Upstream Indicators for the 2025 Foundation Model Transparency Index

Indicator	Definition
Data acquisition methods	What methods does the developer use to acquire data used to build the model?
Public datasets	What are the top-5 sources (by volume) of publicly available datasets acquired for building the model?
Crawling	If data collection involves web-crawling, what is the crawler name and opt-out protocol?
Usage data used in training	What are the top-5 sources (by volume) of usage data from the developer's products and services that are used for building the model?
Notice of usage data used in training	For the top-5 sources of usage data, how are users of these products and services made aware that this data is used for building the model?
Licensed data sources	What are the top-5 sources (by volume) of licensed data acquired for building the model?
Licensed data compensation	For each of the top-5 sources of licensed data, are details related to compensation disclosed?
New human-generated data sources	What are the top-5 sources (by volume) of new human-generated data for building the model?
Instructions for data generation	For each of the top-5 sources of human-generated data, what instructions does the developer provide for data generation?
Data laborer practices	For the top-5 sources of human-generated data, how are laborers compensated, where are they located, and what labor protections are in place?
Synthetic data sources	What are the top-5 sources (by volume) of synthetic data acquired for building the model?
Synthetic data purpose	For the top-5 sources of synthetically generated data, what is the primary purpose for data generation?
Data processing methods	What are the methods the developer uses to process acquired data to determine the data directly used in building the model?
Data processing purpose	For each data processing method, what is its primary purpose?
Data processing techniques	For each data processing method, how does the developer implement the method?
Data size	Is the size of the data used in building the model disclosed?
Data language composition	For all text data used in building the model, what is the composition of languages?
Data domain composition	For all the data used in building the model, what is the composition of domains covered in the data?
External data access	Does a third-party have direct access to the data used to build the model?
Data replicability	Is the data used to build the model described in enough detail to be externally replicable?
Compute usage for final training run	Is the amount of compute used in the model's final training run disclosed?
Compute usage including R&D	Is the amount of compute used to build the model, including experiments, disclosed?
Development duration for final training run	Is the amount of time required to build the model disclosed?
Compute hardware for final training run	For the primary hardware used to build the model, is the amount and type of hardware disclosed?
Compute provider	Is the compute provider disclosed?
Energy usage for final training run	Is the amount of energy expended in building the model disclosed?
Carbon emissions for final training run	Is the amount of carbon emitted in building the model disclosed?
Water usage for final training run	Is the amount of clean water used in building the model disclosed?
Internal compute allocation	How is compute allocated across the teams building and working to release the model?
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?
Code access	Does the developer release code that allows third-parties to train and run the model?
Organization chart	How are employees developing and deploying the model organized internally?
Model cost	What is the cost of building the model?