Large Language Models for Inorganic Synthesis Predictions

Seongmin Kim[1], Yousung Jung[2,3,4*], and Joshua Schrier[5*]

[1] Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291, Daehak-ro, Yuseong-gu, Daejeon 34141, Korea

[2] Department of Chemical and Biological Engineering (BK21 four), Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea

[3] Institute of Chemical Processes, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea

[4] Interdisciplinary Program in Artificial Intelligence, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea

[5] Department of Chemistry and Biochemistry, Fordham University, 441 E. Fordham Road, The Bronx, New York 10458, United States

*Email: jschrier@fordham.edu

*Email: yousung.jung@snu.ac.kr

Abstract

We evaluate the effectiveness of pre-trained and fine-tuned large language models (LLMs) for predicting the synthesizability of inorganic compounds and the selection of precursors needed to perform inorganic synthesis. The predictions of fine-tuned LLMs are comparable to—and sometimes better than—recent bespoke machine learning models for these tasks, but require only minimal user expertise, cost, and time to develop. Therefore, this strategy can serve both as an effective and strong baseline for future machine learning studies of various chemical applications and as a practical tool for experimental chemists.

Synthesizing novel compositions of matter is a pre-requisite for scientific and practical breakthroughs.[1] Discovery would be accelerated if one could predict whether a hypothetical compound could be made and what precursors should be used to make it. Human experts have developed physical theories and heuristic rules for these tasks,[2–6] but increasingly machine learning (ML) is used to predict synthesizability[7–11] and select precursors.[12–15] However, developing and training ML models requires substantial expertise, slowing adoption by experimental chemists.[16,17]

General purpose large language models (LLMs) are a form of generative artificial intelligence (AI),[18] pre-trained on a broad dataset so they can be applied to many different tasks using natural language. Pre-trained LLMs have been investigated for a wide variety of chemical tasks,[19,20] such as extracting structured data from the literature,[21–24] writing numerical simulation software,[25] and education.[26] LLM-based workflows have been used to plan syntheses of organic molecules[27,28] and metal-organic frameworks (MOFs).[21,29–31] Recent work has benchmarked materials science[32,33] and general chemical knowledge[34–37] of existing LLMs, and there are efforts to develop chemistry/materials-specific LLMs.[38,39] Fine-tuning LLMs on modest amounts of data improves performance for specific tasks, while still taking advantage of the general pre-training to provide basic symbol interpretation and output formatting guidance. Chemical applications of LLM fine-tuning have addressed property regression and classification of organic molecules,[40–43] and been used to improve the water-harvesting behavior of MOFs.[29]

Here, we demonstrate that fine-tuned LLMs can predict inorganic synthesizability and precursor selection with performance comparable to bespoke ML models. We used the GPT-3.5 (`gpt-3.5-turbo-0125`) and GPT-4 (`gpt-4-0125-preview`) pre-trained LLMs from OpenAI, but expect similar results for other LLMs. For each task, we compared the results to recently published ML models, closely following the datasets and data processing used in those earlier studies; detailed descriptions are in the **Supporting Information**. The available data was divided into 80% training and 20% test; all models were trained and tested on exactly the same data. Data is prepared for the LLM fine-tuning process by organizing it into pairs of user input and desired response; examples are shown in the **Supporting Information**. Fine-tuning was performed starting from the `gpt-3.5-turbo-0125` base model, with OpenAI's default hyperparameters, using 20% of the test data for validation. All LLM evaluations were performed with the model temperature set to zero to return the most probable response. Data, prompts, and code are available online.

Synthesizability prediction: Given a chemical formula, predict if the compound could be made. This is a positive-unlabeled (PU) learning problem,[44–46] as the available data consists of formulas of known (previously made) compounds and unknown (hypothetical) compounds which may not be synthesizable. Following Jang and Noh et al.,[47] we used the 393,053 unique inorganic compositions contained in the Materials Project (MP)[48] and Open Quantum Materials Database (OQMD)[49] (retrieved 02/2020) to define the set of possibilities; the 40,817 compounds with Inorganic Crystal Structure Database (ICSD) references are positive (synthesized) and the remaining 352,236 are unlabeled (hypothesized). The only chemical input to each model is the formula, in the format `Li1Fe1P1O4`. LLMs were provided with the prompt: `You are an expert inorganic chemist.  Determine if the following compound is likely to be synthesizable based on its composition, answering only "P" (for positive or possible) and "U" (for unknown or unlikely).` LLM fine-tuning on all 32,653 positive compounds and an equal number of randomly selected unknown compounds in the training set requires <2.5 hours and <25 USD (as of 04/2024; these are only approximate as time may vary with server load and cost is expected to decrease in the future). We compared to the stoichiometric convolutional graph neural fingerprint (stoi-CGNF), a composition-based synthesizability classification model trained by semi-supervised PU learning, and a stoichiometric similarity baseline model which classifies materials synthesizability based on similarity cutoff;[47] detailed explanations of these models are in the **Supporting Information**. Unlike traditional binary classifiers with positive and negative data, only the true positive rate (TPR) or recall can be measured unambiguously for PU problems, as one lacks true negative data. However, precision (PREC) and false positive rate (FPR) can be estimated by using the prior knowledge, α, which is the estimated proportion of the positive among the unlabeled dataset.[50,51] The detailed α-estimation process is described in the **Supporting Information**; we computed $\alpha = 0.088$, in agreement with the previous result (8.1%).[47]

As depicted in Figure 1a, GPT-3.5 (FT) and stoi-CGNF metrics are comparable, and both outperform the GPT-3.5, GPT-4, and baseline methods in recall. (The seemingly high recall of the pre-trained GPT-3.5 model is an artifact of always predicting the positive class, resulting in very low precision.) However, their precision is lower than the baseline. These results correspond to choosing the highest probability outcome (i.e., in the binary classification task, labeling P if the model predicts $p(\mathrm{P}) > 0.5$.) The exact metric values of the 0.5-threshold results were tabulated in Table S1 and the probability distributions of each model are shown in Figure S4. Using the 0.5-threshold for stoi-CGNF and GPT-3.5 (FT) results in 15% of the unlabeled data predicted as positive ($p(\mathrm{P}|\mathrm{U})$), which is inconsistent with our estimated α (8.8%).

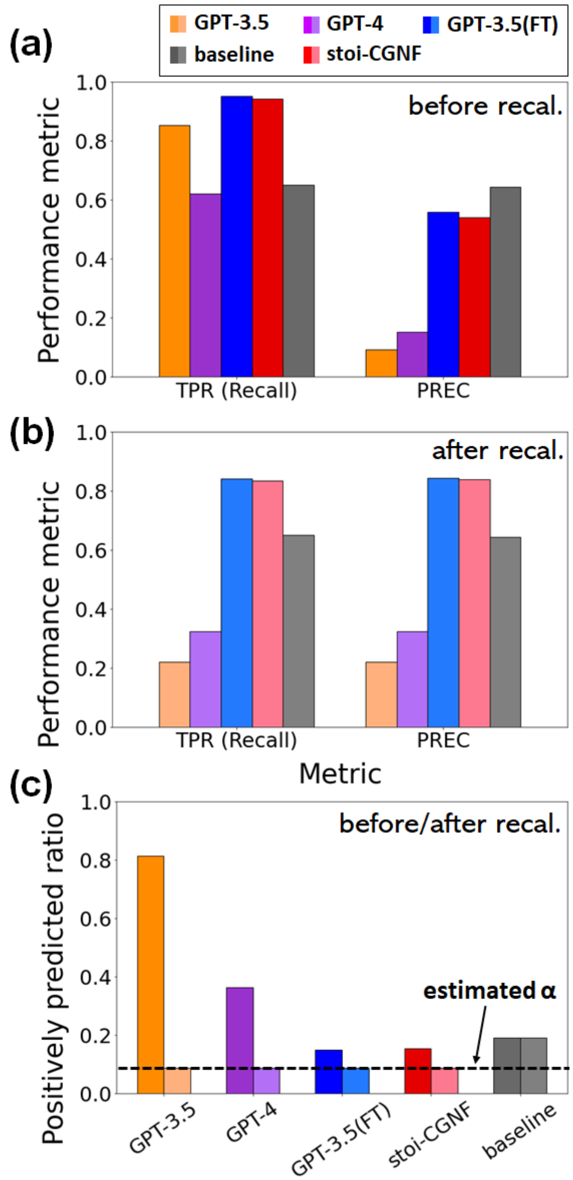

Figure 1: a) Comparison of synthesizability predictions upon 0.5-threshold. The precision (PREC) was calculated from the α-estimation. b) Comparison of synthesizability predictions after the recalibration by using the prior knowledge, α=0.088. Thresholds are 0.723, 0.761, 0.963, and 0.977 for stoi-CGNF, GPT-3.5(FT), GPT-4, and GPT-3.5. c) The positively predicted ratio among the unlabeled materials, before and after recalibration.

To match the assumed structure of our PU dataset, we adopted higher threshold values, which corresponds to the ratio of positively predicted unlabeled entries are well-matched in α value, and recalibrated model performance. (See Figure 1b, 1c, and Table S2.) Probabilities for the GPT models are obtained by querying the log-probabilities assigned to each possible response; Ramos et al. used a similar strategy for in-context Bayesian optimization of catalysts from text descriptions.[52] Although threshold recalibration decreases recall, the models obtained more balanced performance. (Figure 1b) After recalibration, fine-tuned GPT-3.5 outperformed baseline and stoi-CGNF model in both metrics, despite using fewer unlabeled data in training. In contrast, recalibrating the pre-trained GPT-3.5 and GPT-4 models reduces their performance below the similarity baseline. This demonstrates the necessity of fine-tuning, independent of the recalibration strategy.

Precursor Selection: Given the formula for a target compound, predict the complete set of precursors that must be provided. The output must exactly match the entire sets of precursors in a known example synthesis; because the output is restricted to a predefined list of precursors this is a multi-label prediction problem.[53] Following Kim et al.,[15] we began with the text-minded synthesis dataset of Kononova et al.,[54] removing entries with inconsistent or incomplete data, and retaining only reactions that contained precursors used in ≥5 example reactions, which results in 11,923 unique reactions and 311 precursors. (See Figure S1.) The only chemical input to each model was a chemical formula (e.g., `LiFePO4`) and the desired output is of the form `LiFePO4 <- Li2CO3 + FeC2O4 + (NH4)2HPO4`. Between 2 - 8 precursors must be specified for a target (Figure S2); most targets have only one unique reaction, but some have as many as 12 unique reactions (Figure S3). Like the synthesizability task, the test data contains only positive examples; "incorrect" predictions may actually work in the laboratory, so our evaluation underestimates true performance. LLMs were given the following system prompt: `You are an expert solid-state chemist planning a material synthesis. You are provided with a target compound and must select the precursor reagents needed to synthesize the target. Typically two or more precursors are needed. The precursor reagents must provide all of the elements in the target. However, because the reactions are performed in an open system where gases can escape, it is acceptable for some non-metal elements (e.g., H, C, N, O, F, Cl, Br) present in the precursors to be absent in the final target product. Your synthesis task is only to identify the correct precursor to use. Do not provide stoichiometry. Only use precursors from the following candidates: … If asked to generate more than one reaction recommendation for a target, each recommendation should be different, and be separated with a newline. Return only output of the following format for each recommendation: [Target] <- [Precursor] + [Precursor] + [Precursor]` They were then asked in the user prompt `synthesize [Target]` or `provide 5 synthesis plans for [Target]` to generate the prediction. Because of the smaller dataset, we performed a 5-fold cross-validation. We compared the LLMs to the recent Elementwise template model (Elemwise)[15] and a random statistical baseline in which precursors are selected based on their frequency in the dataset for each element. Fine-tuning each GPT-3.5 model required <90 minutes and <11 USD (as of 04/2024).

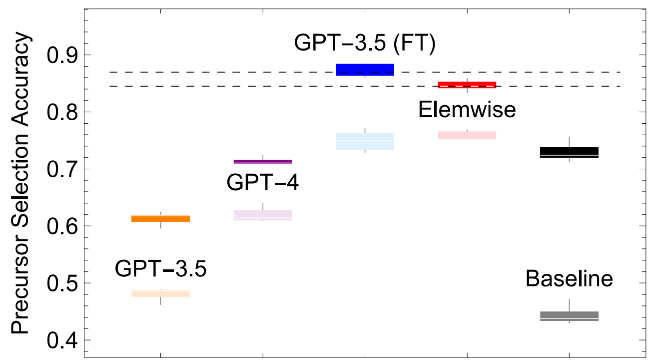

Figure 2: Precursor selection accuracy, evaluated over 5-fold cross validation for each model. Light and dark boxes indicate Top-1 and Top-5 performance, respectively, with the white horizonal lines indicating the median. Dashed lines indicate the minimum and maximum accuracy of a hypothetical "perfect" Elemwise method over the cross validation.

As shown in Figure 2, the top-1 predictions of the fine-tuned GPT-3.5 and Elemwise models are comparable; all other methods are inferior. Although the pre-trained models have better top-1 predictions than the statistical baseline, they make worse top-5 predictions, suggesting that they do not sample adequately diverse possibilities when generating multiple outputs. This might be counteracted by increasing the model temperature when generating multiple outputs. LLMs can generate precursor sets that are outside the domain of the Elemwise model, which assumes that each metal type present in the target corresponds to one precursor. For example, diammonium phosphate does not contain a metal, and thus cannot be predicted as a precursor by the Elemwise model. The performance of a hypothetical "perfect" Elemwise model is shown as the dashed lines in Figure 2; the top-5 predictions of the fine-tuned GPT-3.5 model slightly exceed this limit.

Although the overall prediction qualities are comparable, the models make different predictions. As depicted in Figure 3, the Top-1 and Top-5 predictions of the two models can vary, where one is correct and the other is incorrect. Combining the Top-5 predictions of both models would predict the correct precursors for 93% of target compounds. This suggests the value of including fine-tuned LLM predictions in ensemble methods. We investigated whether combining the top-5 predictions of the Elemwise (84.5% accuracy) and GPT-3.5 FT (86.0%) and then asking the pre-trained GPT-4 model to discuss the feasibility of each plan before selecting the best five syntheses would improve the results. This provides both human-readable explanations and a small, but statistically significant, improvement of the resulting top-5 prediction accuracy to 87.6%. Despite instruction to only use the predefined precursors, the LLMs sometimes hallucinate precursors outside this allowed set (see Table S3). One possible solution is to create a combination model that retains only the first 5 unique reactions from the GPT-3.5 (FT) and Elemwise model that only contain allowed precursors; this increases the top-5 prediction accuracy increases to 90.9%. (See **Supporting Information**.)

Figure 3: Comparison of a) Top-1 and b) Top-5 prediction precursor prediction accuracy by the Elemwise and GPT-3.5 (FT) models summed over all cross-validation splits.

In conclusion, fine-tuned LLMs are comparable to= or better than the latest ML models developed specifically for the synthesizability and precursor selection problems, using only the target chemical formula as input. In the case of the precursor selection problem, the general output allows for answers that are outside the domain of bespoke ML models. Because of their simplicity and low cost, we recommend that fine-tuned LLMs be used as a strong baseline method against which to compare future bespoke ML models. Previous results on organic molecules have come to similar conclusions for regression and classification;[40–43] our results extend that recommendation to inorganic chemistry and PU-learning and multilabel tasks. We also recommend that developers of chemistry- and materials-specific LLMs[39,55–57] prioritize the ease with which users can fine-tune the models to unlock this capability.

A limitation is that this approach relies upon statistical patterns in the training data; biases present in reported syntheses[58,59] may hinder extrapolation to novel or rare chemistry.[60] For example, the 10 most popular precursors comprise 43% of all precursor choices in this dataset (Figure S7); statistical learning approaches like the one in this paper will continue to preferentially suggest these precursors, even if they are suboptimal. Also, commercial LLMs do not disclose their training sets, which may raise the concern of inadvertent leakage of test data into the training set. However, the relatively poor performance of the pre-trained models for these tasks suggests this is unlikely.

As our goal was to illustrate this approach in the simplest possible way, there are many ways the performance might be improved. We made no attempt at prompt engineering[61] or hyperparameter optimization of the fine-tuning process. External function calling of "tools" (e.g., performing numerical or thermodynamic calculations) can be combined with iterative chain-ofthought methods ("think step by step") to further improve problem solving.[26,27,62] Finally, we expect continued advances in LLMs and fine-tuning methodologies to improve performance.

Acknowledgement

YJ acknowledges support from NRF (RS-2023-00283902, 2021R1A5A1030054) and IITP (2021-0-01343) of Korea government. JS acknowledges Fordham University for granting a sabbatical to study LLMs, Seoul National University for a Global Visiting Faculty Fellowship during which the work was initiated, and support by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Heavy Element Chemistry Program under contract KC0302031, subcontracted through Los Alamos National Laboratory.

Supporting Information

This Supporting Information is available free of charge at http://pubs.acs.org/[provided by publisher]

- Description of data preparation. Plots of the distribution of number of unique reactions and number of precursors. Description of model construction and training. LLM prompts. Description for evaluation metrics. Tables of the model performance for the synthesizability task. Description of methods and results for re-evaluating top-5 predictions using GPT-4 and code for associated statistical tests. Description of PU learning prompt modification experiments and table of results. Histogram of top-10 precursors occurrences. (PDF)

Data Availability

The code and data underlying this study are openly available on Github at https://github.com/jschrier/SynthGPT/, with a persistent archival copy deposited in Zenodo at DOI:[to be deposited after article passes peer review]. Access to GPT-3.5 and GPT-4 is commercially available to the public at https://openai.com. The stoi-CGNF source code is available at https://github.com/kaist-amsg/Synthesizability-stoi-CGNF and the Elemwise source code is available at https://github.com/kaist-amsg/ElemwiseRetro/ .

Notes

The authors declare no competing financial interest.

References

(1) Cheetham, A. K.; Seshadri, R.; Wudl, F. Chemical Synthesis and Materials Discovery. *Nat. Synth* **2022**, *1* (7), 514–520. https://doi.org/10.1038/s44160-022-00096-3.
(2) Goldschmidt, V. M. Die Gesetze der Krystallochemie. *Naturwissenschaften* **1926**, *14* (21), 477–485. https://doi.org/10.1007/BF01507527.
(3) Bartel, C. J.; Sutton, C.; Goldsmith, B. R.; Ouyang, R.; Musgrave, C. B.; Ghiringhelli, L. M.; Scheffler, M. New Tolerance Factor to Predict the Stability of Perovskite Oxides and Halides. *Science Advances* **2019**, *5* (2), eaav0693. https://doi.org/10.1126/sciadv.aav0693.
(4) Ouyang, B.; Wang, J.; He, T.; Bartel, C. J.; Huo, H.; Wang, Y.; Lacivita, V.; Kim, H.; Ceder, G. Synthetic Accessibility and Stability Rules of NASICONs. *Nat Commun* **2021**, *12* (1), 5752. https://doi.org/10.1038/s41467-021-26006-3.
(5) Woodward, P. M.; Karen, P.; Evans, J. S. O.; Vogt, T. *Solid State Materials Chemistry*; Cambridge University Press: Cambridge, 2021. https://doi.org/10.1017/9781139025348.
(6) West, A. R. *Solid State Chemistry and Its Applications*, Second edition.; Wiley: Hoboken, NJ, 2022.
(7) Jang, J.; Gu, G. H.; Noh, J.; Kim, J.; Jung, Y. Structure-Based Synthesizability Prediction of Crystals Using Partially Supervised Learning. *J. Am. Chem. Soc.* **2020**, *142* (44), 18836–18843. https://doi.org/10.1021/jacs.0c07384.
(8) Gu, G. H.; Jang, J.; Noh, J.; Walsh, A.; Jung, Y. Perovskite Synthesizability Using Graph Neural Networks. *npj Comput Mater* **2022**, *8* (1), 1–8. https://doi.org/10.1038/s41524-022-00757-z.
(9) Gleaves, D.; Fu, N.; Siriwardane, E. M. D.; Zhao, Y.; Hu, J. Materials Synthesizability and Stability Prediction Using a Semi-Supervised Teacher-Student Dual Neural Network. *Digital Discovery* **2023**, *2* (2), 377–391. https://doi.org/10.1039/D2DD00098A.
(10) Zhu, R.; Tian, S. I. P.; Ren, Z.; Li, J.; Buonassisi, T.; Hippalgaonkar, K. Predicting Synthesizability Using Machine Learning on Databases of Existing Inorganic Materials. *ACS Omega* **2023**, *8* (9), 8210–8218. https://doi.org/10.1021/acsomega.2c04856.
(11) Antoniuk, E. R.; Cheon, G.; Wang, G.; Bernstein, D.; Cai, W.; Reed, E. J. Predicting the Synthesizability of Crystalline Inorganic Materials from the Data of Known Material Compositions. *npj Comput Mater* **2023**, *9* (1), 1–11. https://doi.org/10.1038/s41524-023-01114-4.
(12) Kim, E.; Huang, K.; Jegelka, S.; Olivetti, E. Virtual Screening of Inorganic Materials Synthesis Parameters with Deep Learning. *npj Comput Mater* **2017**, *3* (1), 1–9. https://doi.org/10.1038/s41524-017-0055-6.
(13) Kim, E.; Jensen, Z.; Van Grootel, A.; Huang, K.; Staib, M.; Mysore, S.; Chang, H.-S.; Strubell, E.; McCallum, A.; Jegelka, S.; Olivetti, E. Inorganic Materials Synthesis Planning with Literature-Trained Neural Networks. *J. Chem. Inf. Model.* **2020**, *60* (3), 1194–1201. https://doi.org/10.1021/acs.jcim.9b00995.
(14) He, T.; Huo, H.; Bartel, C. J.; Wang, Z.; Cruse, K.; Ceder, G. Inorganic Synthesis Recommendation by Machine Learning Materials Similarity from Scientific Literature. *Science Advances* **2023**, *9* (23), eadg8180. https://doi.org/10.1126/sciadv.adg8180.
(15) Kim, S.; Noh, J.; Ho Gu, G.; Chen, S.; Jung, Y. Predicting Synthesis Recipes of Inorganic Crystal Materials Using Elementwise Template Formulation. *Chemical Science* **2024**, *15* (3), 1039–1045. https://doi.org/10.1039/D3SC03538G.
(16) Yano, J.; Gaffney, K. J.; Gregoire, J.; Hung, L.; Ourmazd, A.; Schrier, J.; Sethian, J. A.; Toma, F. M. The Case for Data Science in Experimental Chemistry: Examples and

Recommendations. *Nat Rev Chem* **2022**, *6*, 357–370. https://doi.org/10.1038/s41570-022-00382-w.
(17) Back, S.; Aspuru-Guzik, A.; Ceriotti, M.; Gryn'ova, G.; Grzybowski, B.; Gu, G. H.; Hein, J.; Hippalgaonkar, K.; Hormázabal, R.; Jung, Y.; Kim, S.; Kim, W. Y.; Moosavi, S. M.; Noh, J.; Park, C.; Schrier, J.; Schwaller, P.; Tsuda, K.; Vegge, T.; Lilienfeld, O. A. von; Walsh, A. Accelerated Chemical Science with AI. *Digital Discovery* **2024**, *3* (1), 23–33. https://doi.org/10.1039/D3DD00213F.
(18) Anstine, D. M.; Isayev, O. Generative Models as an Emerging Paradigm in the Chemical Sciences. *J. Am. Chem. Soc.* **2023**, *145* (16), 8736–8750. https://doi.org/10.1021/jacs.2c13467.
(19) Jablonka, K. M.; Ai, Q.; Al-Feghali, A.; Badhwar, S.; Bocarsly, J. D.; Bran, A. M.; Bringuier, S.; Brinson, L. C.; Choudhary, K.; Circi, D.; Cox, S.; Jong, W. A. de; Evans, M. L.; Gastellu, N.; Genzling, J.; Gil, M. V.; Gupta, A. K.; Hong, Z.; Imran, A.; Kruschwitz, S.; Labarre, A.; Lála, J.; Liu, T.; Ma, S.; Majumdar, S.; Merz, G. W.; Moitessier, N.; Moubarak, E.; Mouriño, B.; Pelkie, B.; Pieler, M.; Ramos, M. C.; Ranković, B.; Rodriques, S. G.; Sanders, J. N.; Schwaller, P.; Schwarting, M.; Shi, J.; Smit, B.; Smith, B. E.; Herck, J. V.; Völker, C.; Ward, L.; Warren, S.; Weiser, B.; Zhang, S.; Zhang, X.; Zia, G. A.; Scourtas, A.; Schmidt, K. J.; Foster, I.; White, A. D.; Blaiszik, B. 14 Examples of How LLMs Can Transform Materials Science and Chemistry: A Reflection on a Large Language Model Hackathon. *Digital Discovery* **2023**, *2* (5), 1233–1250. https://doi.org/10.1039/D3DD00113J.
(20) Zhang, J.; Fang, Y.; Shao, X.; Chen, H.; Zhang, N.; Fan, X. The Future of Molecular Studies through the Lens of Large Language Models. *J. Chem. Inf. Model.* **2024**, *64* (3), 563–566. https://doi.org/10.1021/acs.jcim.3c01977.
(21) Zheng, Z.; Zhang, O.; Borgs, C.; Chayes, J. T.; Yaghi, O. M. ChatGPT Chemistry Assistant for Text Mining and the Prediction of MOF Synthesis. *J. Am. Chem. Soc.* **2023**. https://doi.org/10.1021/jacs.3c05819.
(22) Thway, M.; Low, A. K. Y.; Khetan, S.; Dai, H.; Recatala-Gomez, J.; Chen, A. P.; Hippalgaonkar, K. Harnessing GPT-3.5 for Text Parsing in Solid-State Synthesis – Case Study of Ternary Chalcogenides. *Digital Discovery* **2024**, *3* (2), 328–336. https://doi.org/10.1039/D3DD00202K.
(23) Lee, S.; Heinen, S.; Khan, D.; Anatole Von Lilienfeld, O. Autonomous Data Extraction from Peer Reviewed Literature for Training Machine Learning Models of Oxidation Potentials. *Mach. Learn.: Sci. Technol.* **2024**, *5* (1), 015052. https://doi.org/10.1088/2632-2153/ad2f52.
(24) Polak, M. P.; Morgan, D. Extracting Accurate Materials Data from Research Papers with Conversational Language Models and Prompt Engineering. *Nat Commun* **2024**, *15* (1), 1569. https://doi.org/10.1038/s41467-024-45914-8.
(25) D. White, A.; M. Hocky, G.; A. Gandhi, H.; Ansari, M.; Cox, S.; P. Wellawatte, G.; Sasmal, S.; Yang, Z.; Liu, K.; Singh, Y.; Ccoa, W. J. P. Assessment of Chemistry Knowledge in Large Language Models That Generate Code. *Digital Discovery* **2023**, *2* (2), 368–376. https://doi.org/10.1039/D2DD00087C.
(26) Schrier, J. Comment on "Comparing the Performance of College Chemistry Students with ChatGPT for Calculations Involving Acids and Bases." *J. Chem. Educ.* **2024**. https://doi.org/10.1021/acs.jchemed.4c00058.

(27) Bran, A. M.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; Schwaller, P. Augmenting Large-Language Models with Chemistry Tools. *Nature Machine Intelligence* **2024**, *6*, 525–535. https://doi.org/10.1038/s42256-024-00832-8.
(28) Boiko, D. A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous Chemical Research with Large Language Models. *Nature* **2023**, *624* (7992), 570–578. https://doi.org/10.1038/s41586-023-06792-0.
(29) Zheng, Z.; Alawadhi, A. H.; Chheda, S.; Neumann, S. E.; Rampal, N.; Liu, S.; Nguyen, H. L.; Lin, Y.; Rong, Z.; Siepmann, J. I.; Gagliardi, L.; Anandkumar, A.; Borgs, C.; Chayes, J. T.; Yaghi, O. M. Shaping the Water-Harvesting Behavior of Metal–Organic Frameworks Aided by Fine-Tuned GPT Models. *J. Am. Chem. Soc.* **2023**, *145* (51), 28284–28295. https://doi.org/10.1021/jacs.3c12086.
(30) Zheng, Z.; Rong, Z.; Rampal, N.; Borgs, C.; Chayes, J. T.; Yaghi, O. M. A GPT-4 Reticular Chemist for Guiding MOF Discovery. *Angewandte Chemie International Edition* **2023**, *62* (46), e202311983. https://doi.org/10.1002/anie.202311983.
(31) Zheng, Z.; Zhang, O.; Nguyen, H. L.; Rampal, N.; Alawadhi, A. H.; Rong, Z.; Head-Gordon, T.; Borgs, C.; Chayes, J. T.; Yaghi, O. M. ChatGPT Research Group for Optimizing the Crystallinity of MOFs and COFs. *ACS Cent. Sci.* **2023**, *9* (11), 2161–2170. https://doi.org/10.1021/acscentsci.3c01087.
(32) Zaki, M.; Jayadeva; Mausam; Krishnan, N. M. A. MaScQA: Investigating Materials Science Knowledge of Large Language Models. *Digital Discovery* **2024**, *3*, 313–327. https://doi.org/10.1039/D3DD00188A.
(33) Deb, J.; Saikia, L.; Dihingia, K. D.; Sastry, G. N. ChatGPT in the Material Design: Selected Case Studies to Assess the Potential of ChatGPT. *J. Chem. Inf. Model.* **2024**, *64* (3), 799–811. https://doi.org/10.1021/acs.jcim.3c01702.
(34) Hatakeyama-Sato, K.; Yamane, N.; Igarashi, Y.; Nabae, Y.; Hayakawa, T. Prompt Engineering of GPT-4 for Chemical Research: What Can/Cannot Be Done? *Science and Technology of Advanced Materials: Methods* **2023**, *3* (1), 2260300. https://doi.org/10.1080/27660400.2023.2260300.
(35) Guo, T.; Guo, K.; Nan, B.; Liang, Z.; Guo, Z.; Chawla, N.; Wiest, O.; Zhang, X. What Can Large Language Models Do in Chemistry? A Comprehensive Benchmark on Eight Tasks. In *Advances in Neural Information Processing Systems*; Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Curran Associates, Inc., 2023; Vol. 36, pp 59662–59688.
(36) AI4Science, M. R.; Quantum, M. A. The Impact of Large Language Models on Scientific Discovery: A Preliminary Study Using GPT-4. arXiv December 8, 2023. https://doi.org/10.48550/arXiv.2311.07361.
(37) Mirza, A.; Alampara, N.; Kunchapu, S.; Emoekabu, B.; Krishnan, A.; Wilhelmi, M.; Okereke, M.; Eberhardt, J.; Elahi, A. M.; Greiner, M.; Holick, C. T.; Gupta, T.; Asgari, M.; Glaubitz, C.; Klepsch, L. C.; Köster, Y.; Meyer, J.; Miret, S.; Hoffmann, T.; Kreth, F. A.; Ringleb, M.; Roesner, N.; Schubert, U. S.; Stafast, L. M.; Wonanke, D.; Pieler, M.; Schwaller, P.; Jablonka, K. M. Are Large Language Models Superhuman Chemists? arXiv April 1, 2024. https://doi.org/10.48550/arXiv.2404.01475.
(38) Wang, Z.; Chen, A.; Tao, K.; Han, Y.; Li, J. MatGPT: A Vane of Materials Informatics from Past, Present, to Future. *Advanced Materials* **2003**, *36* (6), 2306733. https://doi.org/10.1002/adma.202306733.

(39) Zhang, D.; Liu, W.; Tan, Q.; Chen, J.; Yan, H.; Yan, Y.; Li, J.; Huang, W.; Yue, X.; Zhou, D.; Zhang, S.; Su, M.; Zhong, H.; Li, Y.; Ouyang, W. ChemLLM: A Chemical Large Language Model. arXiv February 9, 2024. https://doi.org/10.48550/arXiv.2402.06852.
(40) Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. Leveraging Large Language Models for Predictive Chemistry. *Nat Mach Intell* **2024**, 1–9. https://doi.org/10.1038/s42256-023-00788-1.
(41) Chen, L.; Xie, Z.; Evangelopoulos, X.; Omar, O. H.; Troisi, A.; Cooper, A. Fine-Tuning GPT-3 for Machine Learning Electronic and Functional Properties of Organic Molecules. *Chem. Sci.* **2024**, *15*, 500–510. https://doi.org/10.1039/D3SC04610A.
(42) Zhong, S.; Guan, X. Developing Quantitative Structure–Activity Relationship (QSAR) Models for Water Contaminants' Activities/Properties by Fine-Tuning GPT-3 Models. *Environ. Sci. Technol. Lett.* **2023**, *10* (10), 872–877. https://doi.org/10.1021/acs.estlett.3c00599.
(43) Zhong, Z.; Zhou, K.; Mottin, D. Benchmarking Large Language Models for Molecule Prediction Tasks. arXiv March 8, 2024. https://doi.org/10.48550/arXiv.2403.05075.
(44) Elkan, C.; Noto, K. Learning Classifiers from Only Positive and Unlabeled Data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*; ACM: Las Vegas Nevada USA, 2008; pp 213–220. https://doi.org/10.1145/1401890.1401920.
(45) Li, X.-L.; Yu, P. S.; Liu, B.; Ng, S.-K. Positive Unlabeled Learning for Data Stream Classification. In *Proceedings of the 2009 SIAM International Conference on Data Mining*; Society for Industrial and Applied Mathematics, 2009; pp 259–270. https://doi.org/10.1137/1.9781611972795.23.
(46) Bekker, J.; Davis, J. Learning from Positive and Unlabeled Data: A Survey. *Mach Learn* **2020**, *109* (4), 719–760. https://doi.org/10.1007/s10994-020-05877-5.
(47) Jang, J.; Noh, J.; Zhou, L.; Gu, G. H.; Gregoire, J. M.; Jung, Y. Synthesizability of Materials Stoichiometry Using Semi-Supervised Learning. *Matter* **2024**, *7*, 2294–2312. https://doi.org/10.1016/j.matt.2024.05.002.
(48) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. a. The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Materials* **2013**, *1* (1), 011002. https://doi.org/10.1063/1.4812323.
(49) Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): Assessing the Accuracy of DFT Formation Energies. *npj Comput Mater* **2015**, *1* (1), 15010. https://doi.org/10.1038/npjcompumats.2015.10.
(50) Jain, S.; White, M.; Radivojac, P. Recovering True Classifier Performance in Positive-Unlabeled Learning. *AAAI* **2017**, *31* (1). https://doi.org/10.1609/aaai.v31i1.10937.
(51) Zeiberg, D.; Jain, S.; Radivojac, P. Fast Nonparametric Estimation of Class Proportions in the Positive-Unlabeled Classification Setting. *AAAI* **2020**, *34* (04), 6729–6736. https://doi.org/10.1609/aaai.v34i04.6151.
(52) Ramos, M. C.; Michtavy, S. S.; Porosoff, M. D.; White, A. D. Bayesian Optimization of Catalysts With In-Context Learning. arXiv April 11, 2023. https://doi.org/10.48550/arXiv.2304.05341.

(53) Madjarov, G.; Kocev, D.; Gjorgjevikj, D.; Džeroski, S. An Extensive Experimental Comparison of Methods for Multi-Label Learning. *Pattern Recognition* **2012**, *45* (9), 3084–3104. https://doi.org/10.1016/j.patcog.2012.03.004.
(54) Kononova, O.; Huo, H.; He, T.; Rong, Z.; Botari, T.; Sun, W.; Tshitoyan, V.; Ceder, G. Text-Mined Dataset of Inorganic Materials Synthesis Recipes. *Sci Data* **2019**, *6* (1), 203. https://doi.org/10.1038/s41597-019-0224-1.
(55) Chen, Z.-Y.; Xie, F.-K.; Wan, M.; Yuan, Y.; Liu, M.; Wang, Z.-G.; Meng, S.; Wang, Y.-G. MatChat: A Large Language Model and Application Service Platform for Materials Science. *Chinese Phys. B* **2023**, *32* (11), 118104. https://doi.org/10.1088/1674-1056/ad04cb.
(56) Hudson, N. C.; Pauloski, J. G.; Baughman, M.; Kamatar, A.; Sakarvadia, M.; Ward, L.; Chard, R.; Bauer, A.; Levental, M.; Wang, W.; Engler, W.; Price Skelly, O.; Blaiszik, B.; Stevens, R.; Chard, K.; Foster, I. Trillion Parameter AI Serving Infrastructure for Scientific Discovery: A Survey and Vision. In *Proceedings of the IEEE/ACM 10th International Conference on Big Data Computing, Applications and Technologies*; ACM: Taormina (Messina) Italy, 2023; pp 1–10. https://doi.org/10.1145/3632366.3632396.
(57) Yu, B.; Baker, F. N.; Chen, Z.; Ning, X.; Sun, H. LlaSMol: Advancing Large Language Models for Chemistry with a Large-Scale, Comprehensive, High-Quality Instruction Tuning Dataset. arXiv February 17, 2024. https://doi.org/10.48550/arXiv.2402.09391.
(58) Jia, X.; Lynch, A.; Huang, Y.; Danielson, M.; Lang'at, I.; Milder, A.; Ruby, A. E.; Wang, H.; Friedler, S. A.; Norquist, A. J.; Schrier, J. Anthropogenic Biases in Chemical Reaction Data Hinder Exploratory Inorganic Synthesis. *Nature* **2019**, *573* (7773), 251–255. https://doi.org/10.1038/s41586-019-1540-5.
(59) Beker, W.; Roszak, R.; Wołos, A.; Angello, N. H.; Rathore, V.; Burke, M. D.; Grzybowski, B. A. Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki–Miyaura Coupling. *J. Am. Chem. Soc.* **2022**, *144* (11), 4819–4827. https://doi.org/10.1021/jacs.1c12005.
(60) Schrier, J.; Norquist, A. J.; Buonassisi, T.; Brgoch, J. In Pursuit of the Exceptional: Research Directions for Machine Learning in Chemical and Materials Science. *J. Am. Chem. Soc.* **2023**, *145* (40), 21699–21716. https://doi.org/10.1021/jacs.3c04783.
(61) *Prompt Engineering Guide*. https://www.promptingguide.ai/ (accessed 2024-03-02).
(62) Chiang, Y.; Chou, C.-H.; Riebesell, J. LLaMP: Large Language Model Made Powerful for High-Fidelity Materials Knowledge Retrieval and Distillation. arXiv January 30, 2024. https://doi.org/10.48550/arXiv.2401.17244.

For Table of Contents Only