Algorithm 1 NLPO- Natural Language Optimization

Input: Dataset $\mathcal{D} = \{(\mathbf{x}^i, \mathbf{y}^i)\}_{i=1}^N$ of size N
Input: initial policy parameter π_{θ_0}
Input: initial LM π_0
Input: initial value function parameters V_{ϕ_0}
Input: initialize parameterized masked policy $\pi_{\psi_0}(\cdot|\cdot, \pi_{\theta_0})$ with parameterized top-p policy π_{θ_0}
Input: policy update frequency μ
repeat
    Sample mini-batch $\mathcal{D}_m = \{(\mathbf{x}^m, \mathbf{y}^m)\}_{m=1}^M$ from $\mathcal{D}$
    Collect trajectories $\mathcal{T}_{\tau_i}$ by running policy π_{ψ_n} in for batch $\mathcal{D}_m$ in env
    Compute Preference and KL penalty rewards $\hat{R}_t$
    Compute the advantage estimate $\hat{A}_t$
    Update the policy by maximizing the PPO-Clip objective:

$$\pi_{\theta_{m+1}} = \text{argmax}_\theta \frac{1}{\mathcal{D}_m T} \sum_{\tau \in \mathcal{D}_m} \sum_{\tau=0}^{T} \min(r_t(\theta) A^{\pi_{\theta_m}}, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon) A^{\pi_{\theta_m}})$$

    where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_m}(a_t|s_t)}$.

    Update the value function:

$$V_{\phi_{m+1}} = \text{argmin}_\phi \frac{1}{\mathcal{D}_m T} \sum_{\tau \in \mathcal{D}_m} \sum_{t=0}^{T} \left(V_\phi(s_t) - \hat{R}_t \right)^2$$

    Update the parameterized masked poicy every μ iterations:

$$pi_{\psi_{n+1}}(\cdot|\cdot, \pi_{\theta_{m+1}})$$

until convergence and **return** π_θ