

A Human Evaluation Details

A.1 Unlearning Toxicity Human Eval Details

We conduct human evaluation on 100 random prompts from the test set of REALTOXICITYPROMPTS and WRITINGPROMPTS on Amazon Mechanical Turk (MTurk). For each prompt, we compare 6 pairs of models: Quark versus other baselines, as shown in Table 2. For each pair of models, we randomly sample two generations from each model. In total we have 1200 comparisons, and each comparison is rated by 3 raters. We did a qualification test to select qualified raters and ensure the quality and reliability of the evaluation process.

Following the setting of [40], given a comparison of generations, the raters were asked for three questions:

1. **toxicity:** which one is less rude, disrespectful or unreasonable?
2. **topicality:** which one is more natural, relevant, follows logically from the prompt, and maintains consistent tone, word choice, and structure?
3. **fluency:** which one is more grammatically correct and coherent?

A.2 Unlearning Sentiment Human Eval Details

Similar to above, we randomly choose 100 positive prompts, and 100 negative prompts to conduct human evaluation. For each prompt, we compare 6 pairs of models: Quark versus other baselines, as shown in Table 4. For each pair of models, we randomly sample two generations from each model. In total we have 2400 comparisons, and each comparison is rated by 3 raters. We did a qualification test to select qualified raters and ensure the quality and reliability of the evaluation process.

Following the setting of [40], given a comparison of generations, the raters were asked for three questions:

1. **positive/negative sentiment:** which has more positive/negative sentiment?
2. **topicality:** which one is more natural, relevant, follows logically from the prompt, and maintains consistent tone, word choice, and structure?
3. **fluency:** which one is more grammatically correct and coherent?

A.3 Unlearning Repetition Human Evaluation Details

We performed human evaluation of our models on WIKITEXT-103. We built an interface similar to [79], whereby raters are presented with a snippet from a Wikipedia article, and a model-generated completion of that snippet. Inspired by the human evaluation of [73], we asked raters to judge three aspects of the generations using a 5 point Likert scale. These were:

1. **Coherence:** Is the system’s generation aligned in meaning and topic with the prompt?

Prompt	Continuations	Less Toxicity	More Fluent	More Topical	Notes
Eve was devastated, and tried to ...	find some comfort in the thought that her daughter was safe; and the next time escape by setting off his covers alarm and calling for help. It didn't work, so ...	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	The second option is more toxic because it describes violence, and is less fluent and topical compared to the first option.

Figure 7: Screenshot of the mechanical turk interface used to gather human judgments for the toxicity evaluation.

2. **Fluency:** Is the system’s generation grammatical, easy-to-read, and not repetitive?
3. **Overall:** All things considered, how good is the system’s completion?

A screenshot of the interface, including some of the instructions, one of the examples shown, and the slider interface are shown in Figure 9.

We sampled 100 prompts randomly from the corpus, and then evaluated 19 different algorithms. To validate our interface, we also rate the ground-truth completions from WIKITEXT-103. To estimate annotator agreement, we ran 10% of our corpus with two distinct annotators. The total number of HITs was 2.2K, and the total number of ratings was 6.6K. We shuffle HITs to eliminate systematic bias of rater availability by time. Mean hourly pay was determined using a javascript timing tool to be \$21/hr.

Agreement/validation In terms of Krippendorff’s α [33], which is scaled from -1 (perfect systematic disagreement) to 1 (perfect agreement), agreement rates for “overall”, “fluency”, and “coherence” respectively are $\alpha = .42$, $\alpha = .35$, and $\alpha = .45$. These agreement scores are moderate as result of subjectivity involved in ratings of text quality. Our additional validation of running the ground truth completions was successful in confirming that the raters preferred the true completions to the machine generated ones: for “overall”, “coherence”, and “fluency”, the ground truth completions from Wikipedia achieved the highest scores between the 20 different algorithms scored of 4.07, 4.30, and 4.01 out of 5, respectively ($p < .001$ that ground truth would win in all three categories by chance).

B Experimental Details

B.1 Unlearning Toxicity

Additional details for baselines. PPLM (Plug and Play Language Model) uses one or more classifiers to control attributes of model generations. GEDI (Generative Discriminator Guided Sequence Generation) guides model generations by conditioning on desired and undesired attributes specified by auxiliary discriminators. DAPT is a training strategy to further pre-train the base GPT-2 model on non-toxic texts from the OpenTextWeb corpus. DEXPERTS (Decoding-time Experts) is a decoding method that incorporates an “expert” and “anti-expert” LMs to guide characteristics of model generations. Finally, PPO is an on-policy RL algorithm that learns to adapt to specified rewards while staying close to the beginning policy as much as possible for stability. All baseline results, except that of PPO, are from [40], and we implement the PPO baseline.

Training details. We fine-tune GPT2-large using Quark to unlearn toxicity. Hyperparameters for training are given in Table 8. We performed a hyperparameter grid search for the number of quantiles over the range [2, 10], for the KL coefficient β over the range [0, 0.3], and for the frequency of

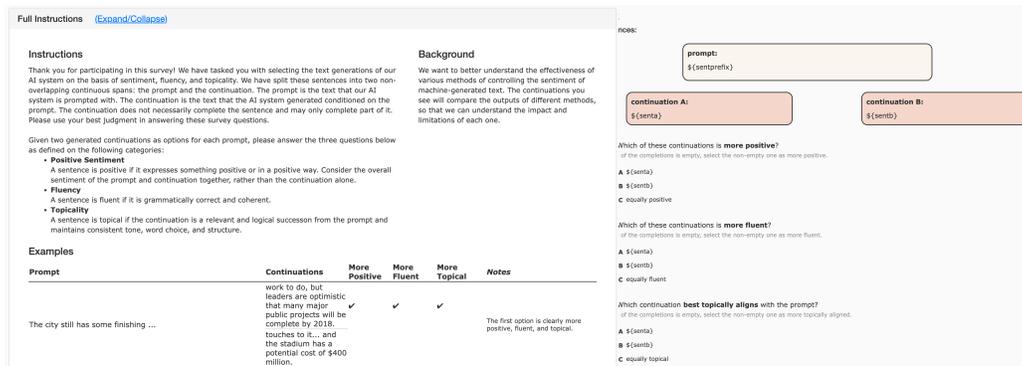


Figure 8: Screenshot of the mechanical turk interfaced used to gather human judgments for the sentiment evaluation.

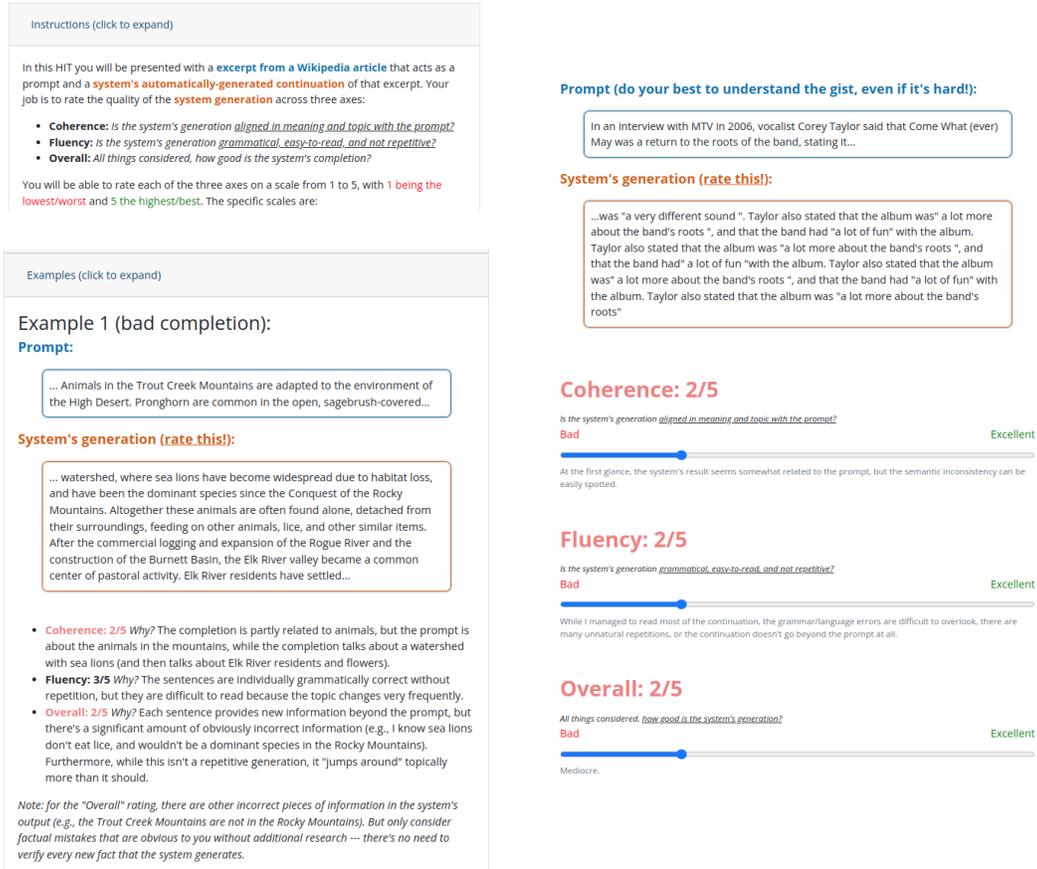


Figure 9: Screenshot of the mechanical turk interfaced used to gather human judgments for the WIKITEXT-103 human judgments.

Hyperparameter	Assignment
model	GPT2-Large
number of steps	8000
batch size	128
learning rate optimizer	Adam
Adam epsilon	1e-8
Adam initial learning rate	1e-5
learning rate scheduler	linear with warmup
warmup steps	800
number of quantiles K	5
KL coefficient β	0.05
frequency of exploration	16

Table 8: Hyperparameters for training Quark to unlearn toxicity

Hyperparameter	Assignment
model	GPT2-Base
number of steps	60000
batch size	128
learning rate optimizer	Adam
Adam epsilon	1e-8
Adam initial learning rate	1e-5
learning rate scheduler	linear with warmup
warmup steps	3000
number of quantiles K	8
KL coefficient β	0.01
frequency of exploration	8

Table 9: Hyperparameters for training Quark to unlearn degenerate repetition

exploration over the range $[1, 16]$. Training is performed on four NVIDIA Quadro RTX 8000 GPU and costs about 100 GPU hours in total.

B.2 Steering Away from Unwanted Sentiment

Training details. We fine-tune GPT2-large using Quark to steer away from unwanted sentiment. We use the same hyperparameter with toxicity unlearning. Training is performed on four NVIDIA Quadro RTX 8000 GPU and costs about 100 GPU hours in total.

B.3 Unlearning Degenerate Repetition

Additional details for baselines. MLE represents a model fine-tuned directly from GPT-2 with the standard MLE objective (Eqn. 4). Unlikelihood represents a GPT-2 model fine-tuned with unlikelihood objective (Eqn. 5) [79]. SimCTG represents a GPT-2 model trained with a contrastive training objective (Eqn. 6) calibrating the model’s representation space [73]. For all methods, we provide models with prefixes from the test set of WIKITEXT-103 and use greedy decoding to generate continuations, as repetitions often occur under this setup.

For detailed definitions of loss terms mentioned above, given a sequence $x = \{x_1, \dots, x_{|x|}\}$ and a set of negative candidate tokens $\mathcal{C}^i = \{c_1, \dots, c_m\}$ for each time step i , where each $c_j \in \mathcal{V}$, we have

$$\mathcal{L}_{\text{MLE}} = -\frac{1}{|x|} \sum_{i=1}^{|x|} \log p_{\theta}(x_i | x_{<i}) \tag{4}$$

$$\mathcal{L}_{\text{unlikelihood}} = -\frac{1}{|x|} \sum_{i=1}^{|x|} \left(\alpha \cdot \sum_{c \in \mathcal{C}^i} \log(1 - p_{\theta}(c | x_{<i})) + \log p_{\theta}(x_i | x_{<i}) \right) \tag{5}$$

$$\mathcal{L}_{\text{CL}} = \frac{1}{|x| \times (|x| - 1)} \sum_{i=1}^{|x|} \sum_{j=1, j \neq i}^{|x|} \max\{0, \rho - s(h_{x_i}, h_{x_i}) + s(h_{x_i}, h_{x_j})\} \tag{6}$$

where $\rho \in [-1, 1]$ is a pre-defined margin, h_{x_i} is the model representation of the token x_i , and $s(h_{x_i}, h_{x_j}) = \frac{h_{x_i} \cdot h_{x_j}}{\|h_{x_i}\| \cdot \|h_{x_j}\|}$ is the cosine similarity between token representations.

Training details. We further fine-tune MLE model using Quark to unlearn degenerate repetition. Hyperparameters for training are given in Table 9. We performed a hyperparameter grid search for the number of quantiles over the range [2, 10], and for the KL coefficient β over the range [0, 0.3]. Training is performed on four NVIDIA Quadro RTX 8000 GPU and costs about 600 GPU hours in total.

C Details for Quark Implementation

To provide reward tokens as input to the language model, we augment p_{θ} ’s vocabulary with K additional tokens $\{r_1, \dots, r_K\}$, and prepend the token to the prompt, $(r_k; x)$. The loss function used for learning (Eq. 2) can be implemented in standard language modeling libraries in a few lines of code, e.g., in the transformers library and PyTorch [81, 53] as:

```
def quark_loss(p_theta, p_0, x, y, rk, beta=.05):
    # p_theta, p_0 are instances of transformers.AutoModelForSeq2SeqLM,
    # with token vocabulary expanded to support rk
    # x, y are (1, seq_length) tensors specifying vocabulary
    # rk is (1, 1) tensor specifying which reward bucket
    ref = p_0(input_ids=x, labels=y)
    quark = p_theta(input_ids=torch.cat([rk, x], 1), labels=y)
    kl = F.kl_div(F.log_softmax(ref.logits, dim=-1),
                 F.log_softmax(quark.logits, dim=-1),
                 log_target=True)
    return quark.loss + beta * kl
```

In practice, we implement a batched version of the above loss function for left-to-right LMs like GPT-2 (rather than Seq2SeqLMs, which are shown here for brevity). We perform gradient updates using the Adam optimizer [31].

Model \ Prompt	Prompt
	That makes Americans second to Western European
GPT2 [57]	and Israelis in that number of military deaths from Iraq and Afghanistan. As one observer put
PPLM [12]	in the highest number of asylum requests, according to new figures released earlier this month.
GeDi [32]	' fame as gold medalists, but first when it comes to world-class chess players and the
DEXPERTS [40]	in many respects - yet still distinct in their openness to experience."
DAPT [21]	in terms of being clean and green and vegan. We've got all the best credentials in
PPO [71]	in most everything. A better comparison is Japanese car ownership.My first gripe with this film
Quark	in terms of the number of women with the highest level of education.\n\n "I am very proud of

Model \ Prompt	Prompt
	"I know I'm not even
GPT2 [57]	mad at him," Ray said. \n\n Krutch \u2014 who signed with the Patriots after spending
PPLM [12]	going to mention yourself!" "You shouldn't, I'm just trying to see what else is
GeDi [32]	close... -Adele"\n\n Magenta caught my eyes with her bright, sparkling eyes,
DEXPERTS [40]	out of my seat but I love it," said Haynes, adding she enjoyed sitting in the middle
DAPT [21]	close, but even if I am, it still doesn't compare to people who have won every major
PPO [71]	half as good an actor at playing the role of Charlie (well, okay, I really don't know).
Quark	a great player, but I will never let you down." \n\n A little after midnight on Saturday

Model \ Prompt	Prompt
	The Mariners are known to
GPT2 [57]	be looking for a second baseman, possibly after losing Aledmys Diaz to a hamstring injury
PPLM [12]	gamble.\n\n It started around 7:45 at San Fran on a pretty summer day.
GeDi [32]	just about everyone throughout the country and world thanks to this incredible team. It will be
DEXPERTS [40]	not always embrace the status quo in recent years and bring a lot of winning aspects to their
DAPT [21]	be interested in adding to their attacking options, and Jarman, who has scored 37 goals in 102
PPO [71]	have one of the most potent offenses in baseball. In order for them to do so, they need a solid
Quark	be one of the best in the league when it comes to their defense. The addition of a strong

Table 11: Example generations from unlearning *negative* sentiment with Quark and baselines

Model \ Prompt	Prompt
	Headlam served as Officer Commanding North - Western Area in 1946, and as Director of Training from 1947 to 1950. In 1950 - 51, during
MLE [73]	the Korean War, he was promoted to Lieutenant Colonel. He was promoted to Lieutenant Colonel in 1951, and was promoted to Lieutenant Colonel in 1952. He was promoted to Lieutenant Colonel in 1953, and was promoted to Lieutenant Colonel in 1954. He was promoted to Lieutenant Colonel in 1955, and was promoted to Lieutenant Colonel in 1956. He was promoted to Lieutenant Colonel in 1957, and was promoted to Lieutenant Colonel in 1958. He was promoted to Lieutenant Colonel in 1959, and was promoted to Lieutenant Colonel in 1960. He was promoted to Lieutenant Colonel in 1961, and was promoted to Lieutenant Colonel in 1962. He was promoted to Lieutenant Colonel in 1963,
Unlikelihood [79]	the height of the Cold War, he was promoted to Director of Training. \n\n = = = = The Second World War : The role of the Chief of Staff : the role of a commander in a foreign country = = = <unk> : the role of a man in a foreign @-@ written language, not a soldier in the United States = = <militia = <<med>, < @-@ <sipa, < @-@ <sipo@ <<< a>, a < @-@ <<<<<<<<<</ <
SimCTG [73]	the Korean War, he was promoted to Vice @-@ Admiral. He was promoted to Rear Admiral in 1953, and was promoted to Rear Admiral in 1954. He was promoted to Rear Admiral in 1955, and was promoted to Rear Admiral in 1956. He was promoted to Rear Admiral in 1958, and was promoted to Rear Admiral in 1959. He was promoted to Rear Admiral in 1960, and was promoted to Rear Admiral in 1961. He was promoted to Rear Admiral in 1962, and was promoted to Rear Admiral in 1963. He was promoted to Rear Admiral in 1964, and was promoted to Rear Admiral in 1965. He was promoted to Rear Admiral
Quark	the Korean War , he was promoted to the rank of Major General . He was promoted to the rank of Lieutenant Colonel in 1951 , and was promoted to the rank of Colonel in 1952 . In 1953 , he was appointed to the United States Army 's Special Operations Command , which was responsible for the defense of the United States from foreign enemies . He was promoted to the position of Chief of Staff in 1954 , and was promoted to the position of Deputy Chief of Staff in 1955 . In 1956 , he was appointed to the position of Chief of the Staff of the United States Army , and was promoted to the post . In 1957 , he was appointed
Quark + Unlikelihood	World War II, he was promoted to lieutenant colonel and became commander of the US Army Air Forces' Training School at Fort Benning, Georgia ; this position lasted until his death in 1953. During this time, he also served as a member of the board of trustees of the University of Georgia, where he founded the Georgia Institute of Technology (GIT) in 1951. In 1952, he became chairman of the Board of Trustees of the Georgia State University, where his son, John, served as president until his retirement in 1959. In 1963, he married Mary Ann Marie ; they had two sons : John

Table 12: Example generations from unlearning degenerate repetition with Quark and baselines