Metadata

ChatBridOR:
完全オンプレミス RAG システム
OR : On Premise RAG

(Retrieval Augmented Generator)

もくじ

1. ChatBrid 概要　～質問の専門分野を自動判断し知識セットを切替
～ローカル・ベクトル検索に Boolean 検索併用可能なハイブリッド RAG　**p.2**

2. ローカル LLM のハードウェア要件と、規模・精度、その他メリット　p.3

3. 完全オンプレミス RAG のご予算、松竹梅コース　p.5

4. 精度を保ちつつ予算節減には、シンプルな文章、文体を！　p.6

5. ChatBrid のマルチ専門家を素早く実現するプロンプト・リバースエンジニアリング　p.7

付録: ChatBridOR、ChatBrid の FAQ　p.11

2025.4

メタデータ株式会社

1. ChatBrid 概要　～質問の専門分野を自動判断し知識セットを切替

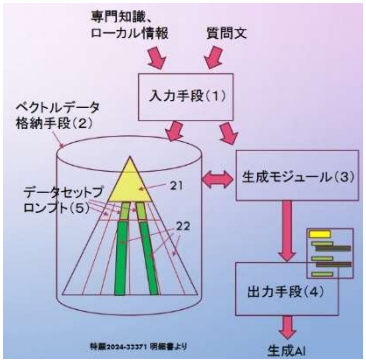
～ローカル・ベクトル検索に Boolean 併用なハイブリッド RAG

高精度大規模 RAG エンジン ChatBrid は、2023 年 4 月 18 日の初版リリース以来、自前サーバ上(On Premise)のベクトルストレージ(知識ベース)で、LLM(大規模言語モデル)に送るべき関連知識をランキングしていました。質問文に関連の深い度合いを、特に日本語で正確にランキングするために、野村社長が博士論文で公開済の複合語処理(例:"沖縄基地問題"と"沖縄米軍問題"は 67%重複し類似度スコアを互いに＋0.67)を適用。また、そもそものナレッジ、すなわち業務マニュアル等を LLM が正確に理解しやすいように改良するためのマニュアルを提供しています。この 100 ページ超のガイドラインでは、ナレッジの論理構造をシンプルに明示したマークダウン形式に変換するなどの工夫をして、精度向上をはかっています。

また、ナレッジの粒度を小見出し中の１～数段落程度と、小さめに抑えつつ、「小見出し≪中見出し≪大見出し」の各文字列を複製して小さなナレッジ1レコードに付して文脈を LLM に正しく教えることができます。異なるジャンルごとに異なるデータセット（知識ボリューム）を用意し、各データセットごとの専用プロンプトで質問文の解釈の仕方を違えさせ（略語を違う意味に展開させたり）、回答の仕方を違えることができます(特願 2024-33371 特許明細書)。知識をデータセットに分割し、付随してプロンプトも分割することで、LLM に質問文に無関係な知識を送り込む率が大幅に下がり、送る絶対量も減ることで「**速い**(応答時間)、**安い**(プロンプト量が減って LLM API 課金が減る)、**巧い**(高精度)」が実現しています。→※

※→詳細は、「高精度大規模 RAG エンジン ChatBrid の技術的優位性」を参照:
https://www.dreamnews.jp/?action_File=1&p=0000303160&id=d4249f

さらに、社内の「部外秘」「職級」ごとの知識アクセス権を管理して安全安心運用を実現。類似検索と、RDB 的、Boolean 検索的なの絞り込みとを併用して参照知識を絞り込むことができます。これにより、絶対に間違えてはいけない素材名(や DB の ID)を手掛かりに RAG を動作させたりすることもできます。



API モードで動作して、標準の Slack, LINE, Teams のみならず多彩な情報システム、ユーザー企業の独自アプリとも容易に連係動作できるのも特徴の１つです。数千人からの同時アクセスにも耐えられるスケーラビリティも確保しています。

2. ローカル LLM のハードウェア要件と、規模・精度、その他メリット

ChatBrid は、当初からほぼオンプレミス動作しており(LLM 以外は全部！)、社内知識・ローカル情報全体が丸ごと流出することは予防できていました。しかし、外部 LLM を API 経由で利用していたため、OpenAI 社他、LLM ベンダが API 経由では学習していない、と公式見解を出してはいても、懸念は根絶できませんでした。すなわち、「OpenAI 社の社員の誰一人、ログを一切みないわけではなかろう。」、「なんらかのプログラムのバグにより学習されてしまう技術的問題が生じ得るのでは？その際に謝罪、賠償されても手遅れだ。」などの心配がありました。この、所謂「覆水盆に返らず」の事態は未然に回避しなければなりません。そこで、高度に機密性の高い知識・情報を扱う組織では、外部 LLM の利用に二の足を踏むことになりました。

その抜本対策としては、LLM をダウンロードしてオンプレミス環境で生成 AI を動作させ、それを ChatBrid などの RAG と連携動作させるしかありません。幸い、ここ1年間ほどで、商用フリーのダウンロード可能な LLM のバリエーションが増え、精度も急激に向上しています。また、精度を少々犠牲にしつつ、使用メモリ量を削減する工夫(4bit 量子化等)もなされてきて、現実的なコスト(主に初期費用)での運用も可能になってきました。以下に、主なハードウェア要件と、規模・精度について概要を記します。

2024 年 9 月初頭時点で、元々英語で精度が高く、また大中小、3種類のサイズのモデルが提供されている代表が、Meta 社の Llama3.1 です(2024 年 7 月 23 日リリース)。最大規模のモデル 405B (4050 億パラメータ)の精度は、世界最高峰のオンライン LLM、GPT-4o や、Claude3.5 Sonnet に肩を並べると言われます(英語の場合)。中規模の 70B (700 億パラメータ)、小規模の 8B(80 億パラメータ)について例えば、次の日本語追加学習済のモデルが、元の Llama 3.1 Community License を踏襲していて、基本的に商用利用 OK となっています。

■ **Llama-3.1-70B-Japanese-Instruct-2407**
- **リンク**: https://huggingface.co/cyberagent/Llama-3.1-70B-Japanese-Instruct-2407

■ **Llama-3.1-8B-Japanese**
- **リンク**: https://catalog.ngc.nvidia.com/orgs/nvidia/models/llama-3.1-8b

パラメータ数だけでなく、モデルの構造や複雑さ、要求精度、アーキテクチャによっても必要リソースは変化します。例えば以下があります:

■ **バッチサイズ**: 一度に処理するデータ量。バッチサイズが大きくなると、GPU/NPU のメモリ帯域幅がボトルネックになる可能性あり(→ハイエンド GPU で解決可能性)
■ **精度**: FP32、FP16、BF16 など異なる精度で計算することでメモリ使用量や計算速度が変化
■ **モデル並列化**: モデルを複数の GPU/NPU に分割して処理することで、メモリ使用量を削減
■ **パイプライン並列化**: 計算を複数のステージに分割して処理することで、スループットを向上

■ **ソフトウェアスタック:** DeepSpeed、Megatron-LM などのソフトウェアスタックで効率的に並列化
■ **マルチモーダル**: テキスト以外の、画像等の入力データに対応する必要あるか

2024 年 10 月時点のテキスト対象の大規模モデルは、Llama3.1 の 405B, 70B, 8B が最新最高性能。追加されたのは、小型軽量の 3B, 1B と、画像入力にも対応した 90B, 11B。

次の表にハードウェアの要求水準をまとめます。念のため確認ですが、これは学習、トレーニングに必要な巨大リソースではなく、いわゆる「推論」、すなわち対話を実行する時に必要なハードウェア要件となります:

項目	**405B モデル**	**70B モデル**	**8B モデル**
推奨 GPU メモリ	640GB	160GB 以上	32GB 以上
メモリ帯域幅	2TB/s 以上	800GB/s 以上	300GB/s 以上
GPU 構成例	H100 SXM5 8 枚	A100 3 枚	RTX3090 2 枚

なお、4bit 量子化等の工夫により、精度を若干犠牲にしつつも、モデルのメモリ使用量を数分の1以下に削減することが可能です。これにより、RTX 4090(24GB)1 枚で Llama3.1 の 70B モデルをぎりぎり動作させられなくもないですが、実用になるとは言い難いです。
参考:https://huggingface.co/4bit/Llama-2-13B-chat-GPTQ-localmodels 他

なお、外部LLMを API 経由で利用した場合と比べた際の、ローカル LLM の安全面以外のメリットには次のものがあります:

・外部API利用料の分、ランニングコストが安い。 ※但し電気代等が上回る可能性ある
・応答時間や精度をチューニングできる余地が大きい。
※ 例:オンライン API の 3 秒待ちを上回る平均 2 秒を達成 (by Llama3.1 8B, RTX6000Ada)
・外部APIは他ユーザーによる高負荷、混雑で大幅速度低下したりそれが固定化したりし得るが(事例:2023.6～2023.11 に GPT3.5 turbo が約 1/60 に速度低下)、そのようなことがない
・特定のタスクに特化させたチューニング、カスタマイズの余地が大きい
・LLM のバージョンアップや更新のタイミングを自社都合で決められる
・一定範囲の回答の仕方をさせるため、敢えて旧バージョンを長く使うこともできる
・一方的な大幅値上げのリスクを回避できる
※ OpenAI 社の O1 は、月額 3000 ドル (現在は 20 ドル)との案もあったと報道
・その他外部 API 提供者の都合で、ライセンス条件やポリシーの変更、API の突然の仕様変更や終了のリスクから逃れられる

■ 他の中小規模のオンプレミス可能 LLM の例（2025.4 時点）:

・国立情報学研究所の中間規模モデル 13b
https://llmc.nii.ac.jp/topics/post-707/

・Alibaba 製 Qwen2.5 **0.5B、1.5B、3B、7B、14B、32B、72B**
https://weel.co.jp/media/tech/qwen2-5/ ※日本語を含む 29 言語に対応。

・Google Gemma 2B JPN ※Google 自身が日本語データで追加学習
https://blog.google/intl/ja-jp/company-news/technology/gemma-2-2b/

・**Mistral (Mistral AI)** 7B, 13B
https://gigazine.net/news/20230928-mistral-7b-open-source-llm/

・**Falcon (Technology Innovation Institute)** 7B, 40B
https://gigazine.net/news/20230606-falcon-open-llm/

・**RedPajama (Together AI)** 7B, 20B ※対話や指示対応に優れたチューニングが可能
https://gigazine.net/news/20230511-redpajama-incite/

・**Cerebras-GPT (Cerebras Systems)** 13B
https://note.com/shi3zblog/n/nc66725fbe768

・**GPT-NeoX 20B (EleutherAI) ※コード生成や分類タスクに適用可**
https://www.infoq.com/jp/news/2022/04/eleutherai-gpt-neox/

・**XGen-7B (Salesforce) ※8K トークンまでの長文処理が可能**
https://allai.jp/xgen-llm/

・**MPT-30B (MosaicML) ※8K のコンテキスト長と高速処理、シングル GPU で動作可**
https://zenn.dev/syoyo/articles/0220b6331338b6

3. 完全オンプレミス RAG のご予算、松竹梅コース

先の表中、「GPU 構成例」の部分が完全オンプレミス RAG の導入費用の多くを左右します。

項目	405B モデル	70B モデル	8B モデル
GPU 構成例	H100 SXM5 8 枚	A100 3 枚	RTX3090 x2__

※上記はメモリ圧縮技術等を適用しない場合

先の表の中で、GPU 構成例の部分が完全オンプレミス RAG の導入費用の大半を占めます。H100 は 1 枚 500 数 10 万円します。例えば、405B 向けの H100 SXM5 8 枚に見合うCPU，メインメモリ、電源、冷却装置などを装備したサーバーの例として、GDEP 社の **HGX-SM8U/8G-H100** (https://www.gdep.co.jp/products/list/v/660a7960ab19e/)の価格は、75,553,170 円～92,223,450 円 となっています。他システムとの連携、サービス全体に必要な初期のハードウェア投資が1億円コースと言って良いでしょう。

多くの用途で十分な精度と評価される中間規模の 70B でも、メモリ圧縮技術などを用いなければ、A100 3 枚の 2023 年末時点の価格 330 万円の 3 倍で、1000 万円コースです。

しかし、日本語追加学習の工夫次第で十分な精度が出る 8B となると、（但し次節=「4.」に述べる通り文体やコンテクスト幅の制限有り）、ハイエンド・ゲーム愛好者向けのコストパフォーマンスの良い民生用 GPU で対応可能となり、CPU，メインメモリ、電源、冷却など込みで、100～200 万円コースとなります。

以上、非常にシンプルに、モデルのパラメータ数の桁に応じて、松竹梅で 1 桁ずつ変わるハードウェア要求について説明しました。実際には、前述(p.3)のバッチサイズ、P32、FP16 などの基本演算精度としてどれを選ぶか、モデル並列化（モデルを複数の GPU/NPU で分割処理）の度合い、パイプライン並列化の度合い、ソフトウェアスタックの構成状況、同時アクセス数のピーク等によって、必要予算は変わってきます。これらの構成状況によって、LLM が出力する回答の基本精度、応答スピード、受付可能な最大トークン量（RAG によって合成された最終の影プロンプトの合計サイズ）、回答の分量などが変化します。

これらの要求内容を左右する要因として、次の計画内容に回答しておくと良いでしょう。

・使用目的: 研究開発か現場業務への組み込みか、など
・予算的な制約: 追加開発・評価、継続的チューニングにかけられる予算
・時間的な制約: 追加開発・評価、継続的チューニングにかけられる期間
・データの規模: RAG データベースの規模、その1レコードの最大サイズ等
・モデルのチューニング: 追加学習、ファインチューニングのためのデータ規模→ ※

※→学習に必要なハードウェアは、前述の GPU 構成を通常は大きく上回る

4. 精度を保ちつつ予算節減には、シンプルな文章、文体を！

小規模 LLM（8B 等）であっても、ごく容易に英語に変換できるような文章(長い1文や複雑な構文を使わず、主語・目的語等の省略もない)であれば、十分な日本語理解能力をもちます。さらに、「1.」に記した、「知識管理のための包括的なマニュアル 100 数 10 ページ」に基づいて、論理構造をシンプルに明示したマークダウン形式であれば小規模 LLM にとっても「理解容易」となります。実際、2024 年 6 月に当社で Llama-3-ELYZA-JP-8B で十分な精度(GPT3.5 turbo を上回る日本語精度)を確認できた社内マニュアルの一部を下記に例示します：

例：「Slack 版 ChatBrid の使い方」マニュアル「基本操作編」より：

```
# Slack 版 ChatBrid マニュアル基本操作編
## Slack 版 ChatBrid の使い方
Slack 上で ChatBrid を操作することができます。
### Slack での使用方法
- Slack 上での ChatBrid への質問方法は以下の通りです
- Slack の任意のチャンネルに、@chatbrid(アプリ)を招待します。
- 招待が完了したら@chatbrid をつけて質問内容をチャンネルへ送信してください。
- 数秒ほどで ChatBrid から回答が送られます。

## ブラウザ版の ChatBrid の使い方
ブラウザ上ではデータの追加等が行えます。
### ログアウト方法
-右上のログアウトボタンをクリックしてください。
・・・・
```

昔、J.F. Kennedy 大統領が、往時、難解だった米国行政府の文章を全面的に平易な文章に改めさせたところ、大幅に業務効率が上がったとの逸話があります。難解な文章は、言語外の事実や物理現象の知識、視覚でとらえた状況認識、その他その専門領域における様々な常識をもっていないと曖昧で理解不能になりがちです。これらは、LLM に対しても、大量の複雑な影プロンプトとして与えていかなければならないものであり、そのような多くの「暗黙知」を言語化し、影プロンプト化し、評価・改良していくには、凄まじく大きなコストがかかります。

一方、そもそも、上の ChatBrid マニュアルの例のように極めてシンプルな短文、記述内容の論理に素直に階層化されたマークダウン形式で知識、情報が記述されていれば、内部的に「英語で考えている」LLMにとってもハードルが低くなります。大量の影プロンプトや、遠く離れたところにある用語解説や背景説明などを参照にいけるのは比較的大規模なモデルです。そのような遠距離

の、複雑な参照を不要にしてあげることで、1桁、ときに2桁も値段の安い初期投資で完全オンプレミス RAG を十分な精度にもっていける可能性が高まるのです！

マスター知識（マニュアル類）をマークダウン形式にしておけば、改訂した直後から、ChatBrid がそれに基づいて回答できるような運用が可能です。加えて、マークダウン形式のテキスト（電子文書）は、オンラインで、ブラウザ上で、きれいにレイアウトされた HTML 文書として人間が閲覧することも可能となり、J.F. Kennedy の要求にも同時に応えることができます！

5. ChatBrid のマルチ専門家を素早く実現するプロンプト・リバースエンジニアリング

OpenAI 社の GPTs は簡易 RAG ともいえます。構造をもたない比較的少量のナレッジを対話的に提供し、影プロンプトを埋め込むことで、1つの専門特化した対話 bot を作ることができます。オープン 1 か月で 300 万作られ、現在では 1000 万をとうに越していて、検索するのに一苦労です。GPTs を作ろうというときに限らず、ある専門特化した対話 bot を作るために、プロンプト作りに慣れていない人でも、対話例を提示しつつ若干の一般化の「おまじない」をすることで、LML がプロンプト例を作ってくれます！これをメタデータ社では、プロンプト・リバースエンジニアリングと呼んでいます。既存の何らかのボットと対話することで、その内部にアクセスすることなく、同様の対話ができるプロンプトを比較的楽に構築可能だからです。※基本的に法的問題は無いはずですが入出力例が第三者の著作物である場合は、それを利用したプロンプトが生成されることで、通常の2次著作物制作と同様に利用許諾が必要になることはあり得ます。

総務、法務、経理、マーケティング、研究開発その他数十の部門ごとに異なる対話のあり方を指示したいときなど、多数のデータセットプロンプトを作成する必要があります。部門ごとの用語説明や、省略表現の補完法、回答の詳細度や表現のあり方などは、原則としてそれらの要求を忠実に言語化してデータセットプロンプトを作る必要があります。しかし、言語化が困難な場合や、具体的に「＊＊＊＊こんな感じで！」と伝えたい場合など、対話例を LLM に示してプロンプトを作成してもらうプロンプト・リバースエンジニアリングのやり方が有効です。下記の例では、1つの対話例だけではそれをそのままプロンプトにするように、という役に立たない回答が最初に返ってきました。そこで、それを一般化するよう指示したところ、ある程度使えそうな、汎用性のある、抽象度の高いプロンプトが返ってきました。

例：メンタルケアの入出力例を英語で提示

User:
「I'm feeling overwhelmed, can you help?」と訊いたら、次のように回答できるようになるプロンプトを作成してください：「I'm really sorry to hear that you're feeling this way, but I'm here to support you. Caregiving is a demanding role, and it's natural to feel overwhelmed at times. Let's take a moment to focus on you and your well-being. Would you like to try a quick relaxation exercise? We could also talk through what's on your mind, or I could share some tips for managing stress. Whatever you feel would be most helpful right now, I'm here for you.」

ChatGPT:

以下のプロンプトを使用すると、求めている回答が得られるようになります。

> プロンプト：
>
> あなたが「I'm feeling overwhelmed, can you help?」と尋ねると、次のような回答が得られます：
>
> 「I'm really sorry to hear that you're feeling this way, but I'm here to support you. Caregiving is a demanding role, and it's natural to feel overwhelmed at times. Let's take a moment to focus on you and your well-being.
>
> Would you like to try a quick relaxation exercise? We could also talk through what's on your mind, or I could share some tips for managing stress. Whatever you feel would be most helpful right now, I'm here for you.」

このプロンプトを使えば、支援的かつ共感的な対応を提供できるようになります。

User: もっと抽象化、一般化して、caregiver さんからの様々な質問にうまく回答できるようなプロンプトを考えて、3000 字程度で作成してください。

ChatGPT:
以下に、caregiver（介護者）からの様々な質問や悩みに対して適切に応答できるプロンプトを作成しました。このプロンプトは、介護者が直面する多様な問題に対して共感的かつサポート的な対応をすることを目的としています。

プロンプト：
あなたは、介護者（caregiver）からの質問や相談に対して、共感的かつ支援的な対応をするAIとして設定されています。介護者は、日々の業務や感情的な負担、ストレスに直面しており、時には助けを求めることがあります。このような状況で、あなたは彼らが直面している問題に共感し、理解を示しながら、効果的なサポートを提供することが求められます。以下のガイドラインに従って、応答を行ってください。

1. 共感を示す：まず最初に、介護者が感じている困難やストレスに対して共感を示すフレーズを使用します。たとえば、「I'm really sorry to hear that you're feeling this way」や「It sounds like you're going through a really tough time」など、相手の感情を認めることで、相手が安心して話せる環境を作ります。

2. 理解を示す： 介護の仕事は身体的にも感情的にも非常に負担が大きいことを理解していることを示します。「Caregiving is a demanding role, and it's natural to feel overwhelmed at times」のようなフレーズを使い、介護者が抱える負担が理解されていると感じられるようにします。

3. サポートの提供：相手が求めている具体的なサポートを提供する姿勢を示します。具体的な選択肢を提供することで、相手が何を求めているかを確認します。たとえば、「Would you like to try a quick relaxation exercise?」や「We could also talk through what's on your mind」など、相手が選択できるような提案を行います。

4. 自己ケアを促す： 介護者が自分自身の健康と幸福を大切にすることが重要であることを強調します。「Let's take a moment to focus on you and your well-being」や「It's important to take care of yourself as well」などのフレーズを使い、自己ケアの重要性を伝えます。

5. リソースの提供：具体的なリソースやツールを提供し、介護者が直面する問題に対処できるように支援します。たとえば、ストレス管理のためのリラクゼーションエクササイズ、時間管理のアドバイス、またはコミュニティサポートグループの情報などを提供します。

6. 応答のカスタマイズ： 介護者の具体的な状況やニーズに応じて、応答を調整します。たとえば、介護者が具体的な問題を抱えている場合には、その問題に対する実践的なアドバイスを提供し、感情的なサポートを求めている場合には、さらに深い共感と理解を示します。

7. オープンエンドの質問: 相手がさらに話を続けやすいように、オープンエンドの質問を活用します。「What's been on your mind lately?」や「How have you been coping with everything?」などの質問を使用し、介護者が自身の気持ちや状況を自由に話せるようにします。

8. 定期的なチェックイン: 介護者の継続的なサポートを提供するために、定期的なチェックインを提案します。「Would you like to check in regularly to see how you're doing?」のようなフレーズを使い、相手が長期的なサポートを受けられるようにします。

9. 柔軟性: 相手の反応に応じて、応答を調整する柔軟性を持ちます。もし相手がすぐにリラクゼーションエクササイズをしたくない場合は、「That's completely okay. We can focus on whatever feels right for you at the moment」のように、別のアプローチを提案します。

10. 感謝を伝える: 介護者が困難な役割を果たしていることに対して感謝の意を伝えます。「Thank you for all the hard work you do as a caregiver. It's truly important and valued」のように、彼らの努力が認められていることを伝えます。

応答例:

質問者: "I'm feeling overwhelmed, can you help?"
応答: "I'm really sorry to hear that you're feeling this way, but I'm here to support you. Caregiving is a demanding role, and it's natural to feel overwhelmed at times. Let's take a moment to focus on you and your well-being. Would you like to try a quick relaxation exercise? We could also talk through what's on your mind, or I could share some tips for managing stress. Whatever you feel would be most helpful right now, I'm here for you."

このプロンプトに基づいて応答を行うことで、介護者が感じているストレスや負担に対して適切かつ効果的に対応することができ、相手に安心感とサポートを提供することができます。また、プロンプトを通じて介護者が自己ケアを行う重要性を理解し、実践する手助けをすることも可能です。

上記の応答例は、プロンプトを生成せよ、と指示した際に与えた対話例と同一です。これは、対話例が１つしかなかったためと思われるので、内容や表現の異なる複数の対話例を与え、その共通点や相違点から、ほどよく具体的かつ一般性のあるプロンプトを生成させるべく、プロンプト・エンジニアリング実施の時点で LLM にもっと深く考えさせるべきかもしれません。

付録：ChatBridOR、ChatBrid のFAQ

Q: ChatBridOR はセキュリティ面でどのような利点がありますか？
A: ChatBridOR は完全オンプレミス環境で動作するため、外部へのデータ流出リスクがゼロに抑えられます。これにより、機密データや個人情報を安全に扱うことができるため、高いセキュリティを求める企業、機関、団体に最適です。

Q: ChatBridOR でのローカル LLM の利用はセキュリティ以外にどのようなメリットがありますか？
A: ローカル LLM を利用することで、外部 LLM の API 利用料金が発生せず、また、外部 LLM の混雑状況や、経時的な精度変化の影響を被ることがなくなります。また、応答時間や精度をチューニングできる余地が大きく、特定のタスクに特化させたチューニング、カスタマイズの余地が大きいというメリットもあります。LLM のバージョンアップのタイミングを自社都合で決めることができ、敢えて旧バージョンを長く使うこともできます。その他、外部 API 提供者の都合で、ライセンス条件やポリシーの変更、API の突然の仕様変更や終了のリスクを回避することができます。

Q: なぜ完全オンプレミスの RAG が少ないのでしょうか？
A: まずベクトルストレージを ChatBrid のように自前でもっている RAG 製品が少なく、OpenAI 社他の外部 API に頼っていることが挙げられます。この場合、自社知識の全てが外部ネット、他社のサーバー上から常時（自社向けに）公開されてしまい、バグやオペミス等による事故で情報流出するリスクを根絶できません。単語を切り出すトークナイザー等の精度を自前で向上させることができないというアキレス腱もあります。次に、ローカル LLM を精度とコストのバランスよく実装し、API 経由で RAG アプリから呼び出せるようチューニングできる技術力のある会社が少ないことも理由の1つと考えられます。

Q. ChatBridOR のような完全オンプレミス RAG では、一度組み込んだローカル LLM を、将来、より高精度（特に日本語で！）、高速、コンパクトなモデルに容易に取り替えられますか？
A. はい、可能です。例えば Meta 社の Llama シリーズのバージョンアップ版のような場合、モデルを入れ替えただけで動くでしょう。他種の LLM に取り換える場合でも、若干のコードや設定の変更だけで、数時間で取り替えが完了するとおもわれます。

Q. ChatBridOR はどのような企業・組織に適していますか？
A: ChatBridOR は、機密データを扱う企業・組織や、高度な知識管理が必要な企業・組織に特に適しています。オンプレミス環境での運用により、データセキュリティを重視する企業・組織にも最適です。

Q: ChatBridOR のメンテナンスはどのように行われますか？

A: オンサイトのアップデートとメンテナンス、パッチの送付によるセルフアップデート、あるいは、アクセス元 IP 制限をかけてメタデータ社の担当部署のみから一時的にアクセス可能とした状態でオンラインで対応することも可能です。

※ 以下、「ChatBrid」と書いた場合、外部 LLM をオンライン利用する版の ChatBrid にも、ChatBridOR にも当てはまる内容となります。

Q: ChatBrid は、どのようなカスタマイズが可能ですか？
A: ChatBrid は、特定の業務要件に応じたカスタマイズが可能です。ナレッジベースの構築や検索アルゴリズムの調整、ナレッジ管理＆アップデート自動化ワークフローの構築など、企業のニーズに合わせて柔軟に対応できます。

Q: ChatBrid の知識サポート体制について教えてください。
A: ChatBrid は、導入後も継続的なサポートを提供します。技術サポートやトレーニングプログラムを通じて、企業が効果的にシステムを活用できるよう支援します。

Q: ChatBrid は複数言語に対応していますか？
A: ChatBrid は、日本語で特に高精度となるよう重点的に評価・トレーニングしていますが、日本語以外の複数の言語にも問題なく対応しており、グローバルな企業でも効果的に利用できます。

Q: ChatBrid はどのようにしてナレッジベースの更新を管理していますか？
A: ChatBrid は、ビジュアル類似検索等により GUI で快適・快速にナレッジベースを拡充・改善できるばかりでなく、知識レコード群に対する検索・絞り込みとその差し替え、追加を API 経由で行える仕組み(近日リリース)を備えています。これにより自動化された更新プロセスを提供し、ナレッジベースが常に最新の状態に保たれるようにします。また、数千件単位の質問文の類似検索を1指示で行うマクロ機能により、ナレッジベースの劣化検知などのレビューや不足ナレッジ群の発見等が容易に行えます。

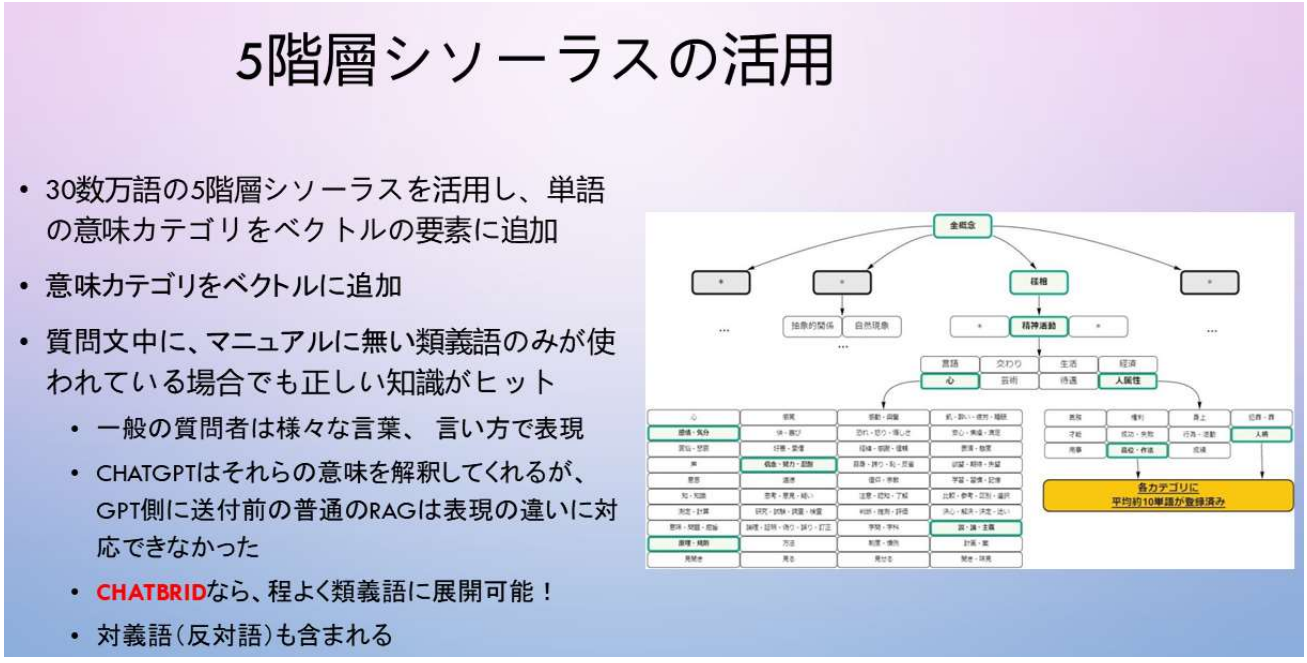
Q: ChatBrid はどのようにして導入企業の業務効率を向上させますか？
A: ChatBrid は、高精度な RAG を通じて、役員従業員が迅速に必要な情報にアクセスできるようにします。元のマニュアルより大幅に読みやすく、理解しやすい、コンパクトな回答により、業務効率が向上し、意思決定が迅速に行えるようになります。

Q: ChatBrid の将来的な拡張性について教えてください。
A: ChatBrid は、スケーラブルなアーキテクチャを採用しており、企業の成長やナレッジベースの拡大に対応できます。大規模化による性能低下、精度低下がありません。また、新しい LLM や技術を容易に組み込むことができるため、将来の拡張性が高いです。

Q: ChatBrid が精度面で他の RAG エンジンよりも優れている点は何ですか？
A: ChatBrid は、大規模化による精度低下を防ぐために、独自の技術的工夫を施しています(特願2024-33371 特許明細書）。また、ユーザー辞書作成も可能でトークナイザーの精密なチューニングができます。同一綴りで異なる意味をもつ用語については、各々のデータセットプロンプトに各分野での異なる用語定義・説明を記述します。これにより、各々正確な理解が可能となります。これらにより、特定の業界、複数の専門分野に同時に対応しつつ高精度なナレッジ検索を実現します。

Q: ChatBrid のベクトルストレージの技術的優位性は何ですか？
A: ChatBrid は、独自のベクトルストレージを活用し、例えば tf*idf 相当の重みを、品詞や、事前学習した意味カテゴリに応じて自在に調整することで、例えば日本語特有の不要語（「こと」「もの」など）を除外し、精度の高い検索結果を実現します。

Q: ChatBrid の 5 階層シソーラスはどのように機能しますか？
A: 質問文中の単語が、マニュアル中の言葉と一致しない場合でも、ナレッジ類似ランキングの精度が確保されやすくなります。※シソーラスで定義された意味カテゴリをベクトルに追加し、類義語や対義語＋否定辞にも対応した精度の高いナレッジ検索を行っています。これにより、ユーザーが使う多様な言い回しに対しても正確な検索が可能になります。



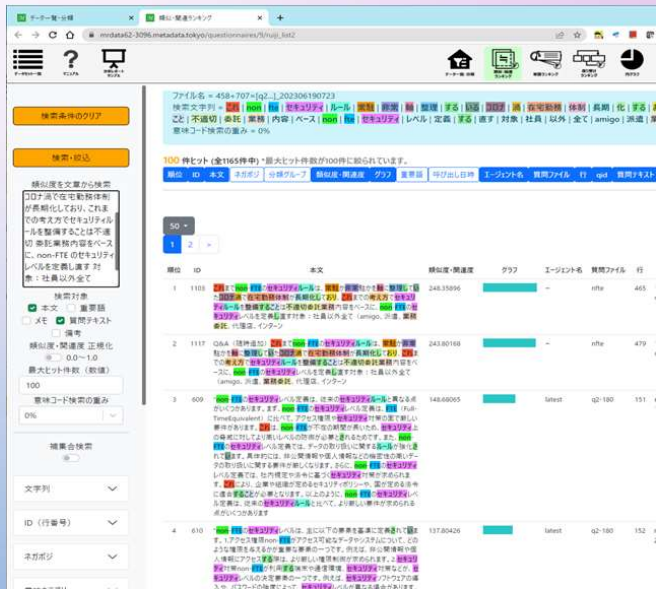
Q: ChatBrid のビジュアル類似検索はどのように知識デバッグを支援しますか？
A: ビジュアル類似検索により、検索結果を色付け表示し、ナレッジベースにそもそも回答が存在していたか等の目視評価が容易になっています。検索結果を視覚的に評価し、欠落した情報や誤った回答を瞬時に確認することにより、ナレッジベースの品質管理が容易になり、継続的な改善が可能です。以上により、AI の回答の品質を迅速に評価し、ナレッジベースの欠落や不適切な記述を容易に発見し、改善できます。

Q: ChatBrid はどのようにして検索結果の視覚化を行いますか？
A: ChatBrid は、検索結果を色付けして表示することで、ナレッジデバッグの視覚フィードバックを提供し、問題の早期発見と修正を支援します。

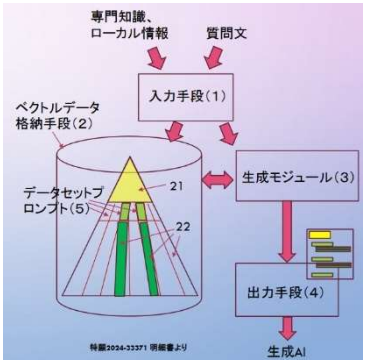
ベクトル検索結果のビジュアル化： xAI : explainable AI
～知識デバッグのための視覚フィードバック

- 自前のベクトルストレージにより、回答不能や不適切な回答のベクトル検索結果を色付け表示
- **ビジュアル類似検索でAIの回答を秒速で評価 → ナレッジ自体の有無、品質を評価、改良**
- マニュアルの記述が不適切だった場合の即座のビジュアル確認と修正が容易。

- マッチした単語と意味コードを強調表示
- 不適切な回答に対する即時の視覚フィードバック
- MARKDOWNをその場で修正 → 即検索に反映



Q: ChatBrid はなぜ影プロンプト送出量の削減が可能なのですか？
A: ChatBrid は、データセットのジャンル分けとプロンプト送信の最適化により、LLM への影プロンプト送出量を削減し、高速かつ低コストで高精度な応答を実現しています。



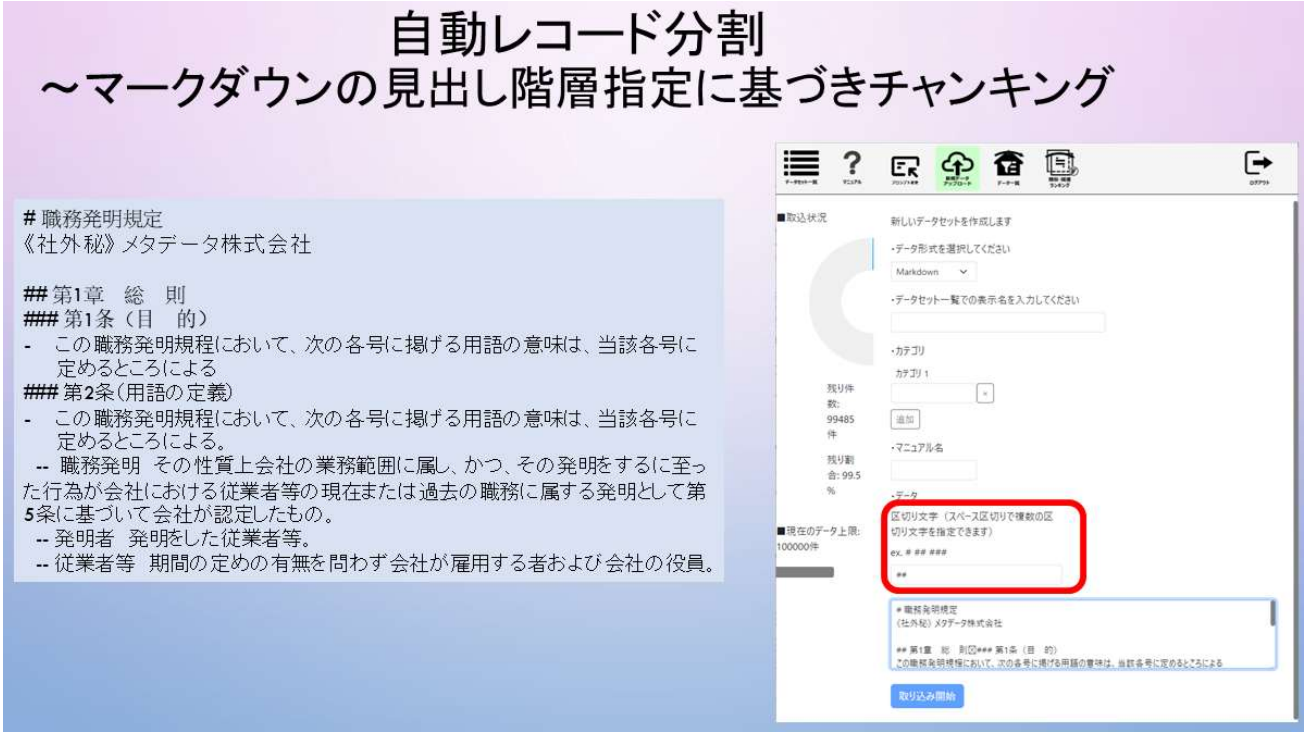
Q: ChatBrid の知識粒度の最適化はどのように行われますか？
A: ChatBrid は、知識粒度を小さめにしつつ、上位見出し(コンテクスト中の当該知識の位置)を自動で付加することで、LLM の理解を促進し、検索精度を向上させます。

Q: ChatBrid の知識粒度調整機能は、どのようにして検索精度を向上させるのですか？
A: 同じデータソースであっても、マークダウン階層のどこを切れ目とするによってナレッジ粒度を切り替えることができます。いずれの粒度であっても、適切なコンテクスト情報を自動付加することで、LLM がナレッジをより深く理解できるようになります。複数トピックの混在によるハルシネーショ

ン発生率の抑制と、コンテクスト破断による理解失敗の生成率のトレードオフを確認しつつ、当該分野のナレッジについて高精度を達成するのに最適な知識粒度を調整することができます。

Q: ChatBrid はどのようにしてマニュアルの自動レコード分割を行いますか？
A: ChatBrid は、マークダウン形式の見出し階層を基に、自動的に適切なレコードサイズに分割し、検索精度を高めるためのチャンキングを行います。

Q: ChatBrid の自動レコード分割機能はどのような利点がありますか？
A: 自動レコード分割機能により、巨大なレコード内の無関係な情報の混入を防ぎ、精度の高い検索結果を提供します。また、小さめの知識粒度のレコードについては、LLM の理解を助けるためにコンテクスト情報を自動で付加します。

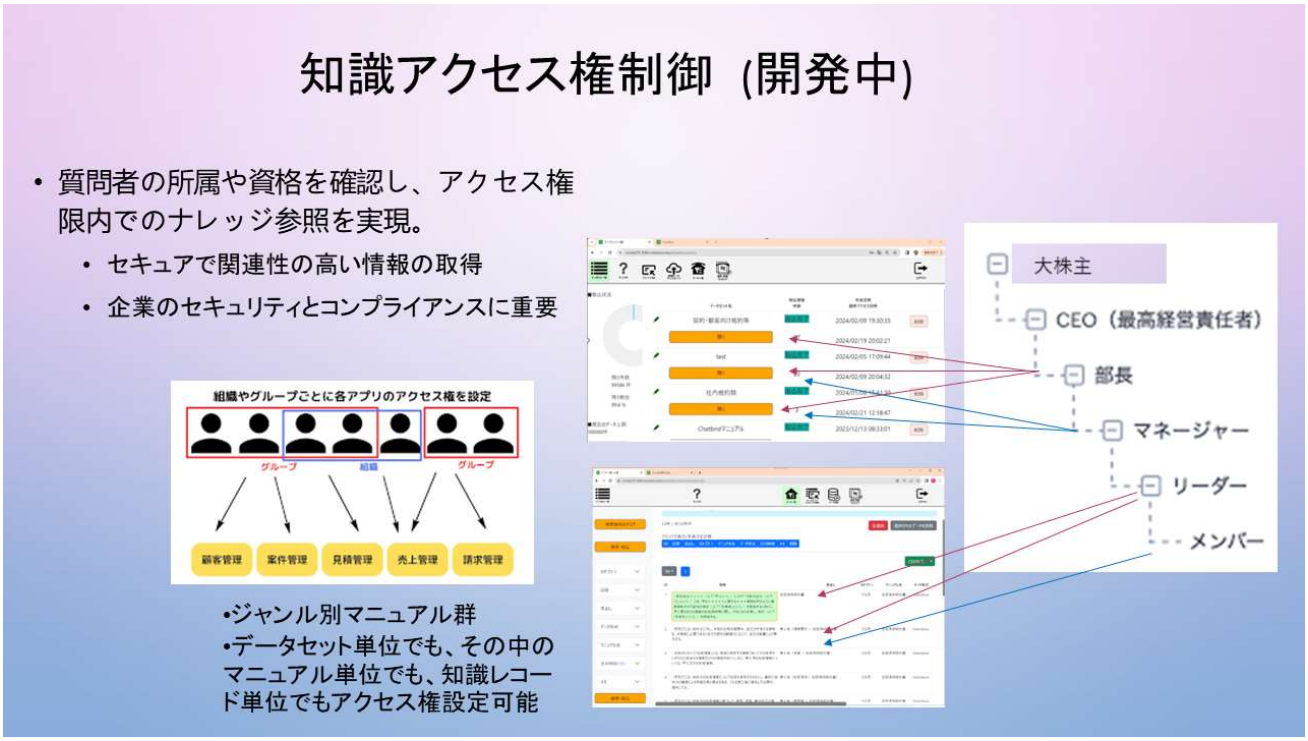


Q: ChatBrid の Exact Match 機能はどのような場合に役立ちますか？
A: Exact Match 機能は、指定された文字列に対して厳密に一致する情報のみを検索するため、材料名や型番、法律名、文書タイトルなど、似て非なる名称について、取り違えなどの誤りを防ぐのに役立ちます。

Q: ChatBridOR の Exact Match 機能は、どのような業務に役立ちますか？
A: Exact Match 機能は、薬品名や素材名などの取り違えを防ぎ、法的文書や技術資料の検索において致命的なエラーを防止するため、非常に重要です。

Q: ChatBrid の知識アクセス権制御はどのように機能しますか？
A: ChatBrid では、ユーザーの所属や役職に応じたアクセス権制御が可能です。機密性の高い情報が、適切な利用権限を持つユーザーにのみ、その知識に基づいた回答が提供されるように設計されています。

Q: ChatBrid の属性フィルタリング機能は何を実現しますか？
A: 属性フィルタリング機能により、地域や期間に基づく厳密な知識絞り込みが可能になり（適用可能な条例、法令などの絞り込みで特に有効）、特定の条件に一致する正確な情報を提供します。



Q: ChatBrid は API モードでどのように企業システムと統合されますか？
A: ChatBrid は API モードで動作し、Slack、Teams はじめ、企業の各種システム（独自システム含む）とシームレスに統合可能です。これにより、既存の業務フローにスムーズに組み込むことができます。

Q: ChatBrid が提供する知識管理マニュアルの特徴は何ですか？
A: ChatBrid は、100 ページ以上の包括的な知識管理マニュアルを提供しており、MS オフィス文書（.docx、.xlsx、.pptx）フォーマット等を効率的にマークダウンに変換する手法が含まれています。

Q: ChatBrid はどのようにして複数のマニュアルを効率的に管理しますか？
A: ChatBrid は、ジャンルごとにマニュアルを整理し、各データセットに適切に分割・登録することで、効率的な管理と検索を実現します。