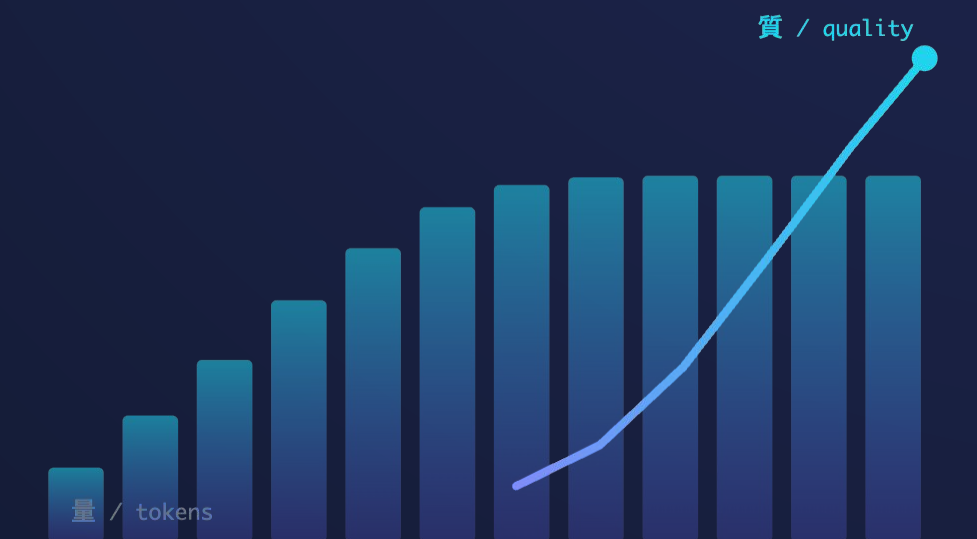


経済産業省 AX時代におけるスキルのあり方検討WG 第3回

トークンマキシムの次、 トークンミニマムの考え方

「AI協働力」が問われる時代へ

葛岡 宏祐 | 株式会社ハイヤールー 代表取締役



PRESENTER



葛岡 宏祐

株式会社ハイヤールー 代表取締役

2018 —



DeNA / AI Research Engineer

自動運転向けのAI（3D空間認識・タクシー乗客需要予測）を研究開発。大手自動車メーカーとの共同研究を本案件化。



2020 —



メルカリ / AI画像検索 テックリード

「写真からさがす」——メルカリのAI画像検索を牽引。秒間100リクエスト超の大規模システムにAIを実践導入。



2020.12 —



ハイヤールー / 創業者・代表取締役

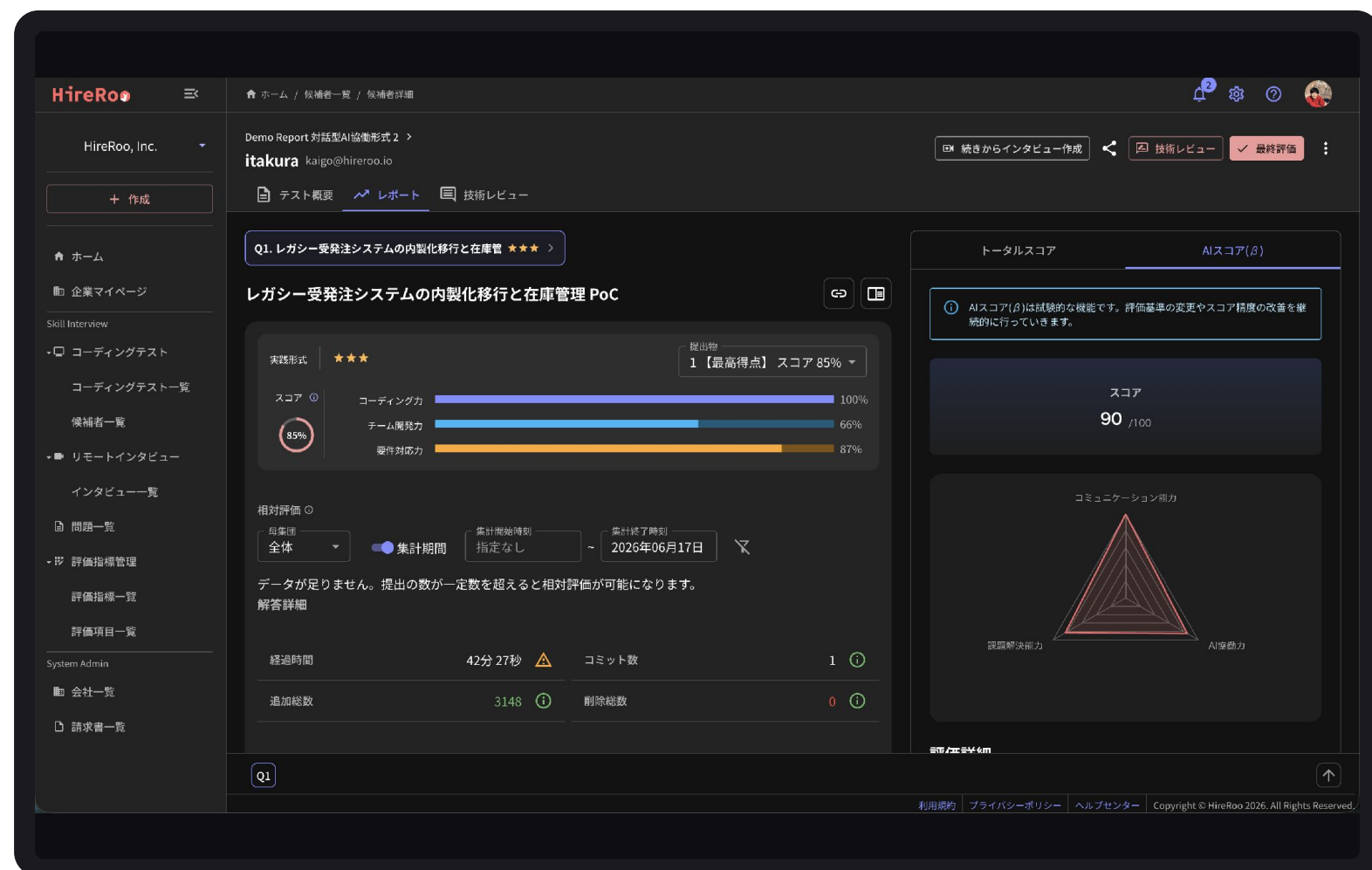
AI時代のエンジニアスキルを測定するプラットフォームを開発。累積300社超に導入。

AI研究 × 大規模AI実装 × プロダクトの立ち上げ — 3つの実務経験

AIと協働して課題を解く力は、測れる。

採用「時」に使う

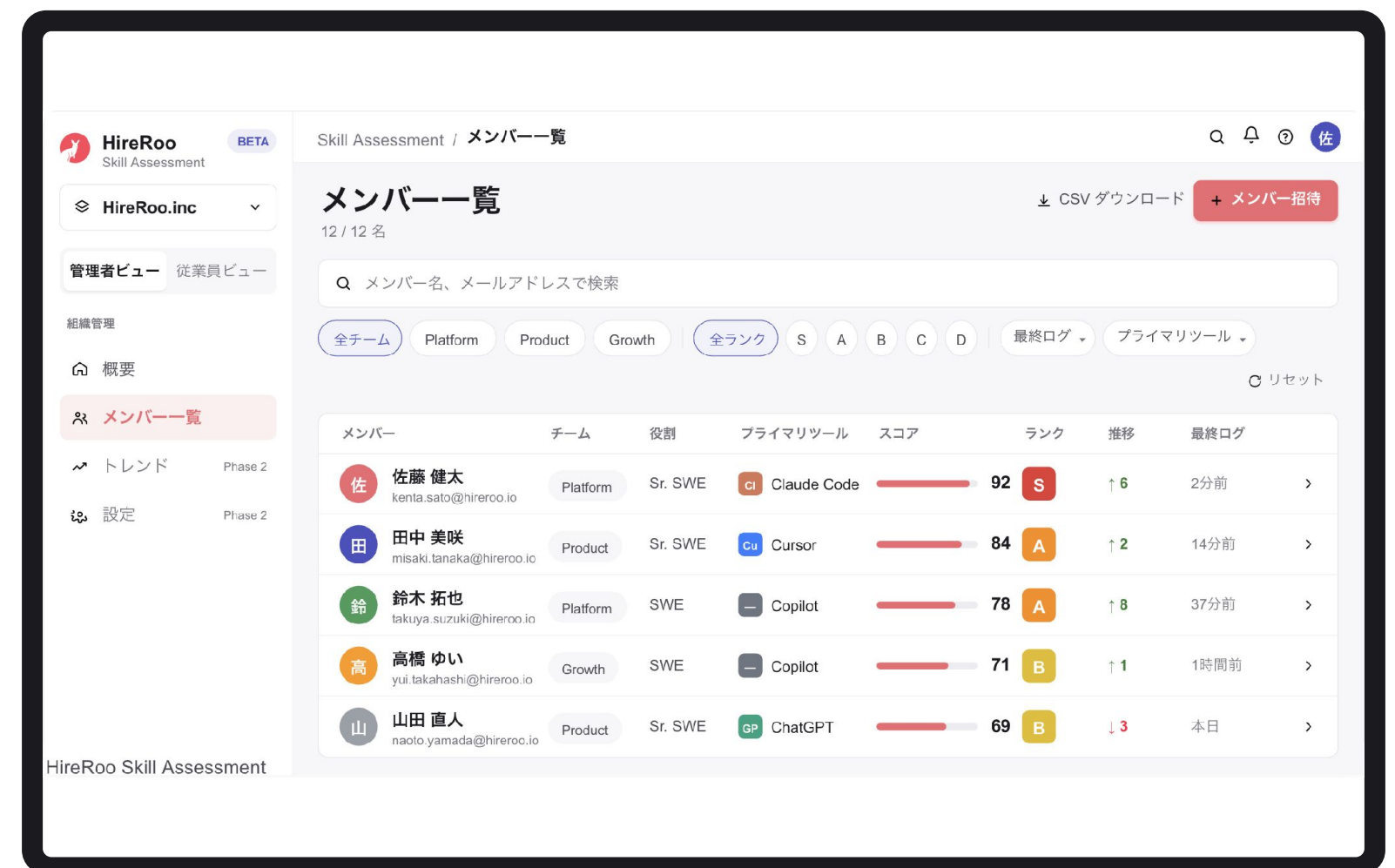
Skill Interview 累計300社+



実践形式で、候補者がAIと協働して課題を解けるかを見極める。

採用「後」に使う

Skill Assessment 30社限定β



組織に入ったメンバーのAI協働スキルを可視化し、伸ばす。

採用の入口から、入った後の組織づくりまで——AI時代のスキルを、一気通貫で。

AI時代の採用テストを、再定義する。

スタートアップ・上場企業

 GREE

 SmartHR

 DMM.com

 GMO
INTERNET GROUP

 freee

 Timee

弁護士ドットコム

 Demaecan

 RevComm

 M3, Inc.

 kakaku.com

 KINTO
Technologies

 dwango

 PKSHA
TECHNOLOGY

 FLYWHEEL

 coconala

 @nifty

 SMS

 REAZON
HOLDINGS

 GA TECHNOLOGIES

SES・受託開発企業

 SHIFT

 NTT Data
株式会社NTTデータ ニューソン

 SOLIZE

 Simplex

累積導入企業

300+

社以上の導入実績

01

トークンマキシム

——熱狂の時代



あるエンジニアは**年収相当のトークン**を使い、
10倍の成果を出した

— Andrew Bosworth Meta CTO

Meta: Claudeonomics — 震源地

Pragmatic Engineer (Gergely Orosz) · 2026.04.23

85,000人

社員を対象にランキング化

60.2兆

トークン / 月 を消費

\$900M

API換算 (実質 \$100M 超)

称号化 (上位250人)

Token Legend / Session Immortal / Cache Wizard
—— 消費量がステータスに

「ダッシュボードが見られるので、意図的にトークンを増やす」
—— 本質ではない最適化、コスト高が課題感に

▶ Goodhart の法則：指標が目標化すると壊れる — LOC → ストーリーポイント → コミット数 → トークン

「測りやすいものを、**測るべきもの**と勘違いした」
— たくさん使うことが優秀さの証だった

Meta 社内リーダーボード "**Claudeonomics**"

¥1,500億 / 月消費

トップ消費者 **single engineer**

¥2.1億 / 月消費

かつての LOC 問題の再来

~2000s

コード行数（LOC）で生産性を測る

→ 冗長なコードを量産しただけ

2025-26

トークン量で生産性を測る

→ 同じ轍を踏むのか？

測定対象を取り違えると、人はその指標を最適化してしまう

02

崩壊と転換

——消費量は、成果に比例しなかった

「使っている」だけでは、頭打ち

90%

のエンジニアがすでに AI を利用

80%

が効果を体感している

しかし
↓

+20%

の生産性向上にとどまる

= 「使う」だけでは頭打ち

消費量 $\neq$ 生産性

タスク 完了率

+34%

↑ 唯一の前進

バグ数 / 開発者

+54%

↑ 悪化

コード レビュー時間

5倍

↑ 悪化

コード チャーン(書き直し率)

+861%

↑ 最大の警告

出典：Faros AI、22,000人調査（2026年 Q1）

熱狂は、現場で噴出した

Uber	2026年のAI予算を4ヶ月で枯渇 COO「成果との連動が見えない」
Meta	リーダーボードが炎上・削除 エージェントをアイドル起動してランキングを操作
Microsoft	複数部門で Claude Code をキャンセル

2026 ・ 06

財布の紐が締まった

「"青天井サブスク"時代の終焉」

Anthropic

2026.06.15

- Background Process が **別枠課金** へ
- 月 **\$200** クレジット上限 → 超過分は API 単価
- Pro が享受した **\$200** 相当の補助が打ち切り

定額制



従量課金

GitHub

2026.06.01

- Copilot が定額制 → **消費ベース従量課金** へ
- あるエンジニアの月額請求が急騰：

\$67 → **\$966**



将来AIは電気や水のように、
使った分だけ払うモデルになる

— Sam Altman OpenAI CEO

QUESTION — 問い

水道の蛇口、出しっぱなしにしますか？

AIも、電気や水と同じ。使った分だけ、
コストは流れ続ける。

01 これからは「何に・どう使ったか」を追う

02 問われるのは量ではなく、質



03

トークンミニマム

—— 質が問われる時代



不要なトークンは、企業の生産性への**直接課税**だ

— Aaron Cheah AI アーキテクト

量から、質へ

01 ~~トーンは「たくさん使うもの」~~

→ 「最小限で価値を出すもの」

02 ~~「AIを使っている」~~

→ 「AIを使いこなしている」

03 ~~「AI活用量の証明」~~

→ 「AI協働の質の証明」

Uber：予算枯渇と ROI ギャップ

The Information / Fortune / Rapid Response Podcast (A. Macdonald, C00)

32% → 84% Claude Code 採用率 (5,000人組織・Q1)

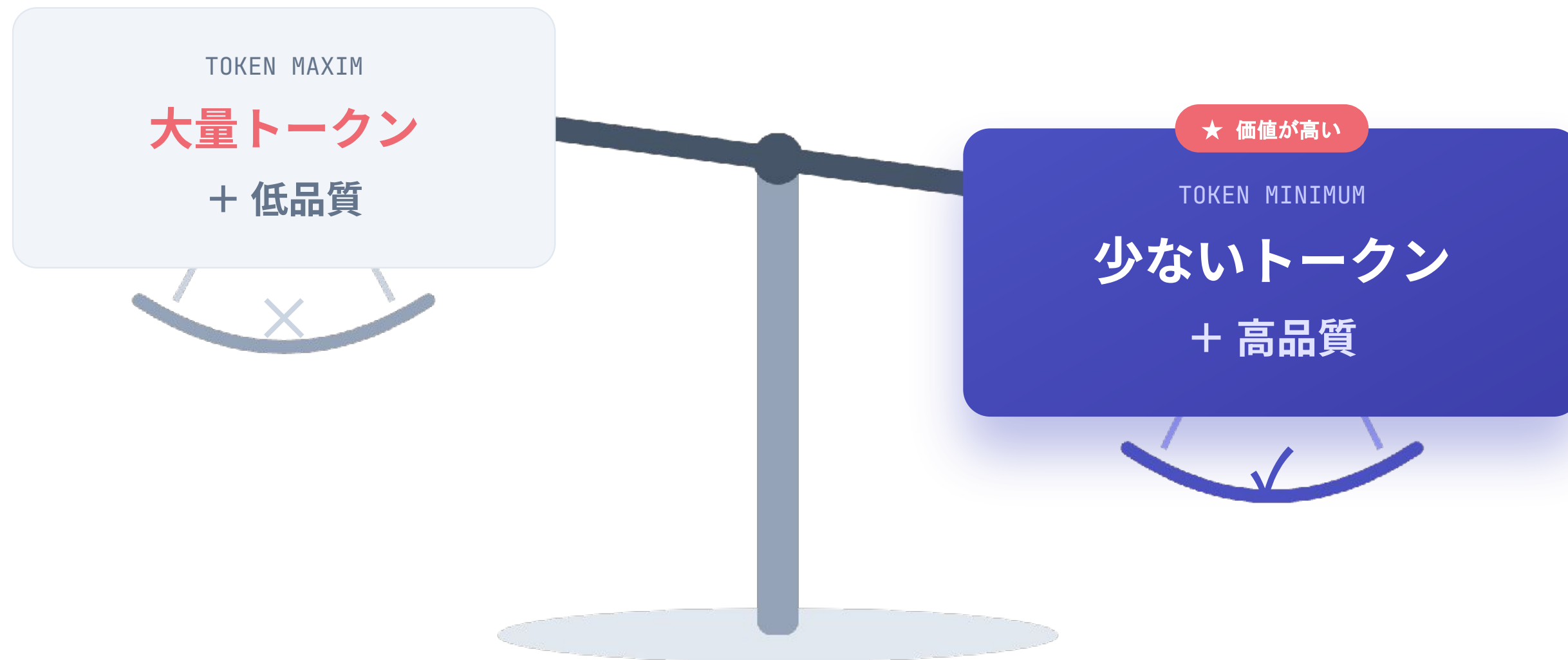
\$500-2,000 AIツールコスト / 人・月

4ヶ月 2026年のAI予算全額を使い切る

「AI支出と、顧客が実際に求めるプロダクトとの
明確なつながりが見えない」

"That link is not there yet." — Uber C00

価値があるのは、高品質



少ないリソースで、**高い品質**を —— これが評価すべき本質的な価値

AI活用の「4つのD」

Anthropic ・ AI Fluency Index

D1

Delegation
役割分担

人とAIの役割分担を決める

D2

Description
言語化

目的を簡潔に言語化してAIを導く

D3

Discernment
見極め

出力の妥当性を見抜く

D4

Diligence
責任

利用と結果に責任を持つ

海外の評価フレームワークは、いずれも **同じ方向** へ収束しつつある

エンジニア視点では、「5つのスキル」

CoderPad ・ 5 Skills of the Future Developer



Anthropic の「4つのD」 とも **大枠が一致** — 業界の収束点

出典：CoderPad "5 Skills of the Future Developer: A Framework for Evaluating AI Fluency" (2026)

トークン使用量を測っても、"質"は見えない

同じ「AI活用の可視化」でも、見るべき指標がまったく違う

従来 – Token Maxing

どれだけ使ったかを測る

トークン消費量・コスト・利用回数

多く使うほど「優秀」とみなす

→ 使いこなしの質はわからない

これから – Skill Scoring

どれだけ使いこなしたかを測る

AI協働の質をスコア化（0-100）

個人・チーム単位でランク付け

→ 誰が上手いかが見える

「使った量」ではなく「使いこなしの質」—— 別の測定とスコアリングが要る

AI協働スキルは、**どう測ればよいのか**

特別な試験は要らない。普段の開発からそのまま、3ステップで測る

STEP 1

集める

普段使うAIツール

(Claude・Cursor・Copilot等) から、**AI活用のログを自動で集める**

軽量アプリを入れるだけ。設定変更・APIキーの差し替えは不要



STEP 2

評価する

ログから「**指示・反復・検証**」などの協働プロセスを評価し、スコア化する

"多く使ったか"ではなく"うまく使えているか"を見る



STEP 3

見える化する

個人・チーム別の**ダッシュボード**で、習熟度を一覧で確認できる

誰がどのツールを、どの程度使いこなしているか

「AI協働力」は、すでに可視化できる

同じ実践課題でも、AI協働の質でスコアは分かれる —— 右の **AIスコア画面** に注目

低スキル

クーポン・割引適用 API の実装

AIスコア 53/100

高スキル

レガシー受発注システムの内製化移行

AIスコア 90/100



同じ実践課題でも、AI協働の質でスコアは大きく分かれる —— 右の「AIスコア (β)」で定量化できる

エンジニアを6軸で捉える

従来の 5 軸

01 CS基礎

02 実装力

03 システム設計

04 課題解決

05 コミュニケーション

+ 新しい 第6の軸

06 AI協働力

NEW

既存の指標に新たに AI協働力を新設

従来の5軸は、エンジニアの実力を語る共通言語
生成AIが開発に組み込まれた今、6つ目の軸——AIを使いこなす力——を、独立した評価対象として再定義

AI協働力の内訳は 4つのスキルの組み合わせ

① Agent Setup

環境エンジニアリング

AGENT.md 整備・ツール設定・静的コンテキストの提供

② Prompting

AIへの指示

タスク設計・指示の明確さ・動的コンテキストの提供

③ Iteration

反復改善

収束効率——少ない往復でゴールに到達する

④ Verification

出力の検証

AI生成コードの確認・修正・採否の判断

Agent Setup → Prompting → Iteration ↺ Verification

上手い使い方 vs 下手な使い方

上手い

- ✓ ゴール・制約・出力形式が明確な指示
- ✓ 関連ファイルを事前に提供
- ✓ 原因分析 → 的確な修正指示
- ✓ 早期に中断・方針転換
- ✓ コードを読み、修正してコミット

下手

- ✗ 「いい感じにやって」と丸投げ
- ✗ AI任せで手当たり次第に探索
- ✗ 「直して」をただ繰り返す
- ✗ AIの暴走を放置
- ✗ 確認なしでそのまま採用

量のトークンマキシムから、質のトークンミニмумへ

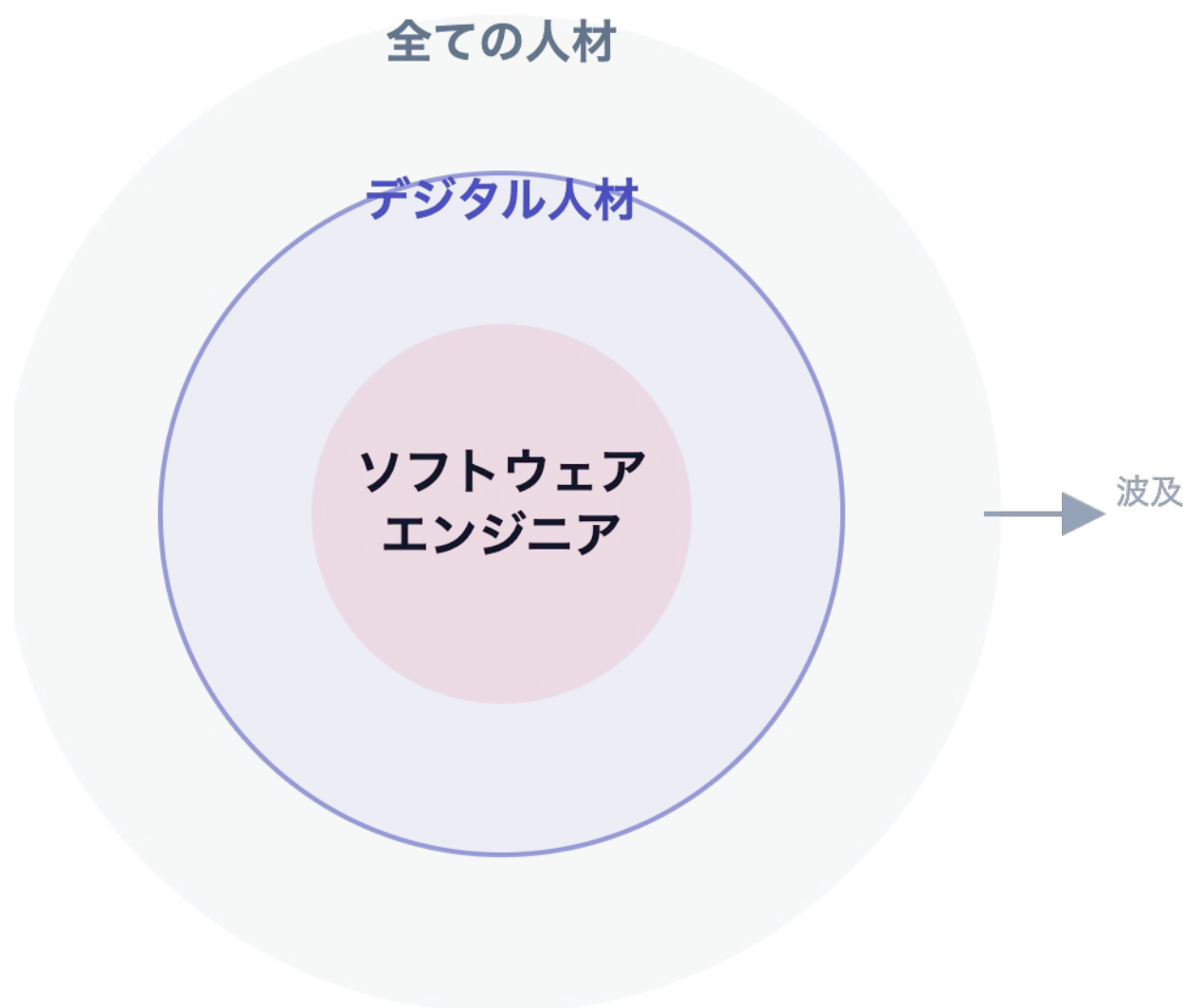
- 01 マキシムの終焉 AIベンダーの料金プラン変更など、複合的な理由でトークンマキシムは終わりを迎えた
- 02 トークンミニмумへ 次に求められるのは、少ないトークンで成果を出す力
- 03 量から質への測定 これまでの量で測る時代から、質を測る時代へ転換する
- 04 物差しは形成途上 指標は未確立だがAI活用の物差しは定義されつつある —— HireRoo はその測定ツール
- 05 全人材へ波及 この変革は エンジニア → デジタル人材 → すべての人材 へ波及していく

RECAP

| エンジニアだけの話ではない

この波は、やがて **全ての人材** に広がる

ソフトウェアエンジニアは **先行指標** —— AI活用が最も顕著に現れる最前線にすぎない



いま

ソフトウェアエンジニア

AI活用が最も顕著な最前線。だから、まずここからはじまる。

次に

デジタル人材

データ・デザイン・企画・マーケ・コーポレート —— 日常業務にAIが浸透。

やがて

全ての人材

ホワイトカラー全体へ。職種を問わず、AI協働力が問われる。

「AIをどう使いこなすか」を測る必要性は、**職種を越えて広がる** —— 最も顕著なエンジニアの現場で、いま方法論を確立する

HIREROO の取り組み

Hireroo Monitor（LLMプロキシ）でAI活用ログ計測

SEI × Four Keys マトリクスで協働力スコアリング

個人・チーム・組織レベルで可視化



「AI使っています」ではなく、
「AIと上手く協働できています」の証明へ

HireRoo  株式会社ハイヤールー

APPENDIX

参考データ・出典アーカイブ

トークンマキシム / トークンミニマム ― リサーチまとめ

A

トークンマキシム：報道・一次情報

Meta ・ Uber ・ Shopify ・ Google I/O

Meta・Amazon・Microsoft の後退

Fortune ・ 2026.05.28 - "Tokenmaxxing is over."

Meta	リーダーボードを削除——ランキングのためにアイドルエージェントを起動していたことが発覚
Amazon	社員が無意味・不要なタスクにエージェントを起動し、消費数を稼いでいた
Microsoft	複数のプロダクト部門で Claude Code をキャンセル

起源：Karpathy / Shopify / Google I/O

Andrej Karpathy

「ゲームの名はトークン。**ループから外れ**、スループットをどう最大化するか」—— 自律動作の最大化という本来の思想が起源

Shopify（2025）

最初のトークンリーダーボード。のちに「競争を煽りたくない」と **usage dashboard** へ改名（F. Thawar）

Google I/O（2026.05）

Sundar Pichai が "tokenmaxxing" に言及し **警戒を呼びかけ**。用語が一般化した転機

出典：Kiny AI（2026.04.12）／Pragmatic Engineer Podcast／Trending Topics（2026.05.27）

B

調査・学術データ

Faros AI ・ Jellyfish ・ Microsoft × Stanford ・ arXiv

Faros AI：22,000人調査（Q1 2026）

Faros AI Blog (Neely Dunlap) ・ 2026.04.23



スループットは「何が Ship されたか」を測るが「何が生き残ったか」は測らない。トークン消費はインプットであって、アウトカムではない。

Jellyfish：12,000人・200社調査

Activated Thinker / Medium ・ 2026.05

トークン消費量（中央値）

51M / 月

90パーセンタイル

380M / 月（中央値の 7.5倍）

マージPR数の比較（上位20% vs 下位20%）

11

下位20%

23

上位20%

PR数は2倍に対し、トークンは7.5倍——高消費帯ほど「コスト当たりPR数」が急低下

▶ トークンには効率カーブがある——低消費帯では実利、高消費帯では逡減

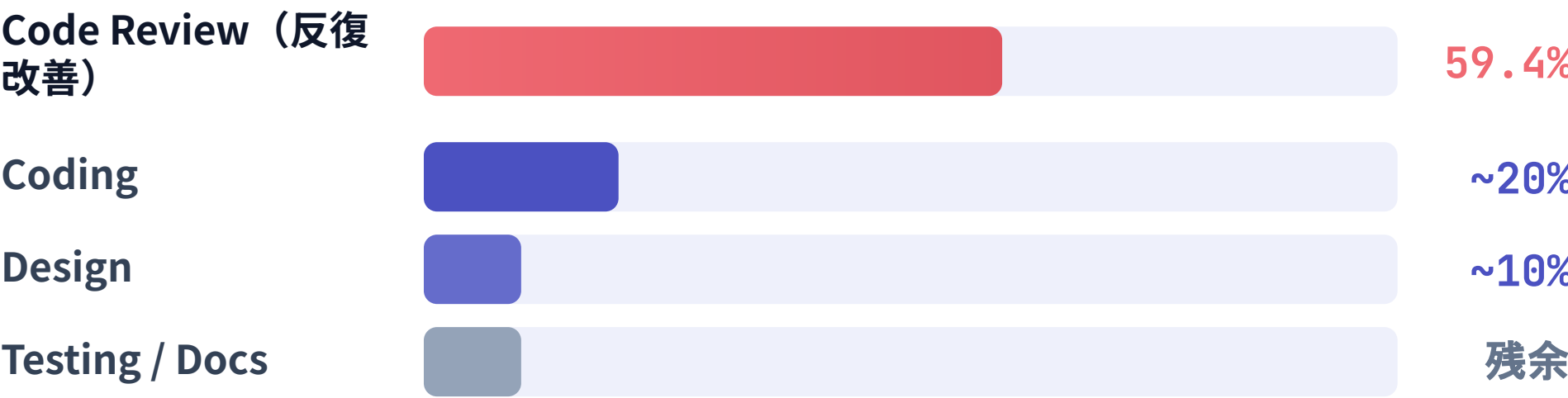
Microsoft Research × Stanford：5つの発見

arXiv 2604.22750 (Brynjolfsson, Pentland, Pei ほか)・2026.04

- 1 エージェントタスクは通常のコード補完・チャットの **1,000倍** のトークンを消費
- 2 高消費 ≠ 高精度 —— 同一タスクで最大 **30倍** の差、精度は**中間コストでピーク**、高コストで飽和
- 3 モデル差：Kimi-K2 / Claude Sonnet 4.5 は GPT-5 より平均 **150万トークン** 多く消費
- 4 専門家の「難易度」評価と実際のトークンコストの**相関は弱い**
- 5 AI は**自分の消費量を予測できない**（相関係数 最大0.39、系統的に過小評価）

Tokenomics：トークンはどこで使われるか

arXiv 2601.14470 (ChatDev × GPT-5・30タスク)・2026.01.20



主要コストは初期のコード生成ではなく、**自動化された改善・検証フェーズ**にある。

▶ HireRoo の **Iteration (収束効率)**・**Verification (検証)** 不足が、最もコストに直結する

Vantage：Input / Output 比率は 25:1

Vantage Blog ・ 2026.04.15 - 50ターンのエージェントセッション

入力トークン **約100万**

出力トークン **約4万**

比率 **≒ 25 : 1**

従来のAI利用では出力トークンが割高だったが、エージェント型では
入力トークンが支配的コストに。

コンテキストが蓄積するほど入力コストは増大し、出力はほぼ一定。

▶ **Agent Setup** の巧拙が、入力トークン（=主コスト）を直接左右する

C 批判と反省

Milestone ・ Falconer ・ Built In

Milestone：「LOC問題の再来」

Milestone Blog ・ 2026.04.24

「この映画は前に見た。LOC もコミット数もストーリーポイントも客観的に見えた。トークン数は同じ罨だ——ただし**直接的なインフラコスト** も伴う」

トークンを使うべき場面

- 広範なりポジトリ移行
- テスト生成・大規模コード検索
- 不慣れなシステムへのオンボーディング
- 評価負荷の高いリリース作業

目的は**ミニマリズムの強制**ではない。
支出を **完成品質・経過時間・下流の信頼性** に整合させること。

Falconer：個人測定が機能しない理由

Falconer Guides ・ 2026.06.02

01

コンテキスト支配

個人スキルより、コードベース・質問の性質・ドキュメントの鮮度が消費量を左右

02

指標のゲーミング

評価に入ると、エンジニアは成果でなく**その指標**を最適化する

03

分散が大きすぎる

上級×慣れ と 若手×不慣れ で**3桁の差**が生まれる

▶ **システム（チーム・サーフェス）レベル**なら、分散が平均化され、インセンティブが個人ゲームからシステム改善へシフトし、最大の非効率が可視化される

自己保存動機が生む「強制トークンマキシム」

「'AI利用率' や '1人あたりPR数' が評価に入るなら、生産性が低いと見られないために、より多くAIを使うのは常識（=自己保存）になる」

— Gergely Orosz

トークンマキシムは「楽しいゲーム化」として語られがちだが、実際には使わなければならない心理的プレッシャーを生み、燃え尽きと無駄を同時に引き起こす。

D トークンミニマムと効率化

Cheah • Glean • Falconer • MindStudio

Aaron Cheah：トークンミニマムの提唱

Trending Topics ・ 2026.05.27 (AI アーキテクト)

「不要なトークンは、企業の生産性への**直接課税**だ—— レイテンシーを遅らせ、ユニットエコノミクスを削る」

実装原則

- 1つの巨大モデルに全てを任せない
- タスク特化の小型モデルを組み合わせる
- 「いかに少ない生成で完了するか」で評価

「最も洗練されたAIフレームワークは、ジョブ完了のために **どれだけ少なくて生成すれば済むか** で評価されるようになる」

永続的知識 vs 一時的リクエスト

Nagaprasad Sathyanarayana / Medium · 2026.04.13

永続的知識（一度書く）

仕様・リポジトリルール・繰り返しワークフローを一度パッケージ化。メモリは持続的な好みのみ保持。

⇨ Agent Setup（静的コンテキスト）

一時的リクエスト（デルタのみ）

ライブプロンプトは「今必要な特定ステップ」だけに使う。

⇨ Prompting（動的コンテキスト）

同時に得られる4つのゲイン

- トークン浪費の削減
- レイテンシー低下
- コンテキスト汚染の防止
- 出力品質の向上

Glean：テキスト編集ではなくシステム設計問題

Glean Perspectives ・ 2026.06.03

高コストはどこに？

計画・ツール使用・エージェント間通信・検証・リトライに分散。システムプロンプト短縮では効かない。

最も高コストなテール（上位10%のラン）は 曖昧な指示・欠けたコンテキスト・無制限ループ が原因。平均でなくテールを潰す方が効果的。

LOOP Skill Engine (Wang et al., 2026)

成功トレースを「ブランチのないレシピ (SKILL.md)」にコンパイル

93.3%

日次反復タスクの削減

99.98%

高頻度オペの削減

Falconer：共有知識層という解

Falconer Guides ・ 2026.06.02 - 冗長な再導出問題

エージェントAが認証ミドルウェアを問われ、14ファイル+3スレッドを読んで回答。2時間後、エージェントBが同じ質問に**同じ14ファイル・同じ推論**を繰り返す
—— **組織は一度払った情報に再度課金**される。

× 個人レベルのトークン数

ノイズが多く、ゲーミングされる

✓ チーム/システムの output-per-token

平均化され、改善可能

MindStudio：「トークン効率」の定義

MindStudio Blog ・ 2026.05.20

トークン効率＝1トークンあたりにモデルがこなす有用な仕事量。非効率なモデルは冗長なCoT・出力・過剰なコンテキスト処理に浪費する。

非効率が引き起こす3つの連鎖問題

① レイテンシー

20回反復ループで **10倍**の時間差が生まれる

② コスト

トークン課金は**線形**に増える

③ コンテキスト汚染

不要情報で**後続ターンの精度が低下**

E 業界の料金体系変化

定額制 → トークン従量課金へ

Anthropic 2026 モデル単価

Anthropic API 公式 / Finout Blog ・ 2026.06 (\$ / MTok)

モデル	Input	Output
Claude Haiku 4.5	\$1.00	\$5.00
Claude Sonnet 4.6	\$3.00	\$15.00
Claude Opus 4.8	\$5.00	\$25.00
Batch 処理 / Prompt Cache		全モデル 50%OFF / Cache 最大 90%OFF

Claude Code の平均コスト：\$6 / 開発者・日（90%が \$12/日 以下）—— Anthropic 公式

報道・一次情報 / 調査・学術データ

01

The Pulse: 'Tokenmaxxing' as a weird new trend

Gergely Orosz / Pragmatic Engineer · 2026-04-23 · [blog.pragmaticengineer.com](#)

03

Tokenmaxxing is over. A flawed way to measure AI ROI

Jeremy Kahn / Fortune · 2026-05-28 · [fortune.com](#)

05

What Is Tokenmaxxing? The AI Workplace Trend Explained

Sara B.T. Thiel / Built In · 2026-04-22 · [builtin.com](#)

07

The Productivity Paradox of Generative AI Consumption

Adnan Masood, PhD / Medium · 2026-04-20 · [medium.com/@adnanmasood](#)

09

Why AI token consumption isn't engineering productivity

Neely Dunlap / Faros AI · 2026-04-23 · [faros.ai/blog/tokenmaxxing](#)

11

Tokenomics: Where Tokens Are Used in Agentic SE

arXiv 2601.14470 · 2026-01-20 · [arxiv.org/abs/2601.14470](#)

02

Tokenmaxxing Is Not AI Productivity

Activated Thinker / Medium · 2026-05 · [medium.com/activated-thinker](#)

04

Productivity Metric or Vanity Trap?

J. Steinschaden / Trending Topics · 2026-04-23 · [trendingtopics.eu](#)

06

Silicon Valley's Most Controversial New Status Game

Kingy AI · 2026-04-12 · [kingy.ai](#)

08

Tokenmaxxing Explained: How It Works & Why It Matters

Milestone Team · 2026-04-24 · [mstone.ai](#)

10

How Do AI Agents Spend Your Money?

Microsoft Research + Stanford · 2026-04-29 · [arxiv.org/abs/2604.22750](#)

12

The Hidden Cost Driver in Agentic Coding Sessions

Vantage · 2026-04-15 · [vantage.sh/blog/agentic-coding-costs](#)

トークンミニマム・効率化 / 料金体系変化

13

"Every Unnecessary Token... a Direct Tax on Productivity"
J. Steinschaden / Trending Topics · 2026-05-27 · trendingtopics.eu

15

Token Efficiency in AI Coding Agents
Nagaprasad Sathyanarayana / Medium · 2026-04-13 · medium.com/@nprasads

17

Token Efficiency vs Model Intelligence
MindStudio · 2026-05-20 · mindstudio.ai/blog

19

Anthropic shakes up its pricing strategy
Jérôme Marin / Cafetech in English · 2026-06 · cafetechinenglish.substack.com

21

The AI Token Pricing Crisis Behind the Revenue Race
Investing.com · 2026-05 · investing.com/analysis

14

Token efficiency when agents share a knowledge layer
Falconer Guides · 2026-06-02 · falconer.com/guides

16

How to optimize token efficiency in agentic systems
Glean Perspectives · 2026-06-03 · glean.com/perspectives

18

Techniques to Reduce AI Token Usage: The 2026 Playbook
Program Strategy HQ · 2026-05 · programstrategyhq.com

20

Canonical ref: Anthropic May 13 \$200 credit policy change
MagnaCapax / GitHub Gist · 2026-05-14 · gist.github.com

22

Anthropic API Pricing in 2026: Complete Guide
Finout · 2026-06 · finout.io/blog/anthropic-api-pricing

※ 本資料は2026年6月22日時点の情報に基づきます。作成：葛岡宏祐（株式会社HireRoo）