RAGによる生成AIの業務利用適応に向けた論点

KPMG Advisory Lighthouse

Contents

RAGの仕組み 2

考慮すべきポイント 4

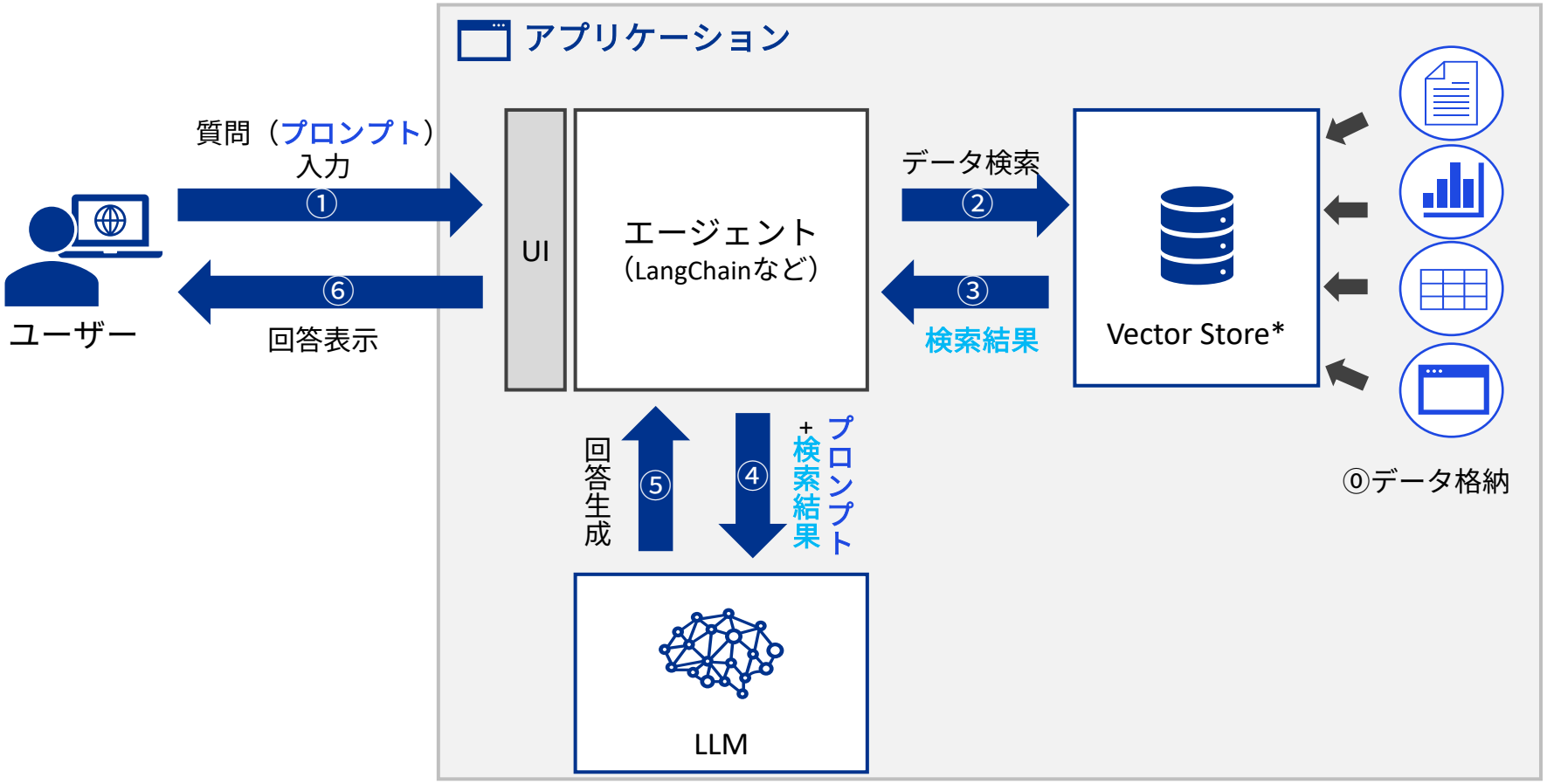
データ準備における論点 5

アプリケーションにおける論点 6

© 2024 KPMG Advisory Lighthouse, Inc., a company established under the Japan Companies Act and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

RAGの仕組み

RAGでは、LLMによるテキスト生成に情報の検索機能を付加することで独自データに基づく回答を可能にしている



*RAGで利用したいデータを格納したデータベース

© 2024 KPMG Advisory Lighthouse, Inc., a company established under the Japan Companies Act and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

事前準備段階で行う作業

『⓪データ格納』はRAG利用の事前準備に相当し、4ステップに分かれている

事前準備段階で行う作業

Ⅰ 対象とするデータの選定	RAGの目的である「特定のユースケースにおいてユーザーが期待する回答を得る」ためのデータを定義し、既存データの選定・新規データの作成を行う。
↓ データファイル	
Ⅱ データの読み込み・抽出	データ検索を可能とするため、RAGで利用したいデータを整形し、一定サイズのチャンク（1つのテキストデータ）に分解する。
↓ データチャンク	
Ⅲ テキストのエンベディング	データ検索時にベクトル検索*を行うために、データチャンクを数百～数千の次元を持ったベクトルに変換する。
↓ チャンク+ベクトル	
Ⅳ Vector Storeへの格納	データチャンクとベクトルはVector Storeに格納し、アプリケーションで検索可能な状態にする。

*テキストや画像などのデータを数値ベクトル(数値の配列)として表現し、ベクトル間の類似性を数学的に計算することで関連する情報を効率的に検索する手法であり、単純なキーワード検索では実現できないコンテクストも加味した検索を可能とする。データと同様に、検索時に入力されたプロンプトもベクトル化し、2つのベクトルの類似度を計算する。

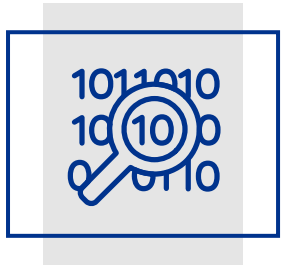
KPMG
© 2024 KPMG Advisory Lighthouse, Inc., a company established under the Japan Companies Act and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

考慮すべきポイント

:データ準備の論点(d)
:アプリケーションの論点(a)

RAGの業務水準適応に向けて、特にRAG特有のプロセスである「データの検索」「検索結果の選定」「検索結果を加味した回答の生成」にフォーカスして論点を見ていく必要がある

プロンプトの意図に沿ったデータ検索の実行

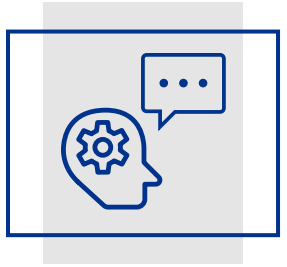


- d1 ユースケースに適したデータの準備
- d6 適切なベクトル次元数の選択
- d7 エンベディング精度の検証

回答の生成に必要な検索結果の選定

- a1 ユースケースに適した類似度の算出
- a2 類似度以外での検索結果のリランク
- a3 LLMに入力するデータチャンク数の決定

検索結果を反映した適切な回答の生成



- d1 ユースケースに適したデータの準備
- a5 生成回答の評価手法の確立
- a4 十分な文章量に対応できるモデルの選択
  - d4 チャンクサイズの決定
  - a3 チャンク数の決定

© 2024 KPMG Advisory Lighthouse, Inc., a company established under the Japan Companies Act and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

データ準備における論点

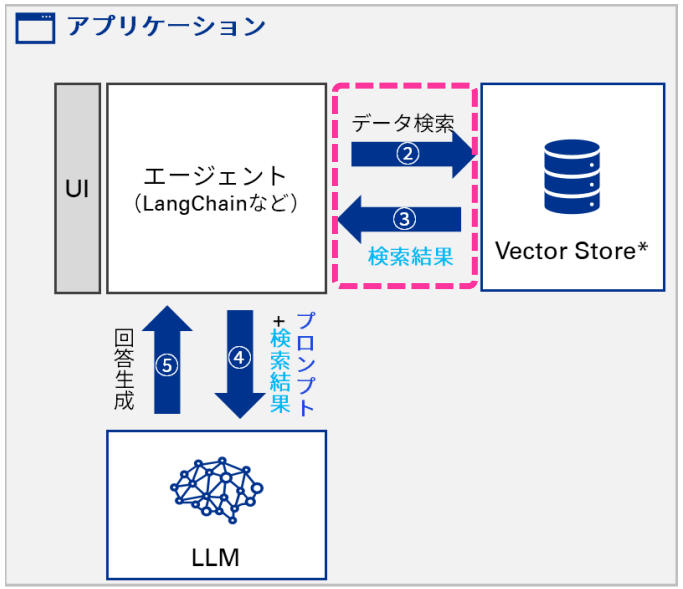
各ステップにおける論点について適切な検討を行うことでRAGの業務効果を最大化できる

作業ステップ	検討すべき論点		
Ⅰ 対象とするデータの選定	d1 ユースケースに適したデータの準備 想定される問いに対して十分な情報を提供できるデータを用意すると同時に、不要な情報・齟齬のある情報など、LLMが回答を生成する際に誤った解釈をしないよう品質にも気を配る	d2 データ整形要否の確認 利用できるデータファイルがない場合、システムで読み込みができないファイルしかない場合には、整形や新規作成を行う。抽出プログラムにより、抽出可否が異なるため併せて検討する	
Ⅱ データの読み込み・抽出	d3 データに適した抽出手段の選択 二段組のテキストファイルのレイアウト崩れや、特殊な文字コードの文字化けなど、抽出の際に不具合が起こる場合がある。 対応可能なファイルレイアウト・文字コードなどには注意する	d4 適切なチャンクサイズの決定 サイズが大きすぎる場合、関係性の低い内容を含むテキストを参考に回答を生成してしまい、回答精度が低下しやすい。他方、小さすぎても必要なコンテクストが取得しきれず、不十分な回答になりかねないため、バランスが重要	d5 メタデータ取得可否の確認 RAGでは情報の出典元ファイル名やページ番号など、メタデータの明示が可能である。利用サービスにより明示の可否が異なるため、データ抽出時のメタデータ取得可否を確認しておく
Ⅲ テキストのエンベディング	d6 適切なベクトル次元数の選択 次元数は「検索条件への一致度」を計算するためのパラメータ数であるため、一般的には多い方が精度は高くなる。利用するデータの多寡、多様性などを考慮して適切な判断を行う	d7 エンベディング精度の検証 ベクトル変換時のロジックも重要な論点である。同じデータチャンクに対しても、モデルごとに生成されるベクトルが異なるため、適切なベクトル検索結果が返ってくるよう、検証が必要である	
Ⅳ Vector Storeへの格納	d8 上限文字数の確認 前述の「チャンクサイズ」「ベクトル次元数」はデータベースであるVector Storeに格納できる範囲で設定する必要があるため、併せて検討する	d9 過不足ないデータ量の維持・管理 Vector Storeのサイズを大きくするほどデータを増やすことができ、検索したい情報を引き出せる可能性は高まるが、検索対象が増えるために精度は低下するリスクがある	

© 2024 KPMG Advisory Lighthouse, Inc., a company established under the Japan Companies Act and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

アプリケーションにおける論点（1/2）

利用段階における論点についても考慮が必要である
回答に必要な情報を取得する、データ検索時の論点は下記のとおり



a1 ユースケースに適した類似度算出手段の選択

類似度の計算方法により、同じプロンプトに対しても検索結果の順位が異なるため、計算方法の選定は重要である。
自社で想定するユースケースに応じて適切に選定する必要がある。

・代表的な計算方法

手法	代表的なユースケース	計算方法
コサイン類似度	• テキスト検索 • ドキュメント検索	ベクトルの角度・大きさを基に類似度を計算
ユークリッド距離	• クラスタリング • 近傍探索	二点間の直線距離で類似度を計算
ジャカード係数	• ドキュメントの単語セットの類似度測定	2つの集合の交わりのサイズを合併で除算し類似度を計算

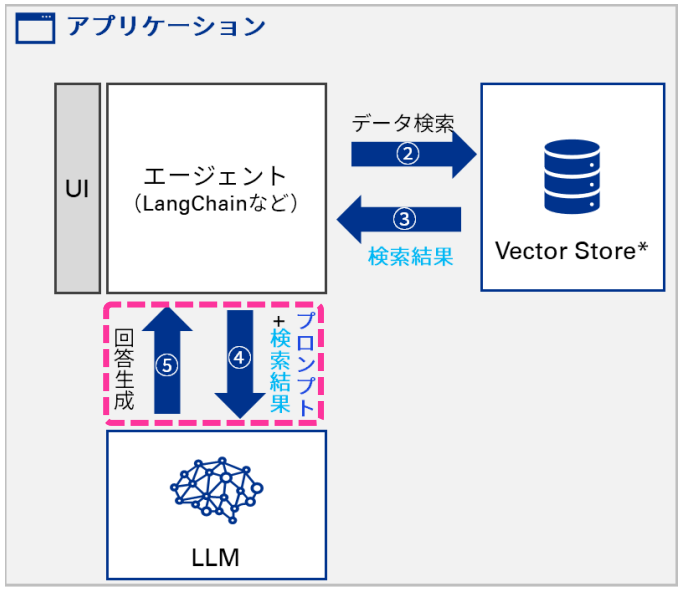
a2 類似度以外での検索結果のリランク

前述のとおり、ユースケースにより、純粋にプロンプトへの一致度のみで検索を行うのではなく、ばらつきを持たせることで別の視点からの情報を取得した方が望ましい場合もある。
その場合には類似度計算でランク付けした後、1～30位、31位～60位などグループを作成した中から各5件をサンプリングするなど、類似度ランクと別のロジックで多様性を持たせることも検討したい。

© 2024 KPMG Advisory Lighthouse, Inc., a company established under the Japan Companies Act and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

アプリケーションにおける論点（2/2）

回答生成時の論点は下記のとおり



a3 LLMに入力するチャンク数の決定

ベクトル検索で取得したデータは、必ずしも最も類似度の高い1件だけをLLMの入力とする必要はない。複数を利用することで回答の精度や多様性を高めることが可能となる。

他方、件数が多過ぎる場合には回答生成に不要な情報も多く混在してしまうほか、検索結果内での情報の齟齬が起こる可能性が高まるため、LLMの回答生成に悪影響を及ぼす可能性がある。

回答生成に利用するチャンク数も慎重に検討したい。

a4 十分な文章量に対応できるモデルの選択

前述のとおり、RAGにおいてLLMに入力されるテキストはプロンプトにデータの検索結果を追加したものである。従って、プロンプトのみであれば文字数は限定的ではあるが、検索結果として入力されるデータチャンクのサイズ、および数によっては入力自体の文字数（トークン数）は大きくなることから、モデルがサポートする上限トークン数に配慮する必要がある。

a5 検索結果の解釈・生成回答の評価手法の確立

検索で取得したデータをLLMが十分に解釈し、過不足のない回答を生成することはRAGにとって必須要件の1つである。プロンプトに対して適切な検索結果を取得できた場合にも、回答生成の段階で必要な情報が欠落したり、誤った解釈による要約が行われた場合には、ユーザーの期待を満たせず業務利用水準には至らない。

プログラムによる評価でよいか、人間が見るべき範囲かを見極め、LLM自体の性能を評価することも欠かせない。

© 2024 KPMG Advisory Lighthouse, Inc., a company established under the Japan Companies Act and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

ここに記載されている情報はあくまで一般的なものであり、特定の個人や組織が置かれている状況に対応するものではありません。私たちは、的確な情報をタイムリーに提供するよう努めておりますが、情報を受け取られた時点およびそれ以降においての正確さは保証の限りではありません。何らかの行動を取られる場合は、ここにある情報のみを根拠とせず、プロフェッショナルが特定の状況を綿密に調査した上で提案する適切なアドバイスをもとにご判断ください。

© 2024 KPMG Advisory Lighthouse, Inc., a company established under the Japan Companies Act and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

The KPMG name and logo are trademarks used under license by the independent member firms of the KPMG global organization.