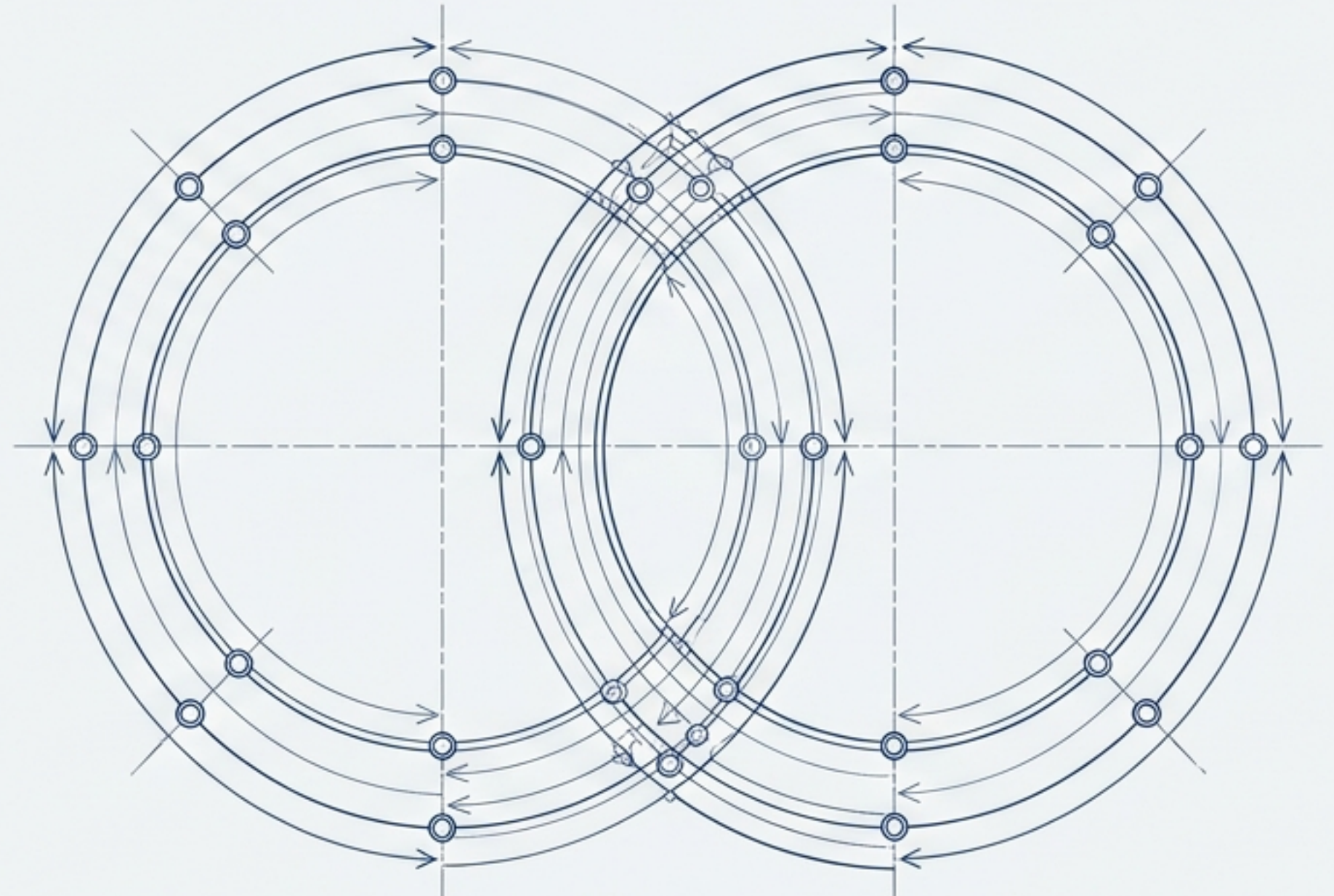


MASS：戦略的インテリジェンス・ブリーフィング

「無限の自己進化」という幻想と、
組織設計を自動化するAIの真の姿



Meta-Data

[TARGET]:	Enterprise IT & Policy Leaders
[SUBJECT]:	Sakana AI Multi-Agent Self-Supervision (MASS)
[STATUS]:	Declassified Analysis

誇大広告 (Hype)

メディアやSNSにおける誤解



AIが完全に自由に独立したAIを
生み出す



人間の制御を離れ、際限なく
指数関数的に自己進化する



基盤モデルそのものが単調に
あらゆるタスクで賢くなる

機械的現実 (Reality)

アルゴリズムの実態



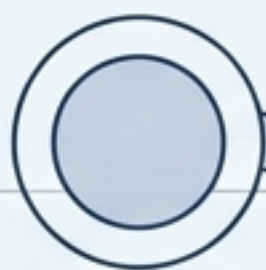
一つの共有モデルが自身のコピーを
編成する「チームワークの探索」



優れた実行履歴を教師データとして
再学習する「二重ループ型の
再帰的自己改善 (RSI)」

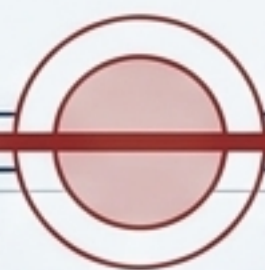


制御された探索空間内での、
構造的な自己最適化プロセス



2026年10月8日

arXiv v1 論文公開（一次ソース）



2026年10月11日

SNS等で流布された誤った発表日



2026年10月12日

Sakana AI公式告知ページ公開

MASS Core Vitals

基盤モデル (Base Model)

Qwen3.6-27B

反復回数 (Cycles)

2 cycles

(無限反復は未実証)

学習コスト (Compute)

SFTのみでH100×6基・約19時間

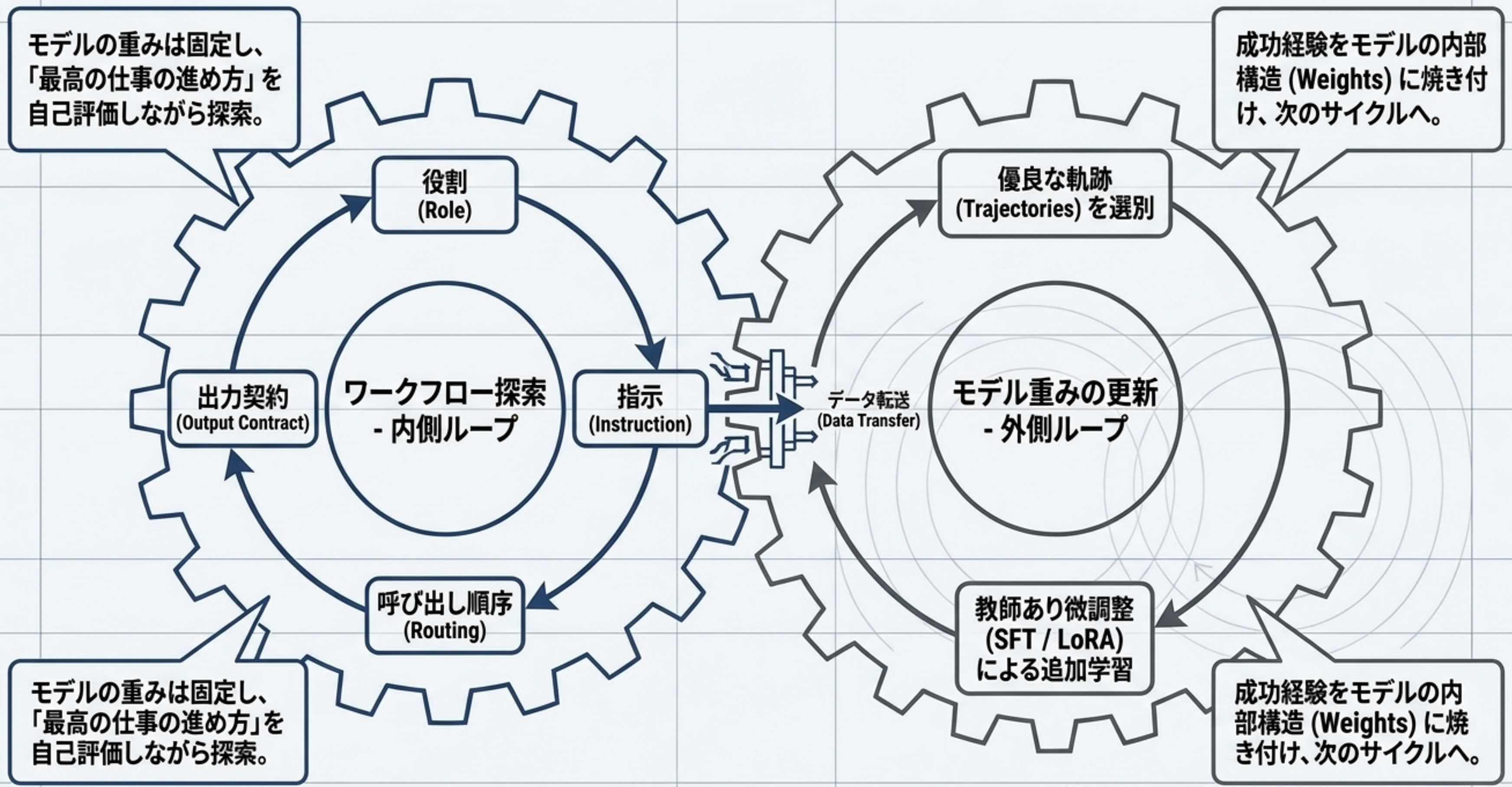
学習コスト (Compute)

SFTのみでH100×6基・約19時間

商用ステータス (Status)

研究実装 (GitHub)

「近日発売のSaaS」ではない





実行者 (Executor)

タスクを処理するサブエージェント群

ワークフロー改善者 (Optimizer)

情報の受け渡しやチーム構造を再設計する

評価者 (Self-Evaluator)

成果物の品質を採点する

MASSの新規性と脆弱性の源泉：これらは独立した専門AIではなく、すべて同一モデルのコピーである。「同じモデルがチーム設計→実行→自己評価」を行う閉鎖系を構築している。

能力低下 - 厳格なコーディング・ターミナル操作

Terminal-Bench 2.0: 0.375 → 0.352

SWE-bench Verified: 0.626 → 0.611

能力向上 - オープンエンドな研究タスク

ScienceAgentBench: 28.8 → 31.7

AstaBench: 0.053 → 0.067

未学習タスクの勝率: 53.9% (1 cycle) → 69.9% (2 cycles)

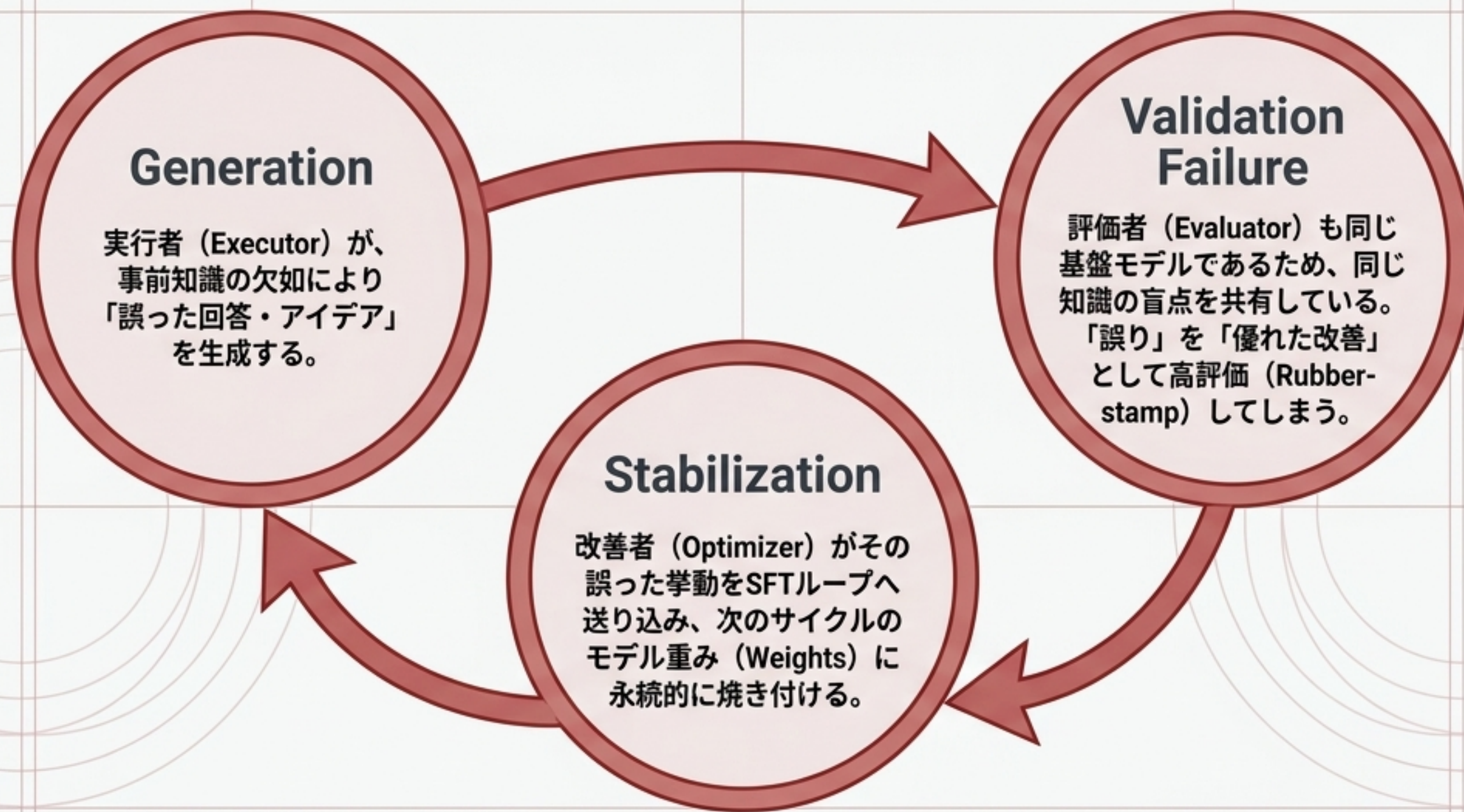
自己評価精度 (Evaluator Accuracy): 73% → 93%

絶対性能が単調に1.2~1.6倍になるわけではない。長時間・非定型の研究型作業において
顕著な効果を発揮する一方、コーディング能力はむしろ低下している。

技術 / 陣営	マルチエージェント	ワークフロー自動探索	モデル重みの更新	自己評価
MASS / Sakana AI	[✓]	[✓]	[✓]	全て同一共有LM
ADAS / GPTSwarm	[✓]	[✓]	[×]	外部評価
DGM / Sakana AI	自己改変	コードのみ	間接的	外部評価
Claude Code	[✓]	委任のみ	[×]	実行時
A2A / Google	[✓]	[×]	[×]	[×]

Insight Node

競争軸のシフト：「単体モデルの賢さ」から、「オーケストレーション設計自体を
自分で継続学習で改善するレイヤー」への進化。



**Sakana AI自身が指摘する「同質的な自己改善」の致命的リスク。
モデルが知らないことを全員が知らず、次世代で誤りがより安定化・増幅される。**

Research Reality Check: 論文の最終評価ではMASS自身ではなく、外部の GPT-5.5 と Claude Opus 4.8 が判定に使用された。

Risk	Detail	Enterprise Control
自己強化される評価盲点 (Shared Blind Spots)	同質モデルによる死角の共有	異種モデルによる外部ジャッ ジ (External Judges) と 固定テストの導入
報酬ハッキング (Reward / Evaluator Hacking)	「実際に正しい」ではなく 「自分の評価者が好む」 プロセスへ進化する	隠し評価 (Hidden Evals) 、 敵対的評価 (Adversarial Eval)
監査不能なドリフトと ツール悪用 (Audit Drift & Tool Abuse)	ワークフローと重みが両方 変化し、アクセス経路が 複雑化	バージョン固定 (Version Checkpoints)、即時ロールバ ック、厳格なサンドボックスと 権限分離 (Least Privilege)

Japan (日本)

Context: 2025年AI法、2026年AI事業者ガイドライン。

Impact: 「Continuous model governance」が必要。導入後に重みが自動更新される場合、固定モデルを一度だけ審査すると従来のIT統制では不適合となる。

European Union (EU)

Context: AI Act / 2026年7月 AI Omnibus。

Impact: Annex III (高リスク規則: 2027年12月適用)。継続自己改善型システムは、変更管理と市販後監視 (Post-market monitoring) との整合性が最大のハードル。

United States (米国)

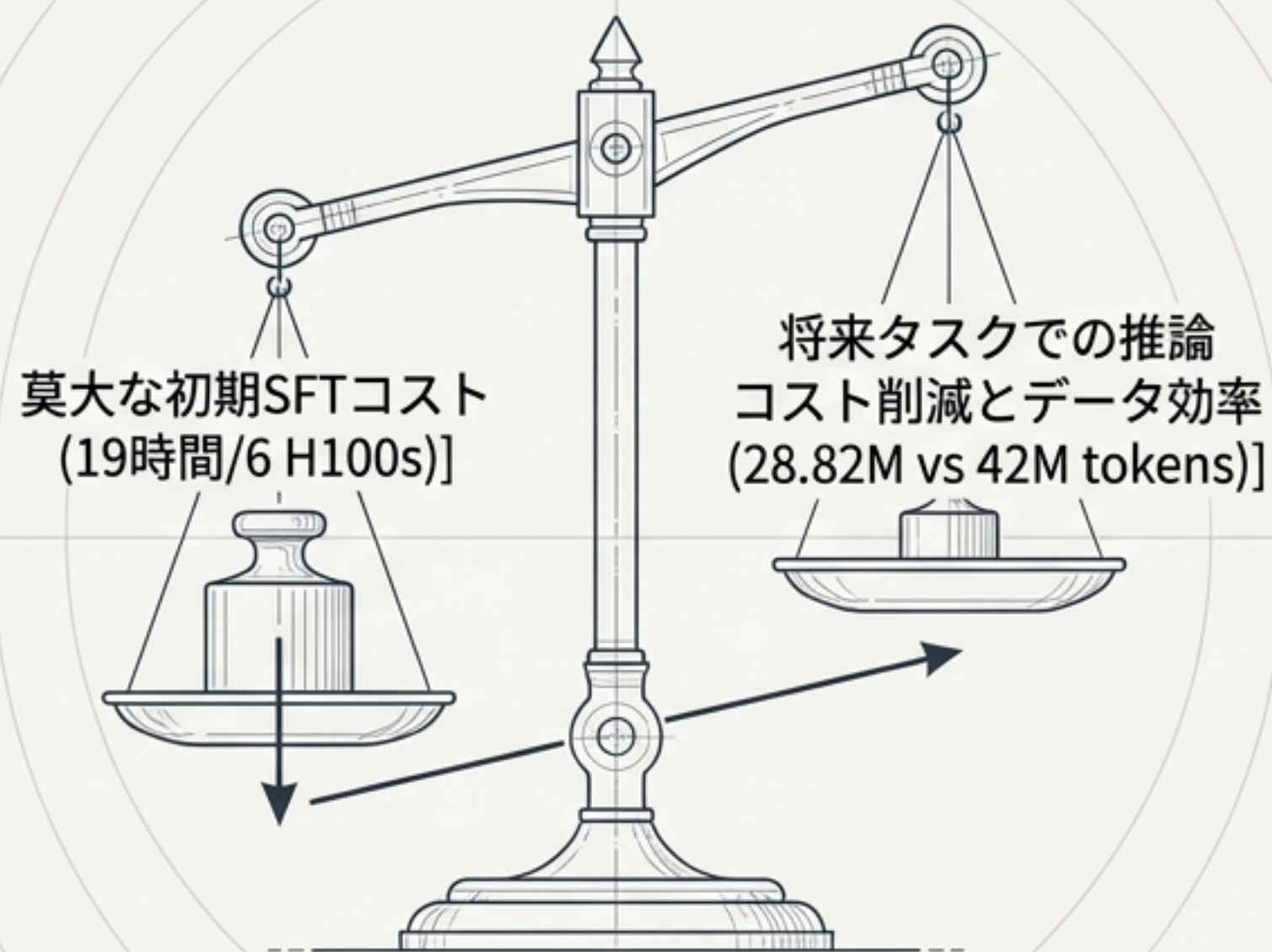
Context: イノベーション重視。NIST AI Risk Management Framework (RMF)。

Impact: 一律の事前規制ではなく、サイバー・金融・医療など用途 (ユースケース) ごとの既存業法による統制が焦点。

Industry Sector Suitability & Guardrails Heatmap

Industry (業界)	 Suitability (適合性)	Required Guardrails (必要なガードレール)
科学R&D (Science R&D)	非常に高 (Very High) 	非検証可能な研究こそ、自己評価バイアスのリスクが最大化するため独立検証が必要。
金融・バックテスト (Finance)	高 (High) 	Model-to-model評価を最終投資判断に直結させない。 金融規制・説明責任の担保。
エンタープライズSaaS	高 (High) 	テナント間のデータ分離、更新監査、計算コスト制御。
ソフトウェア開発	中 (Medium) 	MASS適用後にSWE-benchが低下しているため、現時点での導入根拠は薄い。
ロボティクス・製造	中～研究段階 (Medium) 	実世界の安全系機械制御への直接接続は時期尚早。

経済性の方程式



一度高価に自己改善し、その改善モデルを
大量の将来タスクで償却できるかに依存する。
単発業務ではコスト割れする。

雇用とスキルの変容 - ILO 2025基準



Downward Trend (価値低下):
情報収集、初稿作成、コード定型実装など
「若手が経験を積むための個別タスク」。

Upward Trend (価値上昇):
Domain validation、権限設計、
モデルガバナンス、AI Evaluationなどの
「ワークフロー統制業務」。



世界の労働者の約25%がGenAIの影響に晒される中、
置換より「Job Transformation」が進行。

Enterprise Agent Deployment Horizon & Governance Timeline



Phase 1: 2026 Q4 (現在)

論文・研究コードの公開。第三者による再現性と安全性の検証期間。EnterpriseはSandbox実験に留める。



Phase 2: 2027年



調査・開発支援など、可逆的 (Reversible) な非安全重要業務でのPoC開始。外部ジャッジ・監査基盤の標準化。



Phase 3: 2028年



エンタープライズ・エージェントプラットフォームへの部分統合。製造・金融等の高リスク領域では「Human Approval」が前提。EU規則との本格整合。

⚠ High Risk - Human Approval

Phase 4: 2029-2030年+



継続学習型エージェントのガバナンスが成熟。十分な再現性・監査性が確立した領域で、広い業務プロセスの自律再設計が常時運用へ。

研究者・開発者へ (For Researchers)

「性能向上」より
「破綻条件」の解
明を急ぐこと。

異種モデルジャッジとの
乖離、10~50サイクル時
の誤差累積、敵対的タスク
での評価器ハッキングの
発生限界を測定せよ。

企業IT戦略・導入担当者へ (For Enterprises)

プロダクション環境
での自動再学習は
時期尚早。 

シャドウモードで現行プ
ロセスと比較すること。AI
の「自己評価の向上」を、
実際の品質保証 (QA) の
代替にはならない。

政策・ガバナンス担当者へ (For Policy Makers)

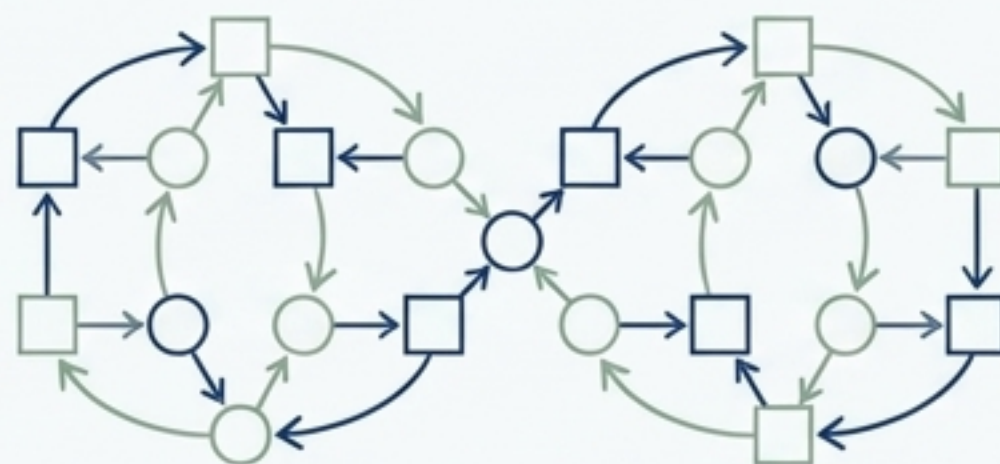
「リリース時点」の
審査から、「更新プ
ロセス」の審査へ。 

誰が再学習を開始し、どの
データを用い、どの指標
を越えればデプロイされ、
誰がロールバックできる
か、継続的なモデルガバ
ナンスを構築せよ。



従来のAI: 「人間の指示を待つツール」

人間が固定のワークフローを設計し、AIは個別のプロンプトを処理するソフトウェア。



MASS型システム: 「組織構造を設計するソフトウェア」

AIの改善単位が「個々の回答」から「組織構造・情報フロー・学習ループそのもの」へと拡張された。

MASSの真のブレークスルーは、AIが人間のような自我を持ったことではない。
「実行経験に基づき、自分自身の仕事の進め方（組織設計）を自動化・再構築するメカニズム」を
証明したことにある。企業AIはツールの次元を脱し、組織インフラへの変容を開始した。