Evaluating Extrapolation Ability of Large Language Model in Chemical Domain

Taehun Cha and **Donghun Lee***
Department of Mathematics
Korea University
{cth127, holy}@korea.ac.kr

Abstract

Solving a problem outside the training space, i.e. extrapolation, has been a long problem in the machine learning community. The current success of large language models demonstrates the LLM's extrapolation ability to several unseen tasks. In line with these works, we evaluate the LLM's extrapolation ability in the chemical domain. We construct a data set measuring the material properties of epoxy polymers depending on various raw materials and curing processes. LLM should predict the material property when novel raw material is introduced utilizing its chemical knowledge. Through experiments, LLM tends to choose the right direction of adjustment but fails to determine the exact degree, resulting in poor MAE on some properties. But LLM can successfully adjust the degree with only a one-shot example. The results show that LLM can extrapolate to new unseen material utilizing its chemical knowledge learned through massive pre-training.

1 Introduction

Marcus (1998) depicted two aspects of the generalization: interpolation and extrapolation. The interpolation targets a problem *within* the training space, while the extrapolation targets the *outside*. Despite the rapid development of machine learning technology, even a modern deep-learning-based model struggles to extrapolate on some tasks that humans find easy (Lake and Baroni, 2018, Barrett et al., 2018 and Saxton et al., 2019).

Human reasoning involves the extrapolation ability (Webb et al., 2020), especially for knowledge discovery. Mitchell et al. (2018) exemplified Halley's prediction on the return of a comet: it was possible thanks to Newton's inverse square law of gravity and would be difficult with pre-Newtonian models. Newton found laws that went *beyond simply maximizing the fit* to the known set of planetary bodies (Mitchell et al., 2018), unlike usual machine learning models.

The current success of large language models (LLMs) shows hints of their extrapolation ability. Conneau and Lample (2019) reported that fine-tuning a multilingual language model on a monolingual classification data set can result in a strong multilingual classifier, which has never seen a multilingual classification data set. Wei et al. (2022) introduced an instruction tuning framework: by training LLM on multiple tasks to follow human instructions, the LLM shows improved zero-shot performance on several unseen tasks. These results suggest an emergent extrapolation ability of LLM utilizing its representation power learned through massive pre-training.

In this paper, we explore the extrapolation ability of LLM in the chemical domain. Our main research question is *Can LLM perform the extrapolation utilizing its internal chemical knowledge?* To examine this question, we suggest a novel task regressing material properties of epoxy polymers when a novel raw material is introduced. LLM should infer the effect of novel raw material on the epoxy polymer from natural language descriptions or SMILES.

2 Related Works

Several researchers adopted an LLM to the chemical domain by training it on a chemistry-related corpus. Fang et al. (2024) introduced a data set for instruction tuning including various molecule/protein-oriented tasks. Cao et al. (2023) and Zhao et al. (2023) integrated the graph structure of molecules into an LLM to improve its representation power. Ye et al. (2023) and Zhao et al. (2024) trained a dialogue model on chemical domain. Our goal is to verify the chemical ability of an existing LLM, not suggesting a new foundation model.

* corresponding author

Guo et al. (2023) verified existing LLMs' ability on eight chemical tasks from name prediction to molecule captioning. They showed that GPT-4 (OpenAI, 2024) showed the best performance on most tasks showing comparable performance with SOTA, task-specific models. Our work is an extension of their work while differing on two points: (1) Their work focused on molecule-level tasks, while our work is compound-level. As more information should be considered, compound-level tasks require more complex reasoning than the molecule-level. (2) Unlike classic tasks, we focus on the extrapolation ability of an LLM, which is also more challenging.

3 Problem Statement

Let $(\mathcal{X}, \mathcal{Y})$ be a domain of independent and dependent variables of train data. We have our train data $\mathcal{D}_{train} = \{(x_i, y_i)\}_{i=1}^{N_{train}}, x_i \in \mathcal{X}, y_i \in \mathcal{Y}$. Let $\mathcal{X}'$ be a domain of additional independent variables. Set a domain of independent variables of test data as $\mathcal{X} \times \mathcal{X}'$. Then we have our test data $\mathcal{D}_{test} = \{(x_i, x_i', y_i)\}_{i=1}^{N_{test}}, x_i \in \mathcal{X}, x_i' \in \mathcal{X}', y_i \in \mathcal{Y}$.

Let $f : \mathcal{X} \times \mathcal{X}' \rightarrow \mathcal{Y}$ be a model trained on $\mathcal{D}_{train}$ with $\mathbb{E}_{(x,y)\sim\mathcal{D}_{train}}[\mathcal{L}(f(x, \phi), y)]$, where $\mathcal{L}$ is a loss function. We measure an *extrapolation ability* of a model f as $\mathbb{E}_{(x,x',y)\sim\mathcal{D}_{test}}[\mathcal{L}(f(x, x'), y)]$, when $X \bigcap X' = \phi$.

A model should infer the relationship between x' and other variables to extrapolate successfully. Our research hypothesis is *Can we utilize LLM's internal chemical knowledge for extrapolation, by providing additional information, e.g. SMILES of an additional raw material?* We test this hypothesis through experiments in the next section.

4 Experiments

4.1 Experimental Setup

We collect 917 data points with lab experiments measuring three dynamic mechanical analysis (DMA) properties, glass transition temperature (T_g), tan delta peak (δ), and cross-link density (v_c). Each data point contains 6 independent variables regarding raw materials (ratio between resin A: resin B_1: resin B_2: resin B_3: curing agent: catalyst) and 4 regarding curing processes (first and second curing temperature and time).

To evaluate the extrapolation ability of LLM, we construct a regression task. Our goal is to predict the DMA properties of test data given train data from a different domain, e.g. different raw materials. LLM should extrapolate the train data utilizing its chemical knowledge.

We test two extrapolation setups: (1) Additional epoxy resin. A model should infer the effect of a new epoxy resin B_i mixed with the original resin A. (2) Replaced epoxy resin. A model should infer the effect of a new epoxy resin B_2 replacing the original resin A. In both settings, train data only utilize the original resin A.

For LLM, we utilize *gpt-4-turbo* (OpenAI, 2024) with 10-shot examples for in-context learning. We select examples based on the cosine similarity of the feature vector between the train set and each test data point.

For baselines, we utilize four regression models, linear regression (**LR**), ridge regression (**RR**), random forest (**RF**, Ho, 1995), and XGBoost (**XGB**, Chen and Guestrin, 2016). To perform extrapolation with baseline regressors, we use the ratio of all epoxy resins ($A + B_1 + B_2 + B_3$) as a proxy variable.

4.2 Additional Epoxy Resin

Here, we evaluate the extrapolation ability of LLM for an additional raw material. Train and test data consists of 4 curing process variables (first and second curing temperature and time). Also, train data consists of 3 raw material-related variables as following:

- Resin A (DGEBA-based oligomer): a standard liquid bisphenol A epoxy resin with SMILES CC(C)(C1=CC=C(C=C1)OCC2CO2)C3=CC=C(C=C3)OCC4CO4
- Curing agent (Dicyandiamide): C(#N)N=C(N)N
- Catalyst: CC1=C(C=C(C=C1)NC(=O)N(C)C)NC(=O)N(C)C

However, test data contains one additional variable, the ratio of resin B_i. Here is a brief explanation of resin B_i:

- Resin B_1: CTBN(Carboyl-Terminated Butadiene Acrylonitrile) modified epoxy resin, where resin A is chemically combined with CTBN, with SMILES O=C(OCC(O)C)CC(C#N)C/C=C/CC(OCC(O)C)=O

	Resin B_1			Resin B_2			Resin B_3		
	T_g	δ	v_c	T_g	δ	v_c	T_g	δ	v_c
LR	4.61	0.0667	0.000347	4.31	0.0539	**0.000225**	8.42	0.0536	0.000329
RR	**4.58**	**0.0666**	0.000349	**4.20**	**0.0537**	0.000228	8.42	**0.0520**	0.000336
RF	5.70	0.0730	0.000310	4.99	0.0760	0.000311	9.79	0.0572	0.000301
XGB	5.61	0.0718	0.000315	4.60	0.0761	0.000237	9.00	0.0559	0.000304
Ours	7.32	0.0859	**0.000299**	5.62	0.0816	0.000288	**6.40**	0.0778	**0.000251**

Table 1: Mean absolute error (MAE) on extrapolative regression results when additional epoxy resin is added. Reported values are the average of MAE on 5 trials.

- Resin B_2: MBS type core shell rubber (CSR) modified epoxy resin, where resin A is physically combined with CSR with a ratio of 65:35 (resin A: CSR).

- Resin B_3: Dimer acid modified epoxy resin with SMILES OC(COC([R]C(OCC(O)COC1=CC=C(C(C)(C)C2=CC=C(OCC3CO3)C=C2)C=C1)=O)=O)COC(C=C4)=CC=C4C(C)(C)C5=CC=C(OCC6CO6)C=C5

LLM's goal is to predict the effect of additional resin on DMA properties only with train data and the chemical information provided above.

As a result, we obtain 385 train data with 7 independent variables and 30 test data (for each i) with 8 independent variables. The example prompt is on Appendix A. The results are on Table 1.

LLM shows superior extrapolation ability on v_c while failing on δ. Performance on T_g highly depends on the type of resin B_i. However, LLM's error shows relatively low volatility (5.62 to 7.32), unlike baseline regressors' which show high volatility (4.20 to 9.79). The result suggests that LLM can be a low-risk extrapolator, unlike utilizing regression models with proxy variables.

By examining the LLM prediction, we find out that LLM tends to adjust its prediction depending on resin type and target value. To quantitatively examine this phenomenon, we compute term frequency adjusting its final prediction. We count the number of tokens indicating its adjustment (*'increase', 'increased', 'higher', 'addition'* for ↑ and *'decrease', 'decreased', 'lower', 'reduction'* for ↓) in sentences mentioning 'resin B'. To verify the validity of the adjustment direction, we also report the average material property values of train and test sets. The results are on Table 2.

		Average			Frequency	
		Train		Test	↑	↓
B_1	T_g	161.3	↓	158.8	92	**358**
	δ	0.68	↑	0.72	**456**	322
	v_c	0.0013	↓	0.0011	67	**483**
B_2	T_g	161.3	↓	156.6	91	**248**
	δ	0.68	↑	0.74	**271**	151
	v_c	0.0013	↓	0.0011	61	**307**
B_3	T_g	161.3	↓	151.0	91	**542**
	δ	0.68	↑	0.74	**350**	527
	v_c	0.0013	↓	0.0009	117	**545**

Table 2: Term frequency analysis when additional epoxy resin is introduced. 'Average' column is the average value of target material properties (e.g. T_g) for each data set. Arrows between two columns represent the required adjustment direction (increase/decrease) from the train to the test set. 'Frequency' column is the term frequency on each word group representing increase/decrease. We mark the frequency in the right direction with bold.

Except for B_3 - δ case, LLM tends to use words mentioning *right direction* (↑ or ↓) more frequently. In other words, LLM captures the right adjustment direction. Though LLM chooses the right direction, LLM tends to overestimate the degree of adjustment and, as a result, shows higher MAEs. Moreover, the ratio between words in the right and wrong directions on δ is relatively low compared to T_g and v_c. These may suggest the reason why LLM's extrapolation ability on δ is relatively low.

An example answer is presented on Table 3.

4.3 Replaced Epoxy Resin

We present the extrapolation ability of an LLM when replacing epoxy resin from A to B_2. It is more challenging as the material properties of a

(...) CTBN is a rubbery polymer that is typically used to improve the toughness of epoxy resins. The incorporation of CTBN into an epoxy resin generally **results in a decrease in $\mathbf{T_g}$** because the CTBN phase is softer and more flexible compared to the rigid epoxy network formed by DGEBA-based resins. (...) The SMILES of Resin B indicates the presence of butadiene and acrylonitrile groups, which contribute to the elastomeric properties of the resin. This further supports the expectation of **a lower $\mathbf{T_g}$** due to increased flexibility and reduced crosslink density. (...)

Table 3: An example answer from LLM for adding resin B_1. LLM utilizes its chemical knowledge of CTBN and its SMILES to extrapolate existing data and predict a decrease in T_g.

	T_g	δ	v_c
LR	12.87	0.1606	0.001094
RR	12.78	0.1573	0.001097
RF	**12.62**	**0.1107**	0.000718
XGB	13.35	0.1199	0.000779
Ours	21.17	0.1116	**0.000467**
Ours (1 shot)	12.14	0.1080	0.000389

Table 4: MAE on extrapolative regression result when the epoxy resin A is replaced by B_2. Reported values are the average of MAE on 5 trials. The last line shows LLM's result with the 1 shot correction.

		Average			Frequency	
		Train		Test	↑	↓
B_2	T_g	161.3	↓	146.26	18	**151**
	δ	0.68	↑	0.71	**176**	106
	v_c	0.0013	↓	0.0007	13	**230**

Table 5: Term frequency analysis when epoxy resin A is replaced by B_2.

product on test data would be much more different from train data. The experimental setting is almost the same with Section 4.2, except that the ratio of resin A is 0 in the test data. We obtain 385 train data and 20 test data. The example prompt is on Appendix A. The results are on Table 4. We also perform the term frequency analysis on Table 5.

Similar to Section 4.2, LLM shows superior performance on extrapolating v_c. LLM also shows the same pattern on term frequency as in Table 2. The result suggests LLM chooses the right adjustment direction utilizing its chemical knowledge. However, MAE on T_g is high compared to baseline regressors, suggesting a similar conclusion: the direction is right, but the degree is wrong.

To check the correctability of the degree, we supply one test data point (with ground truth answer) and the previous LLM's answer for the data point back to LLM. The example prompt is on Appendix A and the result is on the last line of Table 4.

We can verify that LLM can successfully modify its degree of adjustment. As a result, LLM shows the best extrapolation ability with only one-shot correction.

5 Conclusion

In this paper, we evaluate the extrapolation ability of LLM in the chemical domain. We focus on regressing three material properties of epoxy compound when a novel raw material is introduced. We build a data set involving various raw materials and curing conditions from lab experiments. Compared to baseline regressors, LLM shows superior extrapolation ability in predicting cross-link density (v_c), while failing on tan δ peak. By examining the tokens used in LLM prediction, we find out that LLM tends to capture the right adjustment direction while failing to grasp the exact degree of adjustment. We also show that LLM successfully adjusts the degree with only 1-shot example. This result shows the potential applicability of LLM's extrapolation ability in chemical knowledge discovery.

Acknowledgements

This work was supported by the Ministry of Trade, Industry and Energy (MOTIE), Korea Institute for Advancement of Technology (KIAT) through the Virtual Engineering Platform Program (P0022334).

References

David Barrett, Felix Hill, Adam Santoro, Ari Morcos, and Timothy Lillicrap. 2018. Measuring abstract reasoning in neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 511–520. PMLR.

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. 2023. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2024. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *ICLR*. OpenReview.net.

Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2873–2882. PMLR.

Gary F. Marcus. 1998. Rethinking eliminative connectionism. *Cognitive Psychology*, 37(3):243–282.

Jeff Mitchell, Pontus Stenetorp, Pasquale Minervini, and Sebastian Riedel. 2018. Extrapolation in NLP. In *Proceedings of the Workshop on Generalization in the Age of Deep Learning*, pages 28–33, New Orleans, Louisiana. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4 technical report.

David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. Analysing mathematical reasoning abilities of neural models. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Taylor W. Webb, Zachary Dulberg, Steven M. Frankland, Alexander A. Petrov, Randall C. O'Reilly, and Jonathan D. Cohen. 2020. Learning representations that support extrapolation. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Geyan Ye, Xibao Cai, Houtim Lai, Xing Wang, Junhong Huang, Longyue Wang, Wei Liu, and Xiangxiang Zeng. 2023. Drugassist: A large language model for molecule optimization. *arXiv preprint arXiv:2401.10334*.

Haiteng Zhao, Shengchao Liu, Chang Ma, Hannan Xu, Jie Fu, Zhi-Hong Deng, Lingpeng Kong, and Qi Liu. 2023. GIMLET: A unified graph-text model for instruction-based molecule zero-shot learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, Guodong Shen, Xin Chen, and Kai Yu. 2024. Chemdfm: Dialogue foundation model for chemistry.

A Example Prompts

Prompt for Section 4.2
Predict the *[PROPERTY]* of an epoxy product with the following information. You should infer the effect of a new resin B.: Epoxy resin A (DGEBA-based oligomer) is a standard liquid bisphenol A epoxy resin with SMILES CC(C)(C1=CC=C(C=C1)OCC2CO2)C3=CC=C(C=C3)OCC4CO4. Epoxy resin B is a CTBN(Carboyl-Terminated Butadiene Acrylonitrile) modified epoxy resin, where resin A is chemically combined with CTBN, with SMILES O=C(OCC(O)C)CC(C#N)C/C=C/CC(OCC(O)C)=O. SMILES of curing agent (Dicyandiamide) is C(#N)N=C(N)N. SMILES of catalyst is CC1=C(C=C(C=C1)NC(=O)N(C)C)NC(=O)N(C)C. The following is another data point measuring the *[PROPERTY]*. | Ratio ((resin A: resin B): curing agent: catalyst) | First curing condition | Second curing condition | *[PROPERTY]* | | (93.02: 0.0): 6.05: 0.93 | 1.0 hour in 100.0°C | 0.5 hour in 130.0°C | *[PROPERTY1]* | | (93.02: 0.0): 6.05: 0.93 | 1.0 hour in 100.0°C | 1.5 hour in 130.0°C | *[PROPERTY2]* | (...) Fill in the '?'. | (83.72: 9.3): 6.05: 0.93 | 1.5 hour in 100.0°C | 1.0 hour in 130.0°C | ? |
Prompt for Section 4.3
Predict the *[PROPERTY]* of an epoxy product with the following information. You should infer the effect of a new resin B.: Epoxy resin A (DGEBA-based oligomer) is a standard liquid bisphenol A epoxy resin with SMILES CC(C)(C1=CC=C(C=C1)OCC2CO2)C3=CC=C(C=C3)OCC4CO4. Epoxy resin B is an MBS type core shell rubber (CSR) modified epoxy resin, where resin A is physically combined with CSR with a ratio of 65:35 (resin A:CSR). SMILES of curing agent (Dicyandiamide) is C(#N)N=C(N)N. SMILES of catalyst is CC1=C(C=C(C=C1)NC(=O)N(C)C)NC(=O)N(C)C. The following is another data point measuring the *[PROPERTY]*. | Ratio ((resin A: resin B): curing agent: catalyst) | First curing condition | Second curing condition | *[PROPERTY]* | | (93.02: 0.0): 5.12: 1.86 | 1.0 hour in 100.0°C | 1.0 hour in 120.0°C | *[PROPERTY1]* | | (93.02: 0.0): 5.12: 1.86 | 1.0 hour in 90.0°C | 1.0 hour in 130.0°C | *[PROPERTY2]* | (...) Fill in the '?'. | (0.0: 93.65): 4.35: 2.01 | 1.0 hour in 90.0°C | 1.5 hour in 120.0°C | ? |
Additional Prompt for 1-shot Correction
Note that for the data point: | (0.0: 93.65): 4.35: 2.01 | 1.0 hour in 90.0°C | 1.5 hour in 120.0°C |, your answer was *[PREVIOUS ANSWER]*. But the true value was *[PROPERTY3]*.

Table 6: Example prompts for experiments. *[PROPERTY]* can be a glass transition temperature (T_g), tan delta peak (δ), or cross-link density (v_c).