第71回 応用物理学会春季学術講演会 注目講演プレスリリース

2024年 3月 18日

大規模言語モデル用材料分野ベンチマーク作成とそれによるChatGPT・Bardの評価
Benchmark for LLM in Materials Science and the evaluation of ChatGPT and Bard

材料科学分野に特化したベンチマークを作成
LLMを活用することで材料探索の可能性を最大化する

オムロンサイニエックス[1]，阪大[2]，東大[3],
○吉武道子[1]，鈴木雄太[2]，五十嵐亮[1]，牛久祥孝[1]，長藤圭介[3]

【発表概要】

- **材料科学分野に特化した大規模言語モデルの新たなベンチマークを開発。**
- **ChatGPT4が最も高い正答率を示し、無料のChatGPT3.5と試用版のBardとの間にも性能差が見られた。**
- **さらにLLMに共通して見られる正答・誤回答の傾向も観察。特定分野におけるLLMの適用性と精度を理解する上で重要な示唆を提供。**

オムロンサイニエックスの吉武道子らによる研究グループは、材料科学分野に特化した大規模言語モデル（LLM: Large Language Models）の研究開発を推進するための、新たなベンチマークを開発し、最新のLLMにおける性能を評価した。ベンチマークは、大学レベルの材料科学の教科書の問題を元に、文章表現に改変を加えて作成された自由回答（144問）および4択問題（164問）である。対象とするLLMはOpenAIによるChatGPT3.5、ChatGPT4、およびGoogleのBardが用いられた。結果として、有料のChatGPT4が最も高い正答率（0.63）を示し、無料のChatGPT3.5（正答率：0.23）と試用版のBard（正答率：0.28）との間にも性能差が観察された。さらにLLMに共通して見られる正答・誤回答の傾向も観察された。この研究は、特定分野におけるLLMの適用性と精度を理解する上で重要な示唆を提供し、将来のモデル開発に貢献するものである。

【詳細】

言語データで、マテリアルズ・インフォマティクスはさらに進化する

昨今の自然言語処理（NLP）技術、特にLLMの発展と社会への普及が注目を集めている。LLMではGoogleのBERT（Bidirectional Encoder Representations from Transformers ）を皮切りに、2022年11月に登場したOpenAIによるChatGPTなど、さまざまなモデルが生まれている。これらLLMは、膨大なテキストデータを学習することで、文章の生成、質問応答、翻訳、さらにはコーディングなど複数のタスクに対応する能力を持っている。また、特定の分野に特化したモデルの開発も進んでおり、より専門的な知識が必要なタスクにおいても、LLMはその性能を発揮している。

「これまでもマテリアルズ・インフォマティクスでは機械学習が多用されてきましたが、主に扱うのは数値データでした。ここに言語データを加えることができれば、たとえば材料科学における探索などに非常に有用だと考えられます」とオムロンサイニエックスの吉武道子氏は、研究のモチベーションについて話す。これまでも材料科学の研究者の知見を活かしたマテリアルズ・インフォマティクスの研究にも携わってきたという。しかし吉武氏はLLMではなく、材料科学を専門とする研究者であり、応用物理学会でも数多くの発表を行ってきた。

「材料科学の研究者として思うのは、たとえば同じ材料探索でも、着眼点によって利用する知識の範囲がまったく異なるということです。たとえば、機械的な性質と電気的な性質に着眼した材料探索では、そもそも求める指標が異なります。そのため、与えられた数値の内側のみを探索する従来のマテリアルズ・インフォマティクスでは、それぞれの着眼点から見た最適解しか得られません。より包括的で有用な材料探索を実現するためには、数値データと言語データを統合したマテリアルズ・インフォマティクスが必要であると感じています。今回の報告は、材料科学に特化して評価できるベンチマークです。今後のマテリアルズ・インフォマティクスにおける材料探索に役立てたいと思っています」（吉武）

材料科学における有用なベンチマークを作成

吉武氏らの研究グループが取り組んだのは、材料科学に特化したLLMのベンチマークの作成だ。ベンチマークは、LLMの性能を評価するために用いる基準となるものを指し、問題文と解答のセットによって構成される。すでにLLMのベンチマークは数多く存在し、日々多くの論文が発表されている。古くは英語の標準ベンチマークである「GLUE」がある。GLUEは、GoogleのLLMであるBERTに活用されてきた。また2021年には、高校・大学の教科書レベルや各種資格試験問題からつくられた、より広範囲なタスクを評価するベンチマーク「MMLU」が作成・公表された。

研究グループが今回の研究におけるマイルストーンに位置づけたのは、2023年7月に発表されたベンチマーク「SciBENCH」だった。大学の教科書レベルの数学、物理学、化学分野のベンチマー

クであり、これによってChatGPT3.5とChatGPT4が評価されてきた実績がある。吉武氏らが目指したのはSciBENCHの材料科学版だ。

ベンチマークの作成には、大学レベルの材料科学の教科書が用いられた。「広く使われている教科書のうち、PDF化された教材を使用し、直接Web上で読める教材を避ける工夫をしています」と吉武氏は話す。つまりLLMによる自動学習に不向きな教科書を選定し、問題に利用することで、より有用なベンチマークを作成したということだ。

「さらに4択問題の作成においても、まず教科書に掲載されている問題を、熟練の人間の出題者が解き、解答と突き合わせをして、正答に至るプロセスを把握します。その上で人間の出題者は『学生だったらこのように間違うだろう』という予測を立てて、不正解となる選択肢を作成しています」と吉武氏はベンチマークの作成工程を振り返る。

LLMの得意・不得意が見えてきた

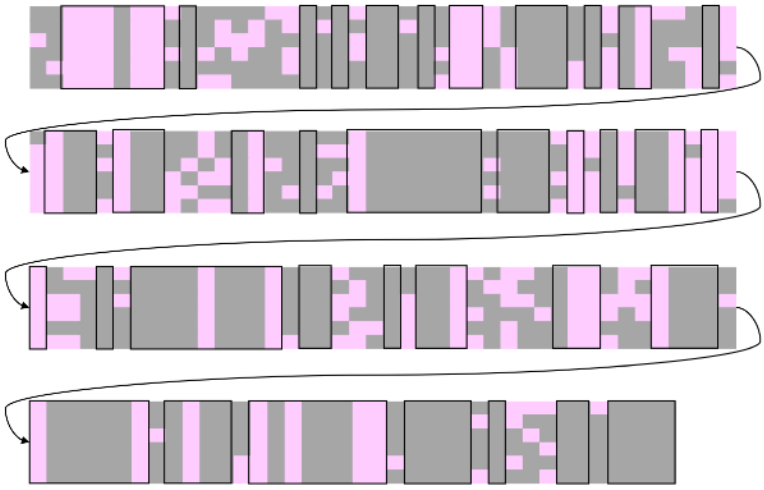
作成されたベンチマークは原子間の結合、欠陥・拡散、転位・強度・破壊、腐食・電気化学、電子的・光学的・熱的性質・磁性、金属・セラミック・高分子・複合材料など、多岐に渡る問題によって構成される自由回答（144問）および4択問題（164問）となった。実験では、これらの問題を対象となるLLM（ChatGPT3.5、ChatGPT4、Bard）に入力し、LLMからの回答を人間の手で解析し、正誤を判定し、正答率を求めた。また、入力の際は毎回、新規のチャットとして入力している。

その結果、無料のChatGPT3.5（正答率：0.23）と有料のChatGPT4（正答率：0.63）の間に顕著な正答率の差が見られた。また、無料で利用可能な試用版Bardは、無料のGhatGPT3.5よりもわずかに正答率が高かった（正答率：0.28）。さらに、各LLMともに問題の一部を抽出し、5回入力を行い、全体の正答率が（入力回数は少ないものの）統計的な平均の正答率を反映しているかを確認している。

「興味深いのは、全問題についてChatGPT3.5では2回、ChatGPT4は1回、Bardで2回の正誤比較を、4択問題についてBardで6回の正誤比較を行ったところ、モデルに共通した正答・誤回答の傾向が見いだせたことでした（図）。つまり、LLMには共通して苦手な問題と得意な問題があるということです。とくに4択問題におけるBardの正誤比較では、6回とも正解あるいは不正解の問題が全問題の62%を占めていました」と吉武氏は結果を振り返る。

今回の研究では状態図や相変化は、図が必須なものが多いため除外していた。今後の展望としては、これらの図をベンチマークに盛り込んでいくことが必要だという。応用物理学会の注目講演では、より詳細な結果の分析が報告される予定だ。

全問題の、ChatGPT3.5x2回、ChatGPT4x1回、Bardx2回の正誤比較



4択問題の、Bardx6回の正誤比較

図: ChatGPT3.5、ChatGPT、Bardそれぞれの正誤比較 : LLMの全問題の正答（ピンクで表示）および誤回答（グレーで表示）を図で表現したもの。正答の箇所と誤回答の箇所がある程度似通っていることがわかる。