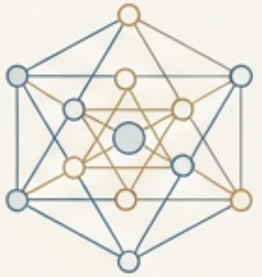
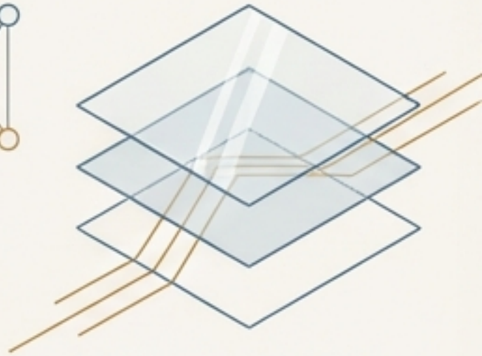
Gemini 3の心臓部：AgenticRLが拓く、次世代AIエージェントの世界

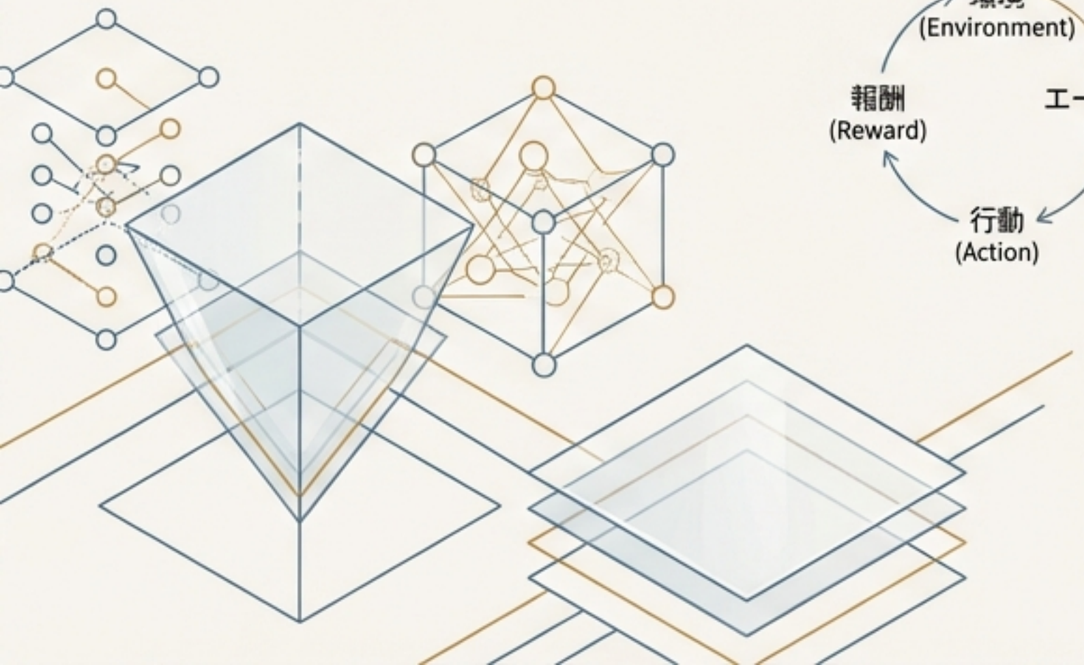
Gemini 3 Flashの画期的な性能を支える技術と、Gemini 3 Proとの比較分析

本プレゼンテーションの要点と構成



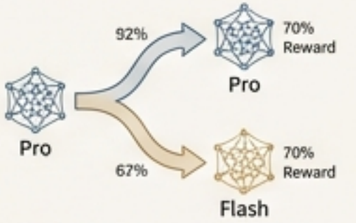


Gemini 3 Flashは、「**Agentic RL**」という新たな強化学習パラダイムを活用することで、フラッグシップモデルであるProに迫る知能を、画期的な速度と効率で実現しました。この技術は、AIが自律的に複数ステップのタスクを計画・実行する「AIエージェント」の実用化を加速させます。



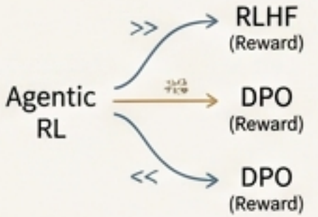
1 2つのモデル (The Models)

Gemini 3 ProとFlashの戦略的な位置づけと性能比較。



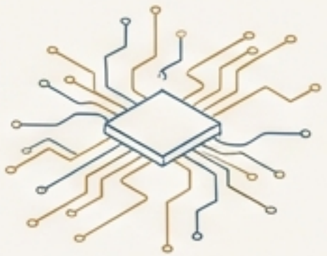
2 核心技術 (The Engine)

Flashの性能を支えるAgentic RLの理論と実装を深掘り。

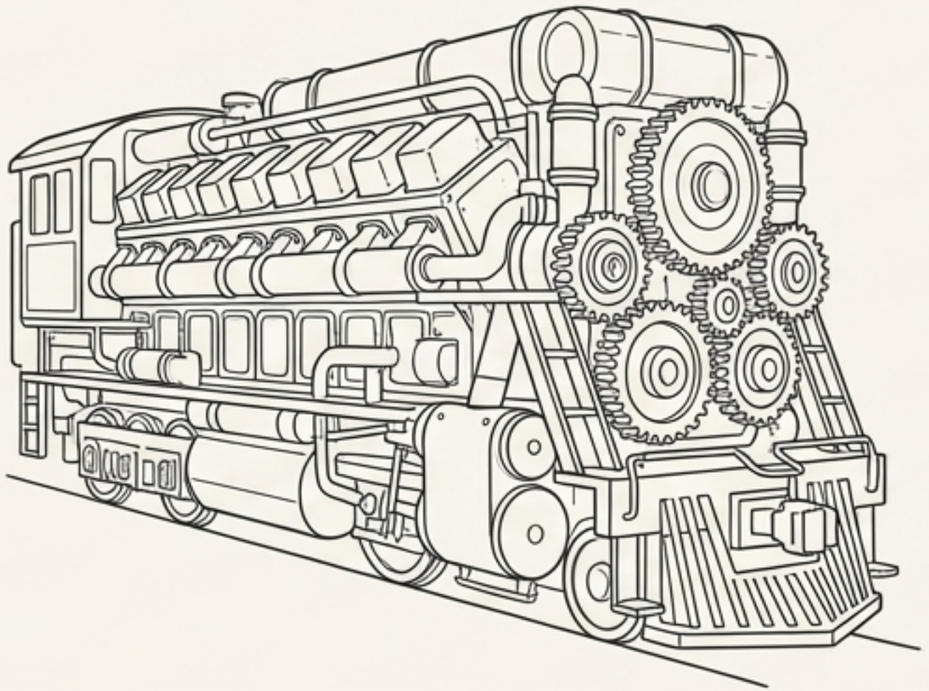


3 技術的文脈 (The Methods)

Agentic RLを、RLHFやDPOといった他の手法と比較分析。



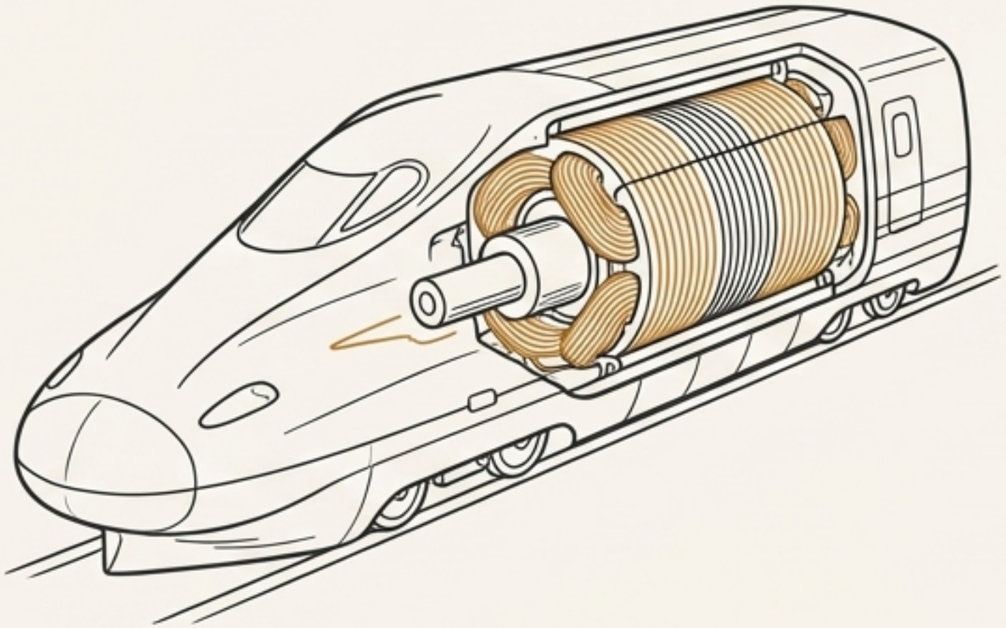
The Gemini 3 Family: 究極の能力 vs. 圧倒的な速度



Gemini 3 Pro

役割: 最高性能を追求したフラッグシップモデル。極めて複雑な推論やマルチモーダル理解が求められるタスクに最適。

キーワード: 高度な推論、フロンティア性能、包括的知識



Gemini 3 Flash

役割: Proの知能を維持しつつ、速度・コスト・効率を極限まで高めたモデル。リアルタイム性が求められる大規模アプリケーションに最適。

キーワード: 高速最適化、低遅延、スケーラビリティ

直接対決: Gemini 3 Pro vs. Flash

「Pro級の知能を、圧倒的な効率で」

	Gemini 3 Pro		Gemini 3 Flash
推論性能	最高水準の応答品質。 複雑なタスクで最先端の精度。 レイテンシは高め。		Pro級の知能を維持しつつ高速。 2.5 Pro比で約3倍高速。 低遅延でほぼリアルタイム応答が可能。
学習効率	大規模学習による最高精度志向。 トレーニングコストは非常に高い。		高効率チューニング。 少ない追加学習でPro級の能力を達成。 iStarなどサンプル効率の高いRL手法を活用。
マルチモーダル能力	最高クラスの理解力。ドキュメント理解や動画要約でSOTA性能。		Proと同等級の精度を維持しつつリアルタイム化。コード実行による画像解析など先進機能も搭載。
ツール使用能力	高度なツール呼び出し能力。 Terminal-Bench 2.0で54.2%のスコア。 慎重に計画・実行する傾向。		ツール使用に特化し軽快な操作。特にコーディングエージェント用途で性能が顕著に向上。 低思考レベルでも連携が破綻しにくい。
エージェント的行動	長大なタスクの遂行力に優れる。 ステップ毎の時間は長い。		リアルタイムエージェント動作に最適化。 高速な思考・行動サイクルで、商用サービスでスケーラブルに運用可能。

注記: Flashの推論コストはProの4分の1以下。

技術的ブレークスルー: AIに「自律性」を与えるAgentic RL

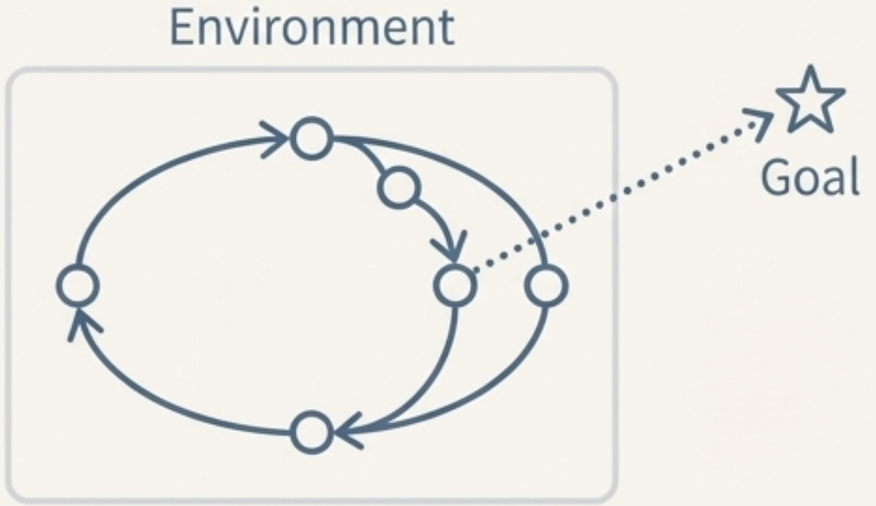
Agentic RLとは

大規模言語モデル（LLM）に、エージェントのような**自主的行動能力**を持たせるための強化学習手法。

従来のRLHF

構造：静的な「1ターン」の出力に対する人間の評価（好き/嫌い）でモデルを微調整。
目的：より人間に好まれる応答を生成する。

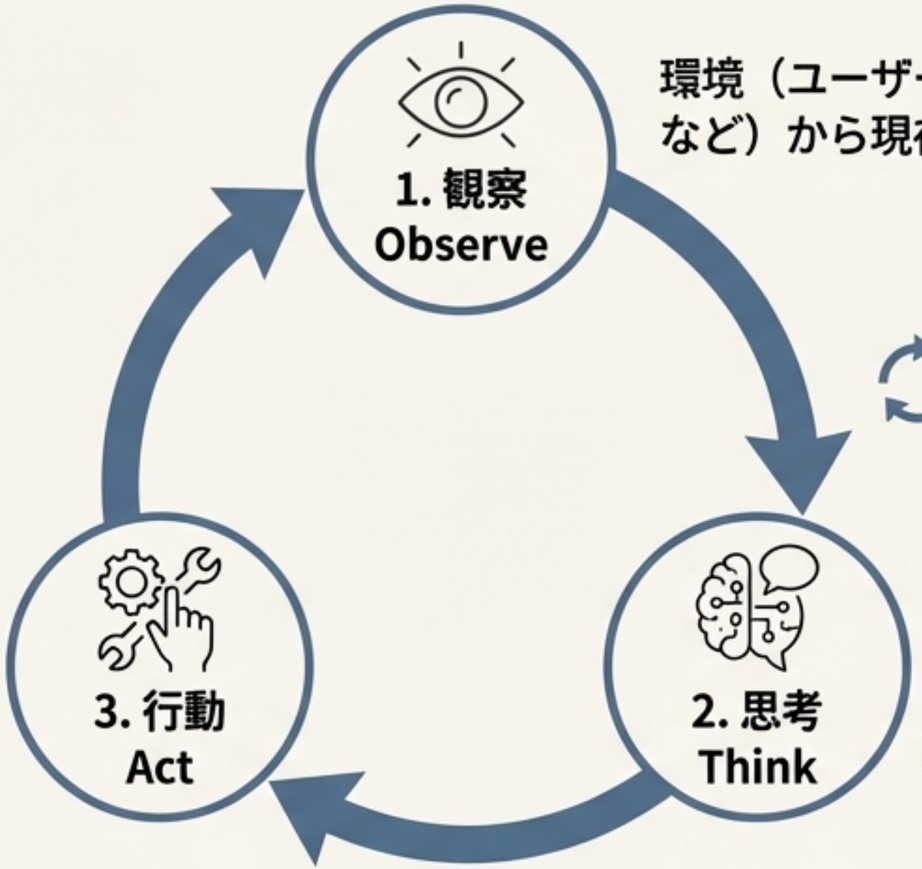
Agentic RL



構造：モデルが対話的な環境で**連続的に意思決定**を行う。
目的：長いシーケンス（マルチステップ）で、長期的なゴールを達成する。

The Agentic Loop: 観察 → 思考 → 行動

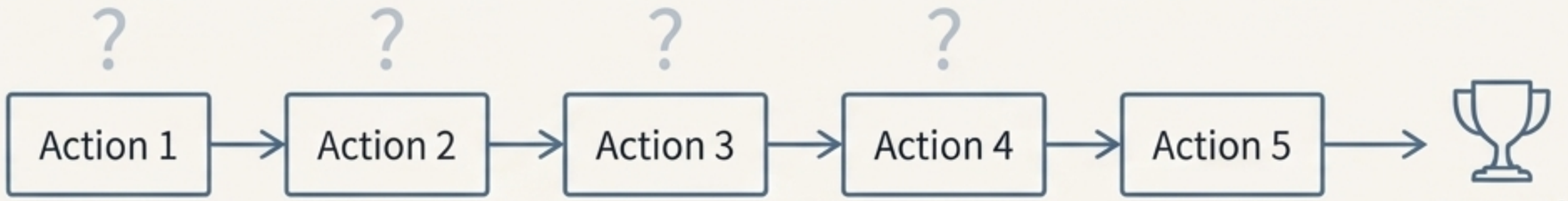
Agentic RLで訓練されたモデルは、長期的な目標達成のために、以下の「エージェント的ループ」を自律的に繰り返します。



ポイント: この思考と行動の二段構えの構造により、LLMはより長い文脈で計画立案と実行が可能になります。

核心的課題: 長期的な意思決定と「クレジット割り当て」問題

エージェントが一連の行動の末に報酬（成功/失敗）を得る場合、どの行動が最終的な成功に本当に寄与したのかを特定するのは極めて困難です。これを**クレジット割り当て（credit assignment）問題**と呼びます。



「どのステップが『決定打』だったのか？」

Agentic RLにおける解決策

- **プロセススーパービジョン :** 最終結果だけでなく、途中経過の良し悪しを評価に組み込む。
- **Implicit Step Reward (iStar) :** 軌跡全体の優劣から、各ステップへの暗黙的な部分報酬を割り振り、学習信号を逐次的に与える。

学習アーキテクチャ: Agenticモデルはいかにして学習するのか

Stage 1: 教師あり微調整（SFT)

内容: エージェントタスクのデモデータを用いて、基本的な振る舞い（ツールの使い方、思考様式）を学習させる。

↓

Stage 2: Agentic RLループ

内容: モデル（ポリシー）が環境と対話し、軌跡を生成。報酬モデルが軌跡を評価し、各ステップに報酬を割り振る。
PPOなどの方策勾配法でモデルを更新。

キーアルゴリズム: iStar（Implicit Step Reward）- 軌跡ペアに対するDPOベースのロスで報酬モデルを学習し、自己強化ループを構築。

↓

Stage 3: 反復的改善

内容: ステージ2を繰り返すことで、方策（LLM）と報酬モデルが相互に改善し合い、性能が向上する。

補足: DPO、GRPO、RLOOなど、LLM向けに調整された最新の最適化手法を組み合わせることで、学習の安定性と効率性を高めている。

アーキテクチャの革新:「思考」と「行動」の分離

Gemini 3では、モデルの内部的な「思考プロセス」と、外部に出力される「行動」を区別するためのアーキテクチャ拡張が実装されています。

革新①: 隠れチェーン・オブ・ソート (Hidden Chain-of-Thought)

- **仕組み: `Thought Signature`と呼ばれる特殊トークン（例: `<thought>...</thought>`）を使用。モデルは内部で思考内容を生成するが、この部分はユーザーには表示されず、次の思考ステップの入力として利用される。**
- **効果: モデル自身が中間ステップを展開し、より複雑な推論が可能に。誤った思考プロセスをユーザーに晒さない安全性を確保。**

```
PROMPT: "今日の東京の天気と、傘が必要か教えて。"
<thought>まず、東京の天気を検索するツールを呼び出す。次に、その結果から降水確率を確認し、傘が必要かどうかを判断する。</thought>

RESPONSE: "今日の東京は晴れのち曇り、降水確率は10%です。傘は必要ないでしょう。"
```

革新②: 高度な関数呼び出し (Advanced Function Calling)

- **仕組み: モデルの出力が、特定のツールをどの引数で呼び出すかを示す構造化されたJSON形式をサポート。**
- **効果: Google 検索、コード実行、URL読み込みといった外部ツールをAPIとして自律的に利用し、自身の知識を補強・検証する。**

```
MODEL OUTPUT:
{
    "function": "google_search",
    "args": {
        "query": "Agentic RLとは"
    }
}
```

理論から現実へ: Googleのエージェント・エコシステム

Agentic RLは研究に留まらず、Google内外のスケーラブルなプロダクトとして展開されています。

Googleのプラットフォーム

Google Antigravity

開発者が対話的にエージェントを構築・テストできる新しい開発環境。

Vertex AI Agent Builder

複雑な業務フローを自動化するエンタープライズ向けエージェント実行基盤。

パートナーによる活用事例 (社会的な証明)

JetBrains

「コーディングAIアシスタントにFlashを採用することで、**Pro同等の賢さでレイテンシを大幅短縮**し、安定したマルチステップAgentループを提供できるようになった。」

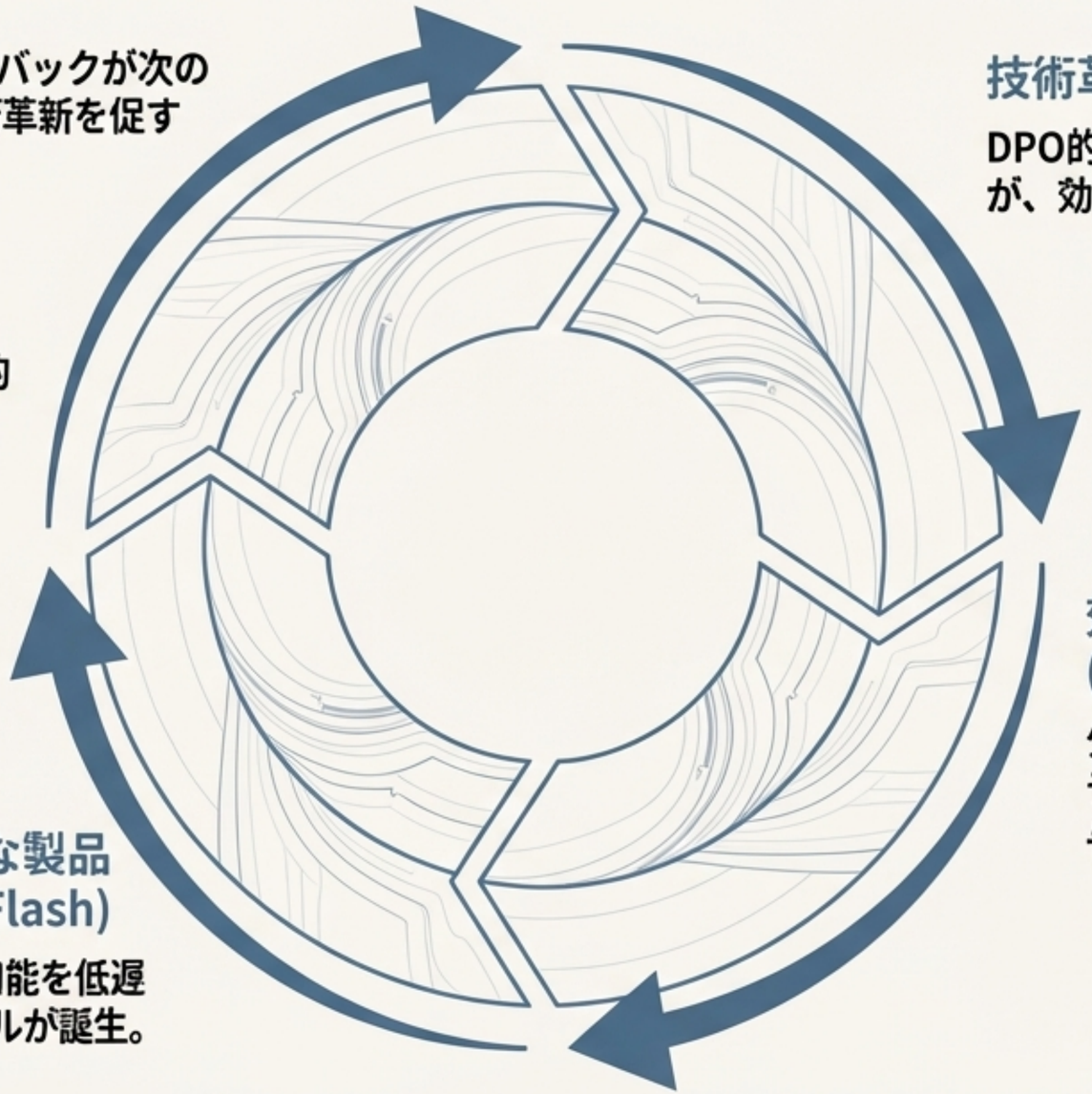
Replit

「Flashは十分な性能と高速性を兼ね備え、**初めてコーディングエージェントの中核ループを実用レベルで駆動できるモデル**だ。特にツール使用能力とコーディングスキルが非常に強力だ。」

手法の比較: Agentic RL vs. RLHF vs. DPO

観点	Agentic RL	RLHF	DPO
学習構造	環境との対話を通じて**マルチターンの軌跡**を生成し、軌跡単位で学習。逐次決定プロセス。	**単発の出力**に対する人間の好みデータで報酬モデルを学習し、RLで最適化。3段階構成。	人間の選好データを用いて、**1段階で方策を直接最適化**。RLステップを省略。
目的関数	**長期的・エピソード全体**の報酬（最終報酬＋ステップ報酬）を最大化。	**人間の好みスコア**を最大化。報酬モデルの出力を高める。	人間の**選好確率**を直接最大化。対数確率差として定式化。
スケーラビリティ	環境とのインタラクションが多く**計算コストが高い**。iStar等で効率は改善されているが、依然として大規模。	**実績はあるが複雑**。3段階のパイプラインとハイパーパラメータ調整が技術的負債に。	**シンプルでスケーラブル**。通常のファインチューニングに近い手間で実行可能。計算コストが低い。
フィードバック	環境からの**自動報酬**（成功/失敗）が中心。必要に応じて人間の介入も可能だが、ラベル効率の向上を目指す。	**人間による詳細なフィードバック**（スコア、ランク付け）が中心。	人間による**二者択一の比較データ**のみを使用。シンプルだが表現力は限定的。

The Virtuous Cycle: すべての要素がいかにして繋がるか



フィードバックが次の
技術革新を促す

技術革新 (Agentic RL + DPO)

DPO的な最適化を組み込んだAgentic RLが、効率的で安定した学習を実現。

効率的な訓練
(Efficient Training)

少ない計算コストとサンプルで、モデルに高度なエージェント能力を付与。

画期的な製品
(Gemini 3 Flash)

効率的な訓練により、Pro級の知能を低遅延・低コストで提供するモデルが誕生。

新たな応用
(New Class of Real-Time Agents)

Flashの登場により、これまで実用的でなかったリアルタイムAIエージェント（コーディング支援、対話型サポート）が大規模に展開可能に。

Key Takeaways

1

Agentic RLは、自律的に複数ステップのタスクを遂行するAIの鍵である。

従来の単発応答型モデルとは一線を画し、AIに計画・実行能力を与える次世代の強化学習パラダイムです。

2

Gemini 3 Flashは、この技術を適用し、Pro級の知能を前例のない速度とコストで提供する。

Flashは単なる高速モデルではなく、Agentic RLによる効率的な学習アーキテクチャの賜物です。

3

この革新は、実用的なリアルタイム・エージェントアプリケーションを大規模に解放する。

コーディングアシスタントから企業の業務自動化まで、これまで不可能だった高度なAIエージェントの社会実装を加速させます。

参考文献と資料

主要論文

- Agentic Reinforcement Learning with Implicit Step Rewards (iStar)
- Simplifying Alignment: From RLHF to Direct Preference Optimization (DPO)

公式ドキュメント＆ブログ

- Gemini 3 Flash | Generative AI on Vertex AI | Google Cloud Documentation
- Gemini 3 Developer Guide | Google AI for Developers
- Build with Gemini 3 Flash: frontier intelligence that scales with you

Q&A

ご清聴ありがとうございました。