

# LLM を土木分野に適応するための基礎的研究

藤井 純一郎<sup>1</sup>・大久保 順一<sup>2</sup>・緒方 陸<sup>2</sup>・天方 匡純<sup>1</sup>

<sup>1</sup>正会員 八千代エンジニアリング株式会社 (〒111-8648 東京都台東区浅草橋 5-20-8)

E-mail: jn-fujii@yachiyo-eng.co.jp (Corresponding Author)

<sup>2</sup>非会員 八千代エンジニアリング株式会社 (〒111-8648 東京都台東区浅草橋 5-20-8)

ChatGPTをはじめとして、大規模言語モデル (Large Language Model, 以下 LLM) を用いた文章生成モデルの発展が目覚ましい。土木分野においても、LLM を活用した業務効率化が期待されるが、LLM の学習は主に Web で集めた文書を中心としているため、土木分野の専門知識に対する学習不足が予測され、正確な文章を生成できない懸念がある。そこで、土木分野において正確な文章生成を実現するための基礎的な研究として、本研究では LLM を土木ドメインに適応することを試みた。正確性の評価手法を提案した上で、LLM の事前学習済み公開モデルと、ファインチューニングによるドメイン適応を行ったモデルで、土木分野に関する文章生成の正確性を評価し、LLM を土木ドメインに適応するための課題について論じた。

**Key Words:** Natural Language Processing, Large Language Model, domain adaptation, fine tuning

## 1. はじめに

ChatGPTをはじめとして、大規模言語モデル (Large Language Model, 以下 LLM) を用いた文章生成モデル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。

土木分野においても、文章生成モデルを活用した業務効率化が期待される。一方で、LLM の学習は主に Web で集めた文書を中心としているため、土木分野の専門知識に対する学習不足が予測され、正確な文章を生成できない懸念がある。また、LLM が事実と異なる文章を生成する Hallucination と呼ばれる現象が発生することが、J<sup>1)</sup> により報告されている。しかし、土木分野においてこれらの文章生成の正確性に関する定量的な性能や改善策は明らかになっていない。

本研究では、土木分野において正確な文章生成を実現するための基礎的な研究として、まず正確性の評価手法を提案する。その上で、LLM の事前学習済み公開モデルと、ファインチューニングによるドメイン適応を行ったモデルで、土木分野に関する文章生成の正確性を評価し、LLM を土木ドメインに適応するための課題について論じる。

## 2. 先行研究

### (1) LLM のドメイン適応に関する研究

一般的に LLM では、大量のテキストで事前学習を行い、個別のタスクに適応するための下流タスクを学習する。代表的な LLM の一つである BERT<sup>2)</sup>を、医療、金融などの分野に適応させた研究がされており<sup>3)</sup>、土木分野においても、箱石ら<sup>4)</sup>が BERT の再事前学習により土木ドメインに適応した土木 BERT を提案している。BERT のドメイン適応においては、ドメイン文書を用いて事前学習を追加で実施し、専門用語がどれだけ生成されるかなどが評価されている。

一方、下流タスクをより少ないデータで学習することを目指し、GPT-1 から GPT-4 まで段階的に開発されてきた<sup>5)6)7)</sup>。GPT では基本的に事前学習は再学習せず、ファインチューニング、few-shot learning、one-shot learning、zero-shot learning により下流タスクを実現する。八幡<sup>8)</sup>らは、GPT-2 を基に、保険に関する文書を集めてファインチューニングによる QA タスクへの適応を行った。この中では、ファインチューニングを行ったパラメータの小さいモデルよりもファインチューニングを行わない、よりパラメータの大きい LLM の性能が良いことが述べられている。GPT を用いることで、下流タスク (ファインチューニング) のみで土木ドメインに適応できる可能性があるが、これを実証した研究は報告されていない。

## (2) 文章生成の正確性評価に関する研究

LLM の評価指標としては、GLUE<sup>10</sup>を用いることが一般的である。しかし GLUE には文章生成を評価するベンチマークデータセットは含まれていない。GPT は文章生成モデルであり、BERT などのように下流タスクを作成して GLUE スコアを出すなどのような、既存の LLM の評価手法による評価が難しい。文章生成のデータセットとしては LAMBADA データセット<sup>11</sup>が用いられている。ただし LAMBADA は、土木ドメインの文章生成を評価するためには適さない。

また、Hallucination の問題に対応するべく Truthful QA データセット<sup>12</sup>が公開されている。Truthful QA では各質問について Wikipedia をソースとして参照していて、質問文、回答文（正答、誤答）、ソースをデータとして公開している。Transformer 登場以後は Ji<sup>13</sup>らによって LLM を用いて作成した QA を用いて評価をするというような手法も提案されている。ただしいずれも、土木ドメイン特有の Hallucination を評価するためには適さない。

GPT による文章生成の評価に関する研究としては、Qin<sup>14</sup>らは、既存の評価タスクをプロンプトエンジニアリングで置き換えて評価を行い高い性能を示した。一方で、松林<sup>15</sup>は、専門性が要求される高等教育において、Hallucination の影響は多大であり、事実確認を行うことが困難であることを指摘している。

## 3. 評価手法

土木分野の特徴として、専門知識に加え現地の条件や自然環境といったローカル情報を反映することが不可欠である。本研究では、土木分野における正確な文章とは、「(1)専門用語やローカル情報を含み(2)事実誤認を含まない(3)自然な文章」と仮定し、以下の指標を定義した。

### (1) Keyword Intersection

GPT の評価としては、質問を生成して、回答の正確さを図る Question Answer (QA) が評価タスクとして近い。この QA の評価手法としては、専門用語やローカル情報を含む単語出現の重複率 Keyword Intersection が一般的である。重複率が高いほど的確な回答を生成できていることを示す。本研究では、評価用データの質問回答について、Ground Truth の回答とモデルが生成した回答の Keyword Intersection を以下の手順で算出した。

- ① それぞれの回答を NLTK で Tokenize
- ② NLTK を用いて各トークンを Stemming
- ③ 両回答のトークン間の IoU を式(1)により算出

$$IoU = \frac{\text{(両回答に共通するトークン数)}}{\text{(両回答の全体ユニークトークン数)}} \quad (1)$$

### (2) Hallucination

生成した文章が事実かどうかを定式化することは困難なため、Hallucination の評価手法はまだスタンダードと呼べるものが確立していない。本研究では、各モデルが生成した土木分野に関する質問回答について、文章単位で人手による事実確認を実施した。事実確認に際して、Wikipedia にある情報に等しい場合は Fact、この情報と異なっている場合は Hallucination として、Wikipedia に掲載されていない情報は Neutral として評価した。生成された質問回答の語数における Hallucination の割合を評価指標とした。評価例を付録に示す。

### (3) Perplexity

Perplexity は生成された文章の流暢さに当たる評価指標であり、土木分野の専門知識に関する評価ではないが、文章生成の精度として最も基本的な指標として採用した。モデルが単語を予測する際の確信度が高いほど Perplexity が低くなるため、Perplexity が低いと自然な文章を生成できている評価となる。

通常、個別の評価用データについて、1つの Ground Truth 文字列に対して複数モデルで Perplexity を算出して比較するが、本研究で評価対象とした GPT-3、GPT-3.5（後述）はソースコードが公開されておらず、Ground Truth に対する Perplexity が算出できない。そこで本研究では、逆に複数モデルが生成した、モデル数分の文字列に対して1つのモデル（評価対象と別モデル）で Perplexity を算出することとした。これにより GPT-3、GPT3.5 も含めて、Perplexity の比較を実現した。

Perplexity は式(2)により算出した。

$$PP(W) = \prod_{i=1}^N P(w_i)^{-\frac{1}{N}} \quad (2)$$

ここで  $w_i$  は生成された文字列、 $N$  はその語数である。本研究では  $w_i$  が生成される確率  $P(w_i)$  を評価対象とは別の gpt-neo-1.3B により算出した。また、参考として GPT-3、GPT3.5 以外は、通常の Ground Truth に対する Perplexity も算出した。

## 4. 文章生成の正確性評価実験

土木分野において LLM による文章生成の正確性を分析するため、河川分野に関する質問回答を例に実験を行った。少量のデータでの下流タスクにより正確性がどのように変化するか検証するため、下流タスクのみでドメイン適応が可能とされる GPT モデルを用いて、事前学習済み公開モデル（GPT-2、GPT-3、GPT3.5）と本研究でファインチューニングを実施したモデル（GPT-2-FT）で正確性を評価した。

表-1 対象文献

文書名	ページ数	備考
Klamath Project January 2023 Temporary Operating Procedure	9	アメリカ河川
Statement of Intent Draft Water Plan (Mary Basin)	38	オーストラリア河川
Report on the current status of the ILISU Hydroelectric Power Plant Project and HASANKEYF	9	トルコ河川
LCAP-chapter-2-Understanding-Elan-Links-Area	94	イギリス河川
The Oroville Dam 2017 Spillway Incident	60	アメリカダム

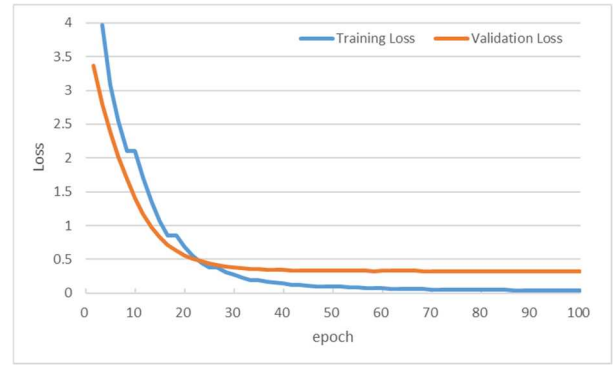


図-1 ファインチューニング学習曲線

表-2 評価対象モデル

モデル名	パラメータサイズ	備考
GPT-2 (large)	774M	
GPT-2-FT (large)	774M	Train データでファインチューニング
GPT-3 (text-davinci-300)	175B	OpenAI 社 API 利用
GPT-3.5 (turbo)	154B	OpenAI 社 API 利用

表-3 正確性評価結果一覧

モデル名	(1) Keywords intersection	(2) Hallucination 割合	(3) Perplexity gpt-neo-1.3B	(3)参考 Perplexity Ground Truth
GPT-2	10.00%	58.10%	12.06	24.47
GPT-2-FT	13.60%	73.20%	24.8	5.76
GPT-3	10.90%	24.20%	15.69	-
GPT-3.5	7.14%	11.50%	15.79	-

(1) 使用データ

河川分野に関する質問回答のファインチューニングおよび評価用データとして、表-1 に示す Web 上で公開されている海外の河川分野での専門文献を収集した。本研究は、少量のデータによる下流タスクのみで土木ドメインへの適応を試みるため、文献数は限られる。

なお、本研究はすべて英語文献を用いて実施した。英語で行った理由は、LLM の事前学習のデータ量が日本語モデルのデータ量より大幅に大きく、より専門性の高い回答が得られる可能性が高いという理由による。

収集した文献から、それぞれの河川に関する質問と回答のペアを作成し、Train(180 件) Validation(98 件) Test(96 件)に分割してデータセットを作成した。このうち Train と Validation はファインチューニングに利用し、Test は各評価指標の算出に用いた。

(2) 使用モデル

実験で用いたモデルを表-2 に示す。ファインチューニングは、ソースコードが公開されており、自前でファインチューニングが可能な GPT-2 (large) モデルでのみ実施した。GPT-2の後継である GPT-3 (text-davinci-300)、それをチャットに特化して強化学習を行った GPT-3.5 (turbo) については、本稿執筆時点でソースコードが公開されていなかったため、開発元の OpenAI 社の API 経由で文章生成を行い正確性の評価を行ったが、ファインチューニングは実施していない。

(3) GPT-2 のファインチューニング

Train データを用いて 100epoch の学習を行った。学習においては Validation データを用いて逐次 Loss を出力した。学習曲線を図-1 に示す。おおよそ 40 エポックで Loss が収束しており、ファインチューニングが正常に行えていることを確認した。

5. 実験結果

実験で得られた 3 つの評価指標の一覧を表-3 に示す。以下では各指標の評価結果と、それらをまとめた考察を示す。

(1) Keyword Intersection

表-3 から、Keyword Intersection は GPT-2-FT が最も高い結果となった。一方、GPT-3、GPT-3.5 とパラメータサイズが増えてもキーワードが増えるという傾向は見られなかった。

図-2 にキーワードの出現回数を示す。ファインチューニングを行った GPT-2-FT が、ファインチューニングを行う前の GPT-2 に比べてキーワードの出現が Ground Truth に近づいていることがわかる。GPT-3、GPT-3.5 は、どちらもキーワードの出現数が少ない。

以上より、ファインチューニングにより専門用語やローカル情報を含んだ文章が生成され、河川分野への適応が進んでいることを確認した。

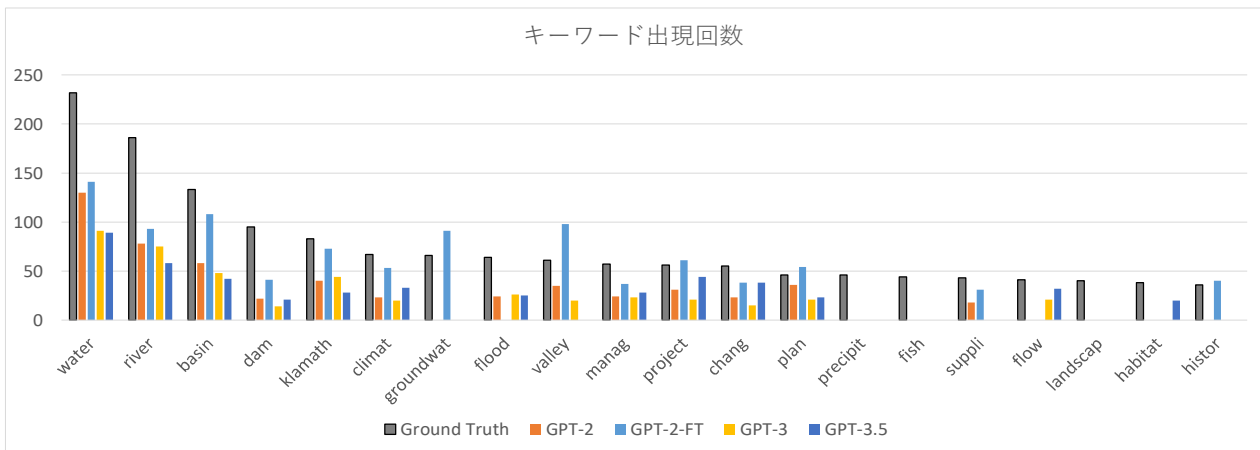


図-2 各モデルキーワード出現回数

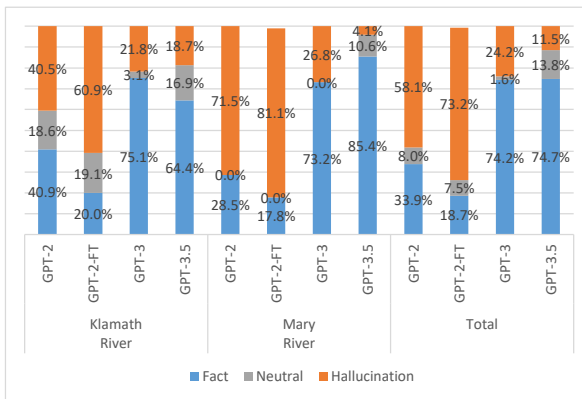


図-3 河川別 Hallucination 発生割合

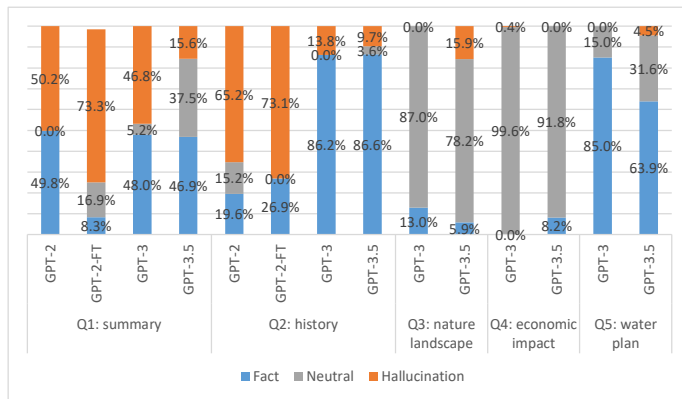


図-4 質問別 Hallucination 発生割合

## (2) Hallucination

Hallucination の評価においては、Klamath River と Mary River の2河川について、Q1 概要、Q2 歴史、Q3 地理的特徴、Q4 経済効果、Q5 水資源計画の5つの分野についての質問を行った。なお、GPT-2、GPT-2-FT については Hallucination が著しかったため、Q1、Q2 のみ評価を行い、Q3 以降の評価は行わなかった。

表-3 から、Hallucination 全体の評価としては GPT-3.5 が最もよい結果となった。

図-3に河川別の Hallucination の評価結果を示す。ここでは4つのモデルを比較するため、Q1 と Q2 の評価結果のみ示している。モデル間の比較としては、GPT-2 系と GPT-3 系では、GPT-2 系の方が Hallucination の割合が大幅に高い傾向を示している。GPT-2 系の中では、Klamath River と Mary River の文献も含めてファインチューンした GPT-2-FT の方が Hallucination の割合が高い結果となり。ファインチューニングによる Hallucination の抑制は確認できなかった。GPT-3 系では、GPT3.5 の方が Hallucination の割合が高かった。河川間の比較では、GPT-2 系では Klamath River の方がどの河川でも当てはまるような一般論の Neutral の割合が多い結果となった。GPT-3 は両河川でほぼ同じ割合を示し、GPT-3.5 では Mary River の方が Hallucination が少ない結果となった。

図4に質問別の Hallucination の評価結果を示す。GPT-2 系は、Q1 概要、Q2 歴史のみの比較となるが、質問内容によらず Hallucination が多い。GPT-3 系では、共通する特徴として、Q3 地理的特徴、Q4 経済効果について Neutral が多く、河川の実情に沿った回答が生成されなかった。Q5 水資源計画については、比較的 Hallucination が少なく Fact の割合も高かった。

## (3) Perplexity

表-3 から Perplexity についてみると、GPT-2 が最も高い性能を示した。逆に GPT-2-FT は最も悪い値を示しており、不自然な文章を生成していることが伺える。ただし Ground Truth に対する両モデルの Perplexity では、GPT-2-FT の方が大幅に低い値を示しているため、ファインチューニングにより Ground Truth の文章を生成できる確率は高まっており、ファインチューニングが効果を発揮していることが示されている。

GPT-2-FT の生成した文章を見ると、専門用語は多く出現するものの、単語の並びが文法的におかしいものや、綴りが不完全なものも多く確認できた。この文章の不自然さが、gpt-neo-1.3B での Perplexity が高いことにつながったと考える。

#### (4) 実験結果の考察

##### a) ドメイン適応前の事前学習済み LLM の性能

本研究で定義した土木ドメインの文章生成に関する3つの評価指標ごとに、最適なモデルが異なる結果となった。Keyword Intersection, Perplexity ではパラメータサイズで劣る GPT-2 が GPT-3, GPT-3.5 とほぼ同等か凌駕する結果を示し、評価指標によってはパラメータサイズのスケールリング則に従わない結果となった。この原因として、本研究で評価した特定の河川に関する質問のようなローカルな文章生成では、必ずしもパラメータサイズが大きなくても生成できないことに起因すると考えた。したがって、土木分野で LLM を活用するためには、下流タスクでのドメイン適応が必要となる。

一方、Hallucination はパラメータサイズが最大の GPT-3 が最良の結果を示した。このことから、LLM を土木分野にドメイン適応する際は、パラメータサイズが大きいモデルをベースとすることが有利なことが示唆される。

##### b) ドメイン適応の効果

GPT-2 と GPT-2-FT の評価結果を見ると、専門用語やローカル情報の Keyword Intersection はファインチューニングによって向上するが、生成される文章が不完全となり Perplexity はむしろ劣化する可能性が示唆される。これは一般的な大量テキストの事前学習で得られた流暢な文章生成の能力が、土木分野の少量テキストでのファインチューニングにより劣化するためと考えた。

また、質問対象の河川の文献でファインチューニングしたにも関わらず、Hallucination が悪化しているが、この原因は判明していない。

## 6. おわりに

本研究では、土木分野において正確な文章生成を実現するための基礎的な研究として、正確性に関する Keyword Intersection, Hallucination, Perplexity の3つの評価指標を提案した。

その上で、文章生成 LLM を少量のデータによる下流タスクのみで土木ドメインに適応する可能性を探るため、複数の GPT モデルで河川分野の質問回答を生成し、3つの評価指標で比較した。その結果から土木分野での文章生成 LLM の活用に向けて以下の知見が得られた。

### (1) 事前学習済み LLM での土木分野の文章生成の限界

事前学習済み LLM のままでは、パラメータサイズが大きい GPT-3, GPT-3.5 でも、専門用語やローカルな回答は十分に生成できず、どこでも当てはまるような一般論の Neutral な文章を生成する傾向が見られた。そのため、土木分野で文章生成 LLM を利用するためには、下流タスクでドメイン適応を図る必要があると判断した。

ただし GPT-4 など、よりパラメータサイズの大きい最新モデルでも同様の傾向かは明らかでない。

### (2) LLM の土木分野へのドメイン適応の難点

ファインチューニングにより専門用語やローカル情報の生成が改善したが、同時に文章が不完全となり Hallucination も悪化する現象を確認した。ドメイン適応によって、必ずしもすべての評価指標が改善しないことが判明した。特に Hallucination が悪化することは、土木分野で LLM を活用する際に問題となる可能性が高い。

ただし本研究では GPT-2 を少量のデータでファインチューニングしたのみであり、他のモデルやデータで同様の傾向かは明らかでない。

### (3) 今後の課題

#### a) 評価指標と評価用データセット

本研究では、独自に評価指標と評価用データを用いて実験を行ったが、現状では土木分野で文章の正確性を評価するための指標やデータセットは定まっていないため、これらの確立が課題である。

#### b) ドメイン適応手法の改善

土木分野の文章生成では、Keyword Intersection, Hallucination, Perplexity の3つを同時に精度よく満たすことが望ましい。ドメイン適応時に Hallucination, Perplexity を低く保ったまま、Keyword Intersection を向上させることが課題である。

そのためには、ドメイン適応に用いるデータ量を増やすことが考えられるが、必要なデータ量は今後の検討が待たれる。また、Hallucination を発生させないような仕組みの採用が望まれる。Hallucination 対策としては、事実確認 (Fact Check) を行う機構を ChatGPT に組み込むという方法<sup>16)</sup>、あるいは、Amerlia ら<sup>17)</sup>のように LLM にそもそもデータソース情報を組み込み、結果を表示する方法などの研究事例があり、これらの知見も考慮してドメイン適応手法を改善することが今後の課題である。

## 付録 Hallucination 評価例

以下の例の場合、河川の長さが明らかに間違い。生態系や水管理に関する問題は Wikipedia に記述がなく、どの河川でも発生するので不明瞭。同様に原住民に関する文章も Wikipedia に記述がなく不明瞭。

Question

Please describe summary information for Klamath River in California

Answer (GPT3.5)

The Klamath River is a major river located in northern California and southern Oregon, stretching approximately 257 miles from its source in the Yolla Bolly-Middle Eel Wilderness to its mouth at the Pacific Ocean. The river is an important source of water for agriculture, industry, and domestic use, and provides habitat for a variety of fish species, including Chinook salmon and steelhead trout. However, the river has a history of major ecological and water management issues, including conflicts over water allocation and dam removal. The Klamath River is also culturally significant to several Native American tribes, who have relied on the river for traditional fishing and cultural beliefs.

赤字: Hallucination 黒字: Neutral 青字: Fact

## 参考文献

- 1) Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Wenliang Dai, Andrea Madotto, Pascale Fung : Survey of Hallucination in Natural Language Generation, arXiv :2202.03629, 2022.
- 2) Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova : BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv : 1810.04805, 2019.
- 3) 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 卷 22 号, 22-22042, 2023.
- 4) 鈴木雅弘, 坂地泰紀, 平野正徳, 和泉潔 : 金融文書を用いた事前学習言語モデルの構築と検証, 人工知能学会第二種研究会資料, 2021 卷 FIN-027 号, pp.5-10, 2021.
- 5) Alec Radford, Ka, rthik Narasimhan, Tim Salimans., Ilya Sutskever : Improving Language Understanding by Generative Pre-Training, Technical Report OpenAI, 2018.
- 6) Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever : Language Models are Unsupervised Multitask Learners, Technical Report OpenAI, 2019.
- 7) Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei : Language Models are Few-Shot Learners, arXiv:2005.14165, 2020.
- 8) Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, Ryan Lowe : Training language models to follow instructions with human feedback, arXiv:2203.02155, 2022.
- 9) 八幡一紀, 桐生佳介, 片柳宏太, 毛利研, 関本和穂 : GPT-2 を活用した質問応答タスクへの適応, 人工知能学会全国大会論文集, Vol.36, Online ISSN: 2758-7347, 2022.
- 10) Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, Samuel R. Bowman : GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding, arXiv:1804.07461, 2018.
- 11) Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, Raquel Fernández : The LAMBADA dataset: Word prediction requiring a broad discourse context, arXiv :1606.06031, 2016.
- 12) Stephanie Lin, Jacob Hilton, Owain Evans : TruthfulQA: Measuring How Models Mimic Human Falsehoods, arXiv :2109.07958, 2021.
- 13) Tianbo Ji, Chenyang Lyu, Gareth Jones, Liting Zhou, Yvette Graham : QAScore -- An Unsupervised Unreferenced Metric for the Question Generation Evaluation, arXiv :2210.04320, 2022.
- 14) Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, Diyi Yang : Is ChatGPT a General-Purpose Natural Language Processing Task Solver?, arXiv :2302.06476, 2023.
- 15) 松林優一郎 : ChatGPT と教育における信頼性, 第 64 回教育機関 DX シンポ, [https://www.nii.ac.jp/event/upload/20230421-05\\_Matsubayashi.pdf](https://www.nii.ac.jp/event/upload/20230421-05_Matsubayashi.pdf), 2023.
- 16) Potsawee Manakul, Adian Liusie, Mark J. F. Gales : Self-CheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models, arXiv:2303.08896, 2023.
- 17) Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, Geoffrey Irving : Improving alignment of dialogue agents via targeted human judgements, arXiv:2209.14375, 2022.

(Received June 30, 2023)

(Accepted August 31, 2023)

# FUNDAMENTAL RESEARCH FOR ADAPTING LLM TO THE CIVIL ENGINEERING FIELD

Junichiro FUJII, Junichi OKUBO, Riku OGATA and Masazumi AMAKATA

The development of text generation models based on Large Language Models (LLMs), such as ChatGPT, has been remarkable. In the field of civil engineering, LLMs are also expected to improve work efficiency. However, since LLMs are mainly trained on documents collected from the Web, there is a concern that they may not be able to generate accurate text due to a lack of training on specialized knowledge in the field of civil engineering. Therefore, as a fundamental study to realize accurate text generation in the civil engineering field, this study attempted to adapt LLM to the civil engineering domain. We proposed an accuracy evaluation method, evaluated the accuracy of text generation in the civil engineering domain using a pre-trained public model of LLM and a model with fine tuning, and discussed the challenges in adapting LLM to the civil engineering domain.