マイクロドメインに向けた LLM における知識活用方法の検討

角掛正弥 [1] 是枝祐太 [1] 薛雅文 [1] 住吉貴志 [1] 永塚光一 [1] 友成光 [1] 山田喬 [1] 十河泰弘 [1]
[1] 株式会社日立製作所 研究開発グループ
{masaya.tsunokake.qu, yuta.koreeda.pb, yawen.xue.wn,
takashi.sumiyoshi.bf, koichi.nagatsuka.mq, hikaru.tomonari.oj,
takashi.yamada.qf, yasuhiro.sogawa.tp}@hitachi.com

概要

大規模言語モデル（LLM）の業務活用では業務に必要な独自知識を LLM が扱う必要がある．個別の業務・製品の知識を LLM に活用させる方法として RAG が主流であるが，複数の業務・製品が複雑な知識体系を持つ小規模なマイクロドメインを形成する場合に有効な知識活用方法は明らかでない．本研究は JP1 というミドルウェア製品を対象ドメインとし，追加学習と RAG を併用したマイクロドメイン特化の有効性を検証する．追加学習ではデータ合成によりデータの量・多様性を補強する．多肢選択問題から成る JP1 の資格認定試験で評価した結果，マイクロドメイン特化により正答率が向上し，最難関の試験では合格点（70%）に達した．

1 はじめに

大規模言語モデル（LLM）の著しい発展 [1, 2, 3, 4] をうけて，その業務活用が進んでいる [5]．多くの業務は組織の規則や業務特性に根差した独自のノウハウや知識が存在する．例えばカスタマーサポート業務では，製品の仕様や不具合に関して回答するために対象製品の知識が求められる．このような業務で LLM を活用する場合，LLM は独自知識を扱う必要がある．一方で，LLM は学習していない知識に基づく生成は行えないため，業務・製品ごとのドメイン知識を LLM に活用させる方法 [6] が求められる．

ドメイン知識の活用を促進する方法には，追加学習と Retrieval-Augmented Generation（RAG） [7, 8] が存在する [5]．医療 [9]，金融 [10]，半導体デザイン [11] などの巨大なドメインでは追加学習によってドメイン特化型モデルの学習が行われているが，学習データが大量には存在しない個別の業務・製品の知識を扱う場合は RAG が主流である．一方でミドルウェアなど，複数の業務・製品が相互作用的に複雑な知識体系を成し，小規模のドメイン（**マイクロドメイン**）を形成する場合もある．しかし，マイクロドメインでの有効な知識活用方法は明らかになっていない．

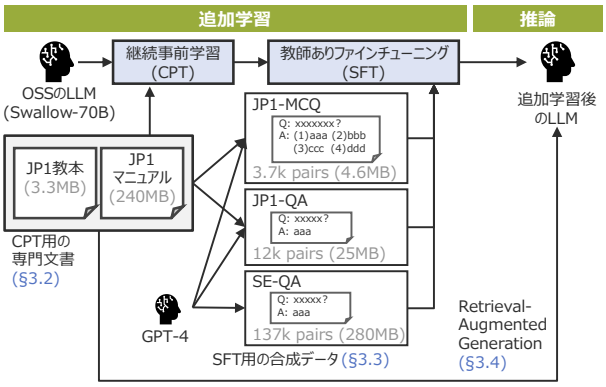

図 1: JP1 ドメインへの LLM のドメイン特化手順

本論文ではミドルウェア製品である JP1 を対象として，マイクロドメインでの知識活用方法（**マイクロドメイン特化手法**）に関するケーススタディを報告する．図 1 に本論文のマイクロドメイン特化手法を示す．JP1 はドメインとしては小規模な一方で独自の仕様に基づく複雑な知識体系を有し，その知識（語義，仕様，使用法など）は複数のマニュアル等で体系的に記述されている．そのため，広範な知識の関係性を学習するために追加学習を行う．しかし，学習に活用可能なデータが Web 上に数多く冗長に存在する主要な OSS と比較して，JP1 はデータの量や多様性に限りがある．そこで，マニュアルを基に複数形式のデータを合成し，学習データを補強する．また，RAG も併用することで質の高い JP1 の文書を推論でも再活用し，効率的な知識活用を図る．追加学習も RAG も行うことにより，下記のような効果が得られる可能性がある．

- 複数文書に跨った明示的/暗黙的な知識の関係性が学べる [5]

表 1: JP1 の資格認定試験 [12]．下段の試験ほど必要な知識が多く，難易度が高いとされる．

名称	説明	開発セット	テストセット
JP1 認定エンジニア	JP1 全般の理解，および運用に必要なテクニカルスキルを修得したエンジニアを認定する．	10 問	24 問
JP1 認定プロフェッショナル (ジョブ管理)	ジョブ管理に関する JP1 製品の導入とシステム構築ができるテクニカルスキルを修得したエンジニアを認定する．	12 問	30 問
JP1 認定コンサルタント (ジョブ管理)	ジョブ管理に関する JP1 製品について，最適なコンサルテーションができるテクニカルスキルを修得したエンジニアを認定する．	20 問	40 問

† 各試験のテスト ID は上段から順に，HMJ-130E，HMJ-1313，HMJ-1223 である．

- RAG で検索を失敗しても適切に生成できる
- RAG の検索文書を効率的に読解・活用できる

JP1 資格認定試験の多肢選択問題（**MCQ**: Multiple Choice Question）を用いて JP1 知識の活用能力を評価した結果，Swallow-70B [13] を用いたマイクロドメイン特化ではどの難易度の試験でも正答率が向上し，その有効性が確認された．特に，最難関の JP1 認定コンサルタント試験ではマイクロドメイン特化後のモデルが合格点（正答率 70%）に達した．また，追加学習と RAG の相補的な関係性も確認された．

2 JP1 について

JP1 は IT システムの運用管理を担うミドルウェアおよびソフトウェア群である．2022 年度の国内市場で同種のミドルウェアとしては最大の売上シェアを記録している．JP1 は相互作用する複数のソフトウェアで構成され，独自の仕様に基づく複雑な知識体系を有するため，小規模ながらドメインを形成していると考えられる．ユーザ数の多さや専門性の高さから，JP1 を扱うスキルの資格認定試験（**JP1 認定試験**）が公式に存在する．JP1 認定試験にはエンジニア，プロフェッショナル，コンサルタントの 3 つの異なるレベルが存在し，コンサルタントが最難関（人間の合格率が 25%）とされる．表 1 に各レベルの説明を示す．いずれも 4 択式の MCQ で構成される．図 2 に MCQ の例を示す．LLM の知識活用能力を MCQ で評価する先行研究 [14, 15, 16] と同様に，我々は JP1 認定試験で LLM を評価する．

3 マイクロドメイン特化手法

3.1 概要

我々の手法は追加学習と RAG を併用してマイクロドメイン特化を行う．本研究の目的は，小規模だが複雑な知識体系を有するドメインへの対応である．巨大なドメイン（医療など）とは異なり，学習データの量も多様性も限られる．そこで，学習データの量・多様性の補強に有効なデータ合成のアプローチ [17, 18] を活用する．すなわち，追加学習ではまず専門文書で継続事前学習（CPT）を行った後に，合成データを用いた教師ありファインチューニング（SFT）[19] を行う．表 2 に追加学習用データの統計を示す．

JP1/AJS3 で PC ジョブの終了判定を次のように設定した場合の動作として、正しいものはどれか

判定結果：しきい値による判定
警告しきい値：５
異常しきい値：１０

1. 終了コードが １０ のときは、異常終了となる．
2. 終了コードが ９ のときは、異常終了となる．
3. 終了コードが ５ のときは、異常終了となる．
4. 終了コードが －１ のときは、異常終了となる．

図 2: JP1 認定プロフェッショナルの問題例 [20]

3.2 継続事前学習

JP1 の専門文書として **JP1 マニュアル**と **JP1 教本** [21] を用いて，次単語予測タスクで学習を行う．

JP1 マニュアルは JP1 の語義や仕様，使用法などを体系的に章立てしつつ整理した文書である．我々は Web 上で配布されている JP1 Version 12 の HTML 版マニュアル[1)]を収集し，本文を抽出した．なお，HTML 版が存在しない場合は PDF 版を収集した．

JP1 教本 [21] は実際に販売されている JP1 エンジニア試験用の参考書である．我々は Microsoft Word 形式の JP1 教本を入手し，python-docx[2)] により段落と表内のテキストを抽出した．LLM は HTML 形式の方が表を読解しやすい [22] ことから，表内のテキストを HTML 形式に変換した．

1） https://itpfdoc.hitachi.co.jp/Pages/document_list/manuals/jp1v12.html
2） https://github.com/python-openxml/python-docx

表 2: 追加学習に用いる学習データの一覧

名称	用途	説明	サイズ（MB）		#文書	#トークン（M）	
JP1 マニュアル	CPT	HTML/PDF 版のマニュアルから抽出したテキスト	232.3		42,694	97.6	
JP1 教本	CPT	JP1 教本から抽出したテキスト	3.3		79	0.9	
JP1-QA	SFT	GPT-4 で合成した JP1 に関する QA	24.9	(9.1)	12,306	6.1	(2.1)
JP1-MCQ	SFT	GPT-4 で合成した JP1 に関する MCQ	4.6	(0.4)	3,661	1.1	(0.1)
SE-QA	SFT	GPT-4 で合成した SE 分野の QA	276.8	(246.6)	137,219	59.9	(53.5)

† JP1-QA/MCQ と SE-QA については，SFT の損失計算で使用される回答部分のみの統計値も丸括弧で併記している．

3.3 教師ありファインチューニング

学習用テキストは多様な形式に変換し併用することで，学習知識の活用能力が向上すると知られている [23, 24, 25]．そこで，複数形式のデータを GPT-4 [3] を用いて合成する．具体的には，JP1 に関する QA（**JP1-QA**）と MCQ（**JP1-MCQ**）を合成する．前者はカスタマーサポート業務への応用を想定して，後者は下流タスクの形式に合わせるために用いる．さらに，JP1 知識の活用にはソフトウェアエンジニアリング（SE）の知識も重要と考え，SE に関する QA（**SE-QA**）も合成する．いずれも質問部分は損失に含めずに次単語予測タスクで学習する．

JP1-QA は付録の図 3 のプロンプトを用いて，JP1 マニュアルから QA を合成する．プロンプトでは，与えたチャンクに基づき JP1 のユーザを模擬した質問とカスタマーサポート職員を模擬した回答を生成するように指示している．チャンクは PDF 版 JP1 マニュアルから作成した．詳細は付録に記す．

JP1-MCQ は付録の図 4 のプロンプトを用いて，JP1 マニュアルから MCQ を合成する．プロンプトでは，与えたチャンクに関する MCQ の質問文と回答を生成するように GPT-4 に指示している．リアルで多様な MCQ 形式の生成結果を得るために，実際の MCQ をプロンプトに与えて合成を行う．プロンプトに含める MCQ は，合成のたびに JP1 教本から 3 つを無作為に選択した．

SE-QA では SE 分野の幅広いトピックをカバーするため，まず “ソフトウェア” に関連する SE 分野のキーワードを GPT-4 に再帰的に生成させ，トピックツリーを形成する．その後，ツリーで親子関係にあるトピック一覧をサンプリングして，そのトピック一覧に基づく QA を合成する．

3.4 推論処理

1 章で述べた通り，推論時に RAG を行う．RAG 用のベクトル DB は JP1 マニュアルと JP1 教本を用いて構築する．各文書を 1,000 トークン毎にチャンキングし[3)]，各チャンクを `multilingual-e5-large` [26] でベクトル化して FAISS [27] に索引付けする．

4 実験

4.1 評価方法

JP1 知識の活用能力と汎用性能の観点で LLM を評価し，マイクロドメイン特化の有効性を検証する．

前者の評価では，2 節の JP1 認定試験を用いる．各試験を開発・テストセットに分割し，テストセットでの完全一致の正答率を算出する．表 1 に各セットの件数を示す．推論時は In-Context Learning（ICL）を行い，LLM が選択肢の番号を出力するように促す．ICL には開発セットから無作為に選択した 5 つの MCQ を用いる．本論文では 10 回の無作為選択における平均正答率を報告する．なお，[37] に基づきテストセットと学習データの重複する部分文字列を確認したが，テストセットのリークはなかった．

後者の評価では，Swallow-70B [13] の公式評価と同じスクリプト[4)]を用い，日本語の一般的な言語理解・生成タスクの性能を評価する．

4.2 評価モデル

マイクロドメイン特化用モデルとして，2024 年 5 月時点で安定した日本語性能を示していた Swallow-70B [13] を用いた．また，参考として `gpt-35-turbo-16k (0613)` と `gpt-4-32k (0613)` [38] でも評価を行った．

4.3 学習設定

CPT ではバッチサイズを 80，学習率を 1.0×10^{-5} で 3 エポック学習した．SFT ではバッチサイズを 256，学習率を 1.0×10^{-5} から線形減衰させ 3 エポック学習した．最適化は AdamW [39] を用い，

3) 隣接チャンクは 200 トークンの重なりを持たせる．

4) https://github.com/swallow-llm/swallow-evaluation with commit hash 04948a0

表 3: JP1 認定試験（多肢選択問題）における正答率．カラムごとに Swallow-70B シリーズで最も高い正答率に下線をひき，全モデルで最も高い正答率を**太字**にしている．

				エンジニア		プロフェッショナル		コンサルタント		マクロ平均	
Model	CPT	SFT	RAG→	No	Yes	No	Yes	No	Yes	No	Yes
Swallow-70B	—	—		62%	83%	51%	57%	43%	63%	52%	68%
Swallow-70B	Yes	—		77%	88%	47%	58%	54%	64%	59%	70%
Swallow-70B	Yes	MCQ のみ		**90%**	90%	41%	60%	**58%**	**70%**	63%	73%
Swallow-70B	Yes	QA のみ		81%	88%	50%	52%	56%	55%	63%	65%
Swallow-70B	Yes	MCQ & QA		85%	**92%**	41%	55%	49%	66%	58%	71%
GPT-3.5	—	—		65%	71%	34%	43%	31%	53%	43%	56%
GPT-4 [2, 3]	—	—		85%	91%	**70%**	**76%**	47%	65%	**67%**	**77%**

表 4: 日本語ベンチマークでの評価結果．Swallow-70B より優る/劣る結果を赤色/青色に色付けしている．

Model	CPT	SFT	JCQA	JEMHopQA	NIILC	JSQuAD	XL-Sum	MGSM	JMMLU	JHumanEval	En-Ja	Ja-En	Avg.
Swallow-70B	—	—	91.7	63.4	69.7	92.1	22.5	47.2	57.6	22.1	30.3	23.0	52.0
Swallow-70B	Yes	—	88.2	53.7	63.6	90.3	22.0	44.4	58.7	21.0	27.4	21.3	49.1
Swallow-70B	Yes	MCQ のみ	93.7	59.9	64.7	90.3	20.4	42.0	57.8	18.7	27.7	21.7	49.7
Swallow-70B	Yes	MCQ & QA	95.5	63.0	63.0	89.7	21.9	47.6	57.6	24.0	28.6	21.5	51.2

† **JCQA** [28]: 常識推論を要する 5 択 MCQ の正答率，**JEMHopQA** [29]: マルチホップ推論を要する QA の文字 F1，**NIILC** [30]: 百科事典に関する QA の文字 F1，**JSQuAD**1 [28]: Wikipedia 記事の機械読解での文字 F，**XL-Sum** [31]: BBC 記事の抽象型要約での ROUGE-2 [32]，**MGSM** [33]: 小学校の数学文章問題（GSM8K）の日本語訳における完全一致の正答率，**JMMLU** [34]: 知識を要する 4 択 MCQ の正答率，**JHumanEval** [35]: コード生成やマルチターン会話での pass@1 の正答率，**En-Ja/Ja-En** [36]: 英語-日本語と日本語-英語の翻訳タスク（WMT'20）の BLEU，**Avg.**: 全スコアのマクロ平均

DeepSpeed ZeRO-3[5] で複数の H100 GPU で学習した．

4.4 JP1 認定試験での評価結果

表 3 に結果を示す．下線部分に注目すると，どの試験でも CPT と SFT の実施後に RAG を行う方法が Swallow-70B で最も高い正答率であり，本論文のマイクロドメイン特化の有効性が確認できる．特にエンジニアとコンサルタントで効果的で，最難関とされるコンサルタントでは合格点の 70%に達した．

プロフェッショナルを除き，CPT のみ実施した Swallow-70B は RAG の有無に関わらず正答率が向上しており，CPT 単体の有効性も確認できる．プロフェッショナルは数学タスクで求められるような演繹的な推論能力が必要な傾向があり，単純な CPT が有効でなかった可能性がある．また，RAG 無しのプロフェッショナルと RAG 有りのエンジニア以外では，SFT（MCQ のみ）が SFT 後モデルの中で最も正答率が高く，下流タスクと同一の形式での SFT の重要性が示唆される．

マクロ平均に注目すると，追加学習後のモデルは RAG によって正答率が向上する傾向にある．同様に，SFT（QA のみ）を除き，RAG 有りでの正答率は追加学習後に向上する傾向にある．特に，CPT 単体や SFT（MCQ のみ）は全試験で向上しており，追加学習と RAG の相補的な傾向が確認できる．

エンジニアとコンサルタントでは RAG の有無に関わらず GPT-4 を上回るケースがある．マイクロドメインであっても追加学習によって，より大規模な LLM の知識活用能力を小規模なオンプレミス環境の LLM で上回ることも可能とわかる．機密情報を扱うことも多い業務活用では有益な傾向である．

4.5 汎用性能の評価結果

表 4 に結果を示す．追加学習後に一般的な NLP タスクの平均性能は落ちている．一方，追加学習後の中では SFT（MCQ&QA）の平均性能が最も高く，複数形式データでの SFT が汎化性能の低下に効果的と示唆される．MCQ である JCQA は SFT（MCQ のみ）により性能が向上しており，下流タスクと同一の形式で学習することの重要性も示唆される．

5 おわりに

本論文では，複雑な知識体系を有する JP1 を対象に，追加学習と RAG を併用したマイクロドメイン特化手法を検証した．本手法によりどの難易度の JP1 認定試験でも正答率の向上が確認できた．今後の課題として，他の LLM での有効性の検証や，DPO [40] 等のアライメントも交えつつ問い合わせ等に対する回答性能も検証することが挙げられる．

5） https://github.com/microsoft/DeepSpeed

謝辞

計算機環境の整備に尽力くださった清水正明氏と，研究にご助言くださった Chetan Gupta 氏，影広達彦氏，鯨井俊宏氏，尾崎太亮氏，笹沢裕一氏に感謝申し上げる．

参考文献

[1] 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, and Dario Amodei. Language models are few-shot learners. In **Advances in Neural Information Processing Systems**, 2020.

[2] OpenAI. GPT-4. **OpenAI Blog**, 2023. https://openai.com/research/gpt-4（2024-08-16 閲覧）.

[3] OpenAI. GPT-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

[4] Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku. **Anthropic Technical Report**, 2024. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf（2024-03-15 閲覧）.

[5] Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K. Qiu, and Lili Qiu. Retrieval augmented generation (RAG) and beyond: A comprehensive survey on how to make your LLMs use external data more wisely. **arXiv preprint arXiv:2409.14924**, 2024.

[6] Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Jian Pei, and Liang Zhao. Domain specialization as the key to make large language models disruptive: A comprehensive survey. **arXiv preprint arXiv:2305.18703v6**, 2023.

[7] 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. **Advances in Neural Information Processing Systems**, Vol. 33, pp. 9459–9474, 2020.

[8] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. **arXiv preprint arXiv:2312.10997v5**, 2024.

[9] 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, and Vivek Natarajan. Towards expert-level medical question answering with large language models. **arXiv preprint arXiv:2305.09617v1**, 2023.

[10] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A large language model for finance. **arXiv preprint arXiv:2303.17564v3**, 2023.

[11] Mingjie Liu, Teodor-Dumitru Ene, Robert Kirby, Chris Cheng, Nathaniel Pinckney, Rongjian Liang, Jonah Alben, Himyanshu Anand, Sanmitra Banerjee, Ismet Bayraktaroglu, Bonita Bhaskaran, Bryan Catanzaro, Arjun Chaudhuri, Sharon Clay, Bill Dally, Laura Dang, Parikshit Deshpande, Siddhanth Dhodhi, Sameer Halepete, Eric Hill, Jiashang Hu, Sumit Jain, Ankit Jindal, Brucek Khailany, George Kokai, Kishor Kunal, Xiaowei Li, Charley Lind, Hao Liu, Stuart Oberman, Sujeet Omar, Sreedhar Pratty, Jonathan Raiman, Ambar Sarkar, Zhengjiang Shao, Hanfei Sun, Pratik P Suthar, Varun Tej, Walker Turner, Kaizhe Xu, and Haoxing Ren. ChipNeMo: Domain-adapted llms for chip design. **arXiv preprint arXiv:2311.00176v4**, 2023.

[12] JP1 技術者資格認定制度 (Version 13 対応). https://www.hitachi-ac.co.jp/service/opcourse/license/jp1.html（2025-01-09 閲覧）.

[13] 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. In **First Conference on Language Modeling**, 2024.

[14] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In **International Conference on Learning Representations**, 2021.

[15] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. **Applied Sciences**, Vol. 11, No. 14, 2021.

[16] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In **Proceedings of the Conference on Health, Inference, and Learning**, Vol. 174 of **Proceedings of Machine Learning Research**, pp. 248–260. PMLR, 07–08 Apr 2022.

[17] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 13484–13508, July 2023.

[18] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. WizardLM: Empowering large pre-trained language models to follow complex instructions. In **The Twelfth International Conference on Learning Representations**, 2024.

[19] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. **arXiv preprint arXiv:2306.11644v1**, 2023.

[20] JP1 試験一覧と出題範囲. https://www.hitachi.co.jp/Prod/comp/soft1/jp1/introduction/cert/exam_jp1/index.html（2025-01-09 閲覧）.

[21] 株式会社日立製作所. IT Service Management 教科書 JP1 認定エンジニア V13 対応. 翔泳社, 2023.

[22] Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In **Proceedings of the 17th ACM International Conference on Web Search and Data Mining**, pp. 645–654, 2024.

[23] Daixuan Cheng, Shaohan Huang, and Furu Wei. Adapting large language models via reading comprehension. In **The Twelfth International Conference on Learning Representations**, 2024.

[24] Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. Instruction pre-training: Language models are supervised multitask learners. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 2529–2550, 2024.

[25] Ingo Ziegler, Abdullatif Köksal, Desmond Elliott, and Hinrich Schütze. CRAFT your dataset: Task-specific synthetic dataset generation through corpus retrieval and augmentation. **arXiv preprint arXiv:2409.02098v1**, 2024.

[26] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual E5 text embeddings: A technical report. **arXiv preprint arXiv:2402.05672**, 2024.

[27] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. **arXiv preprint arXiv:2401.08281v1**, 2024.

[28] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**. European Language Resources Association, June 2022.

[29] Ai Ishii, Naoya Inoue, Hisami Suzuki, and Satoshi Sekine. JEMHopQA: Dataset for Japanese explainable multi-hop question answering. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 9515–9525, 2024.

[30] 関根聡. 百科事典を対象とした質問応答システムの開発. 言語処理学会 第 9 回年次大会 発表論文集, pp. 637–640, 2003.

[31] Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In **Findings of the Association for Computational Linguistics (ACL)**, 2021.

[32] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, 2004.

[33] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. In **The Eleventh International Conference on Learning Representations**, 2023.

[34] 尹子旗, 王昊, 堀尾海斗, 河原大輔, 関根聡. プロンプトの丁寧さと大規模言語モデルの性能の関係検証. 言語処理学会 第 30 回年次大会 発表論文集, 2024.

[35] 佐藤美唯, 野志歩, 梶浦照乃, 倉光君郎. LLM は日本語追加学習により言語間知識転移を起こすのか？ 言語処理学会 第 30 回年次大会 発表論文集, 2024.

[36] Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2020 conference on machine translation (WMT20). In **Proceedings of the Fifth Conference on Machine Translation**, 2020.

[37] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics**, 2022.

[38] Microsoft. Azure OpenAI service models. https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models（2025-01-07 閲覧）.

[39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **International Conference on Learning Representations**, 2019.

[40] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: your language model is secretly a reward model. In **Proceedings of the 37th International Conference on Neural Information Processing Systems**, 2024.

```
以下の文書はソフトウェア製品のマニュアルです。\n
次の手順でこの文書に関する質問、回答、回答の根拠を日本語で作成してください。\n
\n
1.␣この製品を使うユーザーになりきって、Question: という文字列の後に、与えられた文書に基づいて、実際に起きたトラブルや聞きたいこと、分からないことなどを質問として作成して下さい。その際、トラブルに加えて、やりたいことや、状況説明、トラブルシューティングに役立つ情報 (例えば、実行環境、バージョン、実行コマンド、エラー詳細) なども記載するといいかもしれません。\n
2.␣次に、この製品を扱う会社の一流のコンタクトセンターの職員になりきって、Answer: という文字列の後に、文書に基づいて初心者のユーザーの質問に丁寧に答えて下さい。\n
3.␣最後に、Citation: という文字列の後に、回答の根拠となった文書内の記述をそのままコピー&ペーストして書いてください (変更は加えないで抜き出してください)。\n
\n
注意点として、質問や回答を複数個つくったりしないでください。また、Question:, Answer:, Citation: 以外のフォーマットを使わないでください。 \n
\n
では、質問を以下の文書の情報を基に日本語で作成してください。\n
文書:{chunk}
```

図 3: JP1-QA の合成用プロンプト．JP1 マニュアルから作成したチャンクは {*chunk*} で与えられる．

```
次の手順でこの文書に関する選択問題とその回答、回答の解説を日本語で作成してください。\n
\n
1. この製品に関する知識を測るために、Question：という文字列の後に、与えられた文書に基づいて、以下のような例を参考にして選択問題を一つだけ作成してください。\n
2. 次に、Answer：という文字列の後に、選択問題の答えを文書に基づいて答えて下さい。\n
3. 最後に、Explanation：という文字列の後に、問題の解説を書いて下さい。\n
\n
以下が参考となる問題の例です。\n
\n
␣例 1)\n
Question：{Question #1}\n
Choices：\n
{Choices #1}\n
Answer：{Answer #1}\n
Explanation：{Explanation #1}\n
...
\n
注意点として、質問や回答を複数個つくったりしないでください。また、Question：,␣Answer：,␣Explanation：以外のフォーマットを使わないでください。\n
\n
では、質問を以下の文書の情報を基に日本語で選択問題と回答、解説を作成してください。\n
文書：{chunk}
```

図 4: JP1-MCQ の合成用プロンプト．JP1 マニュアルから作成したチャンクは {*chunk*} で与えられる．また，無作為に選択された 3 件の MCQ（図では 2 件を省略）が ICL サンプルとして与えられる．

A　CPT 用データの処理

収集した JP1 マニュアルの加工処理を示す．

1. テキスト抽出：HTML 版はマニュアル本文のテキストのみを抽出する．PDF 版は PDFMiner[6] でテキストへ変換する．
2. ノイズ除去：HTML 版は “目次”，“索引”，“変更内容” などで始まるページを除去する．PDF 版は，ページ番号や目次を示す行を除去する．

B　SFT データの合成の詳細

図 3 と図 4 にそれぞれ JP1-QA と JP1-MCQ の合成に用いたプロンプトを示す．JP1-QA の合成には `gpt-4-1106-preview` を，SE-QA と JP1-MCQ の合成には `gpt-4-32k (0613)` [38] を用いた．JP1-QA の合成に用いるチャンクは，PDF 版 JP1 マニュアルから pypdf[7] で抽出したテキストを 5 ページ単位で分割して作成した．隣接チャンクは 2 ページ分の重なりを持たせ，ページの境界には特殊記号を挿入した．JP-MCQ の合成用チャンクも同様に作成したが，隣接チャンクは 4 ページ分の重なりを持たせた．

C　実験の補足

学習設定　CPT ではコンテキスト長を 2,048 として学習データを分割した．SFT では最大コンテキスト長を 4,096 とした．AdamW [39] は，$\beta_1 = 0.9$, $\beta_2 = 0.95$，weight decay を 0.1 とした．なお，テンソル並列化やパイプライン並列化は使用していない．

表 5: 指示チューニングモデルとの比較．最も高い正答率に下線を引いている．

CPT	SFT	RAG→	ENG No	ENG Yes	PRO No	PRO Yes	CON No	CON Yes
Swallow-70B								
—	—		62%	83%	51%	57%	43%	63%
Yes	—		77%	88%	47%	58%	54%	64%
Yes	MCQ のみ		90%	90%	41%	60%	58%	70%
Swallow-70B-instruct			62%	82%	52%	56%	40%	60%

† ENG: エンジニア，PRO: プロフェッショナル，CON: コンサルタント

推論設定　JP1 認定試験での評価は，最大出力長を 1 とした貪欲的生成で行った．RAG では質問文と選択肢をクエリとして，3.4 節の FAISS から L2 距離が上位 5 件のチャンクを取得する．

指示チューニングモデルとの比較　表 5 に，Swallow-70b の指示チューニングモデルである Swallow-70b-instruct[8] での JP1 認定試験の正答率を示す．RAG 無しのプロフェッショナルを除き，Swallow-70B は指示チューニング後に正答率が向上していない．このことから，マイクロドメイン特化における追加学習は指示チューニングの効果を単に代替したものではないとわかる．

6） https://github.com/pdfminer/pdfminer.six
7） https://pypi.org/project/pypdf/
8） https://huggingface.co/tokyotech-llm/Swallow-70b-instruct-hf