Engineer

検索拡張生成

📖 質問ごとに取得した文書を基に回答を生成する手法

どのようなものか

検索拡張生成は、言語モデルを使い、二つの段階で質問に答える方法です。まずアプリケーションが文書群を検索し、答えを含む可能性が最も高い箇所を探します。次に、その文章を質問とともにモデルの入力に加えます。モデルは、学習中に吸収した知識ではなく、渡された文章を基に答えを書きます。通常RAGと略されるこの名前は、自由形式の質問への回答など、知識を多く必要とする言語処理を扱った2020年の研究論文に由来します。

その論文は、二種類の記憶を説明しています。事前学習済みモデルがパラメーターの中に保持する学習内容を、著者たちはパラメトリックな記憶と呼びました。学習済みの検索器で検索する、独立した文書インデックスは、ノンパラメトリックな記憶です。このシステムは、事前学習済みの系列変換モデルと、Wikipediaの密ベクトルインデックスを組み合わせ、検索結果の使い方を二通り比較しました。一方は、回答全体を同じ検索結果の文章で条件付け、もう一方は、書き出すトークンごとに異なる文章を参照できます。分野を限定しない質問応答の三つのベンチマークで、それまでの最高性能を更新しました。また著者たちは、検索を使わない同等のモデルと比べ、生成する文章が具体的で、多様で、事実にも忠実だと評価しました。

アプリケーションは、すでに持っている部品でも同じ構成を作れます。自社文書の検索インデックスと、APIを通じて呼び出す大規模言語モデルです。検索側は通常の検索であり、二つの系統がどちらも使われます。スパース検索は、BM25などのスコア関数で文書を順位付けする全文検索のように、単語を照合します。密ベクトル検索は、質問と各文章を数値のベクトルに変換し、最も近いベクトルを持つ文章を返します。これがベクトル検索です。密ベクトルによる文章検索を扱った2020年の論文では、比較的少数の質問と文章の組で学習した、密ベクトルを使う検索器が、上位20件に正解の文章を含む割合で、強力なBM25の基準手法を9から19パーセントポイント上回りました。両方を実行し、二つの一覧を統合する構成も取れます。

概念を判断に生かす

「検索拡張生成とは？」を実際の判断に生かしましょう。その概念を、開発または購入するシステムに結び付けて考えるお手伝いをします。 AIエージェントが記録にアクセスでき、顧客ごとの情報が一本化され、権限を超えた情報が表示されず、回答の出所をたどれる状態にします。

サービスを見る お問い合わせ →

用語の背後にある前提を検証する

解説とリンク先の実績を参考に、プロジェクトについての質問を整理してください。ツールを選ぶ前に、前提を一緒に検討できます。 地理情報パイプラインを50倍に高速化 — 50x 地理データパイプラインの高速化

地理情報パイプラインを50倍に高速化 お問い合わせ →

なぜ重要なのか

言語モデルが知っているのは学習に使った文章にあったことだけで、その文章はある決まった日付までのものであり、自社の契約、価格表、先週の議事録は含まれていません。RAGの論文は、パラメーター内の知識に二つの限界があるとしています。回答の根拠を示しにくく、更新もしにくいことです。検索は両方に対処します。回答は渡された文章を出典として示せるため、読者は原文を開いて確認できます。また、文書がインデックスに入った瞬間に、新しい知識が、追加の学習を行わずモデルへ届きます。

また、誰に何を見せてよいかという判断を、本来あるべき場所に戻します。文章の取得は検索であり、質問者の権限で検索すれば、モデルに見せるのは、その人自身が開ける文章だけになります。同じ文書で学習させたモデルには、その境界がありません。学習した内容は、そのモデルを使う全員が利用できます。

代償は、回答の良さが検索結果を超えられないことです。誤った文章を渡されたモデルは、そこから流暢な答えを書き、有用な文章を何も得られなければ、それでも記憶から答えることがあります。そのため、検索拡張型のシステムでは、作業の大部分はモデルよりもデータと検索に向けられます。

実際の活用例

次の記録は、一つの質問が検索拡張型システムを通る過程を示しています。検索は質問した利用者の権限で行い、各文章には出典のファイルとページを添えます。指示は、その文章だけを基に答えるようモデルに求めています。

```
質問: 注文4471にはどの配送条件が適用されますか？
利用者:  sales-3 (ロール: sales, 地域: north)

取得結果（sales-3が読めるものだけ）:
  [1] contracts/framework-2025.pdf, 4ページ
```

```
        "商品は買い手の倉庫に配送します。
         ただし、注文に別の定めがある場合を除きます。"
    [2] orders/4471.txt, 12行目
        "配送条件：基本契約に従う。"

指示：
  以下の文章だけを基に答え、[n]の形式で出典を示してください。
  質問に答える情報がない場合は、その旨を伝えてください。

回答：
  注文4471は基本契約を参照しており[2]、
  商品はその契約に従い買い手の倉庫に配送します[1]。
```

この記録のどの部分も、システムが検査できるデータです。既知の質問に正しい文章が返ったかはテストで確認でき、回答中の各 `[n]` が実際に取得した文章を指すかはスクリプトで確認できます。回答を読んだ人は、契約の4ページ目を開けます。

必要な下準備は、このサイトの会社のデータをAIエージェントが使える状態にするページで説明しています。古いPDF、スキャン、メールのアーカイブを、出典を添えた検索可能なテキストに変換し、各行に、元のシステム、読み込み処理、ファイル、ページを残す方法です。エージェントは数値の根拠となる記録を示し、人はそれを開けます。同ページは、規則をプロンプトではなくデータに持たせることも求めています。エージェントは、決して管理者の権限ではなく、利用者自身の権限で動き、すべての読み取りに、誰のために行ったかを記録します。

問題が起きるところ

検索で見つからなくても、モデルが答える。 正しい文章を取得できなくても、モデルは何かを受け取り、答えを書きます。検索を単独で測定します。実際の質問と、それに答える文章を書き出し、モデルに渡す結果にその文章が現れる割合を調べます。文章だけでは質問に答えられない場合は、そう伝えるようモデルに指示し、そのような質問もテストに含めます。

文章の分割で、答えが二つに切れる。 文書を一定の長さで分けると、表と見出し、条文と例外が切り離され、一方だけを取得すると誤解を招きます。文書がもともと持つ節、ページ、表の行といった構造に沿って分割し、すべての断片に文書名と位置を残します。

インデックスが質問者を考慮しない。 すべてを読めるサービスが作ったインデックスは、質問者が開けない文書の文章も返し、モデルがその人に要約してしまいます。モデルに何かが届く前に、検索処理の中、すなわちインデックスかデー

タベースで、利用者の権限に応じて絞り込みます。プロンプトの規則はアクセス制御ではありません。

文書に指示が仕込まれる。 取得した文章は他人が書いたもので、モデルは、文章と指示を確実に区別できません。NISTの生成AI向けプロファイルは、間接的なプロンプトインジェクションを、取得される可能性が高いデータを通じて、言語モデルを組み込んだアプリケーションを攻撃するものと説明しています。2023年の論文は、実際のシステムへの攻撃を実証しました。取得したすべての文章を信頼できない入力として扱い、コードによる検査なしに、その文章から動作を引き起こせないようにします。また、モデルが呼べるツールに、利用者を超える権限を与えません。

インデックスの更新が遅れる。 文書が変更・削除されても、インデックスに古い文章が残れば、システムは、もう有効でない契約を基に自信を持って答えます。出典が変わるたびにインデックスを更新し、文書を削除したら文章も取り除き、各出典の日付を回答の横に示します。

引用先が、回答の主張を支えていない。 モデルは、文章[2]が裏付けていない文の出典に[2]を挙げることがあります。NISTも、生成した出典自体が作り話になり得ると指摘しています。すべての引用先が、取得した文章を指すことをコードで検査し、定期的なテストの一環として、人が回答を抜き取り、示された出典と照合します。

参考資料

- Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks — RAGモデルを提案し、二つの定式化を質問応答と文章生成で比較した2020年の論文です。
- Dense Passage Retrieval for Open-Domain Question Answering — 質問と文章の組から学習する、密ベクトルを使う検索器を、答えを含む文章の発見についてBM25と比較・測定しています。
- NIST AI 600-1：生成AIプロファイル — 作り話の生成、データのプライバシー、間接的なプロンプトインジェクションを扱う、NISTの生成AIリスクプロファイルです。
- Not what you’ve signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection — 取得データに隠された指示がアプリケーションを乗っ取る仕組みと、実システムへの攻撃を説明します。

用語集

実際の制約に合わせて選ぶ

代表的なデータセット、そこから答えを得たい問い、現在のボトルネックをお知らせください。品質、処理量、復旧の測り方を合意できます。

料金を見る お問い合わせ →

用語を実務に結び付ける

「検索拡張生成とは？」がお客様の組織にどう当てはまるかをご相談ください。現在のシステムの具体例が一つあれば、話を始められます。

サービスを見る

welcome@engineer.company +45 50 10 61 84 デンマーク、コペンハーゲン

このページを持ち歩く

連絡先カード

心を込めて丁寧に制作： D.

CVR 45969703 · Tagensvej 105C, 3. th, 2200 København N

著作権 © 2025 – 現在 · Engineer ApS · すべての権利を留保します。

印刷元： https://engineer.company/ja/glossary/retrieval-augmented-generation/