Appendices for A Simple Contrastive Learning Objective for Alleviating Neural Text Degeneration

Anonymous Author(s)
Affiliation
Address
`email`

A Ethical considerations

In this work, we used publicly available English data to train/validate/test models. As far as we know, the curators of these datasets have taken ethical issues into consideration when creating the datasets. We manually checked some generated texts of the language models trained by CT and did not observe any noticeable traces of concern, such as offensive and malevolent language. We share our source code and trained model weights to support its correct use. To make sure the human workers involved in the data labeling efforts, as part of the human evaluation for this study, are fairly paid, we applied the minimum hourly rate of 10.48 euros, which converts to 11 dollars per hour. However, we warn that generative language models should always be used with caution since the generated texts are usually novel and unexpected wordings may appear when trained on improper data. Especially, generative models can be used maliciously, e.g., to generate fake news articles.

B Using CT in your work

Algorithm 1 Calculate contrastive token loss

Input: Labels $X = (x_1, x_2, \ldots, x_{|X|})$, time t, negative window size M, logits Z_t of time t
Output: Contrastive token loss $\mathcal{L}_{CT}^t$

1: $S_N^t \leftarrow SampleNegatives(X, M, t)$ # according to Eq. (7)
2: $z_{x_t} \leftarrow GatherLogits(Z_t, x_t)$ # positive logits
3: $z_{S_N^t} \leftarrow GatherLogits(Z_t, S_N^t)$ # negative logits
4: $\mathcal{L}_{CT}^t \leftarrow \log\left(1 + \sum_{x_t^- \in S_N^t} \exp(z_{x_t^-} - z_{x_t})\right)$ # Eq. (5)
5: **return** $\mathcal{L}_{CT}^t$

We summarize the steps for calculating $\mathcal{L}_{CT}^t$ in Algorithm 1. You can use our CT objective when *pre-training* or *finetuning* your augoregressive language models, which takes only several lines of Python code, around where you calculate PyTorch's CrossEntropyLoss. Simply use `pip install ct-loss` to install the required packages. Then you can use CT as follows:

```python
import torch

# Suppose we already have the model output logits and labels (sequences
# of token indices). For example when the batch size is 10, sequence
# length is 50 and vocabulary size is 1000:
logits = torch.rand(10, 50, 1000)
labels = torch.randint(0, 999, (10, 50))

# This is how you normally use cross-entropy for a language model:
```

```
from torch.nn import CrossEntropyLoss
ce_criterion = CrossEntropyLoss()
ce_loss = ce_criterion(logits.view(-1, 1000), labels.view(-1))

# This is how you can use our contrastive token loss:
from ct.ct_loss import ContrastiveTokenLoss
ct_criterion = ContrastiveTokenLoss(pad_id=999) # we need pad tokens for masking out tokens in a sequence that should not be used as negative tokens
ct_loss = ct_criterion(logits, labels)

# In our paper, we use CE and CT together
loss = ce_loss + ct_loss
```

C Noise-contrastive estimation for autoregressive language models

We adapted NCE [2] to token-level:

$$\mathcal{L}^t_{NCE} = -\log \sigma(h_t^T W_{x_t}) - \frac{1}{|S_N^t|} \sum_{x_t^- \in S_N^t} \log \sigma(-h_t^T W_{x_t^-}), \tag{8}$$

where $\sigma(\cdot)$ is the *sigmoid* function.

D Gradient functions

To see how loss functions influence the logits during training, we compare the gradient of each loss function. Writing $z_{x_t} = h_t^T W_{x_t}$ for the logit of token x_t, the gradient function is calculated by $\partial \mathcal{L}_* / \partial z_*$, where $\mathcal{L}_* \in \{L_{CE}, L_{UL}, L_{CT}\}$, and $z_* \in \{z_{x_t}, z_{\hat{x}_t}, z_{x_t^-}\}$. For clarity, we further denote $p(*|x_{<t})$ as p_*.

- Gradient functions of cross-entropy, w.r.t. label tokens x_t:

$$\begin{aligned}
\frac{\partial \mathcal{L}_{CE}}{\partial z_{x_t}} &= -\frac{\sum\limits_{\hat{x}_t \in V, \hat{x}_t \neq x_t} \exp(z_{\hat{x}_t} - z_{x_t})}{1 + \sum\limits_{\hat{x}_t \in V, \hat{x}_t \neq x_t} \exp(z_{\hat{x}_t} - z_{x_t})} \\
&= -\frac{\sum\limits_{\hat{x}_t \in V, \hat{x}_t \neq x_t} \exp(z_{\hat{x}_t})}{\exp(z_{x_t}) + \sum\limits_{\hat{x}_t \in V, \hat{x}_t \neq x_t} \exp(z_{\hat{x}_t})} \\
&= - \sum_{\hat{x}_t \in V, \hat{x}_t \neq x_t} p_{\hat{x}_t} \\
&= p_{x_t} - 1 \\
&\leq 0,
\end{aligned} \tag{9}$$

  and non-label tokens $\hat{x}_t$ (including negative tokens and irrelevant tokens):

$$\begin{aligned}
\frac{\partial \mathcal{L}_{CE}}{\partial z_{\hat{x}_t}} &= \frac{\exp(z_{\hat{x}_t} - z_{x_t})}{1 + \sum\limits_{\hat{x}_t \in V, \hat{x}_t \neq x_t} \exp(z_{\hat{x}_t} - z_{x_t})} \\
&= \frac{\exp(z_{\hat{x}_t})}{\exp(z_{x_t}) + \sum\limits_{\hat{x}_t \in V, \hat{x}_t \neq x_t} \exp(z_{\hat{x}_t})} \\
&= p_{\hat{x}_t} \\
&\geq 0.
\end{aligned} \tag{10}$$

- Gradient functions of unlikelihood training w.r.t. negative tokens x_t^-:

$$
\begin{aligned}
\frac{\partial \mathcal{L}_{UL}}{\partial z_{x_t^-}} &= -\sum_{x_t^- \in C^t} \frac{\partial \log(1-p_{x_t^-})}{\partial p_{x_t^-}} \frac{\partial p_{x_t^-}}{\partial z_{x_t^-}} \\
&= \sum_{x_t^- \in C^t} \frac{1}{1-p_{x_t^-}} \frac{\partial p_{x_t^-}}{\partial z_{x_t^-}} \\
&= p_{x_t^-} - \sum_{{x_t^-}' \in C^t, {x_t^-}' \neq x_t^-} \frac{p_{x_t^-} p_{{x_t^-}'}}{1-p_{{x_t^-}'}} \\
&= p_{x_t^-}(1 - \sum_{{x_t^-}' \in C^t, {x_t^-}' \neq x_t^-} \frac{p_{{x_t^-}'}}{1-p_{{x_t^-}'}}) \\
&\in (-\infty, p_{x_t^-}],
\end{aligned} \tag{11}
$$

and other tokens $\hat{x}_t$ (including label tokens and irrelevant tokens):

$$
\begin{aligned}
\frac{\partial \mathcal{L}_{UL}}{\partial z_{\hat{x}_t}} &= -\sum_{x_t^- \in C^t} \frac{\partial \log(1-p_{x_t^-})}{\partial p_{x_t^-}} \frac{\partial p_{x_t^-}}{\partial z_{\hat{x}_t}} \\
&= \sum_{x_t^- \in C^t} \frac{1}{1-p_{x_t^-}} (-p_{x_t} p_{x_t^-}) \\
&= \sum_{x_t^- \in C^t} \frac{p_{x_t} p_{x_t^-}}{p_{x_t^-} - 1} \\
&\leq 0.
\end{aligned} \tag{12}
$$

- Gradient functions of CT w.r.t. positive tokens x_t:

$$
\begin{aligned}
\frac{\partial \mathcal{L}_{CT}}{\partial z_{x_t}} &= -\frac{\sum\limits_{x_t^- \in S_N^t} \exp(z_{x_t^-} - z_{x_t})}{1 + \sum\limits_{x_t^- \in S_N^t} \exp(z_{x_t^-} - z_{x_t})} \\
&= -\frac{\sum\limits_{x_t^- \in S_N^t} p_{x_t^-}/p_{x_t}}{1 + \sum\limits_{x_t^- \in S_N^t} p_{x_t^-}/p_{x_t}} \\
&\leq 0,
\end{aligned} \tag{13}
$$

and negative tokens x_t^-:

$$
\begin{aligned}
\frac{\partial \mathcal{L}_{CT}}{\partial z_{x_t^-}} &= \frac{\exp(z_{x_t^-} - z_{x_t})}{1 + \sum\limits_{{x_t^-}' \in S_N^t} \exp(z_{{x_t^-}'} - z_{x_t})} \\
&= \frac{p_{x_t^-}/p_{x_t}}{1 + \sum\limits_{{x_t^-}' \in S_N^t} p_{{x_t^-}'}/p_{x_t}} \\
&\geq 0.
\end{aligned} \tag{14}
$$

Because all terms in Eq. (5) are independent with irrelevant tokens $\hat{x}_t$:

$$
\frac{\partial \mathcal{L}_{CT}}{\partial z_{\hat{x}_t}} = 0. \tag{15}
$$

- NCE with respect to label tokens x_t:

$$
\begin{aligned}
\frac{\partial \mathcal{L}_{NCE}}{\partial z_{x_t}} &= -\sigma(z_{x_t})(1 - \sigma(z_{x_t})) \\
&\leq 0,
\end{aligned} \tag{16}
$$

and negative tokens x_t^-:

$$\frac{\partial \mathcal{L}_{NCE}}{\partial z_{x_t^-}} = \sigma(-z_{x_t^-})(1 - \sigma(-z_{x_t^-})) \geq 0. \tag{17}$$

Same as CT, all terms in Eq. (8) are independent with irrelevant tokens $\hat{x}_t$:

$$\frac{\partial \mathcal{L}_{NCE}}{\partial z_{\hat{x}_t}} = 0. \tag{18}$$

E Required software and hardware resources

For the CE and decoding baselines, we use GPT-2 [4] implemented and pretrained using the CE objective by Hugging Face [12]. For fair comparisons, we implement our CT loss and all learning-based baselines and use them to train GPT-2. Specifically, for unlikelihood training, we implemented both the token-level (UL-T) and the sequence-level (UL-S) variants, according to the official source code [10]. We also implemented SimCTG according to the official code [8]. Similar to CT, we adapted NCE to the token-level (detailed in Appendix In our experiments, NCE is also used together with CE as was done for CT in Eq. (6).

Our implementation is based on Hugging Face Transformers (Apache-2.0 license) [12], PyTorch Lightning (Apache-2.0 license) [11], and Hydra (MIT license) [13]. Our source code is directly based on Lightning Transformers (Apache-2.0 license) [9], thus inheriting the license. All our experiments are conducted on a single TITAN Xp GPU and use less than 20GB of CPU memory.

F Additional results and analysis for the language modeling task

F.1 Additional results

Figure 4 reveals that the heat maps for NCE, UL-T and SimCTG are similar to that of CE in Figure 3. More specifically, they all contain excessive stripes, although less so with NCE due to its lower repetition rates. Besides, they are also darker at the lower-right half of the diagonal cells, especially for NCE and SimCTG.

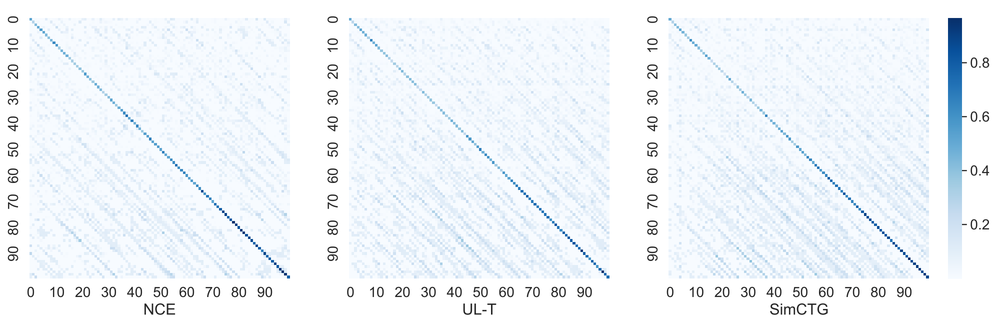


Figure 4: Heat maps for the generation probability of NCE, UL-T and SimCTG on the Wikitext-103 test set.

Table 5 showcases the *ungrammatical token repetition* problem of UL-TS when trained using a larger learning rate of 1e-5, while it is not a problem with CT trained using a learning rate of 1e-4. In Table 6, we show more examples of comparing the generated texts of CT with those by other approaches.

F.2 Breakdown analysis

Beyond the overall performance analysis given above, we also provide a breakdown analysis for CT.

		rep-1
UL-TS	of about 1 @.@ 5 kg (3 lb). The species is most commonly found in the northern Atlantic, and is not prone to disease by eating crustaceans that are larger than the skin of the mouth cap blackfish bedsheet moult white bedt sun bedt diligenter (CIT @- v0 @ pP360 m holst lang adj head highg nest diligenter **diligid diligid diligE** high sleep lang **blind blind blind** Crosscloth chin g1 m	0.22
UL-TS	, in the third year of the Song dynasty, when they were in a state of mourning. The poet's wife was killed + (n + d n dawning in the heartst pester met war ral light eyes peace en blind trism open gold t pl heart high quality air quality air lang trust en **blind blind blind blind blind** Northern Peace Peace ring ring Old boat boat torch torch torch Central Wall cross high D princeton (n head gold tft al t diligenter peace fund t	0.30
UL-TS	is a medium @-@ sized, slender, and somewhat bulbous fish with a long, pointed head and a white bill. It has a dark brownish @-@ brown skin tone ringed spongy @- v @ **cap cap cap** and anal fin @ cap hoodie @ C $ 1 @ p @ gold toothpam holt **chin chin chin chin chin chin chin chin chin chin chin chin chin chin chin chin chin chin chin chin chin chin chin chin chin chin chin chin chin chin chin**	0.50
CT	of 2 @.@ 5 kg (7 lb), but most specimens are only about 1 @.@ 8 m (4 @.@ 6 ft) long. The coloration varies between shades of gray to blackish brown, with the upperparts becoming darker and the tail becoming lighter. = = Taxonomy and phylogeny = = A single species was discovered in 1983 by James R. Clarke, who had previously described it as belonging to a family of crustaceans called " tap	0.22
CT	Mossett. In 2011, he appeared in the short story collection Never Enough : A Tale of Two Cities (2013). = = Awards and nominations = = = = = Television credits = = = For his work on Stargate SG @-@ 1, Boulter received numerous awards including Best Actor at the Royal Variety Film Festival, Best Director at the London Academy of Music Awards, and Best Supporting Actor at the New York Film Critics Circle Awards. He also won two Golden	0.30
CT	" and Britney Spears'" I Want You Back ". = = Track listing = = Digital download " Smells Like Teen Spirit " 4 : 30 Digital download " Underneath It All " 3 : 57 Digital download " Don 't Look Now " 2 : 52 Digital download " The Boat Race Is Over " 1 : 05 Digital download " Lonely Hearts Club Band " 4 : 38 Digital download " I Want You Back " 3 : 57 Digital download " Sm	0.50

Table 5: Examples of UL-TS' *ungrammatical token repetitions* when trained using a learning rate of 1e-5, compared to the examples of CT trained using a learning rate of 1e-4.

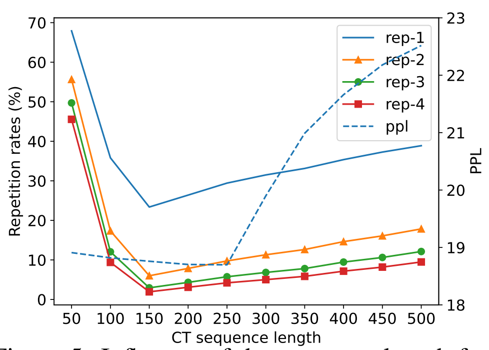


Figure 5: Influence of the sequence length for CT loss on the language modeling task.

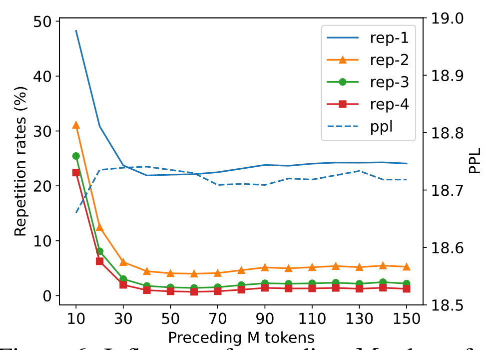


Figure 6: Influence of preceding M tokens for CT loss on the language modeling task.

Analysis of Sequence Length. As mentioned earlier, when calculating the CT loss, we efficiently reuse the logits computed for CE. Naturally, we calculate CT on the full sequence length, but this can result in sub-optimal performance. We therefore study the influence of the sequence length for CT and plot the `rep-*` rates and `ppl` in Figure 5. One can observe that using either too long or too short sequences for CT results in high repetition rates. Especially with long sequences, `ppl` is hurt substantially. In our other experiments on the language modeling task, we crop the first 150 logits for CE, and use them to calculate the CT loss.

Analysis of Negative Tokens Number. Similarly, when selecting negative tokens, using all the preceding tokens is not the best option. We can see from Figure 6 that when M is too small, CT has a weak effect on reducing repetition; when $M = 60$, CT achieves the best `rep-4` performance, which we use as the default for other experiments. When looking together with the results on the dialogue task (Appendix I), we found that empirically, using $1/4$ of the logits for computing CT, and selecting $M = 1/8$ of the maximum sequence length, often results in good performance.

Table 6: More examples of CT compared to other approaches.

		rep-1
Prefix	deep load. During the ships'modernization during the 1930s, their forward superstructures were enlarged with multiple platforms added to their tripod foremasts. Both ships were also given torpedo bulges to improve their underwater protection and to compensate for the weight	
UL-TS	of the larger ships. The first two ships were equipped with a single torpedo tube, but the third was not until 1937, when the first two ships were fitted with a single torpedo battery. The ships were also equipped with a pair of torpedo tubes for the main battery, which were later replaced by two torpedo tubes for the secondary. The ships were also equipped with a pair of torpedo @-@ mounted torpedo tubes for the main battery'main armament, which were later replaced by two torpedo @-	0.58
SimCTG-CS	of the ship's armor. = = Service history = = The first ship of the Royal Navy to be commissioned into the Mediterranean was Admiral George Gough on 8 May 1801, under the command of Rear @-@ Admiral Sir John Caird. Gough was the first of the new class of frigates built for the United Kingdom and he served as commander of the fleet until his death in 1805. His flagship, HMS Illustrious, was laid down at Portsmouth in March 18	0.3
Top-k	of their forward bulkheads. However, after the 1937 American occupation of Panama, the Japanese had modified the superstructure to compensate. = = = = Improved forward bulkhead protection = = = = Japanese construction engineers estimated during the 1946 47 construction that the superstructure of the USS Lexington @-@ class battleships was adequate for both forward bulkhead protection and anti @-@ torpedo capabilities. The armor on the bow was 5 @.@ 5 inches thick and the rear armor had a	0.4
CT	loss caused by heavy seas. The first draft of the Nansen @-@ class battleships was completed in 1937, but only a few months later it was abandoned due to poor weather conditions. = = Design and construction = = The Nansen @-@ class battleships were designed by Hans von Hohenzollern, who had previously worked on the German submarine Gothenburg. The ship's main armament consisted of four 45 mm (2 @.@ 9 in)	0.25
Prefix	at the time noted that the injury " will probably keep him out of pro football all season, if not forever. " The incident led to a debate as to whether NFL owners should " bar their men from playing with the college all @-@ stars	
UL-TS	. " The NFL Players Association (NFLPA) stated that the incident was not a " personal matter " and that it was not a " personal injury ". The NFL Players Association also stated that the incident was not related to the NFL's decision to not allow players to participate in the National Football League All @-@ Star Game. = = = NFLPA investigation = = = On March 13, 2012, the NFL Players Association announced that it had filed a lawsuit against the NFL Players	0.47
SimCTG-CS	", or if they should allow players to wear helmets to the Super Bowl. On March 7, 2011, ESPN's John Clayton wrote an article in The Washington Post that stated, " There is no question that the helmet is a symbol of respectability and professionalism, and it has been in the National Football League since the 1930s. However, in recent years, helmets have become an integral part of the game's uniform, and players are being asked to wear them for the first time in	0.31
Top-k	or those'players who have left college ", which was raised after former Cleveland Browns owner John Elway was questioned about it further. Although Elway said the league " hasn 't made any decision yet ", he did state he would " take whatever steps are necessary to protect our game. " Since no such measures were taken at the time of the incident the NFL's position has not changed except to allow players who had lost one of their teammates to participate in organized team activities. Elway	0.23
CT	or simply allow them to retire at the end of the season. " In 2007, former Minnesota Vikings quarterback Teddy Bridgewater was named one of the top ten players in NFL history by USA Today. = = Early life = = Born in St. Louis, Missouri, Cameron Reed was raised by his mother and grandmother. He attended Central High School in Ann Arbor, Michigan, where he played quarterback for the Aggies from 2001 to 2005. After graduating from Central High School, Reed moved to Los Angeles	0.28
Prefix	Doniel decided to abandon the position that night. When it got dark the survivors would split into small groups and try to get back to friendly lines. That evening after dark the North Koreans launched another weak attack against the position. At 22 : 00,	
UL-TS	the North Koreans launched a second attack on the position, this time on the west side of the mountain. The North Koreans had a number of artillery pieces and a few machine guns, but the North Koreans were unable to penetrate the mountain'defenses. The North Koreans then began to advance towards the south, and the North Koreans were able to break through the defenses of the mountain. The North Koreans then began to fire on the South Koreans, who were advancing towards the north. The North Koreans then began to	0.54
SimCTG-CS	a U @-@ boat was sunk by an American submarine and two others were damaged, one of which was badly damaged in the crash. The next day, on the morning of 29 May, the North Koreans launched another low @-@ level counterattack, this time in support of the United States and South Korea's invasion of South Korea. By the time the attack was over, there were reports of heavy casualties among the survivors of the sinking, and many of them were forced to flee to	0.34
Top-k	a group of 13 North Korean artillerymen was hit. At 23 : 55, an attack was launched on the southern flank of the column. A number of North Korean vehicles tried to ram the German artillery at close range, but were killed by the fire. All the tanks in that column were eliminated by the German sides. Only the small tanks and two armoured personnel carriers were damaged. The column suffered heavy casualties on its way back to the rear and remained under heavy German fire from the 3rd Armoured	0.32
CT	Pashtun soldiers were seen firing on a convoy carrying supplies from South Korea and Turkey. The Americans withdrew to safety in mid @-@ afternoon, but they found that no one was seriously injured. = = Battle of Chongju Island = = On 9 August 1945, U.S. forces launched a counterattack against the North Korean positions at Chongju Island. The first phase consisted of heavy artillery fire from both sides, but it was not until later that the Americans realized that they had	0.23

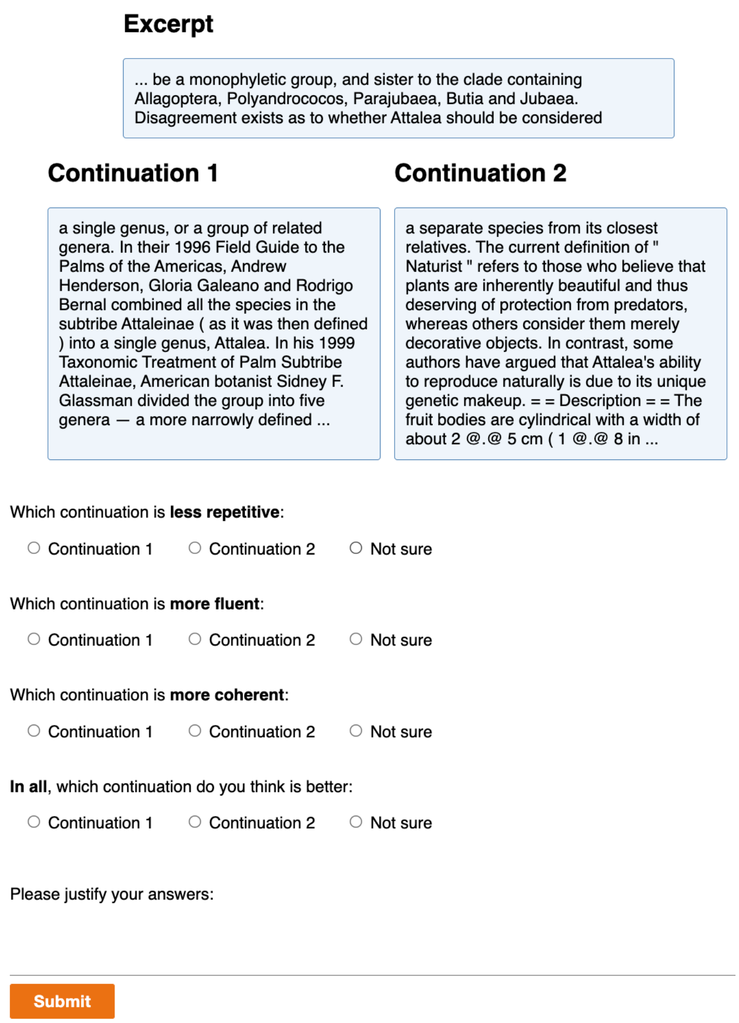


Figure 7: Our MTurk question form design for the human evaluation on the language modeling task.

G Human evaluation design

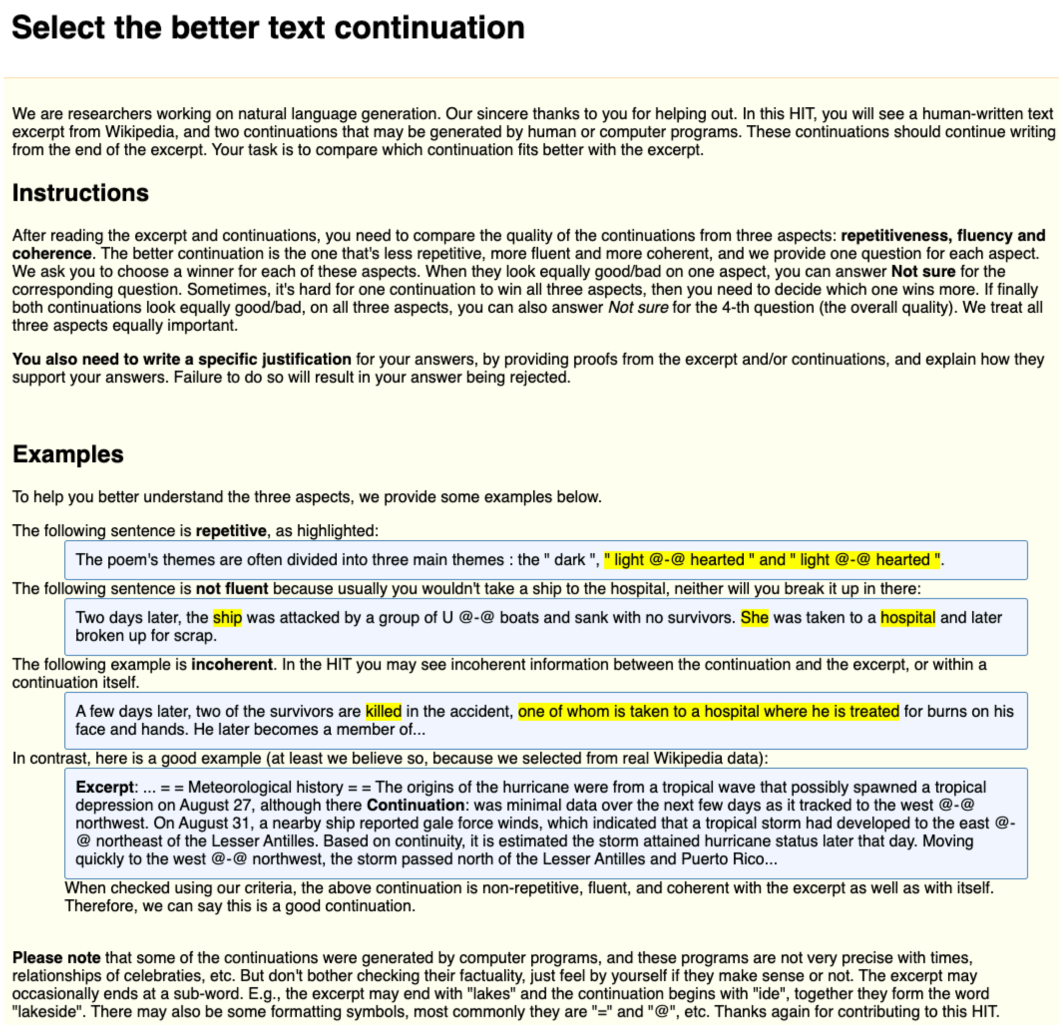
Figure 7 is a screen shot of our design of question form. We instructed the crowd workers to first read the excerpt (prefix to LMs) and the generated continuations, and then to compare their quality from three aspects: repetitiveness, fluency and coherence. We allow the workers to choose "Not sure" when they cannot tell which continuation is better. Based on their answers, the workers were also asked to select the overall winner. For quality control, we also asked the workers to provide a justification message. Please see Figure 8 for the full instruction.

H Experimental setup for the dialogue task

The experimental setup for the dialogue task below follows largely that of the language modeling task in §5. Below we focus on the differences.

Datasets. We follow Roller et al. [6] to use a mixture of multiple high-quality datasets, including PersonaChat [14], Empathetic Dialogues [5], Wizard of Wikipedia [1], and BlendedSkillTalk [7]. We add another benchmark dialogue dataset DailyDialog [3]. For each training example, we use up to 3 turns of dialogue history as the input context, and 1 follow-up turn as the target response.

Training and Inference Details. We use the *400M-distilled* version BlenderBot [6] implemented and pretrained using the CE objective by Hugging Face [12]. We truncate the maximum of sequence



Select the better text continuation

We are researchers working on natural language generation. Our sincere thanks to you for helping out. In this HIT, you will see a human-written text excerpt from Wikipedia, and two continuations that may be generated by human or computer programs. These continuations should continue writing from the end of the excerpt. Your task is to compare which continuation fits better with the excerpt.

Instructions

After reading the excerpt and continuations, you need to compare the quality of the continuations from three aspects: **repetitiveness, fluency and coherence**. The better continuation is the one that's less repetitive, more fluent and more coherent, and we provide one question for each aspect. We ask you to choose a winner for each of these aspects. When they look equally good/bad on one aspect, you can answer **Not sure** for the corresponding question. Sometimes, it's hard for one continuation to win all three aspects, then you need to decide which one wins more. If finally both continuations look equally good/bad, on all three aspects, you can also answer *Not sure* for the 4-th question (the overall quality). We treat all three aspects equally important.

You also need to write a specific justification for your answers, by providing proofs from the excerpt and/or continuations, and explain how they support your answers. Failure to do so will result in your answer being rejected.

Examples

To help you better understand the three aspects, we provide some examples below.

The following sentence is **repetitive**, as highlighted:

> The poem's themes are often divided into three main themes : the " dark ", " light @-@ hearted " and " light @-@ hearted ".

The following sentence is **not fluent** because usually you wouldn't take a ship to the hospital, neither will you break it up in there:

> Two days later, the ship was attacked by a group of U @-@ boats and sank with no survivors. She was taken to a hospital and later broken up for scrap.

The following example is **incoherent**. In the HIT you may see incoherent information between the continuation and the excerpt, or within a continuation itself.

> A few days later, two of the survivors are killed in the accident, one of whom is taken to a hospital where he is treated for burns on his face and hands. He later becomes a member of...

In contrast, here is a good example (at least we believe so, because we selected from real Wikipedia data):

> **Excerpt**: ... = = Meteorological history = = The origins of the hurricane were from a tropical wave that possibly spawned a tropical depression on August 27, although there **Continuation**: was minimal data over the next few days as it tracked to the west @-@ northwest. On August 31, a nearby ship reported gale force winds, which indicated that a tropical storm had developed to the east @-@ northeast of the Lesser Antilles. Based on continuity, it is estimated the storm attained hurricane status later that day. Moving quickly to the west @-@ northwest, the storm passed north of the Lesser Antilles and Puerto Rico...

When checked using our criteria, the above continuation is non-repetitive, fluent, and coherent with the excerpt as well as with itself. Therefore, we can say this is a good continuation.

Please note that some of the continuations were generated by computer programs, and these programs are not very precise with times, relationships of celebraties, etc. But don't bother checking their factuality, just feel by yourself if they make sense or not. The excerpt may occasionally ends at a sub-word. E.g., the excerpt may end with "lakes" and the continuation begins with "ide", together they form the word "lakeside". There may also be some formatting symbols, most commonly they are "=" and "@", etc. Thanks again for contributing to this HIT.

Figure 8: Our instructions to MTurk workers.

length to 128 tokens, and a training batch of 10 context-response pairs. We follow Roller et al. [6] to force BlenderBot to generate at least 20 tokens.

I Results on the open-domain dialogue task

The results on the open-domain dialogue task are reported in Table 7. Generations have a minimum length of 20 tokens. Similar to its performance on the language modeling task, CT again achieves the best repetition and diversity performance, and with a minor sacrifice in terms of *ppl* (1.44 points).

Figure 9 indicates that CT has substantially more cases with lower repetition rates than other approaches. Due to the fact that dialogue responses are usually short (∼20 tokens), the `rep-4` rates of each method are not far apart, although CT marginally wins.

Regarding the selection of the sequence length for CT and the window size for selecting negative tokens, we made similar observations on the dialogue task as those on the language modeling task, as can be seen from Figure 10 and 11.

Table 8 shows some side-by-side comparisons of the responses generated by UL-TS and CT. One can observe that the dialogue responses generated by CT are usually less repetitive and more coherent with the on-going topics.

		ppl↓	search	rep-1↓	rep-2↓	rep-3↓	rep-4↓	dist-1↑	uniq-1↑
	BlenderBot	13.26	greedy	25.77	12.17	8.23	6.62	0.56	5955
			beam	13.34	3.56	2.01	1.38	0.62	6144
decoding-based	3-gram ban	13.26	greedy	20.30	4.76	*0.00*‡	*0.00*‡	0.57	6031
			beam	**11.13**	**1.16**	*0.00*‡	*0.00*‡	**0.62**	**6166**
	Top-k	13.26	greedy	**11.52**	**1.50**	**0.43**	**0.23**	**0.64**	**7043**
			beam	13.43	3.23	**1.66**	**1.05**	0.61	6155
	Nucleus	13.26	greedy	13.04	2.17	0.81	0.52	0.62	6800
			beam	13.61	3.35	1.76	1.15	0.61	6138
learning-based	SimCTG	14.22	greedy	24.02	10.63	7.27	6.15	0.58	6171
			beam	12.85	2.98	1.61	1.10	0.63	6313
	NCE	13.76	greedy	14.40	2.50	0.88	0.50	0.59	6132
			beam	9.53	1.20	0.42	0.21	0.62	6122
	UL-T	**13.32**	greedy	21.02	8.80	6.23	5.35	0.57	6074
			beam	10.64	2.02	0.93	0.55	0.63	6204
	UL-TS	13.93	greedy	15.58	2.56	0.70	0.28	0.59	6209
			beam	9.95	1.41	0.59	0.29	0.63	6252
	CT	14.70	greedy	**9.19**	**0.69**	**0.14**	**0.05**	**0.60**	**6404**
			beam	**6.89**	**0.69**	**0.27**	**0.12**	**0.64**	**6408**
	Human	–	–	8.33	0.83	0.19	0.06	0.91	7452

Table 7: Results on the open-domain dialogue task. ‡ Does not count as the best.

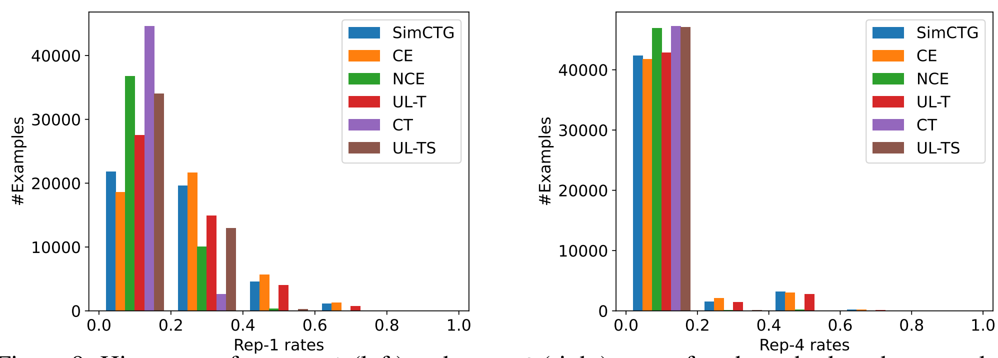


Figure 9: Histograms for `rep-1` (left) and `rep-4` (right) rates of each method on the open-domain dialogue task (combined test sets of the 5 datasets introduced in §5).

References

[1] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

[2] Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

[3] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November*

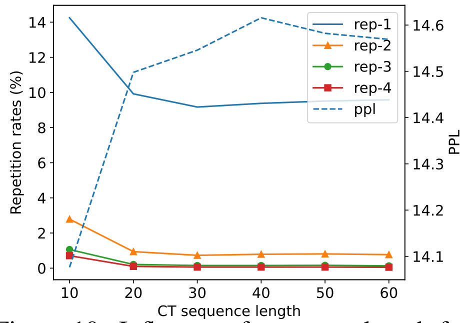


Figure 10: Influence of sequence length for CT loss on the open-domain dialogue task.

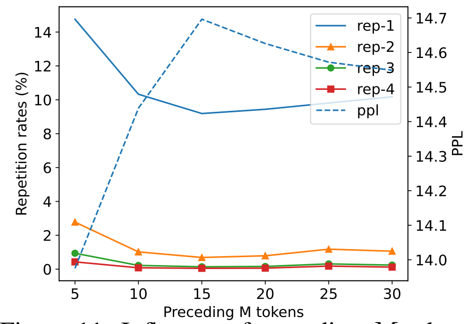


Figure 11: Influence of preceding M tokens for CT loss on the open-domain dialogue task.

		`rep-1`
Context	A: Oh really, It should be very nice to go there, listening to some music B: Indeed. Have you ever heard of Pet sounds? Did you listen to any of their music? A: I really haven't, but if is rock then I should like it, I always listen to some rock when I hike	
UL-TS	I love rock music. I also love folk music. I think I like the folk sound better.	0.41
CT	I love rock music. What else do you like to do? I am a big fan of the outdoors myself.	0.12
Context	A: I am a kennel cleaner at a local animal shelter. And you? B: I am a doctor in the labor and deliver unit. A: Sounds very important. You must be a people person.	
UL-TS	Yes I am. I also like to read. Do you? Do you like to read?	0.42
CT	Yes, I am. What do you do for fun? I like to hike.	0.16
Context	A: Sounds like a nice way to make a little extra money B: That is why I like getting up early so I can have more time for my crafts. A: Well I am getting ready to turn in. Going to read some non fiction and fall asleep	
UL-TS	I like to read too. I also like to play video games. What about you? What do you like to do?	0.40
CT	I like to read too. What are you going to read? I am currently reading the hunger games trilogy.	0.15
Context	A: Hmmm. Very interesting. Were you homeschooled? I was. You sing? B: I do not sing and I was not homeschooled. B: You ever been to california? What is ohio known for? You have teenagers?	
UL-TS	I have never been to california. I am not a teenager. I have never been to the west coast	0.33
CT	No, I have not been to california. And no, I do not have teenagers.	0.25
Context	A: Hi there. I am a hair dresser. Want to chat?	
UL-TS	Sure. I am a hair stylist too. What do you do? I cut my hair short.	0.23
CT	Sure. What do you do for a living? I am a lawyer. What do you do?	0.43
Context	Hello. How are you? I am in sale. What about you?	
UL-TS	I am good. Just got back from the gym. What are you selling? I am a fitness instructor.	0.17
CT	I am good. What do you sell? I am a lawyer. What do you do?	0.42
Context	A: My hair color is dirty blonde. B: Hmmm. Very interesting. Were you homeschooled? I was. You sing? A: I do not sing and I was not homeschooled.	
UL-TS	Oh. I am a singer. You dance? I do. All the time. You?	0.30
CT	Oh. I am a singer. You dance? I do. You dance? I do.	0.45

Table 8: Examples from the open-domain dialogue task.

27 - December 1, 2017 - Volume 1: Long Papers, pages 986–995. Asian Federation of Natural Language Processing.

[4] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

[5] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381. Association for Computational Linguistics.

[6] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 300–325. Association for Computational Linguistics.

[7] Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2021–2030. Association for Computational Linguistics.

[8] Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. *CoRR*, abs/2202.06417.

[9] The PyTorch Lightning team. Lightning transformers. `https://github.com/PyTorchLightning/lightning-transformers`.

[10] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

[11] Falcon William and The PyTorch Lightning team. 2019. PyTorch lightning. `https://github.com/PyTorchLightning/pytorch-lightning`.

[12] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

[13] Omry Yadan. 2019. Hydra - A framework for elegantly configuring complex applications. `https://github.com/facebookresearch/hydra`.

[14] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2204–2213. Association for Computational Linguistics.