



コマース領域での生成AIを活用した 業務効率化

Hidenori Koizumi

Sr. Prototyping Solutions Architect

Public Sector Japan

小泉 秀徳 (Hidenori Koizumi)

Sr. Prototyping Solutions Architect

公共部門向けにプロトタイプ開発と提供を行っています。大学までのバックグラウンドを活かし、Life ScienceでのAWS利用の支援も行っています。

関心のあるフレームワーク

- LangChain
- sst/ion



本セッションのゴール

- デモを通じた生成AIでの業務効率化の可能性のイメージ
- 生成AIの業務適用アプローチの理解
- 生成AI活用における課題解決のヒント

本セッションで取り上げない内容

- AWSサービスの機能詳細
- Fine Tuningや独自モデルの開発方法

Agenda

- 生成AIの業務適用への課題
- Amazon Bedrockを利用した業務効率化
- 業務で生成AIを活用するポイント

生成AIの業務適用への課題



PartyRock Demo

The screenshot displays the PartyRock web application interface. On the left is a dark sidebar with navigation options: PartyRock, Generate app, Apps, Snapshots, Discover, Playlists, PartyRock Guide, What's new, Give feedback, and Backstage. The main content area is titled 'Apps' and includes a 'Sort by last updated' dropdown and a 'Generate app' button. A card at the top shows a preview of an app with the title 'ガイドラインの遵守事項の要約' and a 'Private' status. Below this is a section titled 'Try out featured apps built with PartyRock' with the subtitle 'See what everyone's been up to.' This section contains four featured app cards: 'Vin-du-jour wine pairing', '6 Degrees of Vinyl', 'Good Eats', and 'Movie Recommendation'. Each card includes a brief description of its functionality. At the bottom of the main area is a large 'Generate your next app' button.

PartyRock

生成AIのアプリケーションを簡単に実装できるサービス

The screenshot displays the PartyRock application interface. At the top, there is a header with the text "PDF Content" and navigation icons. Below the header, the main content area is divided into two sections: "PDF Content" and "Summary".

PDF Content

第2章 品名広告

- 「効能又は効果、用法及び用量、警告・禁忌を含む注意事項等情報等については電子添文を参照ください」等と目立つよう見やすい文字【8ポイント以上】で記載すること
- 品名のみを主体とする広告では、キャッチフレーズ、効能又は効果、用法及び用量等の有効性・安全性等に関する情報は記載しないこと。

第3章 記事体広告

- 記事掲載頁には提供企業名を明確に記載すること。
- 掲載するデータは科学的な裏付けがあり信頼性の確保された正確なものであること。
- 有効性に関する臨床成績の記載に際しては、承認された効能又は効果、用法

Summary

以下は、製薬企業が作成する製品情報概要や専門誌広告における遵守事項の要約です。

主な遵守事項:

- 記載内容は科学的根拠に基づき、正確、公平、客観的であること。
- 有効性と安全性の情報をバランスよく記載すること。
- 承認外の効能・効果、用法・用量は記載しないこと。
- 参考情報は明確に区別して記載し、誤解を招かないこと。
- 動物実験結果から臨床への外挿は避けること。
- 虚偽、誇大な表現は使用しないこと。
- 他社や製品の中傷は避けること。
- 例外的データを一般化しないこと。
- 最新の情報と法規制を遵守すること。
- データは科学的裏付けがあり、正確であること。
- 専門誌広告では承認範囲外の情報は記載しないこと。

要約すると、製品情報は科学的根拠に基づき正確で公平な記載が求められ、有効性と安全性の両面を偏りなく伝え、承認範囲外の情報は避けることが重要視されています。

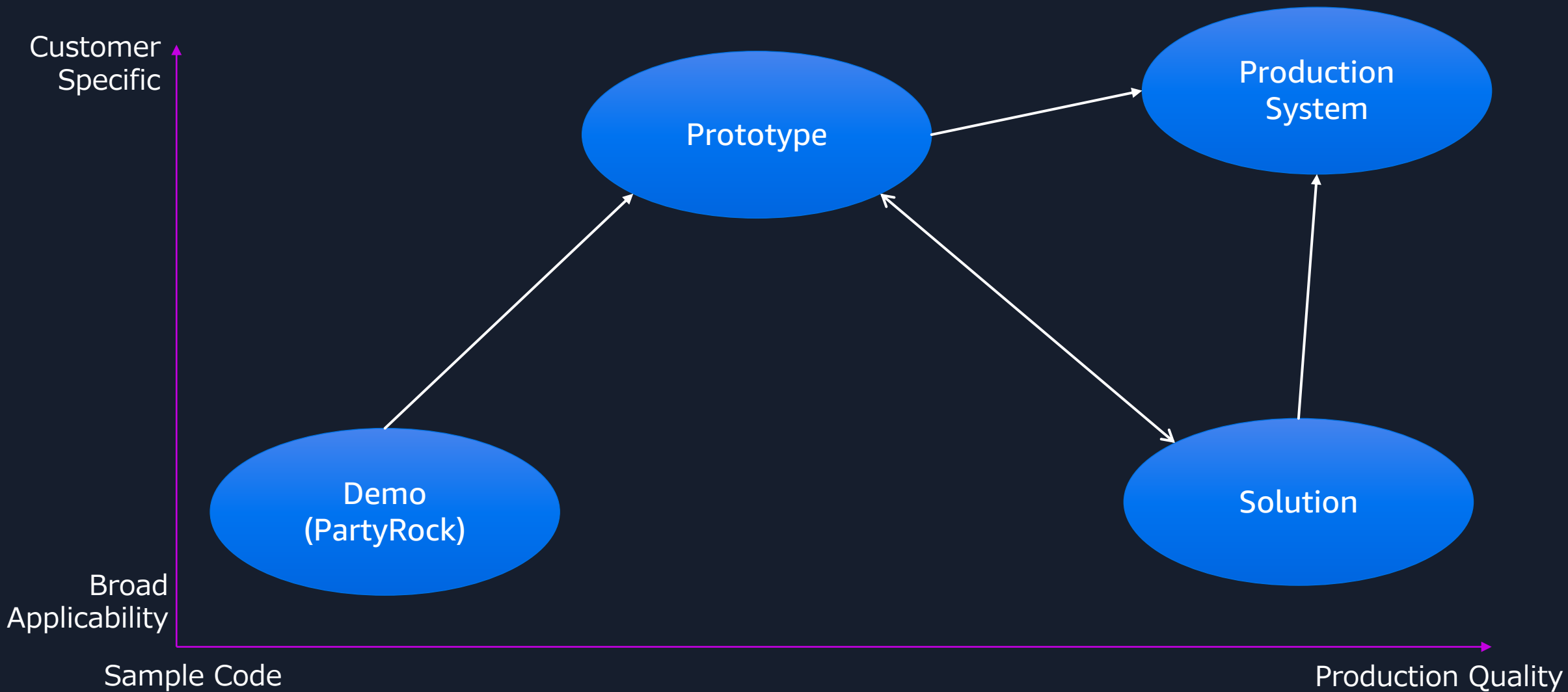
At the bottom of the interface, there is a button with a play icon and the text "Cmd + Enter to replay".

- テキストによる指示と画面操作だけで生成AI (Amazon Bedrock)を組み込んだアプリケーション作成が可能
- 作成したアプリケーションの共有機能あり
- Amazon Bedrockの複数の基盤モデルの選択が可能 (Claude 3, Titan, Llama 2, etc.)

業務適用への課題

- PartyRockなどのデモやサンプルだと要件に合わない
 - 汎用的な機能群で実ワークロードへの課題解決に至らない
- 期待している回答が得られない
 - 外部知識ソース(RAG)の作成と精度
 - Prompt
- 他のAWSサービスとの連携
 - RAG用データの前処理
 - データ基盤
 - AI/MLサービス

本番環境へのロードマップ



Amazon Bedrockを利用した 業務効率化

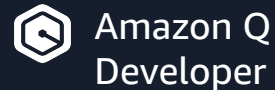


AWS Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs



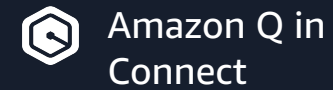
Amazon Q
Business



Amazon Q
Developer



Amazon Q in
QuickSight



Amazon Q in
Connect

TOOLS TO BUILD WITH LLMs AND OTHER FMs



Amazon Bedrock

Guardrails | Agents | Studio | Customization Capabilities | Custom Model Import

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE



GPUs



Trainium



Inferentia



SageMaker



UltraClusters



EFA



EC2 Capacity Blocks

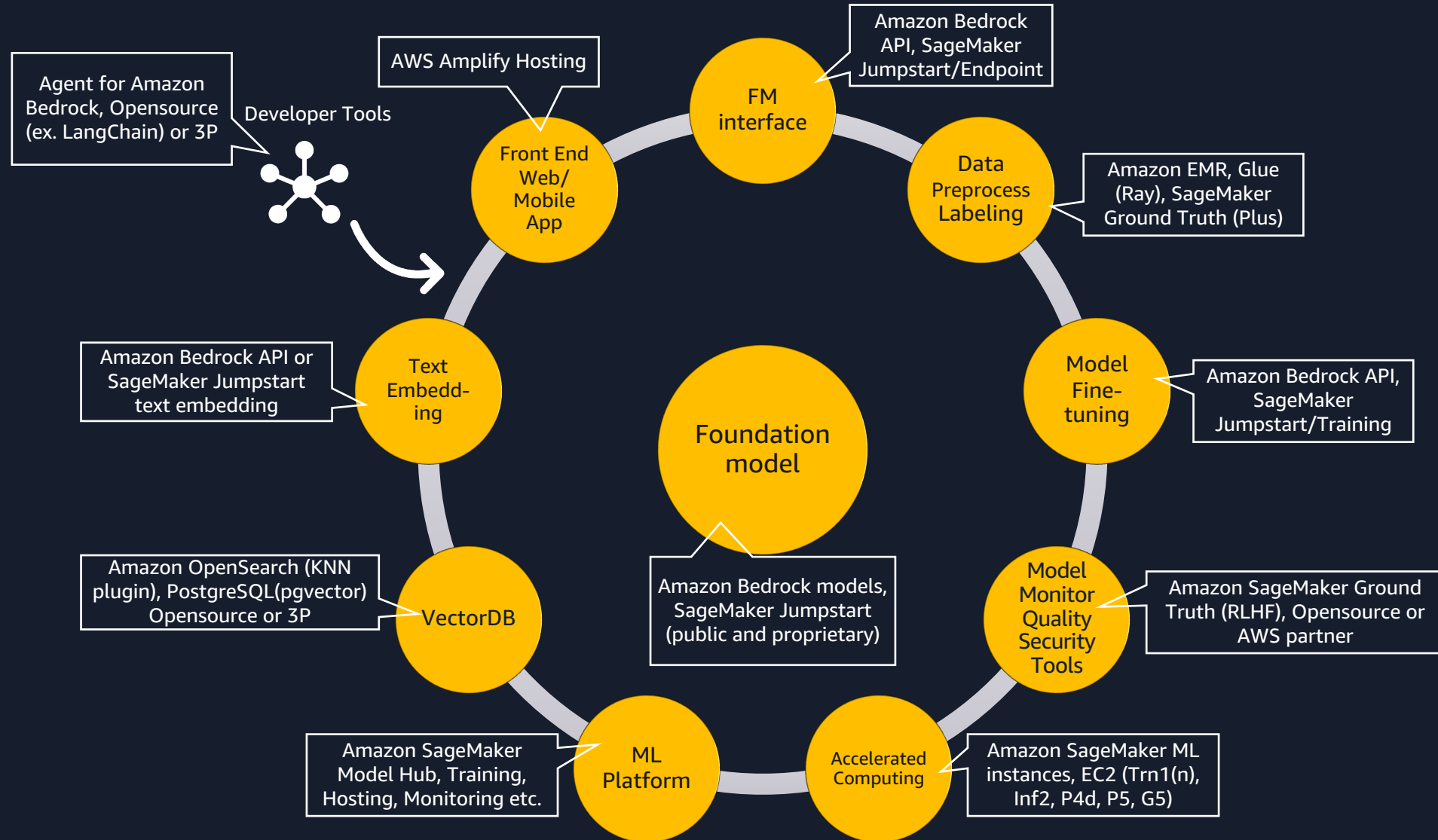


Nitro



Neuron

Components to AWS services mapping



Use case

1. 社内文書を利用したコンテンツレビュー
2. 承認済み図表チェック(画像)

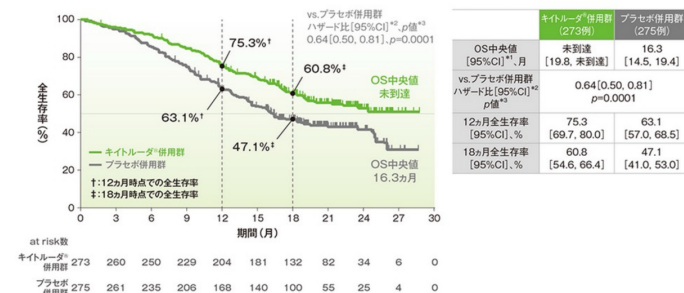
Use case 1

- 社内文書を利用したコンテンツレビュー
 - コンテンツはPDFやJPEGを含む
 - RAG検索用の社内文書登録（医療用医薬品製品情報概要等に関する作成要領）
 - 社内文書(ガイドライン)に基づくバリデーションチェック

主要評価項目：全生存期間（OS）（優越性試験）

■PD-L1発現陽性（CPS \geq 1）患者における全生存期間：OS

OSのKaplan-Meier曲線（ITT集団）



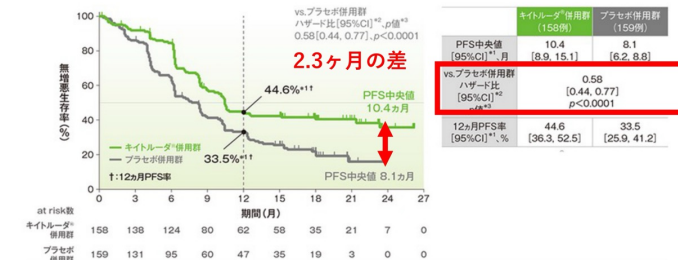
*1 打ち切りデータはproduct-limit (Kaplan-Meier) 法に基づく
*2 投与群を共変量とし、無作為化に用いた層別因子 [初回診断時の遠隔転移 (あり、なし)、ペパシズマブの併用投与 (あり、なし)、PD-L1発現状態 (CPS<1、1 \leq CPS<10、CPS \geq 10)] を層別因子とした層別Cox比例ハザードモデルに基づく
*3 無作為化に用いた層別因子を層別因子とした層別ログランク検定 [片側]、有意水準 $\alpha=0.0054906$ (検証的解析結果) (追跡期間中央値: キートルーダ®併用群 18.3ヵ月、プラセボ併用群 16.3ヵ月、データカットオフ日: 2021年5月3日)

承認時評価資料: 国際共同第III相試験 (KEYNOTE-826試験) Colombo N et al. *N Engl J Med* 2021; 385: 1856-1867
本試験はMSD社の資金提供により行われた。著者のうち、Sarper Toket, Kan Li, Stephen M Keefeは同社の社員である。

主要評価項目：無増悪生存期間（PFS）（優越性試験）、副次評価項目：12ヵ月PFS率

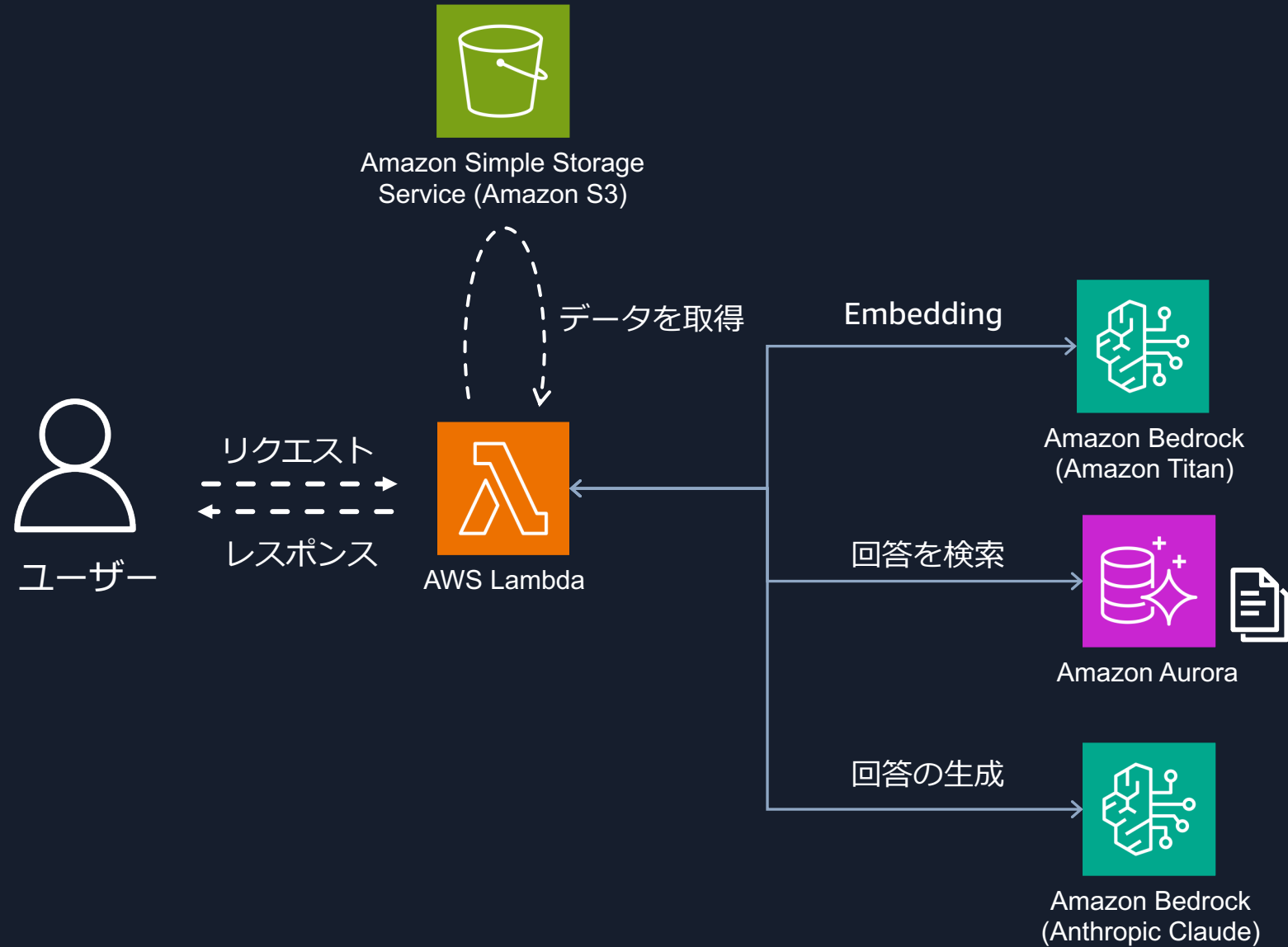
■PD-L1発現陽性（CPS \geq 10）患者における無増悪生存期間：PFS
■PD-L1発現陽性（CPS \geq 10）患者における12ヵ月PFS率

PFSのKaplan-Meier曲線（ITT集団）

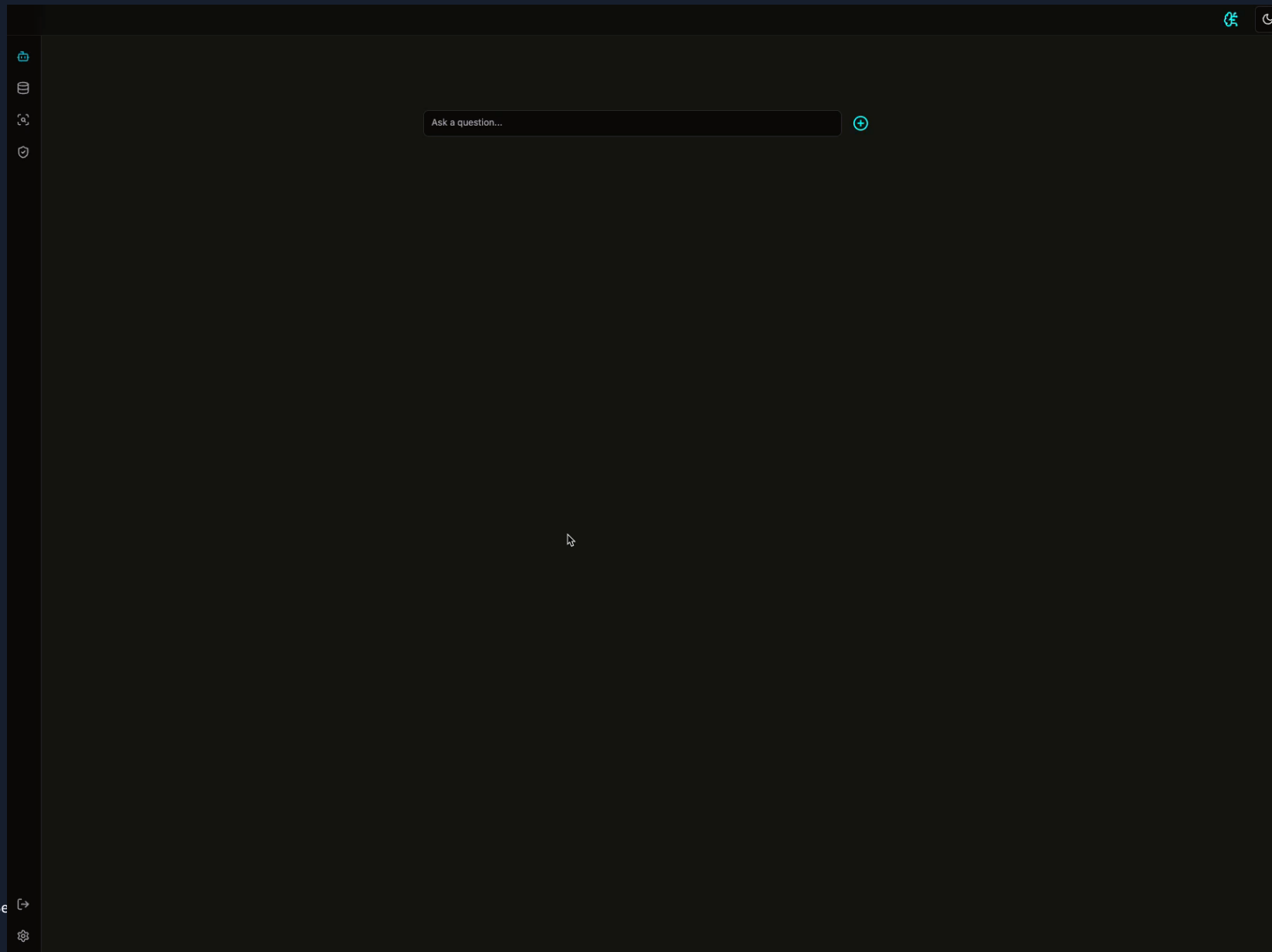


試験担当医師によるRECISTガイドライン1.1版に基づいた評価
*1 打ち切りデータはproduct-limit (Kaplan-Meier) 法に基づく
*2 投与群を共変量とし、無作為化に用いた層別因子 [初回診断時の遠隔転移 (あり、なし)、ペパシズマブの併用投与 (あり、なし)、PD-L1発現状態 (CPS<1、1 \leq CPS<10、CPS \geq 10)] を層別因子とした層別Cox比例ハザードモデルに基づく
*3 無作為化に用いた層別因子を層別因子とした層別ログランク検定 [片側]、有意水準 $\alpha=0.0019593$ (検証的解析結果) (追跡期間中央値: キートルーダ®併用群 18.3ヵ月、プラセボ併用群 16.3ヵ月、データカットオフ日: 2021年5月3日)

Use case 1: Architecture





Use case 1: コンテンツレビュー




Use case 1: コンテンツレビュー

画像は医療用医薬品の臨床成績を示すスライドです。ガイドラインに違反していないか、各画像をレビューしてください。違反があればガイドラインを引用して指摘してください。

 Analyze images and search similarity
Analyzing images with LLM and searching in the document

 Post-processing
Filtering or ranking retrieved documents

 Review the documentation
Reviewing the documentation and generating the result

✔ Check: 画像の2.3ヶ月の差の部分で、対照薬との比較を矢印で強調しています。これは以下のガイドラインに違反する可能性があります。

「I. 製品情報概要 - 20 -」

グラフにおいては縦軸、横軸等の尺度を必要以上に変える等で差を強調した作図をしないこと。「差を強調した作図をしない」対照薬（プラセボを含む）との差や当該薬の投与前後の値の差について、文字（数字等）を大きくして強調したり、色をつけて目立たせること等はできません。グラフ内で差を矢印によって示すことも同様です。

この画像では、対照薬との差を赤い矢印で強調しているため、ガイドラインに違反する可能性があります。差を強調しすぎないよう注意が必要です。

また、以下のガイドラインにも違反する可能性があります。

「II. 専門誌（紙）掲載広告 - (7)」

非臨床比較試験で他社品に関する記載は試験結果の事実のみにとどめ、他社品の試験結果の解説は記載しないこと。また、比較を強調するようなタイトルを避ける等、他社及び他社品の中傷・誹謗とならないようにすること。

対照薬との比較を強調しすぎると、他社品の中傷・誹謗になる可能性があるためです。

関連文書-1

関連文書-2

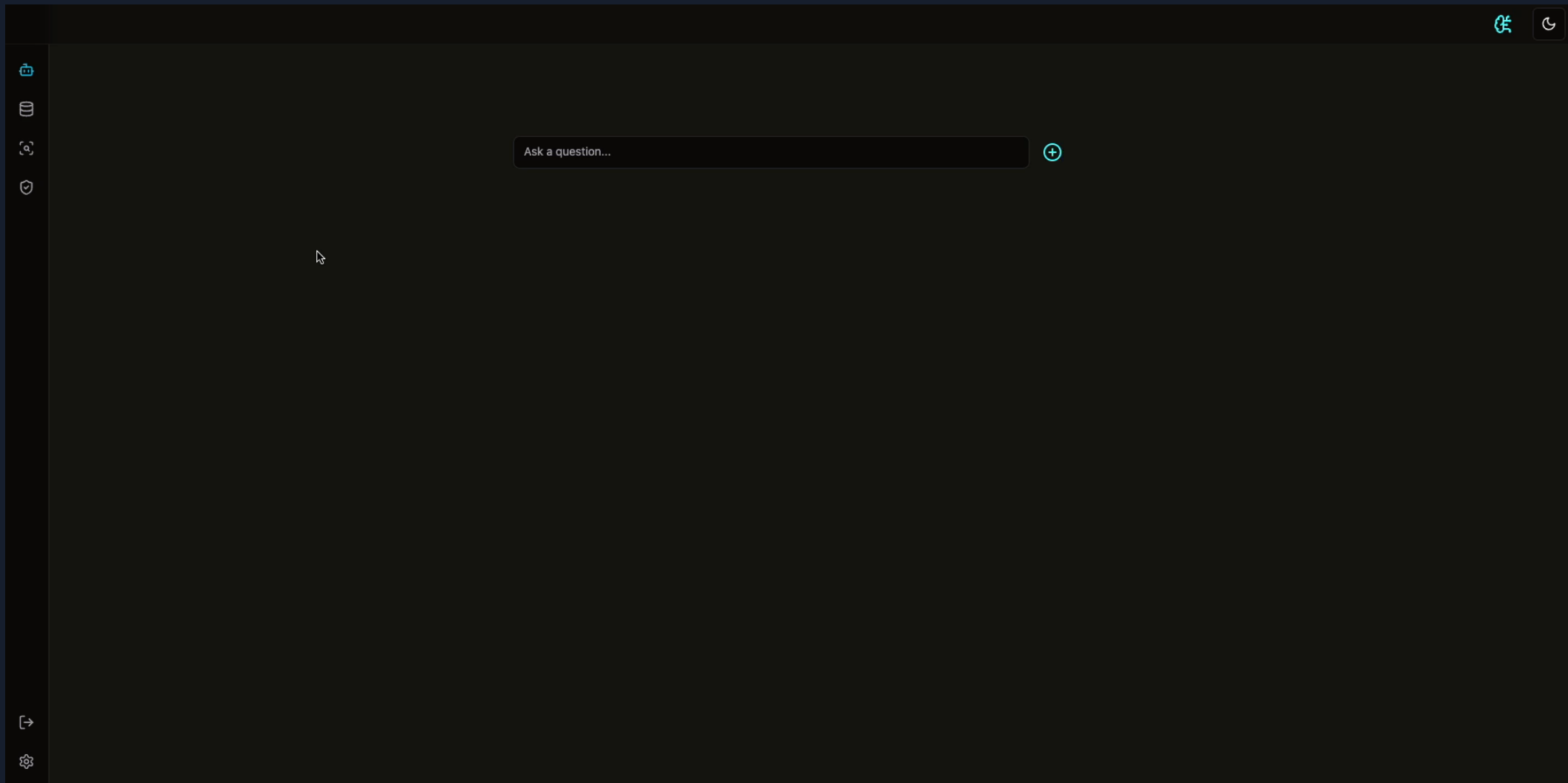
質問: 画像は医療用医薬品の臨床成績を示すスライドです。ガイドラインに違反していないか、各画像をレビューしてください。違反があればガイドラインを引用して指摘してください。



画像の2.3ヶ月の差の部分で、**対照薬との比較を矢印で強調**しています。これは以下のガイドラインに違反する可能性があります。

グラフにおいては縦軸、横軸等の尺度を必要以上に変える等で差を強調した作図をしないこと。「差を強調した作図をしない」対照薬（プラセボを含む）との差や当該薬の投与前後の値の差について、**文字（数字等）を大きくして強調したり、色をつけて目立たせること等はできません**。グラフ内で差を矢印によって示すことも同様です。

Use case 1: コンテンツレビュー(社内文書登録)



Use case 1: コンテンツレビュー(社内文書登録)

Contents Registration

Documents upload

Upload your documents to Amazon S3

Drag and drop files here, or click to select files

Selected Files:

- orosisinki-kisaireiR3.pdf

Upload

Embedding Job

Store the content as embedding in VectorDB with Amazon Bedrock

Table name

Input table name here

(optional) Description

Input a description of the table

(optional) Extract keywords

しないこと,してはいけない

Files

orosisinki-kisaireiR3.pdf

Use case 2

- 承認済み図表チェック(画像)
 - 画像分析
 - 分析した画像と類似する画像が承認済みかチェック
 - S3のapprovedフォルダの画像と照合

approved/ [Copy S3 URI](#)

[Objects](#) [Properties](#)

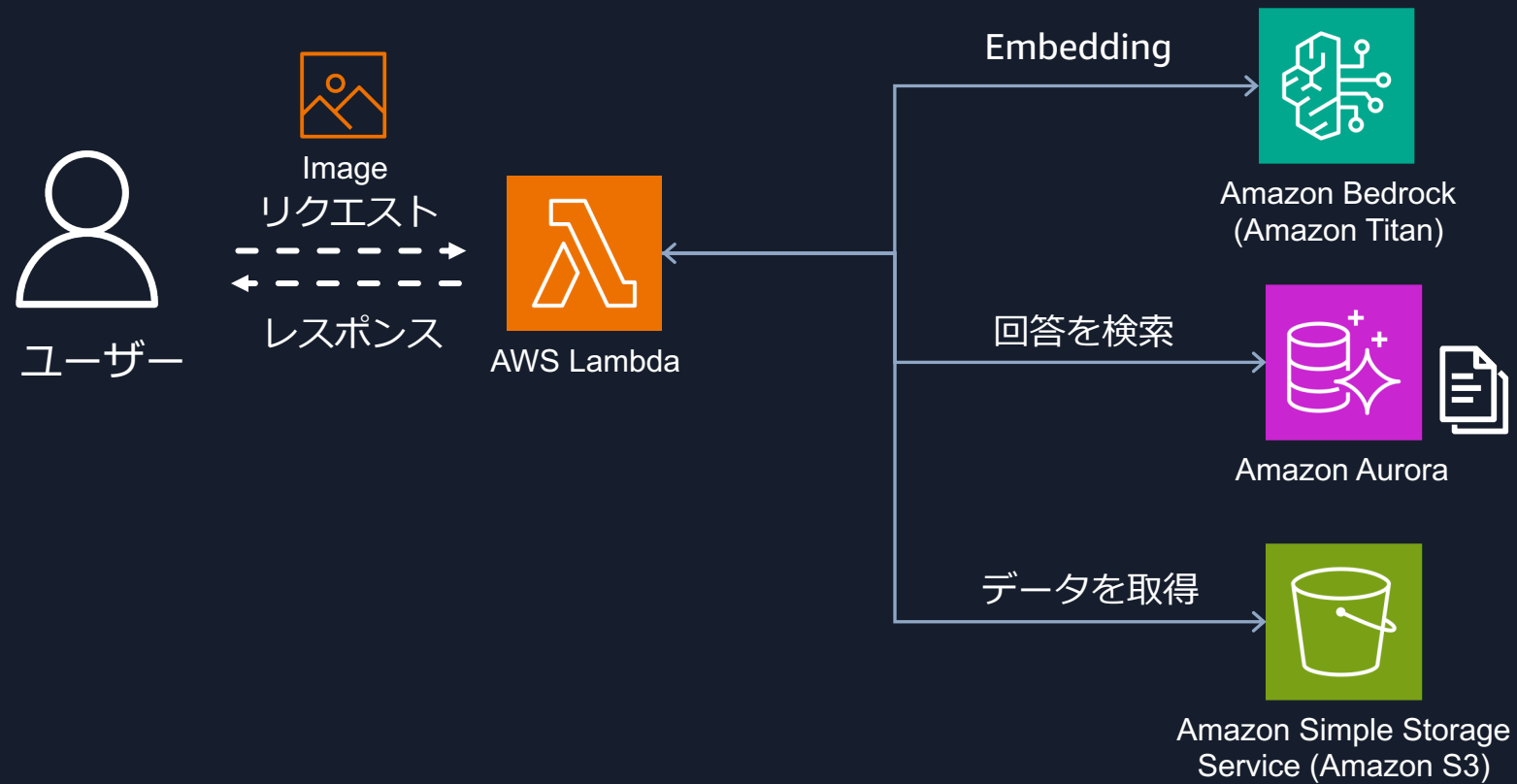
Objects (7) [Info](#) [Refresh](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#) [Create folder](#) [Upload](#)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

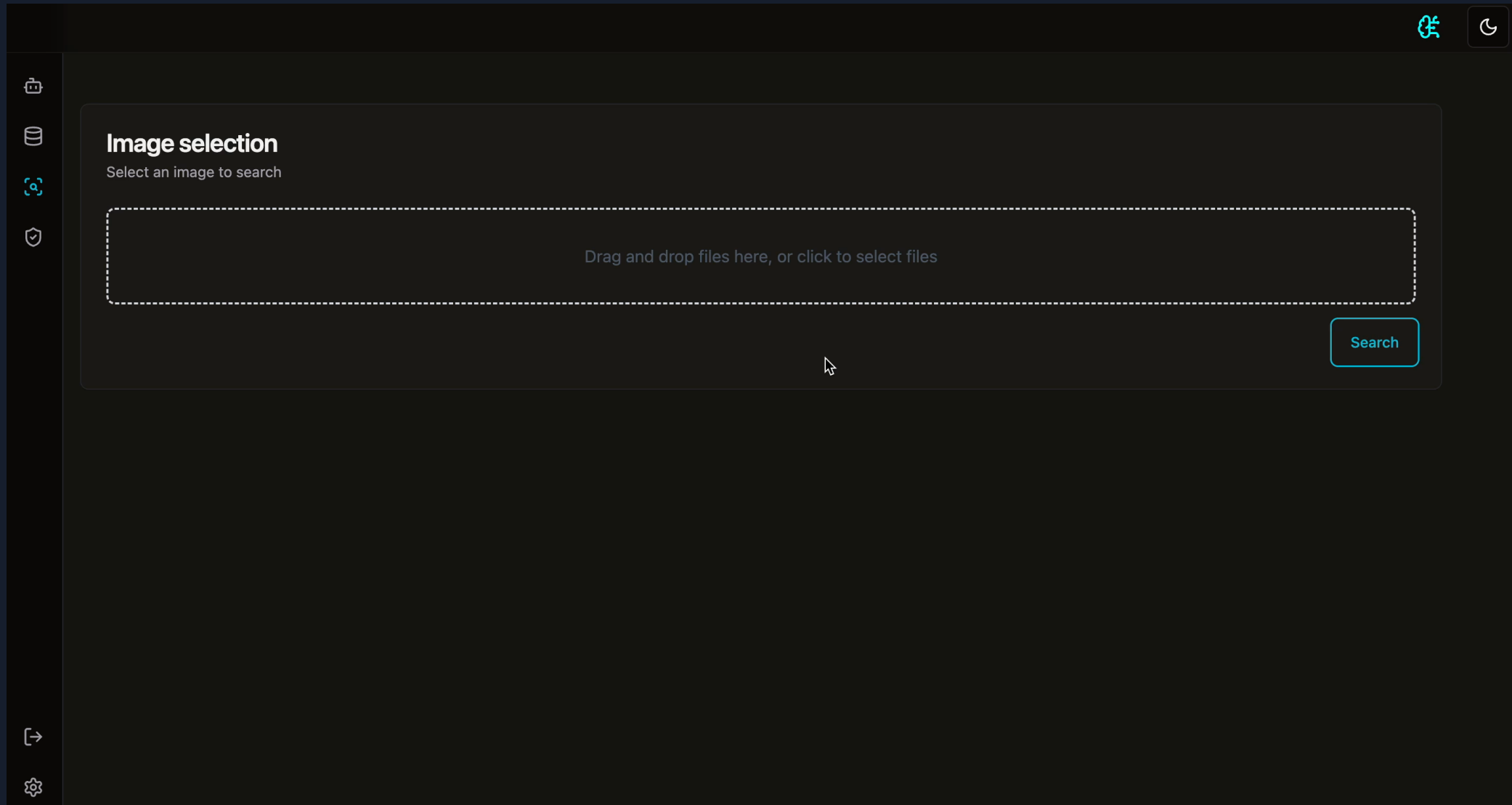
< 1 > [Settings](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	190312-image.jpg	jpg	July 16, 2024, 19:39:41 (UTC+09:00)	591.3 KB	Standard
<input type="checkbox"/>	201001-image.jpg	jpg	July 16, 2024, 19:39:40 (UTC+09:00)	556.3 KB	Standard
<input type="checkbox"/>	211011-image.jpg	jpg	July 16, 2024, 19:39:41 (UTC+09:00)	552.7 KB	Standard
<input type="checkbox"/>	230822-image.jpg	jpg	July 16, 2024, 19:39:42 (UTC+09:00)	582.4 KB	Standard
<input type="checkbox"/>	dominik-lange-BFsm5vldl2l-unsplash.jpg	jpg	July 16, 2024, 19:44:20 (UTC+09:00)	8.5 KB	Standard
<input type="checkbox"/>	laura-college-K_Na5gCmh38-unsplash.jpg	jpg	July 16, 2024, 19:45:36 (UTC+09:00)	83.6 KB	Standard
<input type="checkbox"/>	ray-hennessy-xUUZcpQlqpM-unsplash (1).jpg	jpg	July 16, 2024, 19:46:24 (UTC+09:00)	92.8 KB	Standard

Use case 2: Architecture



Use case 2: 承認済み図表チェック(画像)



Use case 2: 承認済み図表チェック(画像)

Score: 1.0000

Score: 1.0000
Image Path: approved/230822-image.jpg

主要評価項目: 全生存期間 (OS) (優越性試験)
全体集団における全生存期間: OS
OSのKaplan-Meier曲線 (ITT集団)

評価項目	プラチナ治療群 (n=309)	非プラチナ治療群 (n=309)
OS率(95%CI)*, 月	24.4(19.2, 29.6)	16.5(14.5, 19.4)
ハザード比(HR) [95%CI]*, p値	0.67(0.54, 0.84)	p=0.0003
12ヶ月全生存率(95%CI), %	74.8(69.5, 79.3)	63.0(57.9, 68.7)
18ヶ月全生存率(95%CI), %	60.2(54.3, 65.6)	48.3(42.6, 53.0)

※1) 95% CIは95%信頼区間 (Kaplan-Meier) 法に基づき算出された。2) 生存率とは、生存期間の平均値 (平均生存期間) を示す。3) OSとは、全死因による死亡、原因不明の死亡、その他の原因による死亡を含む。4) HRとは、プラチナ治療群と非プラチナ治療群の生存率の差を示す。5) p値は、両群間の生存率の差が偶然によるものである可能性を評価するための指標である。6) 本図表は、2023年8月22日現在のデータに基づき作成されたものである。

Score: 0.8760

Score: 0.8760
Image Path: approved/190312-image.jpg

主要評価項目1: 無増悪生存期間 (PFS) (優越性試験)、副次評価項目: 12ヵ月PFS率
全体集団における無増悪生存期間: PFS
全体集団における12ヵ月PFS率

評価項目	プラチナ治療群 (n=309)	非プラチナ治療群 (n=309)
PFS率(95%CI)*, 月	10.4(8.2, 12.6)	8.2(7.1, 9.3)
ハザード比(HR) [95%CI]*, p値	0.67(0.54, 0.84)	p=0.0003
12ヶ月PFS率(95%CI), %	44.7(39.5, 49.9)	33.0(27.9, 38.1)
18ヶ月PFS率(95%CI), %	33.0(27.9, 38.1)	22.2(17.1, 27.3)

※1) 95% CIは95%信頼区間 (Kaplan-Meier) 法に基づき算出された。2) 生存率とは、生存期間の平均値 (平均生存期間) を示す。3) PFSとは、全増悪、全死因による死亡、原因不明の死亡、その他の原因による死亡を含む。4) HRとは、プラチナ治療群と非プラチナ治療群の生存率の差を示す。5) p値は、両群間の生存率の差が偶然によるものである可能性を評価するための指標である。6) 本図表は、2019年3月12日現在のデータに基づき作成されたものである。

Use case 2: 承認済み図表チェック(画像)

主要評価項目2: 無増悪生存期間 (PFS) (優越性試験)、副次評価項目: 12ヵ月PFS率

- 全体集団における無増悪生存期間: PFS
- 全体集団における12ヵ月PFS率

PFSのKaplan-Meier曲線 (ITT集団)

vs. プラセボ併用群
ハザード比 [95%CI]^{***}, p値^{***}
0.65 [0.53, 0.79], p<0.0001

44.7%^{***}

33.5%^{***}

PFS中央値 10.4ヵ月

PFS中央値 8.2ヵ月

2.2ヶ月の差

1: 12ヵ月PFS率

治験担当医師によるRECISTガイドライン1.1版に基づいた評価
*1 打ち切りデータはproduct-limit (Kaplan-Meier) 法に基づく
*2 投与群を共変量とし、無作為化に用いた層別因子【初回診断時の遠隔転移(あり、なし)、ヘパシスマブの併用投与(あり、なし)、PD-L1の発現状況(CPS<1, 1≦CPS<10, CPS≧10)】を層別因子とした層別Cox比例ハザードモデルに基づく
*3 無作為化に用いた層別因子を層別因子とした層別ログランク検定【片側】、有意水準 $\alpha=0.0012843$ (検証的解析結果) (追跡期間中央値: キイトルーダ®併用群 18.2ヵ月、プラセボ併用群 16.3ヵ月、データカットオフ日: 2021年5月3日)

	キイトルーダ®併用群 (308例)	プラセボ併用群 (309例)
PFS中央値 [95%CI] ^{***} , 月	10.4 [9.1, 12.1]	8.2 [6.4, 8.4]
vs. プラセボ併用群 ハザード比 [95%CI] ^{***} , p値 ^{***}	0.65 [0.53, 0.79]	p<0.0001
12ヵ月PFS率 [95%CI] ^{***} , %	44.7 [38.8, 50.4]	33.5 [28.0, 39.1]

主要評価項目2: 無増悪生存期間 (PFS) (優越性試験)、副次評価項目: 12ヵ月PFS率

- 全体集団における無増悪生存期間: PFS
- 全体集団における12ヵ月PFS率

PFSのKaplan-Meier曲線 (ITT集団)

vs. プラセボ併用群
ハザード比 [95%CI]^{***}, p値^{***}
0.65 [0.53, 0.79], p<0.0001

44.7%^{***}

33.5%^{***}

PFS中央値 10.4ヵ月

PFS中央値 8.2ヵ月

2.2ヶ月の差

1: 12ヵ月PFS率

治験担当医師によるRECISTガイドライン1.1版に基づいた評価
*1 打ち切りデータはproduct-limit (Kaplan-Meier) 法に基づく
*2 投与群を共変量とし、無作為化に用いた層別因子【初回診断時の遠隔転移(あり、なし)、ヘパシスマブの併用投与(あり、なし)、PD-L1の発現状況(CPS<1, 1≦CPS<10, CPS≧10)】を層別因子とした層別Cox比例ハザードモデルに基づく
*3 無作為化に用いた層別因子を層別因子とした層別ログランク検定【片側】、有意水準 $\alpha=0.0012843$ (検証的解析結果) (追跡期間中央値: キイトルーダ®併用群 18.2ヵ月、プラセボ併用群 16.3ヵ月、データカットオフ日: 2021年5月3日)

Score: 1.0000

Score: 0.9521

主要評価項目2: 無増悪生存期間 (PFS) (優越性試験)、副次評価項目: 12ヵ月PFS率

- 全体集団における無増悪生存期間: PFS
- 全体集団における12ヵ月PFS率

PFSのKaplan-Meier曲線 (ITT集団)

vs. プラセボ併用群
ハザード比 [95%CI]^{***}, p値^{***}
0.65 [0.53, 0.79], p<0.0001

44.7%^{***}

33.5%^{***}

PFS中央値 10.4ヵ月

PFS中央値 8.2ヵ月

2.2ヶ月の差

1: 12ヵ月PFS率

治験担当医師によるRECISTガイドライン1.1版に基づいた評価
*1 打ち切りデータはproduct-limit (Kaplan-Meier) 法に基づく
*2 投与群を共変量とし、無作為化に用いた層別因子【初回診断時の遠隔転移(あり、なし)、ヘパシスマブの併用投与(あり、なし)、PD-L1の発現状況(CPS<1, 1≦CPS<10, CPS≧10)】を層別因子とした層別Cox比例ハザードモデルに基づく
*3 無作為化に用いた層別因子を層別因子とした層別ログランク検定【片側】、有意水準 $\alpha=0.0012843$ (検証的解析結果) (追跡期間中央値: キイトルーダ®併用群 18.2ヵ月、プラセボ併用群 16.3ヵ月、データカットオフ日: 2021年5月3日)

	キイトルーダ®併用群 (308例)	プラセボ併用群 (309例)
PFS中央値 [95%CI] ^{***} , 月	10.4 [9.1, 12.1]	8.2 [6.4, 8.4]
vs. プラセボ併用群 ハザード比 [95%CI] ^{***} , p値 ^{***}	0.65 [0.53, 0.79]	p<0.0001
12ヵ月PFS率 [95%CI] ^{***} , %	44.7 [38.8, 50.4]	33.5 [28.0, 39.1]

Score: 0.9519

Score: 0.9519

Image Path: approved/rev-KEYBOTE-826試験_主要評価項目_page-0006.jpg

主要評価項目2: 無増悪生存期間 (PFS) (優越性試験)、副次評価項目: 12ヵ月PFS率

- 全体集団における無増悪生存期間: PFS
- 全体集団における12ヵ月PFS率

PFSのKaplan-Meier曲線 (ITT集団)

vs. プラセボ併用群
ハザード比 [95%CI]^{***}, p値^{***}
0.65 [0.53, 0.79], p<0.0001

44.7%^{***}

33.5%^{***}

PFS中央値 10.4ヵ月

PFS中央値 8.2ヵ月

2.2ヶ月の差

1: 12ヵ月PFS率

治験担当医師によるRECISTガイドライン1.1版に基づいた評価
*1 打ち切りデータはproduct-limit (Kaplan-Meier) 法に基づく
*2 投与群を共変量とし、無作為化に用いた層別因子【初回診断時の遠隔転移(あり、なし)、ヘパシスマブの併用投与(あり、なし)、PD-L1の発現状況(CPS<1, 1≦CPS<10, CPS≧10)】を層別因子とした層別Cox比例ハザードモデルに基づく
*3 無作為化に用いた層別因子を層別因子とした層別ログランク検定【片側】、有意水準 $\alpha=0.0012843$ (検証的解析結果) (追跡期間中央値: キイトルーダ®併用群 18.2ヵ月、プラセボ併用群 16.3ヵ月、データカットオフ日: 2021年5月3日)

業務で生成AIを活用するポイント



LLMを業務適用するアプローチ

	プロンプト エンジニアリング	RAG (Retrieval Augmented Generation)	ファインチューニング	(継続) 事前学習
特徴	既存の大規模言語モデルに適切なプロンプト(入力文)を与えることで、モデルの出力を制御する手法。	大規模言語モデルと外部の知識ソースを組み合わせた手法。情報源に基づいた正確な出力を生成する。	既存の大規模言語モデルを、特定のタスクや応答のデータで追加学習させる手法。	完全に新しい大規模言語モデルを構築する手法。モデルアーキテクチャ、をゼロから設計する。
メリット	既存モデルを活用できるので開発コストが低い	言語モデルの知識不足を補完できる	既存のモデルにユーザの望むふるまいをさせられる	既存のモデルにない新しい知識を獲得できる
デメリット	プロンプトの設計次第で出力の質が大きく変わる	外部知識ソースの構築が必要	学習データセットの作成が必要	莫大な計算資源とデータが必要
開発コスト (金額は目安)	低 (数万円)	中 (数十万円)	高 (数万~数百万円)	大~特大 (数百万~数億)

ファインチューニングや独自モデル(事前学習)のアプローチが必要でしょうか

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE



ファインチューニングや独自モデル作成の前に

- 最新モデルの利用
- プロンプトエンジニアリング
- Advanced RAG

最新モデルの利用

4.3 実験結果

GPT-4とGPT-3.5において、Function calling を利用した分類とプロンプトのみを用いた分類の比較を表 2 に示す。評価指標として、精度、適合率、再現率、f-1 値を使用した。平均はマクロ平均を用いた。どちらのモデルにおいても function calling を用いた手法はプロンプトのみを用いた手法に比べて顕著な精度向上を達成している。特に GPT-4 の Precision では、専用モデルを構築し学習を行った先行研究と同等程度の精度をゼロショットで実現していることが示されている。これは、GPT-4 が豊富な事前学習データを基に、特定のタスクにおいて高い性能を発揮する能力を有していることを示唆している。図1に分類の詳細を示す。さらに、ファインチューニングを行った GPT-3 を用いた分

	GPT-4		GPT-3.5	
	Function calling	Prompt only	Function Calling	Prompt only
Accuracy	0.751	0.711	0.725	0.679
Precision	0.808	0.729	0.761	0.688
Recall	0.703	0.636	0.666	0.528
F1-score	0.728	0.641	0.685	0.562

表 2 Function calling の利用による分類性能

	Training Data: 50, Epoch: 3	Training Data: 250, Epoch: 3	Training Data: 250, Epoch: 9
	Function calling		
Accuracy	0.799	0.814	0.807
Precision	0.777	0.798	0.787
Recall	0.772	0.802	0.793
F1-score	0.774	0.797	0.786
Prompt only			
Accuracy	0.792	0.771	0.794
Precision	0.776	0.756	0.772
Recall	0.786	0.757	0.783
F1-score	0.774	0.748	0.773

表 3 ファインチューニングの効果 (GPT-3.5)

- 旧モデルを学習データでファインチューニングしても、新モデル(ファインチューニングなし)で同程度の精度が出る可能性が高い。
- 新モデルが出るたびにファインチューニングすると、コストと時間が大幅に必要。



最新モデルとLLM周辺機能の改善だけで要件を満たせる可能性あり

https://www.jstage.jst.go.jp/article/pjsai/JSAI2024/0/JSAI2024_1B3GS201/_pdf/-char/en

プロンプトエンジニアリング

- プロンプト分割
 - 画像内容の抽出と、ガイドラインを利用したレビューのPromptを分割

```
new HumanMessage({
  content: [
    ...imagePrompt,
    {
      type: "text",
      text: "画像に書かれている全てのテキストを抽出してください。その際、図や表があれば内容を詳細に記述してください(テキスト、図表、色使い、レイアウトなど)。特に、強調部分(どんな色を用いて、どのように強調しているか)には注意して抽出してください。",
    },
  ],
}),
```

Advanced RAG

Query Optimization (Pre-retrieval)

リクエスト(質問)の品質が悪く、
求める回答が得られない

- Query expansion

検索ドキュメント

検索結果(類似度が高い内容)が
回答に必要な内容と異なる

- ドキュメントの質の向上

Post-retrieval

検索結果(類似度が高い内容)が
回答に必要な内容と異なる

- Re-ranking
- Contextual Compression
- Ensemble



Query Optimization : Query expansion

画像は医療用医薬品の臨床成績を示すスライドです。ガイドラインに違反していないか、各画像をレビューしてください。違反があればガイドラインを引用して指摘してください



```
const retriever = MultiQueryRetriever.fromLLM({  
  llm,  
  retriever: vectorStore.asRetriever(),  
  verbose: true,  
});  
const relevantDocs = await retriever.invoke(input.question);
```



"医療用医薬品の臨床試験結果を示す画像について、製薬業界のガイドラインに準拠しているかを確認する",
"臨床試験の結果を示す画像が、規制当局の要件を満たしているかを評価する",
"医療用医薬品の臨床試験データを示す画像が、倫理的および法的な基準を満たしているかをレビューする"

ドキュメントの質の向上

The screenshot shows the 'Contents Registration' interface in the AWS IAM console. It is divided into two main sections: 'Documents upload' and 'Embedding Job'. The 'Documents upload' section has a dashed box for file uploads. The 'Embedding Job' section contains several input fields: 'Table name', '(optional) Description', '(optional) Extract keywords', and 'Files'. The 'Files' section has a 'Select a file...' button. The 'Extract keywords' field is highlighted with a red box and contains the text 'しないこと,してはいけない'.

Keywordが含まれる箇所を抜粋したRAG用データソースを作成することで、
関連が薄い文書のヒット率を下げる

Post-retrieval

- Ensemble

```
const retriever = new EnsembleRetriever({  
  retrievers: [imageRelevantRetriever,  
              multiQueryRelevantRetriever],  
  weights: [0.5, 0.5],  
});
```

- Re-ranking

- Cohere Rerank 2 Model – Multilingual

```
const rerankedDocuments = await cohereRerank.rerank(docs, query, {  
  topN: 5,  
});
```

```
[{ index: 3, relevanceScore: 0.9871293 }, { index: 1, relevanceScore: 0.29961726 }, ...]
```

The screenshot displays a configuration panel for a RAG system. On the left, there are three main steps:

- Analyze images and search similarity**: Analyzing images with LLM and searching in the document.
- Post-processing**: Filtering or ranking retrieved documents.
- Review the documentation**: Reviewing the documentation and generating the result.

On the right, the **Configurations** section includes:

- Model**: A dropdown menu set to `claude-3-sonnet-20240229-v1:0`.
- Temperature**: A slider set to 1.
- Top K**: A slider set to 250.
- Top P**: A slider set to 0.99.
- RAG DataSource**: A dropdown menu set to `extract_med_pharm_products`.
- Post-processing**: A dropdown menu with options for `Ensemble` and `Re-ranking`.

まとめ

□ 生成AIの業務適用のアプローチは複数あり、正解は一つではない

□ アプローチ(手法)起点ではなく、解決したい課題起点でAWSにご相談ください

	プロンプト エンジニアリング	RAG (Retrieval Augmented Generation)	ファインチューニング	(継続) 事前学習
				
特徴	既存の大規模言語モデルに適切なプロンプト(入力文)を与えることで、モデルの出力を制御する手法。	大規模言語モデルと外部の知識ソースを組み合わせた手法。情報源に基づいた正確な出力を生成する。	既存の大規模言語モデルを、特定のタスクや応答のデータで追加学習させる手法。	完全に新しい大規模言語モデルを構築する手法。モデルアーキテクチャ、をゼロから設計する。
メリット	既存モデルを活用できるので開発コストが低い	言語モデルの知識不足を補完できる	既存のモデルにユーザの望むふるまいをさせられる	既存のモデルにない新しい知識を獲得できる
デメリット	プロンプトの設計次第で出力の質が大きく変わる	外部知識ソースの構築が必要	学習データセットの作成が必要	莫大な計算資源とデータが必要
開発コスト (金額は目安)	低 (数万円)	中 (数十万円)	高 (数万~数百万円)	大~特大 (数百万~数億)

Thank you!

