Fujitsu Kozuchi Enterprise AI Factory ご紹介資料

富士通株式会社

生成AIを取り巻く
市場動向

富士通のアプローチ

uvance

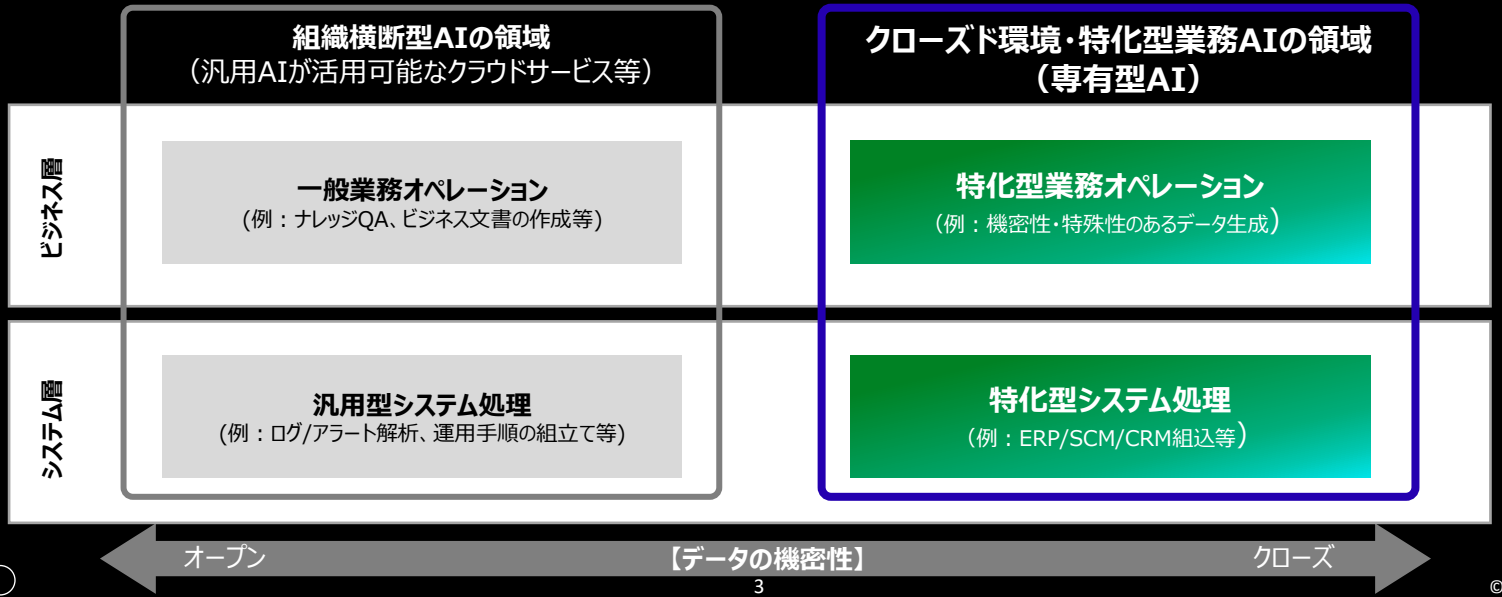
- “機密性の高いデータ”を安心・安全に扱えるプライベート環境としての生成AIプラットフォームの構成が新たな課題

機密性の高いデータが外部に持ち出されるリスク

運用を主体的にコントロールできる環境の確保

自社の業務に特化した生成AIモデル/エージェントの最適化

→ プライベート環境としての生成AIプラットフォームを構成



Fujitsu Kozuchi Enterprise AI Factory
概要

Fujitsu Kozuchi Enterprise AI Factory

uvance

- ソブリン性の高いAIを実現する手段としてオンプレミスでも活用できる専有型AIプラットフォーム

Fujitsu Kozuchi Enterprise AI Factory

3. 高精度生成AIモデルとカスタマイズ・軽量化技術

- 大規模言語モデル Takane
- 内製型ファインチューニング
- モデル量子化

データ

4. AIエージェント開発効率化技術

- AIエージェントフレームワーク
- サンプルエージェント

2. 生成AIトラスト技術

- 脆弱性スキャナー／ガードレール

1. 専有環境を実現するインフラ基盤

Private AI Platform on PRIMERGY／Private GPT

- Multi-LLM
- AI Ready DB
- 生成AIアプリケーション開発支援
- 運用管理基盤

Fujitsu Kozuchi Enterprise AI Factory

uvance

- ソブリン性の高いAIを実現する手段としてオンプレミスでも活用できる専有型AIプラットフォーム

1. 専有環境を実現するインフラ基盤

- オンプレミスに対応し、データを外部に出さないAI活用を実現
- 導入から運用高度化までを一貫して支援
- PRIMERGYベースの高性能GPU基盤を採用*

2. 生成AIトラスト技術

- 7,700種超の脆弱性に対応したLLM脆弱性スキャナー
- ガードレールによる不正入力・不適切出力の抑止
- 非専門家でも安全性を確保できる自動運用

3. 高精度生成AIモデルとカスタマイズ・軽量化技術

- 高精度日本語LLM「Takane」を中核に採用
- 内製型ファインチューニングで業務特化型モデルを作成／改善
- LLM量子化技術により低コスト・高効率なAI運用を実現

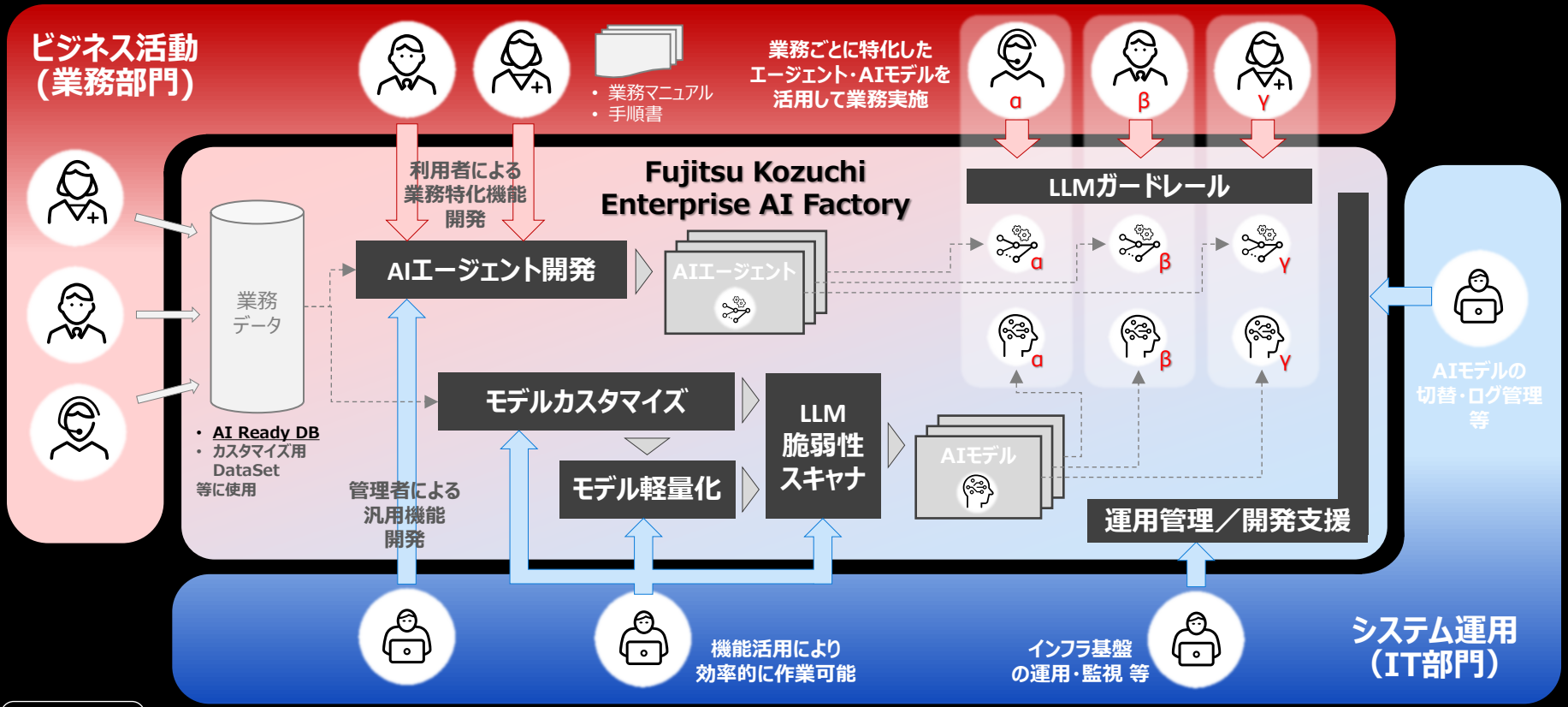
4. AIエージェント開発効率化技術

- ローコード／ノーコードで業務AIを迅速に構築
- MCP対応により既存システムと柔軟に連携
- 複数AIエージェントの協調動作による高度活用

*国内:Private AI Platform on PRIMERGY , 欧州:Private GPT

ビジネス成長を牽引する業務特化型モデルの自社生産

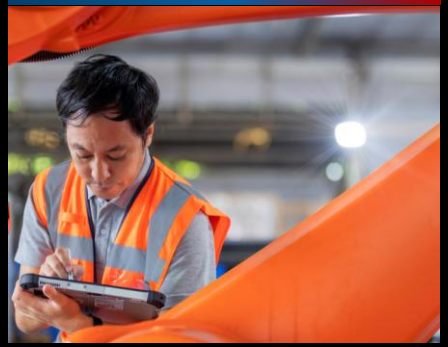
uvance



主要なTarget領域

- 製造業、医療、金融・保険、社会インフラ・公共の4領域を主要ターゲットに設定
- 業界要件に応じたAIの開発・運用・継続的高度化を実現する専有型AIプラットフォーム

製造業



- 設計・生産・保守の各工程で発生する図面、手順書、設備データを横断的に扱い、現場業務へ反復的に展開
- 設計意図理解、品質分析、ロボティクスや自動化設備に関わる知識を工程横断で蓄積・再利用

医療

- 診療記録や検査レポート、院内文書を安全に取り込み、業務プロセス全体で継続的に活用
- オンプレミスやクローズド環境を前提とした、文書生成・検索・要約の標準化や再利用による、医療業務の効率化と運用高度化

金融・保険

- 約款、規程、審査資料、問い合わせ履歴などの業務データを一元的に管理し、業務プロセスに組み込んで継続利用
- 照会対応や文書確認の標準化、業務ルールの反映・更新を通じた、安定運用とスケールを前提としたAI活用

社会インフラ・公共

- 行政文書や運用マニュアル、設備・運行情報を体系的に管理し、組織横断で反復利用
- 防災・交通・エネルギー分野に加え、安全保障や重要インフラ領域における、閉域・専有環境を前提とした持続的な運用

Fujitsu Kozuchi Enterprise AI Factory 詳細

機能スタック（構成イメージ）

uvance

- インフラ・生成AIモデル・アプリケーションの各レイヤーごとに必要な機能をワンパッケージで提供

④ AIエージェント開発効率化技術	生成AIアプリケーション開発支援	エージェントプラットフォーム	Dify 等	コード実行	Jupyterlab	音声データ変換	Whisper	
				ログ管理	Langfuse	話者識別	Pyannote.audio	
				モデル切替	Streamlit	テキスト式図表RAG	docling	
③ 高精度生成AIモデルとカスタマイズ・軽量化技術	Multi-LLM	モデル (LLM/VLM /MLLM)	Takane	gpt-oss gemma3	Embedding	Ruri	OpenAI 互換	Lite LLM
					Rerank	Ruri Reranker	LLM実行基盤	vLLM
	モデルカスタマイズ・軽量化	カスタマイズ	内製型ファインチューニング					
		軽量化	モデル量子化 (量子化PG / 稼働支援PG)					
	AI Ready DB	Weaviate						
② 生成AIトラスト技術	生成AIトラスト	LLM脆弱性スキャナー＆ガードレール						
① 専有環境を実現するインフラ基盤	運用管理	メトリクス監視	Prometheus	メトリクス収集	Node and GPU Exporter	ダッシュボード	Grafana	
	インフラ基盤	コンテナ基盤	Podman					
		OS	Red Hat Enterprise Linux (RHEL)					
		GPUサーバー	PRIMERGY					

Private AI Platform on PRIMERGY[PAPP]

uvance

• 構築済み対話型生成AI基盤

当社AIエンジニアの実践知に基づきバランス調整をした構築済みの
生成AI基盤を、すぐに使えるReady Model としてお届けいたします。

• 設置作業／保守サービス

経験豊富なエンジニアによる確実かつ円滑な設置サービス、
サポートデスクによるハードウェア/OSの保守サービスを提供します。

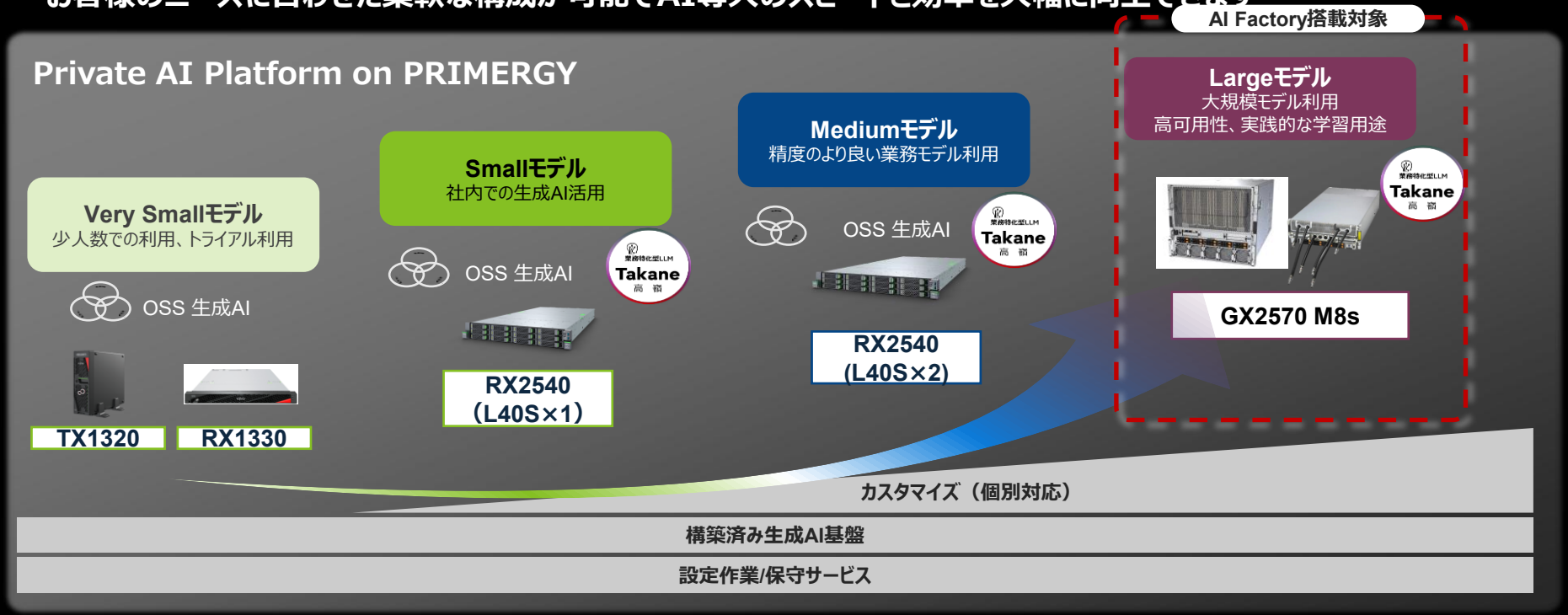
• 関連サービス

導入前のアセスメントから導入後の保守・運用支援まで様々なお客様
課題を解決にむけて、各種関連サービスをご用意しております。

Private AI Platform on PRIMERGY[PAPP]

uvance

- 規模に応じVery Smallから Largeまで多彩なモデルを提供
- お客様のニーズに合わせた柔軟な構成が可能でAI導入のスピードと効率を大幅に向上できます

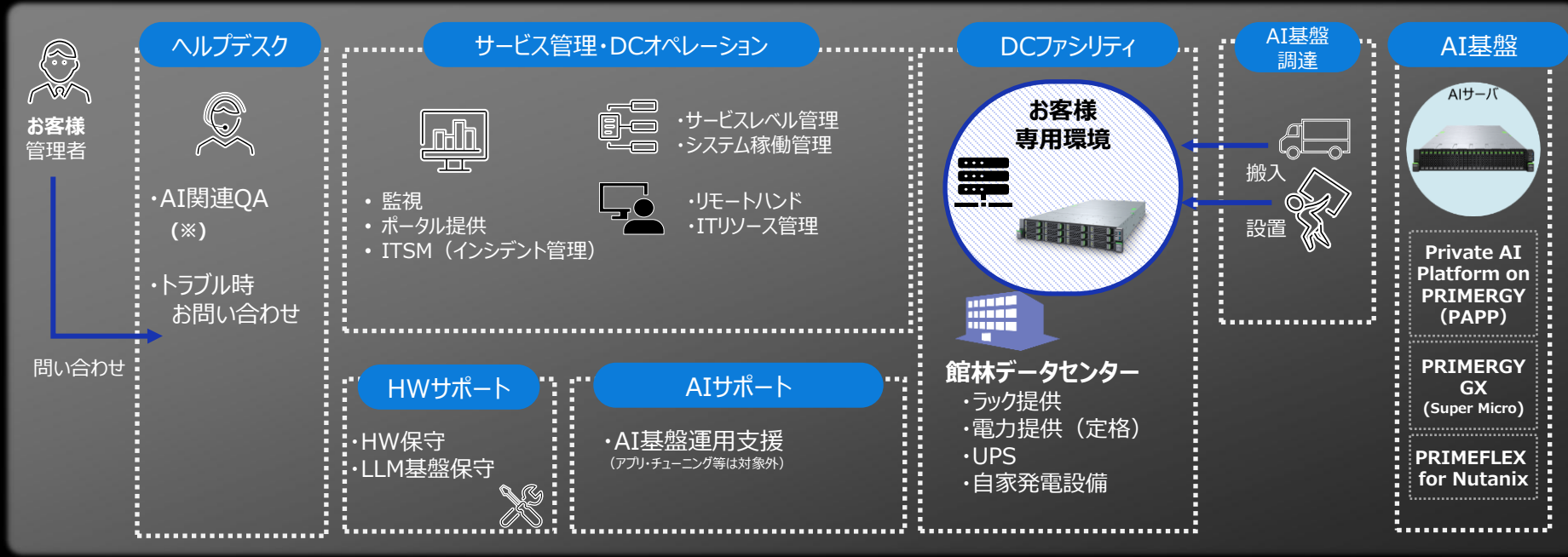


オンプレミス型の生成AIソリューション[Managed]

uvance

- 生成AI活用に必要なHWおよびLLM環境をセットアップし、富士通データセンターから提供
- Dedicated環境として構成し、他社と共用のない機密性の高い環境にてご利用が可能
- ITリソース監視を行い、不具合発生時やAI関連のQAサポートにフルサポート

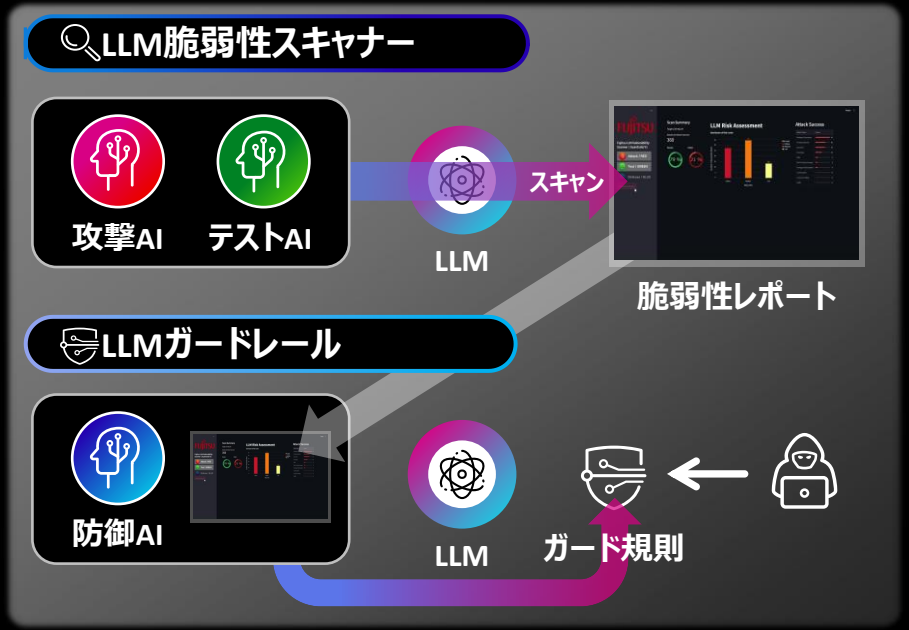
※前頁のSmall/Largeモデルに対応



LLM脆弱性スキャナー＆ガードレール

uvance

- LLMの脆弱性を網羅的な評価で判定し、最適な防御技術を自動的に適用します



業界トップクラスの脆弱性対応

- 富士通独自の手法を含む**7,700種超の脆弱性**に対応
- 従来検出困難だった**対話型の攻撃**も高精度に分析

専門知識不要

- 説明機能により、**非専門家でも**影響を容易に理解可能
- **学習コスト**を抑えたセキュリティ対策強化が可能

防御ルールを自動生成

- レポートに基づき最適な防御ルールを**自動生成**
- LLM本体に手を加えず、**外付けでの防御**を実現

エンタープライズ向け言語モデル「Takane」

cohere

FUJITSU

共同開発
業務特化型LLM
Takane
高 嶺

最新大規模言語モデル
「Command」

- 企業の業務向けに最適化されたLLM
- 企業内データを高精度に利用を可能とする世界トップクラスのRAG技術
- 業界をリードするEmbed、Rerankモデル
- 主要ビジネスタスクへの高カスタマイズ性

- 日本語特化追加学習/ファインチューニング技術
- セキュリティ強化（脆弱性スキャナ&ガードレール）
- プライベート環境での提供
- 基幹業務システム構築で積み重ねてきたナレッジ

なぜTakaneなのか

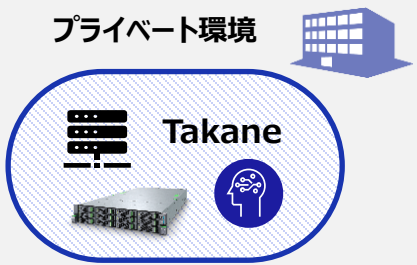
uvance

✓エンタープライズユースに特化したハイセキュリティなLLM

秘匿性の高いデータを LLMで扱いたい

オンプレミス環境にて専有利用が可能
➡秘匿性の高い業務領域のデータを安全にLLMで扱うことが可能

プライベート環境



特定の業務で的確に動作するLLMを オーダーメードしたい

富士通研究所の長年にわたる研究ナレッジをもとに、各業種のお客様向けにさまざまなユースケースを公開。専門知識や業種特有の挙動が要求される場合に**LLMのファインチューニングが可能**

導入からアフターサポートまで ワンストップで提供してほしい

富士通研究所をはじめとして、弊社内の様々なAI関連組織が有機的に連携。**生成AI活用の上流工程からお客様と伴走して導入をサポート**

お客様

FUJITSU

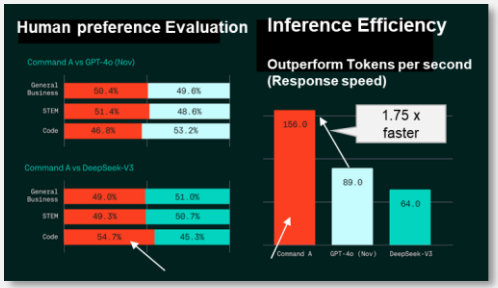
なぜTakaneなのか

uvance

✓少ないインフラリソースで高い日本語性能を発揮

日本語に強く、レスポンスの良いLLMを使いたい

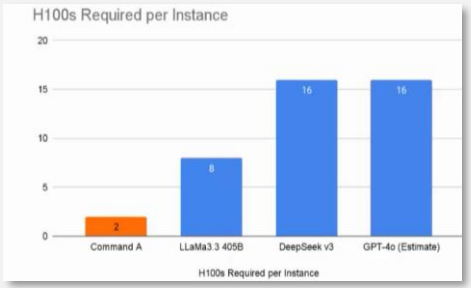
富士通研究所の知見を活かし、**日本語特化のファインチューニングを実施**。JGLUEにおいて世界最高性能を達成



Source: https://cohere.com/blog/command-a

インフラへの投資コストを最小限に抑えたい

NVIDIA H100使用時、GPT-4o比にて、たった**1/8のリソースで動作可能**。最小のインフラ投資で最大のパフォーマンスを発揮

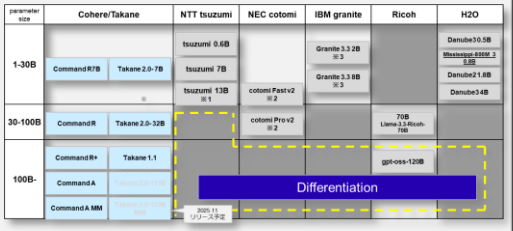


Source: https://cohere.com/blog/command-a

必要なパラメータサイズに合った製品を選定したい

1~30B、30~100B、100B~とバラエティに富んだサービス群を用意。性能の過不足やファインチューニング時のコストを抑制

※2025.9 当社調べ



※1 As of November 2024, it is being evaluated as the medium-sized version of Tsuzumi.
※2 The parameter size has not been disclosed.
※3 In addition to the language model, Granite also provides vision and code models.

Takaneのラインナップ

uvance

- **要件に合わせて適切なサイズのモデルを選択可能**

規模	LLM製品	特長	ご利用例
Small	**Takane 7B** Command R7B	• 小規模モデル • 迅速な応答 • 小中規模のタスク向け	• 企業のチャットボット • コンテンツ生成 • 自動翻訳
Middle	**Takane 32B** Command R	• 中規模モデル • 文脈理解 • 複雑なタスク処理	• 研究機関 • ビジネス分析 • 高度な顧客サポート
Large	**Takane** Command A	• 大規模モデル • 高度な言語処理 • 専門分野対応	• 大規模メディア • 専門分野 • AIアシスタント開発
Large Multi Modal	**Takane 112B Vision** Command A Vision	• マルチモーダルモデル • 図表・画像の読込	• 図表やグラフを含む文章解析 • 設計・製造領域での図面読み取り
拡張用モデル	**Embed 4**	• テキストをベクトル空間に写像、類似性や意味の分析を可能にします	
	Rerank 3.5	• 検索結果の再ランク付け、関連性の高い情報を素早く提供します	

多様な搭載可能モデル

uvance

OpenAI
gpt-oss 20b & 120b

**OpenAI社の高性能な
オープンソースLLM**

**大規模パラメータによる長文入力・複雑な
文脈理解と、商用利用可能な自由度を
兼ね備えた大規模言語モデル**

- gpt-oss-20bは21Bのパラメータを持ち長文処理や複雑な文脈理解に強み
- gpt-oss-120bはの120Bのパラメータを持ちGPT-4に匹敵する精度
  効率と精度を両立

Google
gemma3 12b & 27b

Googleが開発したマルチモーダルLLM

**テキストに加えて画像を処理できる
軽量で高性能なオープンソースAIモデル**

- テキストに加えて画像を入力とした
  テキスト生成が可能
- 日本語だけでなく、140を超える言語を
  サポート
- 量子化技術（GPTQ形式）を採用
  GPU1枚でも動作可能

内製型ファインチューニング＆モデル量子化

uvance

内製型ファインチューニング

- 回答精度が低い
- 業務に特化したカスタマイズにリソースが必要
- 専門知識を持つ人材がいない

▶ **効率的なチューニング手法と実行環境を提供**

モデル量子化

生成AIは膨大な電力を消費するため用途に合わせて軽量化する必要がある

▶ **1ビット量子化で世界最高精度を達成**

量子化：各ニューロン間の結合に割り当てられる重みを圧縮する技術
（重みをより少ないビット数で表現することにより、モデルサイズを削減）

お客様自身でファインチューニング実践可

学習用のDataSetがあれば ファインチューニング可能

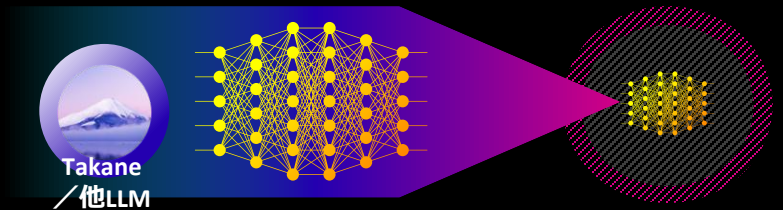
ファインチューニング機能／実行環境

領域／業務特化型モデルの作成・改善サイクルを実現可能

精度を維持しつつ、軽量・低消費電力化

量子化前
（32ビットまたは16ビット）

量子化後
（1ビット）



富士通独自の量子化誤差伝播法を用いて高精度を維持

マルチAIエージェントフレームワーク

uvance

- AIエージェントを領域特化、高品質でセキュアなワークフロー自動化

Multi AI Agent Framework

Building

AI Agent

Connecting

Workflow

Orchestration

Control

領域特化　品質　セキュリティ

領域特化

Agentic Memory

連携

人やロボット、他のエージェントから知識を習得
特定領域で成長しながらワークフローを遂行する

ワークフロー

監視

制約付きルーティング

成長(品質)

ワークフロー全体の複雑なやりとりを適切に監視
多様な制約を考慮しながらワークフロー遂行を支援

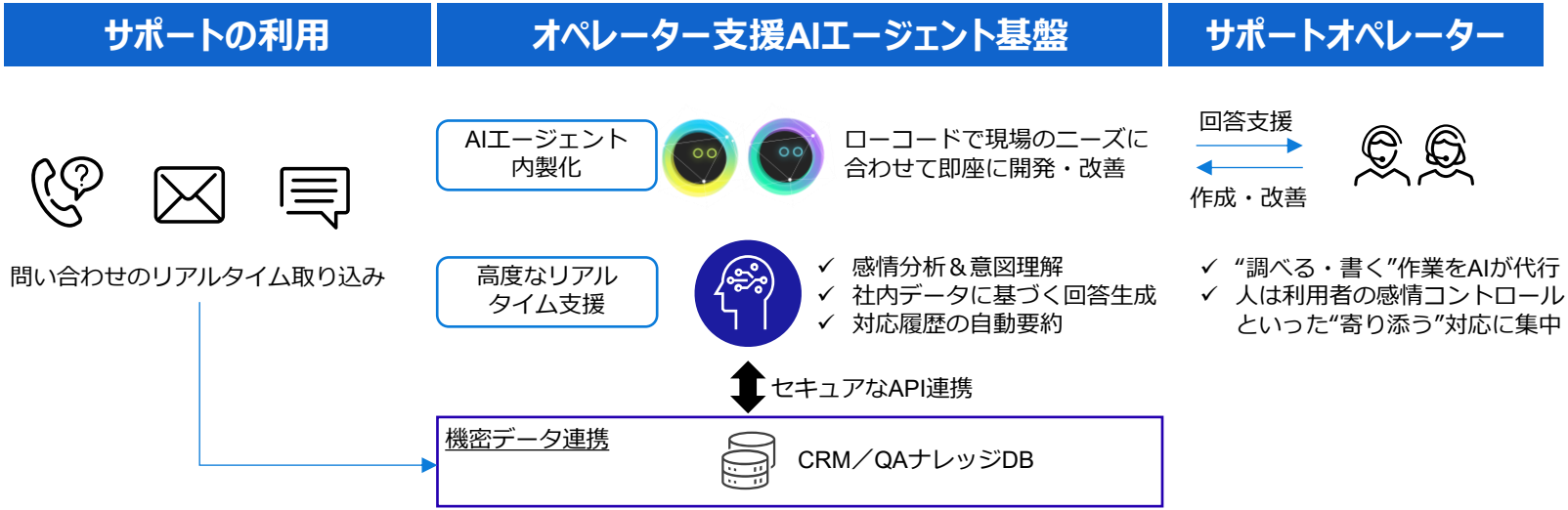
ユースケース

カスタマーサービスの高度化

uvance

背景・目的	• 機密データを守りながら、ローコードで現場主導のAIエージェント開発を実現する次世代コンタクトセンター基盤の確立

活用イメージ



期待効果	• QCD改善：オペレーターの対応品質向上や検索・記録業務の自動化により対応時間の削減 • 業務改革：サポートオペレーター業務を回答の検討作成からエージェントの作成・改善と質問者への“寄り添い”へ

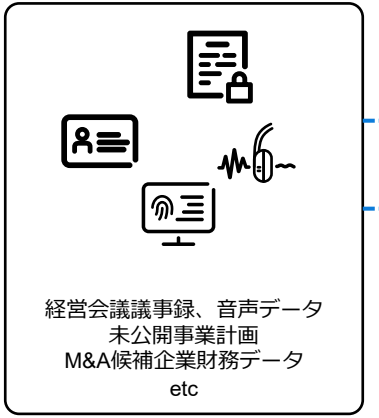
経営の意思決定支援

uvance

背景・目的	• 社内の機密文書やデータソースから情報を抽出し、経営陣の意思決定をサポート

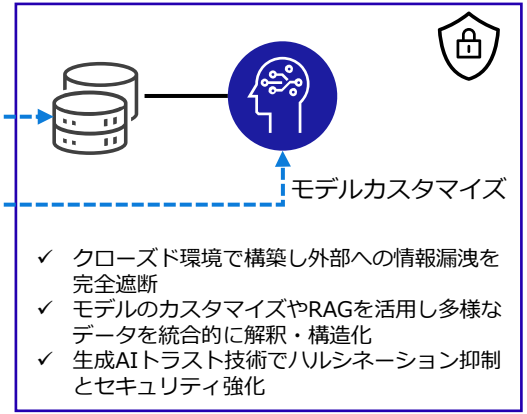
活用イメージ

多種多様な機密情報 → 学習・解析するセキュアなAI基盤 → 経営判断の高度化



経営会議議事録、音声データ
未公開事業計画
M&A候補企業財務データ
etc

RAG



- ✓ クローズド環境で構築し外部への情報漏洩を完全遮断
- ✓ モデルのカスタマイズやRAGを活用し多様なデータを統合的に解釈・構造化
- ✓ 生成AIトラスト技術でハルシネーション抑制とセキュリティ強化

市場ポジション予測、M&A評価、新規事業アイデア

期待効果	• プライベート環境でしか扱えないデータをAIで活用、分析、推論することで経営判断を高度に支援

製造・インフラ現場主導で進化する「エッジAI」

uvance

背景・目的	• 量子化技術による軽量モデル展開と内製型ファインチューニングで、現場に特化したAIを育成

活用イメージ

AIファクトリー（特化、小型化する場所）

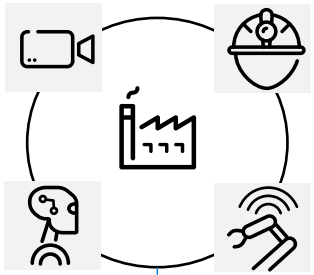
小型デバイス向けにモデルの量子化

現場固有の知識、OT(Operational Technology)領域の学習

AIの配備

再学習による精度向上

エッジ環境（工場・インフラ現場）



- ✓ 予知保全：振動・音から故障の予兆を検知
- ✓ セキュリティ：不審者や危険エリアへの侵入検知
- ✓ 品質管理：外観検査による不良品の排除

フィードバック：誤検知データや新たな異常パターンを安全に連携・蓄積

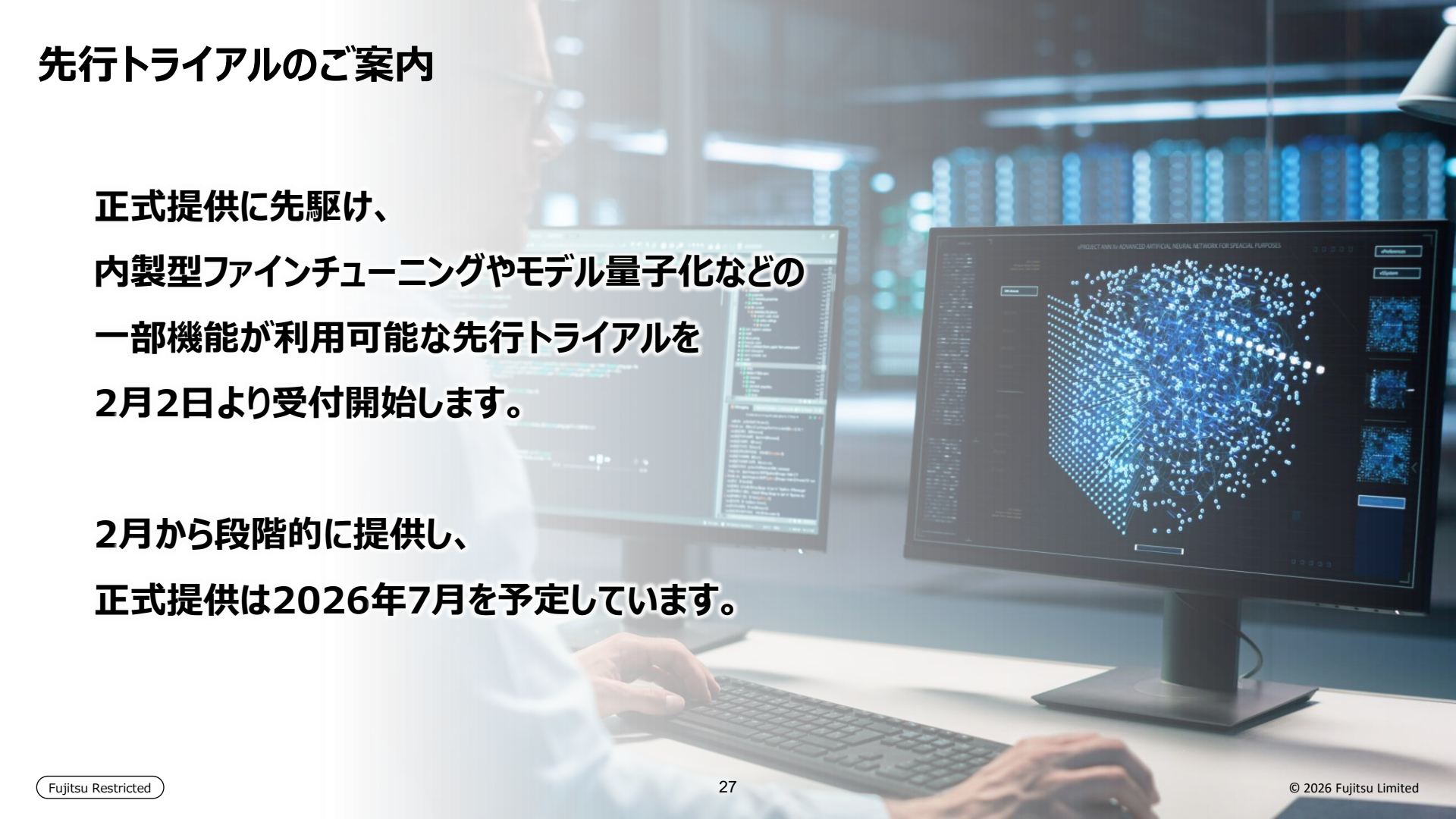
“止まらない現場”を実現する自律型AIエコシステム

期待効果	• 予知保全による設備稼働率向上、外観検査の自動化と品質均一化、危険エリアの常時監視

uvance

先行トライアル

先行トライアルのご案内



正式提供に先駆け、
内製型ファインチューニングやモデル量子化などの
一部機能が利用可能な先行トライアルを
2月2日より受付開始します。

2月から段階的に提供し、
正式提供は2026年7月を予定しています。

トライアル仕様

uvance

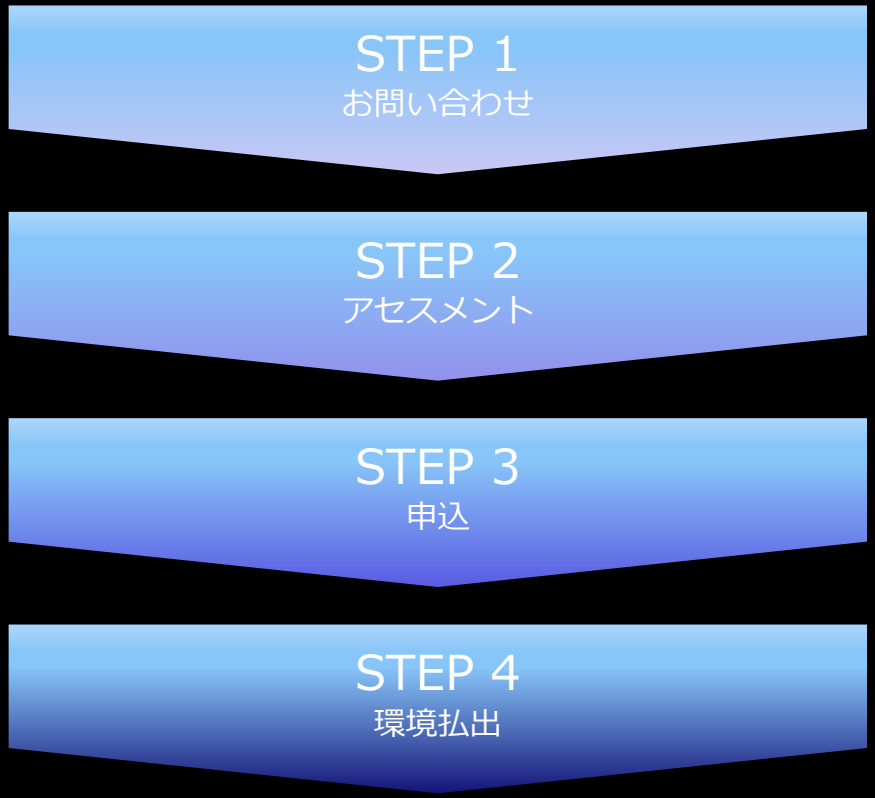
名称	Fujitsu Kozuchi Enterprise AI Factory 先行トライアル
トライアル概要	2026年7月正式リリース予定のFujitsu Kozuchi Enterprise AI Factory の限定機能版をお試し利用いただける環境を無償でご提供します。
対象者	お客様、富士通関係会社
トライアル環境、機能	・Private AI Platform on PRIMERGY ・Fujitsu Generative AI for Cohere (Takane) ・内製型ファインチューニング機能 ・脆弱性スキャナ／ガードレール機能 ・LLM量子化機能(*1) ・AIエージェントフレームワーク(*2)
料金	環境使用料無償（役務支援などは原則有償）
サポート	EmailによるQA
利用期間	利用期間については事前ヒアリングの中で調整させていただきます。(目安：2週間～4週間)
受付開始	2026年2月2日
備考	事前ヒアリング、アセスメントを実施した上で貸出環境、機能の選定と可否判断を実施させていただきます。 貸出環境は数に限りがあるため、スケジュールはご希望に添えない場合がございます。

(*1) サービス提供元でTakaneの量子化を実施し提供いたします。利用者自身でLLMを量子化できる機能については段階的に機能追加予定です。

(*2)Private AI Platform on PRIMERGYに標準搭載のDifyがご利用いただけます。段階的にAI Agent機能を追加予定です。

トライアル開始までの流れ

uvance



- ✓ お問い合わせは担当Sales経由でトライアル窓口へコンタクトお願いします
- ✓ 評価予定内容、利用機能、利用スケジュール、今後の本格利用の予定等をヒアリングさせていただきます
- ✓ ヒアリング結果を基にトライアル実施可否判断およびスケジュールを調整いたします
- ✓ アセスメント結果に基づき正式にお申込みいただきます
- ✓ 環境をセットアップし提供いたします