



IronYun Policy Statement on Use of Datasets for Training AI Models

Effective Date: January 02, 2025

Purpose

This policy establishes guidelines to ensure that datasets used for training artificial intelligence (AI) models are handled ethically, legally, and responsibly. It aims to promote fairness, transparency, and accountability while safeguarding privacy, security, and the rights of individuals. The integrity of AI systems relies on the quality, diversity, and ethical use of the data from which they learn, and this policy provides a framework to guide such practices.

Scope

This policy applies to all personnel involved in the collection, selection, processing, and use of datasets for the purpose of training, validating, or evaluating AI models within the organization. It covers both internal and external datasets, including proprietary, public, and third-party data sources.

Guidelines for Dataset Use

1. Ethical Sourcing

Datasets must be obtained through ethical means, ensuring that data collection and usage respect the rights of individuals. Datasets should not violate privacy, intellectual property, or other rights protected by law. Consent must be obtained where necessary, particularly for personally identifiable information (PII) or sensitive data.

2. Data Privacy and Security

All datasets containing personal, confidential, or sensitive information must adhere to data protection regulations, including, but not limited to, the GDPR, CCPA, and HIPAA. Secure data storage, processing, and transfer mechanisms must be in place to protect datasets from unauthorized access, breaches, or misuse. Pseudonymization and anonymization should be considered when possible.

3. Bias and Fairness

Datasets must be carefully assessed for potential biases that could affect model outcomes and lead to unfair or discriminatory results. The selection of datasets should prioritize inclusivity, ensuring that diverse groups, perspectives, and data types are represented. Regular audits of training data for fairness and bias mitigation are required.

4. Transparency and Documentation

A clear and comprehensive record must be maintained for each dataset used, documenting its source, characteristics, and any preprocessing steps applied. This documentation should also include any known limitations of the dataset, as well as potential risks that may arise from its use. Transparency in dataset sourcing enables accountability and helps identify areas for improvement.

5. Data Quality and Integrity

The quality of data used for training AI models must be assessed for accuracy, consistency, and completeness. Data cleansing processes should be employed to eliminate errors, duplicates, or irrelevant information that could negatively impact the performance or reliability of the model. Continuous monitoring and validation of datasets will ensure their ongoing integrity.



6. Legal Compliance

All datasets must comply with applicable national and international laws governing data collection, use, and sharing. This includes respecting intellectual property laws, such as copyright and licensing agreements, as well as adhering to any specific terms of use set by data providers.

7. Environmental Impact

The environmental impact of data storage and processing should be considered, and efforts should be made to minimize energy consumption. The use of energy-efficient infrastructure and sustainable practices is encouraged to reduce the carbon footprint of AI model training.

8. Continuous Review and Improvement

This policy is subject to regular review to ensure its continued relevance and effectiveness in response to evolving legal, technological, and societal changes. Feedback from stakeholders, including data scientists, ethicists, and legal experts, will be incorporated into the review process.

Consequences of Policy Violation

Failure to adhere to this policy may result in disciplinary action, including suspension or termination of access to datasets, further training, or other corrective measures. Legal consequences may also apply if the use of datasets violates laws or regulations.

Conclusion

The ethical and responsible use of datasets in AI training is critical to fostering trust, ensuring the fairness of AI systems, and minimizing harm. All stakeholders must comply with this policy to ensure that the datasets used in AI training are appropriate, secure, and aligned with the values of fairness, accountability, and transparency.

For further details or inquiries regarding this policy, please contact sales@IronYun.com.