让多智能体协作的进展可积累：面向长程机器学习的AI Agent

Toward Autonomous Long-Horizon Engineering for ML Research

陈国鑫

中国人民大学

Background

AI Box

➢ML Research Engineering: 在固定时间/GPU预算内，把一篇论文或一个研究目标，变成可运行、可验证、可迭代改进的 ML系统

Paper Bench

复现20篇ICML论文（包含环境安装，数据模型下载，以及baseline的复现）

MLE Bench Lite

进行Kaggle竞赛，并与人类榜单比较

21% vs 41%

PaperBench: best agent vs top ML PhDs

24h/task

realistic long-horizon budget

Background

➢四大挑战:

Underspecification

论文不是完整 specification，关键实现细节常常隐含或缺失

System setup

数据、模型、依赖、CUDA、评测脚本都会成为真正瓶颈

Delayed feedback

实验几小时后才返回，失败原因还常常混杂

State continuity

后续决策必须正确继承早期假设、代码变更、实验结果

关键难点：反馈延迟且混杂，进展必须跨阶段、跨角色、跨多轮调用持续累积。

Core Insight: Long-running Agent needs continuity

AI Box

- 智能体调用是transient，而长程问题必须cumulative
- 系统需要保存“未来会用到的证据”

Role continuity

同一智能体下次回来时，能恢复它之前试过什么、假设什么、失败在哪里

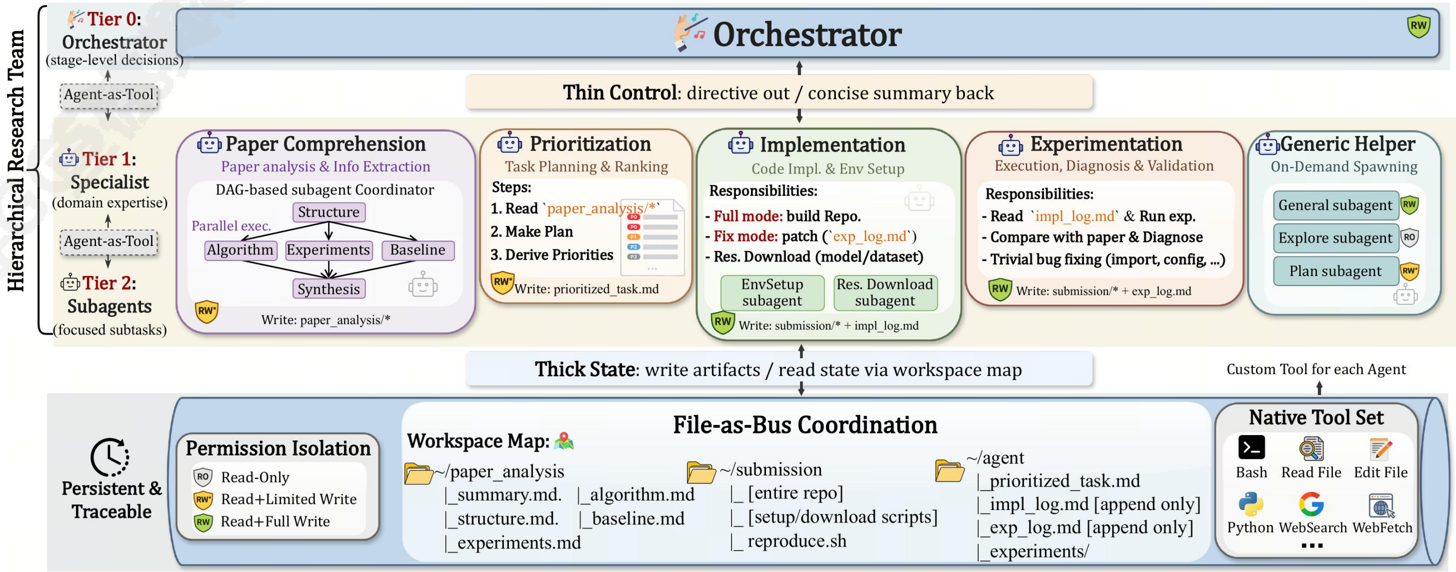
Project continuity

不同角色围绕同一套 evidence 协作，而不是依赖压缩过的 conversational handoff

transient activity → durable progress

AiScientist：thin control over thick state

- ➢Orchestrator 只保留 stage-level control
- ➢详细证据沉淀在 File-as-Bus workspace 中，由智能体按需读取。

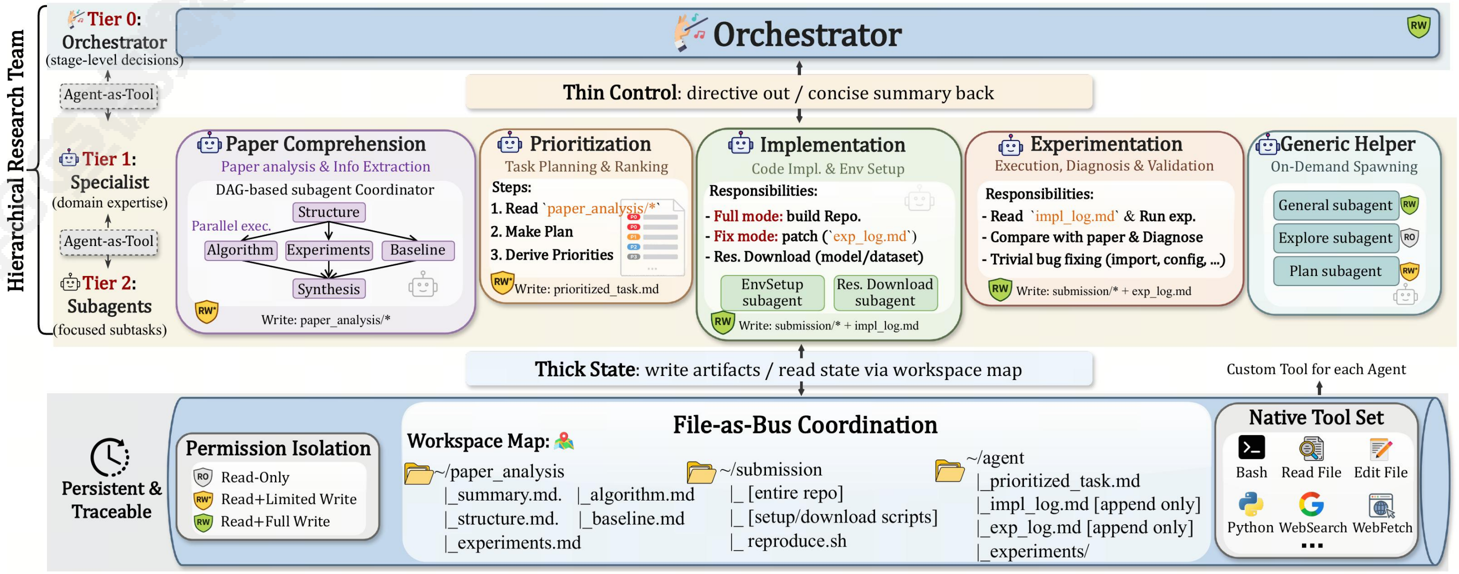


Thin control：concise directives + workspace map

Thick state：durable artifacts + role-readable evidence

AiScientist：thin control over thick state

- ➢Orchestrator 只保留 stage-level control
- ➢详细证据沉淀在 File-as-Bus workspace 中， 由agent按需读取。



Thin control：concise directives + workspace map

Thick state：durable artifacts + role-readable evidence

File-as-Bus：从共享文件夹到 protocol

AI Box

- 每个 artifact 都有 purpose、update mode、writer 和 readers；workspace 因此成为 collaboration contract

Update mode	What it preserves	Examples
Versioned artifact	当前 canonical state + 可追溯修订	paper_analysis/*.md, prioritized_task.md
Append-only log	按时间保留失败、诊断、决策证据	impl_log.md, exp_log.md
Mutable runnable artifact	真实可执行项目状态	submission/, setup/, reproduce.sh

关键点： 不是把所有东西塞进 context，而是让 agent 能通过 workspace map 找到“该读什么证据”

Hierarchical team： delegation 是控制动作

➢Orchestrator 把 specialist invocation 当作 tool action；只有下一步需要局部长上下文或专业能力时才 delegation

Orchestrator	Specialists	Tier-2 subagent	Workspace
stage-level decisions compact control context	paper comprehension prioritization implementation experimentation	bounded subtasks resource lookup local planning	summary back to control artifacts back to state

关键点： 不是把所有东西塞进 context，而是让 agent 能通过 workspace map 找到"该读什么证据"

Evidence loop：delayed feedback 变成证据

AI Box

- 早期先形成 runnable scaffold
- 后续通过 implementation ↔ experimentation ↔ validation 的循环，把失败、metric gap 和修复记录下来

Understand
extract task, metrics, ambiguities

—

Prioritize
rank dependencies, impact, feasibility

—

Implement
build repo, setup resources

—

Experiment
run, compare, diagnose

—

Validate
close gaps, prepare submission

Loop invariant: 每轮输出都必须留下后续能 inspect 的 evidence：code-side decisions、failed runs、metrics、diagnoses、unresolved blockers

Evidence loop：delayed feedback 变成证据

- 早期先形成 runnable scaffold
- 后续通过 implementation ↔ experimentation ↔ validation 的循环，把失败、metric gap 和修复记录下来

Understand
extract task, metrics, ambiguities

—

Prioritize
rank dependencies, impact, feasibility

—

Implement
build repo, setup resources

—

Experiment
run, compare, diagnose

—

Validate
close gaps, prepare submission

Loop invariant: 每轮输出都必须留下后续能 inspect 的 evidence：code-side decisions、failed runs、metrics、diagnoses、unresolved blockers

Main Results

AI Box

- ➢ AiScientist 在 paper replication 和 competition-style ML improvement 两类长链路 (24h) 任务上都有效

+9.92 / +11.15
PaperBench vs strongest matched baselines

81.82%
MLE-Bench Lite Any Medal under both backbones

-31.82 pts
remove File-as-Bus on MLE Any Medal

Long-horizon AI research is not merely a problem of stronger local reasoning; it is a systems problem of maintaining cumulative, inspectable progress.

PaperBench：关键是复用 prior evidence

AI Box

- IterativeAgent 已经增加交互，但 AiScientist 更高分且更低成本；说明 round count 不是充分解释

Backbone	Strongest baseline	AiScientist	Gain	Cost/task
Gemini-3-Flash	20.60	30.52	+9.92	$15.67 vs $27.44
GLM-5	22.58	33.73	+11.15	$12.20 vs $54.90
Frontier ref.	Codex/GPT-5.5: 29.45	Best: 33.73	+4.28	not matched baseline

Takeaway: extra interaction helps only when it is organized around cumulative research-engineering work.

MLE-Bench Lite：valid 只是起点，medal 才难

AI Box

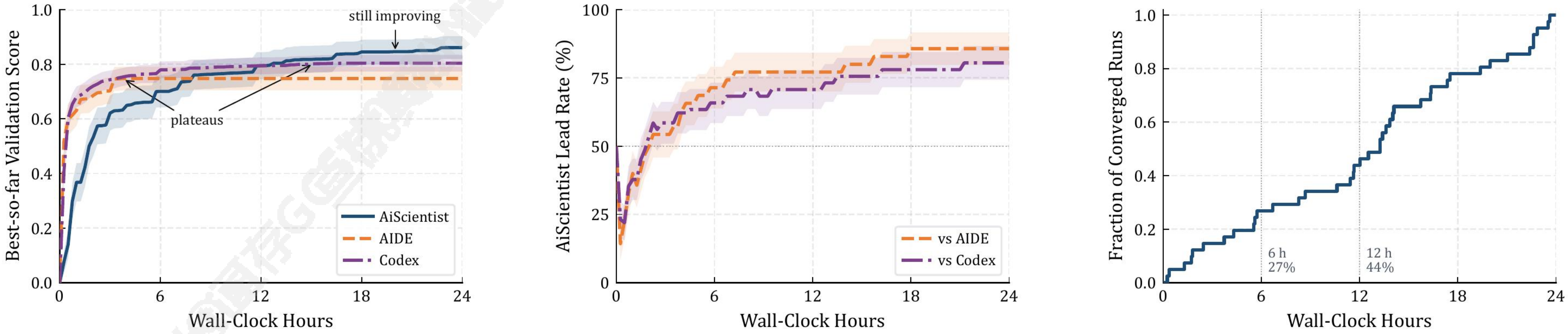
➢优势集中在 competitive outcomes，而不是只生成 valid submission

System	Backbone	Valid	Above Median	Any Medal
AIDE	Gemini-3-Flash	87.88±5.46	59.09±2.62	45.45±0.00
LoongFlow	Gemini-3-Flash	77.27±0.00	77.27±0.00	77.27±0.00
AiScientist	Gemini-3-Flash	100.00±0.00	86.36±0.00	81.82±0.00
ML-Master 2.0	GLM-5	100.00±0.00	80.30±1.52	65.15±1.52
AiScientist	GLM-5	100.00±0.00	89.39±1.52	81.82±0.00

Codex/GPT-5.5 xhigh frontier harness: Any Medal 68.18±2.62; AiScientist exceeds by 13.64 points.

Dynamics：慢启动，但后半程继续涨

➤早期投入理解、规划和 scaffold；优势在后半程逐步扩大

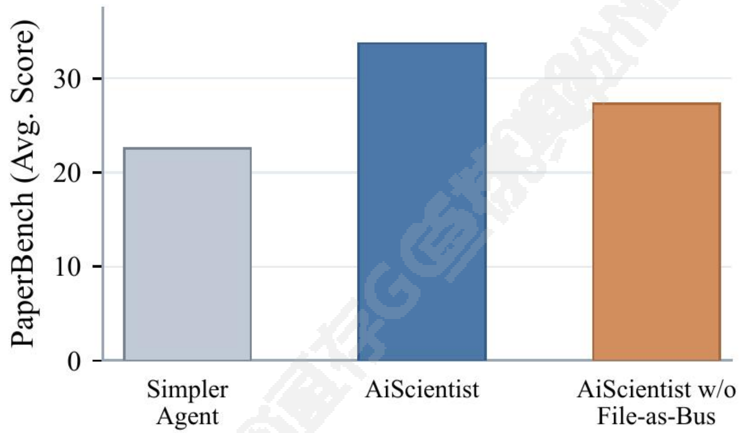


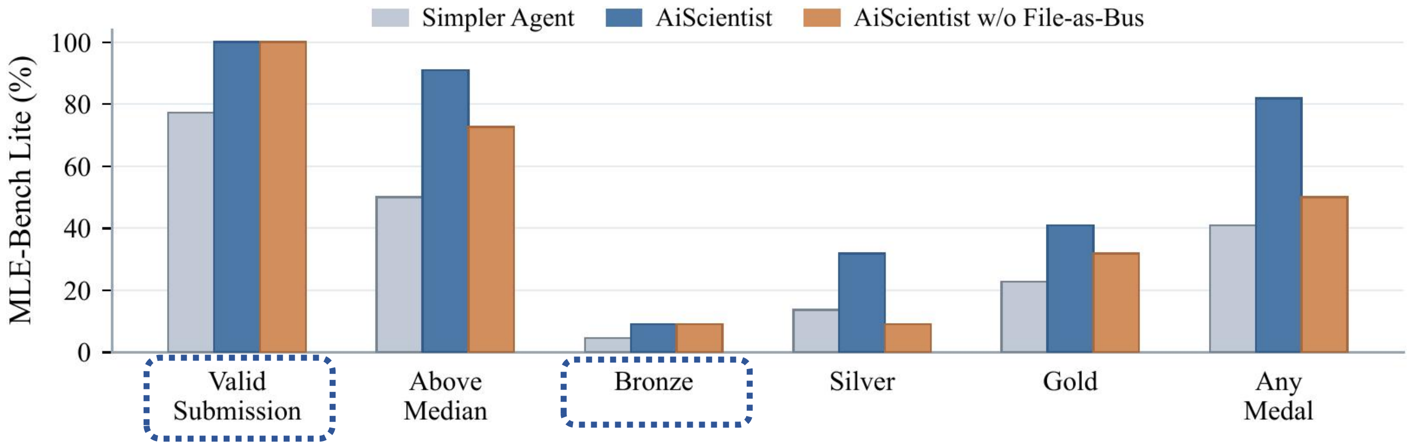
(a) magnitude (b) breadth (c) when final best appears

Insight: long-horizon advantage is not one lucky late task; it broadens across paired runs and strengthens over the full 24h budget.

Ablation：File-as-Bus 支撑后期 refinement

➢去掉 File-as-Bus 后，validity 基本保留，但 medal outcomes 大幅下降





-6.41 pts
PaperBench w/o File-as-Bus

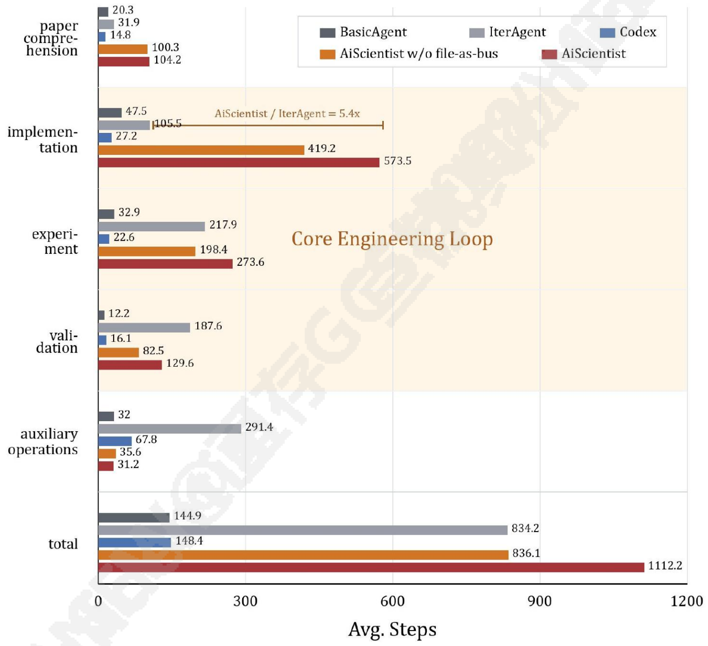
100% valid
MLE validity preserved

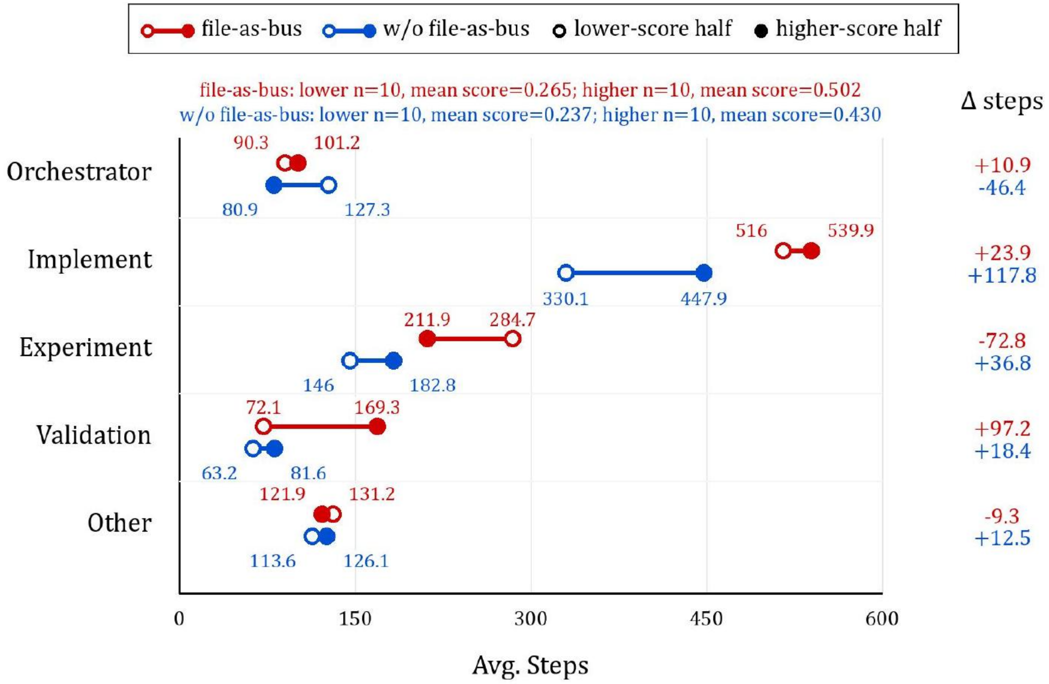
-31.82 pts
MLE Any Medal drop

+22.73 pts
hierarchy helps Above Median

Behavior：成功不是多试，而是更早闭环

➢有 File-as-Bus 时，高分轨迹更早转向 validation/closure



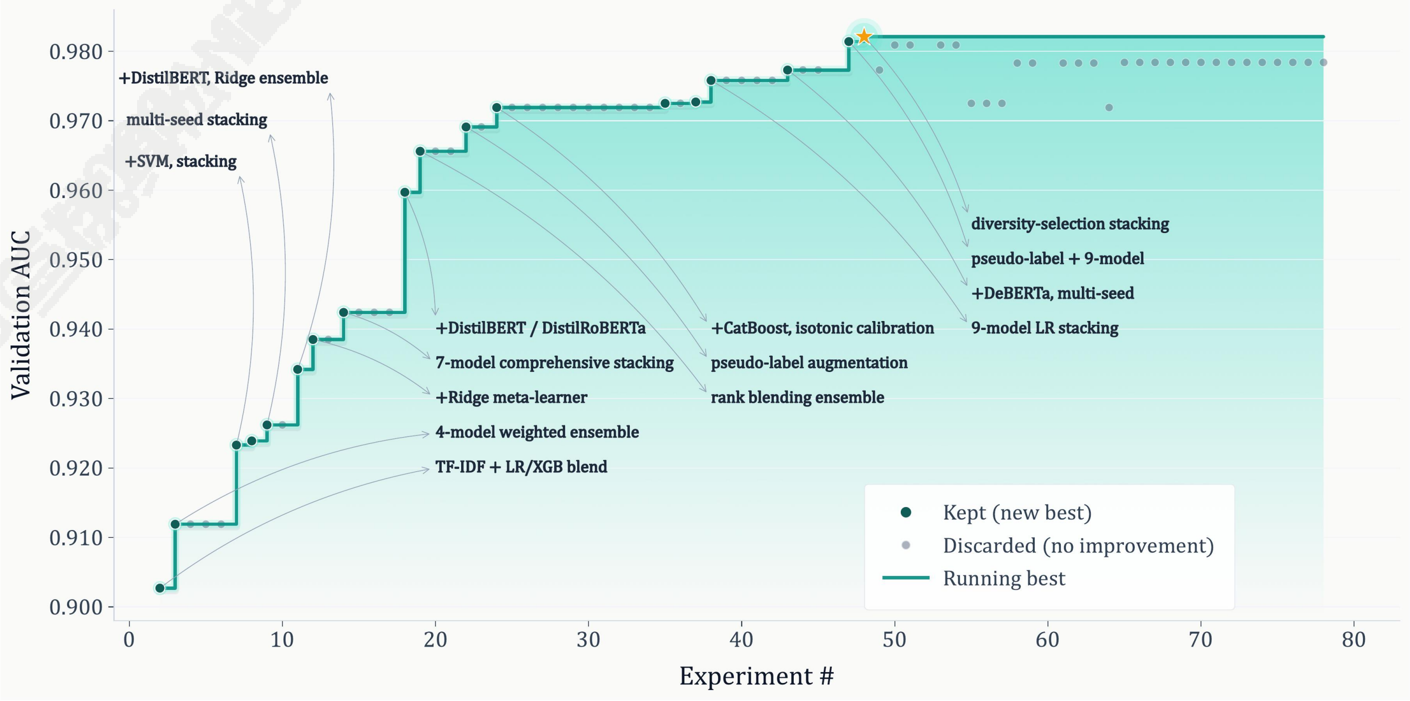


Do not read this as "more steps = better": Codex is a strong low-step reference.
The key question is where effort is spent.

Case：23h 内 74 次实验循环

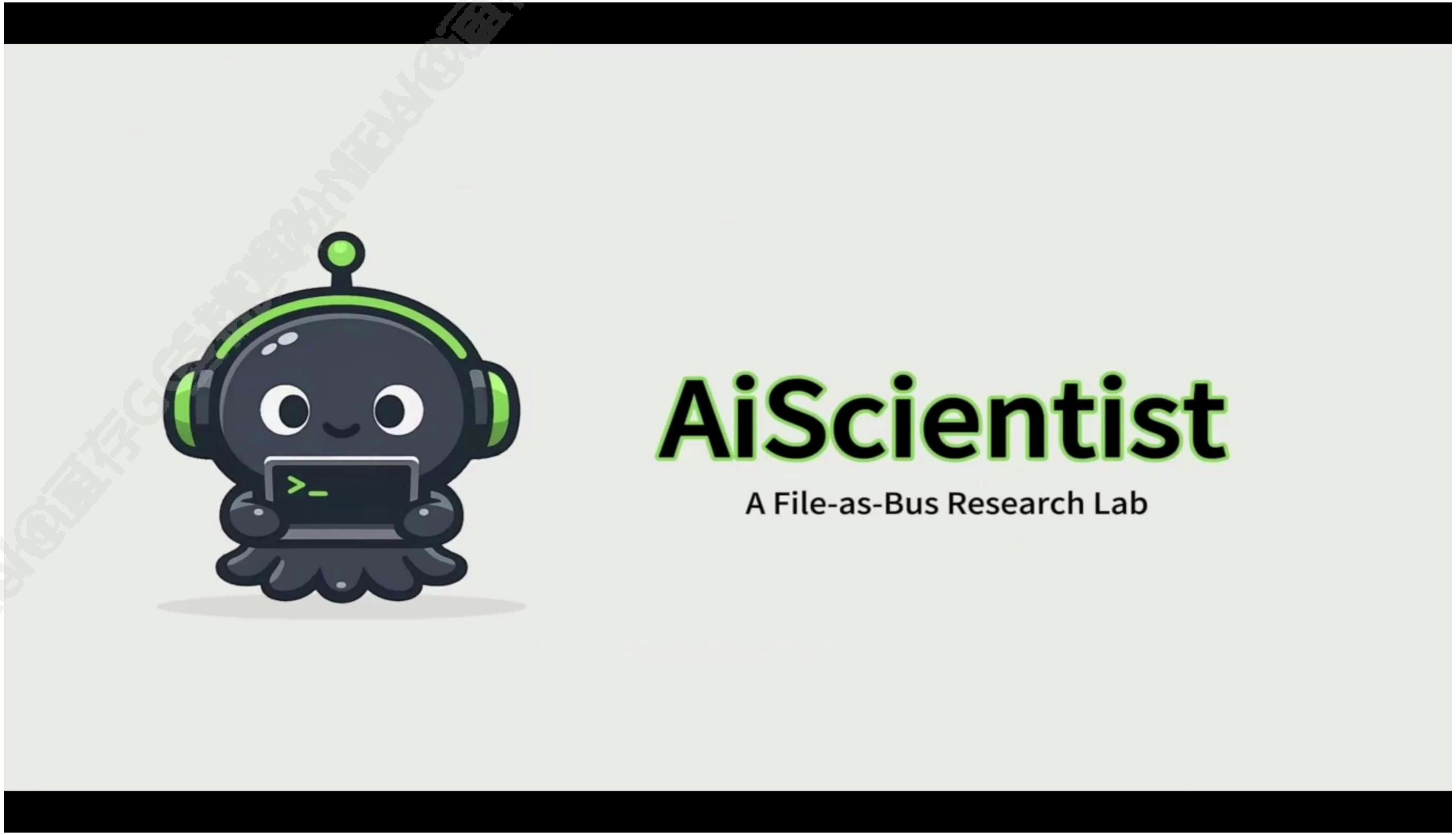
AI Box

➢Detecting Insults：AUC 0.903→0.982，18 次 best-so-far 更新



Case-level lesson: experiment records are memory, not bookkeeping.

Demo



对 multi-agent 系统设计的启示

AI Box

➢如果任务需要跨数小时/数天累积进展，agent framework 应该把 state design 当作 first-class design object

- 把 workspace 设计成 system of record，而不是 chat transcript 的附属品。
- 给 artifacts 明确 purpose、update mode、owner 和 consumer。
- 把 control summary 和 evidence substrate 分离：thin control, thick state。
- 评价 late-budget refinement：只看 first-pass success 会漏掉核心能力。

感谢聆听

Paper

Repo

Demo轻松上手
Paper/MLE bench打榜代码