Towards Creating Benchmark Datasets of Universal Neural Network Potential for Material Discovery

So Takamoto
Preferred Networks, Inc.
Otemachi Bldg. 1-6-1 Otemachi, Chiyoda-ku, Tokyo, Japan 100-0004
takamoto@preferred.jp

Chikashi Shinagawa
Preferred Networks, Inc.
Otemachi Bldg. 1-6-1 Otemachi, Chiyoda-ku, Tokyo, Japan 100-0004
shinagawa@preferred.jp

Nontawat Charoenphakdee
Preferred Networks, Inc.
Otemachi Bldg. 1-6-1 Otemachi, Chiyoda-ku, Tokyo, Japan 100-0004
nontawat@preferred.jp

Abstract

Recently, neural network potentials (NNPs) have been shown to be particularly effective in conducting atomistic simulations for computational material discovery. Especially in recent years, large-scale datasets have begun to emerge for the purpose of ensuring versatility. However, we show that even with a large dataset and a model that achieves good validation accuracy, the resulting energy surface can be quite delicate and the easily reach unrealistic extrapolation regions during the simulation. We first demonstrate this behavior using a DimeNet++ trained on Open Catalyst 2020 dataset (OC20). Based on this observation, we propose a hypothesis that for NNP models to attain the versatality, the training dataset should contain a diverse set of virtual structures. To verify this, we have created a relatively much smaller benchmark dataset called "High-temperature Multi-Element 2021" (HME21) dataset, which was sampled through a high-temperature molecular dynamics simulation and has less prior information. We conduct benchmark experiments on HME21 and show that training a TeaNet on HME21 can achieve better performance in reproducing the absorption process, although HME21 does not contain corresponding atomic structures. Our findings indicate that dataset diversity can be more essential than the dataset quantity in training universal NNPs for material discovery.

1 Introduction

In recent years, neural network potentials (NNPs) have rapidly gained attention owing to the high expressive power of neural networks (NNs) combined with the availability of large-scale datasets. As datasets and models evolve, the scope of NNP applications has gradually expanded. As a benchmark for molecular systems, the QM9 dataset [8, 9], which covers possible patterns of small molecules, has been widely used. Initially, NNPs for organic molecules have focused on H, C, N, and O, which are the major elements in organic molecules. In subsequent studies, NNPs have been extended to include elements such as S, F, and Cl [4, 12]. For NNPs targeting crystal structures [3, 16], the Materials

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

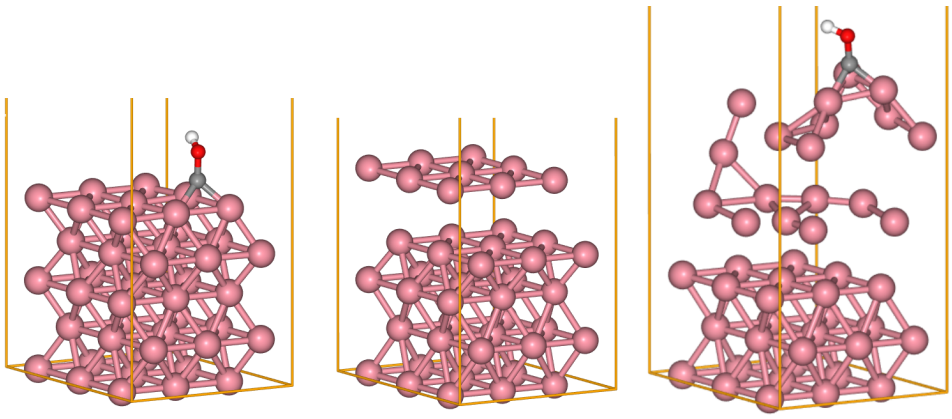

Figure 1: Optimized structures created by dimenetpp_all. The surface and molecule are Co (0001) and COH, respectively. Left: Started from the well-prepared adsorbed structure. Middle: Pure surface structure. Right: Started from the structure which the molecule attached on pure surface structure. The energy of the right structure is lower than that of the left one. The figures were drawn using the VESTA visualization package. [7] Those structures were taken from [15].

Project [6], a large-scale materials database based on DFT calculations, is often used as a benchmark dataset. The Open Catalyst Project, which targets molecular adsorption in catalytic reactions, has constructed a massive surface adsorption structure dataset known as the Open Catalyst 2020 (OC20) dataset [2, 18]. In this way, the area covered by NNPs has gradually expanded.

However, even with the construction of such a large dataset, we found that there is still a significant technical gap in the realization of general-purpose atomic simulations. To demonstrate this, we used the Open Catalyst Project baseline model. We used the publicly available trained model dimenetpp_all from the Open Catalyst Project implementation (https://github.com/Open-Catalyst-Project/ocp/tree/v0.0.3). This model was trained with the DimeNet++ [5] architecture on all data from the S2EF task [2] included in OC20 (approximately 10^8 samples).

The reproduce result is shown in Fig. 1. We confirmed that the simulation worked as expected if we already have the well-prepared adsorbed structure (Fig. 1 Left). However, when we tried to reproduce the corresponding structure generation process by adding molecule on the optimized surface structure (Fig. 1 Middle), the simulation went to the physically unrealistic state. Worse, the broken structure finally obtained through this process was shown to be even more energetically stable than the well-prepared structure (Fig. 1 Right). It indicates that even there is a large dataset and a model that achieves a certain level of validation accuracy for that dataset, and even the simulation target can be considered to be in-domain, the estimated energy surface by the model is still quite delicate and can easily reach unrealistic extrapolation regions during the simulation.

The above result suggests that the conventional way of creating datasets has difficulties in applying them to atomic simulations. All previously proposed datasets were generated based on known structures. However, the above phenomenon can be attributed to the fact that there is no way of knowing that an unrealistic structure is unrealistic for trained models. Therefore, as an opposite idea, we can consider a dataset that is generated to cover as wide a range of phase space as possible, rather than narrowing the domain.

Here, we introduce an atomic structure dataset called the high-temperature multi-element 2021 (HME21) dataset. HME21 dataset contains multiple elements in a single structure and was sampled through a high-temperature molecular dynamics simulation. Thus, the structures are far from stable and contain less prior specific domain knowledge, such as the molecule or crystal structures. The structures produced in this fashion are a class of the most challenging configurations for prediction because of their highly disordered nature. On the other hand, this dataset is expected to provide a highly stringent assessment of the universality of the model. In Section 3, we show that a NNP architecture designed to learn such highly disordered structures can indeed learn HME21 dataset within reasonable accuracy. We also show, surprisingly, that the NNP model trained in this way is

able to reproduce the surface adsorption simulation described above, even though HME21 dataset do not have the corresponding structures.

We believe that the creation of more challenging dataset such as disordered, high-entropy, and less domain specific ones will greatly improve the NNP capability, and HME21 can be one of the standard benchmark for future universal NNP development.

2 High-temperature multi-element 2021 (HME21) dataset

In this section, we describe a highly disordered dataset called High-temperature multi-element 2021 (HME21) dataset. The structures were sampled by high temperature molecular dynamics simulation with NNP. The sampled temperature was from 500 K to 10000 K. Each structure contains up to 20 types of elements. For more detail, see [15]. HME21 is available on [14].

Element types: HME21 contains multiple elements in a single structure and was sampled through a high-temperature molecular dynamics simulation. There are a total of 37 elements in the HME21 dataset, which are H, Li, C, N, O, F, Na, Mg, Al, Si, P, S, Cl, K, Ca, Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Mo, Ru, Rh, Pd, Ag, In, Sn, Ba, Ir, Pt, Au, and Pb.

Data size: There are $24,949$ structures in HME21 dataset. We split the dataset into training ($19,956$ structures), validation ($2,498$ structures), and test ($2,495$ structures) splits at the approximated ratio of $8:1:1$.

Task: The target values are energy and atomic forces. The energy is shifted such that the energy of a single atom located in a vacuum becomes zero. The length is in angstroms (1 Å$= 10^{-10}$ m), and the energy is in electronvolts (eV).

3 Experimental results

In this section, we discuss experimental results of the model trained on HME21 dataset. We begin this section by evaluating the performance of comparison models using mean absolute error (MAE) for energy and force prediction tasks. After that, we demonstrate the ability of reproducing the adsorption process of TeaNet [13], which is the model that performs best in terms of force and energy MAEs on HME21.

3.1 Neural network architecture benchmark using HME21

We conduct benchmark experiments of energy/force prediction task in HME21 using recent NNP architectures. For this benchmark, we selected TeaNet [13], SchNet [10], PaiNN [11], and NequIP [1]. TeaNet treats tensor representations as higher-order geometric features. SchNet uses the bond length for spatial information and employs a convolution with rotationally invariant filters. This has been well examined using various datasets, and its limited representation power has been discussed. PaiNN incorporates a vector representation to resolve the problem of a limited representation of rotationally invariant filters of SchNet. On the other hand, NequIP uses spherical harmonics-based representations. The experimental code for both SchNet and PaiNN is based on the repository found at https://github.com/learningmatter-mit/NeuralForceField, whereas the experimental code for NequiP is based on the repository found at https://github.com/mir-group/nequip. The implementation codes for SchNet, PaiNN, Nequip are under the MIT license, while TeaNet is under the CC BY 4.0 license. The experiments were run on NVIDIA V100 GPUs.

Next, we discuss the choice of hyperparameter. To optimize the performance with respect to the validation set, the hyperparameter selection procedure is based on a grid search and manual hyperparameter tuning. For TeaNet, we use a four-layer model. We first set the energy loss coefficient c_{le} (energy per atom MSE) to 0.0001 and retrained it using $c_{le} = 1.0$ and $c_{le} = 10.0$, whereas the force loss coefficient c_{lf} remained constant at 1.0. The batch size was set to 16, and the learning rate was initialized to 0.001. For SchNet, we use a four-layer model, where the energy loss coefficient was set to 0.05, the batch size was set to 32, and the learning rate was initialized to 0.0005. For PaiNN, we use a three-layer model, where the energy loss coefficient was set to 0.05, the batch size was set to 32, and the learning rate was initialized to 0.0005. For NequIP, we use a five-layer model with different maximum rotation orders $l_{\max} \in \{0, 1, 2\}$. For the five-layer model, the energy loss

Architecture	Energy MAE [meV/atom]	Force MAE [eV/Å]	Force XYZ MAE [eV/Å]
TeaNet	19.6	0.174	0.153
SchNet	33.6	0.283	0.247
PaiNN	22.9	0.237	0.208
NequIP ($l_{\max} = 0$)	52.2	0.249	0.225
NequIP ($l_{\max} = 1$)	53.3	0.233	0.206
NequIP ($l_{\max} = 2$)	47.8	0.199	0.175

Table 1: Benchmark performance of NNP for the force and energy prediction for the HME21 dataset. Energy MAE corresponds to the mean absolute error of the energies of structures divided by their numbers of atoms, Force MAE corresponds to the mean absolute error of the norm of force vectors, and Force XYZ MAE corresponds to the mean absolute error of the force vector component. This table was taken from [15].

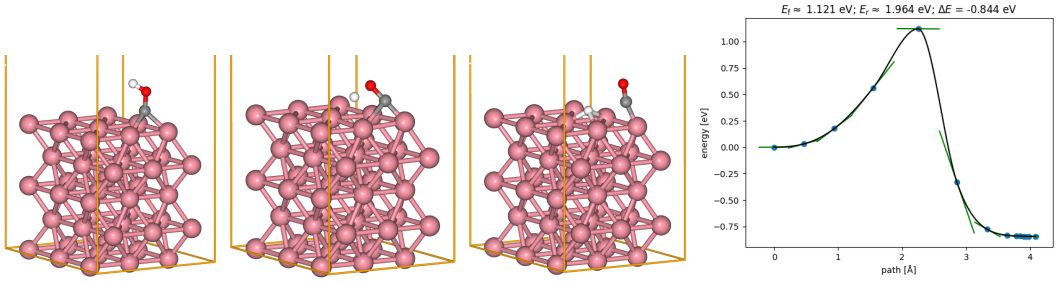

Figure 2: Reaction path of the dissociation of the COH molecule on Co surface obtained by TeaNet trained on HME21. Left: Snapshots of the reaction path (corresponding to initial, transition, and final state, respectively). Right: Energy trajectory. $\mathrm{E}_f, \mathrm{E}_r, \Delta\mathrm{E}$ are the activation energy for forward path, activation energy for backward path, and the reaction energy, respectively. The corresponding energies obtained by DFT calculation is 0.80 eV, 1.81 eV, and -1.01 eV, respectively [17].

coefficient was set to 0.01 and the learning rate was initialized to 0.001. For $l_{\max} \in \{0, 1\}$, we found that setting the batch size to 32 worked best, whereas for $l_{\max} = 2$, setting the batch size to 64 was preferable. We set the cutoff distance to 6.0 Å for all architectures.

The results are presented in Table 1. TeaNet performed well in terms of both the energy and force metrics, which indicates that the TeaNet architecture is suitable for multielement structures which are far from stable coordination. The trained model is available on [15].

3.2 Reproducing adsorption process with TeaNet trained on HME21

Here, we demonstrate the effectiveness of TeaNet trained on HME21 for surface catalytic reaction. We used this model to run the same simulations as those presented in Section 1. We found that both the surface-only structure and adsorbed structure were successfully obtained using structural optimization calculations. More surprisingly, even reaction pathway analysis using nudged elastic band (NEB) calculations of the dissociation reactions of molecules on the surface could be performed as well (Fig. 2). Although there is still room for improvement in accuracy, the activation energy obtained is comparable to that of the DFT calculation. It means that the energy surface along the minimum energy path between two stable points, including the transition state, is smooth enough. This is remarkable result because: 1. Even though it is generally known that the output of neural networks is not always smooth, the behavior of the learned model is smooth enough in this domain of inference and acquires the desired behavior up to the second derivative. 2. No information in the target domain (crystal surface, molecule, adsorption, and dissociation) are included in the dataset. 3. The dataset size is relatively small.

4 Conclusions

We proposed HME21 dataset, a dataset composed of $24,949$ data points with 37 elements with highly disordered structures. In our benchmark experiments, TeaNet performs best on HME21 dataset

in terms of MAE in force and energy predictions tasks and it also succeeded in reproducing the adsorption process for the structure in the extrapolation region. On the other hand, DimeNet++ trained on a large-scale OC20 dataset has difficulty to perform this simulation task. Our results emphasize that dataset diversity is essential for developing universal NNPs. HME21 dataset has also been recently used as one part of the dataset for the development NNPs called Preferred Potential [15] that is capable of handling any combinations of 45 elements and can achieves desirable performance. We believe that HME21 can be one of the standard benchmark datasets to stimulate future research in universal NNP development.

5 Broader Impact

In this work, we propose HME21 dataset for learning NNPs that can support a large number of elements. With this dataset, we believe that it can be useful towards to goal of creating universal NNPs. Once an effective universal NNP can be obtained, we can use it for material discovery. New useful materials can be essential to drive many technology domains such as carbon-neutral energy and renewable energy systems for sustainable technology development, which can improve people's quality of life. On the other hand, similarly to many useful tools in this world, the misuse of material discovery could also lead to harmful applications, e.g., the creation of weapons or dangerous chemical substances. This work does not encourage the use of this technology for such applications that have negative impact to the society. At least, HME21 does not contain radioactive elements. In terms of privacy impact, we note that HME21 does not have direct negative impact in this perspective because it does not contain human-related sensitive information since it is an atomic structure dataset sampled through molecular dynamics simulation.

Acknowledgements

We would like to thank Y. Tsuboi for giving comments to improve the manuscript and I. Kurata for preparing atomic structures in Fig. 1

References

[1] Simon Batzner, Tess E. Smidt, Lixin Sun, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, and Boris Kozinsky. Se(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, 2021.

[2] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, and Zachary Ulissi. Open catalyst 2020 (oc20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072, 2021. doi: 10.1021/acscatal.0c04525. URL `https://doi.org/10.1021/acscatal.0c04525`.

[3] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9):3564–3572, 2019. doi: 10.1021/acs.chemmater.9b01294. URL `https://doi.org/10.1021/acs.chemmater.9b01294`.

[4] Christian Devereux, Justin S. Smith, Kate K. Davis, Kipton Barros, Roman Zubatyuk, Olexandr Isayev, and Adrian E. Roitberg. Extending the applicability of the ani deep learning molecular potential to sulfur and halogens. *Journal of Chemical Theory and Computation*, 16(7):4192–4202, 2020. doi: 10.1021/acs.jctc.0c00121. URL `https://doi.org/10.1021/acs.jctc.0c00121`. PMID: 32543858.

[5] Johannes Gasteiger, Shankari Giri, Johannes T. Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. In *NeurIPS-W*, 2020.

[6] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin a.

Persson. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013. ISSN 2166532X. doi: 10.1063/1.4812323. URL http://link.aip.org/link/AMPADS/v1/i1/p011002/s1&Agg=doi.

[7] Koichi Momma and Fujio Izumi. Vesta 3 for three-dimensional visualization of crystal, volumetric and morphology data. *Journal of applied crystallography*, 44(6):1272–1276, 2011.

[8] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1): 140022, Aug 2014. ISSN 2052-4463. doi: 10.1038/sdata.2014.22. URL https://doi.org/10.1038/sdata.2014.22.

[9] Lars Ruddigkeit, Ruud van Deursen, Lorenz C. Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875, Nov 2012. ISSN 1549-9596. doi: 10.1021/ci300415d. URL https://doi.org/10.1021/ci300415d.

[10] K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K.-R. Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 992–1002, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[11] Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pages 9377–9388. PMLR, 2021.

[12] J. S. Smith, O. Isayev, and A. E. Roitberg. Ani-1: an extensible neural network potential with dft accuracy at force field computational cost. *Chem. Sci.*, 8:3192–3203, 2017. doi: 10.1039/C6SC05720A. URL http://dx.doi.org/10.1039/C6SC05720A.

[13] So Takamoto, Satoshi Izumi, and Ju Li. TeaNet: Universal neural network interatomic potential inspired by iterative electronic relaxations. *Computational Materials Science*, 207:111280, 2022. ISSN 0927-0256. doi: https://doi.org/10.1016/j.commatsci.2022.111280. URL https://www.sciencedirect.com/science/article/pii/S0927025622000799.

[14] So Takamoto, Chikashi Shinagawa, Daisuke Motoki, Kosuke Nakago, Wenwen Li, Iori Kurata, Taku Watanabe, Yoshihiro Yayama, Hiroki Iriguchi, Yusuke Asano, Tasuku Onodera, Takafumi Ishii, Takao Kudo, Hideki Ono, Ryohto Sawada, Ryuichiro Ishitani, Marc Ong, Taiki Yamaguchi, Toshiki Kataoka, Akihide Hayashi, Nontawat Charoenphakdee, and Takeshi Ibuka. High-temperature multi-element 2021 (HME21) dataset. 4 2022. doi: 10.6084/m9.figshare.19658538. URL https://figshare.com/articles/dataset/High-temperature_multi-element_2021_HME21_dataset/19658538.

[15] So Takamoto, Chikashi Shinagawa, Daisuke Motoki, Kosuke Nakago, Wenwen Li, Iori Kurata, Taku Watanabe, Yoshihiro Yayama, Hiroki Iriguchi, Yusuke Asano, Tasuku Onodera, Takafumi Ishii, Takao Kudo, Hideki Ono, Ryohto Sawada, Ryuichiro Ishitani, Marc Ong, Taiki Yamaguchi, Toshiki Kataoka, Akihide Hayashi, Nontawat Charoenphakdee, and Takeshi Ibuka. Towards universal neural network potential for material discovery applicable to arbitrary combination of 45 elements. *Nature Communications*, 13(1):1–11, 2022.

[16] Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, 120:145301, Apr 2018. doi: 10.1103/PhysRevLett.120.145301. URL https://link.aps.org/doi/10.1103/PhysRevLett.120.145301.

[17] Bart Zijlstra, Robin J.P. Broos, Wei Chen, Ivo A.W. Filot, and Emiel J.M. Hensen. First-principles based microkinetic modeling of transient kinetics of co hydrogenation on cobalt catalysts. *Catalysis Today*, 342:131–141, 2020. ISSN 0920-5861. doi: https://doi.org/10.1016/j.cattod.2019.03.002. URL https://www.sciencedirect.com/science/article/pii/S0920586118316158. SI: Syngas Convention 3.

[18] C. Lawrence Zitnick, Lowik Chanussot, Abhishek Das, Siddharth Goyal, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Thibaut Lavril, Aini Palizhati, Morgane Riviere, Muhammed Shuaibi, Anuroop Sriram, Kevin Tran, Brandon Wood, Junwoong Yoon, Devi Parikh, and Zachary Ulissi. An introduction to electrocatalyst design using machine learning for renewable energy storage, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 3.2. We suggest that the accuracy still has room for improvement.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See the end of the first paragraph in Section 3.1.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] See the end of the first paragraph in Section 3.1.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]