A Crowdsourcing Details

Quality Control Since this is a creative annotation task for crowdworkers, rather than a tagging or selection task, we need two groups of crowdworkers for two separate steps: 1) workers to create a counterfactual alternatives for the storylines, 2) workers to create a new story ending that is coherent and logically consistent with the previous context that only changes the original story arc to regain narrative consistency. Crowdworkers with more than 5000 HITs and at least a 99% acceptance rate can take our qualification test, in which we require each crowdworker to do 3 HITs before being approved for the full task. We encourage workers to submit feedback to help us improve our instructions.

Cost We pay \$0.24 to crowdworkers per instance for Step 1 and \$0.36 per instance for Step 2.

B Training Hyperparameters

GPT2 Text is encoded with BPE using a vocabulary size of 50,257. We set the maximum sequence length to 128 tokens, which we found is large enough to contain complete stories. We use Adam optimization with an initial learning rate of 10^{-5} and a minibatch size of 2. We train the models for 10K iterations using early stopping to select the model that does the best on the validation set. At inference time, we generate using the same procedure outlined in Radford et al. (2019): top-k sampling with temperature set to 0.7 and k set to 40.

GPT All models follow the setting of GPT (Radford et al., 2018) that used a 12-layer decoder-only transformer with masked self-attention heads. Text is encoded with BPE using a vocabulary size of 40,000. As above, we set the maximum sequence length to 128 tokens. We use Adam optimization with an initial learning rate of 6.25^{-5}. We train the models for 10K iterations using early stopping to select the model that does the best on the validation set. We use the same generation procedure as for GPT2.