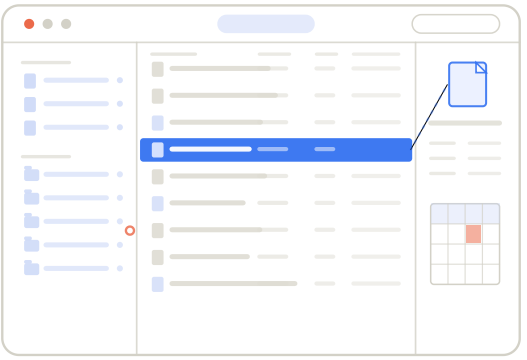
SoTALab
BY ZHIHU

PRODUCT BRIEF
NO.03 / 2026



MEASURABLE LIFT ON APEX-AGENTS

APEX-Agents Training Data

APEX-Agents · Investment Banking

Process-level data that is in-distribution, trainable, and measurable—
built for real, deliverable investment-banking / finance Agent tasks.

THE APEX-AGENTS RULER

EXPERT-CALIBRATED · MULTI-MODEL EVALUATION

THE SHIFT · FROM Q&A TO REAL WORK

Answering questions is not the same as getting work done.

The contest in AI is shifting from "knowledge Q&A" to "can it do real, **economically valuable work**"—and this is not just our view. In launching APEX-Agents, Mercor made the same point: most existing benchmarks test models on isolated prompts or single skills, and **cannot measure whether an agent can take a task from start to finish in a real, messy work setting**.

Measuring this is no longer one question with a fixed answer; it needs **a whole real work environment**: a room of files, emails, and chats, and a task you can only finish by digging through them and reasoning across many steps—**Mercor APEX-Agents** is the recognized ruler for exactly this ability.

> Hundreds of files, threads of email, and one vague prompt— that is the real test.
>
> APEX-AGENTS · REAL ECONOMIC-VALUE WORK

WHY IT MATTERS · A NEW EVALUATION TREND

Realistic Agent evaluation is becoming the consensus.

Agent evaluation is shifting from **isolated Q&A** to **real, long-horizon, professional-workflow tasks**. Mercor's APEX-Agents typifies it—measuring whether frontier Agents finish real work across apps and long chains.

More real
256 PRACTITIONERS

Avg 12.9 yrs' experience (BCG / McKinsey / Morgan Stanley / Citi), across banking / consulting / law.

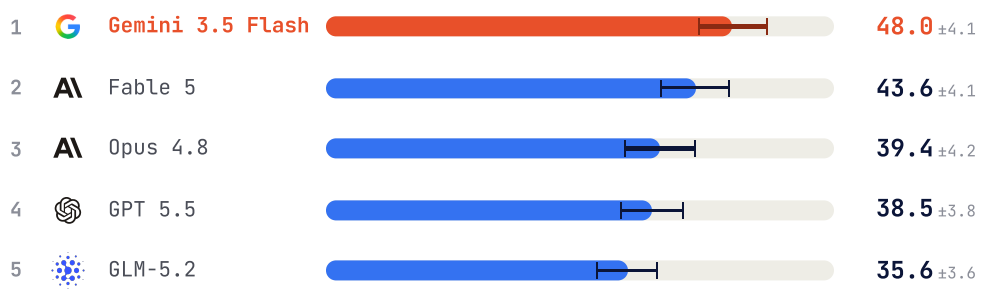
Consensus
TRACKED OVER TIME

Frontier labs keep investing here; Artificial Analysis has **independently reproduced** the results.

Transfers
APPLIED COMPUTE

Applied Compute: models gain **GDPVal +7.7pp, Toolathalon +8.0pp on unseen tasks**—not the board.

MERCOR OFFICIAL BOARD · APEX-AGENTS · PASS@1 · SOURCE MERCOR.COM/APEX



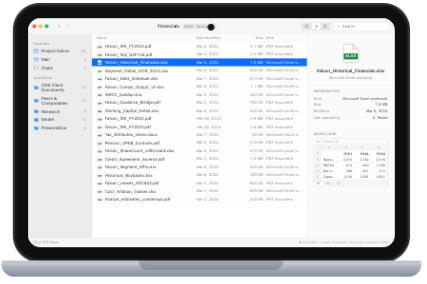
OFFICIAL DATASET: 33 WORLDS · 480 TASKS · SPANNING BANKING / CONSULTING / LAW

On the official board, **every model scores under half on Pass@1**—important yet brutally hard. We build matched in-distribution data **to train the general skill behind this real work**.

WHAT IT IS · DATA STRUCTURE

One World, a full set of Tasks around it.

This SoTALab release targets **banking / finance**. A **World** simulates a banking manager's work computer (files, emails, chats); around it we build **a series of Tasks**—each usually carries several, each with Rubrics and a Gold Response.



WORLD 219 · DEAL ROOM · FINANCIALS — A BANKER'S WORK COMPUTER · SYNTHETIC · EXPERT-REVIEWED

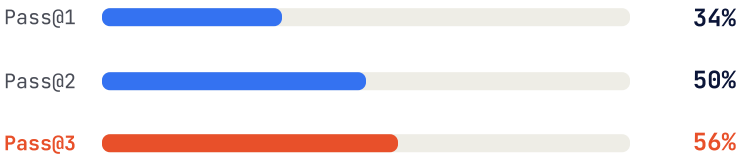
MODULE	ELEMENT	PURPOSE
World	`File system`	A file system, **150+ high-quality attachments** (mostly PDFs / spreadsheets).
	`Email`	Email exchanges between banking managers.
	`Chat`	IM chat between banking managers.
Task	`Task Prompt`	A one-line prompt—usually a numeric conclusion, or create / edit files.
	`Rubrics`	LLM-as-a-Judge criteria—rules checking the final answer's correctness.
	`Gold Response`	Expert reference answer, for humans.
Run result	`Score`	Score for passing / failing the Rubrics.
	`Trajectory`	A model's execution trace on the Task.

Sampled stats **match the official distribution** (150–200 attachments/World); difficulty is **genuinely discriminative**, validated over 90 runs of three models.

WHY IT'S HARD · FAILURE MODES

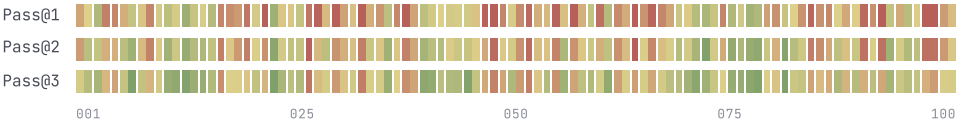
Top models reach halfway—that's the RL headroom.

Even the strongest frontier models pass just once at about **Pass@1 ≈ 34%**; the gap to Pass@3 is ~**20pp**—exactly the RL headroom. Not Q&A but a **"sea of files"**—one task ≈ **149 files** and ~88 tool calls.



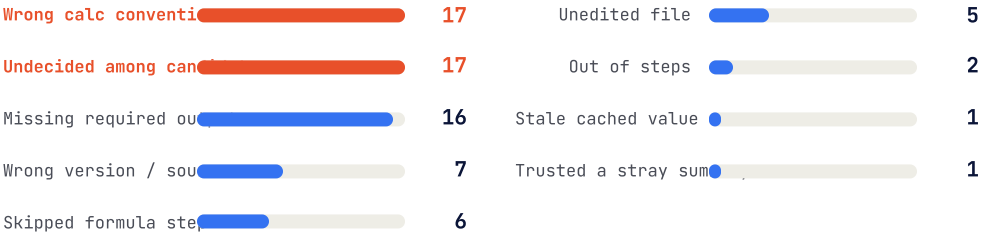
PASS@K CLIMB · SOTALAB SELF-TEST, 100 TASKS · THE HARDER, THE SLOWER THE CLIMB

RUN-SCORE HEATMAP · PASS@1/2/3 × 100 TASKS · ILLUSTRATIVE



FROM PASS@1 TO PASS@3, GREEN (CORRECT) GROWS—MULTI-TRY TASKS ARE THE RL HEADROOM

Failure-mode distribution · sampled counts · highly concentrated



Failures are not noise but concentrated gaps: **missing the authoritative source, misaligned conventions across files, skipped steps, or no single converged answer**.

HOW & WHO · PRODUCTION LINE

Built at scale, and built right.

Public boards forbid training, and few vendors can create **new in-distribution tasks**—we break "hard to build" into a **controllable, inspectable, gated** production line, each step combining system synthesis with expert review.

1 High-quality attachment collection

Collect **200K+** high-quality filings and investment-model attachments, meta-tagged to feed task synthesis.

2 World synthesis

The system synthesizes Worlds and **aligns them to the official statistical distribution**, with realism checked by a three-tier expert system.

3 Task synthesis

Release bar = **Gemini 3.5 Flash fails at least once in 3 tries**; **two experts answer back-to-back** to verify the Gold Response.

4 Distribution gating

Statistically remove too-easy or off-quality tasks, releasing **only high-quality ones**.

WHO BUILDS IT · THREE-TIER EXPERT SYSTEM

Academic
Tsinghua/PKU/Fudan/SJTU/RUC + top-bank internships

Practitioner
3+ yrs full-time IPO / M&A / financing

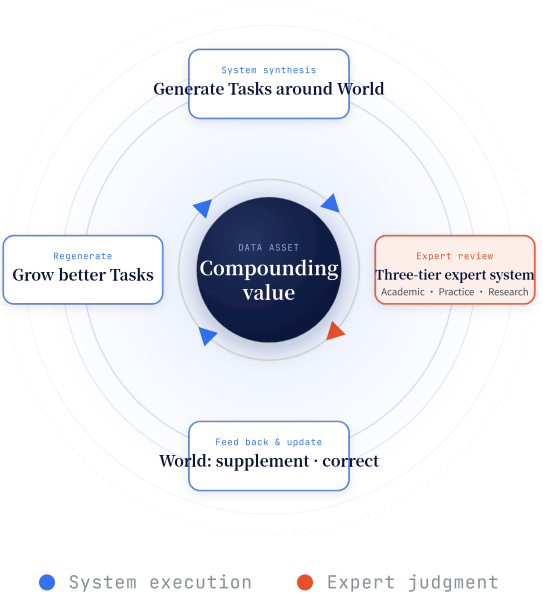
Specialist
CFA / FRM / CPA · quant & risk

System synthesis guarantees **scale and distribution**; expert review guarantees **realism and correctness**—without either, these tasks can be neither built nor built right.

ALWAYS EVOLVING · WORLD & TASK CO-EVOLUTION

Tasks feed back into Worlds, a compounding data asset.

Data is not a one-off output. Each round: the system batch-generates Tasks around a World → **the three-tier expert system reviews Tasks and World together**, cutting weak Tasks and flagging gaps → the World is fixed and enriched → better Tasks grow on a sturdier World. The asset thickens each turn—not old tasks going obsolete, but the line appreciating.



1000+
high-quality tasks built, still growing

∞
synthesizable Worlds · no cap, in-distribution

3 tiers
expert system: Academic · Practitioner · Specialist

Reviewed each round by the **three-tier expert system (Academic · Practice · Research)**—fuller World, sturdier Tasks. A continually appreciating asset.

SAMPLE · LBO MODELING

Sample | Investment Banking | LBO · IRR / MoM

A one-line prompt, a data room's worth of work

> The prompt is one line, but the answer is buried deep in the data room—the model must find the LBO model itself, apply new assumptions, and recompute to reach the right number.

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 model | Email | Chat

✦ STEPS TO FINISH THIS TASK

> **Dig** for the LBO model in the data room (locate that one XLSX among ~150 files).

> **Apply new assumptions**: 15% premium, Term Loan B cap $1,040, conventions aligned across sheets.

> **Recompute and converge** to single numbers: 2029E IRR = **15.8%**, MoM = **2.1x**.

✓ RUBRICS · CRITERIA

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

⚠ COMMON MISSES

- Wrong calc convention, undecided among candidates—no single converged final number.

A one-line prompt → **Multi-step reasoning over a full data room (~150 files • ~88 tool calls)**

One such datum = a trainable prompt + machine-checkable Rubrics + an expert Gold Response + a full World—**the model scores only if it truly gets the work done**.

PROVE IT · ON YOUR TASKS

Evolving with the models— run one real round on your own tasks.

In-distribution & trainable

Public boards forbid training; we build new in-distribution tasks—trainable, measurable, debuggable

Evolves with models

The release bar rises with the best model; difficulty keeps pace with the frontier

Expert loop

Three-tier expert system: review · back-to-back verification · realism check

We prove it

We show you a World demo we built + real trajectories

Tell us your training / evaluation direction, and we will **run one pilot on your tasks**, quickly standing up a matched process-level data and evaluation pipeline for banking Agents.

Book a World Demo · Request a Pilot →

contact@sotalab.ai · WeChat / Feishu to reach us

Measurable lift on APEX-Agents

Truly valuable Agent data measures whether it can get a real job done.

BY ZHIHU