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

緒方 陸[1] 大久保 順一[1] 藤井 純一郎[1] 天方 匡純[1]
[1]八千代エンジニヤリング株式会社
{rk-ogata, jn-okubo, jn-fujii, amakata}@yachiyo-eng.co.jp

概要

土木分野においても自然言語処理技術への期待は大きく，近年実用へ向けた検討が増加している．この技術が充分に機能するためには，文脈を考慮した土木専門用語の言語モデルによる理解が必要であり，その適切な評価が求められる．しかし，既往研究はいかに土木分野へ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．そこで本研究では，土木分野における言語モデルの性能評価のため，評価の際に一度 LLM で要約することで評価を行う自動評価手法を提案する．語彙数が多く回答長が長い場合には，既往手法に比較して本手法が有効に働き，より人手評価に近い結果となることを示した．

1 はじめに

近年，土木分野において従来紙により管理されていた諸元や巡視･点検情報がデジタル化されている．これに伴い，土木技術者の作業効率化・省力化へ向けた自然言語処理活用研究例も増加している[1,2,3,4,5]．一方で，いかに土木分野へドメイン適応させるかに焦点が当てた研究が多く，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，土木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の確立が必要となる．

本研究では，土木分野における言語モデルの性能評価のため，言語モデルが適切に文章を理解し，土木工学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を目指す．QA タスクにおいて，既往評価手法と提案手法を人手評価と比較し，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．

2 既往研究

2.1 ドメイン適応

自然言語生成(Natural Language Generation; NLG)タスクの評価手法として，BLEU [6]や ROUGE [7], BERTScore [8]などがあるが，類義語の判定ができない，専門用語の評価が困難などの問題がある．

土木分野における自動評価例としては，技術文書の分類精度を評価した例[1,3]やキーワードの重なり具合を測る Keyword Intersection，文章生成の流暢さを測る Perplexity を適用したものがある[2]．しかし分類精度は言語モデルの文章生成能力を測るものとしては適さない．また，Keyword Intersection, Perplexity は次の問題がある．日本語の場合，Keyword Intersection による評価を行う場合は形態素解析の必要がある．その際専門用語は辞書登録されておらず，適切に評価できない可能性がある．この辞書整備の課題については箱石ら[1]も指摘している．Perplexity は，文章生成における流暢さを測る指標であり，単語の発生確率から算出するため，藤井ら[2]の指摘の通り専門的な内容を評価するには向かない．

土木以外の特定ドメインにおいては，BERT をファインチューニングすることで特定ドメイン[9,10,11]に適応した例がある．また近年では，ドメイン特化の大規模コーパスで LLM を学習させることで，ドメイン適応した例[12,13]も存在する．これらの例では分類，固有表現抽出（NER），QA タスクなどの精度（Accuracy / Precision / Recall / F1 score）や単語間ベクトル類似度，人手評価などで特定ドメインの性能を評価しているが，ベンチマークデータセットによる評価であり，データ整備が必要である．

2.2 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）を用いた方法も存在する．Wang らはハイパーパラメータの自動チューニングを目的に，パラ

メータサイズ 7B 程度のモデルを学習させることで GPT-3.5 や GPT-4 に近い精度の評価能力を持たせることを可能にした[14]．また，笠原ら[15]も日本語のタスクにおいて, LoRA などによる LLM チューニング手法により，既往の自動評価手法と同等以上の精度を示している．しかしこれらは学習データ整備や学習のコストが発生する．現状データ整備が進んでいない土木分野においては学習不要な自動評価手法が望ましい．

学習不要な自動評価手法として, GPTScore が提案されている[16]．GPTScore は，GPT-3.5 などの LLM が，入力に従って高品質のテキストを生成する確率が高いという前提のもと，条件付き生成確率でスコアを計算する方法である. また Liu ら[17]は, 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を提案した．ただし同著者ら[17]は LLM は LLM が生成した回答をより好む傾向があるなど問題を指摘している．そのほか WebUI 実装の必要はあるが，Web 上でユーザーが入力した質問に対し，複数の LLM の回答を表示させ，より好ましい回答をユーザーが選択することで評価を行うプラットフォームもある[18].

ここで Zheng ら[18]は，LLM には潜在的な限界があるとして，(1) Position bias；選択肢が先のものを好むなど回答の順序に影響を受けるバイアス，(2) Verbosity bias；明確でなく，品質が低く，正確でないとしても，より長く冗長な回答を好むバイアス，(3) Self-enhancement bias；自分で作成した回答を支持するバイアス，を指摘している．同様に Saito ら[19]も Verbosity bias について調査し，人手評価が短い回答を高く評価している場合, LLM の人手評価との一致度は低くなる傾向を確認している．

以上のように，バイアスにも留意しつつ，LLM を用いた評価を行う必要がある．著者らが確認した限りでは土木分野に上記手法を適用した例は無く，その活用が期待される．

2.3 本研究の目的

LLM は大規模なコーパスで学習しており, 土木分野も一部学習している．また，LLM は N-gram ベースの手法のようにキーワードのみではなく，長いコンテキストを踏まえて評価可能と考えた．

本研究では, データ整備が進んでいない状況下で, 土木分野においてより適切な自動評価手法の確立を目指し，LLM を用いた評価手法を提案する．

3 実験方法

3.1 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[20]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．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．参考として例を付表 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．

3.2 実験方法

文章生成モデルには Llama 2 -Chat (7B) [21], GPT-3.5-turbo [22]，GPT-4 [23]，PaLM 2 (Bison) [24] 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．

3.3 評価方法

3.3.1 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. キーワード評価は次の方法で行う．

表 1　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）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

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 1 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．

3.3.2 自動評価

<u>既往手法</u>

既往自動評価手法として，BLEU-4，ROUGE-1 / ROUGE-2 / ROUGE-L，BERTScore の F1 値を算出した．ここで，Zheng ら[18]は LLM を用いた評価手法として次の 3 つを挙げている．(1) Pairwise comparison; LLM に 1 つの質問と 2 つの回答を提示し, どちらが優れているかを判定する方法 (2) Single answer grading; LLM に 1 つの回答を渡し，直接スコアを割り当てる方法 (3) Reference-guided grading; LLM に reference を提供し，優れている回答を判定またはスコアリングする方法．本研究では QA データセットを作成し reference を使用可能なため，(3) Reference-guided grading の方法として G-Eval [17]を採用した．G-Eval について，原著論文[17]では要約タスクで Coherence, Consistency, Fluency, Relevance を採用している．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．なお，プロンプトは Liu ら[17]のものを和訳して使用した．

<u>提案手法</u>

本稿では Liu らの G-Eval [17]をベースに表 1 に示す基準を踏まえてプロンプトを調整した．参考として，使用したプロンプトは付図 1 に示す．

また, 2.2 節で述べたバイアスへの対応も検討する. まず，Self-enhancement bias への対応として，今回文章生成に使用していない Google 社の Gemini-Pro[25]を評価モデルに使用することで Self-enhancement bias の影響を軽減できると考えた．さらに提案手法として，Verbosity bias 軽減を目的とし，生成した文章を要約した上で評価を行う方法を採用した（Proposed-gpt4 / gemi）. なお Position bias について，本稿ではモデル同士の相対評価は実施しないため，Position bias の影響は受けないと考えている．

今回採用した LLM を用いた評価手法一覧を表 2 に示す．プロンプト，要約・評価モデルの違いにより全 5 ケースとなる．既往手法との比較およびバイアスへの対応の貢献は，結果とともに次節で示す．

表 2　LLM を用いた評価手法

手法	プロンプト	要約モデル	評価モデル
G-Eval	[17]を和訳	－	GPT-4
G-Eval-gpt4	付図 1	－	GPT-4
G-Eval-gemi	付図 1	－	Gemini-Pro
Proposed-gpt4	付図 1	Gemini-Pro	GPT-4
Proposed-gemi	付図 1	Gemini-Pro	Gemini-Pro

4 結果と考察

表 3 に結果を示す．上段は既往手法，下段は今回新たにプロンプト等を検討した提案手法である．表中の全体評価は全モデルのスコアをまとめて評価した結果，平均は各モデルの結果の算術平均である．

結果から，全体評価は既往手法のうち BERTScore が最も人手評価に近い結果を示した．ただし後述する今回の結果からは，各モデルは異なる特性を持つと考えられ，モデルを跨いだ全体評価は適切な評価指標になり得ないと推察した．よって，以降は各モデルの評価結果について考察する．

G-Eval および G-Eval-gpt4 は，Llama2 を除いて BERTScore を超えるスコアを発揮した. Llama2 で相関がやや低くなった原因として，G-Eval は関連性の

表 3　実験結果　人手評価と各手法の相関（太字：最高スコア，下線：次点スコア）

手法	Llama2	GPT-3.5	GPT-4	PaLM2	全体評価	平均
BLEU	**0.41**	0.02	-0.35	0.02	0.15	0.02
ROUGE-1	0.34	0.41	-0.09	0.32	0.20	0.25
ROUGE-2	0.38	0.31	0.04	0.15	0.20	0.22
ROUGE-L	0.23	**0.52**	0.00	0.35	0.13	0.28
BERTScore	0.38	0.34	0.00	0.16	**0.38**	0.22
G-Eval	0.15	0.35	0.32	0.47	-0.04	0.32
G-Eval-gpt4	0.24	0.36	0.25	**0.49**	0.05	**0.34**
G-Eval-gemi	0.19	0.31	0.12	0.12	0.12	0.19
Proposed-gpt4	0.02	0.48	**0.41**	0.39	0.02	0.32
Proposed-gemi	0.23	0.29	0.34	0.09	0.04	0.24

強い重要な情報のみで評価しており Llama2 は全体的に低評価となったこと，G-Eval-gpt4 は冗長な回答でも高評価とする傾向があることが挙げられる．また, G-Eval-gemi は G-Eval および G-Eval-gpt4 と比較して同等以下のスコアであった．これらは評価モデルの性能の違いによるものと考えた．Gemini-Pro の性能は多くのテキストベンチマークで GPT-4 には劣っている[25]. よって G-Eval-gemi が低いスコアとなった原因は，Gemini-Pro よりも GPT-4 の方が評価能力が高く, Self-enhancement bias の影響よりもこの差の影響が大きいためと推察した．この傾向は Proposed-gpt4/gemi でも確認できる．

Proposed-gpt4 の GPT-4 の評価は，今回採用した手法の中で最も人手評価との相関が高い結果となった．ここで，表 4 に各モデルが生成した回答の語彙数および回答の平均長さを示す．なお語彙の抽出は，出力として日本語を想定していること，および英語の出力は主に Llama2 のものであり，かつ質問文などの繰り返しが大半であることから，日本語のみを対象としている．表より，Llama2，GPT-3.5，PaLM2 は同程度の語彙数であるのに対し，GPT-4 は回答長が長く，他モデルの 2 倍程度の語彙を持つ．この結果から，語彙数が少ない場合には N-gram ベースの手法などの既往手法である程度の評価が可能だが，今回の GPT-4 の結果のように語彙数が多い場合には既往手法では評価が困難であると考えた．加えて，Proposed-gpt4 が GPT-4 の生成した回答の評価において比較的人手に近い評価となったことから，語彙数が多く回答長が長い場合には回答を要約後に評価する方法が有効に働き，Verbosity bias を軽減できる可能性を示した。一方で Llama2 の評価においては，要約により，英語による質問文繰り返しなど明らかに不要な表現が除かれたことで人手評価との相関が低くなったと推察した．Proposed-gemi に着目すると，要約により GPT-4 の評価スコアは向上しているものの，その他モデルの評価は G-Eval-gemi と同程度となった．この結果から，LLM を用いた評価と人手評価との相関は評価モデルに依存すると考えた．

表 4　各モデルが生成した回答の語彙数と回答平均長さ

モデル	語彙数	回答平均長さ
Llama2	574	630
GPT-3.5	650	205
GPT-4	1249	605
PaLM2	591	253

5 おわりに

本稿では，土木分野の QA タスクにおいて，既往の評価手法と本研究で提案する評価手法につ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．提案手法による評価は既往手法と同等以上のスコアを発揮し，特に語彙数が多く回答長が長い場合においては，生成された回答を要約してから評価することで Verbosity bias を軽減できる可能性を示唆した．

本検討に関して留意すべき点として二点挙げる．まずデータ不足について，今回用いたデータセットは 50 件と小規模であり, このバイアスが結果に表れた可能性がある．土木業界全体としてデータ整備が進んでいないため，業界全体で取り組むべき課題であると考える．また，今回プロンプトは手動で作成した．一方で近年では自動で動的にプロンプトを生成する方法もあり[26]，これにより精度が向上する可能性がある. LLM による評価はプロンプトに敏感であり[17]，この検討も必要であると認識している．

参考文献

[1] 箱石健太，一言正之，萱田大輔．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] 萱田大輔，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] 青島亘佐，宮内芳維．大規模言語モデルの活用による橋梁点検調書作成の省力化に関する検討, pp.274-284, AI･データサイエンス論文集, 4 巻 3 号, 2023.

[5] 稲富翔伍，山根達郎，金崎裕之，全邦釘．大規模言語モデルと画像セグメンテーションによる専門知識融合型土砂災害危険性判断手法，pp.507-514，AI・データサイエンス論文集，4 巻 3 号，2023.

[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] Iz Beltagy, Kyle Lo, Arman Cohan. SciBERT: A pre-trained language model for scientic text. In Pro-ceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th Internation-al Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019.

[10] Kexin Huang, Jaan Altosaar, Rajesh Ranganath. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. arXiv preprint arXiv:1904.05342, 2019.

[11] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinform., Vol. 36, No. 4, pp. 1234–1240, 2020.

[12] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, Gideon Mann. BloombergGPT: A large language model for finance. arXiv preprint arXiv:2303.17564, 2023.

[13] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, Vivek Natarajan. Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617, 2023.

[14] Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, Yue Zhang. PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization. arXiv preprint arXiv:2306.05087, 2023.

[15] 笠原智仁，河原大輔，山崎天，新里顕大，佐藤敏紀. Decoder ベースの大規模言語モデルに基づくテキスト生成の自動評価指標．言語処理学会 第 29 回年次大会, pp.1940-1945, 2023.

[16] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, Pengfei Liu. GPTScore: Evaluate as You Desire. arXiv preprint arXiv:2302.04166, 2023.

[17] 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.

[18] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv:2306.05685, 2023.

[19] Keita Saito, Akifumi Wachi, Koki Wataoka, Youhei Akimoto. Verbosity Bias in Preference Labeling by Large Language Models. arXiv preprint arXiv:2310.10076, 2023.

[20] 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.

[21] 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.

[22] OpenAI. GPT-3.5. (オンライン) (引用日: 2024.1.11.) https://platform.openai.com/docs/models/gpt-3-5.

[23] OpenAI. GPT-4 Technical Report. arXiv preprint, arXiv: 2303.08774v3, 2023.

[24] Google. PaLM 2 Technical Report. arXiv preprint arXiv:2305.10403., 2023.

[25] Gemini TeamGoogle. Gemini: A family of highly capable mul-timodal models. arXiv preprint arXiv:2312.11805, 2023.

[26] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, Eric Horvitz. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. arXiv preprint arXiv:2311.16452, 2023.

付録

付表 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

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:

付図 1　調整したプロンプト（紙面の都合上一部改行を割愛）