建設分野に特化した大規模言語モデルの構築

井　田　慎太郎　　中　林　拓　馬

Training Large Language Model Specialized for Construction

Shintaro Ida　　Takuma Nakabayashi

Abstract

The use of AI powered by large language models (LLMs) to boost productivity through knowledge application is highly anticipated in the construction industry. However, general-purpose LLMs often lack specialized construction knowledge, resulting in incorrect answers and hallucinations. To overcome this problem, we develop a construction-domain-specific LLM. We collected 37.5 billion Japanese characters from construction-related documents and Web data to create a dataset rich in domain knowledge. Using this dataset, we continually pretrained an existing LLM to tailor it for construction applications. This report outlines our dataset creation and LLM training methods and discusses how our construction-specific LLM can support tasks such as planning and safety instructions, as well as its potential for future multi-modal applications in the industry.

概　　要

大規模言語モデル(LLM)を用いた知識活用による生産性向上が建設分野において期待されている。しかし，一般的な LLM は，建設分野に関する専門知識が十分に含まれておらず，もっともらしい誤情報を生成しやすいという課題がある。その課題に対応するため，建設分野に特化した LLM を構築した。はじめに，建設関連の専門文書や Web データなどから約 375 億文字の日本語文章データを収集し，建設分野特有の知識を反映したデータセットを構築した。次に，このデータセットを用いて，既存の LLM に継続事前学習を行い，建設分野に特化した AI モデルの開発を試みた。本報告では，データセットの構築および LLM を学習する手法について紹介する。そして，施工計画書や安全指示書の作成など建設分野特有のタスクへの対応や，画像と文章の両方を扱うことができるモデルへの発展など，建設分野に特化した LLM の利用可能性についての今後の展望を述べる。

1.　はじめに

近年，ChatGPT をはじめとする大規模言語モデル（Large Language Model: 以下, LLM）を用いた生成 AI の発展は著しい。LLM を活用することで，対話形式による AI の利用が可能となり，質問応答やアイデア創出といった業務支援が多様な分野で進みつつある。同様に，建設分野においても，不具合事例の検索や計画書の作成のような，対話型 AI を用いた品質管理能力向上や労働時間短縮などの業務支援への期待が高まっている。

LLM は，利用者が入力した文章の流れを解釈し，次に続く単語を予測することで，利用者が入力した文の続きとなる文章を生成する AI である。これは，膨大な量の文章データから，もっともらしい単語や文章のつながりに関する知識を学習することで実現されている。そのため，どのような文章データで学習するかによって，LLM の入力文の解釈能力や文章生成能力は大きく影響を受ける。これらの解釈能力や文章生成能力を利用し，単純な文章生成だけではなく，質問応答や要約，翻訳といった多様な能力を獲得することができる。

一般に，公開されている多くの LLM は，主としてインターネット上に公開された Web ページから取得したデータを学習データとしている。一方で，それらに含まれる建設分野の情報は限られ，工事概要や新技術のニュースのような一部の情報しか得られない。安全や品質などの工事記録，詳細な設計情報など，建設分野で作成される大部分の文章データは公開されておらず，それらは一般に公開されている LLM の学習のための収集対象から外れている。建設分野の専門知識に関する文章を入力する場合においても，学習に用いた文章データによく見られるパターンに従う形で続きの文章を生成しようとする。しかし，建設会社の社内文書などは学習に用いたデータのパターンと異なるため AI が正しく解釈できず，「事実誤認」や「知識のあるふりをする」，あるいは「未知の専門知識を無視する」といった現象，いわゆるハルシネーション（知識の誤解や誤った論理展開）が発生することがある。これは LLM を技術情報の要約や過去の事例の検索などに用いる際に，誤情報の提示や情

報の欠落などの原因となり得る。こうした課題を解決するため，専門知識を含む参考資料を質問と共に入力する手法がある。しかし，LLM 自体が専門知識について学習していない場合，参考資料として入力された文章内の専門用語を正しく解釈できず，完全な解決には至らない。

このような背景を踏まえ，本研究では建設分野に特化した文章データを収集し，LLM の事前学習の一種である，継続事前学習を実施した。具体的には，施工記録や各種規準類などの専門的な建設関連の文書を中心に学習データセットを構築し，モデルに建設分野特有の知識を埋め込むことを試みた。本手法により，従来のモデルでは理解が困難であった専門用語や業務内容に対する理解度の向上を図り，過去の品質管理や施工計画を正しく解釈した出力が可能となることを目指している。

2.　大規模言語モデル構築の概要

2.1　学習の概要

建設分野に特化した LLM を構築するため，施工記録や各種規準類といった建設分野に特化した文章データを収集し，それらの知識を埋め込む学習を行う。LLM の構築は，一般的に事前学習と事後学習の 2 種類の学習過程で構成される(Fig. 1)。

事前学習とは，大量の文章データを用いて，与えられた文章の続きを生成できる LLM を構築するプロセスである。事前学習の目的は，知識や各言語特有の文法などを LLM に埋め込むことである。そのために，膨大な量の文章データを収集し，与えられた文脈の続きとなる単語の出現確率を学習する。文章は世界中の Web ページのデータを使うことが多いが，日本語のみや建設分野のみなどの特定の領域に偏ったデータを利用することもある。このプロセスにより，利用者が入力した文章を解釈し，文脈に応じて続きとして出てくる確率の高い単語を逐次的に予測することが可能となり，自然な文章生成能力を獲得する。この過程で構築されたモデルは一般にベースモデルと呼称される。全く学習処理を行っていない初期状態のモデルから事前学習を行う場合，言語特有の文法や論理展開など一般知識の学習も必要となる。その際，極めて大量の文章データが必要となり，その学習には膨大な計算資源を消費する。そのデータの収集労力や計算資源を削減するため，継続事前学習という手法の利用が一般的である。継続事前学習とは，初期状態のモデルの代わりに，既存のベースモデルに対して学習処理を行うものである。既存の LLM が既に持っている知識を活用しつつ，追加知識をモデルに反映できるため，追加したい知識に関わる文章データのみを学習データとして用意すればよい。そのため，継続事前学習では事前学習と比べ，文法規則や論理展開など自然言語一般のふるまいを習得させることを目的とした大量の文章を用意することが不要になる。本研究で扱う 8B（80 億パラメーター）規模の言語モデルに対して日本語で継続事前学習を行う場合，200 億から 2,000 億トークン（文章を単語や単語の一部，記号などの最小単位に分割したもの）規模のデータで行った事例が報告されている [2)~4)]。本研究で扱うデータはほとんどが日本語で構成されるため，このトークン数はおよそ 200 億～2,000 億文字に相当する。初期状態からの事前学習では 15 兆トークンの学習データを利用している [5)]ため，これは比較的少量のデータといえる。

一方，事後学習は，事前学習を完了した AI モデルに対話応答や翻訳といった具体的なタスク遂行能力を付与することを目的とする。この段階で構築されるモデルは指示応答モデルと呼ばれる。建設分野には施工計画書や安全指示書の作成など多種のタスクがある。ベースモデルに対してタスクごとの事後学習を行うことで，各タスクに特化したモデルを構築することができる。

本研究では，日本語の事前学習済 LLM を用い，建設分野に関する専門的な知識や用語，文脈に対応できるベースモデルの構築を目指す継続事前学習を行った。具体的には，既存の日本語 LLM が持つ自然な日本語生成能力を活かしつつ，従来の学習データには含まれていない建設分野特有の文章データを追加で学習させることで，これらの専門知識や表現を新たにモデルへ埋め込むことを試みた。

2.2　事前学習済モデルの選定

継続事前学習においては，事前学習済モデルの選定が非常に重要となるため，以下の方針でモデルを選定した。

まず，使用するモデルの構造を決定した。学習計画時点において，継続事前学習の実績が多い Meta 社の Llama 3.1[5)]を採用することとした。Llama 3.1 は，最大で 128,000

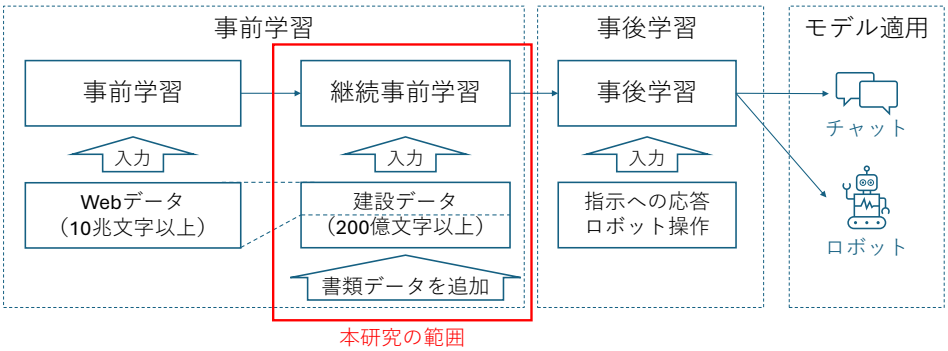

Fig. 1　LLM学習のフロー

LLM Training Workflow

トークン（日本語では約 12 万文字に相当）までの長い文章を一度に入力することができ，また推論速度も速いという特徴を持つモデルである。さらに，建設分野では写真や図面などの画像情報を取り扱うことが多いことから，今後は画像と言語を組み合わせた AI モデルの構築も期待される。Llama 3.1 は，画像も扱えるモデル(Llama 3.2)への発展が容易であることも，選定理由の一つである。

次に，モデルの大きさの選定を行った。Llama 3.1 では，3 種類の大きさのモデルが用意されている。その中で，建設分野に特化したことで学習に用意できるデータが少なめであること，また，完成した AI モデルを実際の業務で使う際に，必要となるコンピューターの性能や運用コストができるだけ小さくなるよう配慮し，最も小さい 8B サイズのモデルを採用した。

最後に，継続事前学習のための既存学習済モデルの選定を行い，東京科学大学の Swallow[6]を採用した。Swallow は Meta の Llama 3.1 の事前学習済モデルに対し，2,000 億トークンの日本語データ（約 2,000 億文字に相当）である Swallow コーパス v2[4]を用いて継続事前学習を行った日本語特化モデルである。Swallow コーパス v2 は収集した Web ページから重複を排除し，更にそこから学術・教養に関連するデータ約 15%を抽出した高品質なデータセットである。Swallow はこれらのデータから日本語特有の文法や日本文化などを学習し，自然な日本語生成能力を有する。以上から，Swallow を継続事前学習のための既存学習済モデルとして採用した。

これらの検討結果を踏まえ，tokyotech-llm/Llama-3.1-Swallow-8B-v0.2[7]に対し継続事前学習を行うこととした。

3.　学習データセットの構築

3.1　建設分野に特化した文章データセットの構成

継続事前学習を行うためには，文章データを大量に準備する必要がある。今回は日本語の事前学習済の LLM に対して継続事前学習を行うため，日本語の文法規則や常識については学習しているが，建設分野に関する専門的な知識は獲得できていない。そのため，建設分野の専門知識を集約したデータセットが必要となるが，このようなデータセットは公開されていない。そのため，本研究では建設分野特化のデータセットを一から構築した。

今回構築する文章データセットのために収集する文章データの目標数量は，継続事前学習の研究報告を参考に 200 億文字以上とした。継続事前学習を実施する際，学習データの内容や分布が，継続事前学習を行うモデルの事前学習時に利用されたデータと大きく異なる場合，事前学習時にモデルが獲得した知識が失われる破壊的忘却と呼ばれる現象が生じやすくなる [8]。そこで，データの約半数にあたる 100 億文字をベースモデルの学習に利用されたデータと同様に Web ページ由来の文章で構成し，破壊的忘却の発生を抑制することを目的としたデータ構成とした。Web ページから収集したデータについて

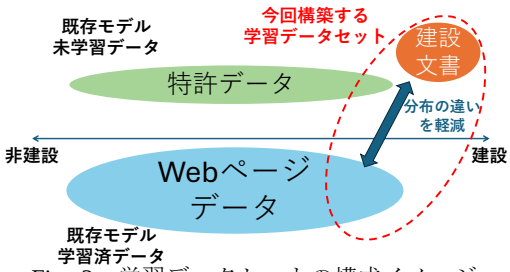

Fig. 2　学習データセットの構成イメージ
Diagram of Training Dataset Structure

は，既存のベースモデルに含まれているデータを重複して利用することで，事前学習済モデルの学習に使用したデータの分布に近づけている。残りの半数については，ベースモデルの学習データには含まれない建設分野の知識を LLM に追加で埋め込むためのデータとして，建設関連の専門文書（施工記録や各種規準類など）および建設関連特許から抽出したデータで構成した(Fig. 2)。

3.2　Web ページからの学習データ作成

最初に，Web ページ由来のデータの収集を行った。その際，商用利用が可能な最大規模の日本語データセットである ABEJA CC JA を利用した。このデータセットは，日本語の Web ページを中心に収集され，重複や不要な部分が除去された約 4,000 億文字分のデータを含んでいる。収集目標量は 100 億文字であるため，ABEJA CC JA 全体の 4,000 億文字から 2.5%を選別する必要があった。前述の通り Web ページ由来のデータセットを構築する目的は，ベースモデルの学習データと，建設関連の独自データの分布を近づけることである。そのため，2.5%の選別はランダムではなく，建設に関連した文章を選別する必要があった。

関連度を判定するため，文章を単語に分割した際の一単語当たりの文字数を指標に定量化して評価を行った。この評価には，建設分野特有の単語はなるべく分割せず，長い一単語として認識できるプログラムを利用し，文章の分割を行った [9]。このプログラムは，建設分野の文章を収集し，文章内で利用されている単語の頻度を分析し，高頻度で出現する単語はなるべく分割しないよう設計されたものである。そのため，建設分野でよく用いられる用語ほど，分割されずに長い単語として出力される傾向にある。例えば，Table 1 に建設分野の文章と，非建設分野の文章を当該プログラムで分割した例を示すが，「日本建築学会」といった語句が 1 つの単語として抽出されている。このように，建設分野に関連する単語が多く含まれる文章ほど，1 単語あたりの平均文字数が多くなる。

この手法を用いて，ABEJA CC JA 全体の 2.5%の分量を目標に，建設分野の文章の選別を行った。単語あたりの平均文字数を用いた文章の選別を行うための閾値の算出には，分割ファイルとして提供されている本データセットから最小の 1 つをサンプルとして用いた。サンプルに格納されている各文章をプログラムで分割し，単語あたりの平均文字数を文章ごとに算出した結果を Fig. 3

Table 1　建設分野の文章および一般の文章の単語あたりの文字数の比較例
Comparison of Characters per Word in Construction and General Texts

元の文章	分割された文章	平均文字数/単語数
一般社団法人日本建築学会は，会員相互の協力によって，建築に関する学術・技術・芸術の進歩発達をはかることを目的とする学術団体です。	一般社団法人 / 日本建築学会 / は / , / 会 / 員 / 相互 / の / 協力 / によって / , / 建 / 築 / に関する / 学 / 術 / ・ / 技術 / ・ / 芸 / 術 / の / 進 / 歩 / 発 / 達 / を / は / か / ることを目的 / とする / 学 / 術 / 団体 / で / す / 。	1.730
トークナイザーは文章を短い文字列（トークン）に分割するものである。	ト / ー / ク / ナイ / ザ / ー / は / 文 / 章 / を / 短 / い / 文 / 字 / 列 / (/ ト / ー / ク / ン /) / に / 分 / 割 / する / も / の / である / 。	1.138

に示す。平均文字数が多い順に各文章を並べた際，上位2.5%の分量達したときの文字数が 1.55 文字であったためこれを閾値として利用することとした。得られた閾値をもとに ABEJA CC JA 全体からのデータ選別を行った。目標は 100 億文字であったが，サンプルに偏りがあったとみられ，結果として 179 億文字が抽出された。

3.3　建設関連の専門文書からのデータの抽出

次に，建設関連の専門文書を解析し，そこから文字データを抽出することで学習用データの作成を行った。対象とする建設分野の書類は，施工記録や各種規準類，論文などで構成した。一般的に，建設分野の文書は表や画像を多く含み，レイアウトが複雑であることが多い。さらに，本報告のような 2 段組のレイアウトを持つ文書の場合，左右の段が誤って結合され，意味がつながらない形で文章が抽出されるケースも少なくない。

データからの文字起こしについては， Microsoft Azure の OCR（光学式文字認識）機能を用いて文書から文字データを取得した。OCR を利用することで，高い精度で文字情報を抽出できる一方で，本文の文章だけでなく，文書のタイトルや図表内の文章，キャプションなども同じように抽出されてしまい，さまざまな要素が混在する結果となった。また，OCR の特性上，文書の上部から段落が抽出されるため，2 段組のような複雑なレイアウトでは，文章の順序が本来意図したものと異なる場合が多かった。

そこで，AI によるレイアウト解析技術を導入し，OCR で抽出した文字データが本文に該当するのか，あるいは図表のキャプションや表内のデータであるのかを判別し，文章の並び順を適切に復元する処理を行った(Fig. 4)。この AI によるレイアウト解析には，Doclayout-YOLO[10]を用いた。本処理により，建設関連の専門書類から本文のみを抽出し，約 112 億文字を抽出することができた。

3.4　特許情報からのデータの抽出

さらに，Web ページに含まれない文章データを得るため，特許情報データを利用した。特許情報には建設関連のデータ以外も含まれているため，Web データと同様に，一単語あたりの文字数を指標とし，建設分野に関連したデータの選別を行った。2004 年以降の特許公報データを対象に選別を行い，約 84 億文字を得ることができた。

3.5　学習データとして集約

3.2 節から 3.4 節で得られたデータを集約し，合計約

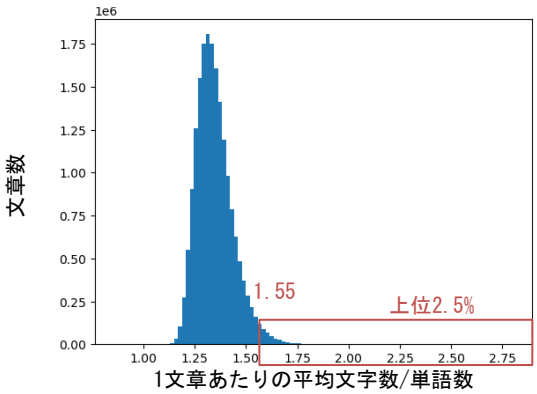

Fig. 3　文字数/単語の集計結果
Number of Characters per Word

Fig. 4　AIによるレイアウト判定結果の例
Sample Results of AI-based Layout Recognition

Table 2　構築した学習データの概要
Overview of Constructed Training Data

母集団	文字数	割合
建設関連の専門文書	11,221,762,273	29.87%
特許データ	8,436,988,784	22.46%
Webページ	17,909,443,954	47.67%
合計	37,568,195,011	100%

375 億文字の文章データセットを構築した(Table 2)。これは，LLM の継続事前学習を行うために目標とした 200 億文字を大きく超過した。しかし，Web ページから抽出したデータが全体の 47.67%と半数程度であり，3.1 節で計画した構成から大きく外れていないため，追加の選別はせず，学習に用いることとした。

4.　AI による言語モデルの学習

4.1　文章データの前処理

収集した文章データを AI の学習に利用するには，トー

クン化処理が必要となる。今回事前学習済モデルとして採用した Swallow で使用されている Llama 3.1 のトークナイザー(文章をトークンに分割するためのプログラム)を用いた。トークナイザーの設定については, Swallow の設定をそのまま引き継いでいる。

さらに，学習に使用する全データのうち，5%を AI の検証用データとして分離した。検証用データは，AI の学習がどの程度進んでいるかを評価するためだけに使い，実際の学習には使用しない。

学習にあたっては，施工記録や不具合事例といった学習データの分類ごとにトークン化処理を行った。データの収集および前処理には時間がかかるため，収集が完了するたびに分類ごとに前処理を行った。学習の安定性や汎化性能を上げるためにはデータに偏りがないことが重要であるため，シャッフルによりデータを並び替える必要があった。これには分類ごとに一定数量でまとめた塊をランダムに並び替えるという，プログラム既定の方法を用いた。

4.2　継続事前学習の実行

3 章で構築した 375 億文字で構成される学習データを，AWS Trainium 1 チップを 32 基搭載したコンピューターを 16 台接続し，合計で 512 基の AI アクセラレーターを使ったコンピュータークラスターを利用し 1 エポック分学習させた結果，約 36 時間で学習が完了した。

学習時の設定値について，ほぼすべての項目を初期設定のまま利用した。ただし，学習の進み方を調整するための学習率のみ変更した。具体的には，学習率の推移曲線は Swallow に従い [3]，Fig. 5 に示すように，直線的に増加して最大値に達した後，コサインカーブで減衰させることとした。また，学習率の最大値については，初期設定を利用したところ破壊的忘却の兆候が見られた。そこで，学習率を減らしながら試行を繰り返し，破壊的忘却が発生しない範囲で設定できる最大の学習率とした。推移曲線の初期で学習率を増加させる理由は，ベースモデルに内容の異なる新しいデータを追加学習する本研究のような場合に, 破壊的忘却の発生を抑制するためである。

モデルの学習の進み具合は，損失で評価した。損失は，モデルの予測と正解との間の誤差を数値化した指標である。LLM では次の単語の予測の際，複数の候補とその出現確率を出し，そのうち正解の単語の出現確率をもとに損失を算出する。この損失が小さいほど，正解の単語の出現確率が高くなるため, 損失を下げるよう学習を行う。学習中の損失の推移を Fig. 6 に示す。

全体として，学習時・検証時の双方で，損失がわずかながら減少する傾向を示すグラフとなった。この際，損失の最大は 2.5 程度にとどまり，すぐに減衰している。損失が極めて高い値をとる場合や，高い値から減少しない場合には破壊的忘却が発生している可能性があるが，本検証ではこのような兆候は見られなかった。学習中に何度か記録された損失の一次的な高い値については，学習開始直後に記録された値よりは高いものの，継続事前学習元のモデルの学習時に記録された値の範囲内であり，問題はないと考えられる。

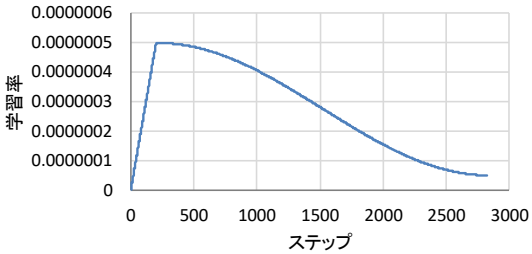

Fig. 5　学習率の推移
Learning Rate

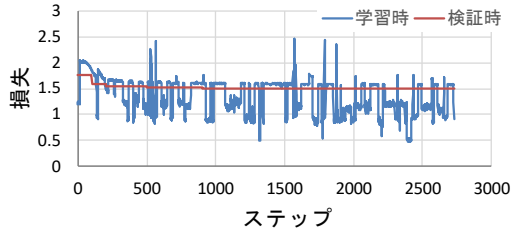

Fig. 6　学習時の損失の推移
Training & Validation Losses

学習時の損失については，グラフ上において，約 1.5 を境に高めの値と低めの値が交互に現れている結果となった。損失が低めとなるタイミングは，継続事前学習元のモデルである Swallow の学習データと分布が近い Web ページ由来のデータを学習している段階と考えられる。これは，Swallow の学習において Web ページ由来のデータへすでに最適化されており，今回の継続事前学習では Web ページ由来のデータを学習している際に，低い損失を記録したと考えられる。

一方，損失が高め傾向を記録するステップでは，事前学習済モデルとは内容が大きく異なる建設分野専門の書類を集中的に学習している段階であると推測される。このステップにおける値はわずかながら減少し，継続事前学習元の Swallow の学習完了時の損失 [3]とほぼ同値になることから，学習は正常に行われたとみられる。

また，学習の途中で損失が極端に低くなる場合が見られた。これは，学習データの中に重複があり，AI モデルが同じデータに対して過度に適応した結果と考えられる。

このような複数の傾向が明確に見られたことから，既存の Web ページ由来のデータと新たに収集した建設分野のデータが分離し，シャッフルが不十分である可能性が示唆される。

5.　課題と今後の展望

5.1　解決すべき課題

本研究では，建設分野特有の知識を LLM に追加学習させるためにデータの収集および継続事前作業を行い，以下の課題が抽出された。

今回の継続事前学習においては，新たな建設分野の

データと Web からの学習済データが分離しており，損失から明確に学習済データの学習時，新たなデータの学習時などが読み取れる結果となった。学習データにこのような偏りがある場合，破壊的忘却が発生しやすくなる。今回その兆候は見られなかったが，今後学習データ量を増加させた場合，効率的な学習のために学習率を上げる可能性があり，その際は学習前の検証で発生したように破壊的忘却が発生する原因となり得る。今後は文章単位でのシャッフルなどシャッフル手法を工夫し，新たなデータと学習済データの偏りのないシャッフルが必要となる可能性がある。

また，学習過程で重複データが多い場合，LLM の出力がそのデータを複製のように出力してしまう可能性がある。これは著作権侵害のリスクにもつながるため，も重複データの排除は重要な課題である。しかしながら，データが膨大であり，人力での内容確認は非現実的なため，過度な偏りがなく適切な分布のデータになっているかを調べる手法を検討する必要がある。

5.2　今後の展望

今回の継続事前学習による LLM の構築により，今後の展望として，以下の 2 点が挙げられる。

第 1 に，今回構築した LLM をベースモデルとし，建設業における各種タスクに対応した指示応答モデルを構築可能なことである。具体的には，建設業の専門用語を解釈し，それらの用語を利用した工程作成や計画書作成といった建設分野特有の各種タスクに対し，事後学習により対応できるようになる。一方で，タスクに特化した事後学習用データの整備には多大な労力と費用を要するため，効率の良いデータ収集方法を検討する必要がある。

第 2 に，画像とテキストを同時に扱う VLM(Vision-Language Model)への展開が挙げられる。建設分野では図面や写真など視覚情報の活用が不可欠であり，VLM モデルの構築は今後の実務支援において重要な役割を果たすと考えられる。今後，VLM の導入により，設計図や現場写真などの視覚情報と文章情報を合わせて入力することが可能となり，例えば現場に即した指摘事項を現場の用語で出力する，などが期待できる。言語のみを扱える LLM で構築した指示応答モデルに蓄積された知識や能力を，画像と言語の両方を扱える VLM へと効率的に転用できる手法なども提案されている [11]。今回学習したモデルも Llama 3.2 へ転用する検証を行いつつ，建設分野での LLM・VLM の早期実用化に取り組んでいきたい。

6.　おわりに

本報告では，以下の点について報告を行った。

1) 建設関連の専門文書の収集や，特許情報および Web データから建設関連の文章を抽出したことにより，約 375 億文字の建設分野に特化した文章データセットを一から構築した。
2) 継続事前学習にて学習率を調整し，破壊的忘却を起こさず LLM に建設分野の知識を追加学習させた。

今後，構築した LLM をベースとし，建設分野の各種タスクに対応した指示応答モデルを続けて構築する。そのモデルを不具合事例の検索や計画書の作成のようなツールに適用することで，品質管理能力の向上や労働時間削減につなげていく。

参考文献

1) Lewis, et. al.：Retrieval-augmented generation for knowledge-intensive NLP tasks, Proceedings of the 34th International Conference on Neural Information Processing Systems, pp. 9459-9474, 2020.12
2) 服部，他：Swallow コーパス v2: 教育的な日本語ウェブコーパスの構築，言語処理学会第 31 回年次大会発表論文集, pp. 94-99, 2025.3
3) rinna 株式会社, “rinna/llama-3-youko-8b”, Hugging Face, 2024-05-01, https://huggingface.co/rinna/llama-3-youko-8b, （参照 2025-05-29）
4) 東京エレクトロンデバイス株式会社, “tokyo-electron-device-ai/llama3-tedllm-8b-v0”, Hugging Face, 2024-10-17, https://huggingface.co/tokyo-electron-device-ai/llama3-tedllm-8b-v0, （参照 2025-05-29）
5) Grattafiori, et. al., “The Llama 3 Herd of Models”, Meta Platforms, 2024-07-23, https://ai.meta.com/research/publications/the-llama-3-herd-of-models/ （参照 2025-05-29）
6) Fujii, et. al., “Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities”, ArXiv, 2024-04-27, https://arxiv.org/abs/2404.17790（参照 2025-05-29）
7) 東京工業大学, “tokyotech-llm/Llama-3.1-Swallow-8B-v0.2”, Hugging Face, 2024-11-11, https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-v0.2 （参照 2025-05-29）
8) 麻場，他：語彙置換継続事前学習による日英バイリンガルモデルの構築と評価，言語処理学会第 30 回年次大会発表論文集, pp.949-954, 2024.03
9) 井田，他：建築分野に特化した文書検索技術の開発，日本建築学会大会学術講演梗概集, pp.123-124, 2024.08
10) Zhao, et. al. , “DocLayout-YOLO: Enhancing Document Layout Analysis through Diverse Synthetic Data and Global-to-Local Adaptive Perception”, ArXiv, 2024-10-16, https://arxiv.org/abs/2410.12628（参照 2025-05-29）
11) Akiba, et. al., Evolutionary optimization of model merging recipes, Nature Machine Intelligence volume 7, pp. 195–204, 2025.01