

緒方 陸*1 岡野 将大*1 大久保 順一*1 藤井 純一郎*1
*1 八千代エンジニアリング株式会社

1. 背景・目的

背景:

- 将来的な技術者不足の問題に対し、資料作成業務等の効率化が求められる
- 土木分野におけるLLMやRAGを活用研究はあるものの、評価に重きを置いた研究は多くない

問題:

- 土木分野で利用可能なデータセットは限定的であり、さらに評価方法も定まっていない
- LLMによる評価は人手評価と相関が低い

目的:

- 実用者目線で評価を行い、現状の課題整理
- ⇒課題に対する精度向上へ向けた検討も実施

既往評価手法と人手評価との相関 [緒方+, 2024]

手法	GPT-3.5	GPT-4	PaLM2
BLEU	0.02	-0.35	0.02
ROUGE-1	0.41	-0.09	0.32
ROUGE-2	0.31	0.04	0.15
ROUGE-L	0.52	0	0.35
BERTScore	0.34	0	0.16
G-Eval	0.35	0.32	0.47
Proposed-gpt4	0.48	0.41	0.39

土木文書の例

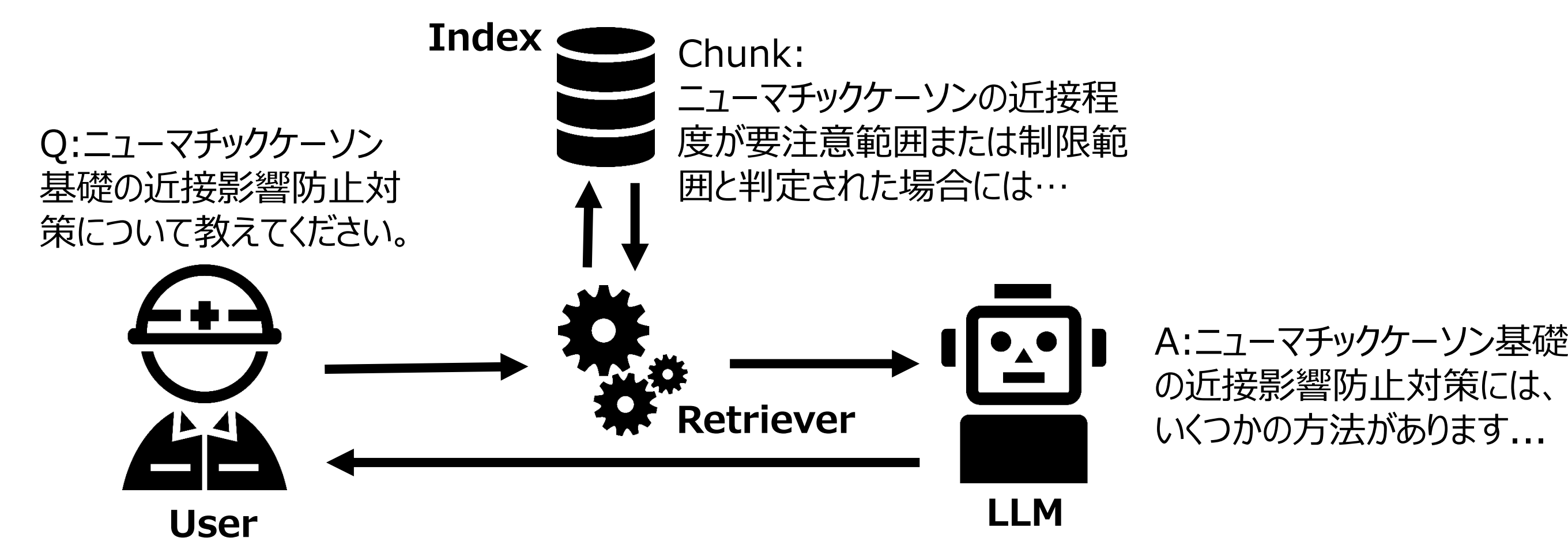
PC斜材付きn型ラーメン橋の斜材は鋼材配置のスペースが限られるため、原則として導入力の大きいB種2号の鋼棒 (SBPR 930/1180) または、SWPR19L 1S28.6 を用いる。
出典: 東北地方整備局 設計施工マニュアル [橋梁編]

設計速度が一時間につき三十キロメートル又は二十キロメートルの道路にあつては、地形の状況その他の特別の理由によりやむを得ない場合においては、十二・五パーセント以下とすることができる。
出典: 道路構造令 (昭和四十五年政令第三百二十号)

土砂災害警戒情報は、土砂災害発生危険基準線 (Critical Line: 以下「CL」という。) をスネークライン (一定時間毎に60分間積算雨量と土壌雨量指数をプロットした点を結んだ線) がおおむね2時間先に超過すると予想された場合に、都道府県と気象庁が共同で発表を行うものである。
出典: 国土交通省 河川砂防技術基準 計画編 (基本計画編)

2. 評価結果と課題の整理

RAG構成



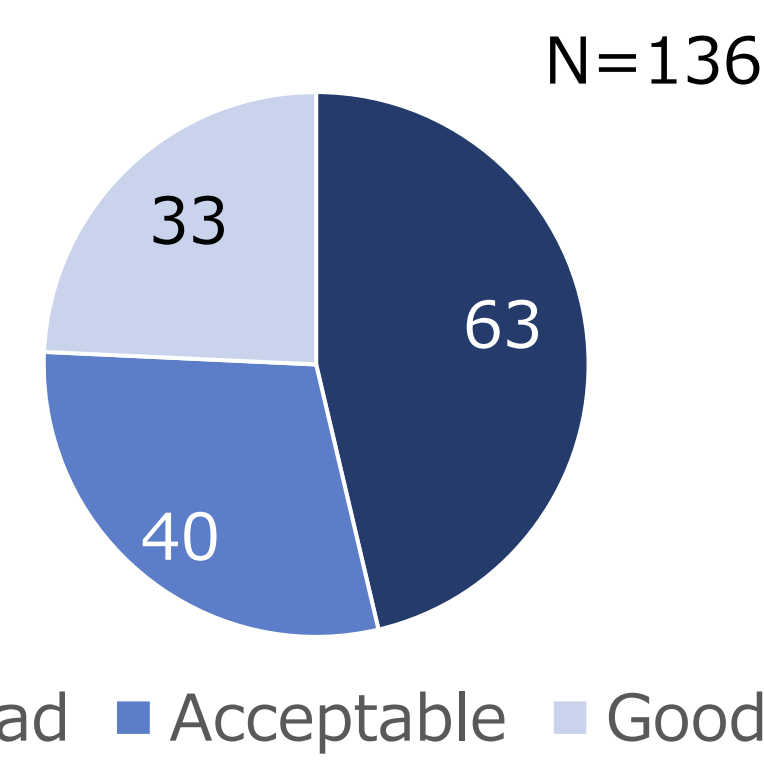
実施内容

- ユーザーは土木技術者であり、実務を想定した質問をシステムへ投げ、回答を得る
- 得られた回答や参照されたテキストなどを見て、3段階の評価("Good", "Acceptable", "Bad")を実施

システム概要

- Indexing: BM25 (1~3-gram) / MiniLM L12 v2 / BM25+reranker*
 - LLM: GPT-3.5-turbo / GPT-4
 - Chunk size: 256, Overlap: 30
 - 対象文書: 橋梁設計分野基準書 15件
- * japanese-reranker-cross-encoder-large-v1

結果



Bad評価の原因と課題の整理

専門用語の理解能力の問題 **課題: データセット作成, LLMのチューニング**

質問	回答
コンクリートのレイタンスとは何ですか	コンテキスト中には「コンクリート表面のレイタンス」という表現があります。したがって、コンクリートのレイタンスとは、 コンクリート表面の滑りやすさや凹凸の程度を表す指標のことです。 × 誤り
桁下空間で確保しなければならない項目は?	桁下空間が確保されているかどうか が重要な項目です。 × 回答になっていない

専門用語に過剰に反応する問題 **課題: 検索方法の改善**

Q: ニューマチックケーソン基礎の近接影響防止対策について教えてください。
Chunk #1:
技術第36巻, 第7号, 同第9号, 同第11号, 1981 2) 福島弘文, 棚村史郎: 周辺地盤のゆるみを考慮したケーソン基礎の設計, 構造物設計資料, NO.79, 1984 3.11 ニューマチックケーソン ニューマチックケーソンの範囲は次の通りとする。(1) 既設構造物が直接基礎・杭基礎・ケーソン基礎の場合 図 3.11.1 ニューマチックケーソンの近接範囲(既設構造物が直接基礎の場合)

Rank	Query	1/2-gram	bm25_score	idf_score	tf_score
1	ニューマチック		6.41	5.71	1.12
2	ニューマチックケーソン		6.41	5.71	1.12
3	ケーソン基礎		5.10	4.54	1.12
4	の近接		4.98	4.44	1.12
5	ケーソン		4.66	4.15	1.12

その他の問題

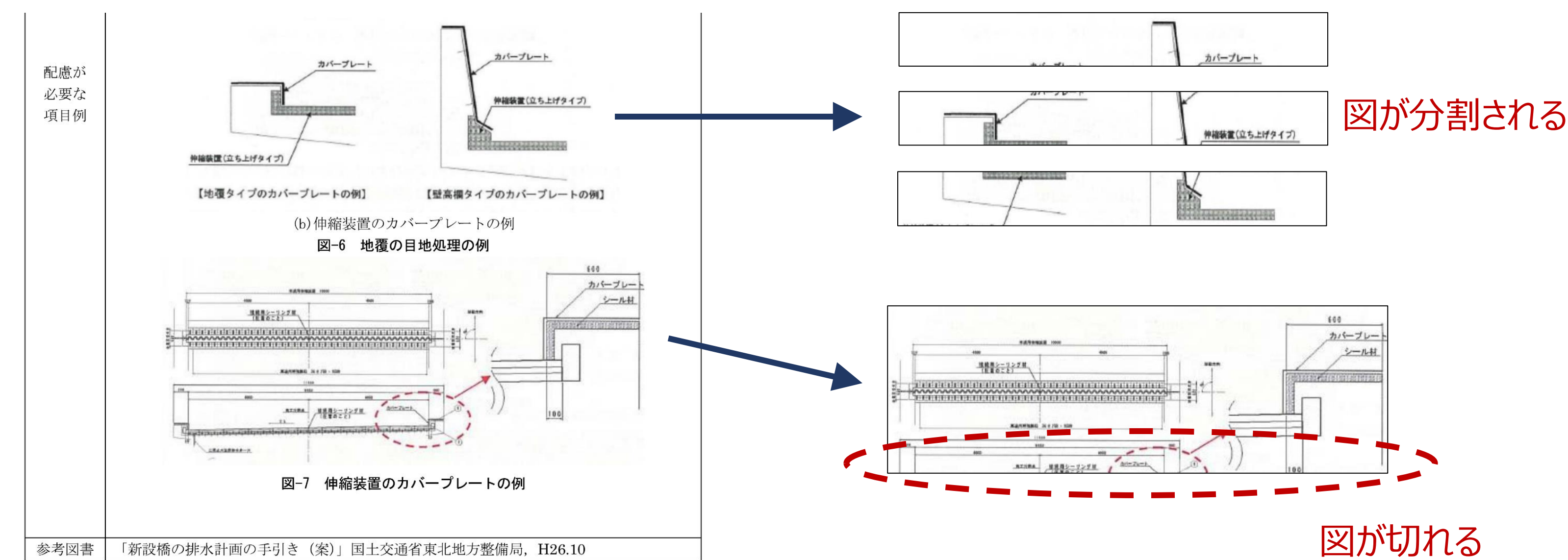
数式への対応や参照すべき文書の分散 (マルチホップ推論), チャンク分割の問題など

文書フォーマットの問題

課題: レイアウト検出, 図表の処理, マルチモーダルモデルの活用検討

今回は図表への対応ができておらず、テキストのみを対象とした。
以下はPythonライブラリUnstructuredによる図検出とGPT-4oの表認識を試行した結果。

Unstructuredの図検出



GPT-4oの表認識

表-3.4.3 工費・工期・社会に与える影響、構造物の状態と、影響度の関係の目安

影響度	工期・費用・社会に与える影響	構造物の状態
大	<ul style="list-style-type: none"> 道路機能回復までに長期間を要し、機能回復に多額の費用が必要となる状態。 人命や財産を脅かす状態。 	<ul style="list-style-type: none"> 構造物の修復に長期間を要するような変状や、場合によっては撤去し再構築しなければならぬ状態。
小	<ul style="list-style-type: none"> 道路機能回復までに比較的短期間で、機能回復に所要の費用が必要となる状態。 人命や財産を脅かす可能性がある状態。 道路機能が確保される状態。 	<ul style="list-style-type: none"> 損傷が限定的で橋としての機能回復が速やかに行い得る状態で、応急および恒久復旧で対応できる状態。 修復を必要としないが、軽微な修復でよい変状が生じている状態。

回答に関連 [国土交通省, 2021]

Q: 影響度が小さい場合の構造物の状態の目安は?

A (GPT-4o): 影響度が小さい場合の構造物の状態の目安は、以下の通りです:
 ・損傷が限定的で縮むとしての機能回復が速やかに行い得る状態: 応急および復旧で対応できる状態。
 ・修復を必要としないが、軽微な修復でより変状が生じている状態。
 これらの状態は、構造物が比較的軽微な損傷を受けているものの、迅速な回復が可能であることを示しています。
 ※誤り箇所

3. 検索精度向上へ向けた検討

- TF-IDFやBM25は用語の意味を捉えられない問題 [Formal+, 2021]
- NN手法はsemanticな検索が可能だが低頻度語に弱い [Scivolino+, 2021]
- ドメイン適応には学習する方法が良い [飯田・岡崎, 2023]
- 一方で土木分野においてはデータが整備されていない
- ⇒学習無しの方法を検討

アプローチ

- クエリの専門用語を活用するため、query vectorとchunk vectorをマージ
- IDFスコアを使用したルーティング (BM25/NN手法の振り分け)

データセット

橋梁設計分野 [国土交通省, 2021]の基準書から作成されたデータセット [緒方+, 2024]を使用

Question	Answer
鋼橋の種類を教えてください	鋼橋には、桁橋、トラス橋、アーチ橋、ラーメン橋、斜張橋、および吊橋等がある
鋼桁橋の構造上の特徴を教えてください	<ul style="list-style-type: none"> 鋼桁橋の主桁は、充腹のI形断面、n形断面及び箱形断面を基本とする。 床版は、鋼床版、コンクリート系床版がある。

結果

いずれのアプローチも性能向上に繋がらない結果となった

Recall@k	BM25 (1,2-gram)	MiniLM	OpenAI-embedding -3-large	E5	Approach ①	Approach ②
1	0.16	0.20	0.54	0.50	0.48	0.46
2	0.30	0.28	0.74	0.58	0.58	0.58
3	0.32	0.32	0.78	0.64	0.62	0.64
5	0.44	0.38	0.84	0.68	0.70	0.68
10	0.60	0.40	0.92	0.88	0.90	0.90
20	0.70	0.46	0.92	0.98	0.96	0.96
30	0.72	0.50	0.92	1.00	1.00	0.98
50	0.88	0.54	0.96	1.00	1.00	0.98

4. 今後の方針

- 土木分野の学習・評価用データセット作成
- 検索性能向上へ向けた土木辞書/知識グラフ作成
- 精度向上へ向けたEmbedding model / LLMのチューニング