

施工技術探索システムの開発

建設分野における Retrieval-Augmented Generation の適用

大政 文和^{*1}

Keywords : RAG, LLM, Generative AI, Digital Transformation

検索拡張生成, 大規模言語モデル, 生成 AI, DX

1. はじめに

建設業界は現在、生産性向上、人材不足、安全性の向上等多くの課題に直面している。これらの課題に対応するためには、デジタル技術の活用が不可欠である。とりわけ、現場配属間もない若手社員へのナレッジ継承や、膨大な専門技術資料や書類の中から必要としている資料や書類を短時間で取り出すようなナレッジ探索についても、デジタル技術の活用による効率化が求められている。特に注目されているのが、最近、開発競争が過熱している生成 AI の一つである大規模言語モデル (Large Language Model : LLM, 以後、分かりやすいように生成 AI と呼ぶ) である。生成 AI は質問に対して、膨大な学習データを基に人間のように回答できることから、そのような課題を解決するツールとしての活用が期待されている。しかし、現状の生成 AI には構造上の問題ともいえるハルシネーション（もっともらしい嘘をつくような現象）があり、現在提供されている生成 AI のサービスを直接使用することが難しい。

ハルシネーションを解決する手段の一つとして、RAG (Retrieval-Augmented Generation : 生成 AI と検索を組み合わせた即応性の高い技術) が普及しつつある。RAG は膨大な専門知識を効率的に活用し、正確で文脈に応じた情報を提供することができる。筆者は 2022 年 11 月に Chat-GPT が公開された直後より、RAG による社内の建築施工技術探索システムの開発を開始し、社員による試用を繰り返している。本報では、開発したシステムの概要とその運用について報告する。

2. システム概要

本システムは、施工に関わる技術的な質問に即時に

*1 技術センター 先進技術開発部 AI 連携技術開発室

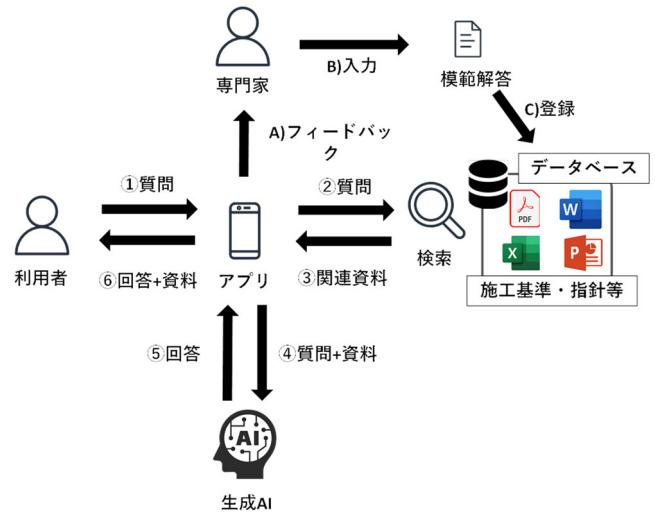


図-1 アプリ概要

Fig.1 conceptual diagram of application

回答し、若手社員の業務を補助する検索システムである。開発した施工技術探索システムの概要図を図-1 に示す。利用者が入力した「質問」を基に自社のデータベースから施工基準・指針・施工事例等の「関連資料」を検索する。アプリケーションは生成 AI に「質問」と「関連資料」を入力として与え、これらの情報を基に適切な回答を生成する。さらに、これらの回答はシステム管理者のみがアクセスできる管理画面において、適切な回答が行われているかチェックすることができ、回答が不正確だった場合は、管理者が専門家に模範解答作成を依頼し、確度が高い優先的に検索される“模範解答データ”としてデータベースに追加する。また、データベースの内容に改善点があると判断されれば、データベースの更新を実施する。

ここで、本システムの根幹技術となる RAG について説明する。RAG (Retrieval-Augmented Generation) は、

生成 AI と情報検索技術を組み合わせた革新的な AI システムである。Lewis et al.¹⁾によって提案されたこの手法は、以下の 2 つの主要コンポーネントから構成される。

- ・検索 (Retrieval) コンポーネント：利用者の質問に関する情報を大規模なデータベースから効率的に抽出する。
- ・生成 (Generation) コンポーネント：抽出された情報を基に、生成 AI を用いて適切な回答を生成する。

3. システム構成

本章では本システムにおける検索コンポーネントと生成の特徴、およびそれらの具体的な実装について述べる。

3.1 システム全体構成

図-2 にシステム構成図を示す。本システムは、Amazon Web Services (AWS) のクラウドサービスを活用し、セキュアで拡張性の高いアーキテクチャを実現し

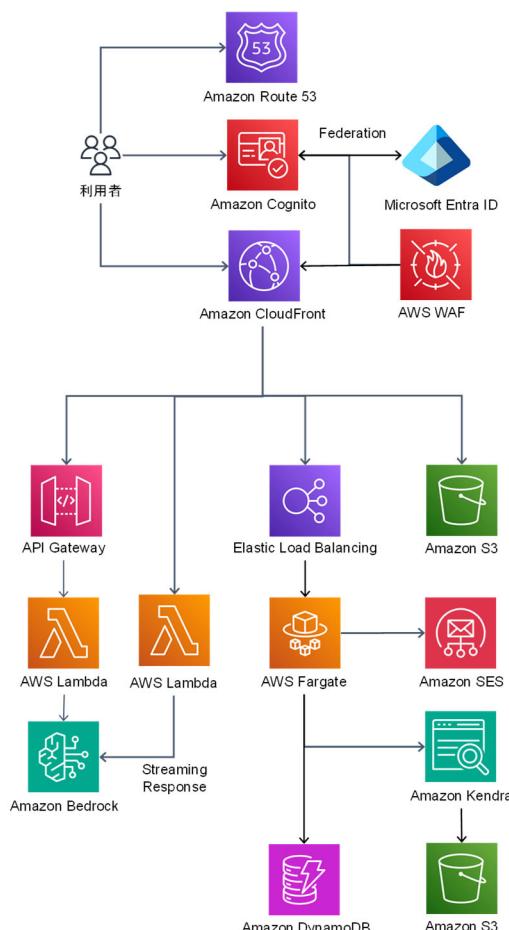


図-2 システム構成図
Fig.2 conceptual diagram of system

ている。主要なコンポーネントとして以下のサービスを採用した。

- ・ユーザー認証：Amazon Cognito と Microsoft Entra ID を使用したフェデレーション認証
- ・ネットワーキングとセキュリティ：Amazon Route 53, AWS WAF
- ・コンテンツ配信：Amazon CloudFront
- ・バックエンド処理：API Gateway+AWS Lambda, Elastic Load Balancing+ AWS Fargate, AWS Lambda(Streaming Response)
- ・ストレージ：Amazon S3
- ・データベースと処理：Amazon DynamoDB
- ・生成 AI：Amazon Bedrock
- ・メッセージング：Amazon SES
- ・検索：Amazon Kendra

3.2 検索システムの構築

RAG の検索コンポーネントとして、Amazon Kendra²⁾ を採用した。Kendra は機械学習を活用した高度な検索機能を提供し、非構造化データに対しても効果的な検索が可能である。本システムでは、検索システムの特徴を最大限に活用するため、以下の実装を行った。

- a) インデックス構築：収集した関連文書を検索システムのインデックスに登録し、全文検索を可能にした。
- b) メタデータ設定：文書タイプやデータソース等のメタデータを各文書に付与し、詳細な検索条件の設定を可能にした。
- c) アクセス制御：文書の機密性に応じたアクセス制御を実装し、セキュリティを確保した。

3.2.1 データ収集と前処理

本システムの構築にあたり、まず包括的なデータ収集と前処理を行った。データ収集では、施工基準、指針・手引き、施工事例等、建設プロジェクトに関連する多様な文書を対象とした。これらは社内データベース、自社のリリース文等データソースから収集した。

3.2.2 カスタム辞書の実装

建設業界の専門用語や略語を含むカスタム辞書を作成し、検索システムに実装した。これにより、「PC 工法」と「プレキャストコンクリート工法」等、建設業界で頻繁に使用される同義語や類語を含むクエリに対しても適切な検索結果を提供することができるようになった。

3.2.3 メタデータの定義

収集したデータを文書タイプごとに最適化されたメタデータ構造を設計した。利用者はメタデータによって検索対象を容易に絞り込むことが可能である。

3.3 生成AIのモデル選択

本システムのRAG (Retrieval-Augmented Generation) の生成AIコンポーネントとしてClaude 3.5 Sonnetを採用した。Claude 3.5 Sonnetは、Anthropic社が2024年に発表した生成AIであり、Claude 3モデルファミリーの一つである。

Claude 3ファミリーには、仕様の異なるHaiku、Opus、Sonnetの3種類がある。Claude 3ファミリーの中で、Haikuは最も軽量で高速なモデルであるが、建設業界の複雑な要求に対応するには能力が不十分である可能性がある。一方、Opusは最も高性能だが、リソース要求が高く、リアルタイムでの応答が求められる場面での使用には適していない。実際の導入を考慮した場合、コスト効率は重要な要素となる。Opusは高性能ながら、利用コストが最も高い。ただし、特に高度な分析等、極めて複雑なタスクに対してはOpusの使用も検討する余地がある。Sonnetは、処理速度とモデル性能のバランスが優れており、本システムの実用的な展開に最適であると判断した。

3.4 RAGの構築

RAGの構築は、複数の段階を経て実施した。検索システムの機能を最大限に活用するため、ファセットの定義を行い、建設業界特有の要求に対応できるよう調整を加えた。

次に、選択したAIモデルのパラメータ調整を綿密に行い、建設業界特有の文脈に適した出力を得られるよう最適化を図った。また、効果的なプロンプトテンプレートを設計し、検索結果と質問を適切に組み合わせる方法を開発した。この過程では、建設業界特有の文脈や要求を十分に考慮したプロンプトを策定した。

検索結果の処理においては、検索システムからの出力を生成AIに入力する前に、関連性や重要度に基づいたフィルタリングと順位付けを実施した。さらに、検索結果の要約を生成する際に入力の処理を最適化することで、生成AIの性能を最大限に引き出す工夫を行った。

システムの出力品質を担保するため、生成された回答の品質のフィードバックを行うモジュールを実装した。必要に応じて、追加情報や参考文献の付加を自動的に行う機能も追加し、ユーザーにとってより有用な情報提供を可能にした。

さらに、継続的な改善を可能にするため、ユーザーからのフィードバックを収集する機能を導入した。

4. 社内展開戦略

4.1 PoC

社内へ普及展開するにあたって、開発部門の従業員、普及展開先の部門の幹部、作業所の従業員を対象としてPoCを行った。以下にその内容について報告する。

4.1.1 実施概要

実施概要について以下に示す。

1)概要

- ・対象人数：60名
- ・対象データ：施工基準、指針・手引き、社内資料
- ・データ数：10万ドキュメント

2)データの準備

各利用部門の責任者にヒアリングを行い業務上必要になる資料・データを収集した。

4.1.2 試行結果と課題

本システムの試行結果から、その有効性が明らかになった。

- ・多様な情報源からの高精度な検索機能により、ユーザーの情報アクセスが大幅に向上したこと。
- ・RAGによる要約等の機能により、情報処理の効率化したこと。
- ・自社の収集したデータのみならず、外部へ公開しているWEBの情報も包括的に取り扱うことで、情報収集の労力を削減していること。

試行対象者から得られた改善点を以下に示す。

- ・関連する情報も得られるが使えない情報も多く表示されること。
- ・AIが関連する資料に使える情報がない場合、「分かりません」とよく回答すること。
- ・正確で明確な質問をしなければ良い回答が得られないこと。
- ・回答までの時間が長く感じること。

4.2 PoCの結果を踏まえた改善策

4.2.1 機能改善

① コンテキストの生成方法の変更

通常のRAGシステムでは、Langchainライブラリを利用してコンテキストを生成することが一般的である。しかし、この手法にはカスタマイズ性の制限や処理速度の問題、さらには外部ライブラリへの依存というデメリットが存在する。これらの課題を解決するため、Langchainに依存しない独自のコンテキスト生成方法を開発した。

具体的には、検索システムによりデータベースから抽出されたデータに対して直接処理を行うことで、よ

り柔軟かつ効率的なコンテキスト生成を実現した。この手法により、システムのカスタマイズ性が向上するとともに、サーバでの処理を軽減できることから処理速度の改善も期待できる。さらに、外部ライブラリへの依存を減らすことで、システムの安定性と保守性も向上する。

② 建設ドメインの知識を優先

建設ドメインに特化した RAG システムの性能を最大限に引き出すためには、ドメイン固有の知識を効果的に活用することが不可欠である。本研究では、プロンプトエンジニアリング技術を駆使し、建設ドメインの知識を優先的に取り込む手法を採用した。

具体的には、建設ドメイン固有の用語と概念に対して、知識を優先的に組み込むプロンプトを工夫した。これにより、生成されるコンテキストの質と関連性を大幅に向上させることができた。

4.2.2 運用体制と模範解答データベースの整備

PoC の結果から、単に RAG システムの改善を行うだけでは、実用に耐えうるシステムとしての維持は困難と判断し、以下に示す運用体制と模範解答データベースおよびその運用システムの開発を実施した。

① サポート体制

RAG システムの効率的な運用を実現するために既存のサポート体制を利用するには重要である。具体的には、現場に対して技術支援を行っている既存の本社チーム（建設業界のドメイン知識を有する専門家によって構成）と連携し、RAG による回答の監視やユーザーからの質問の収集と分類、データベースの更新を行う体制を構築した。合わせて、当サポートチームが効率的に監視やフィードバック作業を行うことができる管理用ユーザーインターフェイス（UI）を開発した。

② 模範解答データベースの構築

上記サポート体制に加え、短時間で正答率を向上させるツールとして、模範解答データベースと模範解答データをサポートチームが効率的に登録できる機能を管理用 UI に追加した。さらに、ユーザーが模範解答データを容易に検索できるように検索機能の実装を行った。将来的には、この模範解答データベースを RAG システムのデータベースに統合することで、システムの回答精度をさらに向上させる予定である。

4.2.3 Web アプリの開発

本 Web アプリは、利用者が質問をアプリに入力し、AI がその質問に対して回答を生成するシステムである。機能の一例について、公開データによるデモ版の実際の画面を図-3 に示す。図-3 内の赤字の説明語の内容は

以下のとおりである。

- ・入力欄：ユーザーが自然言語で質問を入力するための領域
- ・フィルター：検索範囲を調整するためのオプション
- ・AIの回答：システムが生成した回答を表示する領域
- ・フィードバック：ユーザーが回答の質を評価するためのボタン
- ・参考資料：回答の根拠となる情報源を表示する領域

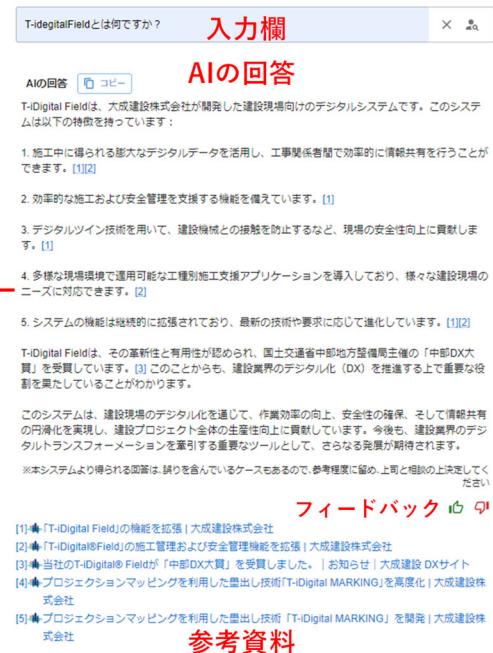


図-3 アプリの画面
Fig.3 screen on application

5. 使用例

本章では施工技術探索システムの使用例を紹介する。開発システムの有効性を確認するため、一般に公開されている生成 AI のみの回答を示し、続いて同じ質問に対する RAG を用いた開発システムの回答を示す。また、開発システムは回答の後に、回答の元となる一次資料の該当ページに即座にアクセスして、回答内容の確認、およびさらに詳細な情報を得ることができるが、ここでは AI の回答のみを表示し比較する。

5.1 社内情報に関する質問

「T-iDigital Field とは何ですか？」と「朝顔の特徴は？」という 2 つの質問で比較検討を行った。

一つ目は、社内の情報に関する質問である。質問にある”T-iDigital® Field”とは、大成建設の土木現場支援ツールとしてデジタル技術を駆使して開発されたシステムの総称である³⁾。

以下に回答を示すが、通常の生成 AI では、質問中の

名詞に似たものについて回答しており、全くの間違いであるのに対し、開発システムは社内のデータに基づいた正確な回答をしている。

質問：“T-iDigitalField とは何ですか？”

a) 生成 AI のみの回答

T-iDigitalField は、○○グループが提供するデジタルトランスフォーメーション(DX)ソリューションです。このプラットフォームは、製造業や産業分野におけるデジタル化を促進し、生産性向上や業務効率化を支援することを目的としています。

T-iDigitalField の主な特徴は以下のとおりです：

データ収集と分析: 工場や設備からリアルタイムでデータを収集し、分析を行います。

可視化: 収集したデータをわかりやすく視覚化し、現場の状況を把握しやすくなります。

AI 活用: 人工知能(AI)を活用して、予測分析や異常検知等を行います。

クラウド連携: クラウドベースのシステムで、場所を問わずアクセスが可能です。

セキュリティ: 高度なセキュリティ対策を施し、重要な産業データを保護します。

b) 開発システム (RAG) の回答

T-iDigital Field は、大成建設株式会社が開発した施工管理および安全管理を支援するシステムです。このシステムの主な特徴は以下の通りです：

- 施工中に取得した膨大なデジタルデータを活用し、工事関係者間での効率的な情報共有を支援します。
- 多様な現場環境で適用可能な工種別施工支援アプリケーションを導入しています。これにより、様々な建設現場での活用が可能となっています。
- デジタルツイン技術を用いて建設機械との接触を防止する機能が追加されました。これにより、現場の安全性が向上しています。

5.2 建設業界の情報に関する質問

二つ目は、建築業界特有の言葉に関する質問である。質問にある”朝顔”とは、「工事中の落下物を防止するため、足場中間に張りめぐらされた仮設棚。特に市街地等の道路側には取り付けなければならない」⁴⁾ものである。

以下に回答を示すが、通常の生成 AI では、”朝顔”を植物として回答しているが、開発システムは建築用語として正確な回答をしている。

質問：“朝顔の特徴は？”

c) 生成 AI のみの回答

朝顔の主な特徴は以下の通りです：

- つる性の一年草です。
- 朝に開花し、昼頃にはしほむ習性があります。
- 花の形はラッパ状で、色は青、紫、ピンク、白など多様です。
- 葉は心臓形で、緑色をしています。
- 種子は黒く、丸い形をしています。
- 日本の夏の風物詩として親しまれています。
- つるを這わせて栽培するが多く、緑のカーテンとしても利用されます。
- 品種改良により、様々な花の形や色が作り出されています。

d) RAG の回答

朝顔は建設現場における重要な安全対策の一つで、主に以下のようない特徴があります：

- 飛来落下物の防止：朝顔は建物や足場から落下する可能性のある物体を捕捉し、地上の歩行者や周辺環境を保護する役割を果たします。
- 危険範囲の縮小：朝顔を設置することで、飛来落下の危険範囲を大幅に縮小することができます。朝顔がない場合、危険範囲は建物高さの 20% (L=H×0.2) とされていますが、朝顔を設置することでこの範囲を小さくすることができます。
- 第三者災害防止：特に市街地での工事において、朝顔の設置は第三者災害・事故防止のための重要な対策の一つとして認識されています。
- 法令遵守：朝顔の設置は「建設工事公衆災害防止対策要綱」(国土交通省)に基づいて行われており、建設現場の安全管理において重要な役割を果たしています。
- 柔軟な対応：朝顔の設置方法や範囲については、「想定される落下物の状況に応じた適切な措置を講ずることにより、これに代えることができる」とされており、現場の状況に応じて柔軟な対応が可能です。

6. 結論と今後の展望

本開発では、建設業界における生産性向上と技術ナレッジの効率的活用を目的として、RAG の技術を応用した施工技術探索システムの開発と実装について報告した。本システムは、AWS のクラウドインフラを基盤とし、Amazon Kendra による高度な検索機能と Claude

3.5 Sonnet による生成 AI を統合したアーキテクチャを採用した。

PoC の結果から、システムの有効性が確認されるとともに、改善すべき課題も明らかになった。特に、検索結果の関連性向上や回答生成の精度改善、応答時間の短縮等が主要な課題として特定された。これらの課題に対し、コンテキストの生成方法の変更や建設ドメインの知識の優先度向上等、具体的な改善策を提案し、実装を進めた。その結果、専門技術に特化して一般的な生成 AI よりも正確な回答を提示できるシステムとして開発することができた。

本開発の成果を踏まえ、以下の点について今後の展望を箇条書きで示す。

- ・システムの継続的改善

ユーザーフィードバックの積極的な収集と分析を通じ、システムの精度と使用性の向上を図る。

- ・データベースの拡充

さらに多様で最新の建設関連資料を追加し、データベースを拡大する。これにより、より様々な質問に対する回答能力の向上が期待できる。

- ・最新 AI 技術の導入

生成 AI の急速な発展に対応するため、定期的にシステムの評価と更新を行う。最新のモデルや技術の導入により、性能向上を継続的に図る。

- ・多言語対応の実現

グローバルな建設プロジェクトや外国語話者の利用に対応するため、多言語でのナレッジ共有や質問が可能なシステムへの拡張を検討する。

- ・法令遵守

建設に関する法令・規制・ガイドラインの変化に対応できる体制を整備する。

- ・新入社員への教育

システムの効果的な活用を促進するため、新入社員の教育・トレーニングへアプリケーションを利用していきたい。

RAG の技術を応用した施工技術探索システムはデジタルトランスフォーメーションを加速し、生産性向上や技術伝承に大きく貢献することが期待される。今後は、より高度な機能の開発と実装を進めていく予定である。

謝辞

本研究の遂行にあたり、AWS Japan の今泉氏、関谷氏には

多大なる支援と技術的助言をいただきました。特に、システムアーキテクチャの設計やAWSサービスの最適な活用方法について、貴重なご意見とご協力をいただきましたことに深く感謝申し上げます。

参考文献

- 1) Patrick Lewis, Retrieval-Augmented Generation for Knowledge-Intensive, NLP Tasks (2020), <https://arxiv.org/abs/2005.11401>
- 2) Amazon Kendra, <https://aws.amazon.com/jp/kendra/>
- 3) 「T-iDigital®Field」の施工管理および安全管理機能を拡張, https://www.taisei.co.jp/about_us/wn/2022/220426_8777.html
- 4) 稲本 稔他, "[図解]建築施工用語辞典", 株式会社井上書院 (1995)