

製造業ドメインにおける日本語 LLM の性能調査

上原大暉¹ 田中宏治¹ 金井健一郎¹ 内出隼人¹ 伍井啓恭¹ 齊藤辰彦¹

¹ 三菱電機株式会社 情報技術総合研究所

{Uehara.Daiki@ak, Tanaka.Koji@bc, Kanai.Kenichiro@dy,
Uchide.Hayato@dy, Itsui.Hiroyasu@ea, Saito.Tatsuhiko@db}
.MitsubishiElectric.co.jp

概要

特定分野を対象とした日本語向けの大規模言語モデル (LLM) に関する性能評価の取り組みが知られている。しかしながら、製造業分野においては日本語 LLM の性能調査が不十分である。本研究では、製造業関連機器についての質問応答タスクに対して複数の事前学習済み日本語 LLM の性能を比較し、製造業分野における専門知識の学習にどの日本語 LLM が適しているかを調査した。結果、教師ありファインチューニング (Supervised Fine Tuning, SFT) を施した LLM において、SFT 前後で Rouge-L が 0.080 から 0.224 に向上することを確認した。一方で、依然としてハルシネーションの残存など、LLM の特定分野の専門知識の定着に課題が見られた。

1 はじめに

近年の大規模言語モデル (LLM) の発展に伴い、LLM へ特定分野の専門知識を学習させる、ドメイン適応に関する調査が進められている。また、日本語に特化した日本語 LLM も複数登場しており、様々なドメイン知識の学習に活用されている。一方で、複数ある日本語 LLM のうち、特定ドメイン向け追加学習に適した LLM の選択方法は自明ではない。特に、製造業ドメインにおける日本語 LLM の性能調査は不十分である。

そこで、本研究では、製造業分野、特にファクトリーオートメーション (Factory Automation, FA) 機器に関する質問応答タスクに対し、複数の事前学習済み日本語 LLM の回答性能を比較した。教師ありファインチューニング (Supervised Fine Tuning, SFT) の有無で LLM の生成結果を比較し、製造業分野のデータを用いたファインチューニングに適した日本語 LLM を調査した。また、LLM の生成結果を分析し、製造業分野でのファインチューニングの有効性

を検証した。

2 関連研究

現在、日本語のコーパスを多く学習した日本語 LLM が多く登場しており [1, 2]、一般ドメインのタスクに対する性能調査 [3] では高い性能を示している。特定ドメインのタスクにおいても、ファインチューニングすることで性能が高くなるのが金融や医療の分野で知られている [4, 5]。

本研究で対象とする製造業分野においても、製造業現場での課題解決のために LLM の活用が検討されている [6]。加えて、製造業ドメイン向けに LLM をファインチューニングすることの有効性が確認されている [7]。また、日本語を含む製造業ドメインのデータセットで LLM をファインチューニングすることで、製造業ドメインタスクに対する性能が向上することも確認されている [8]。

一方で、日本の製造業ドメインタスクをファインチューニングするために有効な日本語 LLM は結論付けられていない。

そこで本研究では、複数の日本語 LLM を対象とし、製造業ドメインの質問応答タスクを用いて、ドメイン適応に適した日本語向け LLM を調査した。

3 日本語 LLM の比較実験

製造業分野における日本語 LLM の能力を検証するために、データセットとして FA 機器の FAQ データを利用し、SFT の有無で比較評価することでドメイン適応の必要性を確認する。また、複数の日本語 LLM を対象に実験し、製造業分野での活用に適した LLM の条件を調査する。

表1 FAQデータの例

質問	ユニバーサルモデル QCPU の Ethernet 内蔵ポートと USB ポートの両方にパソコンを接続して、GX Developer や GX Works2 で同時モニタ可能ですか。
回答	はい、可能です。
質問	QCPU でエラーコード『4100』が発生し、エラーが発生している命令は、[D/ D0 D10 D20] です。対処方法を教えてください。
回答	エラーコード『4100』は、『OPERATION ERROR』です。このエラーが除算で発生している場合、除数 (D10、D11) が 0 になっています。除算実行時に除数が 0 以外になるように修正してください。

表2 実験で使った日本語 LLM

Model	アーキテクチャ	パラメータ数	ITの有無
llama3-youko-8b	Llama3	80億	なし
Llama-3-swallow-8b-v01	Llama3	80億	なし
Llama-3-ELYZA-JP-8B	Llama3	80億	なし
llama-3-youko-8b-instruction	Llama3	80億	あり
Llama-3-Swallow-8B-instruction-v0.1	Llama3	80億	あり
youri-7b	Llama2	70億	なし
youri-7b-instruction	Llama2	70億	あり
ELYZA-japanese-Llama-2-7b-instruct	Llama2	70億	あり
Swallow-7b-instruct-v0.1	Llama2	70億	あり

3.1 データセット

三菱電機の FA 機器に関する¹⁾FAQ (よくある質問) を質問応答タスクのデータとして利用した。データの例を表 1 に示す。各データは、FA 機器に関する質問と、その質問に対する回答が対からなる。

本実験では、2,244 件の質問文と回答文の組を対象とした。本データを用いて、LLM に質問文を入力し、回答文に近い生成結果が得られるかを問う質問応答タスクを実施した。生成結果と回答文の近さの指標としては Rouge-L を用いた。これにより、製造業の FA 機器という専門性の高い分野に対する LLM の回答性能を検証する。

3.2 モデル

本実験では、日本語 LLM として表 2 に示す LLM を使用した。これらは、Meta 社の事前学習済みモデル Llama2[9] 及び Llama3[10] を日本語データで継続事前学習した LLM である。継続事前学習に加えて、指示チューニング (Instruction-Tuning, IT. モデルに対して具体的な指示を与え、その指示に従って動作するようにモデルを調整する手法) が施された LLM も含む。これらの LLM を比較実験し、製造業への

ドメイン適応にあたり、モデルアーキテクチャの違いや事前指示チューニングの有無での性能差を調査する。

3.3 学習設定

FA 機器とは、製造業におけるロボットや制御装置などを指す。本実験で扱うデータセットには、FA 機器に関する用語が多数含まれる。それらを扱うには FA 機器に関する専門的な知識が必要である。そのため、FA 機器に関する知識は一般的ではなく、LLM がその内容を十分に事前学習できていない可能性がある。そこで、本データセットで LLM を SFT し、ファインチューニングの有効性を検証する。SFT の形式は、付録の図 1 のプロンプトを使用した。

学習には、計算リソースを節約できる学習手法である LoRA (Low-Rank Adaptation) [11] を使用した。

本実験では、全データ 2,244 件を学習用、検証用、評価用に 8:1:1 に分割した。学習条件の詳細を付録の表 5 に示す。

1) <https://fa-faq.mitsubishielectric.co.jp/category/show/694>

表3 Rouge-L スコア結果

Model	Model No.	SFTの有無	Rouge-L
llama3-youko-8b	1	なし	0.080
	2	あり	0.224
Llama-3-swallow-8b-v01	3	なし	0.140
	4	あり	0.175
Llama-3-ELYZA-JP-8B1	5	なし	0.127
	6	あり	0.202
llama3-youko-8b-instruction	7	なし	0.058
	8	あり	0.216
Llama-3-Swallow-8B-instruction-v0.1	9	なし	0.073
	10	あり	0.220
youri-7b	11	なし	0.112
	12	あり	0.136
youri-7b-instruction	13	なし	0.091
	14	あり	0.148
ELYZA-japanese-Llama-2-7b-instruct	15	なし	0.095
	16	あり	0.139
Swallow-7b-instruct-v0.1	17	なし	0.115
	18	あり	0.201

4 結果と考察

4.1 実験結果

評価結果を表3に示す。すべてのモデルにおいて、教師ありファインチューニング (Supervised Fine Tuning, SFT) を施すことで Rouge-L が向上した。特に、表3のNo.2のモデルである llama3-youko-8b では、0.080pt から 0.224pt へ大きく向上した。このことから、製造業分野の質問応答タスクにおいてファインチューニングの必要性が示唆される。SFT を施したモデルの中では、No.2 と 12、No.6 と 16 の比較から、Llama2 ベースのモデルよりも Llama3 ベースのモデルの方が一貫して精度が高かった。同様に、No.4 と 10、No.12 と 14 の比較から、事前指示チューニングが施されたモデルの方が精度が高くなる傾向が見られた。このことから、Llama3 ベースで指示チューニング済みのモデルを利用することが効果的であり、ベースモデルの選定の際にはモデルアーキテクチャが重要であることが示唆される。

最も高い精度であった No.2 のモデルの生成例のうち、Rouge-L が 0.760 であったサンプルを表4に示す。表4の「SFT 未実施 LLM の生成文」に示すように、SFT を未実施の LLM は質問の趣旨から逸脱

し、冗長な回答を生成することが確認された。これに対し、表4の「SFT 実施済 LLM の生成文」に示すように、SFT の実施により、LLM は正解文に近い様式で回答を生成することを確認した。これは LLM を SFT することで、回答形式を学習出来ていると考えられる。また、回答の様式だけでなく、FA 機器に関する用語を正確に扱った回答も確認できた。表4の「正解文」と「SFT 実施済 LLM の生成文」内の「RDA」や「SDA」は FA 機器に関連する専門用語である。「質問文」内の用語を引用した回答ではなく、学習データに含まれていた専門用語を活用した回答を生成していたことから、ファインチューニングの有効性を再確認した。

一方で、多数のサンプルで回答にハルシネーションを含むことも確認された (付録の表7)。Rouge-L が大きいサンプルの多くは、「正解文」に FA 機器関連の専門用語を含まない生成例 (付録の表8) であり、専門用語を扱った多数の生成例は Rouge-L が小さかった。また、表4の「SFT 実施済 LLM の生成文」に示すように、「短絡」を「short 絡」のように用語の誤脱字も多く確認された。

これらの結果から、学習データに出現する頻度の低い FA 機器に関する知識や用語は、正確に扱えないという問題が確認された。

表 4 llama3-youko-8b の生成例

質問文	LJ71C24 で相手機器の RS-485 インタフェースが 2 線式の場合の配線方法は?
正解の回答文	SDA と RDA を短絡、SDB と RDB を短絡して、相手機器側の+ (Data-A) - (Data-B) に接続ください。
SFT 未実施 LLM の生成文	2 本の線を接続するだけです。## 例題 2 以下の例はタスクリストの項目です。このタイトルは適当ですか? (中略) R484 を同時に制限すること 26.Python で簡易 R429 を開発する.27.C++で単純な R430 を生成する。
SFT 実施済 LLM の生成文	SDA と RDA を短絡、SDB と RBD を short 絡して、相手側の+ (Data-A) と- (Data-B) に接続ください。

4.2 考察

Rouge-L スコアと生成結果から、質問の趣旨から逸脱し、冗長であった回答文の様式を整えるという点で、SFT の効果を確認できた。一方で、生成された回答内容にハルシネーションが含まれることや、用語を正確に扱えないことが確認された。この結果は、SFT は生成文の様式を整える効果にとどまり新たな知識の注入には継続事前学習が必要である、との報告 [12] とも一致する。

また、LLM の生成内容の正確さの評価は、Rouge-L だけでは不十分である。そこで本実験では、LLM による生成内容を「生成内容の正確さ」を人手で評価した。評価に使用した FAQ データは 224 件で、学習データには含まれないデータである。

「生成内容の正確さ」は、LLM により生成された質問に対する回答が、正解の回答の内容と一致しているかどうかを評価する指標である。結果は、生成文と正解文の内容が完全一致した回答は 224 件の内 14% であった。また、86% は生成内容にハルシネーションを含む回答であった。学習データに含まれない質問への回答能力は、SFT だけでは十分に向上できないと考えられる。

つまり、日本語 LLM における製造業へのドメイン適応においても、LLM への製造業ドメイン知識の注入は SFT だけでは不十分であることが示唆される。また、ハルシネーションへの対策としては、RAG(Retrieval Augmented Generation)[13] により必要なドメイン知識をコンテキストとして LLM の入力に含める方法が考えられるが、コンテキスト内のドメイン知識が不十分な場合を考慮すると、LLM 自身もドメイン知識を扱えた方が望ましい。

したがって今後は、継続事前学習によりドメイン知識を獲得した LLM を構築し、RAG と併用することで精度を改善していく。用語の表記崩れについて

は、製造業ドメインの用語に対して、トークン分割が細かすぎることや、細かいトークンをうまく扱えていないことが原因と考えられる。したがって、製造業ドメインに特化したトークナイザの構築や LLM の語彙拡張により、用語の表記崩れを改善できる可能性がある。

5 おわりに

本研究では、製造業分野における日本語 LLM の性能を調査した。9 種類の LLM に対し、FA 機器に関する質問応答のデータセットを用いて教師ありファインチューニング (Supervised Fine Tuning, SFT) を実施し、その生成結果を比較し分析した。結果として、SFT を施すことで、Rouge-L スコアが最大で 0.081pt から 0.224pt へ向上した。また、SFT 実施後では、質問の趣旨から逸脱し、冗長だった回答文の様式を整えられるようになった。以上の結果から、ファインチューニングの必要性とその効果を確認した。加えて、Llama2 ベースのモデルよりも Llama3 ベースのモデルの方が一貫して精度が高く、モデルアーキテクチャに注意してモデルを選定することの重要性が示唆された。一方で、生成内容にはハルシネーションや用語の誤脱字が多く発生しており、製造業のドメイン知識が十分には学習できていなかった。今後は、製造業のドメイン知識をより適切に扱うために、継続事前学習によるドメイン知識の定着や、RAG との併用によるハルシネーション抑止の効果を検証していく。

参考文献

- [1] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities, 2024.
- [2] Kei Sawada, Tianyu Zhao, Makoto Shing, Kentaro Mitsui, Akio Kaga, Yukiya Hono, Toshiaki Wakatsuki, and Koh Mitsuda. Release of pre-trained models for the japanese language, 2024.
- [3] 大輔河原, 洋平空閑, 禎夫黒橋, 潤鈴木, 祐介宮尾. Llm-jp: 日本語に強い大規模言語モデルの研究開発を行う組織横断プロジェクト. 自然言語処理, Vol. 31, No. 1, pp. 266–279, 2024.
- [4] Masanori Hirano and Kentaro Imajo. Construction of domain-specified japanese large language model for finance through continual pre-training, 2024.
- [5] Issey Sukeda, Masahiro Suzuki, Hiroki Sakaji, and Satoshi Kodera. Jmedlora: medical domain adaptation on japanese large language models using instruction-tuning, 2023.
- [6] Yiwei Li, Huaqin Zhao, Hanqi Jiang, Yi Pan, Zhengliang Liu, Zihao Wu, Peng Shu, Jie Tian, Tianze Yang, Shaochen Xu, Yanjun Lyu, Parker Blenk, Jacob Pence, Jason Rupram, Eliza Banu, Ninghao Liu, Linbing Wang, Wenzhan Song, Xiaoming Zhai, Kenan Song, Dajiang Zhu, Beiwen Li, Xianqiao Wang, and Tianming Liu. Large language models for manufacturing, 2024.
- [7] Yuchen Xia, Nasser Jazdi, Jize Zhang, Chaitanya Shah, and Michael Weyrich. Control industrial automation system with large language models, 2024.
- [8] 岩月憲一. ドメインに特化した比較的少量のデータによる事前学習済み bert の利用可能性: 鉄鋼業における事例. 言語処理学会年次大会発表論文集, 第 28 巻, pp. PH1–8. 日本製鉄株式会社, 2022.
- [9] 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, and et al Vedanuj Goswami. Llama 2: Open foundation and fine-tuned chat models, 2023.
- [10] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, and et al Bobbie Chern. The llama 3 herd of models, 2024.
- [11] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.
- [12] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. Lima: Less is more for alignment, 2023.
- [13] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

A 付録

A.1 実験で使用したパラメータ設定

A.1.1 学習設定

表 5 実験で設定した学習設定

Learning rate	1e-5
Epoch	30
Batch size	1
Gradient accumulation steps	8
Warmup ratio	0.1
Max seq length	2048
LoRA α	64
LoRA r	8
GPU	A6000(48GB) 2 枚

A.1.2 推論設定

表 6 実験で設定した推論設定

Max new tokens	256
No repeat ngram size	2
Do sample	False

A.2 SFT するための学習データの形式

以下は、タスクを説明する指示と、文脈のある入力の組み合わせです。 要求を適切に満たす応答を書きなさい。 ### 指示： MELSEC iQ-Rシリーズのユニット間同期機能とはどのような～。 ### 応答： 複数のユニット間の信号での制御タイミングを合わせる～。

図 1 学習データの形式

A.3 llama3-youko-8b の生成例

A.3.1 ハルシネーションを含む生成例

llama3-youko-8b による生成例のうちハルシネーションを含む生成例を表 7 に示す。「質問文」に対する「SFT 実施済 LLM の生成文」の回答内容が「正解の回答文」の情報と異なる。

表 7 ハルシネーションを含む生成例

質問文	CC-Link IE フィールドは最大いくつのネットワークを構築できますか？
正解の回答文	最大 239 のネットワークを構築できます。
SFT 実施済 LLM の生成文	1 つの CCIE フィールドネット Ver.1 モードマスタ局で最大 64 台まで接続できます。

A.3.2 Rouge-L の高い生成例

llama3-youko-8b による生成例のうち Rouge-L が高かった生成例を表 8 に示す。「SFT 実施済 LLM の生成文」と「正解の回答文」を比較して、バージョンの数値が異なるのみで、他のテキスト部分は一致しているため Rouge-L が高いと考えられる。他の高い Rouge-L の生成例も同様に、「SFT 実施済 LLM の生成文」に示すように、FA 機器関連の専門用語は見られなかった。

表 8 Rouge-L の高い生成例

質問文	Q50UDEHCPU および Q100UDEHCPU に対応している GX Works2 のバージョンを教えてください。
正解の回答文	Ver.1.25B より対応しております。
SFT 実施済 LLM の生成文	Ver.1.15B より対応しております。