Learning from Personal Preferences

Kelly Jiang
Northwestern University
kellyjiang2022@u.northwestern.edu

Berk Ustun
University of California, San Diego
berk@ucsd.edu

Jessica Hullman
Northwestern University
jhullman@northwestern.edu

Abstract

Machine learning practitioners frequently use majority vote to resolve disagreement in multi-annotator datasets. While this approach is natural in settings where a single ground truth label exists for each instance, it hides the presence of disagreement for subjective annotation tasks. In domains such as language modeling, information retrieval, and top-k recommendation, models must avoid suppressing minority views and express when the answer to a query is contentious. We propose personalized error metrics to formalize the requirement of strong performance across a heterogeneous user population. Following this framework, we develop an algorithm for training an ensemble of models, each specialized for a different segment of the population.

1 Introduction

Disagreement is often treated as mere noise by machine learning practitioners. Under the classic Dawid-Skene model [7], for example, each instance in the training set is assumed to have a single ground truth label; when users disagree on a label, it is viewed as a lack of reliability on the part of the individual labelers. The goal is then to calibrate estimates for the reliability of each user by maximum likelihood. The original paper by Dawid and Skene has inspired a host of related works, most notably Zhou et al. [20], which extends their method by allowing instances to be rated as more or less difficult.

Assuming that each instance has a single ground truth label may be natural in some settings like medical coding, where there is a ground truth physical condition that labels aim to capture. However, this assumption is damaging when it masks genuine disagreement. Suppose a user were to query an LLM chatbot with the prompt "Which candidate should I vote for in the next presidential election?" According to a recent opinion poll by Pew Research [11], American voters are becoming increasingly polarized. Would a partisan answer to this query do more harm than good? Such intractable division admits no simple solutions.

In value-laden settings such as search and language modeling, naive solutions —e.g. taking the majority vote or aggregating via a weighted sum —lossily compress away all evidence of disagreement, hiding a rich source of information from the downstream models. By contrast, we view disagreement

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

as a useful signal of user preferences in connection with a stream of recent work including Fleisig et al. [8], Davani et al. [6], and Park et al. [12]. See Uma et al. [17] for a detailed survey of prior work.

Our main contributions include:

1. We propose personalized error metrics to quantify aggregation error induced by using majority vote labels.
2. We present a framework for learning from disagreement. We cluster users by similarity to ensure that personalized risk is reduced across the user population.
3. We validate our method via a data imputation case study constructed from the DICES dataset (section 4).

2 Problem Statement

Our problem setting is a generalization of supervised learning which computes individual errors with respect to each user. Denote the space of instances by $\mathcal{X}$ and the space of labels by $\mathcal{Y}$. In conventional supervised learning, we may formalize the task by defining a probability measure $\mathbb{P}_{XY}$ over $\mathcal{X} \times \mathcal{Y}$ to represent the data generating process.

Then the goal is to learn a classifier h using a sample of pairs $\{\boldsymbol{x}_i, y_i\}_{i=1}^n \sim \mathbb{P}_{XY}$ such that the error $\mathbb{E}_{XY}[\ell(h(\boldsymbol{x}), y)]$ is minimized with respect to the loss function $\ell(\cdot,\cdot)$. We restrict our investigation to classification using the 0-1 loss, $\ell(\boldsymbol{x}, y) := \mathbb{1}[h(\boldsymbol{x}) \neq y]$. This is typically accomplished by empirical risk minimization, converting expectations to sample averages by minimizing $\frac{1}{n}\sum_{i=1}^n \ell(h(\boldsymbol{x}_i), y)$.

To account for personalization, we specify a set of m users. Associate each user $k \in [m]$ with a probability measure $\mathbb{P}_{XY^k}$, and obtain a sample of labels $\{\boldsymbol{x}_i^k, y^k\} \sim \mathbb{P}_{XY^k}$. We then define two optimization objectives that we aim to minimize. The first is utilitarian, minimizing the total expected error of the population of users:

$$\frac{1}{m}\sum_{k=1}^{m} \mathbb{E}_{XY^k}[\ell(h(\boldsymbol{x}), y^k)]$$

. The second is Rawlsian, and forces risk to be distributed as fairly as possible among the m users by minimizing the expected *worst case* error [16]:

$$\max_{k \in m} \mathbb{E}_{XY^k}[\ell(h(\boldsymbol{x}), y^k)]$$

.

3 Methodology

We now describe the information theoretic criteria that enables learning from disagreement. We define a distance metric and sketch its benefits in 3.1.

3.1 Distance Metric

We define a pairwise similarity metric between users based on mutual information [4]. While there are many ways to measure similarity between users, most prior work such as Sap et al. [14] and Wich et al. [19] identify clusters of users by identifying pairs of users with high agreement. This may lead to inflated similarity scores when classes are imbalanced, and fails to take advantage of patterns induced by systemic disagreement. To motivate our choice of distance metric, consider the following toy examples:

Example 1. *We begin with an example to illustrate that simply counting the number of agreements may not indicate a correlation between a pair of users. Let $\mathcal{Y} = \{0, 1\}$ and $m = 3$. Define the ground truth data generating process by: $y_i^0, y_i^1 \sim$ Bernoulli(0.1) and $y_i^2 = 1 - y_i^0$.*

Note that in expectation, users 0 and 1 have 82% agreement. This agreement is specious, however, since the two users are choosing their labels independently from each other —knowing the labels that user 0 supplies gives no information when attempting to predict Y^1. Also note that despite having 0% agreement by construction, user 0's labels can be inferred from user 1's and vice versa.

Example 2. *We continue letting $\mathcal{Y} = \{0, 1, 2\}$ with $m = 2$. Similar to example 1, we construct a data generating process where two users have 0% agreement: let $y_i^0 \sim$ Bernoulli(0.5). Define* $y_i^1 \sim \begin{cases} \text{Bernoulli}(0.5) + 1 & y_i^0 = 0 \\ U\{0, 0\} & y_i^0 = 1 \end{cases}$

By contrast to the previous example, knowing that $y_i^0 = 0$ still leaves us with 1 bit of uncertainty about the value of y_i^1.

Now consider using mutual information:

$$I(A, B) := \sum_{a \in \mathcal{A}, b \in \mathcal{B}} \mathbb{P}_{ab}(a, b) \log(\frac{\mathbb{P}_{AB}(a, b)}{\mathbb{P}_A(a)\mathbb{P}_B(b)})$$

in the place of similarity. Note that this measure naturally corrects for the chance agreement due to class imbalance seen in example 1; $I(A, B) = 0$ if and only if A and B are independent. In the subsequent section, we demonstrate a collaborative filtering-like algorithm based on the mutual information.

Following Rajski [13], we define $d(A, B) = 1 - \frac{I(A,B)}{H(A,B)}$, where $H(\cdot, \cdot)$ is the joint entropy,

$$H(A, B) := - \sum_{a \in \mathcal{A}, b \in \mathcal{B}} \mathbb{P}(a, b) \log(\mathbb{P}(a, b))$$

By normalizing distances within the range $[0, 1]$, we avoid the issue that long sequences will have higher mutual entropy with each other than short sequences by virtue of containing higher total entropy [9]. This situation occurs when some users contribute substantially more data than others.

3.2 Clustering Algorithm

We use the above distance metric in a hierarchical clustering algorithm, avoiding the need to make the number of clusters a hyperparameter. We leverage the hierarchical clustering algorithm described by Dasgupta [5] due to its theoretical guarantees of cluster quality. In this algorithm, each user is viewed as a vertex in a weighted graph, $G = (V, E, w)$ with edge weights between pairs of nodes determined by similarity (here defined as $1 - d$). The goal is to partition the nodes of the graph into two subsets, $(S, \overline{S})$ to minimize the sparsest cut objective:

$$\phi(S, \overline{S}) := \frac{E(S, \overline{S})}{\min(|S|, |\overline{S}|)}$$

$E(S, \overline{S})$ is the total weight of all edges with one vertex in S and the other in $\overline{S}$. This objective is a compromise between cutting the minimum weight and ensuring that the partition is balanced in size. After a cut is found, recurse on both subgraphs. While the sparsest cut problem is known to be NP-hard [15], we opted to use a spectral approximation [1][10] which has runtime log-linear in the number of edges, i.e. $\widetilde{O}(m^2)$.

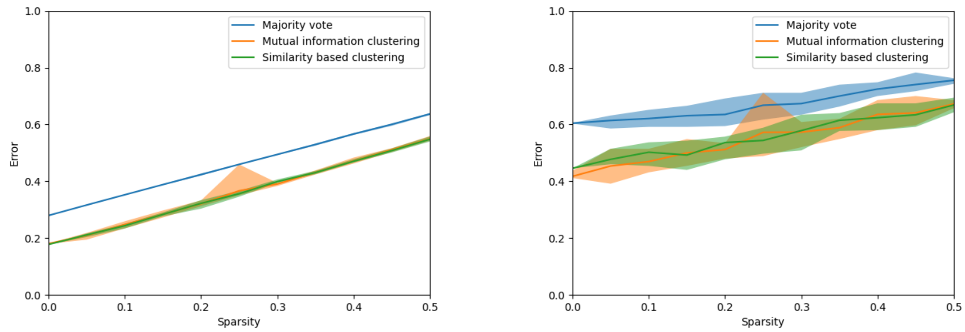
Our per-cluster label imputation strategy is as follows:

1. For each pair of users k, j, compute a maximum likelihood mapping $\pi_{k,j}$ which maps the labels selected by k to those selected by j. In example 1, $\pi_{0,2}(x) = 1 - x$.
2. If user k has failed to label an instance i, impute the missing label by taking majority vote among the labels induced by the mappings, $\{\pi_{1,k}(y_i^1), \dots \pi_{m,k}(y_i^m)\}$.

4 Experiments

4.1 DICES dataset

Setup The DICES dataset [3] was constructed to showcase the dangers associated with naively aggregating labels in a subjective annotation task. In this study, raters were asked to view conversation



(a) Utilitarian error

(b) Rawlsian error

Figure 1: Error of our method compared to the baseline majority vote. Orange and green lines show errors for two different distance metrics. Both figures show averages of 10 runs with error bands.

logs between human users and LLM chat bots and annotate them along five safety dimensions including bias, misinformation, and political partisanship. The dataset is designed to surface disagreement by employing a diverse rater pool and by assigning many raters to each instance.

We hypothesize that our method will lead to improved personalized performance across the survey population by mitigating the aggregation error incurred by majority vote. Note that due to the tightly controlled "lab" conditions that this study was conducted in, we avoid difficulties associated with deployment of a production system such as sampling bias and data drift. We test our hypothesis by dropping data at random according to a sparsity parameter then measuring imputation error.

Results and Discussion Our results are summarized in figure 1. The full data is shown in tabular form in appendix C. We compare the majority vote baseline to our method. We experimented with both the mutual information based metric and a purely similarity based clustering metric (L1 metric). Across all exposures, our method reduced the utilitarian error by an average of 10.0% using the L1 metric and 9.86% using the mutual information metric. The rawlsian error improved by 11.4% using the L1 metric and 12.1% using the mutual information metric. Since the performance of both metrics are similar, we do not make a recommendation about which to prefer. The choice is likely dataset dependent.

Our method succeeds in eliminating a significant amount of aggregation error while leaving room for algorithmic improvements e.g. replacing the spectral cut approximation described in section 3.2 by the more accurate ARV algorithm [2]. The high variance of the rawlsian error can be explained by the fact that the Dasgupta algorithm optimizes for a global (utilitarian) objective. A bottom-up algorithm may be more suitable for optimizing the rawlsian error.

We note that the optimal partitions chosen by the algorithm are highly granular: the most common cluster sizes were two and three. This begs the question of whether regularization is required to avoid overfitting. As noted by Vaughan [18], it may be worthwhile to group relatively dissimlar users together if this obtains a large number of new labels.

5 Concluding Remarks

We defined novel personalized error metrics that quantify the aggregation error incurred by using majority vote labels. These metrics formally justify the intuition that e.g. a 3-2 majority for an instance contains greater label uncertainty than a 5-0 majority. Personalized error for each user could in principle be minimized by deploying a model for each user, but the steep data requirement of this approach makes it impractical. We addressed this issue by proposing a collaborative filtering-like label imputation strategy. Our algorithm was able to achieve 14.9% error reduction of the utilitarian error and 12% reduction of the rawlsian error, avoiding much of the error incurred by majority vote aggregation.

Since our work builds upon supervised learning as a foundation, many of the usual caveats apply. In particular, we require that each pair of users have a significant number of instances labeled in common; to correctly calibrate the mutual information between the pair, we further require that the sub-sample they overlap on be unbiased. This issue can be avoided at dataset construction by carefully determining how instances should be assigned to labelers in advance. We leave the door open to future work improving upon the technical aspects our solution. Our primary aim has been to illustrate several of the algorithmic and statistical challenges associated with the personalized learning setting.

References

[1] Alon, Noga and Vitali D Milman. $\lambda 1$, isoperimetric inequalities for graphs, and superconcentrators. *Journal of Combinatorial Theory, Series B*, 38(1):73–88, 1985.

[2] Arora, Sanjeev, Satish Rao, and Umesh Vazirani. Expander flows, geometric embeddings and graph partitioning. *Journal of the ACM (JACM)*, 56(2):1–37, 2009.

[3] Aroyo, Lora, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. Dices dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36, 2024.

[4] Cover, Thomas M. and Joy A. Thomas. *Elements of information theory Thomas M. Cover; Joy A. Thomas*. John Wiley Sons, 2012.

[5] Dasgupta, Sanjoy. A cost function for similarity-based hierarchical clustering, 2015.

[6] Davani, Aida Mostafazadeh, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022.

[7] Dawid, Alexander Philip and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

[8] Fleisig, Eve, Rediet Abebe, and Dan Klein. When the majority is wrong: Modeling annotator disagreement for subjective tasks, 2024.

[9] Kraskov, Alexander, Harald Stögbauer, Ralph G Andrzejak, and Peter Grassberger. Hierarchical clustering using mutual information. *Europhysics Letters*, 70(2):278, 2005.

[10] Madry, Aleksander. Fast approximation algorithms for cut-based problems in undirected graphs, 2010. URL https://arxiv.org/abs/1008.1975.

[11] Nadeem, Reem. Americans' dismal views of the nation's politics, Sep 2023. URL https://www.pewresearch.org/politics/2023/09/19/americans-dismal-views-of-the-nations-politics/.

[12] Park, Chanwoo, Mingyang Liu, Kaiqing Zhang, and Asuman Ozdaglar. Principled rlhf from heterogeneous feedback via personalization and preference aggregation, 2024.

[13] Rajski, C. A metric space of discrete probability distributions. *Information and Control*, 4(4):371–377, 1961.

[14] Sap, Maarten, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*, 2021.

[15] Shmoys, David B. Cut problems and their application to divide-and-conquer. *Approximation algorithms for NP-hard problems*, pages 192–235, 1997.

[16] Stark, Oded, Marcin Jakubek, and Fryderyk Falniowski. Reconciling the rawlsian and the utilitarian approaches to the maximization of social welfare. *Economics Letters*, 122(3):439–444, 2014. ISSN 0165-1765. doi: https://doi.org/10.1016/j.econlet.2013.11.019. URL https://www.sciencedirect.com/science/article/pii/S0165176513005132.

[17] Uma, Alexandra N, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, 2021.

[18] Vaughan, Jennifer Wortman. *Learning from collective preferences, behavior, and beliefs*. PhD thesis, University of Pennsylvania, 2009.

[19] Wich, Maximilian, Hala Al Kuwatly, and Georg Groh. Investigating annotator bias with a graph-based approach. In *Proceedings of the fourth workshop on online abuse and harms*, pages 191–199, 2020.

[20] Zhou, Dengyong, Sumit Basu, Yi Mao, and John Platt. Learning from the wisdom of crowds by minimax entropy. *Advances in neural information processing systems*, 25, 2012.

Appendix

Table of Contents

A Notation

We provide a list of the notation used throughout the paper in Table 1.

Symbol	Definition
$\mathcal{X}$	Feature space
$\mathcal{Y}$	Label space
m	Number of users
n	Number of items
X	R.V. associated with instances
Y^k	R.V. associated with the k^{th} user's labels
$\mathbb{P}_{XY}$	Probability measure of the joint distribution of X, Y
$H(\cdot)$	Entropy of a R.V.
$H(\cdot, \cdot)$	Joint entropy of a pair of R.V.'s
$I(\cdot, \cdot)$	Mutual information between two R.V.'s
$U\{a, b\}$	The discrete uniform distribution on $[a, b]$
$\phi(\cdot, \cdot)$	The sparsest cut cost of a graph partition
$\pi_{k,j}$	A permutation $\mathcal{Y} \mapsto \mathcal{Y}$ which maps from user k's labeling to user j's

Table 1: Table of Notation

B Hardware specification

All experiments are run using an Ubuntu 23.04 VM with 12 vCPUs virtualized through KVM. The physical server uses an 11th Gen Intel Core i7-11370H CPU with 3.30GHz clock speed. The 12 GB of RAM provisioned to the VM never saturated due to the small data sizes used. A total of 110 trials ran using $\sim$4 hours of CPU time ($\sim$2 minutes per trial).

C Full results of experiment 4.1

All quantities truncated to three significant figures of precision.

Sparsity (%)	Utilitarian error rate (%)	Rawlsian error rate (%)
0	28.0	60.4
5	31.6	61.4
10	35.2	62.1
15	38.8	63.1
20	42.4	63.5
25	46.0	66.8
30	49.4	67.3
35	52.9	70.0
40	56.6	72.5
45	60.0	74.1
50	63.7	75.5

Table 2: Majority vote baseline

Sparsity (%)	Utilitarian error rate (%)	Rawlsian error rate (%)
0	18.1 (-9.9%)	41.7 (-18.7%)
5	20.9 (-10.7%)	45.4 (-16.0%)
10	24.8 (-10.5%)	46.9 (-15.1%)
15	28.4 (-10.4%)	50.0 (-13.1%)
20	32.1 (-10.3%)	51.2 (-12.3%)
25	36.7 (-9.30%)	57.2 (-9.54%)
30	39.1 (-10.4%)	57.3 (-10.1%)
35	43.1 (-9.83%)	58.9 (-11.1%)
40	47.2 (-9.46%)	63.5 (-8.91%)
45	51.1 (-8.96%)	64.1 (-9.97%)
50	54.9 (-8.75%)	67.2 (-8.31%)

Table 3: Our method (mutual information metric)

Sparsity (%)	Utilitarian error rate (%)	Rawlsian error rate (%)
0	17.8 (-10.1%)	44.6 (-15.9%)
5	21.1 (-10.5%)	47.7 (-13.7%)
10	24.4 (-10.9%)	50.2 (-11.9%)
15	28.3 (-10.5%)	49.2 (-13.9%)
20	32.2 (-10.1%)	53.6 (-9.89%)
25	35.6 (-10.3%)	54.4 (-12.4%)
30	39.9 (-9.55%)	57.8 (-9.51%)
35	43.1 (-9.89%)	61.5 (-8.54%)
40	47.0 (-9.59%)	62.3 (-10.1%)
45	50.8 (-9.24%)	63.4 (-10.7%)
50	54.8 (-8.89%)	66.9 (-8.69%)

Table 4: Our method (similarity based metric)

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Error metrics are defined in section 2. Training algorithm is defined at the end of section 3.2.

   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: Limitations are discussed in sections 4 and 5.

   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: This paper includes no proofs.

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: The dataset used in the experimental evaluation is open source. Section 3 supplies a detailed description of the algorithm. Section 4 describes the experimental conditions and evaluation metric.

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
  1. If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  2. If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  3. 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).
  4. 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: [No]

Justification: The source code and documentation are hosted at a private github repository and will be used in a pending conference submission. The repository will be made public after publication of the conference paper.

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: Experimental conditions are detailed in section 4.

   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: Experiment 4.1 was repeated 10 times. Error bars are shown in figure **??**.

   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: Compute resources and execution time are documented in appendix B.

   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: We do not foresee any potential for harm posed by this work.

   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: [Yes]

    Justification: We do not predict any negative social impacts arising from this work.

    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: We do not perceive a high risk for misuse of the methods described i nthis paper.

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: The dataset used for our experimental evaluation is properly credited and is open source.

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: We are not releasing new assets with this paper.

    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: The work presented in this paper did not involve human subjects.

    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: The work presented in this paper did not involve human subjects.

    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.