Are You One of Us? Introducing Mechanism Membership Inference

David S. Hippocampus*
Department of Computer Science
Cranberry-Lemon University
Pittsburgh, PA 15213
hippo@cs.cranberry-lemon.edu

Abstract

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Motivation

The motivation and scope of our paper is as follow:

(a) *What can representation alignment (RSA) tell us about mechanistic alignment?* Consider two models M_A, M_B. Assume that both the circuit and mechanism of M_A is known, but M_B is blackboxed. Let also also assume that there exists an oracle that reveals representational alignment of M_A, M_B. By intervening on M_A and querying this oracle, can we deduce the mechansms of M_B?
(b) *So what?* This may give us much better insight into **brain-LLM alignment**.
(c) *So what?* From a **mechanistic intepretability** perspective, such a procedure may allow us to perform subcircuit/sub-mechanism membership. For example, if M_A is a toy model that implements some mechanism and if M_B is a large model, then deducing membership through representation of alignment between M_A, M_B is much more efficient than reverse enginerring M_B.
(d) *So what?* From a **fairness** perspective, suppose M_B is a large sophisticated model that processes job applications. We want to make sure that M_B is not discriminating based on race/gender/etc. We can construct a toy model M_A (either through Weiss et al. [2021], Lindner et al. [2023] or Friedman et al. [2023]) that exhibits undesirable programs and probe for sub-mechanism membership with M_B.
(e) *So what?* From an **safety and alignment** perspective, similar to above we can rely on illuminated mechanisms Lee et al. [2024] to understand if language models are implementing alignment in a way that is desirable.
(f) *So what?* Checking membership inference is much more powerful than just checking through the potential outcome of the model's outputs. Since once we identify the mechanism we have essentially performed counterfactual inference for infinitely many examples.

Fundamentally, the problem that we are trying to solve is one of **constrained mechanismic membership inference**. We want to infer whether an mechanism is present as a sub-routine in another one.

This work establishes a **sufficient condition** for mechanism membership inference.

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

Preprint. Under review.

2 Background and Related Literature

Background on RASP Programs. The Restricted Access Sequence Programming (RASP) language is a functional programming model designed to capture the computational behavior of Transformer architdectures [Weiss et al., 2021]. RASP programs have shown use in mechanistic interpretability both as an effective benchmarking tool for faithfulness [Conmy et al., 2023, Hanna et al., 2024] and as a method to develop "inherently" interpretable language models [Friedman et al., 2023]. Another line of work uses it (and other similar methods) as a proof technique to reason about the Transformer architecture's generalizability on a host of tasks [Weiss et al., 2021, Merrill et al., 2022, Giannou et al., 2023]. In this paper, we focus on RASP's applications in interpretability.

RASP programs operate on two primary types of variables: *s-ops*, representing the input sequence, and *selectors*, corresponding to attention matrices. These variables are manipulated through two fundamental instructions: elementwise operations and select-aggregate. *Elementwise operations* simulate computations performed by a multilayer perceptron (MLP), while *select-aggregate* combines token-level operations, modeling the functionality of attention heads.

Every RASP program is equipped with two global variables `tokens` and `indices`, essentially primitive *s-ops*. `tokens` maps strings into their token representations:

```
token("code") = ["c", "o", "d", "e"]
indices("code") = [0, 1, 2, 3]
```

On the other hand, `indices` map n-length strings into their indices. That is, a list of $[0, 1, \ldots, n-1]$. Elementwise operations can be computed through composition. That is,

```
(3 * indices)("code") = [0, 3, 6, 9]
(sin(indices)) = [sin(0), sin(1), sin(2), sin(3)]
```

Tokens and their indices can also be mixed through *selection matrices* which are represented through the *s-op select*. This operations captures the mechanism of the QK-matrix. It takes as input two sequences K, Q, representing keys and queries respectively, and a Boolean predicate p and returns a matrix S of size $|K| \times |Q|$ such that $S_{ij} = p(K_j, Q_i)$. Then, the OV-circuit can be computed through the *select-aggregate* operation, which performs an averaging over an arbitrary sequence with respect to the aforementioned selection matrix. For example,

$$\texttt{aggregate}\left(\begin{bmatrix}1 & 0 & 0\\ 0 & 0 & 0\\ 1 & 1 & 0\end{bmatrix}, [10 \quad 20 \quad 30]\right) = [10015]\,.$$

The previous example is directly lifted from Lindner et al. [2023].

Compiling RASP Programs. The power of RASP programming lies in its ability to translate any RASP program into a Transformer, a process known as *compilation*. As described in Lindner et al. [2023], this involves a two-stage approach. First, a computational graph is constructed by tracing the *s-ops* in the program, identifying how these operations interact with and modify the residual stream. Elementwise operations are converted into MLP weights, and individual components are heuristically assigned to Transformer layers. For further details, we refer the reader to Lindner et al. [2023].

As observed by Lindner et al. [2023], this compilation through "translation" introduces inefficiencies. Specifically, the heuristic layer-assignment of RASP components results in Transformers that often contain more layers than they need to have. Moreover, since RASP enforces the use of categorical sequences and hard attention (we only allow Boolean predicates) it requires various *s-ops* to lie orthogonal to each other after embedding as Transformer weights. As a result, this leads to a much larger embedding dimension that is usually observed in actual Transformers [Elhage et al., 2022]. Thus, Lindner et al. [2023] proposes to compress this dimension through a learned projection matrix. The caveat is that this transformation largely not faithful to the original program (measured through cosine similarity of the outputs at individual layers). I don't really understand why we are measuring faithfulness like this. It seems contradictory to the motivation of doing this compression in the first place. That is, we can only achieve cosine similarity of 1 if all of the compressed dimensions are orthogonal in the output space, but that is impossible simply the virtue of performing this compression. Seems that we should be measuring faithfulness differently.

Friedman et al. [2023] takes a different approach, addressing the inherent difficulty of writing RASP programs. To overcome this challenge, the authors propose a method for directly learning RASP programs. This is achieved by constraining the space of learnable weights to those that compile into valid RASP programs, ensuring outputs with categorical variables and hard attention mechanisms. Optimizing over this constrained hypothesis class is performed through a continuous relaxation using the Gumbel distribution [Jang et al., 2017].

RASP Benchmarks. Thurnherr and Scheurer [2024] is a dataset of RASP programs that have been generated by GPT-4. It contains 121 RASP programs. Gupta et al. [2024] provides 86 RASP programs and compiled Transformers. The compiled Transformers are claimed to be more realistic than Tracr compiled ones as instead of performing compression using a linear projection, they leverage *strict interchange intervention training* essentially aligning the intervention effects of the compressed and uncompressed model. This is similar in vein to many existing techniques on causal abstraction Otsuka and Saigo [2022], Zennaro [2022], Massidda et al. [2023].

3 Mechanism Membership Inference

We can frame this problem as one of determining causal abstraction or equivalence under constraints. A important criteria for determine causal equivalence is the consistency of interventions across the two models being compared. This was originally formulated by Verma and Pearl [2022] and stated intuitively as

> *Two causal models are equivalence if there is no experiment which could distinguish one from the other.*

Since this work, there have been several which take different approaches to causal equivalence both in terms of formalism and the actual identification algorithm [Beckers and Halpern, 2019, Otsuka and Saigo, 2022, Zennaro, 2022, Massidda et al., 2023]. These are reviewed in detail in the related literature section. However, all of them have been faithful to the definition given by Verma and Pearl [2022]. Herein, we take inspiration from these approaches as well as the recent advances in mechanistic interpretability to define a method for determining mechanism membership.

In all of the subsequent claims, we assume that model A and B denoted as m_A, m_B are neural networks. Moreover, we also assume that there exists a surjective correspondance between the computational graph of a neural network and the set of causal graphs. Let us also assume that m_A is the whitebox model while m_B is the blackbox one.

3.1 Representation Similarity Analysis, Its Success and Potential Pitfalls

3.2 Component-Level Intervention

Claim 1. *Intervention through noising on a single component (node or edge) in m_A results in decreased representation alignment.*

Claim 2. *Intervention through noising on a single component not in m_A results in no change in representational alignment.*

3.3 Network-Level Intervention

Claim 3. *Intervention through noising on a subnetwork of m_A results in decreased representation alignment.*

Claim 4. *Intervention through noising on a subnetwork that is not in m_A results in no change in representation alignment.*

3.4 System-Level Intervention

Claim 5. *Say m_A was discovered by patching with distribution P, then if we instead patch with distribution Q then the change in representation alignment is proportional to the transport distance between P and Q.*

4 Experimental Methods

4.1 Elliciting different mechanisms from the same task

To evaluate our methods, we need a way to verifiably ellicit different mechanisms on the same task. Let us first fix some task. Then, we proceed with the following steps:

1. Using Friedman et al. [2023], we learn several different explicit Transformer programs (source of randomness). We can check that they are different by looking explicitly at the Transformer programs.
2. Using Gupta et al. [2024] and Geiger et al. [2024] to get different mechanistic realizations of this abstract Transformer program.

References

S. Beckers and J. Y. Halpern. Abstracting causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2678–2685, 2019. Issue: 01.

A. Conmy, A. Parker-Mavor N., A. Lynch, S. Heimersheim, and A. Alonso-Garriga. Towards Automated Circuit Discovery for Mechanistic Interpretability. In *Thirty-Seventh Conference on Neural Information Processing Systems*, Oct. 2023. URL https://arxiv.org/abs/2304.14997.

N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, and others. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

D. Friedman, A. Wettig, and D. Chen. Learning Transformer Programs. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Pe9WxkN8Ff.

A. Geiger, Z. Wu, C. Potts, T. Icard, and N. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In F. Locatello and V. Didelez, editors, *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 160–187. PMLR, 01–03 Apr 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

A. Giannou, S. Rajput, J.-Y. Sohn, K. Lee, J. D. Lee, and D. Papailiopoulos. Looped Transformers as Programmable Computers. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 11398–11442. PMLR, July 2023. URL https://proceedings.mlr.press/v202/giannou23a.html.

R. Gupta, I. Arcuschin, T. Kwa, and A. Garriga-Alonso. Interpbench: Semi-synthetic transformers for evaluating mechanistic interpretability techniques. *arXiv preprint arXiv:2407.14494*, 2024.

M. Hanna, S. Pezzelle, and Y. Belinkov. Have Faith in Faithfulness: Going Beyond Circuit Overlap When Finding Model Mechanisms. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=TZ0CCGDcuT.

E. Jang, S. Gu, and B. Poole. Categorical Reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rkE3y85ee.

A. Lee, X. Bai, I. Pres, M. Wattenberg, J. K. Kummerfeld, and R. Mihalcea. A Mechanistic Understanding of Alignment Algorithms: A Case Study on DPO and Toxicity. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=dBqHGZPGZI.

D. Lindner, J. Kramar, S. Farquhar, M. Rahtz, T. McGrath, and V. Mikulik. Tracr: Compiled Transformers as a Laboratory for Interpretability. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=tbbId8u7nP.

R. Massidda, A. Geiger, T. Icard, and D. Bacciu. Causal Abstraction with Soft Interventions. In M. van der Schaar, C. Zhang, and D. Janzing, editors, *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pages 68–87. PMLR, Apr. 2023. URL https://proceedings.mlr.press/v213/massidda23a.html.

W. Merrill, A. Sabharwal, and N. A. Smith. Saturated Transformers are Constant-Depth Threshold Circuits. *Transactions of the Association for Computational Linguistics*, 10:843–856, Aug. 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00493. URL https://doi.org/10.1162/tacl_a_00493. _eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00493/2038506/tacl_a_00493.pdf.

J. Otsuka and H. Saigo. On the Equivalence of Causal Models: A Category-Theoretic Approach. In B. Schölkopf, C. Uhler, and K. Zhang, editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 634–646. PMLR, Apr. 2022. URL https://proceedings.mlr.press/v177/otsuka22a.html.

H. Thurnherr and J. Scheurer. Tracrbench: Generating interpretability testbeds with large language models. *arXiv preprint arXiv:2409.13714*, 2024.

T. Verma and J. Pearl. *Equivalence and Synthesis of Causal Models*, page 221–236. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022. ISBN 9781450395861. URL https://doi.org/10.1145/3501714.3501732.

G. Weiss, Y. Goldberg, and E. Yahav. Thinking Like Transformers. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11080–11090. PMLR, July 2021. URL https://proceedings.mlr.press/v139/weiss21a.html.

F. M. Zennaro. Abstraction between structural causal models: A review of definitions and properties. *arXiv preprint arXiv:2207.08603*, 2022.

A Appendix / supplemental material

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: **[TODO]**

   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: **[TODO]**

   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 mechanism 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 mechanisms 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: **[TODO]**

   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: **[TODO]**

   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 mechanism, the paper should make it clear how to reproduce that mechanism.
  (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: **[TODO]**

   Justification: **[TODO]**

   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: **[TODO]**

   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: **[TODO]**

   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: **[TODO]**

   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: **[TODO]**

   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: **[TODO]**

    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, algorithms for monitoring misuse, algorithms 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: **[TODO]**

    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: **[TODO]**

    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: **[TODO]**

    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: **[TODO]**

    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: **[TODO]**

    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.