Citadel AI

生成AIを用いるサービス開発の原則

株式会社 Citadel AI

杉山 阿聖

主旨

- 開発者との交流を通じて見えてきた生成AI活用の原則と、残されている課題について共有します

TOC

- **DevOps, MLOps, LLMOps <-**
- 生成AIを利用するシステム開発の原則
- 今後の展望

DevOps, MLOps, LLMOps

- DevOps
- MLOps
- LLMOps

DevOps: Dev VS Ops

- Dev: 顧客に新しい価値を早く提供したい、多少不安定になるかもしれないが運用が頑張れば良い
- Ops: 顧客に安定的に価値を提供したい、新機能の追加で不安定になることは受け入れられない

10+ Deploys Per Day: Dev and Ops Cooperation at Flickr - Slideshare
https://www.slideshare.net/jallspaw/10-deploys-per-day-dev-and-ops-cooperation-at-flickr

DevOps: Dev ❤ Ops

- Dev vs Ops から Dev & Ops に移行しようという提案 (2008)
- 「顧客に価値をすばやく安定的に提供しよう」という提案
- この提案に基づくのが DevOps
- DevOps: Dev と Ops の協調

10+ Deploys Per Day: Dev and Ops Cooperation at Flickr - Slideshare
https://www.slideshare.net/jallspaw/10-deploys-per-day-dev-and-ops-cooperation-at-flickr

DevOps: 継続的改善

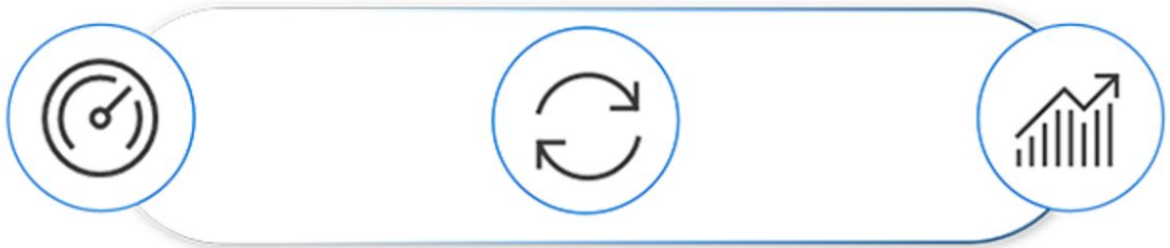
- DevOps の原則のひとつ
- フィードバックサイクルによる改善

What is Continuous Improvement?

Continuously and candidly observing your DevOps process allows teams to identify possible improvement points.

All improvement requires change, but not all change is improvement. This is why measurement is a critical success factor to organizations using DevOps. As Peter Drucker says, “If you can’t measure it, you can’t improve it.”

The lack of an effective feedback mechanism makes it difficult to improve the impact of apps on business. That's why it's important to create an environment that fosters a learning-centric approach for DevOps Improvement, with a focus on making data-based adjustments.



Measurement	Feedback	Improvement
The measurement should lead to a positive behavior change	Evolve to a practice of continuous learning and feedback	Continuous Improvement on software delivery performance

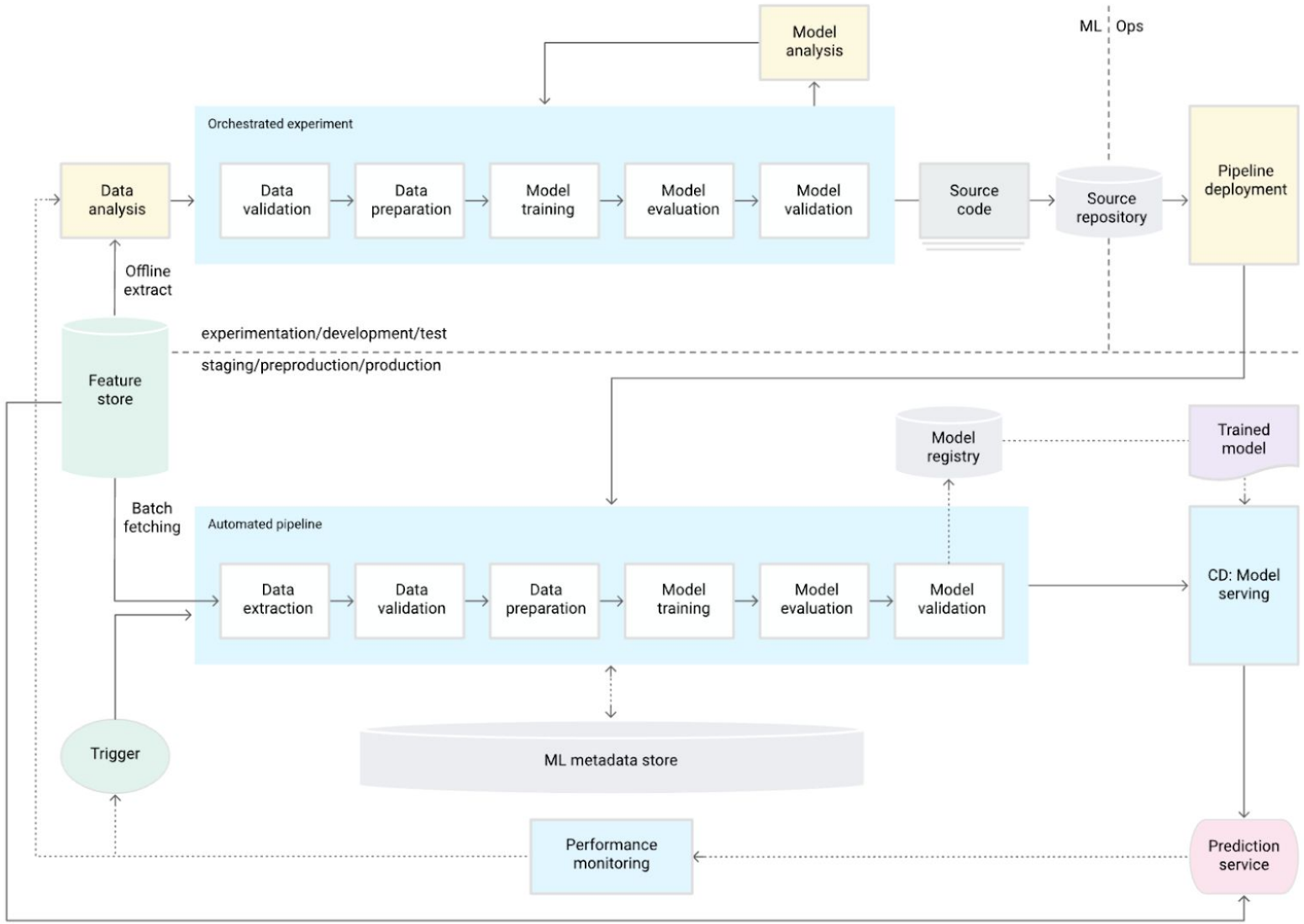
Explore Continuous Improvement - Training | Microsoft Learn https://learn.microsoft.com/en-us/training/modules/characterize-devops-continous-collaboration-improvement/3-explore-continuous-improvement

MLOps: MLOps とは

- 機械学習の成果をスケールさせるためのさまざまな取り組み
- DevOps のプラクティスに基づく
- 機械学習システムを育てる取り組み

MLOps: 継続的な訓練

- MLOps における継続的な改善の実装
- モデルを継続的に訓練して改善



MLOps: Continuous delivery and automation pipelines in machine learning | Cloud Architecture Center | Google Cloud
https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

LLMOps: 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

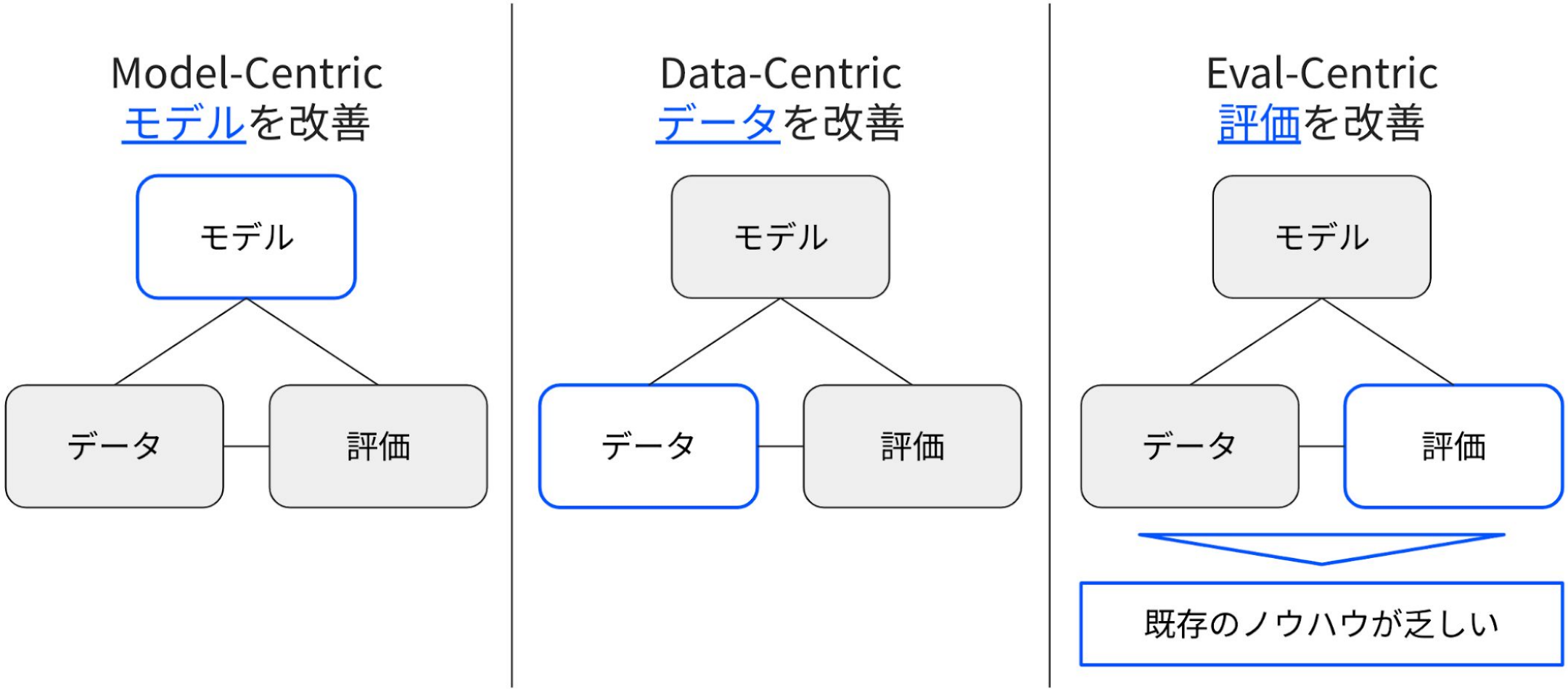
By Albert Lie, Forbes Councils Member.
for Forbes Technology Council, COUNCIL POST | Membership (fee-based)

Published Apr 04, 2025, 07:45am EDT

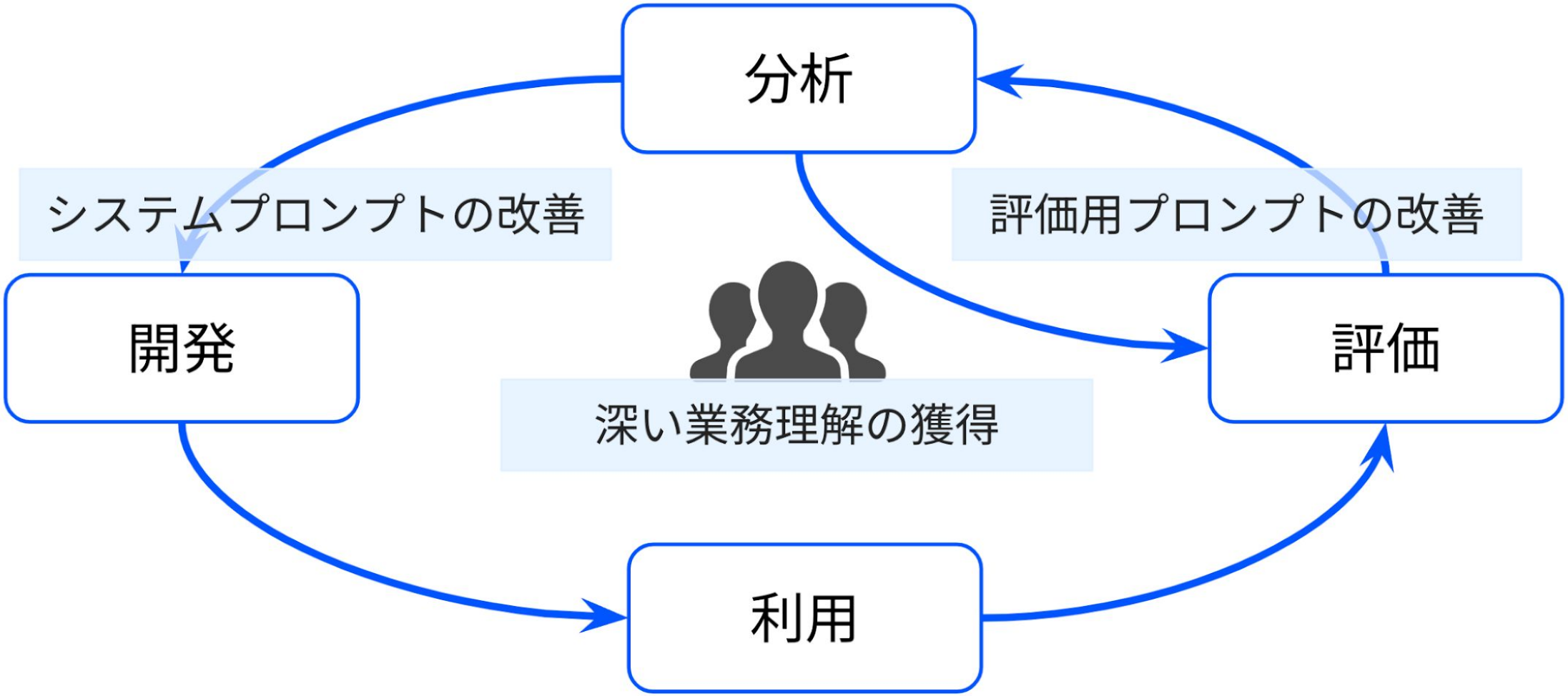
Share Save

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

LLMOps: Eval-Centric AI



LLMOps: 継続的な評価による継続的な改善



TOC

- DevOps, MLOps, LLMOps
- **生成AIを利用するシステム開発の原則 <-**
- 今後の展望

生成AIを利用するシステム開発の原則

1. リスクと効果を考慮し小さく始める
2. 高速にプロトタイピングする
3. 独自のデータを定義し評価データを育てる
4. 専門家を開発チームの一員にする
5. 本番環境でテストする

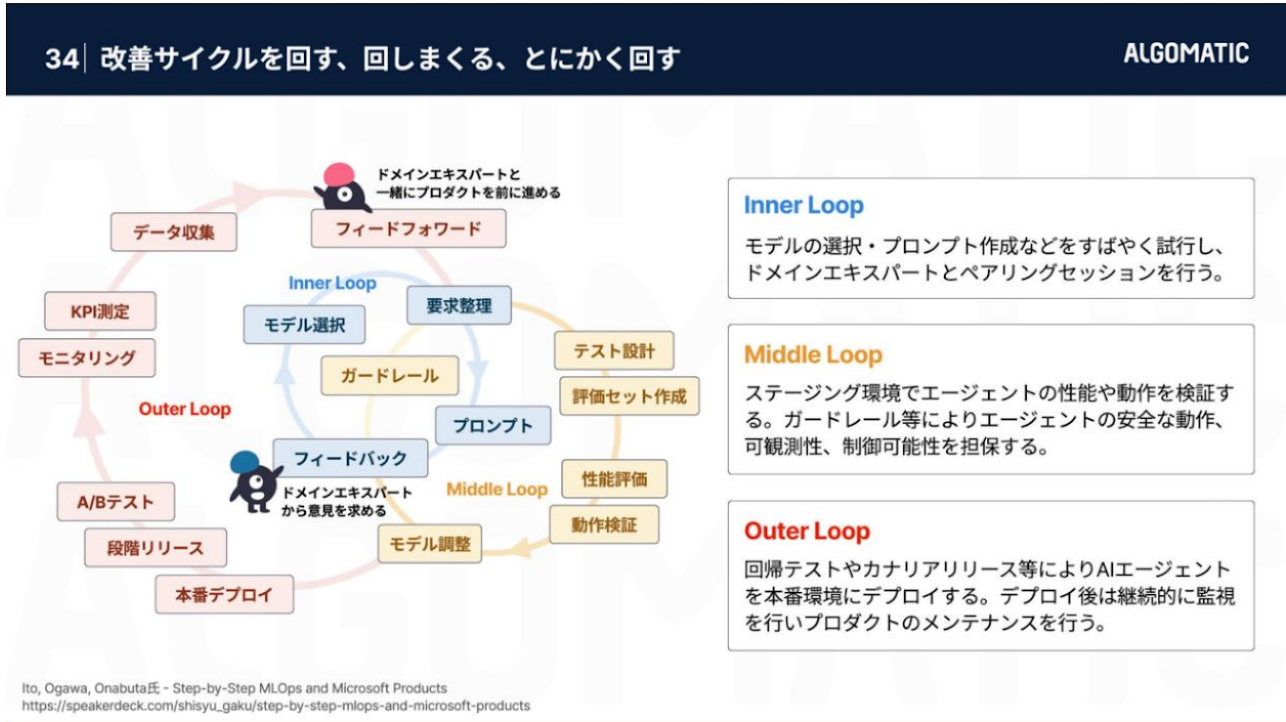
1. リスクと効果を考慮し小さく始める

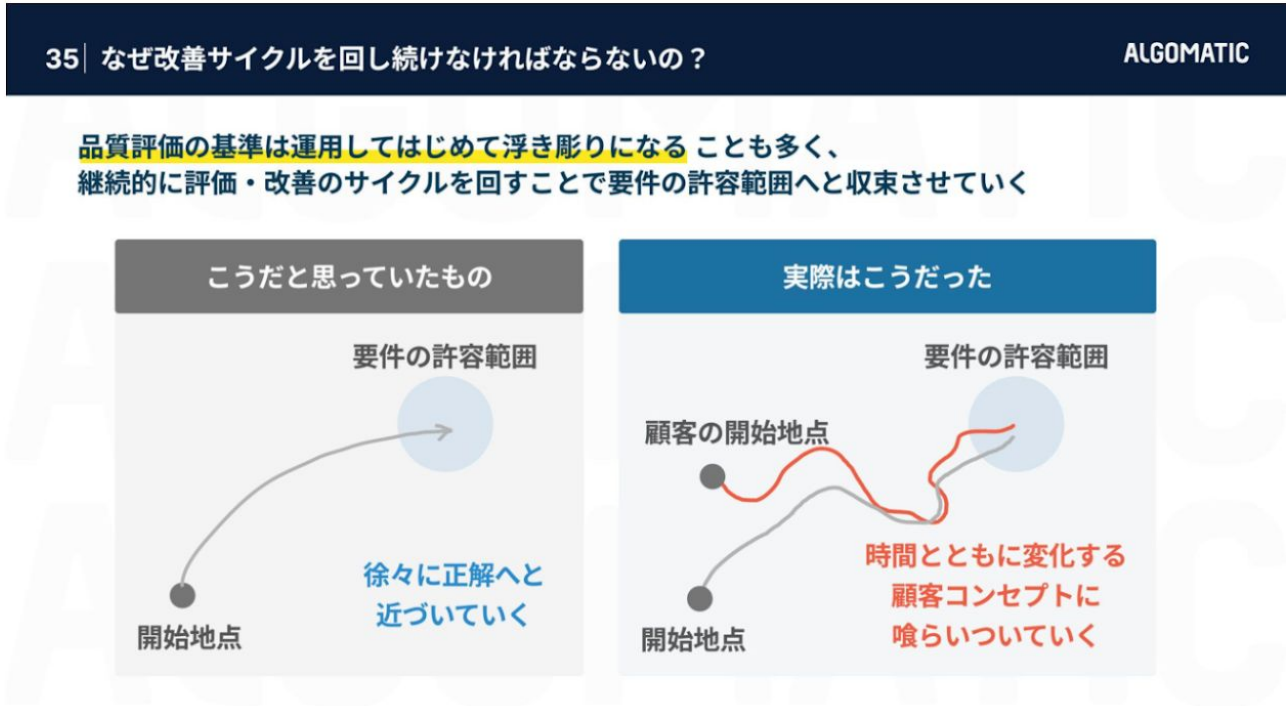
- ユースケースを安全性と効果の2軸で分類
  - 安全性: サービス提供、人にフォールバック、対応不可
  - 効果: システム化が進んでいない、人の経験や勘に頼っている、などで判断
- 安全かつ効果の高いユースケースを特定し推進する

2. 高速にプロトタイピングする

- 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
- 要件定義よりもプロトタイプを優先する
- 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

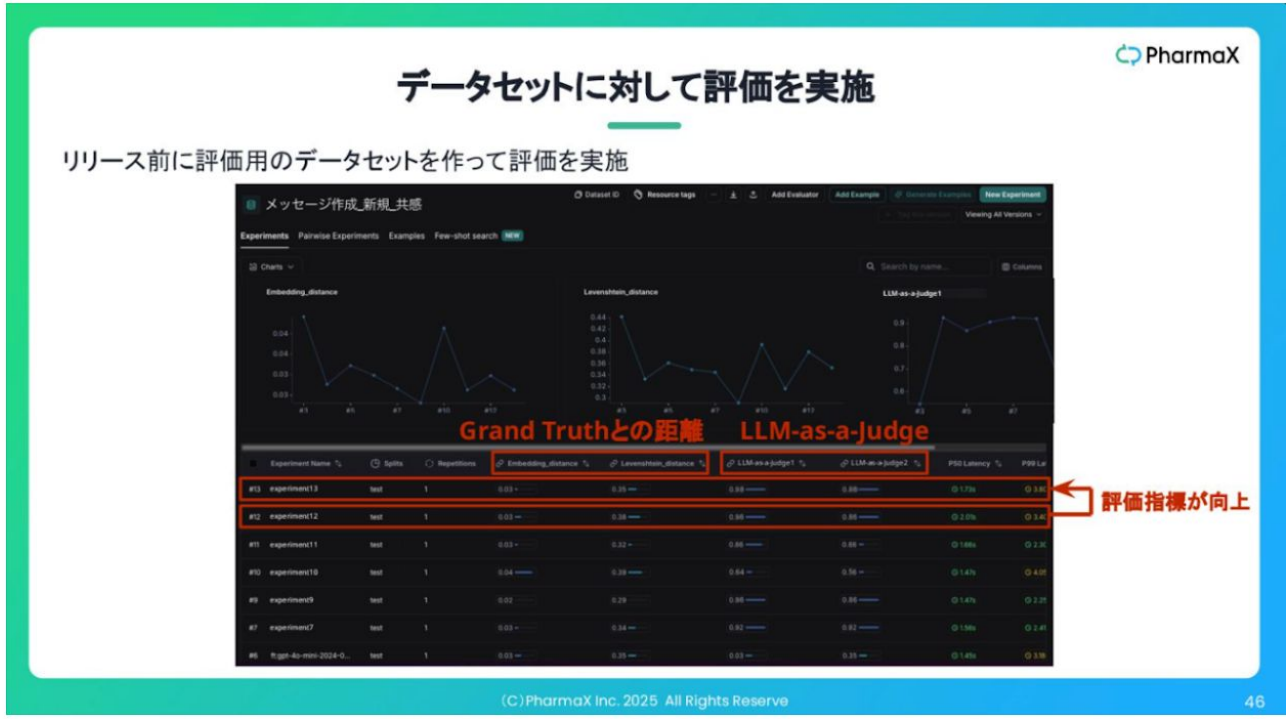




3. 独自のデータを定義し評価データを育てる

- 「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



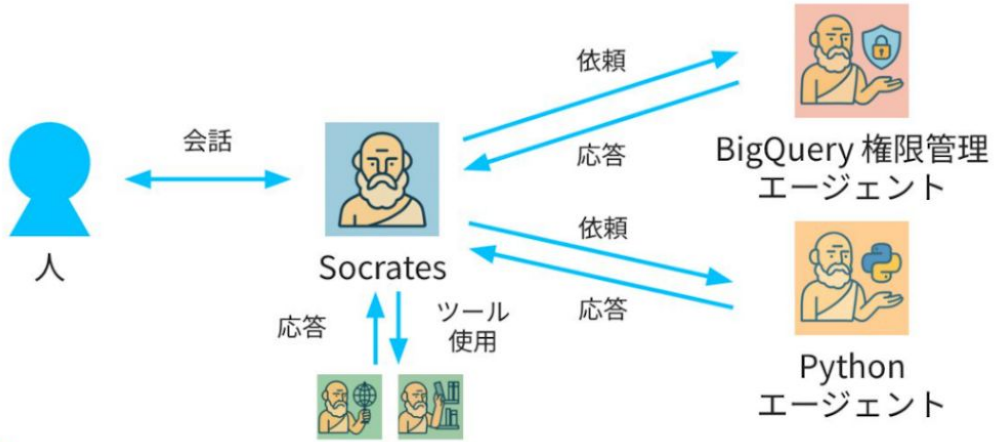
4. 専門家を開発チームの一員にする

- 自然言語は開発者と専門家の共通言語
- 専門家がプロンプトを開発するのは強力なプラクティス
- 業務ごとに小さなエージェントに分割して開発

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。



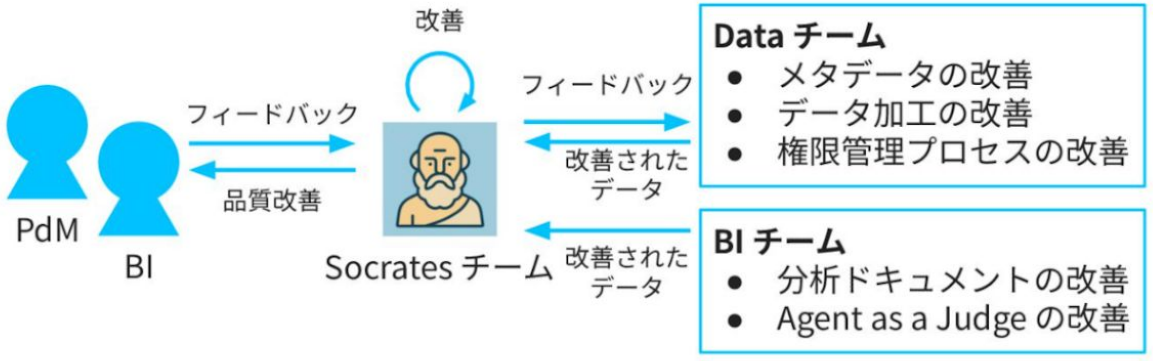
mercari

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



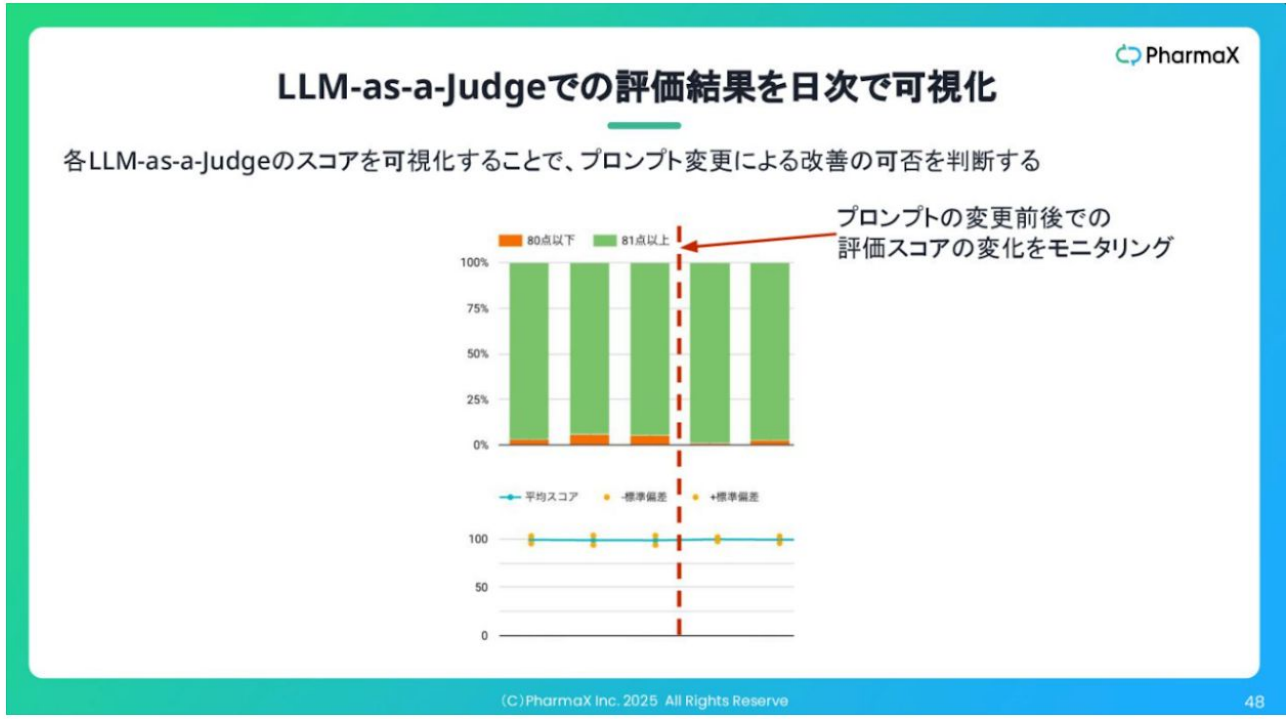
mercari

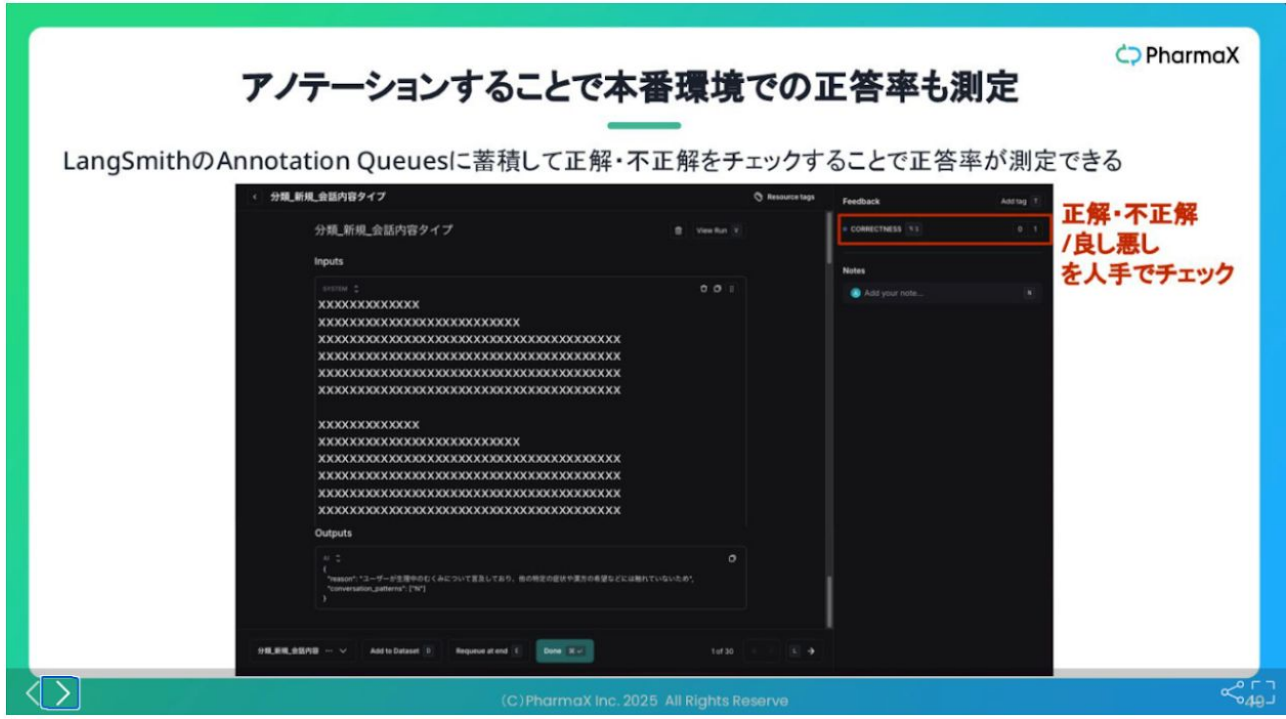
56

5. 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

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





TOC

- DevOps, MLOps, LLMOps
- 生成AIを利用するシステム開発の原則
- **今後の展望 <-**

今後の展望

- AI セーフティ
- AI ガバナンス
- セキュリティ

AI セーフティに関する取り組み

- 現状は「ユースケースの分類を行い、リスクのあるユースケースは避ける」という取り組みが主流
- リスクがあるものの効果の高いユースケースをカバーする方法に苦心している
- 独自データセットでの評価が主流なので、データセット自体の「良さ」を計測する必要がある

AIガバナンス

- AI ガバナンスの3つの要素

1. リスクマネジメント
2. 提供価値の最大化
3. 学習する組織

- リスクマネジメントは取り組まれるものの、提供価値の最大化や学習する組織については知見が少ない

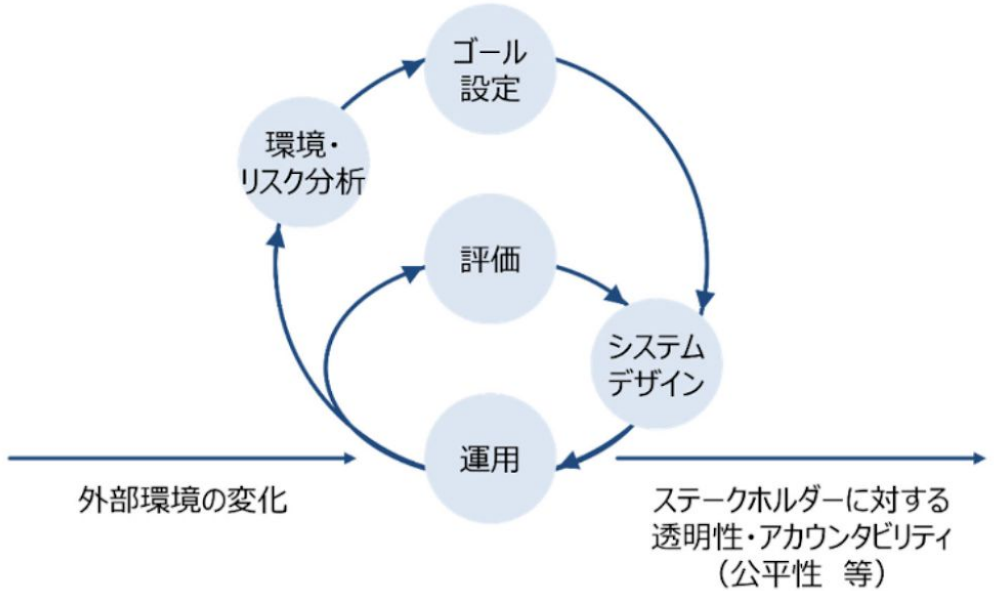


図 6. アジャイル・ガバナンスの基本的なモデル

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

セキュリティ

- Gmail などのサービスに組み込まれた生成 AI への新たな攻撃が開発されている
- MCP や A2A などのプロトコルはまだ仕様や周囲のエコシステムが未成熟
- 自律的なエージェントが主流となった際に何が起きるのか不透明

人間をプロンプトインジェクション攻撃する Gemini
https://zenn.dev/carenet/articles/e7780fc0253d89

AIセーフティ強化に関する研究開発

- 産総研、株式会社コーピー、Citadel AIで共同提案
- 「企業向け実装解説書」としてベストプラクティス集・事例集を作成
- さまざまな企業にヒアリングを実施中

NEDO「AIセーフティ強化に関する研究開発」の採択について - Citadel AI
https://citadel-ai.com/ja/news/2025/04/30/nedo/

まとめ

- DevOps, MLOps, LLMOps といった形で分野は変わったものの、イテレーティブに改善を行う手法は開発された来た
- 生成 AI を用いるサービス開発について、さまざまな取り組みが進んだ結果、いくつかの原則が見えてきた
- 一方で、AI セーフティ、AI ガバナンス、セキュリティについてはまだまだ取り組みは半ば
- いろいろな企業にヒアリングした結果をまとめていくのでご協力お願いします