# Towards Evaluating Robustness of EEG-based Sleep Trackers

**Priyanka Mary Mammen*** **Prashant Shenoy†**
University of Massachusetts Amherst

## Abstract

Recent advancements in mobile health sensing coupled with the availability of large datasets have given rise to large scale adoption of wearable sleep trackers. Wearable sleep trackers enable continuous measurement of sleep in home settings, of which EEG-based trackers are more reliable since they directly measure brain activity. However, EEG-based trackers face challenges in deployment settings when exposed to real-world data which tends to be noisy or out-of-distribution (OOD). In such situations, prediction models may become overconfident, leading to unreliable and inaccurate results. In this study, we explore various scenarios in the deployment settings and measure their impact on the prediction performance of a single-channel based EEG model.

## 1 Introduction

Good quality sleep is one of the key indicators of human well-being. Sleep disorders are becoming commonplace, contributing to many chronic health problems, including cardiovascular diseases, metabolic disorders, immune dysfunction, and mental health disorders. According to the CDC, about 50 to 70 million Americans chronically suffer from a sleep disorder or wakefulness, hindering their daily functioning and negatively affecting their wellbeing and longevity .

The rise in the adoption of commercial sleep trackers, along with advancements in wearable sensing and mobile computing techniques, show promise in longitudinal continuous monitoring in a naturalistic sleep environment when compared with intermittent clinical visits. Unlike trackers, which use surrogate signals such as heart rate, skin temperature, body movements etc, EEG directly measures electrical activity in the brain, which allows more accurate identification of sleep stages. .

Development of a reliable machine learning model for sleep stage classification in real-world settings is a non-trivial task. Sleep stage classification models in trackers are generally trained using datasets curated by clinical experts, with a good majority of data collected from healthy individuals in controlled settings. In reality, data collected in the wild is often noisy, incomplete or affected by sensor misplacement or physiological conditions of the individual. For instance, compared to polysomnography, electrodes used in head-worn EEG trackers often suffer from low SNR (Signal to Noise Ratio) due to motion-artifacts, skin and hair conditions . These differences or mismatch between training and deployment environments leads to distribution shifts (i.e. differences in probability distribution of training data and testing data) which might severely impact the model performance. Prior studies show that health models show degraded performance Wagh et al. [2022] when exposed to noisy data or when there is a change in user behavior/device behavior.

Prior works such as Melnik et al. [2017] tried to understand variance in EEG data quality in different systems, subjects, and sessions. Wagh et al. [2022] did some benchmarking efforts in other EEG-

*pmammen@cs.umass.edu
†shenoy@cs.umass.edu

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Learning from Time Series for Health.

related tasks, mostly instrumentation/sensor-related data-shifts using Monte-Carlo dropout as an uncertainty quantification method. In this paper, we explore various realistic datashift scenarios in the context of EEG-based sleep trackers and measure its impact on the sleep stage classification performance

## 2 Data-shifts in EEG trackers

In this section, we explore different data shift scenarios that the EEG signal might encounter after deployment. These transformations do not change the semantic meaning of the EEG signal. Note that we are focusing only on transformations that can be applied to a single channel EEG, as most of the wearable trackers have fewer channels compared to PSG.

- **Gaussian Noise** - In this transformation, we add a white Gaussian noise to the EEG signal with a standard deviation $(\delta)$ to the EEG signal. The addition of white noise hides the information contained in the high-frequency bands. This transformation is equivalent to a low-pass filter where $(\delta)$ plays the role of cut-off frequency. This transformation thus mimics the physiological and non-physiological artifacts introduced during data acquisition.
- **Frequency Shift** - This datashift helps to capture inter-subject variability. Locations of frequency bands with high power spectrum density are likely to vary between individuals and they have an important part in distinguishing different sleep stages. This transformation shift the frequency $(f)$ of the signal by a factor $\Delta f$ Rommel et al. [2021]. The presence of certain brain rhythms can be characterized by the power spectral density (PSD) of the EEG signals. where peaks occur at specific frequency ranges. Different subjects will have peaks at different locations for each sleep stage. This transformation helps to simulate the potential shifts arising from sensing device, subject characteristics or environmental factors.
- **FT Surrogate** - This transformation preserves the frequency band power ratios while it makes some changes to the signal representation in the time domain. This may cause high misclassification rates on sleep stages such as N2, which are strongly characterized by specific patterns in the time domain. This shift can help to identify the models reliance on specific time-domain patterns versus overall frequency content.
- **Rotations/Amplitude Shift** - Spatial domain transformations include shifts in electrode positions which occur during data collection. This transformation simulates the changes in contact quality, which results in a change in signal amplitude.

## 3 Experiments

We used the publicly available SleepPhysionet dataset Goldberger et al. [2000] to conduct the experiments. It consists of whole-night polysomnographic recordings from 78 healthy subjects. It has EEG data collected from 2 channels: Fpz-Cz and Pz-Oz. We preprocessed the dataset by applying a low-pass filter with a cutoff frequency of 30Hz. We selected the data from Fpz-Cz channel as it is found to be more robust and accurate based on prior works Tsinalis et al. [2016], Sors et al. [2018]

We adopted the CNN-based architecture proposed in Supratak et al. [2017]. The EEG model was trained on small segments of data (mini-batches). We adhered to the original hyperparameter configurations and trained the model on an NVIDIA A100 GPU. For dataset partitioning, we used a 70-20-10 split at the user level for training, validation, and testing, ensuring that no data from the same individual appeared in multiple sets. To address class imbalance, we applied a weighted cross-entropy loss function during training.

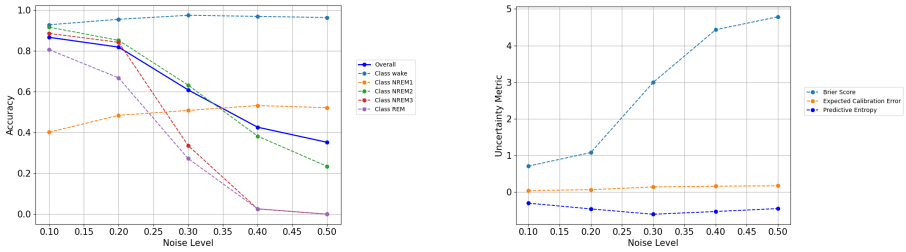
### 3.1 Impact of Data Shifts on Classification Performance

We analyze how different types of data shifts affect the performance of sleep-stage classification models trained on single-channel EEG signals. Building on the taxonomy of shift types discussed in Section 2, we systematically apply a range of distribution shifts,including frequency perturbations, amplitude scaling, additive noise, and temporal distortions to the test data. We then assess the resulting degradation in classification accuracy, precision, recall, and F1 score across different sleep stages.

Furthermore, we compare several uncertainty estimation metrics under these shifted conditions. *Brier Score* quantifies the mean squared difference between predicted probabilities and the true class labels, with values ranging from 0 (perfect confidence and correctness) to 1 (completely incorrect with high confidence). *Expected Calibration Error (ECE)* measures the average gap between predicted confidence and actual accuracy across bins, ranging from 0 (perfect calibration) to 1 (poor calibration). *Entropy* captures the uncertainty of the predicted probability distribution, ranging from 0 (confident prediction) to $\log K$ for a $K$-class problem (maximum uncertainty under uniform distribution) Below we summarize our observations for each data shift scenario:

1. **Gaussian Noise**

   Figure 1 illustrates the degradation in classification accuracy (left) and the behavior of uncertainty metrics (right) under increasing Gaussian noise. While the wake class remains robust, most stages, particularly REM and NREM3, show rapid performance drops. No consistent variability in entropy score, although Brier Score and ECE showed an increase in trend.

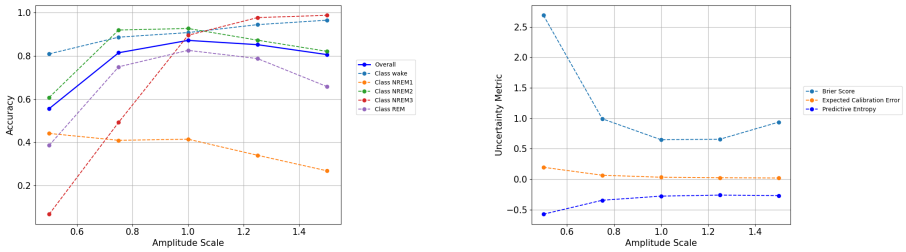



(a) Accuracy Metrics (b) Uncertainty Metrics

Figure 1: Model performance metrics under varying levels of gaussian noise.

2. **Amplitude Scaling**

   Amplitude scaling analysis (Figure 2) reveals that model accuracy peaks around the original signal amplitude (1.0), with marked performance drops at both low and high scaling. Notably, NREM3 benefits from increased amplitude, reflecting it's strong slow-wave components. However, uncertainty metrics show that both Brier Score and ECE are minimized at native scale, while predictive entropy remains low across all scales, indicating that the model remains confident even under mild scaling, but is most reliable near the original amplitude.



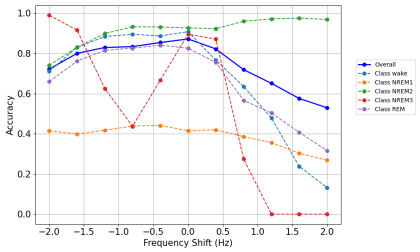

(a) Model Performance Metrics (b) Uncertainty Metrics

Figure 2: Model performance metrics under varying levels of amplitude scale.

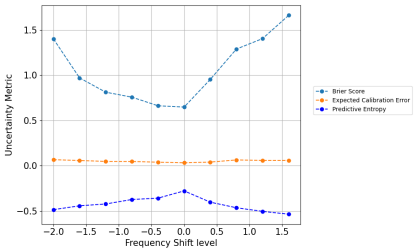
3. **Frequency Shift**

   We can observe that NREM2 accuracy shows an increasing trend with an increase in the value of frequency shift. Similarly, NREM3 accuracy shows a mixed trend and finally drops to zero at higher frequencies. Whereas, the remaining sleep stages show a consistent decreasing pattern. NREM2 and NREM3 are characterized by more certain brain rhythms NREM2 is characterized by sleep spindles, and N3 is characterized by slow waves in the delta band.

   Figure 3 shows classification accuracy (top) and uncertainty metrics (bottom) under increasing frequency shift perturbations. The model performs best near the original frequency




(a) Model Performance Metrics

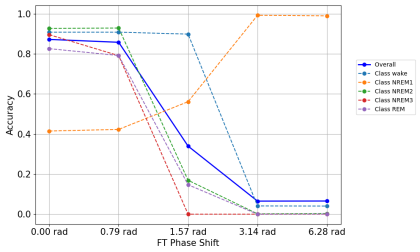



(b) Uncertainty Metrics

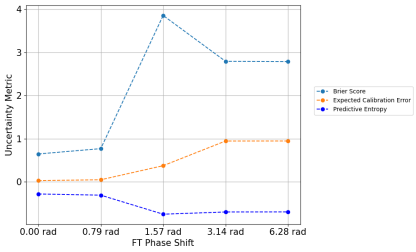
Figure 3: Model performance metrics under varying levels of frequency shift.

distribution (0 to -1.5 Hz), with a sharp accuracy drop observed under positive shifts. Notably, the model exhibits overconfident misclassifications as shown by decreasing predictive entropy and increasing Brier Score, reflecting a lack of uncertainty awareness under spectral distribution shifts.

4. **FT Surrogate**




(a) Model Performance Metrics




(b) Uncertainty Metrics

Figure 4: Model performance metrics under varying levels of frequency-phase shift.

Figure 4a shows the effect of applying increasing FT phase shifts to EEG signals on sleep stage classification accuracy. As the phase shift increases, the performance of the model degrades sharply across most sleep stages, with near-zero accuracy observed beyond 1.57 rad. This suggests a high dependence of the classifier on phase-sensitive spectral patterns. Interestingly, NREM1 shows an unexpected increase in accuracy, potentially indicating class confusion or misclassification bias under distortion.

Figure 4a shows the variation in uncertainty metrics under increasing FT phase shifts. The Brier Score and Expected Calibration Error rise sharply, indicating decreased predictive reliability and calibration under phase distortion. Interestingly, predictive entropy drops, highlighting the model's overconfidence in incorrect predictions, a sign of poor robustness to phase-based distribution shifts

## 4 Conclusions and Future Work

In this paper, we make an attempt to understand the impact of data shifts on sleep stage classification models using EEG data. Our analysis shows that while the baseline model achieves strong performance on in-distribution EEG data, it lacks robustness to signal distortions and provides unreliable uncertainty estimates under data shifts. We plan to extend this work to enhancing both robustness and uncertainty quantification of the classification model. We will also investigate whether incorporating additional EEG channels can further improve robustness of the models.

## 5 Acknowledgements

This work is supported in parts by NSF grants 1722792, 2211302, 2211888, 2213636, and 2105494.

## References


Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Dagmar Krefting, Christoph Jansen, Thomas Penzel, Fang Han, and Jan W Kantelhardt. Age and gender dependency of physiological networks in sleep. *Physiological measurement*, 38(5):959, 2017.

Hannah McCann, Giampaolo Pisano, and Leandro Beltrachini. Variation in reported human head tissue electrical conductivity values. *Brain topography*, 32:825–858, 2019.

Brice V McConnell, Eugene Kronberg, Peter D Teale, Stefan H Sillau, Grace M Fishback, Rini I Kaplan, Angela J Fought, A Ranjitha Dhanasekaran, Brian D Berman, Alberto R Ramos, et al. The aging slow wave: a shifting amalgam of distinct slow wave and spindle coupling subtypes define slow wave sleep across the human lifespan. *Sleep*, 44(10):zsab125, 2021.

Andrew Melnik, Petr Legkov, Krzysztof Izdebski, Silke M Kärcher, W David Hairston, Daniel P Ferris, and Peter König. Systems, subjects, sessions: To what extent do these factors influence eeg data? *Frontiers in human neuroscience*, 11:150, 2017.

Cédric Rommel, Thomas Moreau, Joseph Paillard, and Alexandre Gramfort. Cadda: Class-wise automatic differentiable data augmentation for eeg signals. *arXiv preprint arXiv:2106.13695*, 2021.

Arnaud Sors, Stéphane Bonnet, Sébastien Mirek, Laurent Vercueil, and Jean-François Payen. A convolutional neural network for sleep stage scoring from raw single-channel eeg. *Biomedical Signal Processing and Control*, 42:107–114, 2018.

Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE transactions on neural systems and rehabilitation engineering*, 25(11):1998–2008, 2017.

Orestis Tsinalis, Paul M Matthews, Yike Guo, and Stefanos Zafeiriou. Automatic sleep stage scoring with single-channel eeg using convolutional neural networks. *arXiv preprint arXiv:1610.01683*, 2016.

Neeraj Wagh, Jionghao Wei, Samarth Rawal, Brent M Berry, and Yogatheesan Varatharajah. Evaluating latent space robustness and uncertainty of eeg-ml models under realistic distribution shifts. *Advances in Neural Information Processing Systems*, 35:21142–21156, 2022.


## A Technical Appendices and Supplementary Material

### A.1 Sources of Data Variability in EEG-based Trackers

The accuracy of the sleep stage classification models relies heavily on the quality and distribution of input EEG signal. In this section, we try to understand the sources of various data shifts that can occur in EEG-based sleep models.

- User Demographics: Age is a critical factor that affects the activity of the brain. Neural connectivity decreases with age and results in lower amplitude EEG waves. This decline particularly causes discrepancies in identifying especially N3 stage sleep for older populations McConnell et al. [2021]. In addition to age, other demographic factors, such as gender and individual physiological differences, also affect sleep patterns Krefting et al. [2017]. These differences stem from physiological differences such as skull shape, cortical folding, tissue conductivity, and brain tissue shapes between individuals McCann et al. [2019], all of which impact neural activity propagation in the individuals. Therefore, due to these individual differences, sleep-stage classification will be difficult.

- Data Collection Ecosystem: Quality of an EEG signal is influenced by a multitude of factors such as hardware variability, environmental conditions, and sensor degradation. Motion is the major reason for signal corruption in wearable devices. Other factors such as placement of electrodes, quality of gel mains interference, electrical noise interference, and baseline wander of the electrodes can further make the signal noisy. While band-pass filters are employed to counter these, motion artifacts and electrical noise interference can still persist, posing challenges to the accurate sleep stage classification.
- Physiological Conditions: The health condition of an individual can impact the quality of the EEG signal. For instance, individuals with sleep disorders can exbhibit noisy artifacts which can affect the quality of sleep labels. Similarly, consumption of some drugs or medication can also cause disruptive sleep behavior. Also, certain pathologies can fog the patterns in a person's sleep.

## NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer:[Yes]

   Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We use a publicly available dataset to conduct the experiments. We also provide sufficient instructions to reproduce the main experimental results. we will provide the code repository upon acceptance.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.