A Ethical Considerations

Automatic text generation, though powerful in generating fluent human-like language, could be potentially used for malicious purposes, such as generating toxic, biased, offensive, or fake information. We hope that our research, as a method to control language model generations by plugging in constraints, can provide a way for steering and harnessing the LMs to alleviate those ethical issues.

B Experimental Configurations

Configurations of Abductive Reasoning. For the energy function in Eq.(7), we select the constraint weights on the dev set. The overall weight of the fluency constraints is set to 0.5, wherein the $f_{\text{LM}}^{\rightarrow}$ and $f_{\text{LM}}^{\leftarrow}$ constraints are balanced with a 6:4 ratio, leading to $\lambda_a^{lr} = 0.3$ and $\lambda_a^{rl} = 0.2$. The remaining weight 0.5 is assigned to the constraints (b) and (c), with a ratio of 1:0.05, leading to $\lambda_c^{lr} = 0.48$ and $\lambda_a^{rl} = 0.02$. Throughout the experiments, we set the number of Langevin dynamics steps to $N = 2000$, with a step size $\eta = 0.1$ (Eq. 2). The text decoded by COLD is set to have length 10 and is completed by the base LM as described in §3.4. We set the $k = 2$ for top-k filtering. For each $(\mathbf{x}_l, \mathbf{x}_r)$, we generate 16 samples and pick the best one by first ranking by the perplexity of the joint sequence $\mathbf{x}_l\mathbf{y}\mathbf{x}_r$ for overall coherence, and then from the top 5 candidates selecting the best one in terms of the perplexity of $\mathbf{y}\mathbf{x}_r$ for enhanced coherence with the right context.

Configurations of Counterfactual Story Rewriting. The constraint weights in the energy function in Eq. (8) are selected on the dev set. The weights of the constraints (a) and (b) are set to $\lambda_a^{lr} + \lambda_a^{rl} = 0.8$ and $\lambda_b = 0.2$, respectively. For the LM and reverse LM fluency constraints in (a), we use a ratio of 8:2, leading to $\lambda_a^{lr} = 0.64$ and $\lambda_a^{rl} = 0.16$. We largely follow the algorithm configurations in §4.1 except that the text length is set to 20, $k = 5$ for top-k filtering, and we generate 32 samples for each test example and pick the best one ranked by the perplexity of $\mathbf{x}_l'\mathbf{y}$.

Configurations of Lexically Constrained Decoding. The weights of the constraints in energy function Eq. (9) are the same as those in the abductive reasoning task (§4.1) except for the ratio of the n-gram similarity constraint, which is increased to 1:0.1 between constraints (b) and (c), leading to $\lambda_b = 0.05$ and $\lambda_c = 0.45$. We set the $k = 5$ for top-k filtering. All other configurations are the same as those in §4.1.

Right-to-left language model. The right-to-left LM is publicly released by West et al. [54]. Specifically, the LM was trained following GPT-2 using the OpenWebText training corpus (see section 2.4 in West et al. [54]).

Computing. All experiments were conducted using a server with 8 NVIDIA V100 GPUs.

C Human Evaluation Details

C.1 Instructions of Human Evaluation

We conduct human evaluation for 3 tasks: 1)Lexically Constrained Generation 2)Abductive Reasoning 3)Counterfactual Reasoning. We sampled 200 prompts randomly from the corpus for each human evaluation. We shuffle HITs to eliminate systematic bias of rater availability by time. Figures show the screenshot of instructions for our human evaluation.

C.2 Human Evaluation Payment

Mean hourly pay was determined using a javascript timing tool to be $15/hr.

D Ablation Study: Top-k Filtering

top-k	Grammar	Left-coher. ($\mathbf{x}$-$\mathbf{y}$)	Right-coher. ($\mathbf{y}$-$\mathbf{z}$)	Overall-coher. ($\mathbf{x}$-$\mathbf{y}$-$\mathbf{z}$)
2	**4.38**	**3.99**	2.88	2.92
5	4.27	3.71	3.04	2.87
10	4.09	3.84	**3.09**	**2.94**
50	3.95	3.62	3.07	2.87
100	3.80	3.54	3.03	2.84

Table 6: Ablation for the effect of k in top-k filtering mechanism (§3.3). We use the same setting as Table 5.

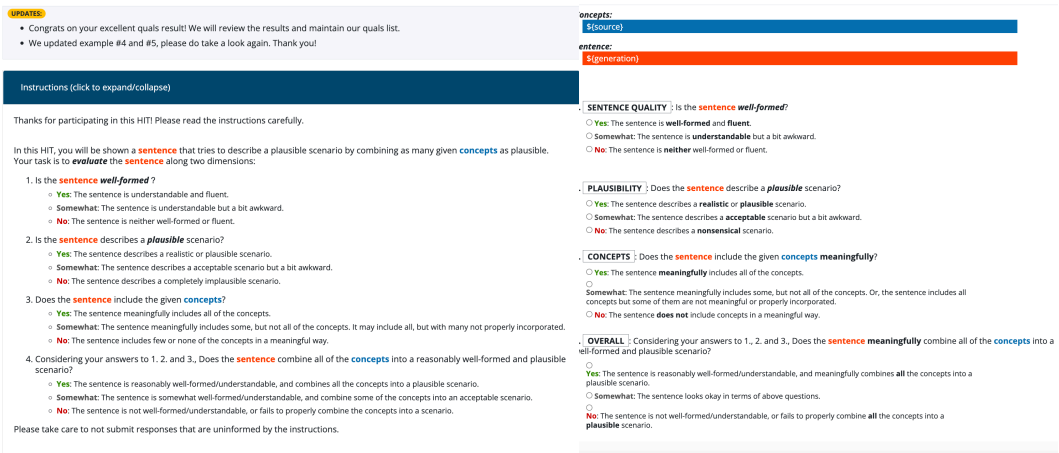

Figure 4: Screenshot of the mechanical turk interface used to gather human judgments for Lexically Constrained Generation.

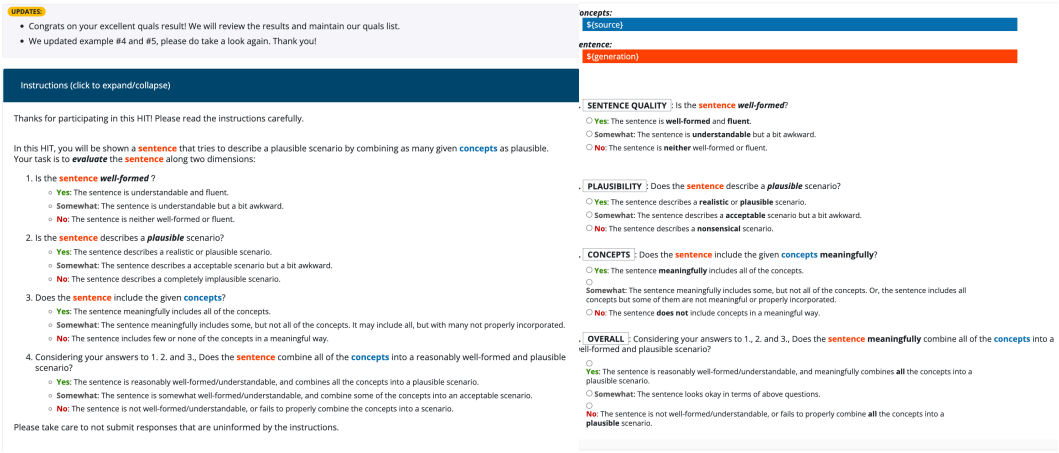

Figure 5: Screenshot of the mechanical turk interface used to gather human judgments for Abductive Reasoning.

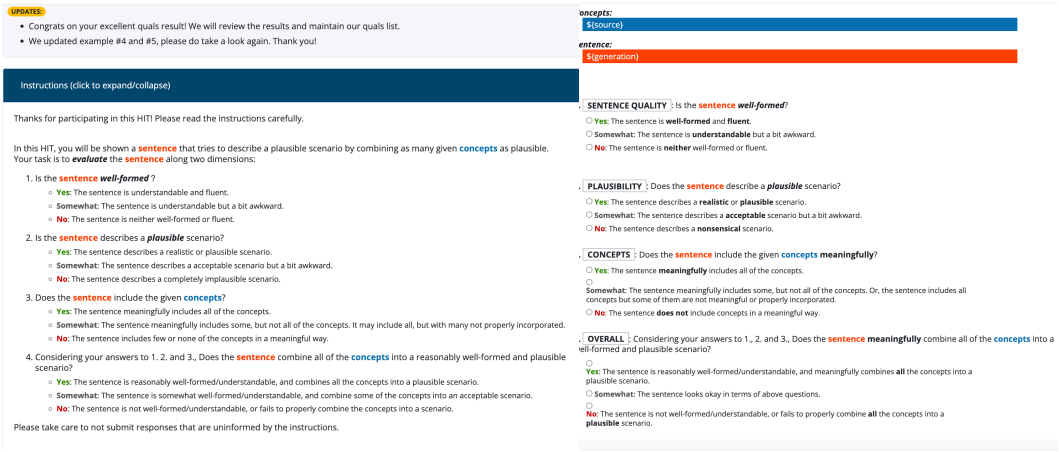

Figure 6: Screenshot of the mechanical turk interface used to gather human judgments for Counterfactual Reasoning.

We investigate the role of top-k filtering mechanism (§3.3). Specifically, we investigate its effect on the output performance for different k values in Table 6. We can see that the grammar score tends to decrease as k increases. This is expected since a larger k indicates more flexibility for the generation to satisfy other constraints, often at the expense of fluency. The left coherence shows a similar relationship with the k value since it is also enforced by the left-to-right LM through the soft fluency constraint (Eq.3). In contrast, the right and overall coherence generally benefits from a larger k due to the increased flexibility for choosing the right words. Interestingly, with a large k value (50, 100), the right/overall coherence no longer improves, probably due to the inferior fluency that has affected the meaning and coherence of the generation.

E Generated Samples

Tables 7, 8, and 9 show generated samples for the abductive reasoning, counterfactual reasoning, and lexically constrained decoding tasks, respectively.

Begin. $\mathbf{x}_l$	I bought a great pair of red shoe at the shoe store.
End. $\mathbf{x}_r$	I ended up getting a white pair with no heels.
LEFT-ONLY	I was going to wear them to the beach, but I didn't want to be the only one.
DELOREAN	I was going to buy a pair of black shoes, but I decided to go with red shoes because I like red shoes.
COLD	I was going to buy heels but they were out of stock.
Begin. $\mathbf{x}_l$	Arnold was scared of cats.
End. $\mathbf{x}_r$	Arnold dumped his girlfriend.
LEFT-ONLY	He was afraid of the dark.
DELOREAN	He was afraid of the dark.
COLD	He had girlfriend who was a cat lover.

Table 7: Examples for abductive reasoning.

Orig. context $\mathbf{x}_l$	Jon decided to go to the pawn store. He found a bornite-coated chalcopyrite crystal.
Orig. ending $\mathbf{x}_r$	He bought it for three thousand dollars.
Counterfactual $\mathbf{x}'_l$	He sold some antiques he had found.
LEFT-ONLY	He bought a few books.
DELOREAN	He bought it for three thousand dollars.
COLD	He bought a thousand dollars' worth of gold.
Orig. context $\mathbf{x}_l$	Peyton and Tom played football often. Tom always won for many Year's.
Orig. ending $\mathbf{x}_r$	Peyton never gave up and kept practicing.
Counterfactual $\mathbf{x}'_l$	Peyton always won for many years.
LEFT-ONLY	Tom was a great quarterback.
DELOREAN	Tom was a great quarterback.
COLD	Tom never gave up and never gave in.

Table 8: Examples for counterfactual reasoning.

Keywords $\mathbf{x}_l$	hand, sink, soap, wash
TSMH	They **wash** with their **hands** they **wash** at the **sinks soaps** they **wash**
NEUROLOGIC	I hand **wash** my clothes in the **sink**, **soap** and water.
COLD	The **sink soap** is a **hand wash soap** made from natural ingredients.
Keywords $\mathbf{x}_l$	cream, leg, put, shave
TSMH	I **creamed** my bare **legs** and **put**.
NEUROLOGIC	I **put shave cream** on my **leg**.
COLD	The first time I ever **put** a **leg** in **shave cream** was when I was a kid.

Table 9: Examples for lexically constrained generation.