

マルチモーダル大規模言語モデル用材料分野ベンチマーク作成

Benchmark for multi-modal LLM in Materials Science

オムロンサイニクエックス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

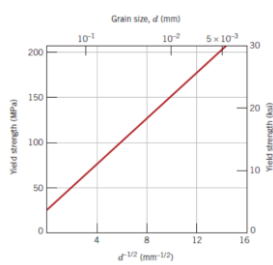
OMRON SINIC X¹, Osaka Univ.², Univ. Tokyo³ ○Michiko Yoshitake¹, Yuta Suzuki², Ryo Igarashi¹, Yoshitaka Ushiku¹, Keisuke Nagato³

E-mail: michiko.yoshitake@sinicx.com

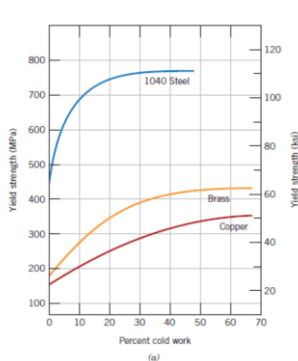
昨今の大規模言語モデル (LLM) をはじめとする自然言語処理技術の発展は目覚ましく、2022年11月の [CharGPT](#) 登場以来様々な LLM が作成され、2024年5月13日にはマルチモーダル化(テキスト・音声・画像を組み合わせると入力・出力ができる)されたモデル ChatGPT-4o が発表された。言語モデルの性能を評価するために用いられるのがベンチマークと呼ばれる問題文と解答のセットで、2021年に高校・大学の教科書レベルや各種資格試験問題から広範囲なベンチマークである MMLU[1]、2023年7月には大学の教科書レベルの数学・物理学・化学の分野のベンチマークとして SciBENCH[2]が作成・発表され、我々も材料科学の大学教科書[3,4]の問題をベースに材料科学分野に特化したベンチマークを作成し[5]、前回の春の応用物理学会で発表した。最近、マルチモーダルに対応した広範囲ベンチマーク MMMU[6]も発表されている。

今回、マルチモーダル化に対応した材料科学分野用に、先と同じ教科書から、図を使用しないと解答できない問題をベンチマークとして作成した。著作権の問題を回避するために、問題の本質や正答を変えない配慮をした上で独自に作画を行ってベンチマーク用の図面とした。下記に示した一例は、「If it is assumed that the plot in Figure 7.15 is for noncold-worked brass, determine the grain size (d) of the alloy in Figure 7.19; assume its composition is the same as the alloy in Figure 7.15.」という問題文に対する作図例である。①②は、教科書記載の Fig.7.15 と Fig.7.19、③④は独自に作画した①②に相当する図 (③は線の勾配を変え、④は線を部分的に改変し、brass などの材料名を A,B,C としている) である。なお、当日は ChatGPT-4o に回答させた結果についても報告予定である。

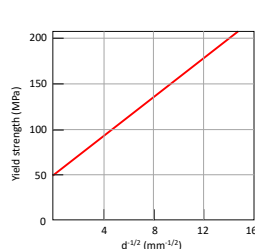
①



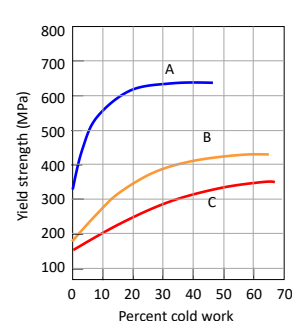
②



③



④



[1] arXiv:2009.03300v3.

[2] arXiv:2307.10635.

[3] Rolf E. Hummel, Understanding Materials Science, 2nd ed. 2004, Springer.

[4] William D. Callister, Jr. and David G. Rethwisch, Materials Science and Engineering An Introduction, 8th ed. 2010, John Wiley & Sons.

[5] arXiv:2409.03161v2.

[6] arXiv:2311.16502v4.