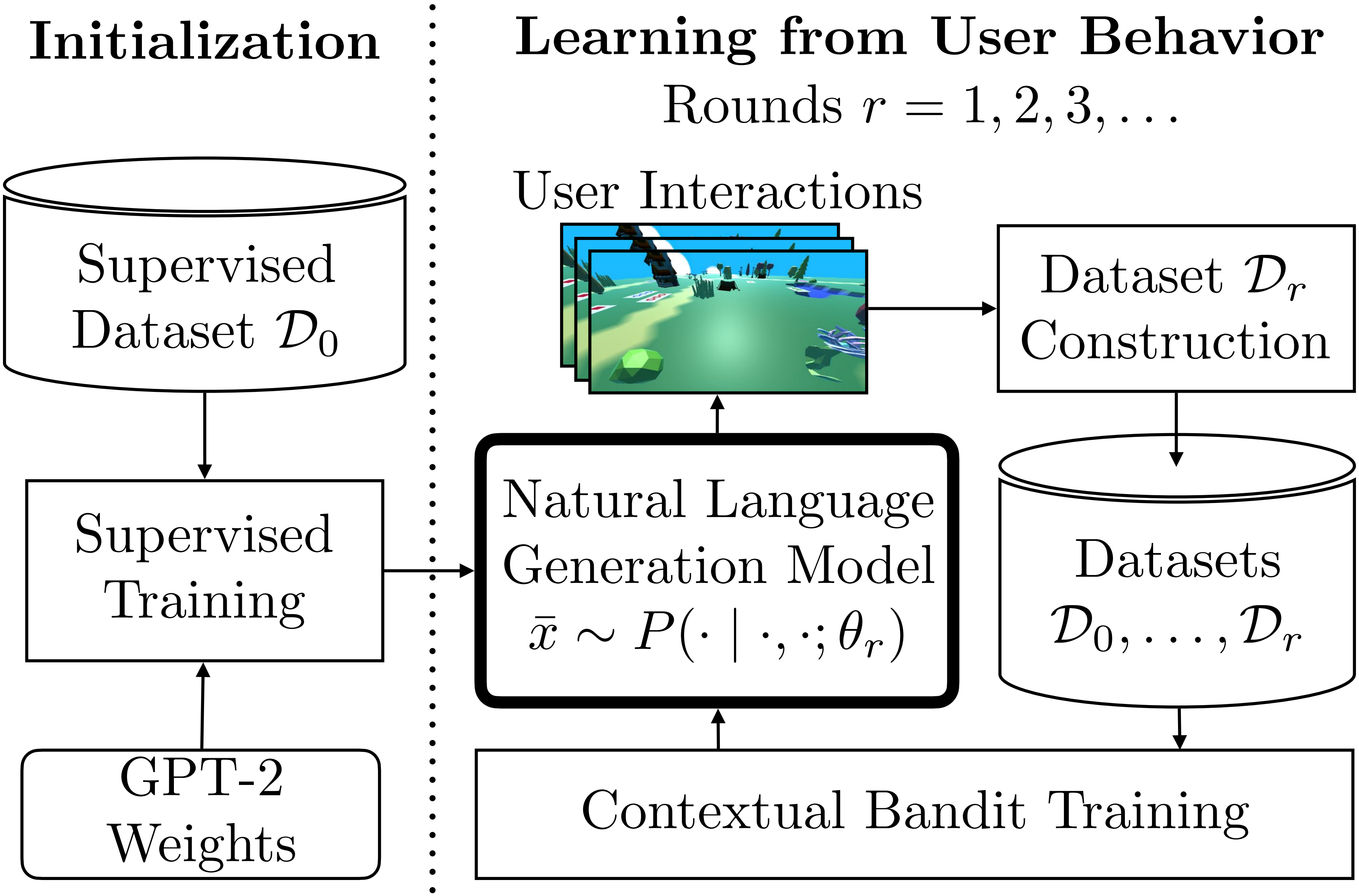
Initialization
Learning from User Behavior
Rounds $r = 1, 2, 3, \ldots$
User Interactions
Supervised Dataset $\mathcal{D}_0$
Dataset $\mathcal{D}_r$ Construction
Supervised Training
Natural Language Generation Model
$\bar{x} \sim P(\cdot \mid \cdot, \cdot; \theta_r)$
Datasets $\mathcal{D}_0, \ldots, \mathcal{D}_r$
GPT-2 Weights
Contextual Bandit Training