特集≫≫　建設 DX と生成 AI

建設分野における言語モデルの自動評価に向けた研究

藤　井　純一郎・緒　方　　陸

大規模言語モデル（Large Language Model, 以下 LLM）の発展を受けて，建設分野においても LLM を活用した業務効率化が期待されている。活用に当たっては業務要件に基づき定量評価を行った上で適切な LLM を選択することが重要となるため，LLM の自動評価手法の確立が求められる。

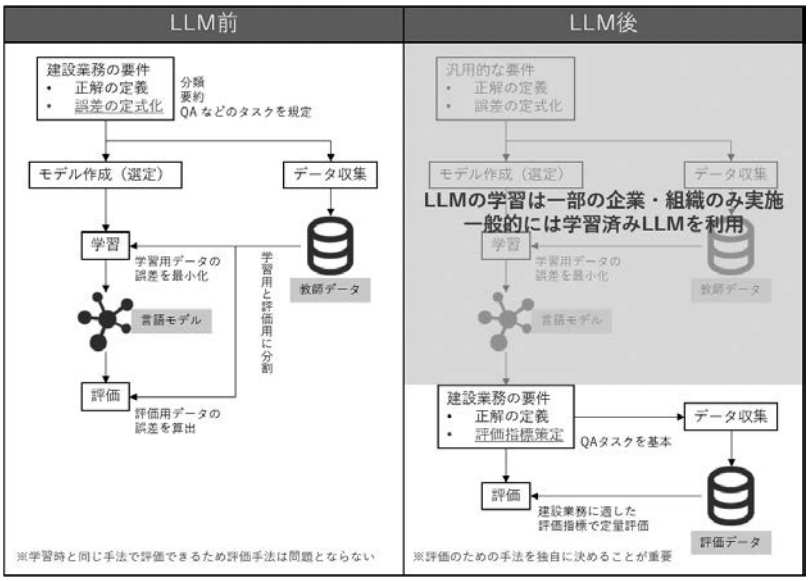
しかし LLM で生成した文章の自動評価には課題が残っており，特に建設分野の専門知識を含む文章の自動評価は，まだ黎明期にある。本研究は，建設分野において専門知識の文脈を考慮して文章を生成できるかを測る評価手法の確立に向けた初歩的な研究である。QA タスクにおいて，本研究で提案する手法も含めて複数の自動評価手法について人手評価と比較し，自動評価の現状と今後の展望について述べる。

キーワード：自然言語処理，大規模言語モデル，文章生成，自動評価

1. はじめに

ChatGPT をはじめとして，LLM の発展が目覚ましい。LLM は特定のドメインのテキストでなく，web 等から収集した膨大なラベルなしテキストを学習した汎用モデルであり，自然な文章を生成できる。そのため，メール自動作成，要約作成，チャットボットなど，さまざまな応用が進んでいる。LLM 登場以前は，タスクに応じて教師データを作成しタスクに特化したモデルの学習やファインチューニングを行うことが一般的であったが，特に ChatGPT 登場後は LLM 自体は変更せずにプロンプトエンジニアリングや Retrieval-Augmented Generation などにより LLM の能力を活用する研究や事例が多くなった。すなわち，言語モデルを「どう作るか」から「どう使うか」にパラダイムが大きく変わったと感じている。

業務に自然言語処理を適用することを考えると，従来はタスクに適用させる段階で教師データを作成するため，それらのデータの一部を用いることで精度評価が可能であった。一方で LLM を業務に適用する場合，エンドユーザーは教師データを保有していないため，精度評価を行うための手法やデータはユーザーが用意



図―1　言語モデルを業務適用するフローのイメージ（LLM 前後で評価手法の位置づけが大きく異なる）

する必要がある。従ってLLMのパラダイムにおいては，言語モデルや教師データよりも，（特にLLMを活用する上で一般的なQAタスクでの）評価指標や評価データがより重要になると考える（**図―1**）。

ただし，QAタスクで生成された回答は正解が一意に決まるものではないため，定量評価は簡単ではない。それが建設分野の専門知識を含む場合は尚更である。

建設分野における自然言語処理活用研究例は増加している[1)～5)]が，いかに建設分野へドメイン適応させるかに焦点を当てた研究が多く，言語モデルの文章生成能力の評価には重きが置かれていない。また，建設用語を含む文章の正確性を評価する手法の確立やデータセットの整備は課題としても指摘されている[2)]。評価手法の一つとして，人手評価はよく用いられるが，時間的・金銭的コストが高い。そのため，研究・開発の加速へ向けては自動評価手法が求められる。自動評価のためのデータや指標が確立すれば，新たなLLMが公表された際に自動で評価を行い，最適な言語モデルを選択する助けとなるであろう。

本研究は，建設分野において，言語モデルが適切に文章を理解し，専門知識の文脈を考慮して文章を生成できるかを測る評価手法の確立に向けた初歩的な研究である。QAタスクにおいて，本研究で提案する手法も含めて複数の自動評価手法について人手評価と比較し，自動評価の現状と展望について論じる。

2. 自動評価手法

(1) LLMを用いない評価手法

QAタスクに適用可能な自動評価手法として，以下が挙げられる。誌面の都合上，手法の詳細は参考文献を参照されたい。ただし，建設分野の専門的な文章に適用する場合，既存研究では類義語の判定が困難なことや，専門用語を多く含む文章の評価が難しいという問題が指摘されている。

・BLEU[6)]

単語の一致度に着目し，生成文章の単語のうちどの程度が正解に含まれるか評価する指標

・ROUGE[7)]

単語の一致度に着目し，正解の単語のうちどの程度が生成文章に含まれるか評価する指標

・BERTScore[8)]

単語そのものでなくBERTの埋め込み表現を利用して，正解と生成文章の類似度を評価する指標

(2) LLMを用いた自動評価

近年では自動評価手法として大規模言語モデル（LLM）を用いた手法の研究が進められている。発展途上の分野であるため，標準的な手法は確立していないが，LLMの進展に伴い（1）で挙げた問題を解決できる可能性がある。建設分野の少量のQAデータセットを用いて自動評価を行うことを想定すると，学習不要かつ正解データを参照して評価を行える手法が望ましい。このような手法としてはG-Eval[9)]と呼ばれる手法が挙げられる。ただし，G-EvalもLLMが生成した回答をより好む傾向があることや，より長く冗長な回答を好むといった問題が指摘されている点には留意が必要である。

・G-Eval

GPT-3.5やGPT-4の出力スコアの確率を使用し，それらの加重合計を最終結果として得る方法

3. 自動評価手法の比較実験

(1) 概要

現時点では建設分野の専門知識を含む文書の評価は，人手評価が最も信頼度が高いという前提に立ち，人手評価を正解とした。その上で各種自動評価指標がどの程度人手評価に近いかを比較する実験を行った。なお，実験内容は筆者らの先行研究[4)]から一部抜粋したものである。

(2) データセット

評価用データセット作成にあたり，国土技術政策総合研究所の資料[10)]を用いた。当資料は橋梁設計分野の実務においても使用され，橋梁計画における基本条件設定やリスク評価について記載された資料である。

表―1 QAデータセット サンプル

Question	Answer
鋼橋の種類を教えて下さい	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下さい	・鋼桁橋の主桁は，充腹のＩ形断面，π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。
鋼橋でかつコンクリート系床版を有する桁橋の構造上の特徴を教えて下さい	・コンクリート系床版を有する桁橋は，鋼の主桁と，床版を接合して桁とした構造。 ・鋼主桁は，充腹のＩ形断面，π形断面及び箱形断面を基本とする。 ・コンクリート系床版には，RC床版，鋼コンクリート合成床版，PC床版などがある。

この資料から人手により，自然言語処理分野で一般的なタスクである Closed-book QA タスクのデータセット（全50件）を作成した。参考として例を**表―1**に示す。なお，作成においては資料の文章から抜き出したものを正解の回答とし，LLM 自体が知識を保有（学習してパラメータに暗黙的に保持）していれば回答できる内容となっている。また前提として，ユーザーが Question を入力した際に欲しい回答として Answer を想定している。作成したデータセットは J-STAGE Data に公開した[11]。

(3) 実験方法

文章生成モデルには Llama 2 -Chat（7B）[12]，GPT-3.5-turbo[13], GPT-4[14], PaLM 2（Bison）[15] を使用した。また入出力は日本語とし，以下をプロンプトとして与えた。なお，{q} は QA データセットの質問を表す。

```
### 指示：
橋梁設計技術者として，以下の質問に“日本語で”回答してください。
### 制約：
- 単語や文章を3回以上繰り返す出力は禁止します。
- 日本語以外の言語の出力は禁止します。
質問：{q}
### 回答：
```

上記プロンプトで生成したテキストを用い，次節で述べる人手評価と各種自動評価手法を比較した。両結果の比較には Spearman の順位相関係数を用い，どの自動評価手法が人手評価に近いかを評価した。なお，回答の生成は全てのモデルで一度のみ行った。

(4) 評価方法

(a) 人手評価

人手評価は半自動のキーワード評価を組み合わせる形で実施した。自動でキーワード評価を実施する場合は専門用語の抽出が問題となるため，まず著者が人手で正解の回答からキーワードを抽出することでこの問題に対処した。次に，抽出したキーワードが生成した回答にどの程度含まれるかを算出し，その割合を keyword score とした。なお，正解および生成した回答に含まれるキーワードはユニークなものを用いた。キーワード評価後，土木工学系大学院生 4 名によりモデルが生成した文章を 5 段階で評価した。評価基準は**表―2**に示す。4 名の作業者間の評価結果に相関があることを確認し，これらの平均を正解スコアとした。

(b) 自動評価

本研究では既往手法として，2 章に挙げた自動評価手法を用いる。LLM を用いない手法としては BLEU-4, ROUGE-1 / ROUGE-2 / ROUGE-L，BERTScore の F1 値を算出した。LLM を用いる手法の G-Eval については，原著論文では Coherence, Consistency, Fluency, Relevance の 4 種類の値を提案しているが。本研究では「言語モデルが適切に文章を理解し，土木分野の文脈を考慮して文章を生成できるか」を測るため，生成されたテキストの内容を評価する Relevance のみを採用した。

既往手法に加え，本研究では G-Eval をベースに下

表―2　人手評価基準

SCORE	評価基準
1	キーワードを含まず（keyword score < 0.3），回答の大半（>=50%）は誤り，または致命的な（橋梁設計業務に支障をきたす可能性のある）誤りがある，もしくは不要な文章（質問文や単語の繰り返しなど）を含む
2	キーワードを含まず（keyword score < 0.3），回答の一部（<50%）は誤り
3	キーワードを含まない（keyword score < 0.3）が，誤りではない
4	キーワードを一部（keyword score < 0.5）を含み，誤りではない回答
5	キーワードの大半（keyword score >= 0.5）を含み，正解と同じ意味の回答

表―3　実験対象の評価手法

	手法	プロンプト	評価モデル
既往手法	BLEU	−	−
	ROUGE-1	−	−
	ROUGE-2	−	−
	ROUGE-L	−	−
	BERTScore	−	−
	G-Eval	原著論文のものを和訳	GPT-4
提案手法	提案 -gpt4	下記	GPT-4
	提案 -gemi	下記	Gemini-Pro

あなたには質問と正解，およびモデルが生成した回答の一つのセットが与えられます。
あなたの課題は，モデルが生成した回答を 1 つの指標で評価することです。
以下の指示をよく読み，理解してください。レビュー中はこの文書を開いておき，必要に応じて参照してください。
評価基準：
スコア（1-5）- モデルが生成した回答が正解に，意味的に類似していること。モデルが生成した回答には，正解の土木工学的な観点で重要なキーワードを含めること。モデルが生成した回答が正解のキーワードを含んでいなければ，スコアは 1 となる。モデルが生成した回答が正解のキーワードを多く含み，誤りを含まなければスコアは 5 となる。評価者は，質問文や英語の文章，誤りなど不要な情報を含む回答にはペナルティを課すよう指示されている。
評価ステップ：
1. 正解とモデルが生成した回答を形態素解析し，キーワードを特定する。
2. 特定したキーワードの重なり具合を評価する。
3. モデルが生成した回答が正解に土木工学的観点で意味的にどの程度類似しているかを評価する。
4. ステップ 2 およびステップ 3 の評価をもとに，評価基準に従いスコアを 1 から 5 の間で割り当て，スコアの根拠も特定する。
5. ステップ 4 で割り当てたスコアを float 型の数値を回答する。
質問：
{{Question}}
正解：
{{Reference}}
モデルが生成した回答：
{{text}}
評価様式（スコアのみ）：
- スコア：

記の改善を試み，**表―３**に示す評価手法を実験した。
・**表―２**に示す基準を踏まえてプロンプトを調整
・評価モデルを GPT-4 から Gemini-Pro に変更

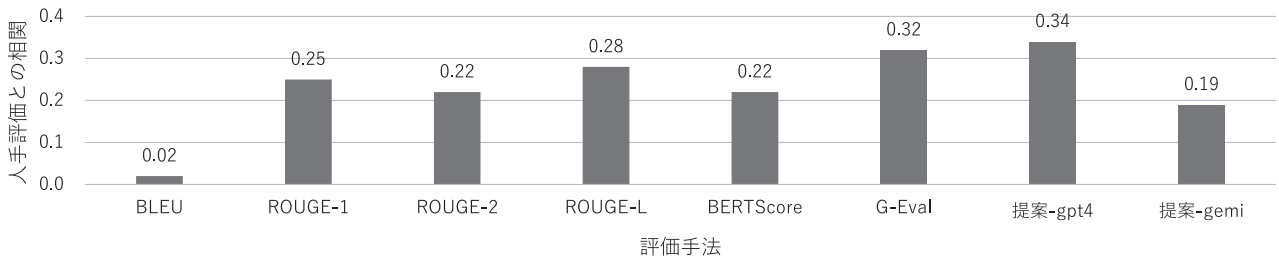
4. 実験結果と考察

実験対象とした 8 種類の自動評価指標の人手評価との相関の平均値を**図―２**に示す。全体として，各種自動評価結果の人手評価との類似度は低く，最も高いスコアは提案-gpt4 の 0.34 に留まる。従って現状では，いずれの自動評価手法も人手の評価には及ばない。その中で LLM を用いた評価手法，特に GPT-4 を用いた手法（G-Eval，提案 -gpt4）が比較的人手評価との相関が高かった。一方で評価モデルを GPT-4 よりもパラメータ数の少ない Gemini に変更するとスコアが下がる結果となった。このことから，自動評価に用いる LLM の性能が上がれば，人手評価に近づく可能性があると考えられる。

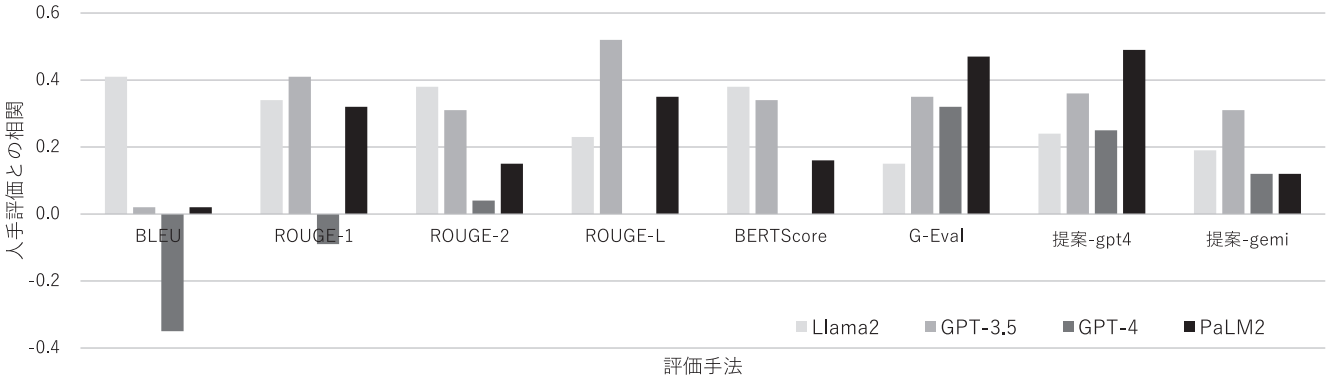
また，回答生成に用いた Llama 2，GPT-3.5，GPT-4，PaLM 2 ごとの違いを見ると，ベストの評価指標がすべて異なる結果となった（**図―３**）。GPT-4 を用いた手法も Llama2 の回答に対しては他手法に比べてスコアが劣っている。従って現時点で必ずしも GPT-4 を用いた手法が優位とは言えない。これは LLM により，文章の長短や冗長な繰り返しの有無，文体の違いなど，生成される回答の特徴が異なるが，自動評価ではそれらの影響を受けてしまうことが理由と考えている。

5. おわりに

本稿では，建設分野の QA タスクにおいて，既往の自動評価手法と本研究で提案する自動評価手法につ



図―２　評価手法別の人手評価との相関



図―３　回答別・評価手法別の人手評価との相関

いて人手評価と比較し，どの手法が土木分野において適切な評価が可能かを検討した。現状では人手評価に対して自動評価の信頼度は劣るため，すぐには自動評価を実用することは難しいと判断している。

その中で LLM を用いた自動評価手法は相対的に人手評価との相関が高く，特に GPT-4 を用いた手法はスコアが高かったことから，自動評価に LLM を活用することの有効性が示唆される。今回の結果から，将来的に LLM 自体の性能が上がることで，自動評価の信頼度も上がることが期待できる。

建設分野での自然言語処理の活用はまだ緒に就いたばかりである。今後，活用を加速していくためには，LLM の自動評価手法を確立し，業務に応じて適切な LLM を定量的に評価できることが重要と考える。そのため，今後も継続して自動評価手法の研究を進める所存である。

JCMA

《参考文献》

1）箱石健太，一言正之，菅田大輔．土木分野における事前学習モデル BERT による精度検証．土木学会論文集特集号（土木情報学），79 巻 22 号，22-22042，2023.
2）藤井純一郎，大久保順一，緒方陸，天方匡純．LLM を土木分野に適応するための基礎的研究，pp.779-785，AI･データサイエンス論文集，4 巻 3 号，2023.
3）菅田大輔，箱石健太，一言正之．土木・建設分野における大規模言語モデルの利活用に向けた検証と考察，pp.670-676，AI・データサイエンス論文集，4 巻 3 号，2023.
4）緒方陸，大久保順一，藤井純一郎，天方匡純．土木分野における LLM を用いた言語モデル評価手法の提案，pp.2079-2084，言語処理学会第 30 回年次大会発表論文集，2024.
5）緒方 陸，大久保 順一，藤井 純一郎，天方 匡純．土木分野における言語モデル評価指標の検討，pp.66-76，AI･データサイエンス論文集，5 巻 1 号，2024.
6）Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU：A method for automatic evaluation of machine ranslation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318, 2002.
7）Chin-Yew Lin. ROUGE：A package for automatic evaluation of summaries. In Proceedings of the ACL Workshop：Text Summarization Branches Out, pp. 74–81, 2004.
8）Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore：Evaluating text generation with BERT. In Proceedings of the Eighth International Conference on Learning Representations, 2020.
9）Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, Chenguang Zhu. G-Eval：NLG Evaluation using Gpt-4 with Better Human Alignment. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 2511–2522, Singapore. Association for Computational Linguistics, 2023.
10）国土技術政策総合研究所国土交通省．道路橋の設計における諸課題に関わる調査（2018-2019）．国土技術政策総合研究所資料，No.1162，2021.
11）Question Answering（QA）for bridge design.（オンライン）（引用日：2024.6.5.）https://doi.org/10.50915/data.jsceiii.25459144.v2
12）Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, Thomas Scialom. Llama 2：Open foundation and fine-tuned chat models. arXiv preprint arXiv：2307.09288, 2023.
13）OpenAI. GPT-3.5.（オンライン）（引用日：2024.1.11.）https：//platform.openai.com/docs/models/gpt-3-5.
14）OpenAI. GPT-4 Technical Report. arXiv preprint, arXiv：2303.08774v3, 2023.
15）Google. PaLM 2 Technical Report. arXiv preprint arXiv：2305.10403, 2023.

［筆者紹介］

藤井　純一郎（ふじい　じゅんいちろう）
八千代エンジニヤリング㈱
技術創発研究所　AI 解析研究室
室長

緒方　陸（おがた　りく）
八千代エンジニヤリング㈱
技術創発研究所　AI 解析研究室
主任研究員