 据此补充修正 → 基于更扎实的 World，再长出更好的 Task。转得越多，资产越厚——不是模型变强让旧题过期，是这条产线本身在持续增值。

1000+
已构造高质量任务，仍在持续增长

∞
可合成 World · 同分布无上限

3层
专家体系：学术型 · 实务型 · 研究型

每一轮迭代都由**三层专家体系（学术 · 实务 · 研究）**质检把关——World 更完整、Task 更扎实。这是持续增值的数据资产，不是一次性交付的快照。

样本 · LBO 建模

样本 | Investment Banking | LBO · IRR / MoM

一句话的题面，一间数据室的活

> 题面只有一句，但答案埋在数据室深处——模型要自己找到 LBO 模型、套用新假设、重算，才能给出正确数字。

TASK PROMPT

From the LBO, assume a **15% premium** is offered to CNS holders and max transaction leverage of Term Loan B is **$1,040**. What is the revised **IRR and MoM for 2029E**, rounded to one decimal place?

World ~150 files | PDF | XLSX 模型 | Email | Chat

✦ 干成这道题要几步

> **翻找**数据室里的 LBO 模型（约 150 个文件中定位到那张 XLSX）。

> **套用新假设**：15% 溢价、Term Loan B 上限 $1,040，跨表口径对齐。

> **重算并收敛**到单一数字：2029E IRR = **15.8%**、MoM = **2.1x**。

✓ RUBRICS · 判据

- States 2029E IRR is **15.8%**
- States 2029E MoM is **2.1x**

⚠ 常见失手

- 计算口径错、多候选未定——给不出唯一收敛的最终数字。

一句话的题面 → **一间完整数据室的多步推理（~150 文件 · ~88 次工具调用）**

一条这样的数据 = 可训练的题面 + 可机检的 Rubrics + 专家 Gold Response + 完整 World ——**模型必须真把活干完，才能得分**。

邀请验证 · PROVE IT ON YOUR TASKS

随模型进化——
邀请你用自己的任务实测一轮。

同分布可训练

公开榜禁训练，我们造同分布的新题——可训练、可评测、可追错

随模型进化

准出线随最强模型上移，难度持续跟进前沿

专家闭环

三层专家体系：质检 · 背靠背验证 · 真实性把关

敢自证

给你看我们造的 World demo + 真实 trajectory

告诉我们你的训练 / 评测方向，我们用**你的任务做一轮 pilot**，快速组建对口的投行 Agent 过程级数据与评测管线。

预约 World Demo · 申请 Pilot →

contact@sotalab.ai · 微信 / 飞书联系对接人

Measurable lift on APEX-Agents

真正有价值的 Agent 数据，衡量的是能不能干成一件真实的工作。

SoTALab

BY ZHIHU