物性値に関する相対感覚の LLM への導入の効果

早川和磨 [1] 旭良司 [2] 佐々木裕 [1]
[1] 豊田工業大学 知能数理研究室　[2] 名古屋大学
{sd22072,yutaka.sasaki}@toyota-ti.ac.jp, asahi.ryoji.d9@f.mail.nagoya-u.ac.jp

概要

本稿は，大規模言語モデル（Large Language Model; LLM）に物質に関する専門知識を導入する際に，物性値自体を追加学習するよりも，物性値間の相対関係を追加学習する方が効果的であることを示す．近年の LLM の飛躍的な発展にともない，対話や質問応答，翻訳など言語理解・言語生成に関する技術の実応用が加速している [1]．しかし，学習データの不足および物理現象の複雑さに起因する要因により，高度な専門性を要する分野において LLM の性能は依然として限定的である．本報告では，バンドギャップ等の物性値を LLM に導入する際に 2 つの化学組成の持つ物性値の相対関係を導入する．物性値の相対関係に関するファインチューニングを行うことにより，相対関係に関する正解率が向上することを示す．さらに，燃料電池に関する QA データセットにおける性能が，物性値の相対関係の学習により向上することも示す．

1　はじめに

大規模言語モデル（Large Language Model; LLM）は様々な実用的な言語生成タスクにおいて利用されている [2]．しかし，高度な専門性を要するドメインにおいては，汎用的な LLM の性能は依然として限定的である [3]．その主な要因は，LLM の事前学習データにおける専門情報の不足にある．特定の科学技術分野における専門的文書の量は限られており，また企業や研究機関が保有するクローズドな実験データなどは学習対象に含まれていない [4]．そのため，専門領域特有の高度専門的な概念を言語モデルとして獲得できていないという課題がある．このような未学習の専門知識を補完し，特定のドメインへ適応させるための主要なアプローチとして検索拡張生成 [5]（Retrieval-Augmented Generation; RAG）とファインチューニング [6]（Fine-tuning）という 2 つの方法がある．

RAG に関しては，新素材や新材料に関する生成において，RAG の検索対象に直接生成の参考にできる情報が含まれていないという問題がある [7]．また，検索された断片的な情報を統合して深い論理的推論を行う能力は，依然として LLM のその分野における事前学習済み言語モデルの能力に依存している [8]．

ファインチューニングにおいても，専門領域の学習データは一般的なテキストと比較して極めて少量であることが多く，大規模なパラメータを持つモデルを十分に最適化するには至らない場合がある [9]．特に材料科学などのドメインでは，テキスト情報だけでなく，化学組成や物性値といった数値データの相関や物理的な法則性をモデル内部に正確に埋め込むことは難しく，単なる言語的なパターンの学習に留まってしまうという限界がある [10]．

そのため，本研究では LLM に物性値といった数値情報や物性知識をどのように追加学習することが効果的であるかを明らかにすることを目的とする．材料固有の物理的な性質をより直感的に捉えるためのアプローチとして，物性に関する相対感覚の概念を導入する．具体的には，特定の物性値の絶対値を直接予測させるのではなく，複数の物質間における物理量の比較タスクを学習プロセスに組み込む．このような比較学習を通じて，モデルは単なるテキストとしての専門用語だけでなく，物質間に存在する相関関係や物理的なポテンシャルの高低といった感覚的な関係性を獲得することが期待される．

2　関連研究

2.1　絶対的な数値情報の学習

LLM は，数値であっても文字列とみなしてトークン単位の処理を行う [11] ため，材料科学において極めて重要な物理量や組成比を数値として正確に理解・出力することができない．これに対し，特定のドメイン知識を反映した数値情報の学習手法 [12] が

研究されている．例えば，材料の組成式や結晶構造パラメータを直接トークン化するのではなく，数値としての意味を保持したまま学習させる手法 [13] や科学データセットを用いた継続事前学習（Continual Pre-training）[14] が提案されている．材料科学に特化した LLM である MatSci-BERT[15] や，より大規模な Llama ベースの調整モデルでは，材料特性の絶対値を回帰タスクとして解く能力が評価されており単なるテキスト生成を超えた物性予測への応用が進んでいる [16]．具体的は応用例として，Xie らは材料科学文献で継続事前学習を行った Darwin を提案し，組成式から結晶構造やエネルギーなどの数値を直接予測するタスクにおいて，従来の機械学習モデルに匹敵する精度を達成している [17]．

2.2 相対的な順序情報の学習

材料探索においては絶対的な数値の予測精度もさることながら相対的な順序関係の把握が実用上重要となる．これは実験条件の最適化やベイズ最適化の初期探索において，有望な候補を絞り込む指針となるためである．近年の研究では材料の物性値を直接予測させる代わりに，順序関係を学習させる手法が有効であることが示されている．Li らは，分子選択やベイズ最適化の文脈において，数値を直接予測するよりも順序関係をターゲットとする方が，ノイズに強く，探索効率が大幅に向上することを実証している [18]．このような順序情報の学習により，LLM は材料間の相関ルールや物理的な傾向をよりロバストに獲得することが可能となる．これは，高精度な教師データが不足している環境下での材料設計において，LLM を物理的妥当性に基づいたヒューリスティックな探索エンジンとして活用する有効なアプローチである．

2.3 Battery Device Data QA

本研究では，電池分野におけるドメイン知識の理解度を評価するためのベンチマークとして，「Battery-Device-Data-QA」データセット [19]（以下，バッテリー QA）を用いる．このデータセットは，リチウムイオン電池を中心とした蓄電デバイスに関する専門的な知識を問うために構築されたものである．その特徴は，単なる文献情報の要約にとどまらず，テキストマイニング技術に基づき論文中の数値データや技術仕様が精緻に構造化されている点にある．

本データセットは，与えられた文脈の中から正解となる該当箇所を直接抜き出す**抽出型 QA（Extractive QA）**のタスク形式を採用している．データは「文脈（Context）」，「質問（Question）」，および文脈内における「正解（Answer）」の 3 組で構成されるが，ここで正解は生成されたテキストではなく，文脈中の特定の**スパン（部分文字列）**として定義されている．つまり，モデルには回答をゼロから生成する能力ではなく，文脈内から適切な情報の所在（開始位置と終了位置）を正確に特定する能力が求められる．具体的なデータの例を表 1 に示す．同一の文脈から正極（Cathode）や負極（Anode）の情報を個別に特定するものなど，多岐にわたる質問が含まれている．

本研究においてバッテリー QA を採用する意義は，比較タスクによる感覚値の学習によって得られた物理的な感覚が，具体的な専門知識を問う QA タスクにおいて，どの程度回答の正確性や論理的妥当性の向上に寄与するかを検証する点にある．

3 提案手法

3.1 提案手法の概要

本研究では，LLM である GPT-4o[20] を基盤モデルとして採用し[1)]，2 つの材料間の物性値の大小を比較する比較タスクを通じて学習を行う． 本手法は，単一材料の物性値を個別に予測させるのではなく，ペア間の大小関係を判定させることで，材料データ全体に潜む相関関係や順序構造をモデルに学習させる．このプロセスを通じて，特定の比較タスクの精度向上のみならず，材料間の類似性や隠れた物理的規則性の理解が深まり，結果として未知の組成に対する物性の推論や他の関連タスクへの高い汎用性能を獲得することを目指す．

3.2 データセット

本研究では，材料科学におけるデータセットの MatBench[21] を採用する．MatBench に含まれる多様な物質群の中から，材料の組成式および対応する物性値を抽出して使用する．LLM に材料間の相対的な関係性を学習させるため，元のデータセットを加工し，以下の 3 要素で構成されるペア比較プロンプトを生成する．

1） 予備実験の結果，GPT-5 系にしても本タスクにおいては性能差はなく，コストが増加するため GPT-4o を採用した．

表 1　バッテリー QA データセットに含まれる質問応答ペアの例

Context (文脈)	Question (質問)	Answer (回答)
The blended slurry was then cast onto a clean current collector (**Al foil** for the cathode and **Cu foil** for the anode) and dried at 90 °C under vacuum overnight.	What is the cathode?	Al foil
(Same context as above)	What is the anode?	Cu foil

- **入力（Input）**：2 つの材料の組成式
- **指示（Instruction）**：提示された 2 つの材料のうち，より高い [物性名] を持つ材料の組成式を回答してくださいという自然言語による英語の指示
- **出力（Label）**：物性値の大小関係に基づき，正解となる材料の組成式

この形式のデータを多数作成し学習させることにより，LLM が数値そのものではなく，物性に関する相対的な順序関係を習得できるかを検証する．

具体的な実験設定として，本稿では Matbench の `[matbench_expt_gap]` を対象とした．ここからランダムサンプリングにより合計 5,000 件のペアデータセットを構築し，学習用データとして 3,000 件，検証用・テスト用データとして 1,000 件にそれぞれ分割した．

3.3　比較タスクによるファインチューニング

GPT-4o のファインチューニングには，OpenAI が提供する API を用いた教師あり学習（Supervised Fine-tuning）[22] を適用する． 学習プロセスでは，上述のプロンプトを入力し，LLM が正しい比較結果をテキストとして生成するように最適化を行う．具体的には，モデルが生成するトークンの次トークン予測確率に基づき，正解のラベルテキスト（例："Material A"）の出現確率を最大化するように学習が進められる．

この比較学習を経たモデルは，単なる数値の記憶ではなく，材料間の相対的な優劣関係を言語的な文脈として理解するため，未知の材料に対する予測においても，物理的に妥当な関係を維持した回答が可能となることを期待する．このようにして養われた専門家のような直感的な順序感は，学習に用いた特定の物性比較に留まらず，未知の材料に対する性質推論や，2.4.1 項で述べたバッテリー QA のような高度な専門知識を問うタスクにおいても，物理的整合性の高い回答を導き出すための基盤となると期待される．

4　結果と考察

本章では，提案手法の有効性を評価するために行った，比較タスクでの実験とバッテリー QA の実験について説明する．4.1 節では，比較タスクでの実験について説明し，4.2 節では，バッテリー QA の実験結果と結果に対する考察を述べる．

4.1　相対比較タスクでの学習・実験

4.1.1　実験設定

本実験では，2 つの化学組成の物性値の大小比較を答えるタスクに関する評価を行った．大規模言語モデル（GPT-4o）をベースラインとし，提案手法であるファインチューニングを施したモデルとの比較検証を行った．評価用データセットには，1,000 組の比較対象ペアと提示順序に依存して回答を決定しないようにするため、提示順を入れ替えた同一ペアを追加し、計 2,000 件の評価サンプルを構築した．実験の比較対象として，以下の条件でモデルを作成した．

1. **ベースライン (GPT-4o)**: ファインチューニングを行っていない初期モデル．
2. **数値予測チューニング**: 数値理解能力の向上を目的としたチューニングモデル．
3. **Targeted Property Comparison (bg 比較チューニング)**: バンドギャップ（bg）の大小比較データを用いたモデル．データ数（500〜3000 件）による正解率の推移を検証した．
4. **Cross-Property Comparison (誘電率比較チューニング)**: ターゲットとは異なる物性（誘電率）の比較データを用いたモデル．タスク特化の影響を検証するため，データ数（500〜3000 件）で実施した．
5. **Mix / Mix チューニング**: バンドギャップと誘電率の二つを 1500 件ずつ学習させたモデル．

4.1.2　実験結果

各モデルにおける比較タスクの正解率（Accuracy）を表 2 に示す．

表 2 bg 比較タスクにおける各モデルの実験結果

Model / Method	Data Size	Accuracy (%)
ベースライン (GPT-4o)	-	63.75
数値予測チューニング	3000	71.50
bg 比較チューニング	500	79.45
bg 比較チューニング	1000	80.10
bg 比較チューニング	1500	81.55
bg 比較チューニング	2000	81.30
bg 比較チューニング	2500	82.10
bg 比較チューニング	3000	**82.80**
誘電率比較チューニング	500	38.30
誘電率比較チューニング	1000	29.85
誘電率比較チューニング	1500	39.10
誘電率比較チューニング	2000	60.80
誘電率比較チューニング	2500	53.15
誘電率比較チューニング	3000	42.80
Mix チューニング	3000	80.60

ベースラインである GPT-4o の正解率は 63.75%であった．これに対し，数値予測を行ったモデルは 71.5%となり，一定の精度向上が見られた． 最も高い精度を示したのは「bg 比較チューニング」であり，学習データ数 3,000 件のモデルで 82.8%を記録した．また，「Mix チューニング」および「Sepa-mix チューニング」においても，それぞれ 80.6%，80.05%と高い精度を維持した．

一方で，「誘電率比較チューニング」に関しては，多くの条件でベースラインを下回る結果となり，特にデータ数 1,000 件では 29.85%まで低下するなど，著しく不安定な挙動を示した．

4.2 Battery QA での実験

4.2.1 実験設定

本実験では，バッテリー材料の物性値（バンドギャップ，誘電率など）に関する知識を問うデータセットを用いて，モデルの回答精度を検証した．評価指標には，正解と完全に一致した割合を示す EM (Exact Match) を採用した．

比較対象は先の実験で使用したものと同じである．

4.2.2 実験結果

Battery QA タスクにおける各モデルの実験結果を表 3 に示す．

実験の結果，bg 比較チューニング 3000 モデルが最も高い EM スコア（45.20）を記録した．これは

表 3 Battery QA における Exact Match(EM) スコア

Model / Method	Data Size	EM (%)
ベースライン (GPT-4o)	-	42.62
数値予測チューニング	3000	43.56
bg 比較チューニング	1500	44.26
bg 比較チューニング	3000	**45.20**
誘電率比較チューニング	1500	43.56
Mix チューニング	3000	40.28

ベースラインである GPT-4o（42.62）を 2.58 ポイント上回る結果である．一方で，学習データをバンドギャップと誘電率で混合した mix チューニングモデルは最も低いスコア（36.30）となった．バンドギャップに関する相対比較のファインチューニングを行うことで一見直接影響がなさそうなバッテリー QA タスクの正解率が向上するのは予想外の発見であった．直接物性値を教えるファインチューニングよりも相対比較を教える方が効果があることも興味深い．数値予測チューニングではスコアがあまり改善されていないことから，単に化学組成の用語に LLM が領域適合したということではない．

5 おわりに

本研究では，LLM においてバッテリー分野の専門的なドメイン知識をより正確に扱うため，物理量の比較タスクを用いたファインチューニング手法を提案し物性に関する相対感覚値を学習することの有効性を明らかにした．

実験の結果，以下の 3 点が明らかとなった．(1) ターゲットとする物性の比較学習を行うことで，LLM は物性値の相対感覚を向上させることができることが示された．bg 比較チューニングモデルは，比較タスクにおいて 82.8%の正解率を記録した．(2) この比較学習によって得られた相対感覚は，専門的な「Battery-Device-Data-QA」タスクにおいても有効であり，GPT-4o を超える 45.20%の EM スコアをを達成した．(3) 学習データの一貫性の重要性である．異なる物性（誘電率）を混合して学習させた場合，正解率が低下する現象が観測された．

今後の課題は，タスク間に干渉を起こさず，むしろ相乗効果を生むようなプロンプトエンジニアリングや学習スケジュールの最適化することと，モデルが「なぜその数値を妥当と判断したのか」という推論の根拠を，学習した比較知識に基づいて説明可能にすることである．

謝辞

この成果は，NEDO（国立研究開発法人新エネルギー・産業技術総合開発機構）の委託業務（JPNP25002）の結果得られたものです.

参考文献

[1] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, et al. A survey of large language models. **arXiv preprint arXiv:2303.18223**, 2023.

[2] OpenAI. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

[3] Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Zilong Liu, Payal Chandak, Shengchao Liu, Hans Vanhaesebrouck, Marinka Zitnik, et al. Scientific discovery in the age of artificial intelligence. **Nature**, Vol. 620, No. 7972, pp. 47–60, 2023.

[4] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. **arXiv preprint arXiv:2211.09085**, 2022.

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

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, 2019.

[7] Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. Seven failure points when engineering a retrieval augmented generation system. In **Proceedings of the 3rd International Conference on AI Engineering - Software Engineering for AI (CAIN)**, Lisbon, Portugal, 2024. ACM.

[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. 2024.

[9] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In **International Conference on Machine Learning**, pp. 15696–15715. PMLR, 2023.

[10] Guo Leonid, et al. Do large language models understand chemistry? a conversation with chatgpt. **Journal of Chemical Information and Modeling**, Vol. 63, No. 11, pp. 3343–3351, 2023.

[11] Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do NLP models know numbers? probing numeracy in embeddings. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 5307–5315, Hong Kong, China, November 2019. Association for Computational Linguistics.

[12] Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do NLP models know numbers? probing numeracy in embeddings. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 5307–5315, 2019.

[13] Yury Gorishniy, Ivan Rubachev, and Artem Babenko. Embeddings for numerical features in tabular deep learning. In **Advances in Neural Information Processing Systems**, Vol. 35, pp. 24291–24303, 2022.

[14] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8342–8360, Online, July 2020. Association for Computational Linguistics.

[15] Tanishq Gupta, Mohammad Zaki, N. M. Anoop Krishnan, and Mausam. Matscibert: A materials domain language model for text mining and information extraction. **npj Computational Materials**, Vol. 8, No. 1, p. 102, 2022.

[16] Amalie Trewartha, Nicholas Walker, Haoyan Huo, Sanghoon Lee, Kevin Cruse, John Dagdelen, Alexander Dunn, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Quantifying the advantage of domain-specific pretraining on named entity recognition tasks in materials science. **Patterns**, Vol. 3, No. 4, p. 100488, 2022.

[17] Tong Xie, et al. Darwin: A language model for physics-informed materials synthesis. **arXiv preprint arXiv:2308.13413**, 2023.

[18] Cheng-Hao Li, Kevin Cherry, Anirban Deshwal, and Janardhan Rao Doppa. Ranking over regression for bayesian optimization and molecule selection. **arXiv preprint arXiv:2106.12644**, 2021.

[19] Shu Huang and Jacqueline M Cole. Batterybert: A pretrained language model for battery database enhancement. **J. Chem. Inf. Model.**, p. DOI: 10.1021/acs.jcim.2c00035, 2022.

[20] OpenAI. GPT-4o API. https://platform.openai.com/docs/models/gpt-4o, 2024. Accessed: December 2025. Model version: gpt-4o.

[21] Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: The matbench test set and automatminer reference algorithm. **npj Computational Materials**, Vol. 6, No. 1, p. 138, 2020.

[22] Yu Song, Siyuan Zhang, Bowen Li, Zhaorun Wang, et al. Large language models for materials science: A comprehensive survey. **arXiv preprint arXiv:2307.02239**, 2023.

A　プロンプトの例

物性値の相対比較に関するプロンプトの例を表 4 に示す.

表 4　Example of the pair-wise comparison prompt format constructed from the Matbench dataset. The model is instructed to identify the material with the higher band gap.

Component	Content
Input	• $Li_1Ga_1O_2$ • $Ag_8Ge_1S_6$
Instruction	Which of the following two materials has the larger band gap? Consider the band gap in units of electron volts (eV). You must output only the chemical formula of the material with the larger band gap.
Label	$Li_1Ga_1O_2$

B　相対比較に関するチューニング結果に関する考察

実験結果より以下の知見が得られた.

- ターゲットタスクに即した学習データの有効性である．bg 比較チューニングの結果を見ると，学習データ数が 500 件の時点で 79.45%とベースラインを大きく上回り，その後もデータ数の増加に伴い，概ね単調増加で精度が向上している．これは，モデルがバンドギャップ比較というタスクの論理構造を適切に学習し，データ量に応じたスケーリング則が働いていることを示唆している.
- ドメイン外データの干渉である．「誘電率比較チューニング」の結果は，ベースラインよりも大幅に低い精度となった．これは，同じ「比較タスク」であっても，対象となる物性（誘電率）が異なると，モデルが異なる分布や論理パターンを学習してしまい，バンドギャップ比較の推論能力を阻害したためと考えられる．この結果は，LLM のファインチューニングにおいて，タスク形式だけでなくドメイン知識の整合性が重要であることを示している.
- 数値理解とタスク特化の関係である．「数値予測チューニング」はベースラインより優れていたが，「bg 比較チューニング」には及ばなかった．これは，単に数字の扱いを強化するだけでは未知の物質での比較は不十分であり，比較推論を学習させることが最高精度を達成するために有効であることを示している.

C　バッテリー QA の結果についての考察

実験結果に基づき以下の観点から解析を行う.

- **データサイズとタスク特化の効果**
  bg 比較において，データサイズを 1500 から 3000 へ増加させることで，スコアが 44.26 から 45.20 へと向上した．これは，ドメイン特化データの増加がモデルの回答精度向上に寄与することを示唆している．また，数値予測チューニングよりも比較チューニングの方が高い性能を発揮する傾向が見られた.
- **GPT-4o との比較**
  bg 比較や誘電率に関連する単一タスク特化型モデルの多くが，GPT-4o のスコア（42.62）を上回った．これは，バッテリー材料のような専門性の高い領域においては，汎用モデルよりも，特定ドメインにファインチューニングされたモデルの方が高い正確性を発揮できることを示している.
- **混合データ（Mix）の影響**
  mix 比較や mix わけてのスコアは，単一タスクモデルや GPT-4o と比較して低調な結果（40.28）となった．これは，性質の異なるタスクやデータを無差別に混合して学習させることで，モデルの推論能力に干渉が生じ，かえって精度が低下した可能性が考えられる.