A. Derivations

Derivation of the gradient of loss *w.r.t.* logit We follow the same notation as in the main paper. At time step t, assuming that the pre-softmax scores (*i.e.,* logits) are denoted as $\boldsymbol{o}^t$ over the vocabulary $\mathbb{V}$, where o_i^t denotes the score for the token with index i in the vocabulary. Similarly, we have $p_i^t = [\mathrm{softmax}(\boldsymbol{o}^t)]_i$. Let k denote the index of the ground truth token at step t.

The cross entropy loss at step t is given as (we omit t for notational simplicity):

$$\mathcal{L} = -\sum_i y_i \log p_i \tag{8}$$

where $y_i = 1$ if $i = k$, otherwise $y_i = 0$. Thus the loss function can be rewritten as:

$$\mathcal{L} = -\log p_k = -\log(\frac{e^{o_k}}{\sum_j e^{o_j}}) = \log(\sum_j e^{o_j}) - o_k \tag{9}$$

Therefore, we can derive the partial derivative of the loss *w.r.t.* the logit o_i as follows.

$$\begin{aligned}\nabla_{o_i}\mathcal{L} &= \nabla_{o_i}\log(\sum_j e^{o_j}) - \nabla_{o_i} o_k \\ &= \frac{1}{\sum_j e^{o_j}} \cdot \nabla_{o_i}(\sum_j e^{o_j}) - \mathbb{1}(i=k) \\ &= \frac{e^{o_i}}{\sum_j e^{o_j}} - \mathbb{1}(i=k) \\ &= p_i - \mathbb{1}(i=k)\end{aligned} \tag{10}$$

Derivation of the gradient of ScaleGrad *w.r.t.* logit We first denote $\mathbb{S}_{\text{novel}}$ as the novel token set at current time step and $\mathbb{V}' = \mathbb{V} \setminus \mathbb{S}_{\text{novel}}$.

Suppose the current target token belongs to the novel token set, *i.e.,* $k \in \mathbb{S}_{\text{novel}}$. The scaling equation for target token can be rewritten into the function of logits as follows.

$$\begin{aligned}\tilde{p}_k &= \frac{\gamma \cdot p_k}{\gamma \sum_{j\in\mathbb{S}_{\text{novel}}} p_j + \sum_{j\in\mathbb{V}'} p_j} \\ &= \frac{\gamma \cdot \frac{e^{o_k}}{\sum_{m\in\mathbb{V}} e^{o_m}}}{\gamma \sum_{j\in\mathbb{S}_{\text{novel}}} \frac{e^{o_j}}{\sum_{m\in\mathbb{V}} e^{o_m}} + \sum_{j\in\mathbb{V}'} \frac{e^{o_j}}{\sum_{m\in\mathbb{V}} e^{o_m}}} \\ &= \frac{\gamma \cdot e^{o_k}}{\gamma \sum_{j\in\mathbb{S}_{\text{novel}}} e^{o_j} + \sum_{j\in\mathbb{V}'} e^{o_j}}\end{aligned} \tag{11}$$

For the notational simplicity, we notate $a = (\gamma \sum_{j\in\mathbb{S}_{\text{novel}}} e^{o_j} + \sum_{j\in\mathbb{V}'} e^{o_j})$. The loss function can be rewritten accordingly as:

$$\begin{aligned}\mathcal{L} &= -\log(\tilde{p}_k) \\ &= -\log\frac{\gamma \cdot e^{o_k}}{a} = \log a - \log(\gamma \cdot e^{o_k})\end{aligned} \tag{12}$$

We thus have the gradient of the SG loss *w.r.t.* the logit (o_i) as follows:

$$\begin{aligned}\nabla_{o_i}\mathcal{L} &= \nabla_{o_i}\log a - \nabla_{o_i}\log(\gamma \cdot e^{o_k}) \\ &= \frac{1}{a}\cdot\nabla_{o_i} a - \frac{1}{\gamma \cdot e^{o_k}}\cdot\nabla_{o_i}(\gamma \cdot e^{o_k}) \\ &= \frac{1}{a}\cdot(\gamma \cdot e^{o_i}\mathbb{1}(i\in\mathbb{S}_{\text{novel}}) + e^{o_i}\mathbb{1}(i\in\mathbb{V}')) \\ &\quad - \mathbb{1}(i=k) \\ &= \begin{cases} \dfrac{\gamma \cdot e^{o_k}}{a} - 1, & \text{if } i = k \text{ and } i \in \mathbb{S}_{\text{novel}} \\ \dfrac{\gamma \cdot e^{o_i}}{a}, & \text{if } i \neq k \text{ and } i \in \mathbb{S}_{\text{novel}} \\ \dfrac{e^{o_k}}{a} - 1, & \text{if } i = k \text{ and } i \notin \mathbb{S}_{\text{novel}} \\ \dfrac{e^{o_i}}{a}, & \text{if } i \neq k \text{ and } i \notin \mathbb{S}_{\text{novel}} \end{cases} \\ &= \begin{cases} \lambda_i \cdot p_k - 1, & \text{if } i = k \text{ and } i \in \mathbb{S}_{\text{novel}} \\ \lambda_i \cdot p_i, & \text{if } i \neq k \text{ and } i \in \mathbb{S}_{\text{novel}} \\ \alpha_i \cdot p_k - 1, & \text{if } i = k \text{ and } i \notin \mathbb{S}_{\text{novel}} \\ \alpha_i \cdot p_i, & \text{if } i \neq k \text{ and } i \notin \mathbb{S}_{\text{novel}} \end{cases}\end{aligned} \tag{13}$$

Similarly, it is easy to derive the same results when current target token does not belong to the novel token set.

B. Novel token set illustration

Figure 4 shows an example of how the novel token set changes when the model is learning to predict the sentence "people who are interested ..". At beginning, the novel token set $\mathbb{S}_{\text{novel}}$ is equivalent to the vocabulary $\mathbb{V}$. The size of the novel token set shrinks as the decoding proceeds.

C. Undesired property of UL training

We use the same notation as Welleck et al. (2020) to explain the undesired UL property. From their paper (page 4):

With a single negative candidate, the (negative) gradient is:

$$\begin{aligned}&\nabla\mathcal{L}_a = x^* - m \odot p, \\ &\quad\text{where } m = \begin{cases}(1 - \alpha\dfrac{p_{\text{neg}}}{1-p_{\text{neg}}}) & \text{if } i \neq i_{\text{neg}} \\ (1+\alpha) & \text{if } i = i_{\text{neg}}\end{cases}\end{aligned} \tag{14}$$

where $x^* \in \{0,1\}^{\mathcal{V}}$ is a one-hot ground-truth vector, $m \in \mathbb{R}^{\mathcal{V}}$, $p = p_\theta(\cdot|x_{<t})$, and p_{neg} is the probability of the negative candidate at index i_{neg}.

As the paper says (page 5):

".... At the ground-truth token index i^*, the unlikelihood gradient is positive, increasing the ground-truth token's prob-

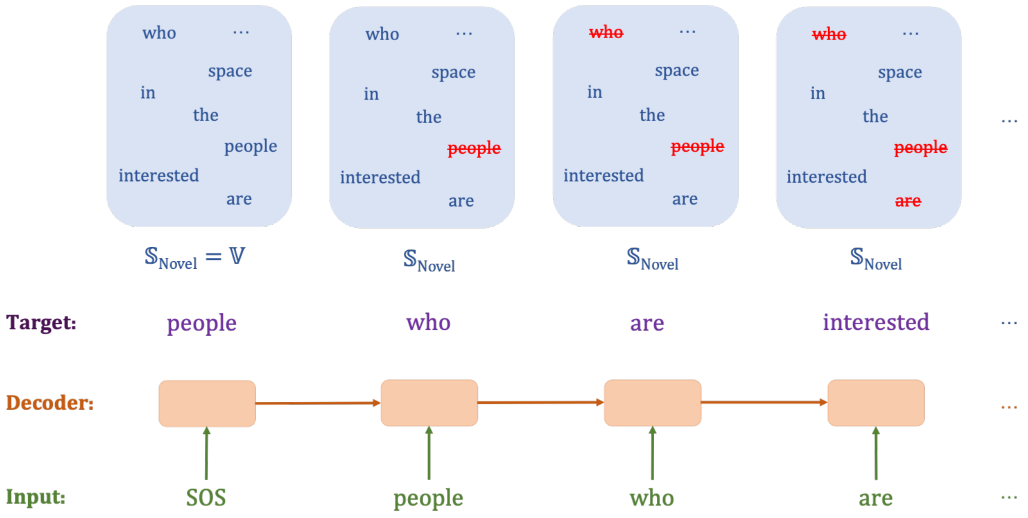

who ⋯ space in the people interested are
$\mathbb{S}_{\text{Novel}} = \mathbb{V}$
who ⋯ space in the ~~people~~ interested are
$\mathbb{S}_{\text{Novel}}$
~~who~~ ⋯ space in the ~~people~~ interested are
$\mathbb{S}_{\text{Novel}}$
~~who~~ ⋯ space in the ~~people~~ interested ~~are~~
$\mathbb{S}_{\text{Novel}}$
⋯
Target:
people
who
are
interested
⋯
Decoder:
⋯
Input:
SOS
people
who
are
⋯

Figure 4: An illustration of how the novel token set changes as decoding proceeds for the sentence "people who are interested ...". The words marked in purple are the target words that the model is learning to predict at each decoding step.

ability with a magnitude that grows with p_{neg}. Conversely, at the negative candidate index i_{neg} the gradient is negative. At all other token indices $i \notin \{i^*, i_{\text{neg}}\}$, the gradient moves from negative to positive as p_{neg} increases. For instance, with $\alpha = 1.0$ the gradient increases the probability of each token x_i when the model assigns high probability to the negative candidate ($p_{\text{neg}} > 0.5$). ”

We notice that at the ground-truth token index i^*, with $\alpha = 1.0$ and $p_{\text{neg}} > 0.5$, the gradient norm is $|\nabla \mathcal{L}_a| = 1 + |m| \cdot p^*$. The model will therefore decrease p^* to reduce $|\nabla \mathcal{L}_a|$, which is against our optimization principle.

D. Human evaluation details

We conduct the human evaluation for two pairs of systems *i.e.,* SG vs. MLE and SG vs. UL. For each pair, the models generate their own continuations based on the same 100 randomly chosen prefixes. Two native speakers of English are then asked to evaluate the generated texts independently. During the study, users are instructed to judge which generated text is a better continuation of the prefix based on the overall quality (*e.g.,* readability, relevance to the prefix, grammar, and fluency).

The **Win Rate** in Table 3 is calculated as the total number of times that two users prefer the texts produced by the winner divided by the total number of cases in the evaluation ($2 \times 100 = 200$). To get a reliable human study, we also compute the percentage agreement and the chance correlated measure, Gwet's AC1/gamma coefficient (Gwet, 2008) as the inter-rater agreement. Gwet's AC1/gamma coefficient overcomes the issue where traditional measures, such as Cohen's Kappa, are not robust to skewed distributions of rankings. Figure 5 shows the interface for human evaluation study.

E. Hyper-parameter search domain for directed generation

During decoding, we apply length normalization following previous works. For the hyper-parameter in length normalization (beam search decoding), we use $\beta \in \{0.0, 0.5, 1.0, 1.5, 2.0\}$ for text summarization and $\beta \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ for image paragraph captioning.

F. Experimental results on open-ended generation

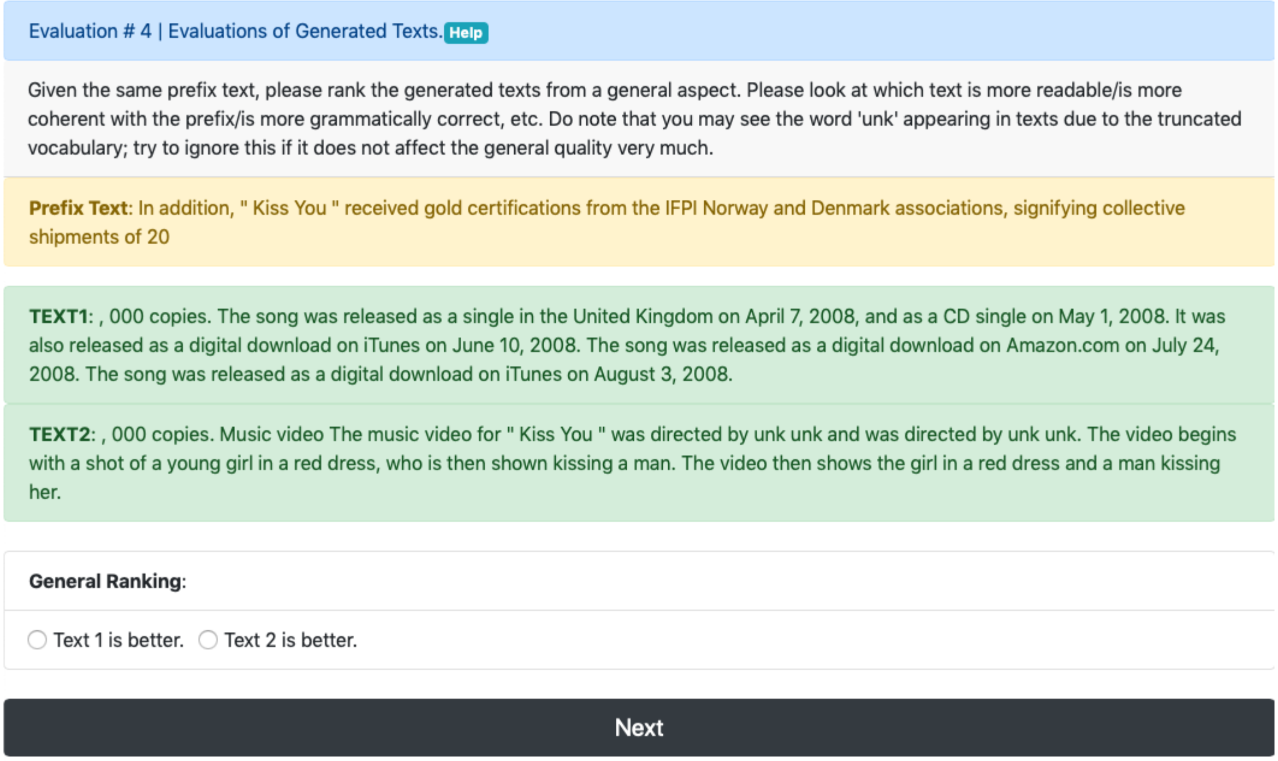
F.1. Full experimental results on WikiText-103

We present the full experimental results on WikiText-103 (Merity et al., 2017) test set for open-ended generations in Table 10. All the numbers are averaged over 3 runs with different random seeds and shown together with standard deviations.

In addition, we provide the full results *w.r.t.* different hyper-parameters for UL and SG on the WikiText-103 validation set in Table 11.

F.2. Open-ended generations results on IMDB dataset

Table 12 shows the open-ended generation results on movie revies from IMDB dataset.



Evaluation # 4 | Evaluations of Generated Texts.

Given the same prefix text, please rank the generated texts from a general aspect. Please look at which text is more readable/is more coherent with the prefix/is more grammatically correct, etc. Do note that you may see the word 'unk' appearing in texts due to the truncated vocabulary; try to ignore this if it does not affect the general quality very much.

Prefix Text: In addition, " Kiss You " received gold certifications from the IFPI Norway and Denmark associations, signifying collective shipments of 20

TEXT1: , 000 copies. The song was released as a single in the United Kingdom on April 7, 2008, and as a CD single on May 1, 2008. It was also released as a digital download on iTunes on June 10, 2008. The song was released as a digital download on Amazon.com on July 24, 2008. The song was released as a digital download on iTunes on August 3, 2008.

TEXT2: , 000 copies. Music video The music video for " Kiss You " was directed by unk unk and was directed by unk unk. The video begins with a shot of a young girl in a red dress, who is then shown kissing a man. The video then shows the girl in a red dress and a man kissing her.

General Ranking:

○ Text 1 is better. ○ Text 2 is better.

Figure 5: Human evaluation interface

G. Experimental details

In this section, we present the details of the datasets used in our experiments as well as the necessary experimental setup. All the experiments were conducted with a single GPU on our machine (CPU: Intel(R) Xeon(R) Gold 6240 CPU @ 2.60GHz; GPU: NVIDIA RTX 2080Ti).

For each task in our experiments, we use the same model architecture and train it with different objectives (*i.e.,* MLE, ScaleGrad and unlikelihood). The hyper-parameters that are used for different training objectives in the same task are exactly same, except for the ones described in Appendix E. We list the key hyper-parameters in this section.

G.1. Open-ended generation

Dataset The WikiText-103 (Merity et al., 2017) is a collection of over 100 million tokens extracted from the set of verified Good and Featured articles on Wikipedia. The training, validation and test sets contain 104m, 218k and 245k tokens, respectively.

Experiments For all the experiments, we use the same setup and the same hyper-parameters as listed in Table 13, except for the method-specific hyper-parameters. We load the GPT-2 medium and fine-tune it on WikiText-103 with a maximum of 35k iterations and select the model based on the validation perplexity.

G.2. Summarization

Dataset We use CNN/DM (Hermann et al., 2015; Nallapati et al., 2016) and NYT50 (Durrett et al., 2016) in our experiments for text summarization. Table 14 shows the dataset statistics in details.

Experiments The models are taken from (Liu & Lapata, 2019) and we train the models for the abstractive summarization with MLE, unlikelihood training and ScaleGrad on CNN/DM and NYT50. We list the hyper-parameters that we used in Table 15.

G.3. Image paragraph generation

Dataset We use the image paragraph captioning corpus Visual Genome dataset, introduced by Krause et al. (2017). The dataset contains 14,575 training, 2,487 validation, and 2,489 testing images. The average length of description paragraph is 67.50 tokens.

Table 10: Results for open-ended generations on the **Wikitext-103 testset**. **ppl**, **uniq** and **Rep/l** are computed at BPE-level and the rest are at word-level. The "↑" denotes higher value for better performance and "↓" is the opposite. Number marked with * are estimated based on the testset.

	Language Modeling					Auto Completion			
Models	**ppl** ↓	**uniq** ↑	**Rep/16** ↓	**Rep/32** ↓	**Rep/128** ↓	**Rep-1** ↓	**Rep-2** ↓	**Rep-3** ↓	**uniq-w** ↑
MLE	**13.24**$_{\pm 2e-4}$	12.54k$_{\pm 4e-3}$	0.234$_{\pm 5e-6}$	0.380$_{\pm 8e-6}$	0.619$_{\pm 7e-6}$	0.661$_{\pm 1e-5}$	0.500$_{\pm 3e-5}$	0.424$_{\pm 7e-5}$	16.83k$_{\pm 1e-1}$
UL ($\alpha = 1.0$)	16.06$_{\pm 2e-2}$	13.18k$_{\pm 6e-3}$	0.212$_{\pm 1e-6}$	0.341$_{\pm 1e-7}$	0.558$_{\pm 9e-6}$	0.559$_{\pm 6e-5}$	0.363$_{\pm 2e-4}$	0.291$_{\pm 3e-4}$	19.11k$_{\pm 7e-2}$
SG ($\gamma = 0.2$)	14.20$_{\pm 2e-2}$	**13.61k**$_{\pm 2e-3}$	**0.197**$_{\pm 6e-7}$	**0.317**$_{\pm 1e-6}$	**0.522**$_{\pm 4e-6}$	**0.443**$_{\pm 9e-7}$	**0.215**$_{\pm 2e-6}$	**0.143**$_{\pm 4e-6}$	**22.25k**$_{\pm 2e-2}$
Human	-	18.27k	0.177	0.285	0.480	0.382*	0.096*	0.037*	27.55k*

Table 11: Results for open-ended generation tasks on the **Wikitext-103 validation set**. **ppl**, **uniq** and **Rep/l** are computed at BPE-level and the rest are at word-level. The "↑" denotes higher value for better performance and "↓" is the opposite. Number marked with * are estimated based on the testset. The results are averaged over 3 runs with different random seeds.

	Language Modeling					Auto Completion			
Models	**ppl** ↓	**uniq** ↑	**Rep/16** ↓	**Rep/32** ↓	**Rep/128** ↓	**Rep-1** ↓	**Rep-2** ↓	**Rep-3** ↓	**uniq-w** ↑
MLE	13.17	12.52k	0.236	0.384	0.621	0.665	0.510	0.428	16.71k
UL($\alpha = 0.5$)	14.91	12.45k	0.217	0.350	0.579	0.601	0.424	0.348	18.02k
UL($\alpha = 1.0$)	16.52	12.77k	0.210	0.336	0.552	0.551	0.359	0.289	19.14k
UL($\alpha = 1.5$)	19.63	13.41k	0.201	0.315	0.523	0.489	0.267	0.205	22.00k
SG($\gamma = 0.2$)	14.43	13.73k	0.195	0.316	0.518	0.451	0.237	0.175	22.29k
SG($\gamma = 0.5$)	13.53	13.25k	0.218	0.352	0.576	0.561	0.389	0.331	19.13k
SG($\gamma = 0.8$)	13.27	12.79k	0.229	0.369	0.603	0.625	0.443	0.365	17.59k
Human	–	17.68k	0.173	0.278	0.470	0.376	0.097	0.032	27.63k

Experiments We follow the same experimental setup as in (Melas-Kyriazi et al., 2018). We train the model with different objectives and choose the model for testing based on the validation loss. During generation, tri-gram blocking and length-normalization are applied. Hyper-parameters that are used in our experiments are listed in Table 16.

H. Experimental results of different decoding strategies for auto-completion.

Table 17 shows the results for the auto-completion task when we train the model with ScaleGrad and infer with different decoding strategies.

I. Stochastic decoding for image paragraph captioning

We apply different stochastic decoding strategies for the MLE baseline on image paragraph captioning and report the results in Table 18. The experimental results demonstrate that stochastic decoding strategies do not work well in directed generation tasks, which is consitent with our findings in summarizaiton experiments.

J. Hyper-parameter sensitivity

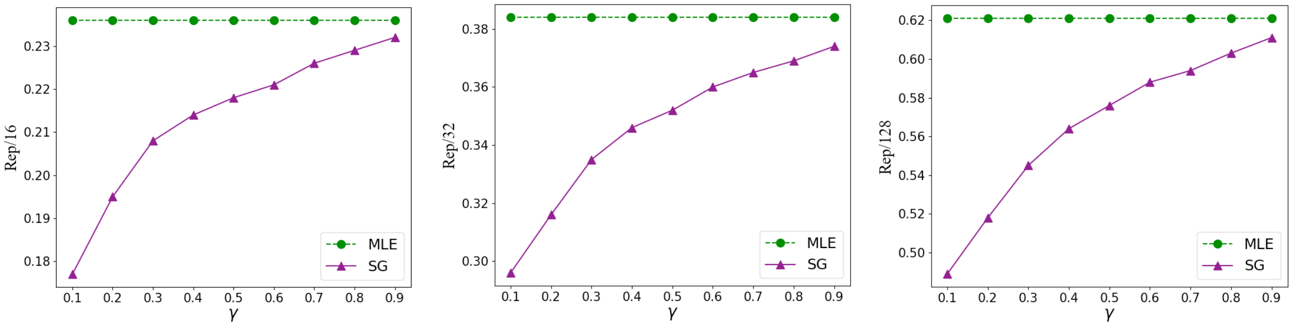
To fully present the sensitivity of Rep/l to the hyper-parameter, we further show how the Rep/l (*i.e.,* l=16, 32 and 128) change with γ in Figure 6.

K. Examples

In the following, we show the examples of generated texts in three tasks: auto-completion (Table 19 and Table 20), image paragraph captioning (Table 21 and Table 22) and text summarization (Table 23, Table 24, Table 25 and Table 26). In addition, Table 27 and Table 28 show the example of auto completion on PTB testset and movie reviews from IMDB dataset.

Table 12: Results for open-ended generations on movie reviews from **IMDB** dataset. **ppl**, **uniq** and **Rep/l** are computed at BPE-level and the rest are at word-level. Numbers marked with * are estimated based on the movie reviews from IMDB.

	Language Modeling					Auto Completion			
Models	**ppl**	**uniq**	**Rep/16**	**Rep/32**	**Rep/128**	**Rep-1**	**Rep-2**	**Rep-3**	**uniq-w**
MLE	100.764	7.48k	0.153	0.254	0.449	0.662	0.499	0.429	7.70k
UL ($\alpha = 1.0$)	108.334	8.09k	0.123	0.205	0.373	0.545	0.346	0.274	9.31k
SG ($\gamma = 0.2$)	110.451	8.14k	0.114	0.187	0.344	0.383	0.142	0.081	10.42k
Human	-	14.49k	0.118	0.208	0.378	0.329*	0.084*	0.009*	*19.11k



(a) Rep/16

(b) Rep/32

(c) Rep/128

Figure 6: Hyper-parameter (γ) sensitivity in the language modeling task on Wikitext-103 development set.

Table 13: Hyper-parameters for open-ended generation. **M** denotes the model-specific hyper-parameters. $\mathbf{lr}_0$ is initial learning rate.

Models	$\mathbf{lr}_0$	**M**	**batch**
MLE	2×10^{-5}	–	300
UL	2×10^{-5}	0.5/1.0/1.5	300
ScaleGrad	2×10^{-5}	0.2/0.5/0.8	300

Table 14: Dataset statistics for summarization.

Dataset	**Training Size**	**Validation Size**	**Test Size**
CNN/DM	287,227	13,368	11,490
NYT50	96,834	4,000	3,452

Table 15: Hyper-parameter lists for text summarization. **M** denotes the model-specific hyper-parameters. $\mathbf{lr}_0^{\mathbf{BERT}}$ and $\mathbf{lr}_0^{\mathbf{dec}}$ stand for initial learning rate for BERT and Transformer decoder. β is the hyper-parameter in length normalization.

Models	$\mathbf{lr}_0^{\mathbf{BERT}}$	$\mathbf{lr}_0^{\mathbf{dec}}$	**M**	**batch**	β	**Beam Size**
CNN/DM						
MLE	0.002	0.2	–	140	1.0	5
UL	0.002	0.2	0.5	140	2.0	5
ScaleGrad	0.002	0.2	0.8	140	1.5	5
NYT50						
MLE	0.002	0.2	–	140	1.5	5
UL	0.002	0.2	0.5	140	2.0	5
ScaleGrad	0.002	0.2	0.8	140	1.5	5

Table 16: Hyper-parameter lists for image paragraph captioning. **M** denotes the model-specific hyper-parameters. $\mathbf{lr}_0$ is initial learning rate.

Models	$\mathbf{lr}_0$	**M**	**batch**	β (w/o & w/ 3-blocking)
MLE	5×10^{-4}	–	10	0.0/0.2
UL	5×10^{-4}	0.5	10	0.0/0.3
ScaleGrad	5×10^{-4}	0.5	10	0.6/0.6

Table 17: Results of different decoding strategies for auto-completion.

Approaches	ppl	Rep-1	Rep-2	Rep-3	uniq-w
ScaleGrad					
Greedy Search ($\gamma = 0.2$)	14.20	0.443	0.215	0.144	22.25k
Beam Search (b = 3)	14.20	0.422	0.210	0.134	8.75k
Beam Search (b = 6)	14.20	0.453	0.250	0.171	8.32k
Beam Search (b = 10)	14.20	0.489	0.298	0.214	8.00k
Top-p (p = 0.3)	14.20	0.356	0.107	0.049	30.48k
Top-p (p = 0.9)	14.20	0.217	0.027	0.008	52.76k
Top-k (k = 40)	14.20	0.254	0.039	0.012	39.50k
Top-k (k = 100)	14.20	0.234	0.031	0.010	44.27k
UL					
Greedy Search ($\alpha = 1.0$)	16.06	0.559	0.363	0.291	19.11k
Beam Search ($b = 6$)	16.06	0.577	0.418	0.325	7.49k
Top-p ($p = 0.3$)	16.06	0.444	0.176	0.070	24.45k
Top-k ($k = 40$)	16.06	0.336	0.067	0.021	31.89k
MLE					
MLE	13.24	0.661	0.500	0.424	16.83k
Beam Search ($b = 6$)	13.24	0.697	0.566	0.463	6.11k
Top-p ($p = 0.3$)	13.24	0.558	0.210	0.116	20.13k
Top-k ($k = 40$)	13.24	0.485	0.154	0.076	24.26k
Human	–	0.382	0.096	0.037	27.55k

Table 18: Image paragraph captioning results for stochastic decoding on Visual Genome testset.

Models	CIDEr
Top-p (p=0.3)	19.54
Top-p (p=0.9)	18.67
Top-k (k=40)	18.73
Top-k (k=100)	18.05
MLE w/ 3-block	22.77

Table 19: Example 1 of generated continuations for different training objectives.

Prefix
Robert Boulter is an English film, television and theatre actor. He had a guest starring role on the television series The Bill in 2000. This was followed by a starring
MLE
role in the film The Last of the Mohicans in 2002. He has appeared in the television series The X Factor, The X Factor Australia, The X Factor New Zealand, The X Factor South Africa, The X Factor New Zealand, The X Factor South Africa, The X Factor New Zealand, The X Factor New Zealand, The X Factor New Zealand, The X Factor New Zealand, The X Factor New Zealand, The X Factor New Zealand, The X Factor New Zealand, The X Factor New Zealand
Unlikelihood
role in the film The Last of the Mohicans, which was released in 2002. He also appeared in the television series The X Factor, where he played the role of a DJ. In 2005, he starred in the film The Last of the Mohicans, which was released in 2006. He also appeared in the television series The X Factor, where he played the role of a DJ. In 2008, he starred in the television series The X Factor, where he played the role
ScaleGrad
role in the BBC's comedy series The Office in 2002. In 2003, he appeared in the BBC's crime drama series The Yard, playing a police officer who is caught up in a murder investigation. He also appeared in the BBC's crime drama series The X Factor, playing a man who is accused of murdering his wife. In 2004, he starred in the BBC's comedy series The Office, playing a fictional version of himself. He has appeared in several

Table 20: Example 2 of generated continuations for different training objectives.

Prefix
You's patriotism, and Mei's reflections on the quotidian are a few examples. More broadly, Du Fu's work in transforming the lushi from mere word play into "a vehicle for serious poetic utterance"
MLE
is a major contribution to the genre. The first of the three major works of the genre, was written in the early 1920s, and was the first of the three to be published in English. It was published in the United States in the United States, and was translated into French in the 1930s. The work was the first of the three
Unlikelihood
is a major influence on the modern Chinese novel. The modern Chinese novel is often characterized as being "literary in nature", but it is not limited to literary genres. The modern Chinese novel has been described as "a literary genre that has been influenced by the Western literary tradition". The modern Chinese novel has been described as "a literary genre that has been influenced by the Western literary tradition"
ScaleGrad
is a powerful example of how the Chinese literary tradition can be used to explore themes of nationalism and self determination. In addition to his own poetry, Du Fu wrote several works of prose. The first was a collection of short stories called "The Story of the Three Sisters", which was published in 1885. It was followed by a series of novels, including the first two volumes of the Song dynasty epic poem "The Three Sisters"

Table 21: Example 1 of generated paragraphs for different training objective on Visual Genome testset.

Reference
Three adult horses are in a field and a small horse. A horse color light brown has the neck low and is eating grass. This horse has the mane and the tail color tan. Two dark brown horses also are eating grass from the ground. The baby horse has a white spot on the front and is tiny. On back there is a wall of stones.
MLE w/o tri-gram blocking
There are two horses in a field. The horses are brown and white. The horses are brown and white. The horses are brown and white. The horses are brown and white. The horses are brown and white. The horses are brown and white. The horses are brown. The horses are brown. The horses are brown. The horses are brown. The horses are brown. The horses are brown. The horses are brown. The horses are brown. The horses are brown . The horses are brown. The horses are brown. The horses are brown. The horses are brown. The horses are brown. The horses are brown.
MLE w/ tri-gram blocking
There are two horses in a field. The horses are brown and white. The horse on the right is brown and brown. The horse is brown. The other horse is white. The grass is green and green. The sky is very and green. There are trees in the background. The trees are green and healthy. The horses are black. The sun is shining brightly.
ScaleGrad w/ tri-gram blocking
Two horses are standing in a field. The field is covered with green grass and weeds. The horse is brown and white. The horses are brown and white. There is a large tree behind the horses. The sky is blue and clear. There are many trees in the background. The grass is green. The sun is shining brightly. The sky is mostly cloudy.

Table 22: Example 2 of generated paragraphs for different training objectives on Visual Genome testset.

Reference
There is a white and yellow train with red doors next to a red and yellow train. They are on top of the railroad tracks. There is a brick walkway next to the trains.
MLE w/o tri-gram blocking
A train is on the tracks. The train is yellow and yellow. The train is red and white . The train is red and white. The train is red and white . The train is on the train. The train is yellow and yellow. The train is blue and white . The train is on the tracks. The train is yellow and yellow . The train is blue and white. The train is blue and white. The train is on the tracks . The train is yellow. The train is blue. The train is blue. The train is blue. The train is blue. The train is blue. The train is blue. The train is blue. The train is blue. The train is blue. The train is blue.
MLE w/ tri-gram blocking
A train is on the tracks. The train is yellow and yellow. The train has a yellow stripe on the front. The front of the train is red. The sky is blue. There are many windows on the train. There is a train on the platform. The platform is made of metal. The is a platform on the side of the train. The are many tracks on the train. There are wires on the ground. There is a building behind the train tracks. There is a large building behind the train.
ScaleGrad w/ tri-gram blocking
A train is on the tracks. There are two sets of tracks next to the train. The train is white and yellow. There is a large white building behind the trains.

Table 23: Example 1 of summaries generated by the model trained with different training objective on CNN/DM testset.

Reference
report by dick marty, lead investigator for council of europe, gives bleak description of secret prisons run by cia in poland and romania from 2003 to 2006. marty says he relied heavily on information from cia agents. critics in poland and romania attack marty’s use of anonymous sources.
MLE
swiss sen dick marty reports that central intelligence agency operates secret prisons run by american in poland and romania from 2003 to 2006. says prison were operated exclusively by americans in poland and.
Unlikelihood
swiss sen dick marty reports that secret prisons run by central intelligence agency in eastern europe, with information he says is gleaned from anonymous intelligence agents. report is prepared by swiss senator investigating cia operations for council of europe, 46 - nation rights group. scathing report says prison were operated exclusively by americans in poland and romania from 2003 to 2006.
ScaleGrad
dick marty, swiss senator investigating cia operations for council of europe, gives bleak description of secret prisons run by central intelligence agency in eastern europe, with information he says is gleaned from anonymous intelligence agents. report says prisons were operated exclusively by americans in poland and romania from 2003 to 2006.

Table 24: Example 2 of summaries generated by the model trained with different training objective on CNN/DM testset.

Reference
middleton wrote about eating whale for a national newspaper travel story. she described eating it at the #123-a-night juvet landscape hotel. whale and dolphin conservation group said the news was ‘disappointing’. the wdc said ms.middleton should have gone whale watching instead.
MLE
the 31-year-old has been slammed for ‘promoting the cruel and unnecessary whaling industry’. the 31-year-old boasted of eating minke whale in norway, to the horror of environmental campaigners who say she is helping promote an ‘unimaginably cruel’ trade. the hunting and importation of whale meat is banned in britain , as it is in most parts of the world.
Unlikelihood
the 31-year-old boasted of eating minke whale at juvet landscape hotel in norway. the 31-year-old has been slammed for ‘promoting the cruel and unnecessary whaling industry ’¡q¿environmental campaigners say she is helping promote an ‘ unimaginably cruel ’ trade.
ScaleGrad
duchess of cambridge’s brother-in-law has led a personal crusade against poaching and protecting wildlife. pippa middleton boasted of eating minke whale in norway, conservation group said she is helping promote ‘promoting the cruel and unnecessary whaling industry’.

Table 25: Example 1 of summaries generated by the model trained with different training objective on NYT50 testset.

Reference
protesters angry over bribery scandal involving state-run oil company petrobras. brazilian president dilma rousseff also is struggling with an economic downturn.
MLE
protesters are calling for president dilma rousseff to be impeached. rousseff's supporters call for the president to be impeachment¡q¿they say there hasn't been any evidence she was involved in the corruption scandal.
Unlikelihood
protesters are calling for president dilma rousseff to be impeached. there is a number of issues at play. one of the biggest is an investigation into a multimillion-dollar kickback scheme at the state-run oil company petrobras.
ScaleGrad
president dilma rousseff is facing an economic downturn and a massive bribery scandal. rousseff has defended her right to protest and acknowledged the need to clean up corruption at petrobras.

Table 26: Example 2 of summaries generated by the model trained with different training objective on NYT50 testset.

Reference
wollemi pine, which flourished 200 million years ago, is available to gardeners and can be grown indoors or outdoors. thought to extinct, it was discovered in 1994 in wollemi national park in australia. botanists took cuttings from wild colony of trees and began propagating baby pines and sending them to botanic gardens around world.
MLE
new york city plant - o - ramachan pine, which grew among dinosaur 200 million years ago, is named after dr david noble, who rappelled into deep gorge in blue mountains of wollemi national park, ten miles west of sydney, australia. photos. map. pine was discovered until 1994, when bush delay and national parks officer david noble tried rappelled into woods of wollemi, which he had never seen before. scientists had evidence of wollemi pine's existence, but only from fossils.
Unlikelihood
nicholas d kristof op-ed column on plight of wollemi pine, one of oldest on earth who grew among dinosaur 200 million years ago. notes that it was not discovered until 1994, when bush rescue officer david noble rappelled into deep gorge in blue mountains of sydney, australia, and came back up with branch of tall tree that he never saw before.
ScaleGrad
nicholas d kristof op-ed column on new species of wollemi pine, which grew among dinosaur 200 million years ago and was discovered until 1994, when bush delay and national parks officer david noble rappelled into deep gorge in blue mountains of wollemi national park, 120 miles west of sydney, australia, and discovered new species, wollemia nobilis, named after noble.

Table 27: Example of generated continuations for different training objectives on PTB testset.

Prefix
last resort who were criticized after the 1987 crash – once again couldn’t handle the selling pressure. Big investment banks refused to step up to the plate to support the beleaguered floor traders by buying big blocks of stock, traders say. Heavy selling
MLE
pressure was also felt by the Wall Street banks. The Wall Street banks were forced to sell more stock to hedge funds and other investors, and to buy more stock to hedge against losses. The Wall Street banks were forced to sell more stock to hedge funds and other investors, and to buy more stock to hedge against losses. The Wall Street banks were forced to sell more stock to hedge funds and other investors, and to buy more stock to hedge against losses.
Unlikelihood
pressure was also felt by the banks themselves. In 1987, the New York Federal Reserve Bank, which had been the largest bank in the United States since the 1930s, was forced to sell $1.5 billion worth of stock. The bank’s stock price fell to $0.001, and it was forced to close its New York branch. The bank’s stock price fell to $0.001, and it was forced to close its New
ScaleGrad
by the big banks led to a sharp drop in prices, and the market collapsed. In 1992, the Federal Reserve Bank of New York issued a report on the financial crisis, which concluded that the Fed’s actions had not been sufficient to prevent the collapse of the housing market. The report recommended that the Fed increase its balance sheet by $1.5 trillion (equivalent to about $2.5 trillion today), but this was not implemented.

Table 28: Example of generated continuations for different training objectives on movie reviews from IMDB dataset.

Prefix
One of the other reviewers has mentioned that after watching just 1 Oz episode you’ll be hooked. They are right, as this is exactly what happened with me. The first thing that struck me about Oz was its brutality and unflinching scenes of violence.
MLE
I was not expecting anything like that in a show about a family. I was expecting a show about a family, but I was expecting something more. I was expecting a show about a family, but I was expecting something more. I was expecting a show about a family, but I was expecting something more. I was expecting a show about a family, but I was expecting something more. I was expecting a show about a family, but I was expecting something more. I was expecting a show
Unlikelihood
I was immediately hooked. The violence was so graphic that I was completely taken aback by how it was handled. I was also struck by the fact that the show was so well-written that it was almost impossible to not be entertained. The second thing that struck me about Oz was its ability to make me laugh. I was so taken aback by the violence that I couldn ’t help but laugh. The violence was so graphic that I was completely taken aback by how it
ScaleGrad
It was also very funny, but not funny enough to make me laugh out loud. The second thing that struck me was the fact that the show’s characters were so well developed that they were able to carry their own weight without being annoying or annoying. In its original broadcast, “Oz” was watched by 9.5 million viewers, according to Nielsen ratings. It was the highest rated