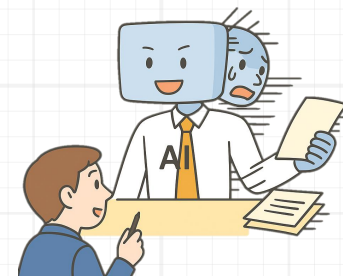


2025年5月15日（木）

“動く” から “使える” への **ステップアップ**！

RAGの精度は、ここまで上げられる

～生成AIをビジネスで本気で使うために必要な“精度の上げ方”を伝授～



業務効率化に結びつくRAG活用の考え方が分かる



0 | 当社について

NTT Data 株式会社NTTデータ数理システム

「数理学とコンピュータサイエンスにより現実世界の問題を解決する」をミッションに

累計**8,000**件以上のプロジェクトに真摯に向き合い

技術力を研ぎ澄ましてきた、**数理学のプロフェッショナル集団**です

会社概要

- 会社名** 株式会社NTTデータ数理システム
- 所在地** 東京都新宿区信濃町35 信濃町煉瓦館1階
- 資本金** 5,600万円（NTTデータ100%出資）
- 従業員数** 約130名（うち75%以上が技術者）

主要事業

パッケージソフトウェア開発

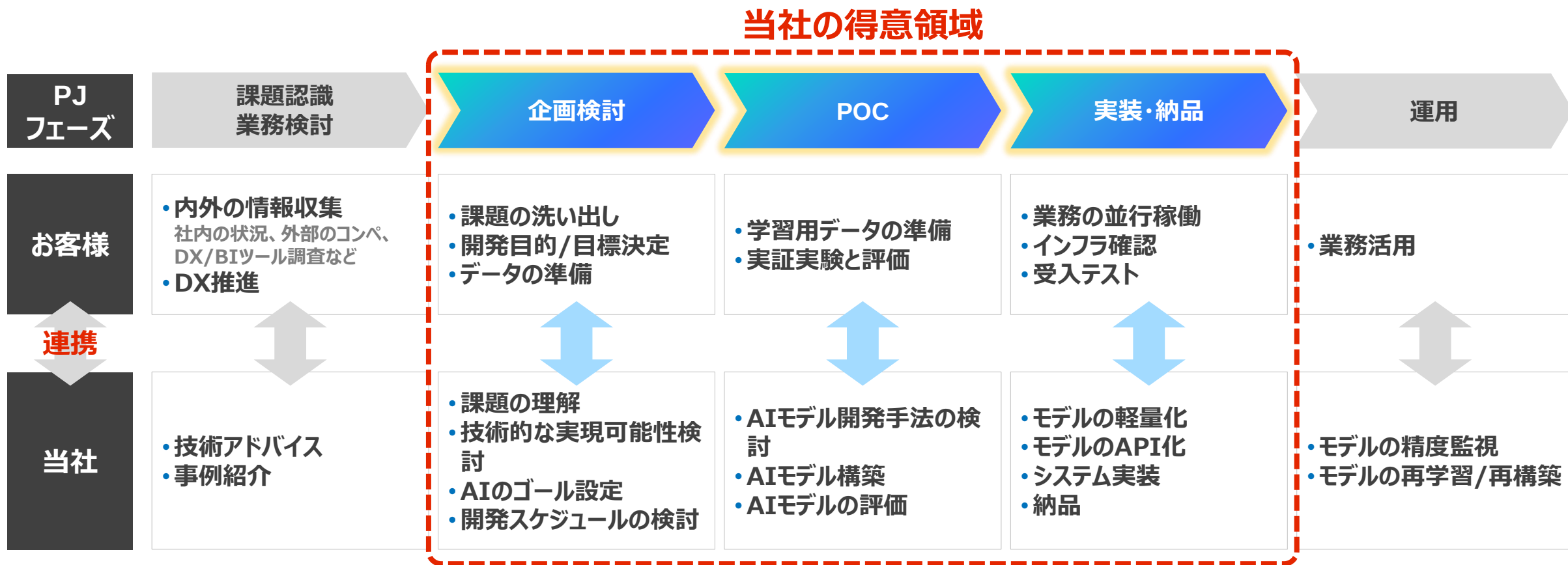


分析
コンサルティング

データサイエンス
教育

受託分析・開発

受託・コンサルティングサービス プロジェクトの進め方イメージ



当社がコアロジック部分を実装し、スクリプト、実行ライブラリ・モジュール、GUIを伴うアプリケーションとしてご提供します
お客様のシステム部門（子会社）やグループ会社のSIerと連携してシステム組み込みを進めることも多くございます

1 | イン트로ダクション

——なぜ、いまRAGに注目が集まるのか？

ついに東大にも合格する生成AI

生成AIはますます賢さを増しており、OpenAI社の推論系モデルChatGPT-o1は
共通テスト（旧センター試験）・2次試験ともに**最難関大学入試合格圏内**との調査^[1-3]も

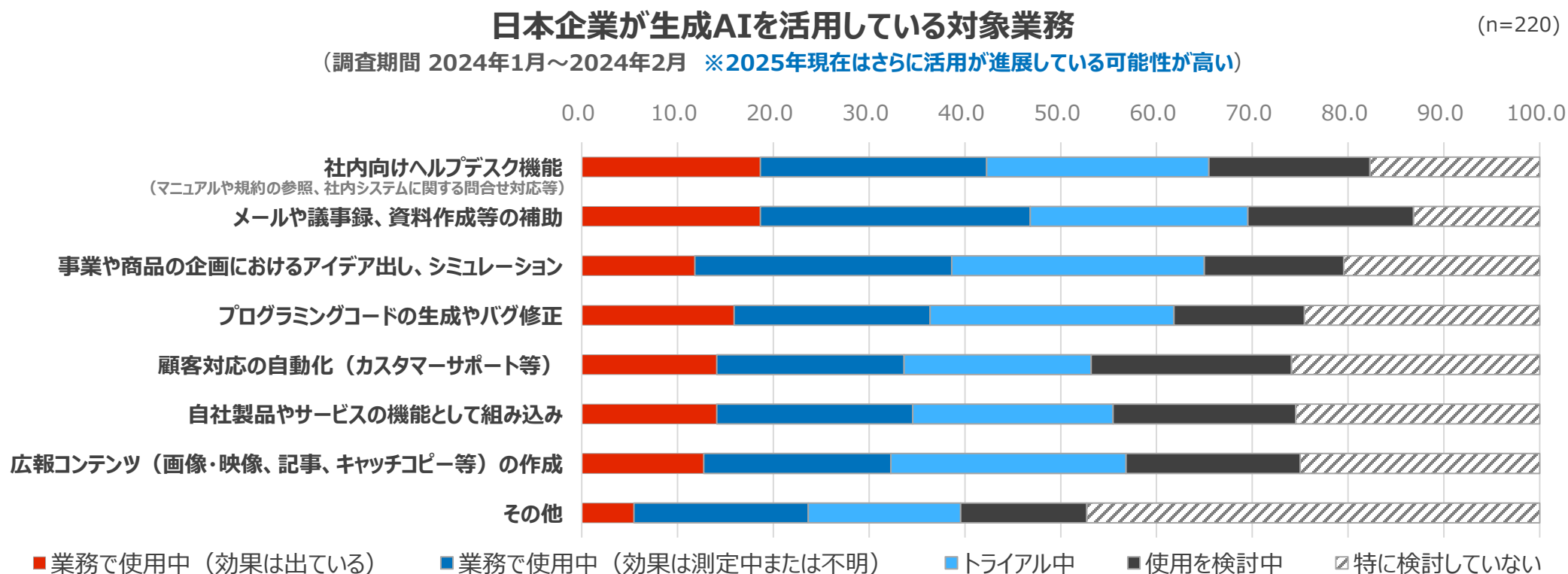


最難関大学入試合格者に匹敵する、幅広い知識・深い思考力・高い表現力を持つ

[1] 日経クロステック／日経コンピュータ 2025年1月29日『特集 大学入試 vs. 生成AI 共通テスト・東大入試に挑戦——第1回 「o1」が2025年度共通テストの得点率で9割超、新科目「情報I」でも高得点』 <https://xtech.nikkei.com/atcl/nxt/column/18/03078/012800002/>
[2] 日経クロステック／日経コンピュータ 2025年2月14日『特集 大学入試 vs. 生成AI 共通テスト・東大入試に挑戦——第3回 o1が2025年東大入試の1次選抜を突破、苦手の数学を克服し「合格」の可能性も』 <https://xtech.nikkei.com/atcl/nxt/column/18/03078/012800005/>
[3] 日経クロステック／日経コンピュータ 2025年4月7日『特集 大学入試 vs. 生成AI 共通テスト・東大入試に挑戦——第4回 OpenAI o1が2025年東大文系入試を余裕で突破、理三合格レベルにも到達』 <https://xtech.nikkei.com/atcl/nxt/column/18/03078/012800006/>

日本企業でも進んできた生成AI活用

社内業務効率化を中心に、生成AI活用のトライアル・試用が進んでいる
2024年時点で「**効果は出ている**」と回答した企業は僅かだが、**今後増加が予想される**



総務省『国内外における最新の情報通信技術の研究開発及びデジタル活用の動向に関する調査研究』（2024）より。グラフは株式会社NTTデータ数理システムが作成

生成AIを社内活用する上での課題

生成AIは公知情報のみで学習されているので、そのままでは公知情報しか答えられない

ChatGPTに当社の社内規則（非公知情報）について尋ねても...

NTTデータ数理システムの社員です。出張したら日当はいくらもらえるのですか？

NTTデータ数理システムの出張時の日当（出張手当）について、具体的な金額や支給条件は、公開されている情報からは確認できませんでした。

一般的に、出張手当の金額や支給条件は、企業の就業規則や出張規程に基づいて定められており、出張先の地域や出張の目的、期間、職位などによって異なる場合があります。

詳細な情報を知りたい場合は、社内の人事部門や総務部門に直接お問い合わせいただくことをおすすめします。

答えてくれない！

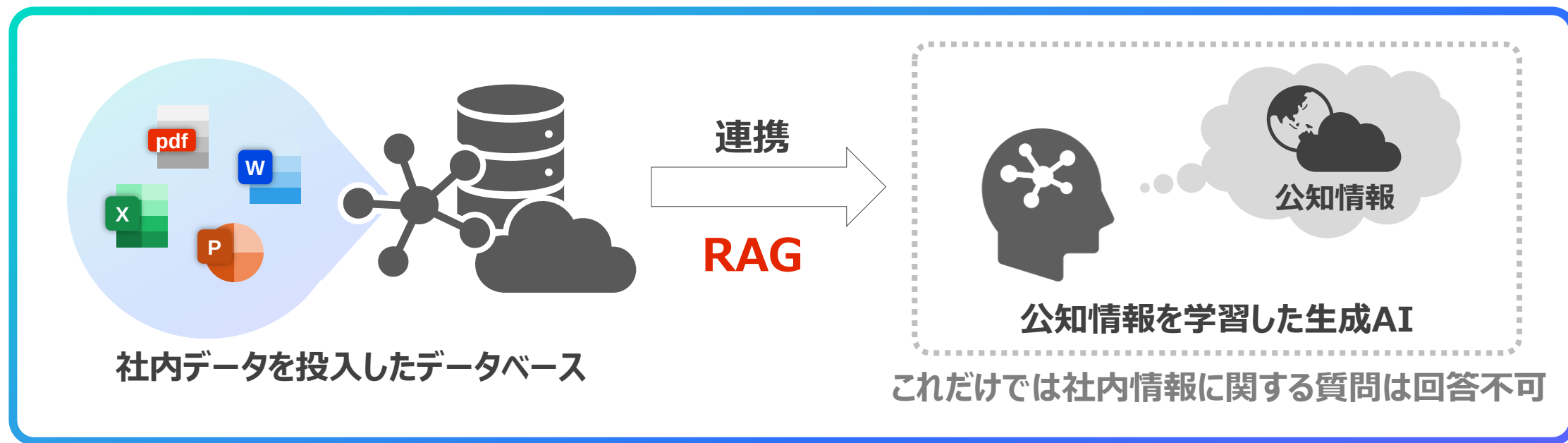
ChatGPTを作っているOpenAI社からは
当社の社内文書は確認できないので
AIが学習しようがない！

**生成AIのビジネス活用で重要な「社内情報に基づく回答」を実現するためには
別の仕組みで社内データと連携する必要がある！**

※株式会社NTTデータ数理システムがChatGPTを利用して生成した結果および、OpenAI社のChatGPTの画面の一部を含みます。

データ連携を実現する仕組みがRAG

各企業個別の情報を活用するためには、**データ連携の仕組み**が別に必要
データ連携の仕組みとして**最も幅広く使われているのがRAG**



社内情報に関する質問が回答できる

生成AI活用のキーとして、RAG（Retrieval-Augmented Generation）への注目が高まる

RAGを使うと…

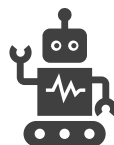
各種データと連携して、そのデータについての知識を補って回答させることが可能に

RAG無し



NTTデータ数理システムの社員です。
出張したら日当はいくらもらえますか？

公開されている情報からは確認できませんでした。



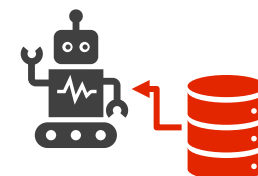
知識不足で回答を拒否

RAG有り



NTTデータ数理システムの社員です。
出張したら日当はいくらもらえますか？

旅費規程第14条によると、日当は1
日当たり5,000円支給されます。



RAGで規程類などの
データと連携

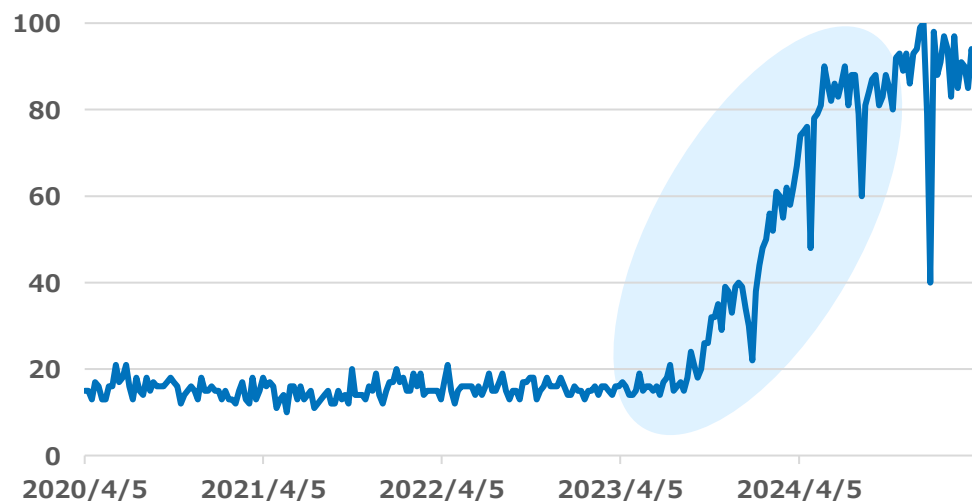
データと連携して知識を補って
正答を導ける

RAGへの注目

2023年4月ごろから、急激にRAGへの注目度が上昇
多くのRAGソリューションが市場に投入され、データとの連携が容易になりつつある

Google検索ワード「RAG」のインタレストの変遷

2023年4月ごろから急速にインタレストが上昇



NTTデータのRAGソリューション

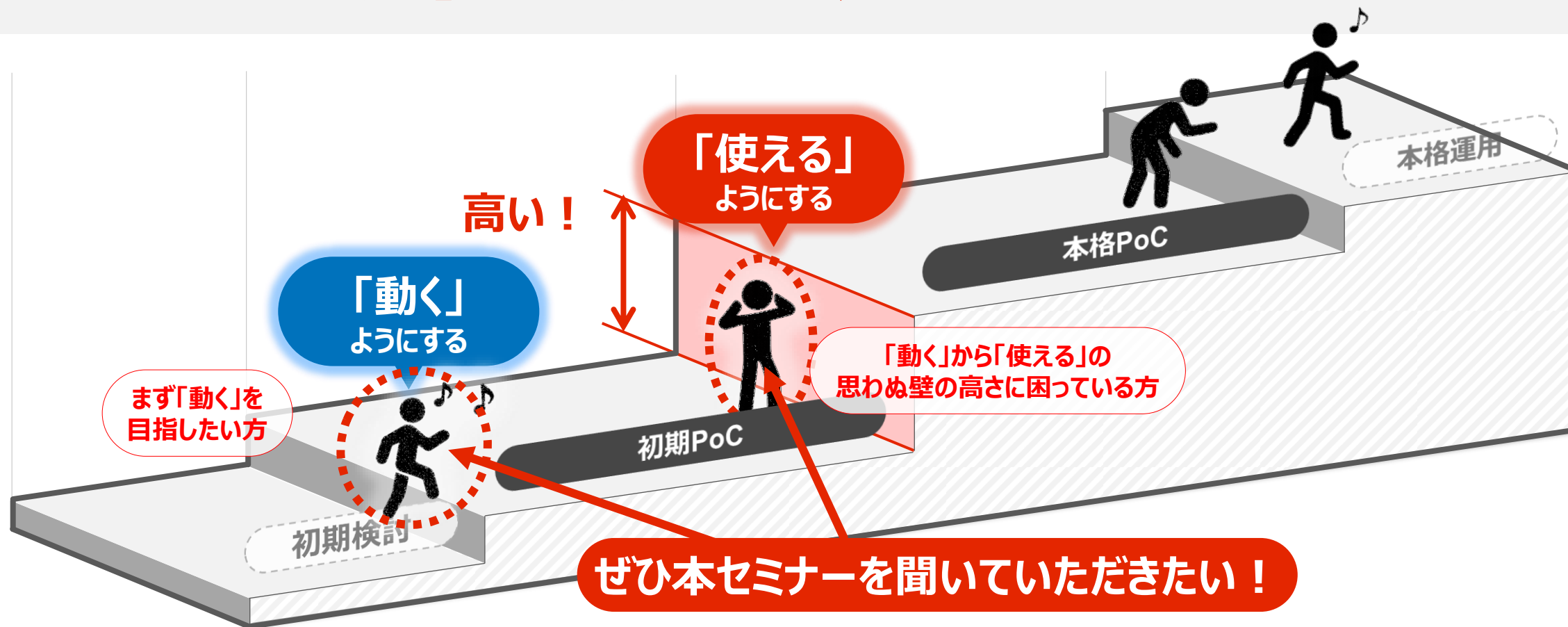
2023年6月末、データ連携機能をもった
LITRON® Generative Assistantリリース

The screenshot displays the LITRON Generative Assistant interface. On the left, a dark blue banner reads '多様なデータを連携させて根拠ある回答文を作成する 生成AIサービスを提供開始' (Start providing generative AI services that create evidence-based responses by linking diverse data). Below this, a section titled '2023年6月29日 株式会社NTTデータ' (June 29, 2023, NTT Data Corp.) provides details about the service. The main part of the interface shows a flow: 'データストア (お客様固有データ)' (Data Store (Customer-specific data)) containing '旅費規程一式.xlsx' and 'FAQ_交通費.txt', leading to 'データ検索' (Data Search), then '回答作成' (Response Creation), and finally '生成AIサービス 大規模言語モデル' (Generative AI Service Large Language Model). On the right, a '利用イメージ：出張旅費' (Usage Example: Business Travel Expenses) section shows a user asking '出張費の宿泊上限はいくらですか？' (What is the accommodation limit for business travel expenses?). The assistant responds with a detailed answer based on the '旅費規程一式.xlsx' file, mentioning specific limits and company policies.

社内プロジェクトとしてRAG構築のPoCを進めている企業も増えてきたが...

「動く」と「使える」の間の壁は意外に高い！

各種サービスのリリースにより、RAGを「動かす」ところまでは手軽になってきたでも、「使える」までの壁が意外と高く、悩んでいる企業が多い...？



※図中の一部に、株式会社NTTデータ数理システムが、ChatGPTを利用して生成した絵や図柄が含まれます。

「動く」から「使える」の壁とは？——こんなお悩みありませんか？



とりあえずRAGが動くところまで進んだけど、社内で「これ微妙じゃない？」「これじゃ業務で使えないよ」と言われ、PoC倒れになりそう...

ユーザーの期待値が高すぎて、もはや「使えない」ものは出せない...



クライアントにRAGを提案したいけど、対象文書が複雑すぎて、思うように精度が出せそうにない...

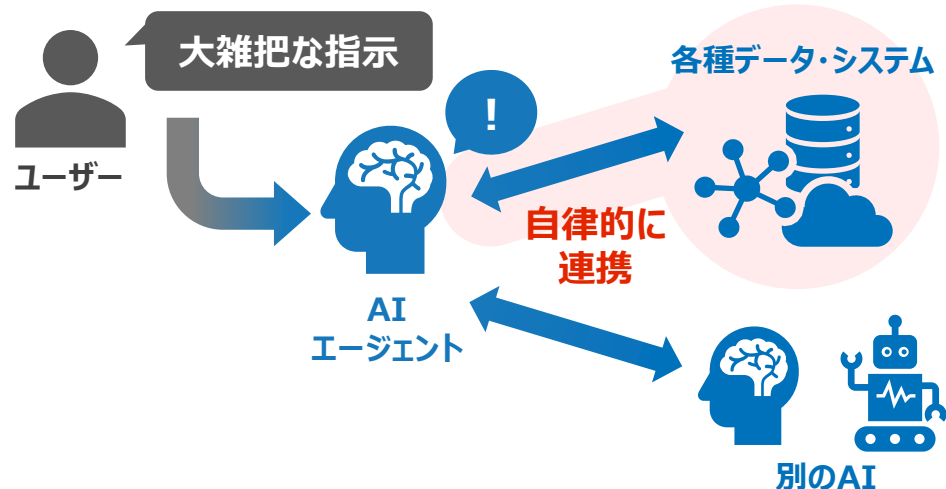
“なんとなく動く”じゃ許されない皆様に向けて、できるだけ具体的にノウハウを共有！

ところで、「AIEージェント」も登場してきたけど…

2024年末ごろから、「AIEージェント」というキーワードも急速に注目を集めるが、
RAGはAIEージェントとデータを連携するための重要な要素技術

よくあるAIEエージェントの説明

周囲のシステムと自律的に連携し臨機応変に動作するAI
大雑把な指示から、必要な情報を自律的に集めて課題解決



**AIEエージェントが大量のデータと連携するためにも
RAGは要素技術として重要！**

- コンテキスト長が数百万tokenと極めて長くなっているとはいえ、まだまだ大規模なデータは入りきらない
- ビジネス用途では、データ自体の管理（文書バージョン管理や権限付与など）も必要であり、その仕組みとしてのデータベースは必要になる

データを活用したいと思う限り、RAGの重要度は依然として高い

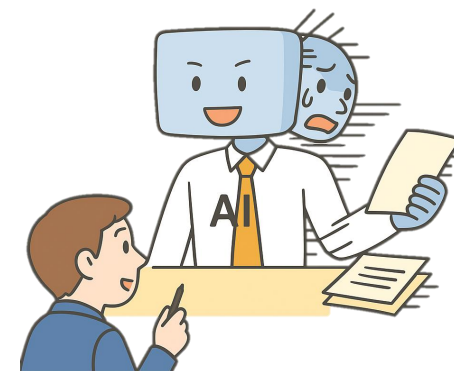
（AIEエージェントがあるから不要というものではなく、むしろRAGとAIEエージェントは地続きの技術）

本セミナーについて

生成AI活用プロジェクトの大きな壁である、**RAGの「動く」から「使える」の壁**を超えるための「考え方」と「進め方」のコツを、具体的事例も交えて共有します！

本日のトピック

1. イントロダクション——なぜ、いまRAGに注目が集まるのか？
2. 「考え方」のコツを知る
RAGの仕組みを完全解説——非エンジニアでも分かる！実装と改善の勘所
3. 「進め方」のコツを知る
RAG精度向上への道——つまずきやすい課題と、現場で使える解決策
4. 事例で知る
ケーススタディ——「教えて！Alkano先生」で実現した精度改善の舞台裏



※NTTデータ数理システムがChatGPT 4oを用いて生成したイラストを、加工して利用しています。

2 | RAGの仕組みを完全解説 ——非エンジニアでも分かる！実装と改善の勘所

なぜ、RAGの「仕組み」が大事なのか？

RAG改善の考え方——全体像

精度・コストが常に優れた「万能な方法」があるわけではない
観察→**仮説**→改善を繰り返す、地道な取り組みが重要！

考え方のステップ

入出力の観察

重要！

仮説思考

改善策立案

ステップの詳細

- RAGの入出力や中間アウトプットを観察し、どういうケースで精度が低くなるのか特定する
- 現在の出力とあるべき出力の差を観察・言語化し、不十分な点を特定していく

- なぜ現在の出力が不十分なのか、その原因について仮説を立てる

- 仮説に基づき、どうしたら精度が向上するのか、施策を立案する

考え方の例



こういうタイプの質問の場合に精度が低くなっているな… 本来はこれぐらいのクオリティまで欲しいのに、現状はこれぐらいのクオリティしか出ていない。〇〇という観点が入っていないな。



本来のクオリティの出力を得るためには、この情報を基にこう考えなきゃいけないはず！でも、このRAGでは別の情報を基に、こういう別の考え方をしているから、「精度が低い」のでは？

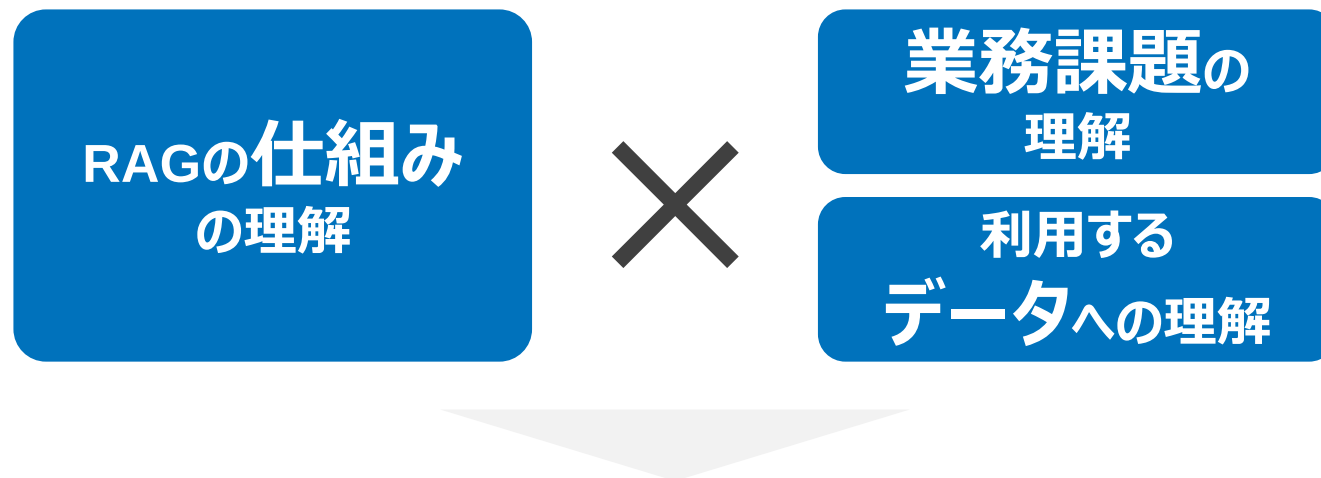


この部分を直したら…、確かに精度が向上しそうだ！だから、RAGの仕組みをこのように変えたら、このタイプの質問の出力は改善できそう！

改善を実現するには、「的を射た仮説」を立てることが重要

的を射た仮説を立てるためには…

①RAGの仕組みの理解と、②業務課題・利用するデータについての理解の両方が重要



これらの理解があってはじめて、**的を射た仮説**が立てられる



本来のクオリティの出力を得るためには、この情報を基にこう考えなきゃいけないはず！
でも、今の仕組みでは別の情報を基に、こういう別の考え方をしているから、「精度が低い」のでは？

そのために、RAGの仕組みをわかりやすく解説！

RAGの仕組み①——なぜ、RAGにより知識補間が実現できるのか？

1. ヒントにより、「知識問題」から「読解問題」へ転換
2. 「ヒント」は文書検索システムを利用して探してくる

ヒントにより、「知識問題」から「読解問題」へ転換

「ヒントとなる文書」を指示文中に加えることで
AIは文書から知識を読み取って答えを導けるようになる

ヒント無し

知識問題

NTTデータ数理システムの社員です。
出張したら日当はいくらもらえますか？



指示文
(プロンプト)

AIの回答

公開されている情報からは確認できませんでした。



知識不足で回答を拒否

ヒント有り（指示文にヒントとして旅費規程）

読解問題

NTTデータ数理システムの社員です。
出張したら日当はいくらもらえますか？

ヒントとなる文書

ただし、以下を参考情報として回答してください。
旅費規程
第14条 日当は、1日あたり5,000円を支給する。

※説明のために作成した仮の内容であり、実際の社内規定とは異なります。

旅費規程第14条によると、日当は1日当たり5,000円
支給されます。



ヒントをもとに読解して、正解を導ける

ヒントが無い場合



ChatGPT ▾



NTTデータ数理システムの社員です。
出張したら日当はいくらもらえますか？

NTTデータ数理システムの社員として出張した際の日当（出張手当）について、具体的な金額は公開されていません。一般的な企業の出張手当の相場を参考にすると、以下のようになります：
出張支援クラウド「BORDER」

国内出張（日帰り）の相場

- 一般社員：約1,950円～2,200円
- 課長クラス：約2,300円～2,600円
- 部長クラス：約2,500円～2,800円
- 取締役：約3,100円～3,500円 出張支援クラウド「BORDER」+1

国内出張（宿泊あり）の相場

- 一般社員：約2,200円～2,500円
- 部長クラス：約2,800円～3,000円

海外出張の相場（北米の場合）

- 一般社員：約4,900円～5,200円
- 部長クラス：約5,600円～6,000円



回答を拒否し
一般論を回答

※株式会社NTTデータ数理システムがChatGPTを利用して生成した結果および、OpenAI社のChatGPTの画面の一部を含みます。

ヒントがある場合（指示文に旅費規程も載せた場合）



ChatGPT ▾



NTTデータ数理システムの社員です。
出張したら日当はいくらもらえますか？

ただし、以下を参考情報として回答してください。
旅費規程
第14条 日当は、1日あたり5,000円を支給する。

ヒントアリだと…

ご質問ありがとうございます。

NTTデータ数理システムの旅費規程第14条に基づく、出張した場合の日当は1日あたり5,000円が支給されます。

したがって、出張日数に応じて、

- 1日出張：5,000円
 - 2日出張：10,000円
 - 3日出張：15,000円
- …と日数×5,000円が支給額となります。

不明点があれば、経理部門や上司に確認するとより確実です。

出張にかかる交通費や宿泊費の扱いもご確認されますか？

📄 📌 🗣️ 🔄 📄

ヒントを基に
正しく回答



ChatGPT ▾



このヒントを自動的に取得してくれば
「知識補間」が実現できる

NTTデータ数理システムの社員です。
出張したら日当はいくらもらえますか？

ただし、以下を参考情報として回答してください。
旅費規程
第14条 日当は、1日あたり5,000円を支給する。

ご質問ありがとうございます。

NTTデータ数理システムの旅費規程第14条に基づくと、出張した場合の日当は1日あたり5,000円が支給されます。

したがって、出張日数に応じて、

- 1日出張：5,000円
 - 2日出張：10,000円
 - 3日出張：15,000円
- ...と日数×5,000円が支給額となります。

不明点があれば、経理部門や上司に確認するとより確実です。

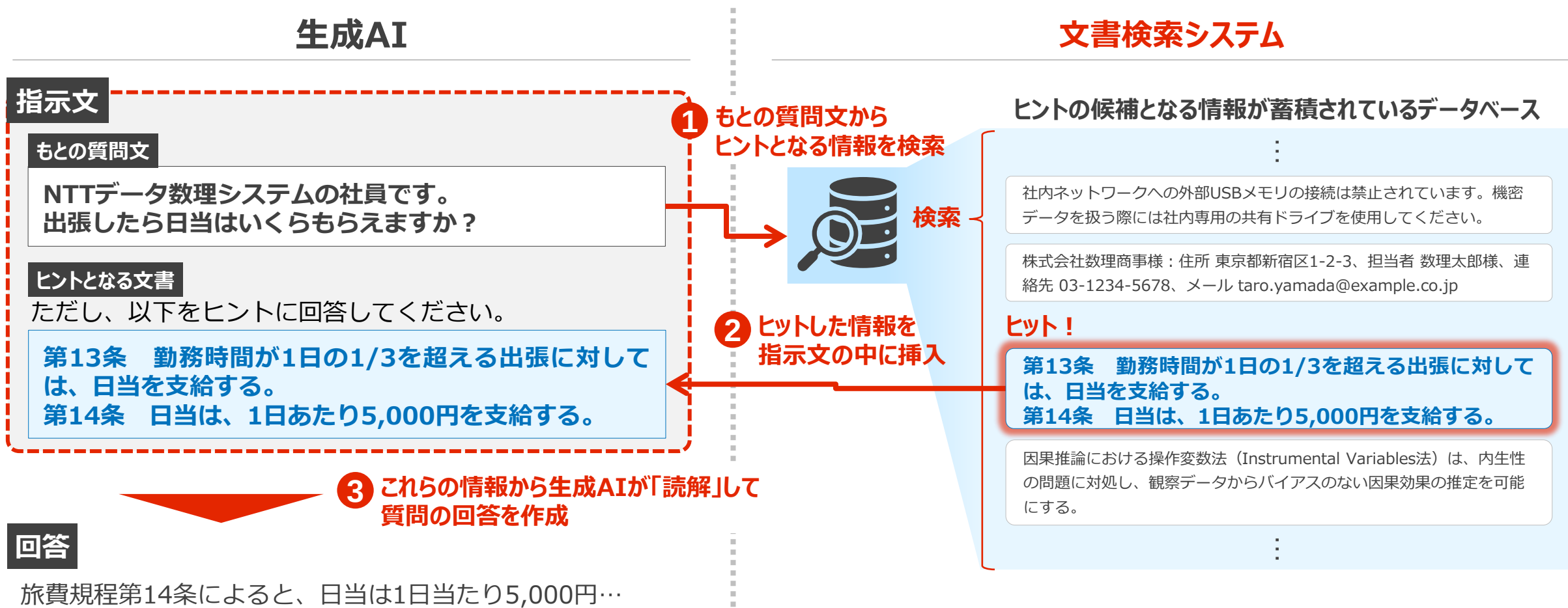
出張にかかる交通費や宿泊費の扱いもご確認されますか？



※株式会社NTTデータ数理システムがChatGPTを利用して生成した結果および、OpenAI社のChatGPTの画面の一部を含みます。

「ヒント」は文書検索システムを利用して探してくる

質問文をもとにヒントとなる情報を検索し、プロンプトに追記することで「読解問題」にする

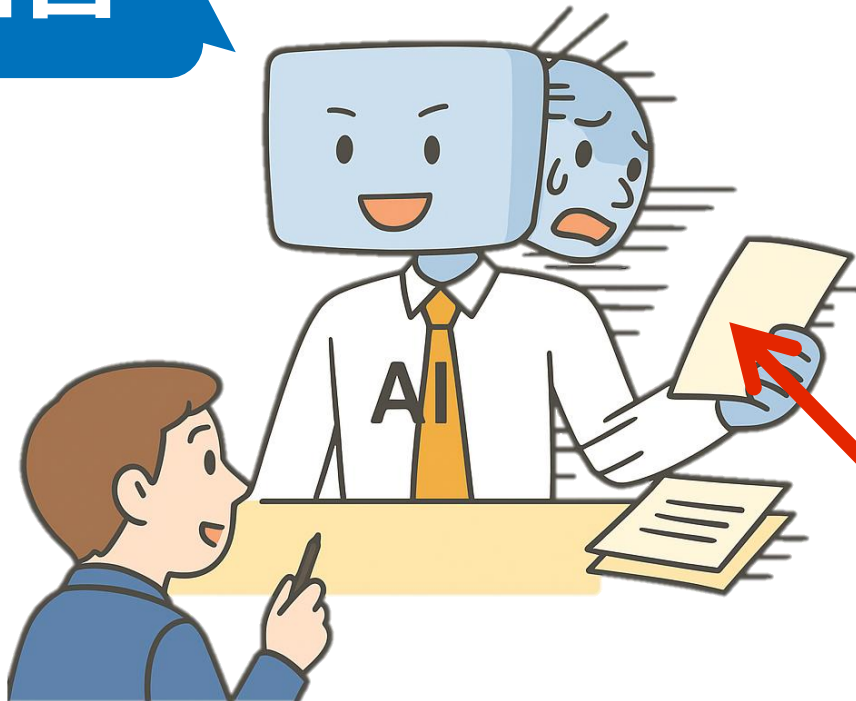


つまり、RAGのイメージは…

回答

何も見なくても、
一流大学に入れるぐらいの知識がある賢いAIだが…

質問



個社業務については何も知らないので、
資料をその場で理解し、
回答を組み立てている…

検索システムが探してきた文書

※NTTデータ数理システムが、ChatGPT 4oを用いて生成したイラストを加工して利用しています。

RAGの仕組み②——類似検索システムの利用と構築

RAGと連携する検索システム

RAGと連携するデータベース・検索システムには様々な種類がある
それぞれ得意・不得意が異なるため、データにあわせて選定することが重要だが…

ベクトルDB



データをベクトル形式で格納
類似検索

リレーショナルDB



データをテーブル形式で格納
構造化データの管理

グラフDB



データをグラフ形式で格納
複雑な関係の解析

ドキュメントDB



データをJSON形式で格納
柔軟なデータ構造の保存

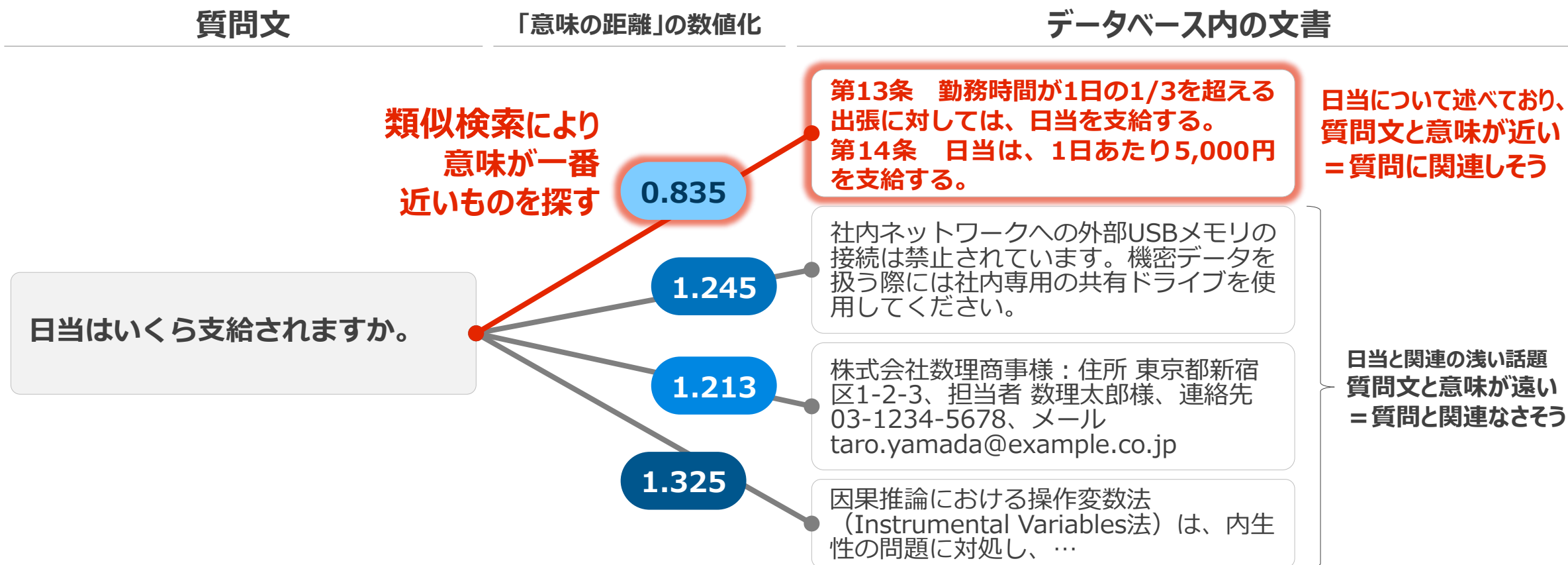
RAGでよく使われている！

まずは万能で使いやすいベクトルデータベースに基づく、類似検索を利用するのが定石

文書の類似検索とは

候補となる文書群の中から、「意味の近い」似た文書を探す仕組み

キーワード検索よりも柔軟に、質問文と「意味の距離」が小さい＝関連しそうな文書を検索可能



【参考】類似検索を支える技術

「意味の距離を計算」（数値化）できるのは不思議だが…
文書埋め込み（sentence embedding）とよばれる自然言語処理技術で実現される

【参考】文章埋め込み（sentence embedding）

文章埋め込み = 「文章」を「ベクトル」（数値の列）に変換する操作
ただし、意味の近い文章同士は、埋め込み後も距離※の「近い」ベクトルになる

「文章」の世界

年次有給休暇の付与日数は、社員の勤続年数に応じて以下の通りとする。入社後6か月経過時点：12日付与……繰り越し上限日数は最大で48日とし、これを超える部分は消滅する。

埋め込み

「ベクトル」（数値の列）の世界

3072次元のベクトル
text-embedding-3-largeの場合

意味の近い文章

距離※の「近い」ベクトル

逆に、似ていない意味・ジャンルの文章同士は、距離の「遠い」ベクトルに埋め込まれる

※ベクトル $x = (x_1, \dots, x_d)^T$ と $y = (y_1, \dots, y_d)^T$ の距離としては、たとえばユークリッド距離 $\|x - y\|_2 = \sqrt{(x_1 - y_1)^2 + \dots + (x_d - y_d)^2}$ を用いる。他、コサイン類似度や内積などの値を用いることもある。

© 2025 NTT DATA Mathematical Systems Inc. 77 NTT DATA 株式会社NTTデータ数理システム

【参考】なぜ言葉を「ベクトル」に変換できるの？

似た意味の言葉・文は、似たような使われ方をする（分布仮説）
膨大なコーパスから言葉の「使われ方」のパターンを数値化することで、意味を捉えた「ベクトル」に

「野球」と意味的に近い

野球

バット ボール グローブ

サッカー ラケット テニスコート

コンピュータ 料理 日本酒

「野球」と共に使われることが多い言葉
例）野球はバットとボールとグローブを使うスポーツです。

「野球」と共に使われることが少ない言葉

言葉の使われ方のパターンを、統計手法や深層学習等の機械学習手法で捉えて数値化している

※最近の「文章埋め込み」では、単にここで説明している「言葉ごとの共起関係」だけではなく、GPTやBERT等の注意機構を備えたモデルで文脈を含めた学習をすることがキモになっています。
※あくまでもイメージの説明です。詳細は専門書を参照してください。

© 2025 NTT DATA Mathematical Systems Inc. 78 NTT DATA 株式会社NTTデータ数理システム

資料末尾の「補足」にて説明していますので、ご興味ありましたら後ほどご覧ください！

© 2025 NTT DATA Mathematical Systems Inc.

31

NTT DATA 株式会社NTTデータ数理システム

文書データからベクトルデータベースを構築する

各種文書データは、テキスト化→チャンキングという手順を経てベクトルデータベースに保存する



2 チャンキング

長いテキストデータを、チャンクと呼ばれる単位に分割

第1章 総則

第1条（目的）

本規程は、株式会社サンプル（以下「当社」という。）の社員が、業務上の必要により出張を行う際に支給される旅費、宿泊費、日当その他必要な経費の支給基準及び手続について定めることにより、適正かつ公平な費用支出を図り、もって社員の職務遂行を円滑にすることを目的とする。

第2条（適用範囲）

本規程は、当社に在籍するすべての正社員、契約社員、嘱託社員に適用する。ただし、出張に関する特別な契約または別途規程が存在する場合は、当該契約または規程を優先するものとする。

チャンク1

第3条（出張の定義）

本規程における「出張」とは、当社の業務を遂行する目的で、通常の勤務場所を離れて一定の距離を越える場所に赴き、かつ一定の時間を要する移動を伴う場合を指す。これには、社外での打ち合わせ、現地調査、顧客訪問、研修、会議等を含むものとし、当該業務に関連して移動した全行程を含めて出張期間とみなす。

第4条（出張命令）

出張を行う社員は、事前に所定の出張申請書を提出し、所属長または上位管理職の承認を得なければならない。出張命令は、口頭によることなく、原則として書面または社内システムにおける承認をもって行うものとする。緊急の場合を除き、無断での出張は認めない。

チャンク2

第2章 旅費等の支給

第5条（旅費の区分）

旅費の支給対象となる費用は、以下のとおりとする。

- 一 交通費：鉄道、航空機、バス、タクシー、船舶等の運賃及び関連手数料
- 二 宿泊費：ホテル、旅館その他の宿泊施設に係る費用
- 三 日当：出張に伴い発生する雑費の補填として支給する定額手当
- 四 通信費・資料代等：出張先で業務上必要となる通信、資料、消耗品等の実費
- 五 その他会社が業務上必要と認めた費用

チャンク3

3

チャンクを埋め込み
ベクトルDBに保存

文書中の図表はどうする？

文書の中には、図表が掲載されているものがあり、うまく活用すると検索精度が上がる
ビジョンモデルを利用して、図表の内容をテキスト化することで容易に活用できる

複雑な図表を含むスライドでも、単にテキストを抽出するだけでなく、ビジョンを使って図表の細かい内容もテキストに起こす

RAG（Retrieval-Augmented Generation）への注目

2023年中旬から、急激にRAGへの注目度が上昇
多くのRAGソリューションが市場に投入され、データとの連携が容易になりつつある

Google検索ワード「RAG」のインタレストの変遷



NTTデータのRAGソリューション

2023年6月末、データ連携機能をもった
LITRON® Generative Assistantリリース



社内プロジェクトとしてRAG構築のPoCを進めている企業も増えてきたが...

© 2025 NTT DATA Mathematical Systems Inc.

11

NTT DATA 株式会社NTTデータ数理システム

RAG（Retrieval-Augmented Generation）への注目

2023年中旬から、急激にRAGへの注目度が上昇

多くのRAGソリューションが市場に投入され、データとの連携が容易になりつつある

Google検索ワード「RAG」のインタレストの変遷

2023年中旬から急速にインタレストが上昇（グラフ有）

グラフ：2020/4/5から2024/4/5までの期間

Y軸：0から100

2023年中頃に数値が20～30から80～100へ急激に上昇し、それ以降高い状態を維持
（100が最大値、変動有）

NTTデータのRAGソリューション

2023年6月末、データ連携機能をもったLITRON® Generative Assistantリリース

画面のサンプルイメージ

LITRON Generative Assistant 利用イメージ・出典表示

ユーザー「〇〇について教えてください？」

アシスタントが利用元や出典を示して回答

社内プロジェクトとしてRAG構築のPoCを進めている企業も増えてきたが...

© 2025 NTT DATA Mathematical Systems Inc.

NTT DATA 株式会社NTTデータ数理システム

※株式会社NTTデータ数理システムがChatGPT 4.1を利用して生成した結果を含みます。

RAGの仕組みのまとめ

RAGとは

もとなる
文書データ



に基づいて、

ユーザーの元の質問

に対する

最終回答

を得るために、

まず1度ベクトルデータベースの構築を行って類似検索システムを構築したあと、
質問ごとに類似検索・文書に基づく回答生成を行う

質問があるごとにやる 類似検索

RAGの仕組み①——類似検索システムの構築と活用 (1/2)

文書の類似検索とは

検索とは文書群の中から、「意味の近い」と似た文書を探す仕組み
キーワード検索よりも柔軟に、質問文と意味の距離が小さい＝関連しそうな文書を検索可能

質問文 「日本の経済の発展は？」

データベース内の文書

類似検索により
意味が最も
近いものを返す

0.835
0.845
0.815
0.825

目的は、類似検索によって、類似した文書を探し出すこと。類似した文書は、類似した意味を持つ。

目的は、類似検索によって、類似した文書を探し出すこと。類似した文書は、類似した意味を持つ。

質問があるごとにやる 文書に基づく回答生成

生成AIへの指示文

ユーザーの元の質問

以下をヒントにしてください

質問に関連しそうなチャンク

ベクトルサーチにより
元の質問文と
類似するチャンクを検索

生成AIが
回答を作成

最終回答

RAG構築時に 1 回だけやる
(文書の更新に応じて、適宜やり直す)

ベクトルデータベースの構築

もとなる
文書データ



テキスト抽出
OCR

図表抽出

テキスト

図表

クリーニング

ビジョン

テキストデータ

txt

チャンキング

チャンク

文章埋め込み
ベクトルDBへの投入

ベクトルDB

文書中の図表はどうする？

文書の中には、図表が埋め込まれているものがあり、うまく活用すると検索精度が上がる
ビジョンモデルを利用して、図表の内容をテキスト化することで容易に活用できる

図表の内容をテキスト化することで、検索精度が上がる

RAGの仕組み②——類似検索システムの構築と活用 (2/2)

文書データからベクトルデータベースを構築する

各種文書データは、テキスト化→チャンキングという手順を経てベクトルデータベースに保存する

チャンキング
長いテキストデータを、チャンクと呼ばれる単位に分割

文章埋め込み
チャンクを埋め込みベクトルに変換

【参考】文章埋め込み (sentence embedding)

文章埋め込み = 「文章」を「ベクトル」(数値の列)に変換する操作
ただし、意味の近い文章同士は、埋め込み後も距離が「近い」ベクトルになる

「文庫」の世界

「ベクトル」の世界

意味の近い文章

意味の近いベクトル

3 | RAG精度向上への道

——つまずきやすい課題と、現場で使える解決策

【再掲】RAG改善の考え方——全体像

精度・コストが常に優れた「万能な方法」があるわけではない
入出力の観察に基づき**仮説を立てて**精度向上施策を立てる、**地道な方法こそが王道**！

考え方のステップ

入出力の観察

重要！

仮説思考

改善策立案

ステップの詳細

- RAGの入出力や中間アウトプットを観察し、どういうケースで精度が低くなるのか特定する
- 現在の出力とあるべき出力の差を観察・言語化し、不十分な点を特定していく

- なぜ現在の出力が不十分なのか、その原因について仮説を立てる

- 仮説に基づき、どうしたら精度が向上するのか、施策を立案する

考え方の例



こういうタイプの質問の場合に精度が低くなっているな… 本来はこれぐらいのクオリティまで欲しいのに、現状はこれぐらいのクオリティしか出ていない。〇〇という観点が入っていないな。



本来のクオリティの出力を得るためには、この情報を基にこう考えなきゃいけないはず！でも、このRAGでは別の情報を基に、こういう別の考え方をしているから、「精度が低い」のでは？



この部分を直したら…、確かに精度が向上しそうだ！だから、RAGの仕組みをこのように変えたら、このタイプの質問の出力は改善できそう！

良い改善策を立案するためには、「的を射た仮説」を立てることが重要

良い改善策立案につながる仮説を出すためには…

1. よくある「落とし穴」の傾向と対策を知ること

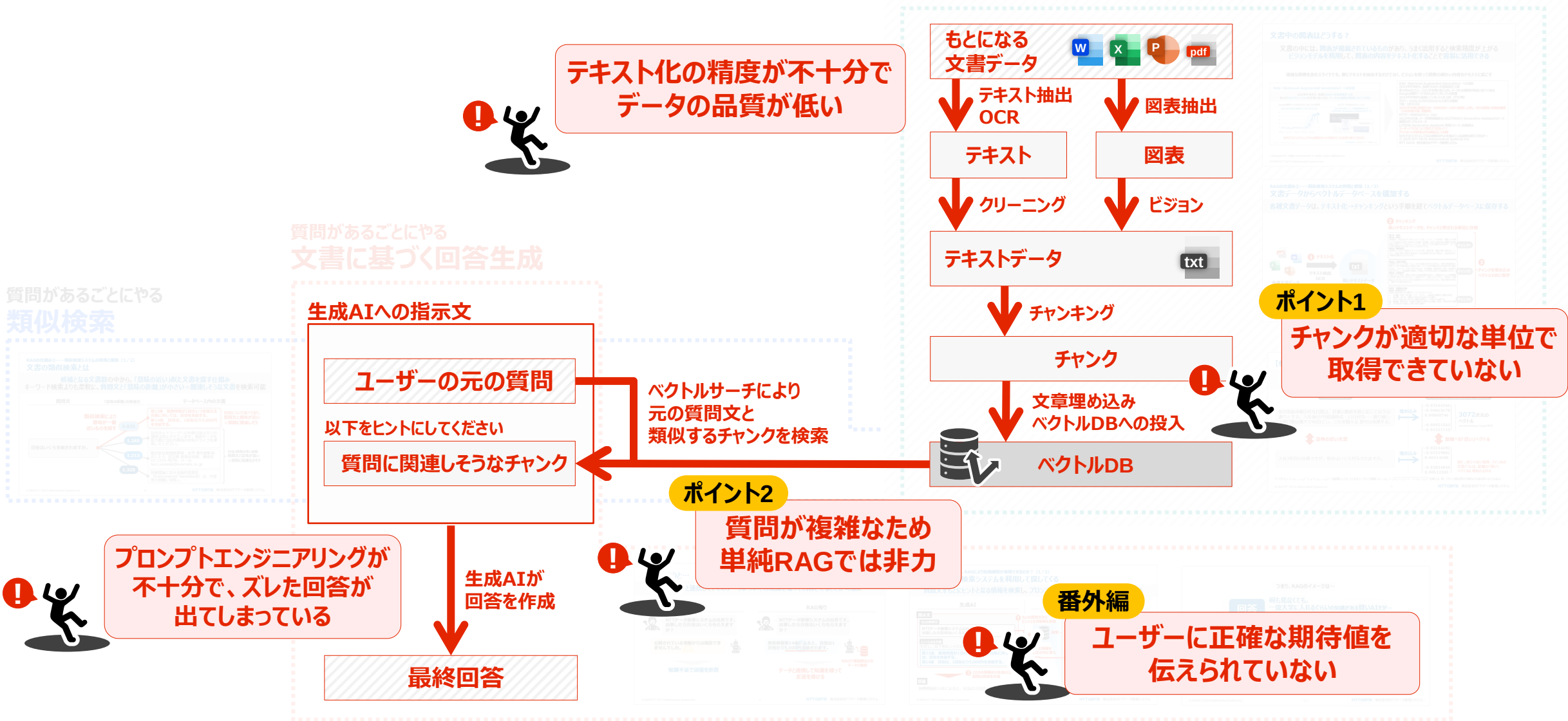


そのうえで…

2. 入出力・中間出力物を、 実際の業務知識に照らし合わせながら、よく観察すること

RAG各所に潜む、代表的な「落とし穴」

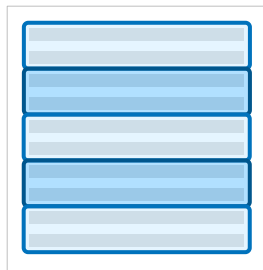
※他にもいろいろな「落とし穴」があるが、よくあるものを掲載



チャンクが適切な単位で分割できていない



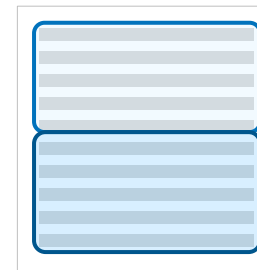
チャンクの分割単位は、細かすぎても粗すぎても精度低下を招く



チャンクを
細かく分ける



チャンクを
粗く分ける



検索の
ヒットしやすさ

各チャンクの意味が単純になるので
類似検索で目的のチャンクが
ヒットしやすくなりやすい



各チャンクの意味が複雑になるので
類似検索で目的のチャンクが
ヒットしにくくなりやすい

ヒットした情報で
生成した結果の
正確さ

各チャンクの中に含まれる文脈が少なくなり
判断の材料が減るので
結果が不正確になりやすい



各チャンクの中に含まれる文脈が増え
判断の材料が増えるので
結果が正確になりやすい

バランスの良いチャンクサイズを見つけることが重要（複数試して良いものを見つける）

【参考】より高度なチャンキング手法

実務的には、発展的なチャンキング手法を用いてRAG全体の精度を改善することが多い
文書の量や許容される計算コスト等、様々なバランスを踏まえて利用する手法を決定する

手法の例	説明
ドキュメントベースチャンキング 階層的チャンキング	クエリの複雑さに応じた検索を可能にするため、ドキュメントを複数の粒度（要約、セクション、段落など）でインデックス化する
セマンティックチャンキング	固定サイズではなく、文の意味的類似性（埋め込み間のコサイン距離など）に基づいてテキストを分割し、一貫した情報単位を作成する。計算量は増えるが、関連性が向上する
生成AIベースのチャンキング エージェントックチャンキング	LLM等の生成AIを利用して、意味的に切れ目の良いステートメントをチャンクとして識別・作成。 計算コストが大きいですが、最も高品質で適応的なチャンクを生成可能
文脈ウィンドウ検索 （チャンク利用時の後処理）	検索は細かいチャンクで行い、生成の際にはそのチャンクの周囲のチャンクもあわせて付加することにより、より多くのコンテキストを取り込む
リランキング （チャンク利用時の後処理）	初期検索後、より高度なモデル（Cohere Reranker等）を使用して、検索されたチャンクをクエリとの関連性に基づいて再評価・並べ替える

参考
Shivansh Kaushik: **Advanced Text Retrieval with Elasticsearch & LlamaIndex: Sentence Window Retrieval** medium. Jan 11, 2024 <https://medium.com/primastat/advanced-text-retrieval-with-elasticsearch-llamaindex-sentence-window-retrieval-cb5ea720aa44>
Anurag Mishra: **Five Levels of Chunking Strategies in RAG**. medium. Jan 14, 2024 https://medium.com/@anuragmishra_27746/five-levels-of-chunking-strategies-in-rag-notes-from-gregs-video-7b735895694d

そもそも、類似検索に基づく単純なRAGだけでは難しい質問も…



課題によっては、そもそも類似検索に基づく単純なRAGでは解決が難しいケースも

難しいパターン1

文書全体を俯瞰しないと答えられない質問

NG例

〇〇に関連する資料を**すべて**要約してください



類似検索に基づく単純なRAGでは、**生成AIが検索システムから渡されたチャンク**だけしか読み取れないので、「**全体**」を俯瞰するような質問回答は苦手

難しいパターン2

文書のメタデータを利用しないと答えられない問題

NG例

いちばん**最近更新された**ファイルはどれ？



そもそも単純な類似検索では、**文書の更新日時**などのメタ情報を考慮しきないので、**回答にメタ情報が必要な質問**をされても回答ができない

より高度なRAGアーキテクチャの採用を検討したり
問題によっては、新しいアーキテクチャをデザインする場合も
具体例は、この後の事例でご紹介します

【参考】「実は」正答のためにはメタデータが必要だったケース…

〇〇部の〇〇課の課長の下の
名前ってなんだっけ？



体制表_20250401.pdf

体制表_20250301.pdf

体制表_20240701.pdf

体制表_20240401.pdf

おそらくこの質問は、現在の課長について聞いているので
「最新の体制表」を見て回答する必要があるが…



データベース内に
更新履歴が別ファイルとして格納されていた場合、
単純な類似検索ではどの体制表が最新のもののなか
分らない

うっかり大昔の課長の名前を
答えてしまうかも？

全体を俯瞰するメタデータをうまく扱うためには？

文書全体を俯瞰する「大域的な文脈」情報や、メタデータをうまく扱うためにはベクトルデータベース以外のデータベースを併用して検索システムを構築することが重要

ベクトルDB



データをベクトル形式で格納
類似検索

リレーショナルDB



データをテーブル形式で格納
構造化データの管理

グラフDB



データをグラフ形式で格納
複雑な関係の解析

ドキュメントDB

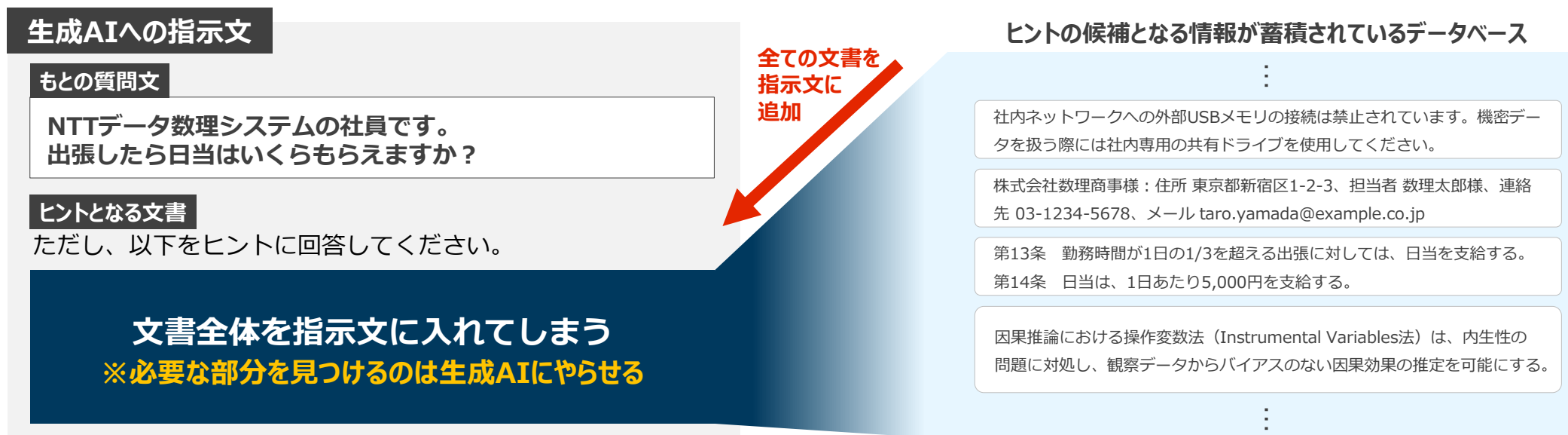


データをJSON形式で格納
柔軟なデータ構造の保存

これらも併用して、大域的な文脈やメタデータをうまく扱う

【参考】最近の動向ーロングコンテキストの利用

候補となる全文書を指示文に加え、回答に必要な部分を見出すのは生成AIにやらせる手も
AIモデルのロングコンテキスト化（入力長の増加）に伴い、実務的にも有力な手段に



- ・ 検索システムを構築せずに済むので、**全文書がモデルの入力長に収まる場合（目安として数十万文字程度）には有効な手段**
 - ・ 全体の俯瞰が必要な問題にも対応しやすい。
 - ・ その分トークンを多く消費するので、コスト効率は良くはない場合がある
 - ・ 精度は一般に出やすいが、中間忘却※（lost in the middle）などの問題もあり、RAGより精度が良いかどうかは場合によりけり
※生成AIが指示文の最初と末尾はよく覚えているが、中間については忘れたような出力を示す現象のこと
- ・ 一方で、**全文書の量が多ければ使えない手段**なので、その場合は**RAGのような検索システムとの連携が必要**になる

ユーザー部門とのコミュニケーションがうまくいかない



例えば、**ユーザー部門**に展開する段階になって、**用意したデータの範囲では答えられない想定外の質問を連発**されてしまう…

ユーザーヒアリング段階で、
ユーザーサイドに参照したい文書を聞き取りして
同意したように思えても…

…… **実際にユーザー部門に展開する段階で…**

チャットインターフェースは意外と
何でも聞いてしまうので、
**まったく開発サイドが想定していないような質問をされ、
AIが素っ頓狂な回答をしてしまうことで失望される**
→「使えない」という噂だけ独り歩きしてしまうことも…

RAG開発側部門

ユーザー側部門

普段あなたの部署ではどういう文書を
参照して作業されていますか？

いま〇〇フォルダに置いてある文書で
ほとんど全部だと思うんですよね…

では〇〇フォルダ以下で
RAG構築しますね

お、何でも聞けるのか。
他部署の業務について聞きたいんだよな

え、何にも答えてくれないじゃん！
じゃあ、何に使えるって言うのさ…

その部署のデータ
入ってないだけにな…

「正確な期待値」は、3段階で共有していく工夫が重要

ユーザーに正確な期待値を伝えるための仕掛けを
①PJ、②UI／システム仕様、③AI の3段階それぞれで考えることが重要

段階	期待値の伝え方の実践例
①PJ	・ PoC段階ではあくまでも「意見収集」「ユースケース収集」を目的に据え、いまは一部の文書のみで試験的に実施していることを強調する
②UI／システム仕様	・ 参照し得る文書の一覧が、UIからわかるようにする ・ 回答に当たり参照した文書のリファレンスを、UI上から確認できるようにする ・ 「こういう質問だったらうまく回答してくれる」テンプレート例を、UI上に表示する ・ AIはハルシネーションなどの問題があるということを、UI上で分かりやすく伝える ・ AIの思考プロセスやログなども、UI上から確認できるようにする ・ 参照する文書をユーザーが指定できるようにする
③AI	・ 関連度が低い文書が出てきたときに、「関連しそうな情報は見つかりませんでした、〇〇という情報ならば見つかったのでこちらに関する質問は回答できます」などと明示的に回答させるようにする（プロンプトエンジニアリングによる）

4 | ケーススタディ

——「教えて！Alkano先生」で実現した精度改善の舞台裏

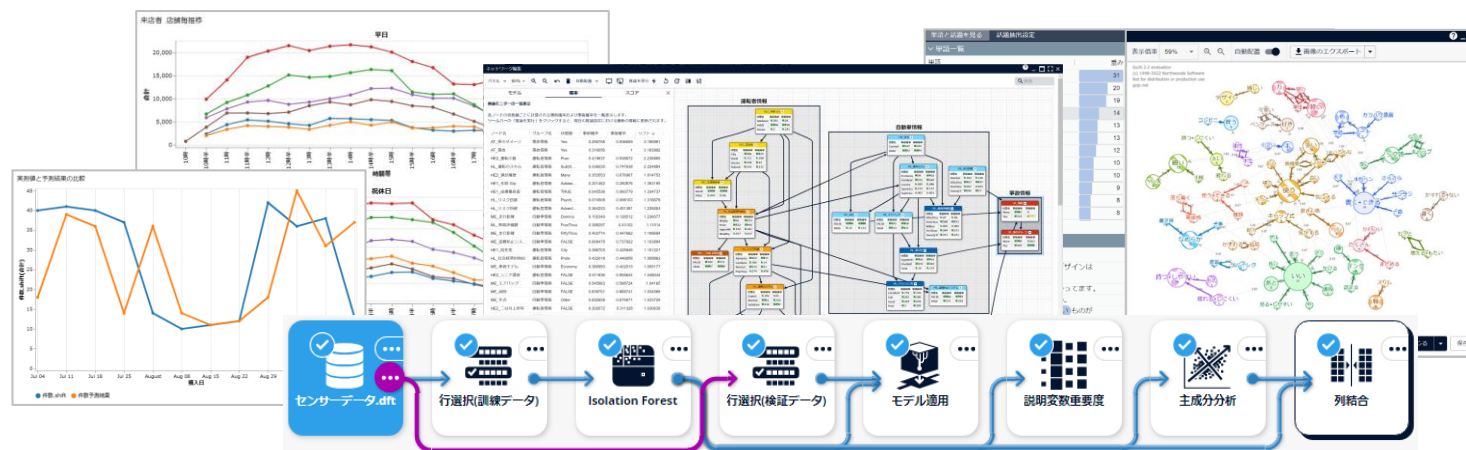
データ分析プラットフォームAlkanoと、「教えて！Alkano先生」のご紹介

背景：データ分析プラットフォームAlkanoのご紹介

プログラミング無しで高度なデータ分析を可能にする、ノーコードデータ分析プラットフォーム
20種類超のデータ前処理機能 + 70種類超のデータ分析機能 であらゆるデータを分析



- ◇ 統計・機械学習×ベイジアンネットワーク×テキストマイニングと多面的な分析
- ◇ データ前処理、データ分析、モデル構築・精度評価の機能を標準搭載
【20種類超の前処理機能 + 70種超の分析機能（一例）】
 - ・ 統計解析、機械学習、ディープラーニング、Auto Modeling、...
 - ・ ベイジアンネットワーク関連（因果関係の可視化）：構造学習、確率推論、感度分析、...
 - ・ テキストマイニング関連：形態素・構文解析、特徴語抽出、共起ネットワーク、...
- ◇ PythonやRのコード拡張、Web API、ODBCによるシステム連携



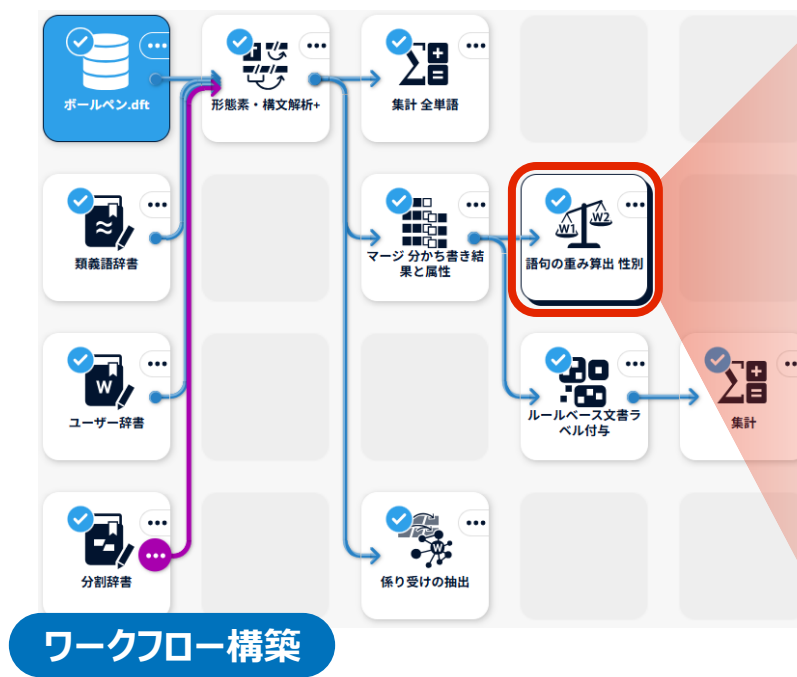
Alkanoでのデータ分析イメージ

多彩な分析アイコンを組み合わせて、高い自由度でワークフローを構築・設定

多様な分析ニーズに答える
多彩な分析アイコン



組み合わせ



語句の重み算出 性別

列選択

列名	列型	対象語句列	キー列
RowID	整数	☐	☐
SntID	整数	☐	☐
TokenID	整数	☐	☐
form	文字列	☐	☐
lemma	カテゴリ	☐	☐
released	カテゴリ	☐	☐

抽出指標

☒ 補完類似度	☐ Fisherの直接確率
☐ χ^2 乗値	☐ Yates補正 χ^2 乗値
☐ Jaccard係数	☐ TF-IDF
☐ Dice係数	☐ cosine
☐ 相互情報量	

抽出キー

☐ キーを分割

実行して閉じる 保存

分析パラメータ設定

幅広い分析ニーズに合わせて、高い自由度でワークフローの構築が可能

2025年4月リリース新機能、「教えて！Alkano先生」

データ分析プラットフォームAlkano上で使える生成AIエージェント Alkanoを使ったデータ分析について質問すると、具体的な分析方法について教えてくれる

「教えて！Alkano先生」利用イメージ



アナリスト

現在、Alkano を使った欠損値処理の部分で悩んでいます。分析観点からどのような処理方法がベストでしょうか？



シニアアナリスト

欠損値は“その他”で置換して、欠損でない“その他”と区別するために欠損有無の情報を別途作成するのはどうでしょう。

@Agent Alkano でこの処理はどう行いますか？

ユーザーがAIエージェントに質問すると…

「欠損値を別のカテゴリーに置き換える」には、Alkano のクリーニング アイコンを使用することで簡単に対応できます。必要に応じて 列追加 アイコンを活用し…

AIエージェントによる回答がすぐに（数分程度で）得られる



生成AIエージェント
Alkano先生

実際の画面



※「教えて！Alkano先生」機能をご利用の際は、Microsoft社の「Azure OpenAI Service」のAPIを利用できる環境が必要です。

5 | まとめ

【再掲】RAG改善の考え方——全体像

精度・コストが常に優れた「万能な方法」があるわけではない
入出力の観察に基づき**仮説を立てて**精度向上施策を立てる、**地道な方法こそが王道！**

考え方のステップ

入出力の観察

重要！

仮説思考

改善策立案

ステップの詳細

- RAGの入出力や中間アウトプットを観察し、どういうケースで精度が低くなるのか特定する
- 現在の出力とあるべき出力の差を観察・言語化し、不十分な点を特定していく

- なぜ現在の出力が不十分なのか、その原因について仮説を立てる

- 仮説に基づき、どうしたら精度が向上するのか、施策を立案する

考え方の例



こういうタイプの質問の場合に精度が低くなっているな… 本来はこれぐらいのクオリティまで欲しいのに、現状はこれぐらいのクオリティしか出ていない。〇〇という観点が入っていないな。



本来のクオリティの出力を得るためには、この情報を基にこう考えなきゃいけないはず！でも、このRAGでは別の情報を基に、こういう別の考え方をしているから、「精度が低い」のでは？

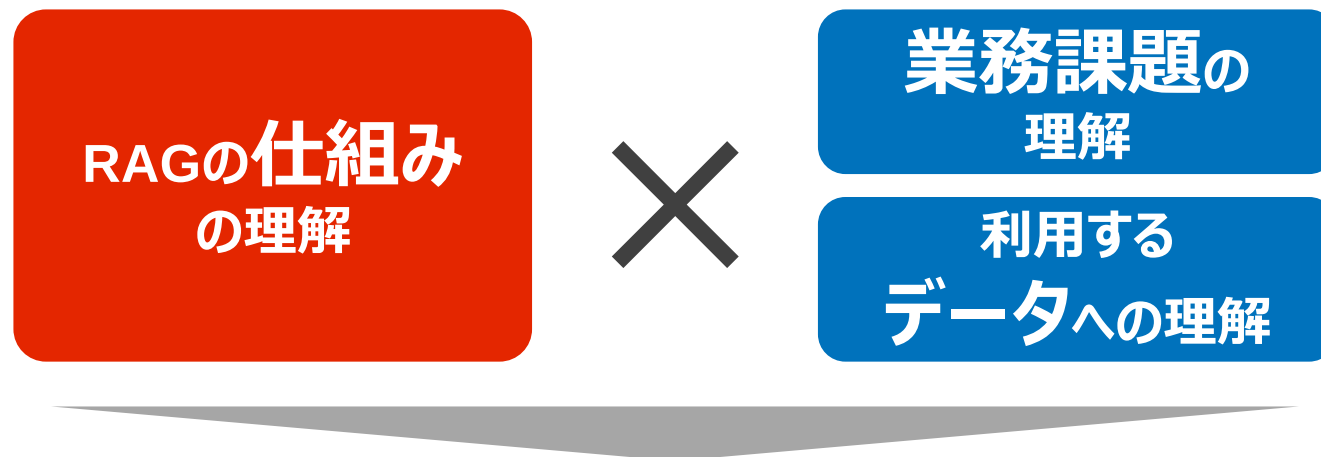


この部分を直したら…、確かに精度が向上しそうだ！だから、RAGの仕組みをこのように変えたら、このタイプの質問の出力は改善できそう！

良い改善策を立案するためには、「的を射た仮説」を立てることが重要

まとめ——RAGの精度改善をするには

RAGの仕組みを理解と**業務課題・データへの理解**の両方が大事！



技術領域のプロフェッショナルとして業務課題・データを理解し、貴社RAG構築PJにご参画させていただきます
NTTデータ数理システムにご相談ください！

当社の豊富な知見に基づき、貴社のRAG精度向上のご支援をさせていただきます！

ご支援
タスク例

- ・ 精度のボトルネック箇所の特定
- ・ より高度なRAGアーキテクチャの提案・実装・精度検証

【参考】より高度なRAGアーキテクチャの例

必要に応じて論文等も参照しながら、課題に応じたベストなRAGアーキテクチャを提案

当社が実際に提案したことのある技術の例（論文で提案されているものを組み合わせて実装）

ナレッジグラフを利用して、RAGの回答正確性向上 [1-3]

製品マニュアルや手順書などの文書群から、概念の関係を抽出
関係性をグラフで表現して蓄積することで、回答の正確性が向上

文書群 (文書A, B, C, ...)

抽出

数理システム

文書A

文書B

Auto Modeling

Alkano

線形回帰

文書C

形態素解析

has feature

made by

ナレッジグラフを参照した質問応答例

Q. Alkanoにはどのような機能がある？

A. Alkano には、AutoModeling [文書B]、線形回帰 [文書C] などの機能があります。

通常のRAGが苦手な「複数文書を正確に踏まえて答えるべき」質問でも、正確な回答を生成できる

応用先: 医療、法律など正確性や根拠が特に要求される分野

RDBなどの表データも、RAGの入力に [4]

入力の表の変換を、中間結果を作りながら繰り返すことで
高精度な応答を実現 (Chain-of-table)

列選択、ソート、カウントなどの基本操作を
生成AIが判断しながら繰り返す

表データ

中間データ

質問に関する部分のみ残す

例 膨大な気象・地象ログデータからの情報探索

震度計番号	観測点名称	緯度	経度	震度	最
0	47,835 日南市油津	31.5784	131.4081	5強	
1	47,829 都城市昌蒲原	31.7299	131.0618	5弱	
2	42,519 新富町上富田	32.0689	131.4574	5弱	
3	47,830 宮崎市錦島	31.9384	131.4142	5弱	
4	41,515 鹿屋市新栄町	31.376	130.8504	5弱	
5	47,827 鹿児島市東元	31.5552	130.5486	4	
6	42,517 佐伯市瀬江瀬江浦	32.8009	131.9216	4	
7	42,512 芦北町芦北	32.299	130.4931	4	

震度が5弱の観測点を挙げてください

都城市昌蒲原, 新富町上富田, 宮崎市錦島, 鹿屋市新栄町

質問を理解したうえで、「震度」列を選択してフィルタするクエリを実行

応用先: POSデータなど、各種表データに基づくQAシステム

プログラムのソースコードを自動解析、RAGに組み込む [5]

レポジトリ全体を解析、コールグラフに基づきドキュメントを生成
ソースコードを読まなければ答えられない難しいQAも可能に

repository

directory

fileA

function

ドキュメント

fileB

class

ドキュメント

例 ソースコードに基づきトラブルシュートするQAチャット

強化学習モデル作成時に "reward is not implemented." というエラーが発生しました。どのように解決すればよいですか。

"reward is not implemented." というエラーは、RLModel クラスのサブクラスで reward メソッドが実装されていないために発生しています。RLModel クラスは抽象クラスのようなものであり、具体的な強化学習モデルを作成するには、以下のメソッドをオーバーライドする必要があります:

```
1. reward(self, simulator, item, prev_time)
2. observation(self, simulator, item)
3. action(self, simulator, obs, act, item)
```

クラス構造をコードから読み取り
エラー原因を特定

これらのメソッドを具体的な強化学習モデルに応じて実装する必要があります。以下に、これらのメソッドを実装する例を示します。

応用先: ソフトウェア開発・サポート業務の効率化

参考文献

[1] Y. Gao *et al.*, Retrieval-Augmented Generation for Large Language Models: A Survey, arXiv:2312.10997 (2023)

[2] Z. Rackauckas, RAG-Fusion: a New Take on Retrieval-Augmented Generation, arXiv:2402.03367 (2024)

[3] Y. Lu *et al.*, Unified Structure Generation for Universal Information Extraction. ACL-long papers, 5755–5772 (2022)

[4] Z. Wang *et al.*, Chain-of-Table: Evolving Tables in the Reasoning Chain for Table Understanding, arXiv:2401.04398 (2024)

[5] Q. Luo *et al.*, RepoAgent: An LLM-Powered Open-Source Framework for Repository-level Code Documentation Generation, arXiv:2402.16667 (2024)

プロジェクトを始めるまで

必要に応じて**無料データアセスメント**を行い、**プロジェクトの進め方をご提案**いたします

お打ち合わせ（要件整理）

弊社スタッフが貴社の課題を伺い、プロジェクト化に向けた要件をまとめます
「実現したいこと」「現状・問題」「課題」「課題解決見込み」を整理します

おススメ

無料アセスメント(分析) ※1

活用したいデータを弊社に送付いただける場合※2、
アセスメント（事前分析）を行い、課題の洗い出し・実現方式の検討をいたします

ご提案・概算お見積り

プロジェクトの進め方について、ご提案をいたします
ご提案書の中に、アセスメント（分析）の結果を盛り込むことも可能です
何度かご提案の機会をいただき、要件をすり合わせて実現イメージを固めていきます

お見積り

プロジェクトの実現イメージが固まりましたら、正式にお見積りをいたします
弊社のご提案が課題解決のお役に立てそうか、ご検討ください

プロジェクト開始

ご発注をいただけたら、プロジェクトを開始いたします
弊社スタッフが、お客様の課題解決に向けて真摯に伴走いたします！

※1：無償の範囲で対応させていただきます。 ※2：NDA締結のうえ、データのやり取りをさせていただくことも可能です。



NTT data

株式会社NTTデータ数理システム

A | 補足

【参考】文章埋め込み (sentence embedding)

文章埋め込み = 「文章」を「ベクトル」(数値の列) に変換する操作
ただし、意味の近い文章どうしは、埋め込み後も距離※の「近い」ベクトルになる

「文章」の世界

年次有給休暇の付与日数は、社員の勤続年数に応じて以下の通りとする。入社後6か月経過時点：12日付与……繰り越し上限日数は最大で48日とし、これを超える 部分は消滅する。



意味の近い文章

入社3年目の社員ですが、有休はいくら付与されますか。

「ベクトル」(数値の列) の世界

埋め込み


$$\begin{bmatrix} -0.03344366 \\ -0.00865678 \\ 0.00080746 \\ \vdots \\ -0.00921563 \\ -0.01351115 \end{bmatrix}$$

3072次元の
ベクトル

text-embedding-3-largeの場合



距離※の「近い」ベクトル

埋め込み


$$\begin{bmatrix} -0.02244202 \\ -0.02339881 \\ 0.00314588 \\ \vdots \\ -0.01834844 \\ 0.00512303 \end{bmatrix}$$

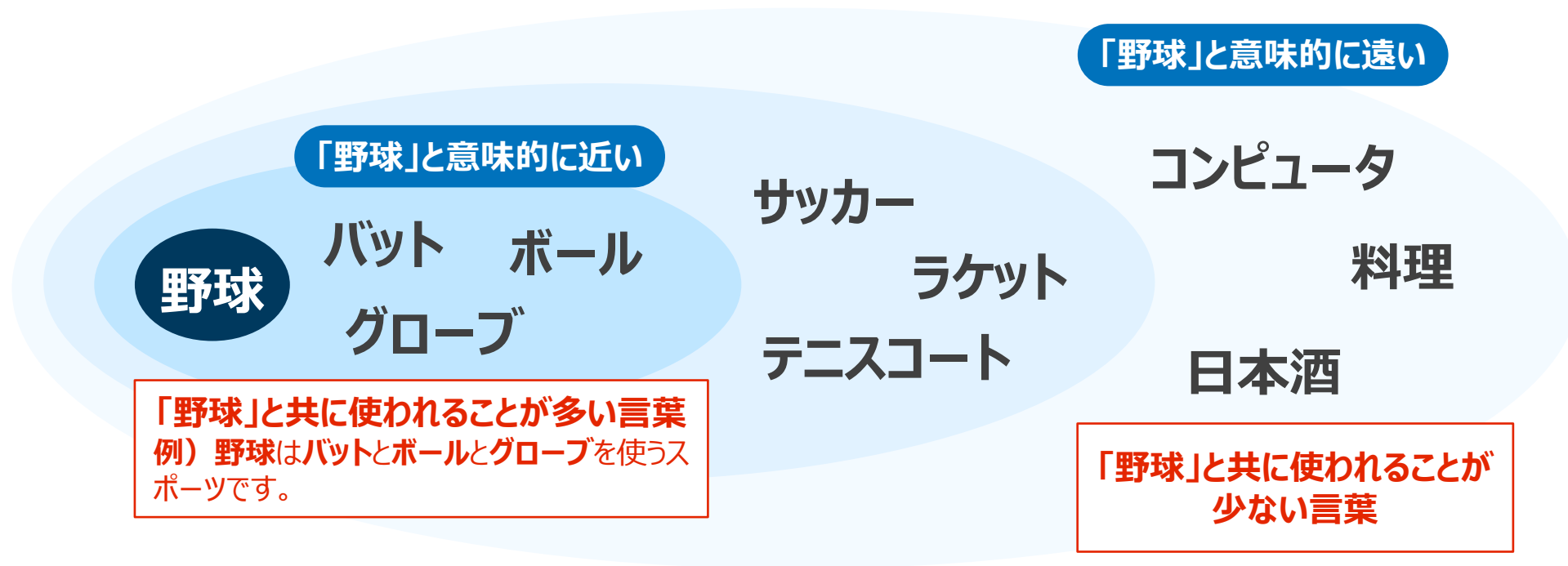
逆に、似ていない意味・ジャンルの
文章どうしは、距離の「遠い」
ベクトルに埋め込まれる

※ベクトル $\mathbf{x} = (x_1, \dots, x_d)^T$ と $\mathbf{y} = (y_1, \dots, y_d)^T$ の距離としては、たとえばユークリッド距離 $\|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{(x_1 - y_1)^2 + \dots + (x_d - y_d)^2}$ を用いる。他、コサイン類似度や内積などの値を用いることもある。

【参考】なぜ言葉を「ベクトル」に変換できるの？

似た意味の言葉・文は、似たような使われ方をする（分布仮説）

膨大なコーパスから言葉の「使われ方」のパターンを**数値化**することで、意味を捉えた「ベクトル」に



言葉の使われ方のパターンを、統計手法や深層学習等の機械学習手法で捉えて数値化している

※最近の「文章埋め込み」では、単にここで説明している「言葉どうしの共起関係」だけではなく、GPTやBERT等の注意機構を備えたモデルで文脈を含めた学習をすることがキモになっています。
※あくまでもイメージの説明です。詳細は専門書を参照してください。