Examining Data Compartmentalization for AI Governance

Nicole Mitchell
Google Research
San Francisco, CA, USA
nicolemitchell@google.com

Eleni Triantafillou
Google DeepMind
London, UK
etriantafillou@google.com

Peter Kairouz
Google Research
Seattle, WA, USA
kairouz@google.com

Abstract

The fusing of a vast corpus of data into model parameters poses a challenge for AI governance, particularly with regards to concerns over the appropriate use of specific examples. We investigate how partitioning data into semantically meaningful groups may allow for training and serving models with finer-grained control over subsets of data. Data compartmentalization can help isolate data groupings with differing levels of risk, permitted usages and expiry dates, and may provide a path towards data attribution. We propose data compartmentalization as a unifying framework across a number of existing technical approaches, and present hypotheses and open questions around the suitability of these approaches for addressing policy concerns related to AI governance.

1 Introduction

Most ML pipelines do not explicitly exploit any structure or hierarchy of training data – all sources are mixed and consumed by training algorithms that are agnostic to their structure. As a result, information from all data sources is typically fused in the model parameters. This poses a challenge for AI governance, as legal and policy restrictions may not apply uniformly to the entire training data corpus. Though monolithic training is by and large the status quo, there is reason to question the merits of this approach on account of both the desire for finer-grained control over data as well as the possible performance enhancements that might come from leveraging data structure.

Non-uniform data requirements may stem from the dynamic nature of data and context in which a model is deployed (e.g., availability, relevance, and licensing), or the inherent risk of some subsets of data and their influence on model capabilities (e.g., privacy, bias, and harms). Because data usage constraints can be time-, place-, and context-dependent, there is a need for training and/or serving models in a way that is aware of and respects these dependencies (see fig. 1). In cases where restrictions on data usage can be met by simply updating the training data corpus, a conservative approach is to retrain the model, even if the majority of data is unchanged. While foolproof, retraining billion-parameter models from scratch is costly, inefficient, and impractical.

The need for effective and compliant approaches that offer non-uniform treatment of data has made relevant ML techniques that leverage *compartmentalized data*: data which is partitioned into semantically meaningful groups. We discuss means of compartmentalizing data for measuring, limiting or isolating the influence of subset(s) of data on the model, and characteristics that describe the corresponding structured data and task settings. While the scope of our work intersects with

2nd Workshop on Regulatable ML at NeurIPS 2024.

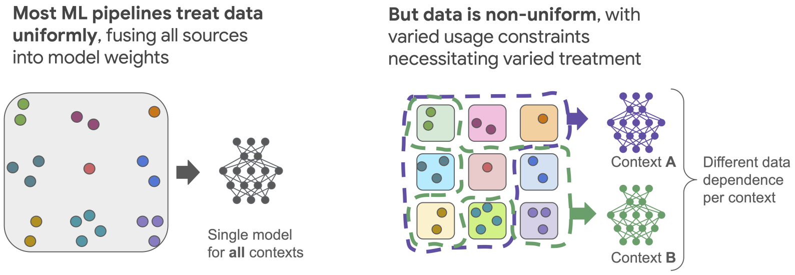
Most ML pipelines treat data uniformly, fusing all sources into model weights
Single model for **all** contexts
But data is non-uniform, with varied usage constraints necessitating varied treatment
Context **A**
Context **B**
Different data dependence per context

Figure 1: Data compartmentalization can allow for training and serving models that are context-aware.

the ideas of “model disgorgement” (Achille et al., 2024), we consider settings in which data can be explicitly grouped according to some aim and consider motivations beyond strictly removing the effects of data. We unify a number of existing techniques at the data, modeling, algorithmic, and inference levels under the framework of facilitating data compartmentalization. Despite their varied motivations and settings, all offer mechanisms for providing finer-grained control over subsets of data. We examine their effectiveness, practicality, and relevance to governance, and present open questions to better align policy motivations with technical approaches and inform future work.

2 Opportunities for AI Governance

Growing interest in better controlling large models has spurred research and led to voluntary commitments and nascent regulatory frameworks (Bommasani et al., 2022; Shevlane et al., 2023). Some motivations stem from practical constraints on data access (e.g., regulatory and licensing compliance), while others relate to risks of AI (e.g., bias, harms, and privacy). By strategically partitioning and managing data within AI systems, practitioners may be better equipped to align their models with overarching principles of responsible development and deployment (UK Department for Science, Innovation and Technology, 2024).

Enhancing traceability of model outputs. Attributing model outputs to the sources that were most influential is needed for interpretability, grounding, factuality, and mitigating harms. Data compartmentalization can make it easier to identify, isolate, and address subsets of the data that are found to be erroneous or problematic. When paired with influence functions (Koh et al., 2019), data compartmentalization may provide a path towards credit assignment.

Allowing efficient data deletion. When subsets of data have been identified as problematic (either due to explicit labeling, or as a result of measuring influence), one may want to remove this data from the model. AI model disgorgement Approaches that enable data compartmentalization may support more efficient deletion from trained models, compared to naively retraining a monolithic model from scratch, which may help facilitate addressing “right to be forgotten” requests under the European Union’s General Data Protection Regulation (European Parliament & Council of the European Union, 2016).

Enabling domain-specific models for regulatory compliance. Compartmentalizing sensitive domain-specific data may facilitate compliance with regulations on model use in particular contexts. To comply with securities regulations, a model trained on financial data could exclude insider information when used for investment recommendations. This would prevent the misuse of privileged information while still allowing for the use of other relevant data for analysis.

Facilitating compliance with licensing terms. Maintaining data source separability will allow for using each source according to its associated license, rather than using the most restrictive terms among all data sources in the mix. Though efforts to attribute licenses to data are underway (Longpre et al., 2023), their feasibility is uncertain given the evolving nature of licensing terms and data interdependencies. Nevertheless, on principle, data compartmentalization strategies may serve as a promising tool for addressing copyright concerns, providing an avenue for further exploration.

Fostering collaborative model development. Organizations could contribute to a joint model without revealing their data, by training separate modules that are combined only at inference time based on access policies. This could enable extensible models trained on data from multiple organizations in a privacy-preserving way that respects requirements on data locality (Rieke et al., 2020).

3 Compartmentalizing Data

Despite how standard ML pipelines treat all data uniformly, in practice data often has meaningful structure. This structure may occur naturally in the data, or may be imposed to yield groups that correspond to subsets of data with uniform usage requirements or qualities of interest. Using relevant existing metadata or generated annotations to specify groupings and leveraging this structure throughout training can help provide traceability of corresponding subsets of data.

There is a long history of embedding structure in data storage systems to specify relations and constraints. Most database management systems are designed to store data in an organized way that preserves relational, hierarchical, or network structure between examples in the database. This makes possible storing relationships between entities, compartmentalizing data, and controlling information flow (Robling Denning, 1982). Historically, access-control lists (ACLs) have been used in computer security to limit data access to particular users according to policy requirements (Daley & Neumann, 1965) and ensure non-interference where there should be no leakage of information between entities (Goguen & Meseguer, 1982).

Specifying appropriate groupings. Compartmentalizing data is important for several reasons, ranging from facilitating traceability or limiting the influence of different subsets of data on the model; efficiently coping with changes in the data distribution due to updated access or relevance; enabling diverse treatment of different groups of data for compliance with licensing terms or other restrictions; to name a few. However, enjoying these benefits from compartmentalization is only possible if we have compartmentalized data according to the right criteria. In this section, we discuss considerations for defining these compartments.

Using natural structure. Structure can arise naturally in data, yielding an inherent partitioning that may be relevant for addressing the concern of interest. Each group might refer to the data owner (e.g., an individual or an institution). This ownership structure might even correspond to physical placement of data across distributed hardware, either on-device (e.g., mobile phones) or on-premise (e.g., hospitals in a network). Groupings might be made according to content creators (e.g., by artist), which can be used for attribution of examples. The source (e.g., a particular text) of each example can also yield groupings, relevant for scenarios in which fluctuation in availability of some source might be expected (e.g., due to opt-out or license terms specifying appropriate use). Groupings may also correspond, more generally, to consistent usage constraints (e.g., as dictated by access policies or licensing). Data structure might be hierarchical, with nested groupings (e.g., categorically grouping sources by content type).

Imposing structure. While inherent structure in data can yield natural partitions, groupings can also be imposed. This is relevant when the metadata attributing each example to a specific case of concern is not given. In such a scenario, structure can be imposed by inferring the groupings of data and annotating them accordingly. Examples include clustering by topic, subject, domain, attribute, or concept. Structure might also be imposed if some artificial subset of the data is known to be risky or subject to change.

Principle characteristics. At a higher level, irrespective of the specific attribute(s) data is grouped upon, the result of data compartmentalization can be described by a number of principle characteristics that capture the statistical properties of the groupings. These characteristics help specify the particular partitioned data setting and inform what ML techniques for compartmentalized data are appropriate.

• *Granularity* specifies how large the data groupings are with respect to the size of the dataset: has the entire corpus been partitioned finely (into groups of few examples) or coarsely (into large groups)?

• *Specificity* indicates whether data groupings can be made completely with clear boundaries segmenting each group: is there a discrete mapping of group to all representative examples, or might there be some uncertainty as to whether groupings are complete?

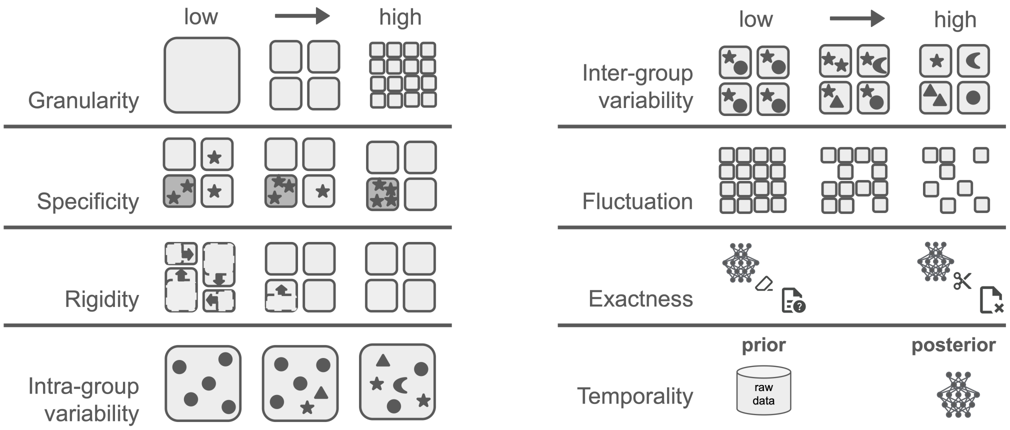

Figure 2: Visualizing characteristics of a compartmentalized data setting.

• *Rigidity* refers to the extent to which the groupings are static: how fixed are the partitions? Might there be a need to vary these groupings over time?

• *Intra-group variability* captures the degree of heterogeneity across examples within a group: how similar is the data within a group?

• *Inter-group variability* considers the degree of heterogeneity across groups: how similar are the distributions of data in each group?

• *Fluctuation* captures the frequency of change to any group's inclusion or exclusion: how often might usage concerns arise about a particular data grouping?

• *Exactness* specifies the strictness with which data groups must be able to be isolated from the model: is it required that each group be fully dissociable from model parameters or might approximate measures that limit the influence of groups suffice?

• *Temporality* specifies when the data groupings of interest are known relative to the overall point in the ML pipeline: are groupings known prior or posterior to training?

Figure 2 showcases the range of values each of these principle characteristics can take. The value of each of these data compartmentalization characteristics emerges from careful partitioning of data aligned with specific concerns of the problem setting. We note that this is not a complete list of characteristics, but we highlight these as some key factors that may determine the success of different ML approaches on compartmentalized data.

4 Strategies for Compartmentalized Data

Various existing techniques across the ML pipeline can be seen as facilitating the use of compartmentalized data. These strategies may be help tease apart model dependencies from data dependencies to address AI governance concerns.

Model architectures. Model architectures that take into account data compartmentalization tend to be *modular*, that is, composed of specialized sub-networks, each responsible for a specific subtask or functionality; see (Pfeiffer et al., 2024) for a survey. These modules can be trained, fine-tuned, or even swapped out independently without affecting the entire model. A simple example is a flat Mixture of Experts architecture, where each expert is trained on a different group of data (Jacobs et al., 1991).

Tiwari et al. (2023) present a case for modular architectures allowing for "Information Flow Control" in machine learning, where particular modules can be included or excluded depending on constraints on downstream data usage.

Training algorithms. *Federated learning* (FL) limits data sharing by training across siloed data in a distributed manner (McMahan et al., 2017). Prototypical FL algorithms bake information across clients into a shared model parametrically through iterative averaging (Reddi et al., 2020). Such an approach is not suitable for traceability or exclusion of some data source. However, the data placement aspect of FL allows for data owners to keep raw data on premise and (in theory) opt-in to participating in training.

Frequently used with FL, *group-level differential privacy* (DP) extends DP to *groups* of examples, where a grouping refers to all examples attributed to an individual, institution, domain, or source Dwork et al. (2006). By operating on compartmentalized data, group-level DP bounds the influence of any group on the model, treating all groups as sensitive.

By contrast, *machine unlearning* (MU) removes (the influence of) a specified subset of training data (the "forget set") from models (Nguyen et al., 2022). An unlearning method can be either *exact*, if it entirely eliminates the influence of the requested training data, or *approximate*, leading to imperfect removal, in exchange for increased efficiency or model utility. Exact unlearning is done via re-training (portions of) a model (Bourtoule et al., 2021), often a modular architecture. A plethora of diverse training algorithms have been proposed for approximate unlearning (Golatkar et al., 2020; Graves et al., 2021; Thudi et al., 2022; Liu et al., 2024; Izzo et al., 2021; Kurmanji et al., 2024; Fan et al., 2023), but designing robust and principled evaluation methods for approximate unlearning is an open problem.

Retrieval and inference. Non-parametric access of data sources through *retrieval-augmented generation* (RAG) allows for maintaining full separability of those sources from model weights (Lewis et al., 2020). The approach presented in SILO (Min et al., 2023) provides such an example. Recently, there has been work advocating for retrieval augmentation in FL (Muhamed et al., 2024), where clients maintain private data stores accessed only at inference. The merging of FL's ownership-based data partitioning with retrieval yields an approach for owner-based selection of data sources divorced from shared model weights.

Note that the strategies we cite do not comprise a complete list, but demonstrate a range of varied approaches across the ML pipeline that operate on compartmentalized data.

5 Hypotheses on Settings and Suitability

We have reviewed several strategies that leverage data compartmentalization and offer finer-grained control of data. None of them is a panacea; they have different strengths and weaknesses that make them suitable to different settings.

The success of leveraging data compartmentalization and associated ML techniques at addressing a particular AI governance concern hinges upon several interdependent factors: 1) the underlying distribution of the training data (i.e., whether it is "compartmentalizable" in a useful way for a particular goal); 2) the choice of data compartmentalization (i.e., whether a partitioning can be defined that fully matches the concerns of the setting); and 3) the ML technique used (i.e., whether a technique or composition of techniques matches the compartmentalization characteristics and meets the aims).

An example where appropriate data compartmentalization paired with a suitable strategy successfully aids in adhering to licensing restrictions is the following setting: data is partitioned by source, a modular architecture where each data source trains a separate sub-network is used, allowing for excluding a data source when its license expires. However, the solution is not always so clear, and in practice there are trade-offs: a particular choice of compartmentalization made for some priority might facilitate one application at the expense of others.

We present hypotheses on how data should be compartmentalized in alignment with various governance concerns, and the characteristics of compartmentalized data that each ML technique is best suited to operate on. We note that additional considerations dependent on the specific data setting should be taken into account when defining compartments, and techniques should be chosen according to trade-offs in addressing additional problem objectives beyond suitability to the compartmentalization characteristics and governance aims (e.g., preserving utility, or maximizing efficiency).

Mapping AI governance concerns to characteristics. To be effective, data compartmentalization should be done in accordance with the objective of the motivating governance concern, so that the subsets of data of interest are grouped. Acknowledging that there may be competing priorities that suggest alternate groupings of data to protect as well as problem-specific factors, here we consider the likely characteristics (defined in section 3) of data compartmentalized for the individual governance concerns introduced in section 2.

• *Enhancing traceability of model outputs:* To ablate the impact of some subset of the data on model capabilities, groupings should be made according to the attribute of interest. Inherently, these groupings may not have high specificity or rigidity, given the difficulty of drawing boundaries around all examples that influence the model in a particular way. Ideally there should be some cohesion to the group of interest (low intra-group variability) and distinction from the remaining training data (high inter-group variability).

• *Allowing efficient data deletion:* The partitioning of data to remove a group at the request of a particular owner or data provider should be specific, rigid and defined prior to training. To remove a concept found to be problematic, the compartmentalization will have limited specificity, as these groupings are inferred. Granularity might vary from a single example to a large portion of the corpus. Depending on the objective, deletion might need to be exact, or approximate removal might suffice.

• *Enabling domain-specific models for regulatory compliance:* Compartmentalizing data for use in training domain-specific models for regulatory compliance should yield groupings that are coarse and rigid. Regulations on context-dependent data likely call for specific groupings and require that groups be exactly separable.

• *Facilitating compliance with licensing terms:* License and contract constraints yield coarse, specific and rigid groupings. These groups are not expected to fluctuate with high frequency. In terms of temporality, licensing terms of data are likely known prior to training, but the usage of the model might not be known, which may preclude the use of some data at inference time. Additionally, data licenses may evolve over time. Licensing terms likely necessitate exact means of separating groups.

• *Fostering collaborative model development:* Data compartmentalization for collaborative model development is naturally defined by data ownership among organizations participating. This partitioning is coarse, specific and rigid. Likely there is low intra-group variability and high inter-group variability. The groupings are known prior to training. The exactness with which these groups need to be separable from model parameters is variable and dependent on the concerns of the organizations participating.

Defining and characterizing the partitioning for each problem setting is the first step towards examining what strategies might meet the motivating aims and concerns.

Formulating hypotheses on where strategies apply. Several considerations influence the suitability of each data compartmentalization strategy for different applications, including trade-offs in computational complexity, model performance, and application-specific priorities. Here, we focus on key characteristics we identify in section 3, posing hypotheses on the suitability of each strategy to data partitioning settings that we invite research to investigate:

• *Modularity* may be most applicable to settings in which data compartmentalization is coarse, rigid and specific. Modular architectures support the training of separate modules on groupings known prior to training, as well as the addition of new groups of data through subsequent training of separate modules. It is not applicable, however, if groupings in the pre-training corpus are known posterior to training. This strict compartmentalization lends itself well to addressing frequent fluctuation in inclusion or exclusion of modules to meet deletion needs and conditional usage.

If there is significant variance in the granularity of some groups (e.g., one group has very few data points assigned to it but others have more) and sufficiently low inter-group variability (i.e., groups are similar to one another), then modular architectures may be a poor choice. Utility may suffer as some subnetworks will be trained with insufficient data. By contrast, a monolithic model could better benefit from positive transfer, allowing data-rich components to influence and aid in learning data-poor components.

Modularity can aid in achieving higher utility (while readily enabling deletion, or conditioning based on relevance for new tasks) when there is high inter-group variability. This is because there may be interference issues associated with training all parameter weights on highly-heterogeneous data.

• *DP* is well-suited to address risks associated with specific, rigid groupings known prior to training, in problem settings where approximate separability suffices. DP uniformly bounds the risk of all groups at the expense of utility. For the same privacy guarantee, a larger group size yields worse utility, making DP suitable for relatively fine granularity partitions (e.g., example-level or user-level) (Ponomareva et al., 2023). Such an approach is tolerant of frequent inclusion or exclusion of groupings (done approximately) as each group has bounded limited influence on a model trained with group-level DP. Critically, DP implicitly assumes that these groupings are made with high specificity and strict boundaries, such that each group maps to a specific piece of private information. This is a challenging, if not prohibitive, requirement for natural language data where private information may occur repeatedly and boundaries are hard to define (Brown et al., 2022).

• *FL* operates on specific, rigid groups known prior to training. Granularity varies from relatively fine (e.g., cross-device) to relatively coarse (e.g., cross-silo). FL typically assumes high inter-group variability and low intra-group variability. While FL offers a means of ownership-based participation, prototypical FL algorithms do not offer any exact separability of group(s) from the model resulting from compartmentalized training. If instead of iterative averaging, models are trained separately on owner data then ensembled or souped, this would make possible more exact separability.

• *MU* may be poised to address specific and limited data groupings that are not rigid, frequently fluctuate, and may be defined posterior to training. Approximate MU yields approximate separation of the group of data of interest (the "forget set") from the model parameters. Preliminary results show that intra- and inter- variability between the group that is requested to be removed and the rest of the training data affects the success of approximate unlearning methods. Several approaches struggle to remove forget sets that are "more similar" to the rest of the dataset (Zhao et al., 2024).

• *RAG* is similarly suitable for specific and limited data groupings that frequently fluctuate, but these groupings must be rigid and should largely be defined prior to training. RAG is not a good solution if there is low rigidity, because if the groupings change, the desired partition of which groups can affect parameters (versus which groups of data can only be retrieved through inference) will also change, potentially necessitating retraining the model, which is expensive. For the same reason, RAG requires groupings known prior to training. However, additional data for inference-time retrieval can be added posterior to training.

6 Open Questions

A number of open questions remain surrounding the appropriate use of data compartmentalization and associated strategies to address the needs of AI governance.

Technical considerations.

• *Suitability*: What is a robust set of principles to inform the choice of strategy for a particular application? Research is needed to investigate the hypotheses we make, and compile criteria for assessing the relevance of each strategy.

• *Composability*: How can these techniques be effectively combined to achieve multiple goals simultaneously? For instance, can FL be used in conjunction with MU to remove data from specific clients while preserving the model?

• *Evaluation*: How can we rigorously evaluate the effectiveness of these techniques, particularly for MU, where defining and measuring "forgetting" is a challenge?

• *Limitations*: To what extent might compartmentalizing data pose challenges for management and transparency of the entire training corpus? For example, multiple data owners and/or model owners may lead to questions of liability and increased security risk.

Legal and policy alignment.

• *Considering alternatives*: How do data compartmentalization techniques compare to alternative strategies (e.g., careful data curation, output filtering, representation engineering) in terms of optimality, efficiency, and effectiveness across different AI governance challenges and contexts?

• *Targeting the right intervention*: Given a specific policy goal (e.g., mitigating bias), what part of the ML pipeline should be targeted for data compartmentalization? Are methods that only process outputs of models sufficient?

• *Metrics of success*: What are the appropriate metrics for measuring the success of data compartmentalization in achieving legal and ethical objectives? How can we balance these metrics with traditional model performance metrics?

• *Regulatory permissiveness*: Are data compartmentalization strategies that involve distributed compute, modularity and collaborative model development sanctioned under the EU AI Act? These approaches may limit the number of co-located model parameters while increasing the total parameter count, which makes it challenging to calculate the "risk" of the model according to EU AI Act parameter count metrics.

Outlook. As the field of AI continues to evolve, so too will the legal and ethical landscape surrounding data usage. The above strategies provide a flexible framework for addressing these evolving needs. By engaging with open questions through interdisciplinary dialogue, we pave the way for the development of responsible and compliant AI systems.

Acknowledgments and Disclosure of Funding

We want to thank Sébastien Krier and Jon Small for engaging in thoughtful discussion to inform and shape the AI governance and policy considerations we present in this work. We also want to thank Karolina Dziugaite, Daniel Ramage, Kaylee Bonawitz and Marc'aurelio Ranzato for their feedback on initial drafts.

References

Achille, A., Kearns, M., Klingenberg, C., and Soatto, S. Ai model disgorgement: Methods and choices. *Proceedings of the National Academy of Sciences*, 121(18):e2307304121, 2024. doi: 10.1073/pnas.2307304121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2307304121.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. On the opportunities and risks of foundation models, 2022.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE, 2021.

Brown, H., Lee, K., Mireshghallah, F., Shokri, R., and Tramèr, F. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 2280–2292, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3534642. URL https://doi.org/10.1145/3531146.3534642.

Daley, R. C. and Neumann, P. G. A general-purpose file system for secondary storage. In *Proceedings of the November 30–December 1, 1965, Fall Joint Computer Conference, Part I*, AFIPS '65 (Fall, part I), pp. 213–229, New York, NY, USA, 1965. Association for Computing Machinery. ISBN 9781450378857. doi: 10.1145/1463891.1463915. URL https://doi.org/10.1145/1463891.1463915.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating Noise to Sensitivity in Private Data Analysis. In *Proc. of the Third Conf. on Theory of Cryptography (TCC)*, pp. 265–284, 2006.

European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council, 2016. URL `https://data.europa.eu/eli/reg/2016/679/oj`.

Fan, C., Liu, J., Zhang, Y., Wei, D., Wong, E., and Liu, S. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*, 2023.

Goguen, J. A. and Meseguer, J. Security policies and security models. In *1982 IEEE Symposium on Security and Privacy*, pp. 11–11, 1982. doi: 10.1109/SP.1982.10014.

Golatkar, A., Achille, A., and Soatto, S. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9304–9312, 2020.

Graves, L., Nagisetty, V., and Ganesh, V. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11516–11524, 2021.

Izzo, Z., Smart, M. A., Chaudhuri, K., and Zou, J. Approximate data deletion from machine learning models. In *International Conference on Artificial Intelligence and Statistics*, pp. 2008–2016. PMLR, 2021.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Koh, P. W. W., Ang, K.-S., Teo, H., and Liang, P. S. On the accuracy of influence functions for measuring group effects. *Advances in neural information processing systems*, 32, 2019.

Kurmanji, M., Triantafillou, P., Hayes, J., and Triantafillou, E. Towards unbounded machine unlearning. *Advances in Neural Information Processing Systems*, 36, 2024.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Liu, J., Ram, P., Yao, Y., Liu, G., Liu, Y., Sharms, P., Liu, S., et al. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36, 2024.

Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., Wu, X., Shippole, E., Bollacker, K., Wu, T., Villa, L., Pentland, S., and Hooker, S. The data provenance initiative: A large scale audit of dataset licensing attribution in ai, 2023.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Min, S., Gururangan, S., Wallace, E., Hajishirzi, H., Smith, N. A., and Zettlemoyer, L. Silo language models: Isolating legal risk in a nonparametric datastore, 2023.

Muhamed, A., Thaker, P., Diab, M. T., and Smith, V. Cache me if you can: The case for retrieval augmentation in federated learning. In *Privacy Regulation and Protection in Machine Learning*, 2024. URL `https://openreview.net/forum?id=MKd1SkDbbz`.

Nguyen, T. T., Huynh, T. T., Nguyen, P. L., Liew, A. W.-C., Yin, H., and Nguyen, Q. V. H. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.

Pfeiffer, J., Ruder, S., Vulić, I., and Ponti, E. M. Modular deep learning, 2024.

Ponomareva, N., Hazimeh, H., Kurakin, A., Xu, Z., Denison, C., McMahan, H. B., Vassilvitskii, S., Chien, S., and Thakurta, A. G. How to dp-fy ml: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research*, 77:1113–1201, July 2023. ISSN 1076-9757. doi: 10.1613/jair.1.14649. URL `http://dx.doi.org/10.1613/jair.1.14649`.

Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečnỳ, J., Kumar, S., and McMahan, H. B. Adaptive federated optimization. In *International Conference on Learning Representations*, 2020.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020.

Robling Denning, D. E. *Cryptography and data security*. Addison-Wesley Longman Publishing Co., Inc., USA, 1982. ISBN 0201101505.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S., Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., Christiano, P., and Dafoe, A. Model evaluation for extreme risks, 2023.

Thudi, A., Deza, G., Chandrasekaran, V., and Papernot, N. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pp. 303–319. IEEE, 2022.

Tiwari, T., Gururangan, S., Guo, C., Hua, W., Kariyappa, S., Gupta, U., Xiong, W., Maeng, K., Lee, H.-H. S., and Suh, G. E. Information flow control in machine learning through modular model architecture, 2023.

UK Department for Science, Innovation and Technology. Frontier AI Safety Commitments. *AI Seoul Summit*, 2024.

Zhao, K., Kurmanji, M., Bărbulescu, G.-O., Triantafillou, E., and Triantafillou, P. What makes unlearning hard and what to do about it. *arXiv preprint arXiv:2406.01257*, 2024.

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: We state the claims and contributions of the paper in the abstract and introduction, and include a section dedicated to open questions which characterize the limitations of this work.

   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: We include a secion dedicated to "Open Questions" that call out the limitations of this work and opportunities for future work.

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: The paper does not include theoretical results.

   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: [NA]

   Justification: The paper does not include experiments.

   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: [NA]

   Justification: The paper does not include experiments requiring code.

   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: [NA]

   Justification: The paper does not include experiments.

   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: [NA]

   Justification: The paper does not include experiments.

   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: [NA]

   Justification: The paper does not include experiments.

   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: This paper conforms to the NeurIPS Code of Ethics. It contains no experiments involving data or human subjects, so there is no potential harm from the research process or societal risk posed.

   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 include a section "Opportunities for AI Governance" which points out the potential positive societal impacts of data compartmentalization. In "Open Questions" we present limitations and examine possible negative societal impacts of data compartmentalization.

    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: The paper poses no such risks as it does not release data or models.

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: [NA]

Justification: The paper does not include experiments that use existing assets.

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: The paper does not release new assets.

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 paper does not involve crowdsourcing nor research with 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 paper does not involve crowdsourcing nor research with 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.