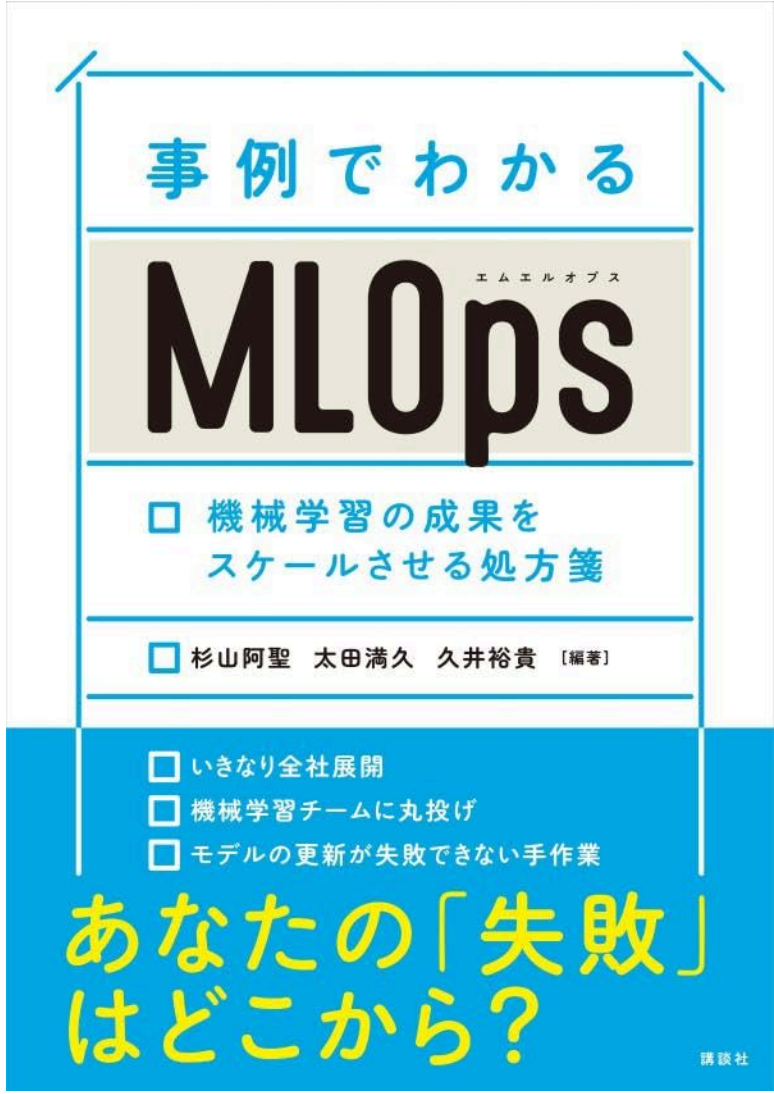
# 生成AI活用の実践解説 (速報版)

Asei Sugiyama

# 自己紹介

- 杉山 阿聖 (@K_Ryuichirou)
- **Software Engineer @ Citadel AI**
- **Google Developer Expert @ Cloud AI**
- MLSE GenAIOps WG
- 機械学習図鑑 共著
- 事例でわかる MLOps 共著



# TOC

- **背景 <-**
- 継続的評価による継続的改善
- AI セーフティ
- AI ガバナンス
- 発展的な話題

# 背景

- AI ネイティブカンパニー
- Agent の流行
- 評価に基づく LLMOps
- 市場調査: NEDO AI セーフティ強化に関する研究開発プロジェクト
- LLMOps の取り組み


2025-8-12

## メルカリが本気で始めた「AI-Native」化。100名規模のタスクフォースが立ち上がるまで

AI/LLM　エンジニアリング

Share

2025年5月、メルカリは全社で「AI-Native」な会社への転換を宣言しました。
その発信源は、グループCEOの山田進太郎（@suadd）によるSlackメッセージでした。

> これは一部の部門や技術職だけの話ではなく、全社員に関わる話です。
> 私たちは「AIを導入する企業」ではなく、「AIを前提に再設計された組織」になります。

メッセージの投稿からわずか数週間で、社内から100名を超えるメンバーが選出され、「AI Task Force」という横断組織が立ち上がりました。中でもAI技術に精通した40名のエンジニアは、主務としてこの組織に異動しています。

この約100名が中心となり、全社の業務を棚卸し、AIを前提に再構築する取り組みをリードしていきます。

事業・サービス

## DeNAが挑む「AIネイティブカンパニー」への全社的取り組み。人をエンパワーし、イノベーションを創出する—南場智子

2025.07.29

# 南場智子　# AI

2025年7月10日に開催された日本最大級のHRラーニングイベント「HR SUCCESS SUMMIT 2025」（株式会社ビズリーチ主催）より、DeNA会長・南場智子の基調講演を全文お届けします。

# AIネイティブカンパニー

DeNAが挑む「AIネイティブカンパニー」への全社的取り組み。人をエンパワーし、イノベーションを創出する—南場智子 | フルスイング by DeNA
https://fullswing.dena.com/archives/100169/
メルカリが本気で始めた「AI-Native」化。100名規模のタスクフォースが立ち上がるまで | mercan (メルカン)
https://careers.mercari.com/mercan/articles/53708/

# AI Agent がビジネスに

- Algomatic: アポドリ、リクルタAIなどエージェントを利用したサービスを次々展開
- LayerX: バクラクシリーズに加え、生成AIプラットフォーム Ai Workforce を展開

複数 AI エージェント同時立ち上げで「打席数」を重視——DMM が用意した20億円でスタートアップスタジオ展開する Algomatic #取材会 - BRIDGE（ブリッジ）
https://thebridge.jp/2025/06/algomatic-launches-startup-studio-with-¥2-billion-investment-from-dmm
LayerX、総額150億円の資金を調達。エンジニアを中心とした人材採用強化やさらなる事業拡大へ https://aismiley.co.jp/ai_news/layerx-ai-saas-fintech/

**複数 AI エージェント同時立ち上げで「打席数」を重視——DMM が用意した20億円でスタートアップスタジオ展開する Algomatic #取材会**

ミートアップに参加してくれた Algomatic 代表取締役の大野峻典氏

1 B!ブックマーク
1 いいね！
4,821 シェアする
Post
Follow

**LayerX、総額150億円の資金を調達。エンジニアを中心とした人材採用強化やさらなる事業拡大へ**

最終更新日:2025/09/08


LayerXは、TCVをリード投資家とする総額150億円のシリーズB資金調達を発表しました。エンジニアを中心とした人材採用強化やさらなる事業拡大に向け活用する方針です。

# Demo hell

- デモまでは行き着くものの、本番化が著しく困難
- 品質を評価し、担保することが極めて困難

Escaping AI Demo Hell: Why Eval-Driven Development Is Your Path To Production
https://www.forbes.com/councils/forbestechcouncil/2025/04/04/escaping-ai-demo-hell-why-eval-driven-development-is-your-path-to-production/

## Escaping AI Demo Hell: Why Eval-Driven Development Is Your Path To Production

By Albert Lie, Forbes Councils Member.
for Forbes Technology Council, COUNCIL POST | Membership (fee-based)

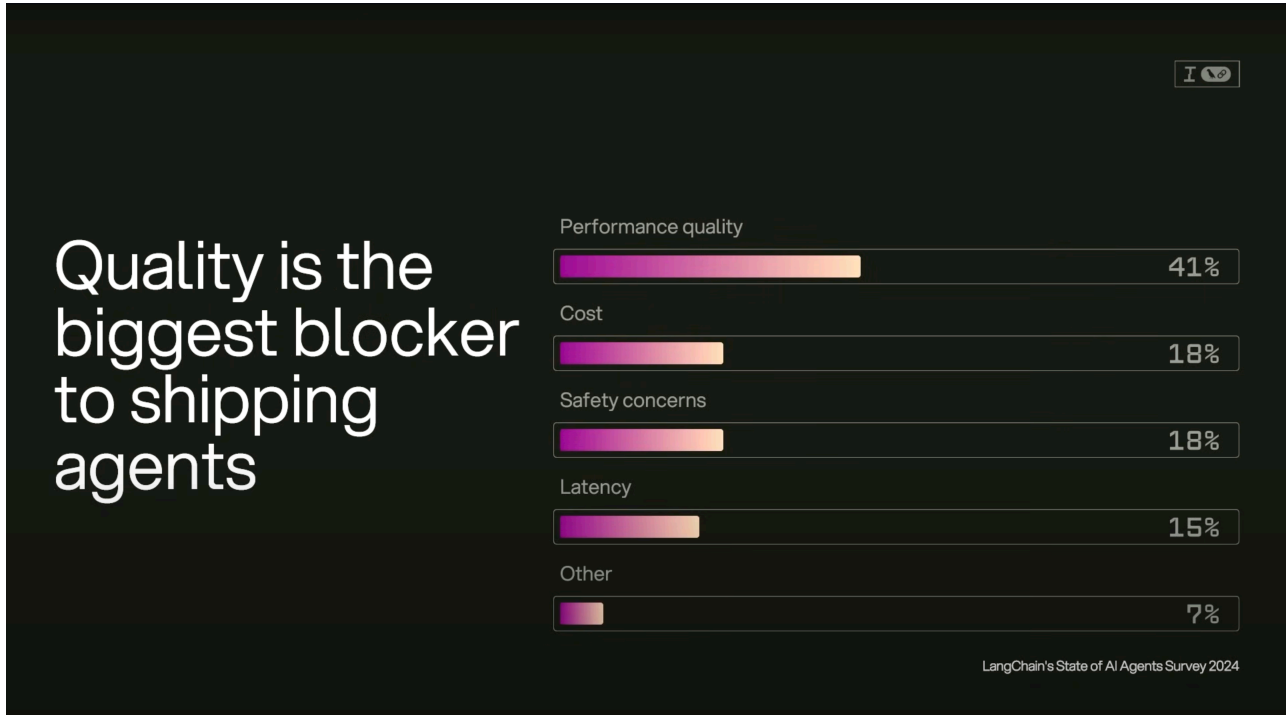
Published Apr 04, 2025, 07:45am EDT

Share Save

*Albert Lie, Cofounder and CTO at Forward Labs, next-gen AI-driven freight intelligence for sales and operations.*

# 評価に基づく LLMOps

- 最近の潮流
- 不確実性の高い分野に対する開発のノウハウ集
- MLOps というよりは「高速 DevOps」





How to Solve the #1 Blocker for Getting AI Agents in Production | LangChain Interrupt https://interrupt.langchain.com/videos/building-reliable-agents-agent-evaluations

# AI セーフティ強化に関する研究開発プロジェクト

- Citadel AI で「企業向け実装解説」としてベストプラクティス集・事例集の作成を担当
- 「デモは簡単にできるもののサービス化や本番化は難しい」というテーマでヒアリング
- 似たような課題・対策を行っていることが見えてきた

## AI セーフティ強化に関するヒアリング依頼書

国立研究開発法人新エネルギー・産業技術総合開発機構 (以下 NEDO)による "AIセーフティ強化に関する研究開発" に関連して、生成 AI の企業内での利活用に関するヒアリングをお願いいたします。

### ヒアリングの背景

### AI セーフティ強化に関する研究開発プロジェクト概要

NEDOは、研究開発とSociety 5.0との橋渡しプログラム（BRIDGE）における施策として、生成AIを適切に管理・利用するために必要となる安全性評価技術の開発・普及を目的に、「AIセーフティ強化に関する研究開発」を行います。具体的な研究開発内容は以下の通りです。

〔1〕リスクベースアプローチの基になる安全性の“ものさし”（基準）となる技術（評価・管理技術）の開発
〔2〕人間拡張など暮らし領域での評価手法の開発と実証およびテスト環境構築技術の開発
〔3〕国際標準化および普及のためのガイダンス等の整備

Citadel AIは、本プロジェクトを産業技術総合研究所(以下、産総研)、株式会社コーピーと共同で進めていきます。

### Citadel AI の担当範囲

弊社は上記の「AI セーフティ強化に関する研究開発」のうち、「〔3〕国際標準化および普及のためのガイダンス等の整備」の一部を担当します。具体的な担当範囲は次のとおりです。

- 企業等へのヒアリングに基づく AI セーフティ評価基準・評価手法の整理
- 企業等へのヒアリングに基づく企業向け実装解説の作成及びその技術的有効性の検証

これらの項目について、生成 AI を活用するためのベストプラクティス集・事例集を「企業向け実装解説書」として作成・公開することを予定しています。

### 目的

「企業向け実装解説書」においては、企業等の各組織で行われている生成 AI を活用するための具体的な取り組みについてベストプラクティスや事例を記載します。これは、現行の各種ガイドラインを補完する形で、ヒアリングに基づき各組織のAI セーフティ評価基準・評価手法事例についての取り組みをボトムアップに収集してまとめる取り組みです。本ヒアリングの結果は「企業向け実装解説書」に記載する内容に用います。

# LLMOps の取り組み

- 継続的評価による継続的改善
- AI セーフティ
- AI ガバナンス

# TOC

- 背景
- **継続的評価による継続的改善 <-**
- AI セーフティ
- AI ガバナンス
- 発展的な話題

# 継続的評価による継続的改善

- 評価は難しい
- Who Validates the Validators?
- プロンプトエンジニアリング
- LLM-as-a-Judge
- 継続的な評価
- ハッカソン
- エージェント
- トレース

# 評価は難しい

- システムの正しい振る舞いを、誰も明確に記述できない
- 品質評価の観点を事前に列挙することは困難で、出力から事後的に得られることが大半

機械学習による言語パフォーマンスの評価 - Speaker Deck
https://speakerdeck.com/langstat/ji-jie-xue-xi-niyoruyan-yu-pahuomansunoping-jia

- 自動評価システムの問題
  - 人間の**構成概念の全てをカバーできていない**

    vs.
- 人間の評価者の問題
  - 自分の**構成概念の全てを言語化することができない**
    - 評価項目の**網羅的なリスト**（can-doリストでは不十分）
    - 個々の評価項目ごとの**重みづけ**（全ての評価項目が同程度の重要性を持っている訳ではない）

    ↓
- 実際のデータから構成概念を推定
  - 大量の学習者データ（**学習者コーパス**）
  - データからの言語情報抽出（**自然言語処理**）
  - 推定のための統計手法（**機械学習**）

17

- **残された問題**
  - どのような評価項目を機械学習に基づく推定実験に用いるか？

    ↓
  - **機械学習による変数選択**
    - **分析に使用する全ての評価項目（説明変数）の重要度を推定し、重要度の低い項目を分析から除外**
    - しかし、分析者による「分析に使用する全ての評価項目」の定義が必要

    ↓
  - （前述のとおり）人間は構成概念の「全て」を言語化することができない

# Criteria Drift

- Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences
- LLM の出力に対する評価基準が、評価を進めるにつれてユーザー自身によって変化または洗練されていく

[2404.12272] Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences https://arxiv.org/abs/2404.12272



## Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences

Shreya Shankar
UC Berkeley
Berkeley, California, USA
shreyashankar@berkeley.edu

J.D. Zamfirescu-Pereira
UC Berkeley
Berkeley, California, USA
zamfi@berkeley.edu

Björn Hartmann
UC Berkeley
Berkeley, California, USA
bjoern@eecs.berkeley.edu

Aditya G. Parameswaran
UC Berkeley
Berkeley, California, USA
adityagp@berkeley.edu

Ian Arawjo
Université de Montréal
Montréal, Québec, Canada
ian.arawjo@umontreal.ca

### ABSTRACT

Due to the cumbersome nature of human evaluation and limitations of code-based evaluation, Large Language Models (LLMs) are increasingly being used to assist humans in evaluating LLM outputs. Yet LLM-generated evaluators simply inherit all the problems of the LLMs they evaluate, requiring further human validation. We present a mixed-initiative approach to "validate the validators"—aligning LLM-generated evaluation functions (be it prompts or code) with human requirements. Our interface, EvalGen, provides automated assistance to users in generating evaluation criteria and implementing assertions. While generating candidate implementations (Python functions, LLM grader prompts), EvalGen asks humans to grade a subset of LLM outputs; this feedback is used to select implementations that better align with user grades. A qualitative study finds overall support for EvalGen but underscores the subjectivity and iterative process of alignment. In particular, we identify a phenomenon we dub *criteria drift*: users need criteria to grade outputs, but grading outputs helps users define criteria. What is more, some criteria appears *dependent* on the specific LLM outputs observed (rather than independent criteria that can be defined *a priori*), raising serious questions for approaches that assume the independence of evaluation from observation of model outputs. We present our interface and implementation details, a comparison of our algorithm with a baseline approach, and implications for the design of future LLM evaluation assistants.

### CCS CONCEPTS

• **Human-centered computing → Interactive systems and tools**; • **Computing methodologies → Natural language processing**.

### KEYWORDS

language models, auditing, evaluation, interfaces, prompt engineering, active learning

### 1 INTRODUCTION

Large Language Models (LLMs) make mistakes—they hallucinate, ignore instructions, and generate outputs that require validation [25]. But validating the behavior of LLMs is challenging. In response, researchers and industry developers have created tools for prompt engineering and auditing that help people with testing outputs more systematically [1, 16, 23, 24, 27, 36, 41, 52]. Such approaches require *metrics*—a set of functions to automatically score LLM outputs, each typically an assertion with *true* or *false* values. These metrics increasingly include calls to "evaluator" LLMs (e.g., [1, 27, 52, 58]) that act as "judges," grading outputs on qualities hard to articulate in code; for instance, the "conciseness" of an output.

The problem is that, just like the LLMs they evaluate, LLMs that perform evaluations cannot be trusted. These "grader" prompts suffer from the same problems as any other prompt—they are unintuitively sensitive to seemingly minor changes in wording or structure [43]. Yet many existing systems do not include support for verifying the quality of LLM-generated evaluations, asking users to simply trust these outputs. How can users reap the efficiency benefits of LLM-assisted evaluation of LLM outputs, while minimizing or avoiding misalignment? How can we help users validate the validators?

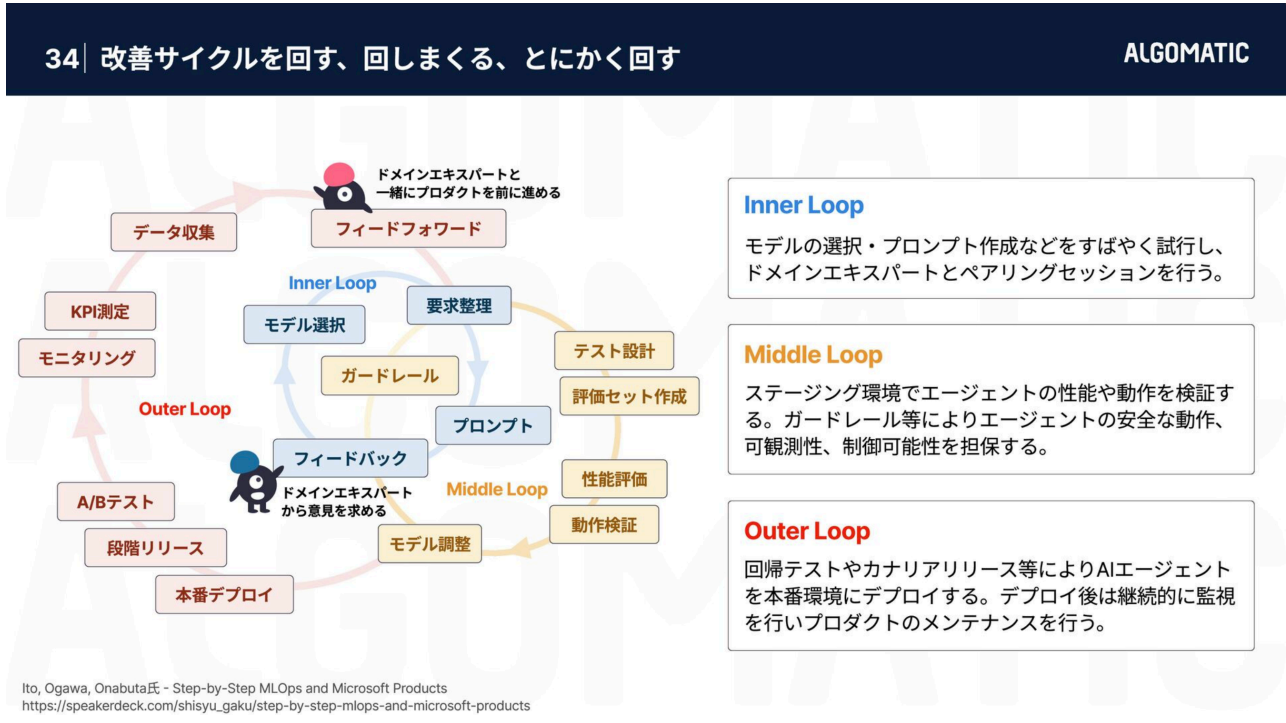
In this paper, we propose a mixed-initiative approach, EvalGen, to address this automated-evaluation alignment problem in the context of prompt engineering. Our approach streamlines the selection of metrics under practical constraints of user effort and latency. Specifically, an LLM suggests criteria in natural language, based on user context (e.g., the prompt under test), that the user can modify. An LLM then generates a pool of candidate assertions for each criterion—either code or LLM grader prompts that output "true" or "false." While the user waits for the LLM to generate candidates, they are asked to grade outputs with a simple "good" (thumbs-up) or "bad" (thumbs-down) voting scheme. These grades then guide the automatic selection of assertions that optimize for alignment with user preferences (Section 4.2). After assertion selection, a final report card reveals the alignment between the chosen assertions and the user's grades. Our approach generalizes beyond the particulars of our specific design, and could be extended to, for instance, update metric implementations with feedback from human preferences, or query the user for finer-grained individual grades.

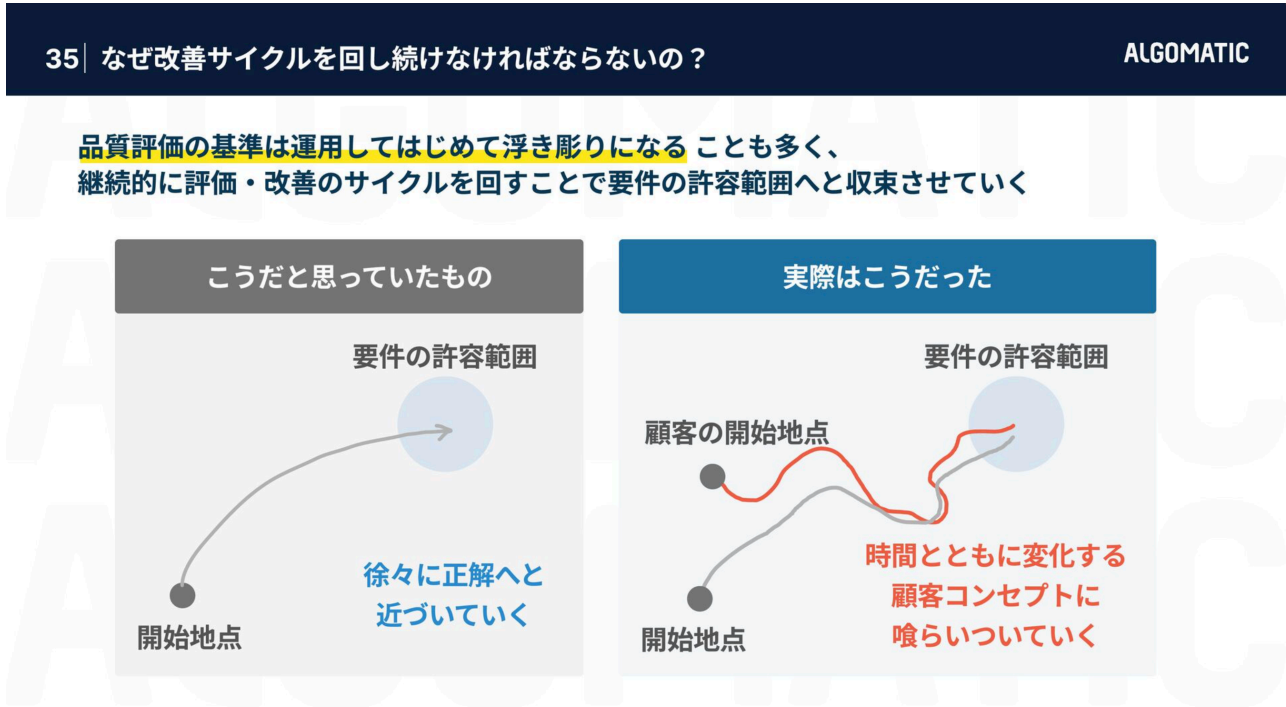
EvalGen is embedded inside an existing open-source interface for prompt engineering and auditing, ChainForge [1]. Our alignment algorithm adapts SPADE [45], a fully-automated algorithm for generating Python assertions from the revision history of a prompt. We performed an off-line verification of our human-guided alignment algorithm with SPADE as a baseline [45], then ran a qualitative user study with nine (9) industry practitioners that use

# 発想の逆転: 高速プロトタイピング

- 専門家も自分の行っていること・やりたいことを明確にできない
- 評価を繰り返すことで専門家の知識を明文化する
- 手戻りを恐れるのではなくイテレーションを回す

AIエージェントの地上戦 〜開発計画と運用実践 / 2025/04/08 Findy ランチセッション #19 https://speakerdeck.com/smiyawaki0820/08-findy-w-and-bmitoatupu-number-19



# プロンプトエンジニアリング

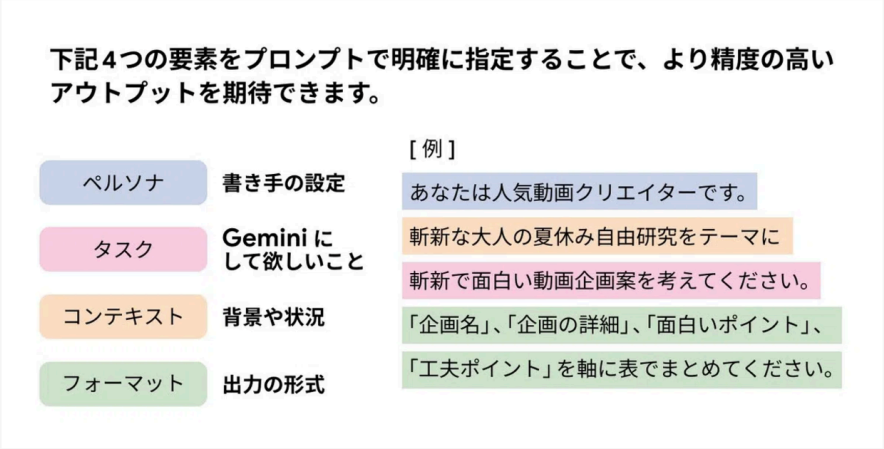
- 入力文章を調整して、言語モデルを効率的に使おうとする手法群
- 指示文を人が見たときにわかりやすくなるよう、明確に記述することが基本
- Gemini の記事が参考になる

Gemini から欲しい回答を引き出すプロンプト術｜Gemini - Google の AI
https://note.com/google_gemini/n/n60a9c426694e

## プロンプトで意識してほしい 4 つの要素

Gemini に何か質問をしたり指示を出したりする際には、以下の 4 つの要素を意識してみましょう。必ずしも全てが含まれている必要はありませんが、意識していただくだけでも、Gemini からの回答の精度がもっと高くなることが期待できますよ。

- ペルソナ（書き手の設定）
- タスク（Gemini にしてほしいこと）
- コンテキスト（指示の背景や状況）
- フォーマット（出力の形式）



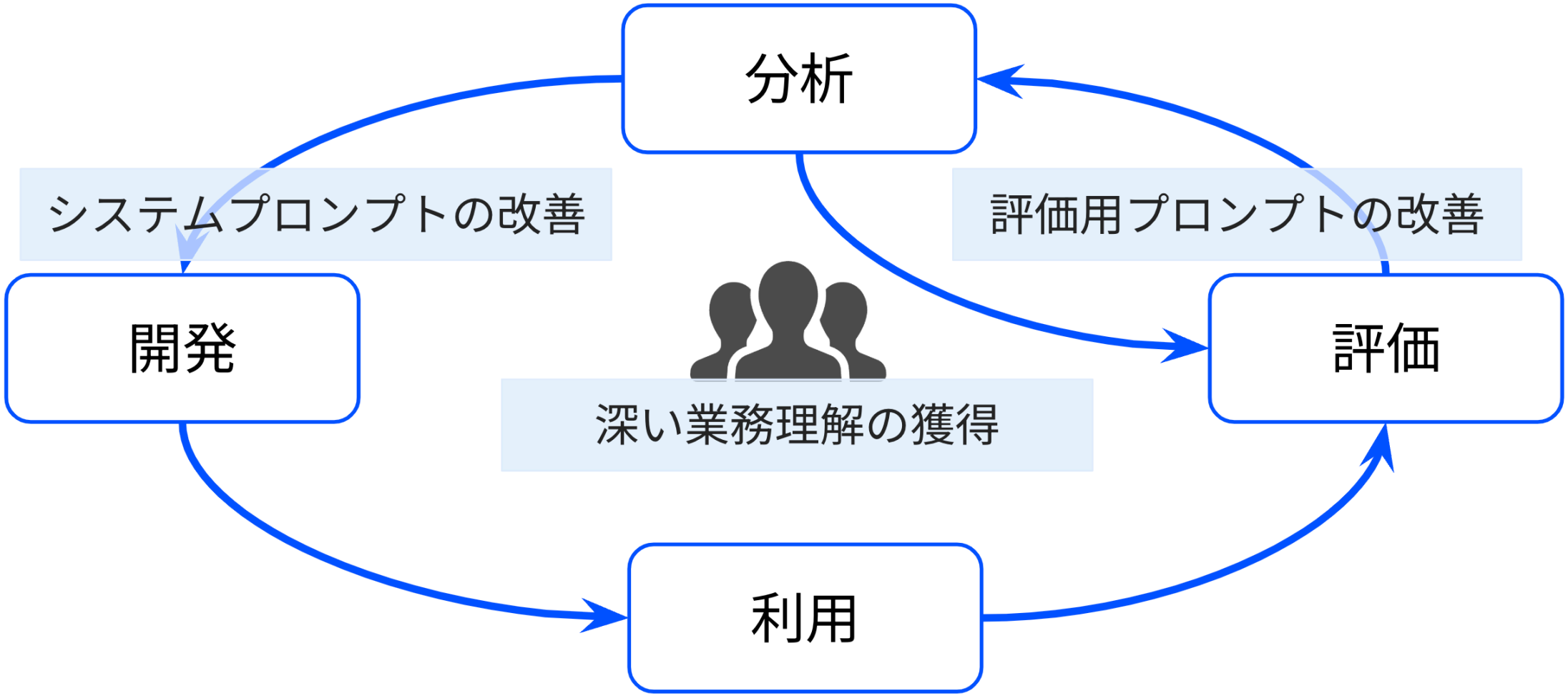

ここでは、この 4 つの要素について、プロンプト例を交えながら詳しく解説していきます。
紹介したプロンプトはコピーもできるので、実際に Gemini を開いて、ご活用いただけたらうれしいです！

# LLM-as-a-Judge

- プロンプトを用いて LLM に出力の良さを評価させる手法
- 新たな評価観点が得られた場合、その評価観点に基づく評価方法の手順書を書くと、その観点に基づく評価が LLM で可能

LLMによるLLMの評価「LLM-as-a-Judge」入門〜基礎から運用まで徹底解説 https://zenn.dev/pharmax/articles/2d07bf0498e212

# 継続的な評価による継続的な改善



# ハッカソン: デジタル庁

- ハッカソンは専門家を巻き込むために有効
- ハッカソンにより「5時間という短い開発時間の中で、38個のプロトタイプ」
- ハッカソンの成果物をOSSとして公開

第三弾：「法令」×「デジタル」ハッカソンを開催しました

公開日:2025年3月14日
最終更新日:2025年3月25日

デジタル庁

「法令」×「デジタル」ハッカソン

2025
2/5 水 – 3/6 木

Lawsy

法令 Deep Research ツール Lawsy を OSS として公開しました

146

Tatsuya Shirakawa
2025年3月6日 07:38

「法令」×「デジタル」ハッカソンの成果物として開発した日本の法令に特化した Deep Research ツール Lawsy をデジタル公共財として OSS 公開しました。本記事では Lawsy の詳細と開発で得た Tips を公開します。

第三弾：「法令」×「デジタル」ハッカソンを開催しました｜デジタル庁
https://www.digital.go.jp/news/9fb5ef8e-c631-4974-96d9-0b145304c553
法令 Deep Research ツール Lawsy を OSS として公開しました｜Tatsuya Shirakawa https://note.com/tatsuyashirakawa/n/nbda706503902

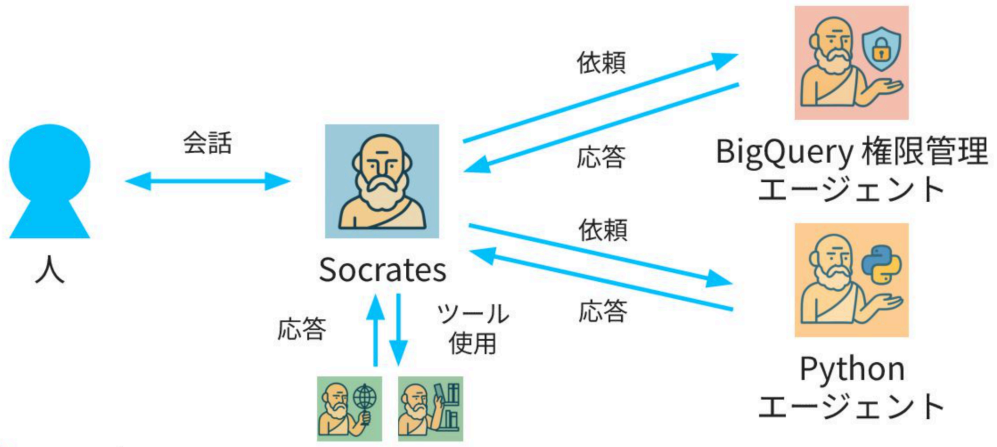
# Agent

- チューニングにより「1つのこと(ドメイン)をうまくやる」ものができる
- 各ドメインを協調させる取り組みがエージェント
- Microservices Architecture の再発見

メルカリにおけるデータアナリティクス AI エージェント「Socrates」と ADK 活用事例
https://speakerdeck.com/na0/merukariniokerudetaanariteikusu-ai-eziento-socrates-to-adk-huo-yong-shi-li

## Socrates との他の社内ツールとの共存関係

**Socrates は、ADK を基盤としたマルチエージェントシステムとして設計。各エージェントが専門タスクを担当し、協調動作。**



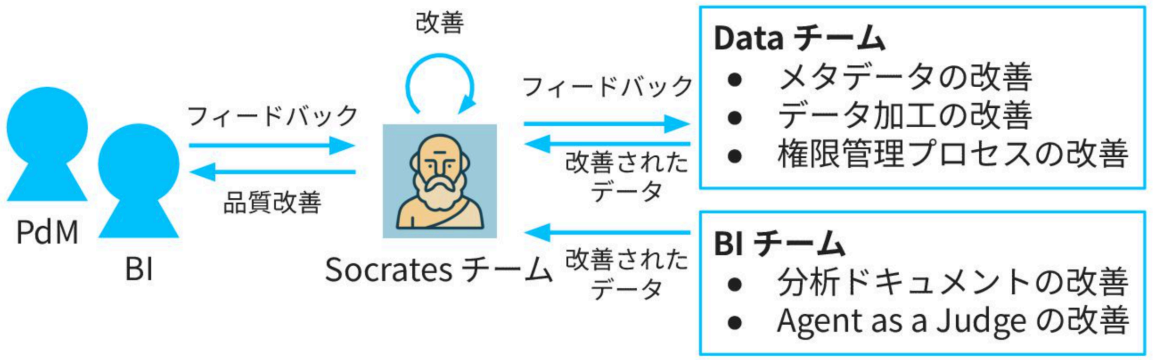

mercari

36

Skip

## Socrates を育むエコシステム：チーム

- **チームの連携：部門横断的な開発・運用体制とフィードバックループ**
  - 利用状況モニタリング、ユーザーヒアリング、エラーログ分析、Agent as a Judge を通じて、プロンプトやツールの迅速な改善を実施




mercari

56

# トレース

- Agent はさまざまなシステムを組み合わせて使うため、望ましくない結果が得られたときにその原因追及が困難
- 最終結果を生成するまでの途中で何が起きているのかを記録し、分析できるようにする
- LangSmith や Langfuse は Trace 機能を実装している

LangSmithによるLLMアプリケーションのトレーシング入門
https://zenn.dev/pharmax/articles/61edc477e4de17

## LangSmithのトレーシング

LangSmithのトレーシング機能を使うことで、実行ログのトレースが非常に簡単に行えます。以下は、実際のトレースログのサンプルです。


*LangSmithのトレースログ*

# TOC

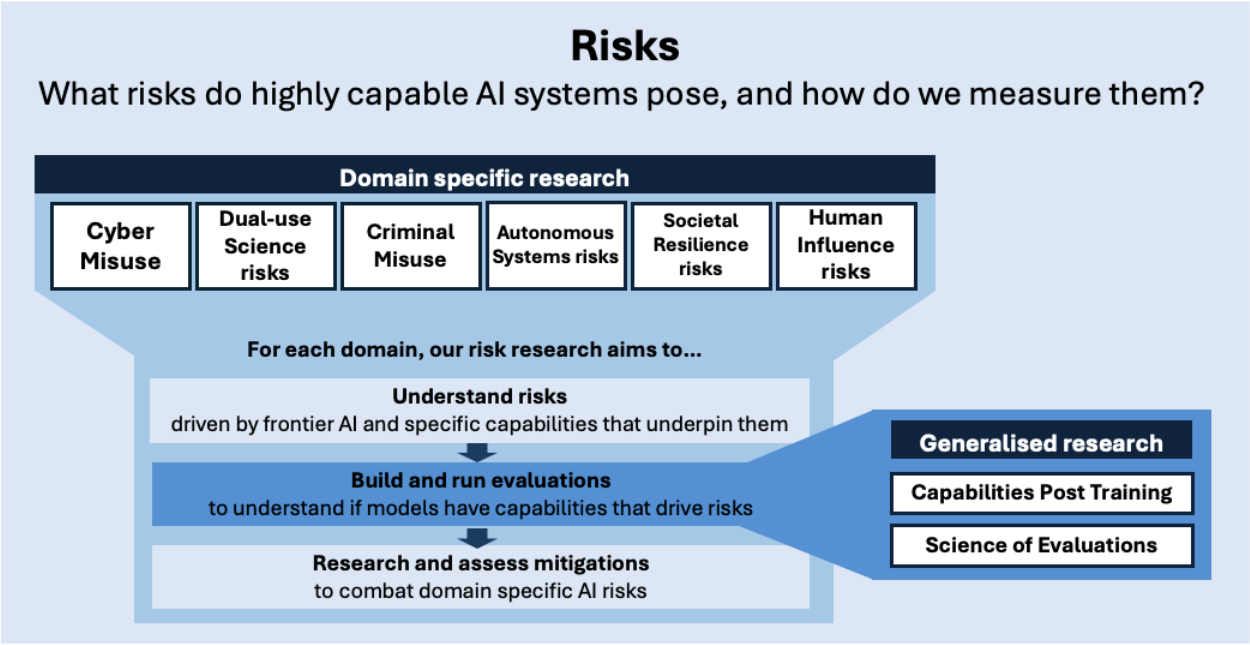
- 背景
- 継続的評価による継続的改善
- **AI セーフティ <-**
- AI ガバナンス
- 発展的な話題

# AI セーフティ

- AI セーフティとは
- 実践 AI セーフティ
- リスクと効果を考慮し小さく始める
- リスクに対策する
- 独自のデータを定義し評価データを育てる
- 専門家を開発チームの一員にする
- 本番環境でテストする

# AI セーフティとは

- 定義自体の議論が進行中
- AI 事業者ガイドラインでは「安全性」を定義
- AISI UK の Research Agenda では 6 種類のリスクを定義




Research Agenda https://www.aisi.gov.uk/research-agenda

# AI セーフティに関する評価観点ガイド

**AI セーフティ評価**

AI システムが AI セーフティの観点で適切であるかどうか見定めること。

※AI セーフティの観点とは、「人間中心」、「安全性」、「公平性」、「プライバシー保護」、「セキュリティ確保」及び「透明性」を重要要素とした見方である。

AIセーフティに関する評価観点ガイドの公開 - AISI Japan https://aisi.go.jp/output/output_information/240918_2/

# 実践 AI セーフティ

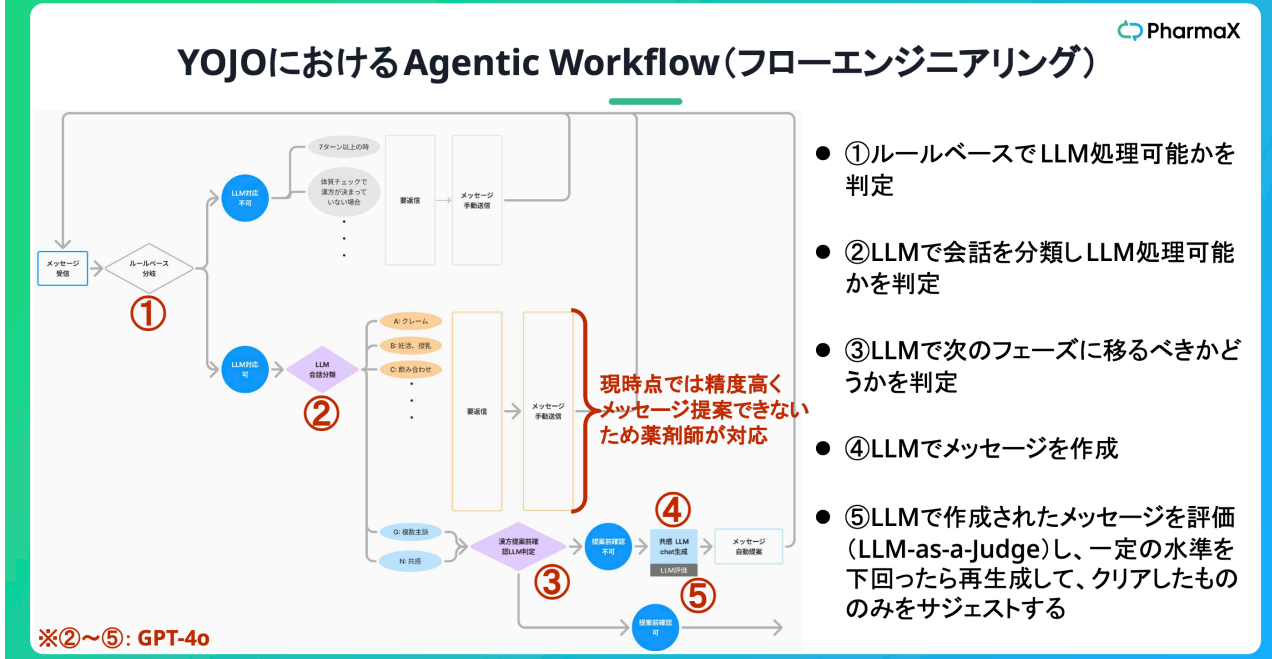
リスクマネジメントの手法を応用

1. ユースケースを列挙
2. ユースケースごとにリスクを分析
3. ユースケースごとに対応 (回避・低減・移転・受容) を決定
4. 安全だと判断できるユースケースに限ってサービスを提供
5. サービスの利用状況をモニタリング

# 事例: PharmaX (YOJO)

- LINE でユーザーが OTC 薬を購入できるオンライン薬局
- 問い合わせをルールと LLM で分類
- ワークフローで安全に対応できるユースケースのみを対応

AIエージェントの継続的改善のためオブザーバビリティ
https://speakerdeck.com/pharma_x_tech/aiezientonoji-sok-de-gai-shan-notameobuzababiritei

# リスクと効果を考慮し小さく始める

- ユースケースを安全性と効果の2軸で分類
  - 安全性: サービス提供、人にフォールバック、対応不可
  - 効果: システム化が進んでいない、人の経験や勘に頼っている、などで判断
- 安全かつ効果の高いユースケースを特定し推進する
- ユースケースを特定し、ホワイトリスト形式で安全なユースケースのみサービス提供

# OpenAI: AI 活用ユースケースの発見と展開

**01**

AI が得意とすることを理解し、ビジネスでの活用機会を見つける— 自社の業務の中で、AI を活かせる場面を見極めます。

**02**

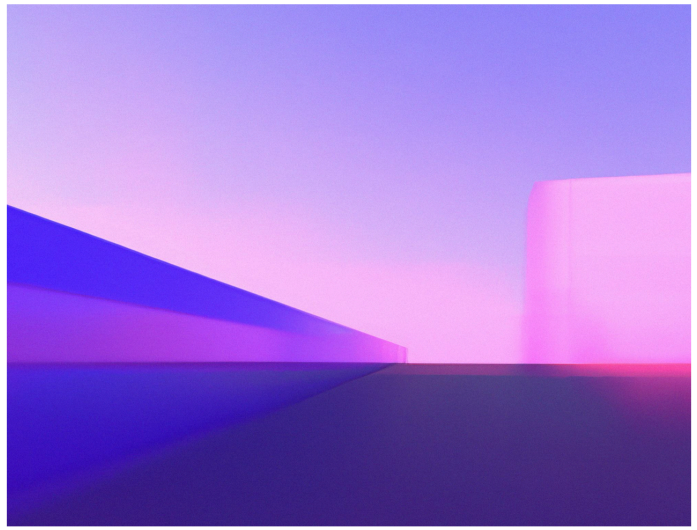
全社的な気づきを促すために、基本的なユースケースを従業員に伝える— 各部門で活用のヒントが広がるよう、社員が AI 活用に慣れる機会をつくります。

**03**

効果が大きいユースケースを収集し、優先順位をつけて取り組む— 成果が見込めるものから順に導入することで、活用効果を最大化します。

OpenAI

## AI 活用ユースケースの発見と展開



先進企業が注力する AI の活用領域とは

ビジネス向けガイドとリソース | OpenAI https://openai.com/ja-JP/business/guides-and-resources/

# リスクに対策する (1/2)

## Human-in-the-Loop

- エンドユーザー向けではなく、社内の専門家向けにサービス提供
- 出力結果を直接提供することを禁止し、一度人手で編集して提供

## 安全な環境を用意

- 第三者に社内情報が漏洩しないように、オプトアウト
- セキュリティや監査に堪え、安全に使える環境を用意して展開

# リスクに対策する (2/2)

## 小規模に始めて徐々に拡大

- リスクを特定して徹底的に検証し、小規模にリリース
- 全面的には利用を禁止し、小規模にパイロットとして導入しテスト

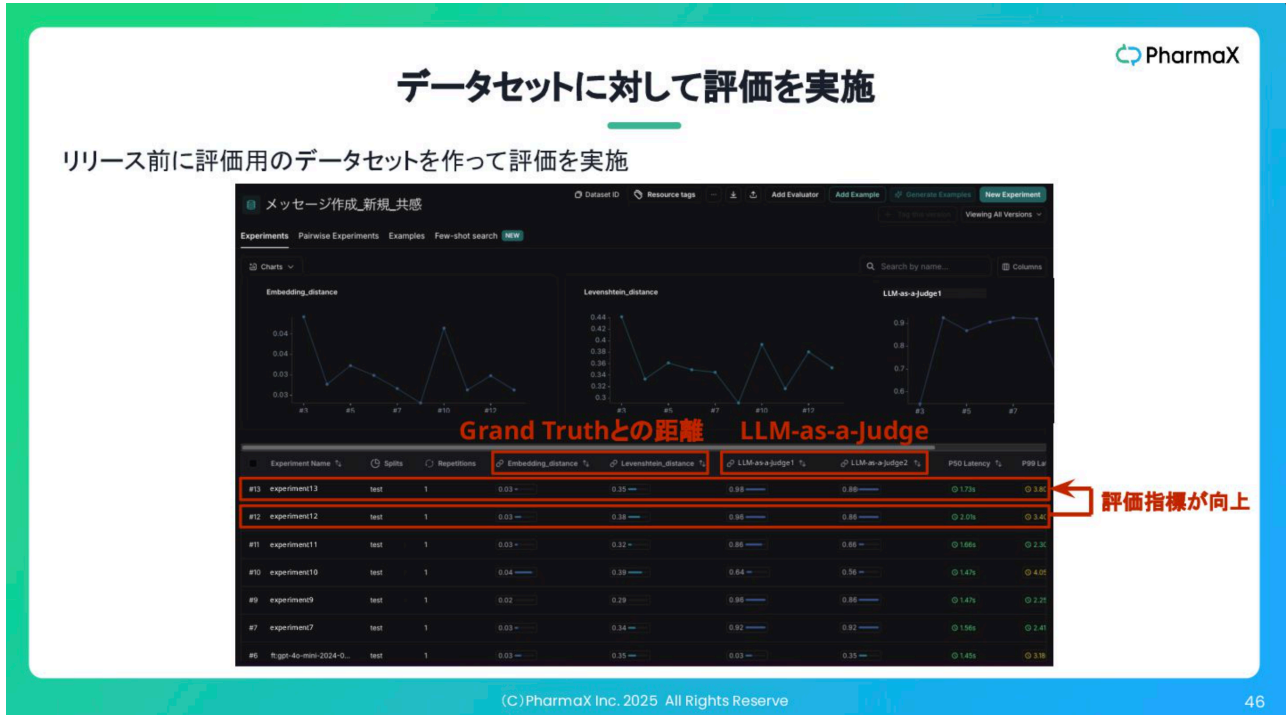
## 社内の体制整備

- ガイドラインやチェックリストを用意し、安全な利用方法を推進

# 独自のデータを定義し評価データを育てる

- 「自分の業務」というベンチマークはない
- 生成 AI に対するユニットテストのように扱う
- 専門家によるレビュー結果を評価データに追加する

AIエージェントの継続的改善のためオブザーバビリティ
https://speakerdeck.com/pharma_x_tech/aiezientonoji-sok-de-gai-shan-notameobuzababiritei




# 事例: ダイキン工業

- 独自のマルチモーダルデータセットを構築
- インターネット上には存在しないデータを用いて、自社の業務知識を持つモデルを作成

ダイキンがサービス支援AIを内製、活用進めた4度の「偶然の出会い」 | 日経クロステック（xTECH）
https://xtech.nikkei.com/atcl/nxt/column/18/00001/10313/

ニュース解説　＋ 連載をフォロー

## ダイキンがサービス支援AIを内製、活用進めた4度の「偶然の出会い」

近岡 裕　日経クロステック

2025.02.26
有料会員限定

全6356文字



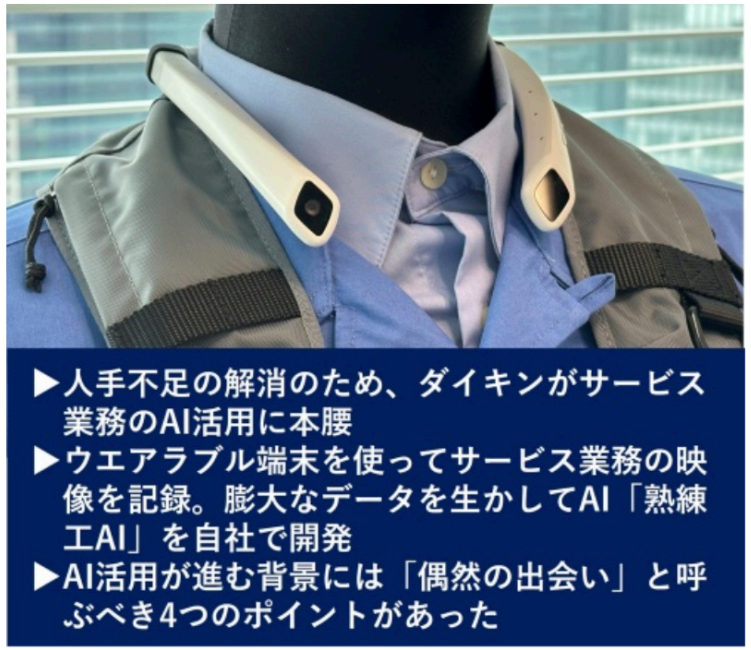
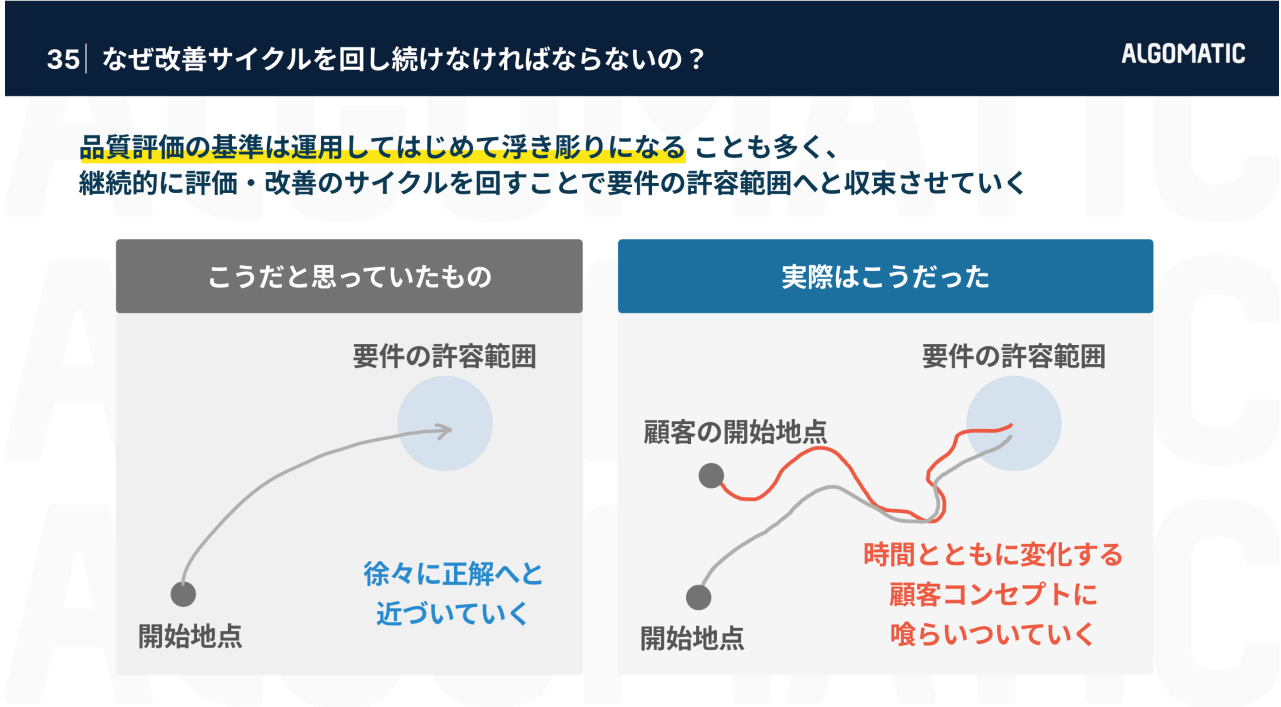



**図1　AI活用に力を入れるダイキン**
狙いはずばり、人手不足の解消。技術的に難易度の高いサービス業務をAI活用の対象に選んだ。実務に生かすために自社でAI「熟練工AI」も開発している。（出所：日経クロステック）
[画像のクリックで拡大表示]

# 専門家を開発チームの一員にする

- 専門家によるレビューや議論の結果をプロンプトに反映
- 専門家もいきなりは自分のノウハウを言語化できないので、イテレーションを回して徐々に言語化

AIエージェントの地上戦 〜開発計画と運用実践 / 2025/04/08 Findy ランチセッション #19 - Speaker Deck
https://speakerdeck.com/smiyawaki0820/08-findy-w-and-bmitoatupu-number-19




# 本番環境でテストする

- どうしても「やってみないとわからない」
- 限定的にリリースして想定外の事象が発生しないか確認する
- モニタリングでリスクと効果を確認

ペアーズにおける評価ドリブンな AI Agent 開発のご紹介
https://speakerdeck.com/fukubaka0825/heasuniokeruping-jia-torihunna-ai-agent-kai-fa-nokoshao-jie

pairs

## AI Workflowを、オンライン評価する

- あるプロダクションの分類タスクのユースケースでやっている例を紹介していく（具体的な使用用途は非公開）
- サンプリングして、ドメインエキスパートに管理画面経由して分類をラベリングしてもらい、それらを非同期BatchでLangfuse カスタムスコアとしてトレースに付与しています
  - 正解率スコアの推移で、性能のデグレをチェックする
  - LLMの判断とドメインエキスパートの判断で不一致のものを週次で確認し、評価データセットに追加しつつ、プロンプトを改善するサイクルを回している

pairs

## AI Workflowを、オンライン評価する

- ペアーズは現在３ヶ国（日本/韓国/台湾）で展開しているが、全ての国でドメインエキスパートにチェックしてもらう体制は取れないことが多く、コストや工数の問題でアノテーションの量の確保も難しいことが多い
- 上記の問題に対応するために、LLM-as-a-Judgeでオンライン評価する仕組みも整えている
  - 独自指標の評価で用いる評価プロンプトの項目は、命令プロンプト本体と似通っていても問題ない
    - 評価の性能と、判断の性能は同一モデルだとしても異なる
  - コストの関係で使えない上位モデルに、サンプリングしてチェックさせるのが一番簡単で効果が出る方法の一つ

# TOC

- 背景
- 継続的評価による継続的改善
- AI セーフティ
- **AI ガバナンス <-**
- 発展的な話題

# AI ガバナンス

- AI ガバナンスとは
- 代表的なガイドライン
- AI ガバナンスにおける典型的な方法論と課題
- 実践 AI ガバナンス

# AI ガバナンス

- リスク管理 + 提供価値の最大化
- アジャイルガバナンス: 組織として学習し続けることを求める
- リスクマネジメントに関するISO標準 (ISO 31000:2018) でもアジャイルの考え方を取り入れている

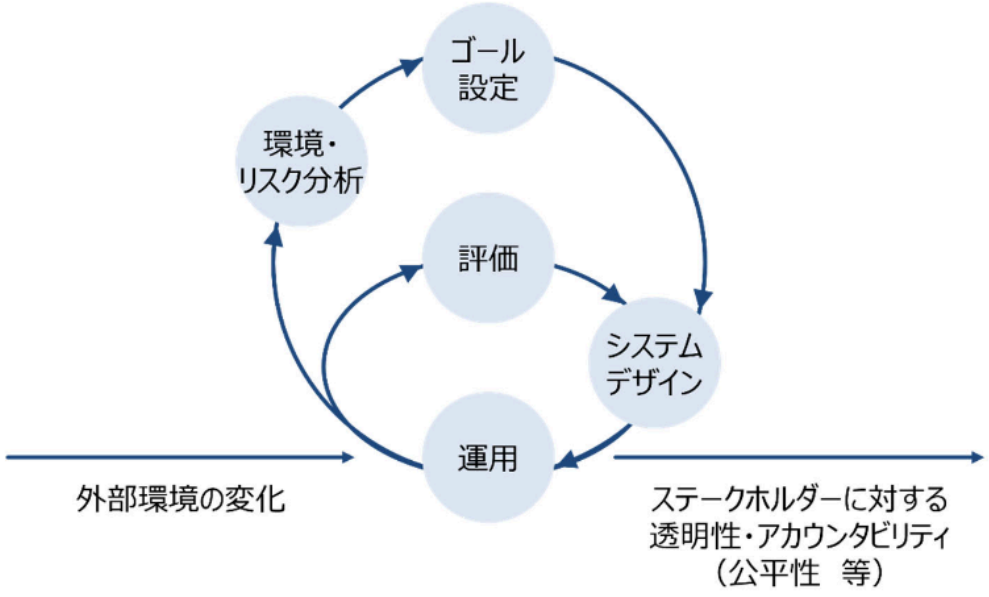



図 6. アジャイル・ガバナンスの基本的なモデル

AI事業者ガイドライン（METI/経済産業省）
https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/20240419_report.html

# AI 事業者ガイドライン

> AI の利活用によって生じるリスクをステークホルダーにとって受容可能な水準で管理しつつ、そこからもたらされる正のインパクト（便益）を最大化することを目的とする、ステークホルダーによる技術的、組織的、及び社会的システムの設計並びに運用。

AI事業者ガイドライン（METI/経済産業省）https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/20240419_report.html

# AI ガバナンスにおける典型的な方法論と課題

- ガイドラインの整備
- 組織体制の構築
- 典型的な課題

# ガイドラインの整備

- AI の活用にあたって遵守すべき事項をまとめたドキュメントを作成
- 組織全体に適用する活用指針をもとに、実業務に適用するためのガイドライン、具体的なチェックリストといった階層構造をもたせることが一般的
- 活用指針は透明性のために外部公開することも

# リクルートAI活用指針

1. ユーザーの機会拡大や社会の発展に繋がるAIの活用
2. 公平性・公正性の追求
3. プライバシー保護とセキュリティ強化
4. アカウンタビリティの向上
5. ステークホルダーや有識者との対話

## リクルートAI活用指針

この「リクルートAI活用指針」は、株式会社リクルートおよびその子会社ならびに株式会社リクルートホールディングス配下のHR関連事業子会社（以下「当社グループ」という）が、AIを活用するにあたり、実現したいことや取り組みの指針を示すものです。

---

### 序章

私たちは、当社グループの基本理念にある「一人ひとりが輝く豊かな世界の実現」に向け、就職、転職、いろいろな働き方の支援、結婚、住まい探しといった人生の大きな決断から、旅行や買い物、食事などの日々の楽しみまで、さまざまなサービスを通じて「まだ、ここにない、出会い。」をお届けしています。

いつでもどこでも情報を得られるようになった今だからこそ、「まだ、ここにない、出会い。」を、桁違いに速く、驚くほどシンプルに、もっと身近にしていくために、AIを活用していきたいと考えています。

一方で、AIの活用には、社会に負の影響をもたらす懸念が存在し、さまざまな議論が継続的に行われていることを認識しています。私たちは、ユーザーの機会拡大や社会の発展に向け、以下に定める指針のもと、AIを活用していきます。

---

### 第１条 ユーザーの機会拡大や社会の発展に繋がるAIの活用

私たちは、基本理念として、「新しい価値の創造を通じ、社会からの期待に応え、一人ひとりが輝く豊かな世界の実現を目指す」ことを掲げています。ユーザーの利便性の向上、機会や可能性の拡大、そして社会の健全な発展のためにAIを活用します。

---

### 第２条 公平性・公正性の追求

公平性・公正性は、明確で統一された定義やゴールを定めることが難しく、不断に追求するものであると認識しています。私たちは、AIの活用においても、人権侵害、不当な差別の助長または多様性の排除などの問題に、真摯に取り組み続けます。

リクルートAI活用指針 | 株式会社リクルート https://www.recruit.co.jp/privacy/ai_policy/

# インテージグループAI利活用指針

1. 安全性の重視・セキュリティの強化
2. 公平性の確保
3. プライバシー保護・コンプライアンス順守
4. 透明性の追求
5. 品質の追求

## インテージグループAI利活用指針

インテージグループは、「Know today, Power tomorrow 知る、つなぐ、未来を拓く」というビジョンのもと、お客様と生活者を繋ぐ架け橋として、豊かで可能性の広がる社会の実現を目指しています。

こうしたビジョンの実現に向けて、社会に多様なイノベーションをもたらすと期待されるAIの利活用を積極的に進めていきます。

また、インテージグループは「人間中心のAI原則」に賛同しています。人間の尊厳と個人の自律を尊重し、社会の持続可能性を重視する理念を守るため、以下に定める指針のもとAIを活用していきます。

### 1．コンプライアンス

AIの利活用に係る法令を遵守するとともに、社会規範・社会道徳、ステークホルダーの利益・要請に従います。

### 2．安全性の重視

利用者および第三者の生命・身体・財産、精神及び環境へ危害を及ぼすことがないよう安全性を重視します。

### 3．セキュリティの確保・プライバシーの保護

AI の安全な利活用を行うため、「インテージグループ情報セキュリティ基本方針」のもと、技術の進化に即した合理的な対策を講じ、データの適切な扱いとセキュリティの確保・維持に努めます。特に個人情報については「個人情報保護方針」のもと、プライバシーの保護に努めます。

### 4．公平性の確保

AI利活用により、人種、性別、国籍、年齢、政治的信念、宗教などの多様な背景を理由とした不当な差別がおきないよう公平性の確保に努めます。

### 5．透明性の追求

利用するAIの特性や用途を考慮し、出力判断結果の説明可能性の追究等、透明性の確保に努めます。また、ステークホルダーに対して適切な情報提供に努めます。

### 6. 品質の追求

AIモデル・AIに投入する学習データ・AI生成結果それぞれについて品質基準・評価指標・検証プロセスを整備するなど、AI利活用サービスの品質維持向上に継続的に取組みます。

日本国および国際的に議論が継続しているAIガバナンス原則を尊重すること等を考慮し、このAI利活用指針は適宜見直しを行います。

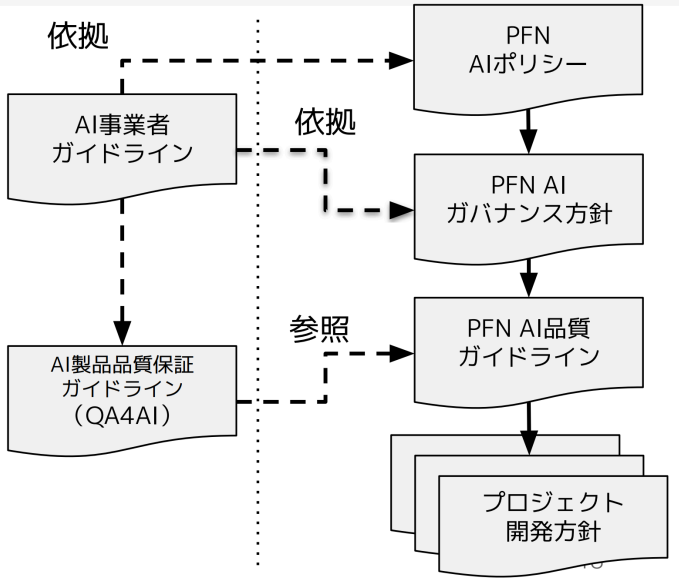
制定　2024年4月1日

インテージグループAI利活用指針　|　企業情報　|　インテージホールディングス
https://www.intageholdings.co.jp/company/ai_policy/

# PFN 規定・ガイドライン体系

- 階層的なドキュメントを用意し、段階的に個々のプロジェクトにあわせて具体化
- 最終的には「プロジェクト開発方針」を個々のプロジェクトで策定

## 規程・ガイドライン体系




PFNのAIに関する基本方針

AIに関する適切な企業活動のための体制・規程

プロジェクトが守るべきルールや手続き

ガイドラインの対応可否・選択理由などを記述

Preferred Networks

## AI Policy

https://www.preferred.jp/ja/company/aipolicy/

### 1. 透明性（正しく伝える）

私たちは、その技術自体が何をするか、何ができるか、何ができないかを可能な限り明確にします。それがどのような仕組みで動くかを、論文その他の出版、ブログ、ドキュメント等を通して、様々なレベルのステークホルダが理解できるよう、丁寧に説明します。

### 2. リスク（適切に怖がる）

私たちは、その技術を使うことによって、社会の様々な場面で起きうるリスクを、私たちの想像力が及ぶ限り考慮します。リスクには、定量化できないもの、想定しにくいものがあり、また、ある時点での社会通念ではリスクではなかったものが、その後にリスクと考えられるようになることもあります。新技術は常に便益とリスクを伴います。私たちは、社会との対話を通して、そのようなリスクに対応していく努力を惜しみません。

### 3. インテグリティ（見たくないものを、見る）

新技術の開発や利用にあたっては、時として私たちに都合の悪い発見があるかもしれません。私たちは、そのようなものから目を背けることなく、誠実に対応します。

Preferred Networksにおける大規模言語モデル開発と活用での AI安全性の考え方 https://www.digiarc.aist.go.jp/event/4th_grand_canvas/pdf/20250304-4th-grand-canvas-03-ohno.pdf



Preferred Networks

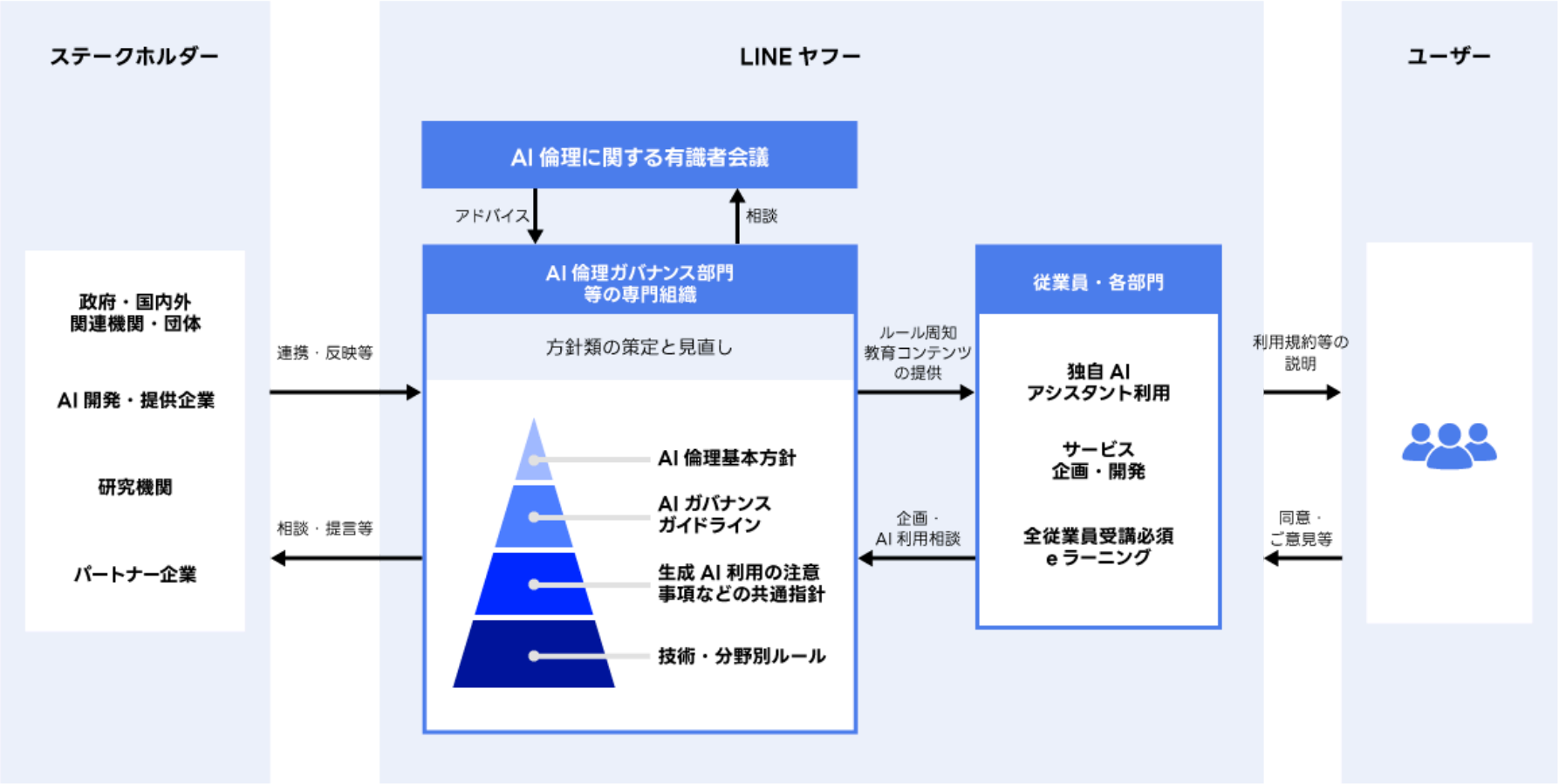
# 組織体制の構築

- AI ガバナンスを推進するための体制構築
- 推進するための専門チームを設けることが多い

# 株式会社リクルートの例


AIガバナンスの取り組み | 株式会社リクルート https://www.recruit.co.jp/privacy/ai_policy/ai_governance/

# LINEヤフー株式会社の例




責任あるAIへの取組み｜LINEヤフー株式会社 https://www.lycorp.co.jp/ja/sustainability/esg/social/responsible-ai/

# 典型的な課題

## リソース不足

- AI ガバナンス専門のチームが組織に1つなので忙しい
- AI という変化の激しい分野で全方位に目を配る必要がある

## ルールの周知と徹底

- AI ガバナンス専門のチームはタイムリーにガイドラインを書き換えるものの、すべてのチームが見てくれるわけではない
- 誰もが取り組めるようになった結果、リスクを知らずに企画立案して推進してしまうことも

# 実践 AI ガバナンス

- うまくいく/うまくいかないチームの違い
- AI ガバナンスのミッション
- AI ガバナンスの提供する機能

# うまくいく/うまくいかないチームの違い

- AI ガバナンスのミッション
- AI ガバナンスの提供する機能

# AI ガバナンスのミッション

- どんな使い方をしても大丈夫なAIを用意して展開しようとすると失敗する (実現できない)
- AI セーフティについては組織内の各チームに任せる
- AI ガバナンスを専門で行うチームでのみ低減できるリスクは「AI を活用しない機会損失リスク」
- うまくいっている AI ガバナンスチームのミッションは「AI の組織的な活用」

# セガでの AI ガバナンス

- CEDEC 2025 講演資料より引用 (pp.11-12)
- もはや AI を使うのが当たり前で気がついたら誰もが使っているという前提
- 「AIを使わないことはありえない」

## 生成AIとの共創に向けて

**利便性と創造性の追求**

- 開発現場での効率化・省力化は現実的な価値
- 一方で、生成AIは新たな発想・体験を生む創造の触媒にもなり得る

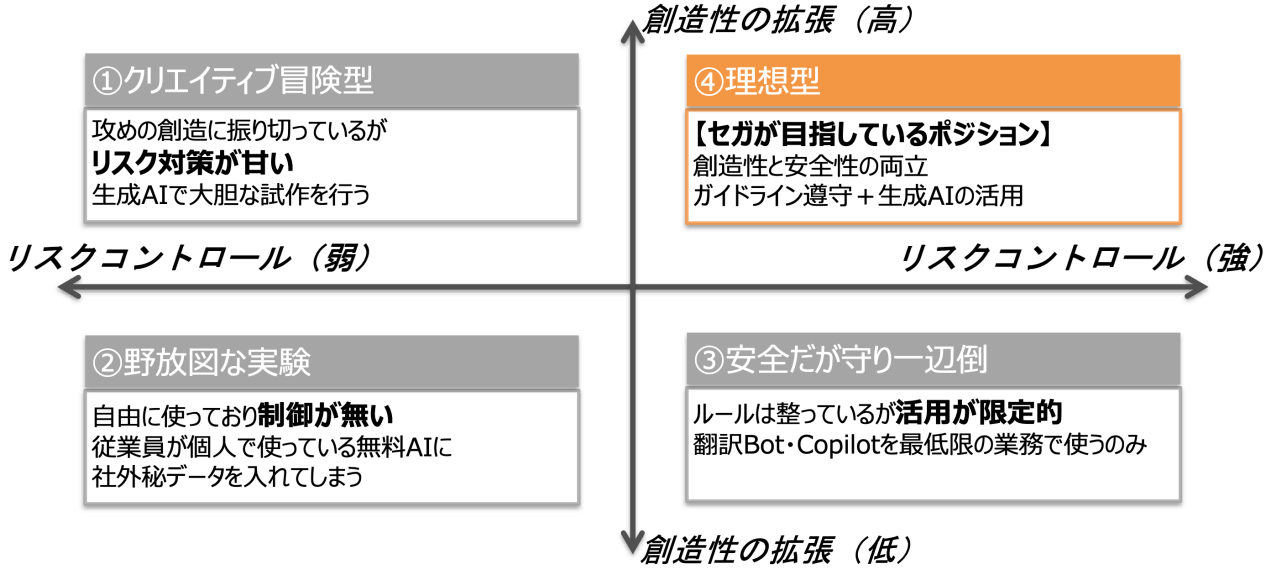
**だからこそ、「誰でも安心・安全に使える環境」が必要**

- ガイドラインやルールの整備だけでなく、
  実態の可視化、ログ管理、判断支援など「運用設計」が要
- 一部の専門家だけでなく、**全員が迷わず使える世界へ**

**創造を支え、リスクに備える**

CEDEC2025

## 「創造性の拡張」と「リスクコントロール」




CEDEC2025

安心安全に生成AIを使おう！社内で運用中の生成AIのガバナンスをご紹介
https://cedil.cesa.or.jp/cedil_sessions/view/3147

# AI ガバナンスの提供する機能

1. 統合的なAI相談窓口
2. AI 活用支援
3. 教育

# 1. 統合的なAI相談窓口

- CEDEC 2025 講演資料より引用 (pp.17, p.19)
- AI ガバナンス担当チームは、組織の中でもっとも活用事例に詳しいチーム
- どんな問い合わせにも答えることで信頼を勝ち取る

安心安全に生成AIを使おう！社内で運用中の生成AIのガバナンスをご紹介
https://cedil.cesa.or.jp/cedil_sessions/view/3147

## 生成AI委員会の設立と構成

**生成AI委員会とAIタスクフォースがグループを支援**

**生成AI委員会**

事務局
■窓口/調査
有志が対応

構成
- IT部門
- 教育
- セキュリティ
- 法務部門
- セガ AIタスクフォース

・技術調査
・ナレッジの蓄積
・ガイドライン更新

**セガ AIタスクフォース**

複数部門を横断したメンバーで構成

・各署からの相談対応
・各種AIツール開発
・最新AI情報取集及び調査
・デューデリジェンス

CEDEC2025

## ガイドライン概要

**ガイドラインのポイント**

**利用審査ベースで現状把握・情報収集**
・各部門からの問い合わせを通じて実情を把握
・実運用の課題やニーズを反映した運用改善に活用
・ガイドライン改定の根拠としてフィードバックを蓄積

**外部ツールの利用について柔軟な判断を支援**
・ツールの種類や変化が早く、画一的な基準では対応しづらい
・過度な「使用禁止」ではなく、都度線引きを行う
・「利用不可」とは書かずに、適切な利用の支援に注力

**企業間契約の社内向けサービスは利用審査不要**
・法務やセキュリティを経て社内契約済のツールは例外扱い
・社員が安心して利用できる「明示的なホワイトリスト」対象
・問い合わせ不要、教育や周知にリソースを集中可能

CEDEC2025

## 2. AI 活用支援

- AI ガバナンス担当の持つ専門知識は、組織内の別チームは持っていない
- チームトポロジーの「イネイブリングチーム」
- 短期間の密なコラボレーションで組織内のチームが AI 活用できるよう支援

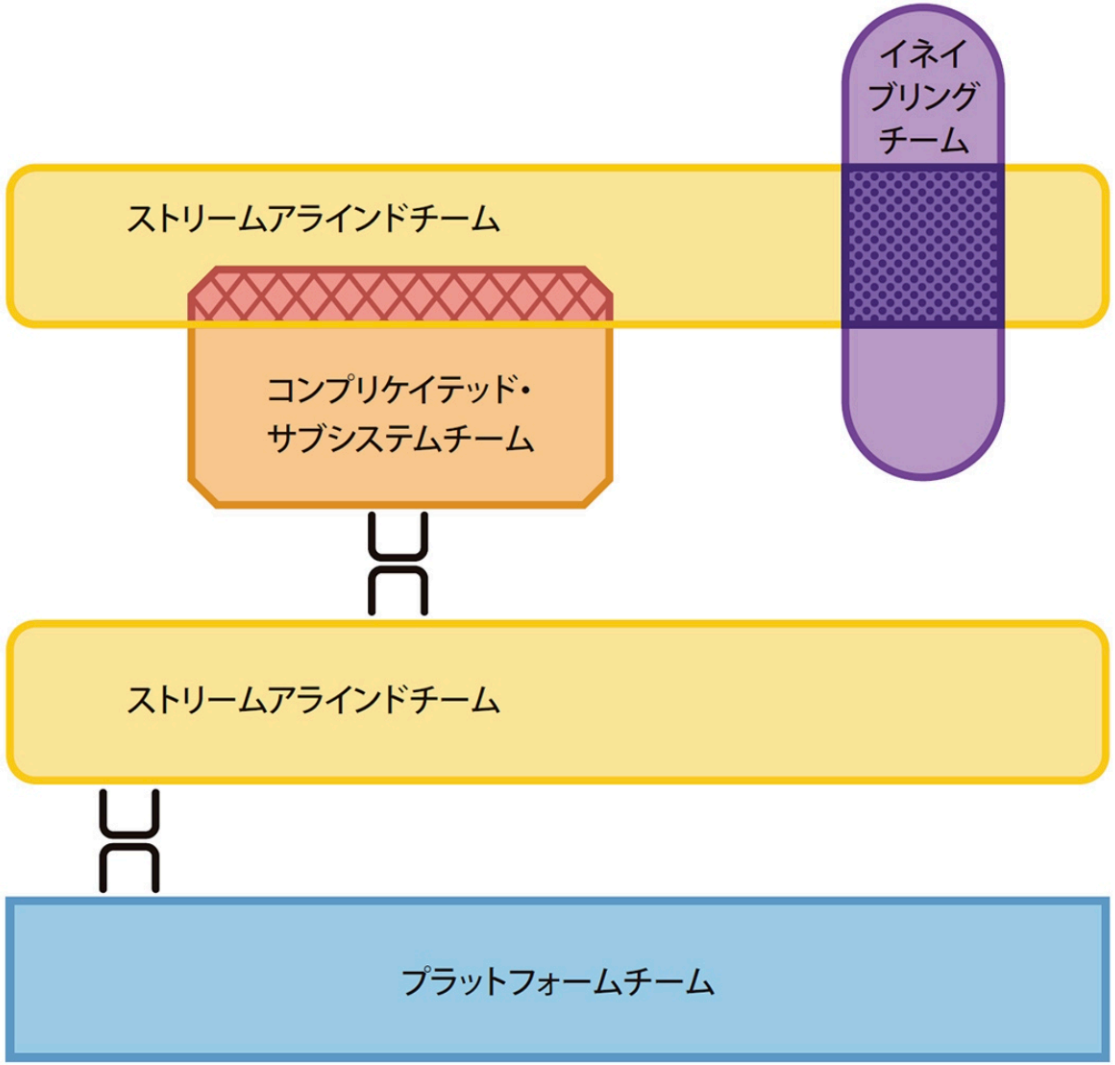


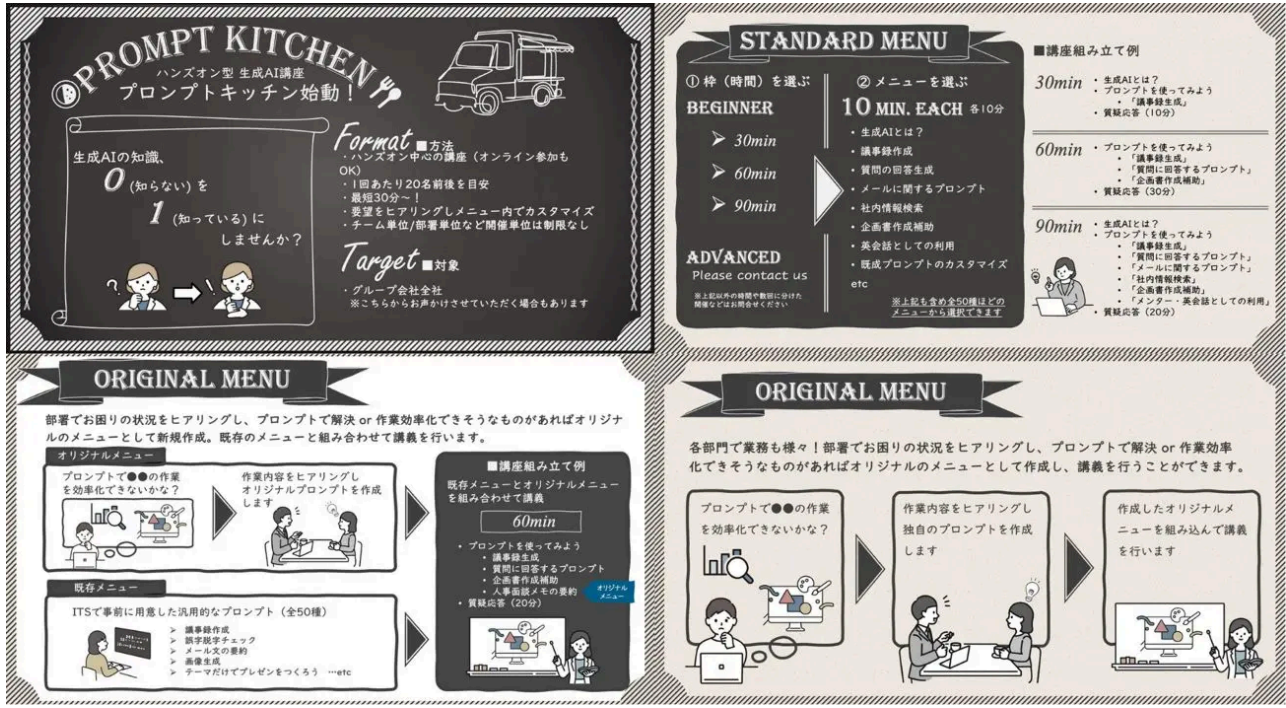

図 7.5　4つの基本的なチームタイプのためのインタラクションモードの基本形

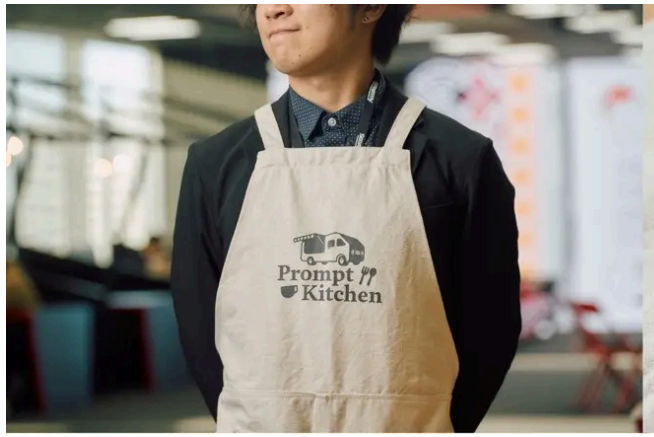
ストリームアラインドチームは X-as-a-Service またはコラボレーションを使用する。イネイブリングチームはファシリテーションを使用する。プラットフォームチームはプラットフォームを利用するチームのために X-as-a-Service を使用する。

マシュー・スケルトン, マニュエル・パイス 著 原田　騎郎, 永瀬　美穂, 吉羽 龍太郎 訳「チームトポロジー」日本能率協会マネジメントセンター 2021年

# プロンプトキッチン

- 「生成AIの活用スキルを従業員に定着させるための実践的な研修プログラム」
- 各部署に行ってニーズを掴み、業務で活用できるプロンプトを一緒に書く



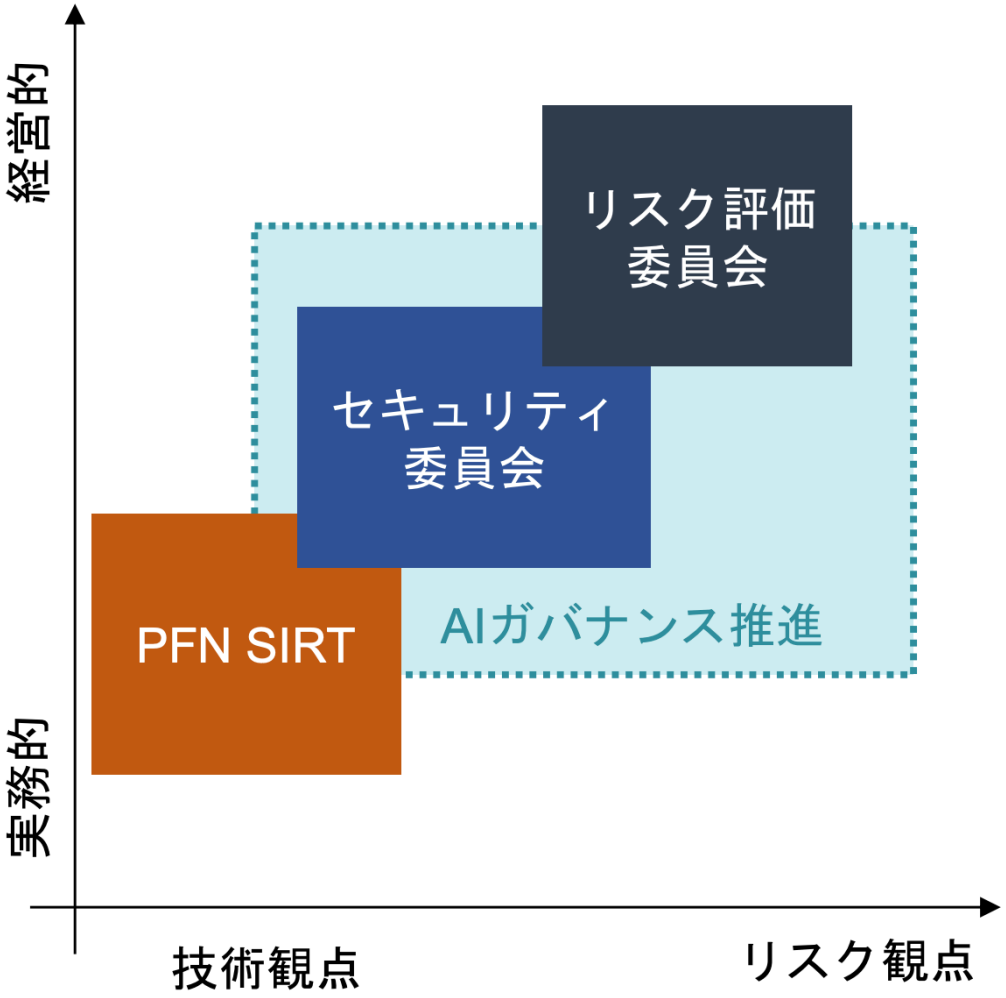





セガサミーG、生成AI徹底活用の仕掛け人・石森拓郎氏に訊く ～IT部門統合から"攻め"への転換、感動体験を創る技術戦略～ | データで越境者に寄り添うメディア データのじかん https://data.wingarc.com/sega-sammy-ishimoritakuro-80385

# PFN のガバナンス体制

- AI リスクを評価する「リスク評価委員会」
- プロジェクトにおける AI 活用を推進できるよう「AIガバナンス推進チーム」が支援




AIベンダーにおけるAIセキュリティ・ガバナンスへの取組 - Speaker Deck
https://speakerdeck.com/pfn/20250717-awt-2025-tokyo-pfn-ai-governance

## 3. 教育

- AI セーフティを組織内の全チームに実施してもらう必要がある
- 権利侵害などのリスク感度を組織的に向上させる必要がある
- 全組織に対して教育プログラムを提供し、リテラシーを向上させる

# LINEヤフーでの取り組み

- 生成AIの利用に伴うリスクを把握するため、全従業員に研修を実施
- 社内のAIアシスタントを使うためには必ず受講しなければならない


| | |
|---|---|
| **主要リスクの学習** | 情報漏えい、権利侵害、不正確な出力、差別的な出力、プライバシー |
| **プロンプトの学習** | 出力品質の向上（プロンプトの良い例・悪い例、テンプレートの提示） |

# TOC

# 発展的な話題

- ソフトウェアエンジニアリングの再発見
- Vibe Coding / Spec Driven Development
- KISS (Keep It Simple Stupid) & リーンソフトウェア開発
- コンウェイの法則

# ソフトウェアエンジニアリングの再発見

| Agent 開発手法 | ソフトウェア開発手法 |
|---|---|
| 小さくはじめて育てる | リーンスタートアップ |
| 高速プロトタイピング | アジャイルソフトウェア開発宣言 |
| 独自データセット定義 | テスト駆動開発 |
| 領域特化エージェント | ドメイン駆動開発 |
| 本番環境でのテスト | DevOps のカナリアリリース |

# Vibe Coding

- 自然言語で指示するだけでコーディングを行わないスタイル
- 非専門家がアプリケーション開発可能に
- プロトタイピングのための試行錯誤が容易に

**Andrej Karpathy**
@karpathy

ポストを翻訳

There's a new kind of coding I call "vibe coding", where you fully give in to the vibes, embrace exponentials, and forget that the code even exists. It's possible because the LLMs (e.g. Cursor Composer w Sonnet) are getting too good. Also I just talk to Composer with SuperWhisper so I barely even touch the keyboard. I ask for the dumbest things like "decrease the padding on the sidebar by half" because I'm too lazy to find it. I "Accept All" always, I don't read the diffs anymore. When I get error messages I just copy paste them in with no comment, usually that fixes it. The code grows beyond my usual comprehension, I'd have to really read through it for a while. Sometimes the LLMs can't fix a bug so I just work around it or ask for random changes until it goes away. It's not too bad for throwaway weekend projects, but still quite amusing. I'm building a project or webapp, but it's not really coding - I just see stuff, say stuff, run stuff, and copy paste stuff, and it mostly works.

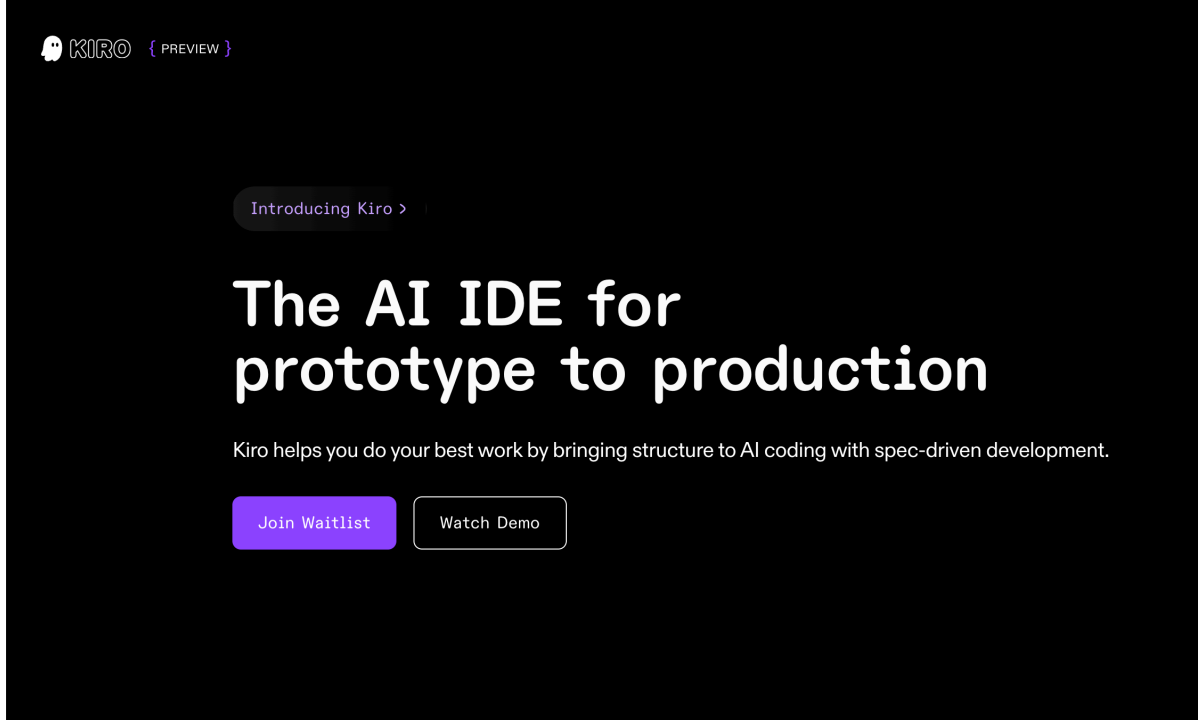
午前8:17 · 2025年2月3日 · **514.1万** 件の表示

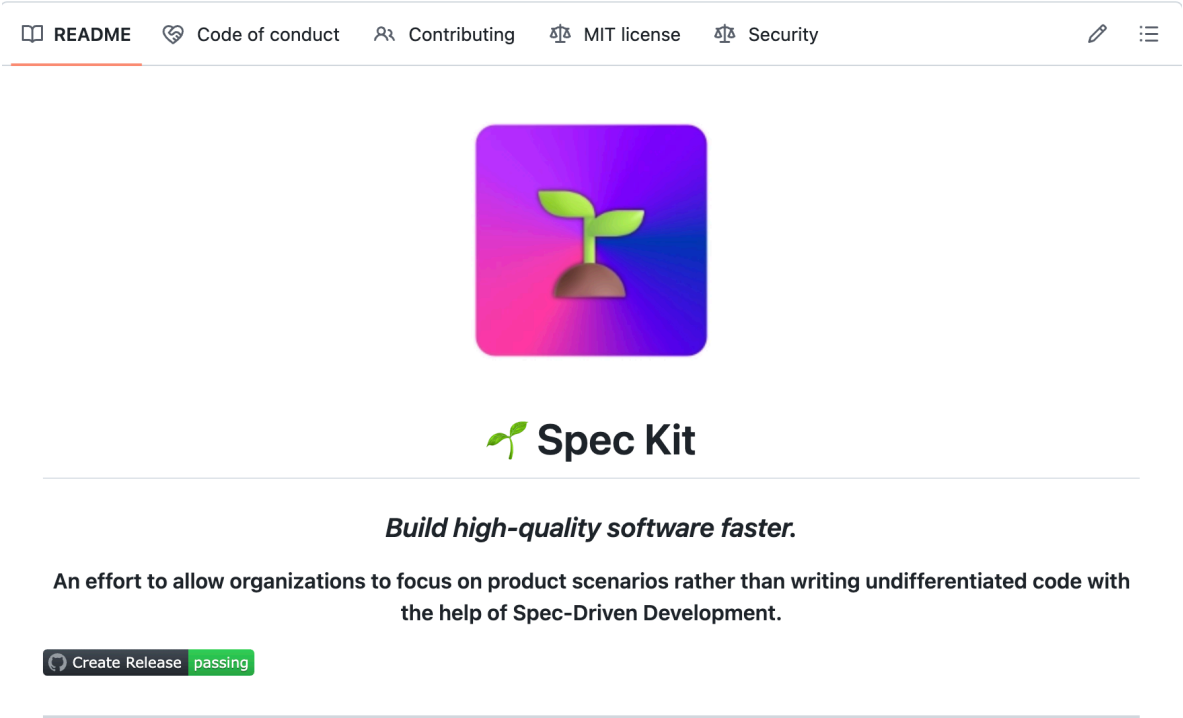
Andrej Karpathy / X https://x.com/karpathy/status/1886192184808149383

# Spec Driven Development

- AWS が開発した Kiro で導入された開発手法
- プロンプトからまずは要件を記述
- 要件に基づきコード生成 (Plan first, then build.)
- GitHub からも Spec Kit が発表された

Kiro: The AI IDE for prototype to production https://kiro.dev/
github/spec-kit: 🌱 Toolkit to help you get started with Spec-Driven Development https://github.com/github/spec-kit



# ウォーターフォール

- Winston W. Royce による開発プロセスの整理
- オリジナルは反復とフィードバックを含む
- SDD という新たな名前でそのように再定義できるか

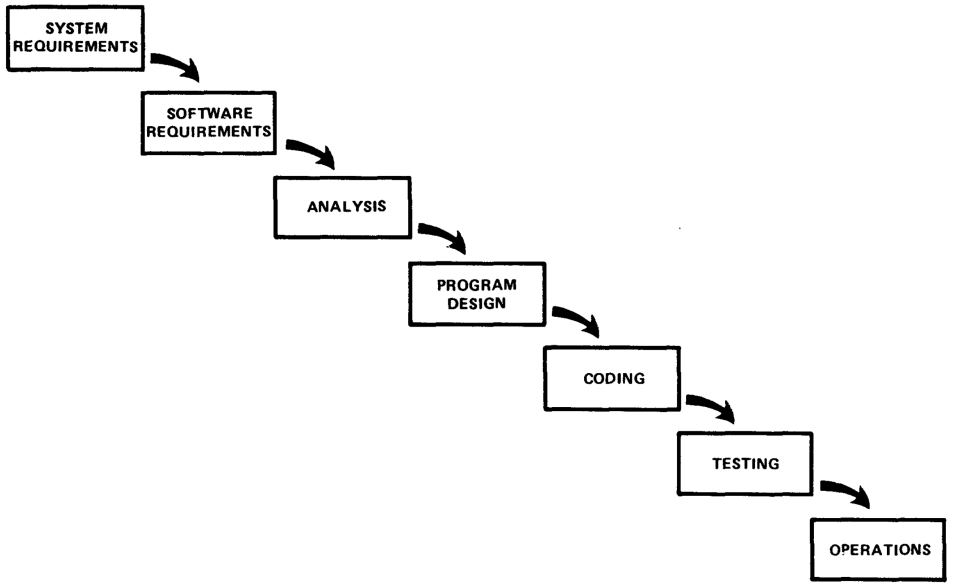



Figure 2. Implementation steps to develop a large computer program for delivery to a customer.

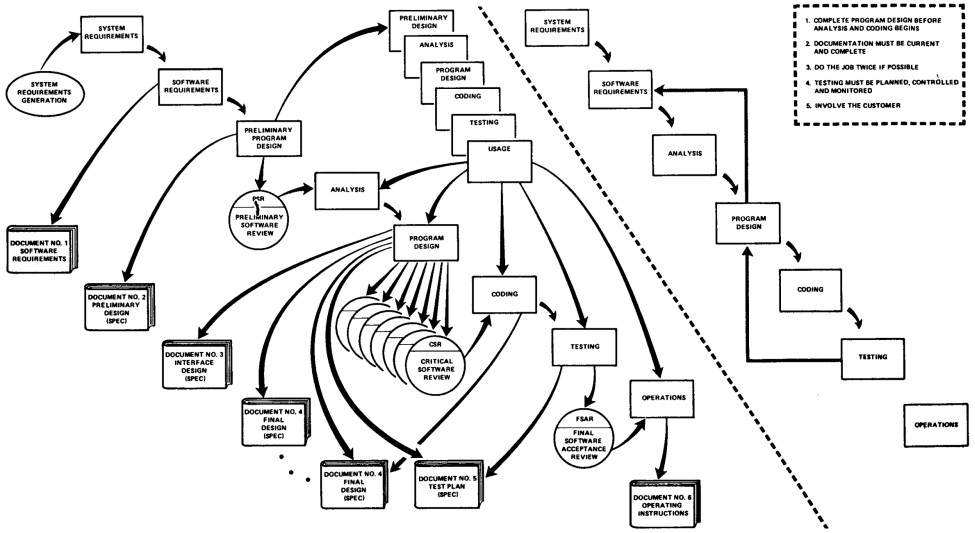



Figure 10. Summary

338

Managing the development of large software systems: concepts and techniques | Proceedings of the 9th international conference on Software Engineering
https://dl.acm.org/doi/10.5555/41765.41801

# KISS (Keep It Simple Stupid) & リーンソフトウェア開発

## オーケストレーション

基本となるコンポーネントが揃ったら、エージェントが効果的にワークフローを実行できるように、オーケストレーションのパターンを検討しましょう。

最初から複雑なアーキテクチャで完全自律型エージェントを作りたいという気持ちを抑えて、少しずつ段階的なアプローチをとる方が成功する確率が高くなります。

一般的に、オーケストレーションのパターンは、大きく 2 つのカテゴリーに分類されます。

エージェント構築実践ガイド https://openai.com/ja-JP/business/guides-and-resources/

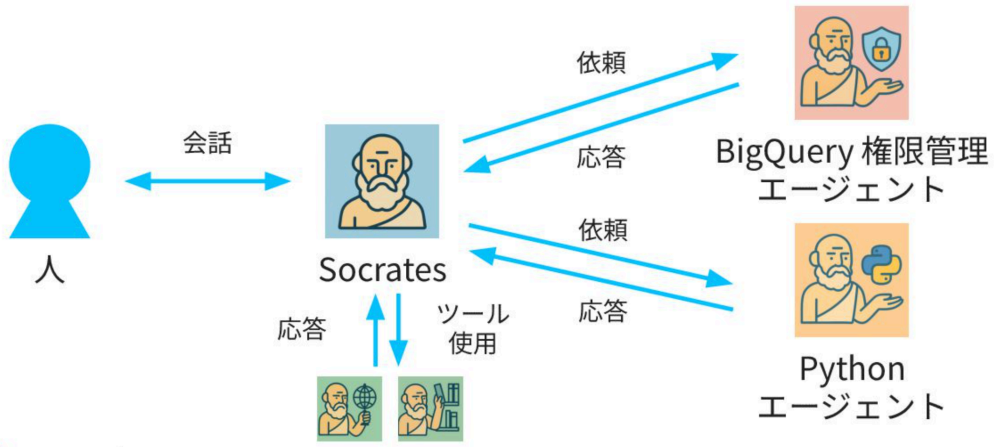
# コンウェイの法則

- チューニングにより「1つのこと(ドメイン)をうまくやる」ものができる
- 「システムを設計する組織は、そのコミュニケーション構造をそっくりまねた構造の設計を生み出してしまう」

メルカリにおけるデータアナリティクス AI エージェント「Socrates」と ADK 活用事例
https://speakerdeck.com/na0/merukariniokerudetaanariteikusu-ai-eziento-socrates-to-adk-huo-yong-shi-li

## Socrates との他の社内ツールとの共存関係

**Socrates は、ADK を基盤としたマルチエージェントシステムとして設計。各エージェントが専門タスクを担当し、協調動作。**



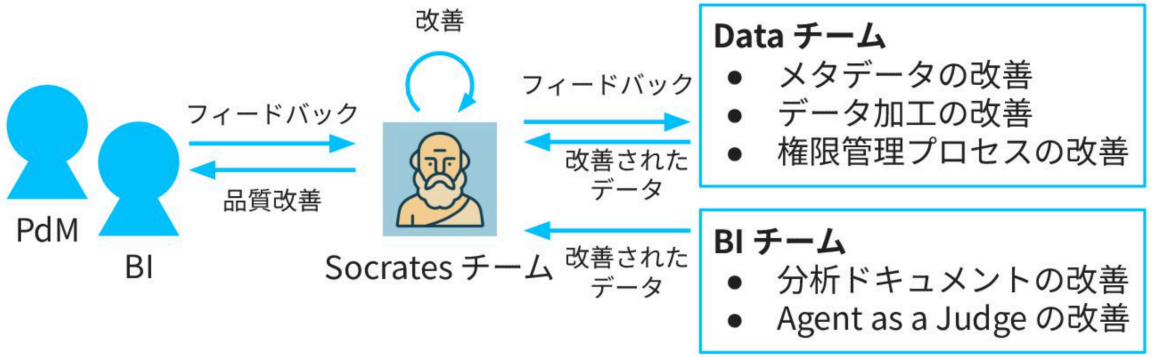

mercari

36

Skip

## Socrates を育むエコシステム：チーム

- **チームの連携：部門横断的な開発・運用体制とフィードバックループ**
  - 利用状況モニタリング、ユーザーヒアリング、エラーログ分析、Agent as a Judge を通じて、プロンプトやツールの迅速な改善を実施




mercari

56

# まとめ

- LLMOps では評価を中心とした継続的な改善により、高速に改善サイクルを回すことで提供したい価値を発見する
- AIセーフティとは、リスクと効果を考慮して安全なユースケースから小さく始め、専門家を巻き込みながら継続的にテスト・改善していく実践的なリスクマネジメント
- AI ガバナンスがうまくいっている組織では AI ガバナンスのミッションを「AI の組織的な活用」としている
- 生成 AI の活用における Software Engineering の再発見が散見されるようになってきた、歴史から未来予測ができるかもしれない