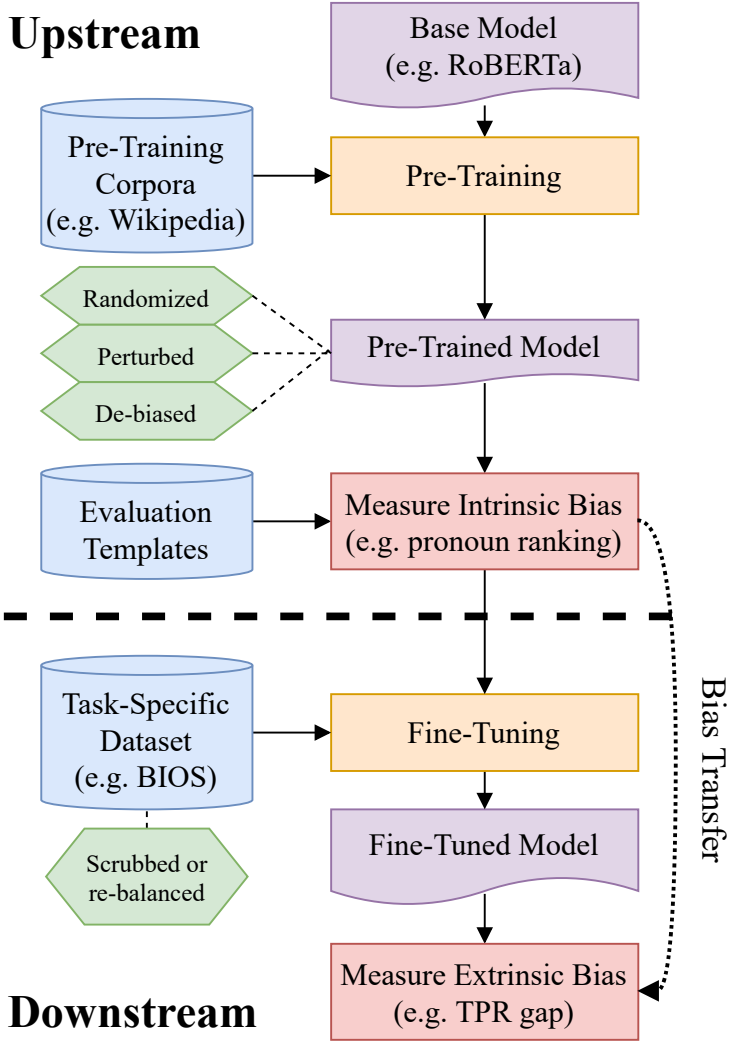
Upstream
Base Model
(e.g. RoBERTa)
Pre-Training
Corpora
(e.g. Wikipedia)
Pre-Training
Randomized
Perturbed
De-biased
Pre-Trained Model
Evaluation
Templates
Measure Intrinsic Bias
(e.g. pronoun ranking)
Task-Specific
Dataset
(e.g. BIOS)
Fine-Tuning
Scrubbed or
re-balanced
Fine-Tuned Model
Measure Extrinsic Bias
(e.g. TPR gap)
Downstream
Bias Transfer