Upstream Indicators for the 2023 Foundation Model Transparency Index

Upstream
Data size: For the data used in building the model, is the data size disclosed?
Data sources: For all data used in building the model, are the data sources disclosed?
Data creators: For all data used in building the model, is there some characterization of the people who created the data?
Data source selection: Are the selection protocols for including and excluding data sources disclosed?
Data curation: For all data sources, are the curation protocols for those data sources disclosed?
Data augmentation: Are any steps the developer takes to augment its data sources disclosed?
Harmful data filtration: If data is filtered to remove harmful content, is there a description of the associated filter?
Copyrighted data: For all data used in building the model, is the associated copyright status disclosed?
Data license: For all data used in building the model, is the associated license status disclosed?
Personal information in data: For all data used in building the model, is the inclusion or exclusion of personal information in that data disclosed?
Use of human labor: Are the phases of the data pipeline where human labor is involved disclosed?
Employment of data laborers: Is the organization that directly employs the people involved in data labor disclosed for each phase of the data pipeline?
Geographic distribution of data laborers: Is geographic information regarding the people involved in data labor disclosed for each phase of the data pipeline?
Wages: Are the wages for people who perform data labor disclosed?
Instructions for creating data: Are the instructions given to people who perform data labor disclosed?
Labor protections: Are the labor protections for people who perform data labor disclosed?
Third party partners: Are the third parties who were or are involved in the development of the model disclosed?
Queryable external data access: Are external entities provided with queryable access to the data used to build the model?
Direct external data access: Are external entities provided with direct access to the data used to build the model?
Compute usage: Is the compute required for building the model disclosed?
Development duration: Is the amount of time required to build the model disclosed?
Compute hardware: For the primary hardware used to build the model, is the amount and type of hardware disclosed?
Hardware owner: For the primary hardware used in building the model, is the owner of the hardware disclosed?
Energy usage: Is the amount of energy expended in building the model disclosed?
Carbon emissions: Is the amount of carbon emitted (associated with the energy used) in building the model disclosed?
Broader environmental impact: Are any broader environmental impacts from building the model besides carbon emissions disclosed?
Model stages: Are all stages in the model development process disclosed?
Model objectives: For all stages that are described, is there a clear description of the associated learning objectives or a clear characterization of the nature of this update to the model?
Core frameworks: Are the core frameworks used for model development disclosed?
Additional dependencies: Are any dependencies required to build the model disclosed besides data, compute, and code?
Mitigations for privacy: Are any steps the developer takes to mitigate the presence of PII in the data disclosed?
Mitigations for copyright: Are any steps the developer takes to mitigate the presence of copyrighted information in the data disclosed?