# ClimFormer – A Spherical Transformer Model for Long-term Climate Projections

**Salva Rühling Cachay**
UC San Diego

**Peetak Mitra**
Xerox PARC

**Haruki Hirasawa**
University of Victoria

**Sookyung Kim**
Xerox PARC

**Subhashis Hazarika**
Xerox PARC

**Dipti Hingmire**
University of Victoria

**Phil Rasch**
University of Washington

**Hansi Singh**
University of Victoria

**Kalai Ramea***
Xerox PARC

## Abstract

Clouds play an important role in balancing the Earth's energy budget. Research has indicated a rise in global average temperatures will lead to thinning of stratocumulus low clouds acting as a positive feedback on warming. Current state-of-the-art Earth System Models do not resolve cloud physics appropriately due to spatial resolution limitations, making it harder to model the cloud-climate feedback. In this study, we propose to learn this feedback with a transformer. To better respect the spatial structure of Earth, we transform the data to a spherical grid. Our resulting spherical transformer called ClimFormer–based on Fourier Neural Operator mixing–trained on climate simulations, is able to model this important energy exchange mechanism, and performs strongly on an out-of-distribution evaluation on ERA5 data.

## 1 Introduction

The Earth mainly receives energy from the sun in the form of radiation. Clouds play a critical role in modulating the transfer of radiation through the atmosphere. Cloud-climate feedbacks are important drivers of atmospheric radiation changes and play a key role in variations of key climate prognostics such as temperature, pressure, and precipitation. As clouds occur on scales smaller than typical Earth System Model (ESM) spatial resolutions, they are typically represented by semi-empirical parameterizations. Uncertainties these cloud processes and their representation are the dominant source of inter-ESM uncertainty in the projected climate response to $CO_2$ increases [25], limiting the precision with which ESMs can be used to make long-term climate projections.

Even with the promise of more computation and the recently unveiled GPU accelerated climate models [11], the physics of the cloud-climate processes are sometimes poorly understood. The advent of Fourier Neural Operators (FNO) offers a promising alternative to learning complicated Partial Differential Equations (PDE) by learning the mapping between functional spaces directly from data [15]. It has shown great promise in learning operators for multi-scale systems including geophysical turbulence [19, 3, 4]. In addition to FNO, Transformers [22] has revolutionized many domains of applied machine learning research [12, 17]. At the core of the Transformer's success is the mechanism of attention, which makes this particular framework well suited to model long-range interactions and teleconnections [7] for climate simulations. The concept of applying Transformers to model weather and climate is not entirely novel. In the recent past, Temporal Fusion Transformers [16] have

*corresponding author, `kramea@parc.com`

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

been used to model geophysical turbulence [4], seasonal climate forecasting [24], meteorological prognostics such as wind-speed [23], and extreme weather[5]. Pathak et al. proposed FourCastNet, a transformer that builds on Adaptive Fourier Neural Operator (AFNO) mixing [9], and applied it to model global weather [21]. Bai et al. [1] proposed Rainformer for precipitation nowcasting, which is based on an architecture that combines CNN and SwinTransformer [18]. Gao et al.[8] proposed EarthFormer, which introduces a novel attention block, called *Cuboid Attention*, and is shown to outperform the Rainformer model for nowcasting applications.

### 1.1 Our contribution

We adapt the global weather forecasting, euclidean grid-based, AFNO Transformer [9, 21] to learn the cloud-climate feedback operator on a spherical grid that better respects the inherent structure of Earth. Through it, we are able of making long-term climate projections regarding intervention scenarios. For the spherical mapping, we use a geodesy-aware spherical co-ordinate system to uniformly sample raw data.

## 2 Background

### 2.1 Climate Dynamics and Fluctuation Dissipation Theorem

The goal of this model is to learn a mapping between cloud changes and their consequent impacts on key climate variables.

$$\Omega(S) : \delta C(\overrightarrow{x}, h, \tau) \rightarrow \left\{ \delta\Psi_{atm}\overrightarrow{(x)}, \delta\Psi_{ocn}\overrightarrow{(x)}, \delta T_s\overrightarrow{(x)}, \delta P\overrightarrow{(x)} \right. \tag{1}$$

However, as there is no existing library of cloud perturbation ESM simulations from which to learn this mapping, we instead adopt an approach informed by the Fluctuation Dissipation Theorem (FDT). The FDT posits that in certain systems, the linear response of the system to a perturbation can be expressed in terms of the fluctuations of the system at equilibrium [13]. The climate system is one such system, wherein chaotic internal variability can be used to estimate externally forced climate response [20]. Thus, by training on a large pool of natural ESM fluctuations, the model is able to learn mappings such that it is able to project the climate response to changes in cloud radiative properties.

In a conventional application of FDT, a linear response function (LRF, $L$) that maps between a set of predictors ($\delta f$) to a set of predictands ($\delta X$). For example, Liu et al., 2018 formulated a FDT operator for the equilibrium response to a constant forcing $\delta f$ as

$$\delta X = L_{FDT}^{-1}\delta f \tag{2}$$

where

$$L_{FDT} = -\left[\int_0^\infty C(\tau)C(0)^{-1}d\tau\right]^{-1} \tag{3}$$

for covariance matrices $C$ and time lag $\tau$. To effectively estimate the LRF, it is critical to have a large sample of internal variability with which to estimate the covariance matrix. However, because the climate system can respond non-linearly to forcing, there are limitations to the LRF-based approach. Thus, we replace the LRF with an ML model, allowing a more comprehensive determination of the relationships between variables. As we are learning based on natural internal fluctuations, we must assume that the relationships between cloud radiation and surface climate are not substantively modified by the levels of climate change we study herein (historical and near future warming).

## 3 Methods

The data are taken from monthly simulations of the CESM2 model. CESM2 is a fully-coupled, community, global climate model that provides state-of-the-art computer simulations of the Earth's past, present, and future climate states [6]. As described in the following section, we first map the raw ESM data from its native euclidean grid to a spherical grid and then apply various climate pre-processing techniques.

### 3.1 Spherical-grid aware modeling

Most ESMs use a Cartesian grid to represent the Earth. Typically, models use climate data in a uniform 2D rectangular gridded pattern. While this may suffice local/regional modeling attributes, they do not capture the physical/geodesy properties of the Earth, particularly as the focus moves away from the equator. For this reason, we developed a geodesy-aware sampling that converts 2D rectangular gridded coordinates to a geodesic grid type. There are several ways a geodesic grid can be manifested and in this study we present results from using icosahedral grids [2].

We first develop a 'backbone' structure of a spherical coordinate system (icosahedron, healpix, etc.). The properties of the spherical coordinates, such as levels or sub-divisions, are given as input. At this point, the coordinates are simply graph networks. In the next step, we assign latitude and longitude values to the graph network (x, y) so that they can be manifested in a geographical coordinate system. Finally, we use the raw data from reanalysis or ESM output and perform bilinear interpolation to obtain the final spherically-sampled data. The resulting data contain 40962 icosahedral pixels per time snapshot, which corresponds to 110 km spatial (or 1 degree) resolution.

Having mapped the CESM2 data to a spherical grid, it is further pre-processed as follows:

- **Remove seasonal cycle (Deseasonalize)**: We perform this process to remove any trends in the season to prepare a seasonal stationary time series data.
- **Remove trend (Detrend)**: We fit a third degree polynomial to remove any trend in data over time. This removes secular trends (for example, rising temperatures as atmospheric $CO_2$ increases) and allows the model to be trained on fluctuations due to internal variability, rather than the forced response.
- **Normalized anomalies**: The anomaly at each grid point is calculated relative to a running mean that is computed over a centered 30-year window for that grid point and month. Anomalies are normalized by dividing by the standard deviation of the anomaly over the same 30-year window for that grid point and month.
- **Remap data to Sphere-Icosahedral**: Use Climate Data Operators to bilinearly remap disparate grids to uniform level-6 Sphere-Icosahedral grid.

### 3.2 Problem formulation

After transforming the raw data as described above, we can now consider inputs $X_t \in \mathbb{R}^{s \times d_{\text{in}}}$ and corresponding outputs $Y_{t+\Delta t} \in \mathbb{R}^{s \times d_{\text{out}}}$, where $t$ is the timestep of the inputs, $\Delta t$ the prediction horizon (i.e. number of timesteps to predict into the future), $s$ is the spatial dimension of the icosahedral, and $d_{\text{in}}, d_{\text{out}}$ are the number of input and output features, respectively.

In our experiments we focus on learning simultaneous responses to cloud forcings, that is $\Delta t = 0$. As shown in Table 1 (in the appendix), we have that $d_{\text{in}} = 7$ and $d_{\text{out}} = 3$, and when using the level-6 sphere icosahedral data grid it holds that $s = 40962$.

### 3.3 Transformers for spherical data

Motivated by the recent successes of transformer-based architectures in various domains, we explore their potential for our problem and spherical data grid. However, prior work and applications usually focused on sequential or euclidean data – applying transformers to spherical Earth data is a novelty. For the present exploratory study we choose to adapt transformers in the simplest possible way to spherical data. To do so, we simply view the spatial icosahedral pixel dimension, $s$, as the token dimension of a transformer. With this problem formulation, we now only need to embed/project the input data into a higher dimensionality, $d$, before passing it through any standard transformer encoder architecture. On the output side, we apply a linear layer that projects the $d$ channels of the transformer output to the $d_{\text{out}}$ predicted output features.

| Category | Variable | Description |
|---|---|---|
| Input | cres | TOA cloud radiative effect in shortwave |
| Input | cresSurf | Surface cloud radiative effect in shortwave |
| Input | crel | TOA cloud radiative effect in longwave |
| Input | crelSurf | Surface cloud radiative effect in longwave |
| Input | netTOAcs | TOA radiation without clouds (clear-sky) |
| Input | netSurfcs | Net clearsky surface radiation and heat flux |
| Input | lsmask | Land-sea binary mask |
| Output | tas | 2-metre air temperature |
| Output | ps | Surface pressure |
| Output | pr | Precipitation |

Table 1: Input and output variables in our dataset.

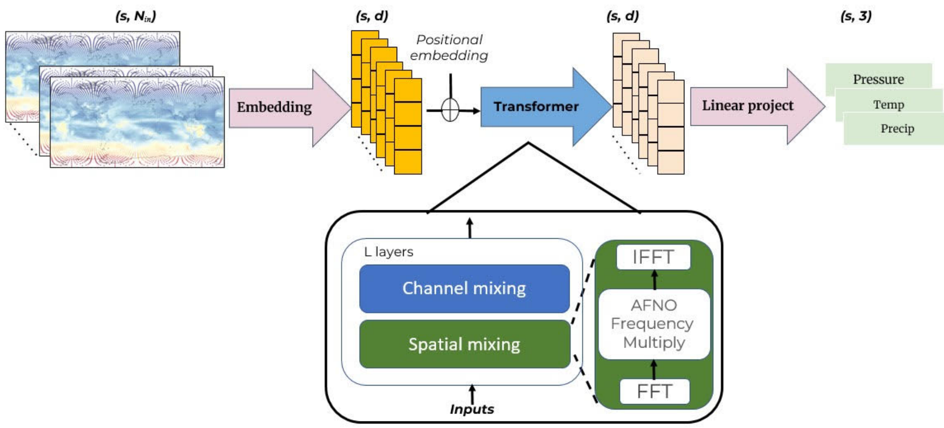


Figure 1: Schematic of the workflow (image partially adopted from [9]).

.

As for the concrete transformer architecture, we choose to build upon the Adaptive Fourier Neural Operator (AFNO) transformer [9]. It has proven successful for relevant applications such as weather forecasting (using an euclidean grid) [21], and enjoys a better complexity of $\mathcal{O}(Nd^2/k + Nd\log N)$ over the $\mathcal{O}(N^2d + 3Nd^2)$ complexity of the standard self-attention transformers, where $N := s$ is the number of input tokens and $k$ a hyperparameter of the AFNO. This is especially important for our problem where the token size is around $40k$. The resulting modeling workflow is shown in Figure 1.

## 4 Results and Discussions

We train the ClimFormer model using CESM2 data, prepared as described in Section 2.1. In this section we present results from the learned operator for inference on the ERA5 dataset [10], which is a reanalysis product produced by ECMWF. Reanalyses are models which ingest large quantities observational data to estimate the historical evolution of the atmosphere, thus providing an estimate of a wide range of atmospheric variables over the entire globe. While these data are not exactly the same as observational data, they are the best method of obtaining physically consistent and complete climate data representing the recent past of the Earth's atmosphere.

Figure 2 shows the predictive ability of the model at a global level. The model is able to predict accurately for most regions of the Earth and the deviance is fairly minor (5-7%) from the ground

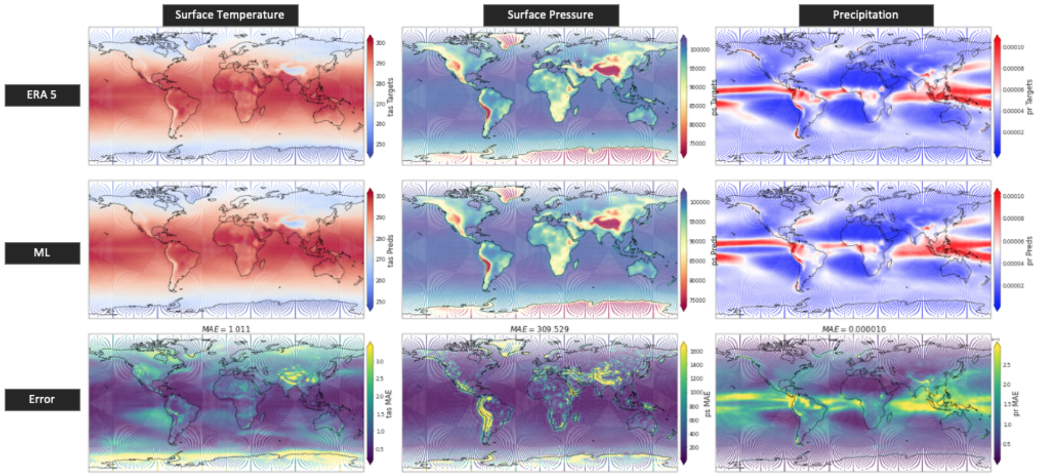


Figure 2: ClimFormer predictions on key climate prognostics shown on ERA5 data. Top row is ERA5 data, middle row are the corresponding ClimFormer predictions, and bottom row is the pixelwise error difference.

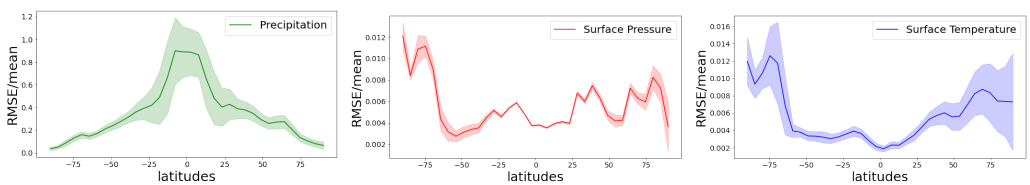


Figure 3: Longitudinally averaged data showing model error, spread is the uncertainty in ground truth data

truth. In Figure 3, the longitudinally averaged RMSE is plotted for the model output variables to ascertain the prediction trends across major zones of the Earth system. It appears that the precipitation skill is poorest (relative) in the tropics (around latitude 0), which is the zone of highest activity in terms of global moisture budget (the interplay between evaporation and precipitation). The surface temperature and pressure errors spike in the poles, but perform exceedingly well in the tropics which receives the most energy, thereby building confidence in the model's ability to learn the functional operator.

## 5 Conclusions and Summary

In conclusion, this study proposes a spherical Transformer model, ClimFormer, to accurately represent the cloud-climate feedback. We test it on out-of-distribution ERA5 datasets and show it is able to generalize under moderate distributional shifts. In the ongoing phase of this study, we are using this emulator to make climate projections under intervention scenarios such as Marine Cloud Brightening [14]. This is especially useful since a model year run on ClimFormer takes about 2 seconds on a single V100 GPU, whereas a comparable CESM2 simulation takes over 20 hours on 200 cores, enabling larger design space exploration which, subsequently, can be validated by running targeted ultra high-resolution CESM2 simulations.

## 6 Broader Impact Statement

The Earth is warming at an unprecedented rate, and the increasing frequency of extreme weather events is a testament to that. While reducing emissions and net zero pledges is at the center of most attention in terms of climate change mitigation efforts, an increasing focus is being placed onto climate intervention techniques such as Stratospheric Aerosol Injection and Marine Cloud Brightening. The scope of this study is to not endorse any particular technique over another (or climate intervention as a necessary step overall) but to simply conduct a scientifically thorough academic/research study in terms of scenario planning, should these steps be necessary as a last resort in the future.

## 7 Acknowledgments

The development of AIBEDO is funded under the DARPA AI-assisted Climate Tipping-point Modeling (ACTM) program under award DARPA-PA-21-04-02. The computing resources were courtesy a generous grant from Amazon Web Services. More information on the project can be found at https://aibedo.readthedocs.io/en/latest/index.html

## References

[1] Cong Bai, Feng Sun, Jinglin Zhang, Yi Song, and Shengyong Chen. Rainformer: Features extraction balanced network for radar-based precipitation nowcasting. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[2] John R Baumgardner and Paul O Frederickson. Icosahedral discretization of the two-sphere. *SIAM Journal on Numerical Analysis*, 22(6):1107–1115, 1985.

[3] Johannes Brandstetter, Rianne van den Berg, Max Welling, and Jayesh K Gupta. Clifford neural layers for pde modeling. *arXiv preprint arXiv:2209.04934*, 2022.

[4] Ashesh Chattopadhyay, Mustafa Mustafa, Pedram Hassanzadeh, and Karthik Kashinath. Deep spatial transformers for autoregressive data-driven forecasting of geophysical turbulence. In *Proceedings of the 10th International Conference on Climate Informatics*, pages 106–112, 2020.

[5] Daniel Salles Civitarese, Daniela Szwarcman, Bianca Zadrozny, and Campbell Watson. Extreme precipitation seasonal forecast using a transformer neural network. *arXiv preprint arXiv:2107.06846*, 2021.

[6] Gokhan Danabasoglu, J-F Lamarque, J Bacmeister, DA Bailey, AK DuVivier, Jim Edwards, LK Emmons, John Fasullo, R Garcia, Andrew Gettelman, et al. The community earth system model version 2 (cesm2). *Journal of Advances in Modeling Earth Systems*, 12(2):e2019MS001916, 2020.

[7] Henry F Diaz, Martin P Hoerling, and Jon K Eischeid. Enso variability, teleconnections and climate change. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 21(15):1845–1862, 2001.

[8] Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. *NeurIPS*, 2022.

[9] John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Efficient token mixing for transformers via adaptive fourier neural operators. In *International Conference on Learning Representations*, 2021.

[10] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

[11] Johann Dahm, Eddie Davis, Florian Deconinck, Oliver Elbert, Rhea George, Jeremy McGibbon, Tobias Wicky, Elynn Wu, Christopher Kung, Tal Ben-Nun, Lucas Harris, Linus Groner, and Oliver Fuhrer. Pace v0.1: A python-based performance-portable implementation of the fv3 dynamical core. *egusphere preprint https://doi.org/10.5194/egusphere-2022-943*, 2022.

[12] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, 2021.

[13] R Kubo. The fluctuation-dissipation theorem. *Reports on Progress in Physics*, 29(1):255, 1966.

[14] John Latham, Keith Bower, Tom Choularton, Hugh Coe, Paul Connolly, Gary Cooper, Tim Craft, Jack Foster, Alan Gadian, Lee Galbraith, et al. Marine cloud brightening. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 370(1974):4217–4262, 2012.

[15] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

[16] Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021.

[17] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *arXiv preprint arXiv:2106.04554*, 2021.

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[19] Björn Lütjens, Catherine H Crawford, Campbell D Watson, Christopher Hill, and Dava Newman. Multiscale neural operator: Learning fast and grid-independent pde solvers. *arXiv preprint arXiv:2207.11417*, 2022.

[20] Andrew J. Majda, Rafail Abramov, and Boris Gershgorin. High skill in low-frequency climate response through fluctuation dissipation theorems despite structural instability. *Proceedings of the National Academy of Sciences*, 107(2):581–586, January 2010.

[21] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[23] Binrong Wu, Lin Wang, and Yu-Rong Zeng. Interpretable wind speed prediction with multivariate time series and temporal fusion transformers. *Energy*, 252:123990, 2022.

[24] Feng Ye, Jie Hu, Tian-Qiang Huang, Li-Jun You, Bin Weng, and Jian-Yun Gao. Transformer for ei niño-southern oscillation prediction. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

[25] Mark D. Zelinka, Timothy A. Myers, Daniel T. McCoy, Stephen Po-Chedley, Peter M. Caldwell, Paulo Ceppi, Stephen A. Klein, and Karl E. Taylor. Causes of Higher Climate Sensitivity in CMIP6 Models. *Geophysical Research Letters*, 47(1), January 2020.

## Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [No] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

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]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [N/A]
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)? [No]
   (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)? [No]
   (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)? [No]
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? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (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]