Look-back Decoding for Open-Ended Text Generation

Nan Xu*, Chunting Zhou*, Asli Celikyilmaz*, Xuezhe Ma*  
*University of Southern California, Meta AI  
{nana, xuezhema}@usc.edu, {chuntinz, aslic}@meta.com

Abstract

Given a prefix (context), open-ended generation aims to decode texts that are coherent, which do not abruptly drift from previous topics, and informative, which do not suffer from undesired repetitions. In this paper, we propose Look-back, an improved decoding algorithm that leverages the Kullback–Leibler divergence to track the distribution distance between current and historical decoding steps. Thus Look-back can automatically predict potential repetitive phrase and topic drift, and remove tokens that may cause the failure modes, restricting the next token probability distribution within a plausible distance to the history. We perform decoding experiments on document continuation and story generation, and demonstrate that Look-back is able to generate more fluent and coherent text, outperforming other strong decoding methods significantly in both automatic and human evaluations.

1 Introduction

Despite the impressive success on generating fluent and accurate sentences for low-entropy tasks such as summarization or translation, large-scale language models (LLMs) still suffer from serious degeneration problems, such as undesired repetitions (Holtzman et al., 2019) and unnatural topic drifts, under open-ended settings (Eikema and Aziz, 2020). Open-ended neural text generation aims to generate coherent and diverse text from LLMs, given contextual prefix (Nadeem et al., 2020; Dhamala et al., 2022), and has spawned a wide range of natural language applications, including contextual text completion (Radford et al., 2019), story generation (Fan et al., 2018), and review generation (Cho et al., 2019).

To alleviate the degeneration problem in open-ended text generation, a number of techniques have emerged over the recent years, which can be categorized into two directions: i) improved learning proposing new learning objectives, e.g., likelihood training (Welleck et al., 2019), contrastive training (Su et al., 2022) and sequence likelihood calibration (Zhao et al., 2022), to compensate for the rooted deficiency of the conventional Maximum Likelihood Estimation (MLE); ii) improved decoding remedying tedious and repetitive generations in decoding search (Su et al., 2022; Li et al., 2022), or combating topic drifts in sampling procedures (Hewitt et al., 2022).

In this work, we propose a new decoding algorithm, named Look-back, which pays particular attention to the probability distribution disparity between continuation and history text. Unlike contrastive search (Su et al., 2022; Su and Xu, 2022) which uses cosine similarity between the hidden representation, Look-back leverages the Kullback–Leibler (KL) divergence to track the distribution distance between current and historical decoding steps. The main motivation of Look-back is that

---

1Code and resources are available at https://github.com/xunannancy/lookBackDecoding.

2The correlation between sequence probability and its quality for MLE trained models can be low (Liu et al., 2022).
KL divergence defines a distance between the probability distributions of decoding steps, which arguably better aligns with the decoding practice. As shown in Figure 1 (a), as the greedy algorithm repeatedly outputs single sentences, the distance with the closest past token distribution decreases towards 0. Besides, when the continuation switches to another topic in Figure 1 (b), the distribution distance of continuation with prefix obtains much higher levels compared with topic-relevant human continuation. Based on our prior observations, for informative and coherent generation, the probability distribution should not be too close to history to guarantee diversity, but relatively close to prefix to maintain coherence.

Experimentally, through two tasks of open-ended text generation, including document continuation and story generation, we demonstrate that Look-back outperforms a variety of open-ended decoding algorithms under different scales of pre-trained LLMs (GPT2-XL and OPT-6.7B) by producing much more coherent texts – high mauve score compared with human continuation and high similarity score measured against prefix, while maintaining similar level of diversity.

2 Related Work

Improved Learning Algorithms Yang et al. (2018); Adiwardana et al. (2020) observed that increasing number of candidates in beam search or sampling leads to worse quality of generated data. They attribute this to the predominant training objective (i.e., Maximum Likelihood Estimation) that might not accurately rank generated sequences by quality (Zhao et al., 2022). Besides, Holtzman et al. (2019) found that searching for the probable sequences always results in short and repetitive texts, which further motivated recent efforts to improve generation via revised learning objectives. Welleck et al. (2019) proposed unlikelihood training to force unlikely generations to be assigned lower probability by the model. To alleviate degeneration, SimCTG (Su et al., 2022) introduced a contrastive training objective to preserve sparseness of the token similarity matrix of the generated text. To avoid unintentionally boosting the probability of other irrelevant tokens in unlikelihood training, Jiang et al. (2022) leveraged contrastive token learning to explicitly teach the LLM to assign negative tokens with a lower probability than positive tokens through more focused contrast between the two. Based on a BERTScore-style similarity metric between model decodes and targets measured in the model’s latent space, Zhao et al. (2022) calibrated model-generated sequences with sequence likelihood calibration to better align with reference sequences via different types of losses (e.g., rank and margin loss).

Improved Decoding Algorithms Liu et al. (2022) observed that search methods (e.g., greedy and beam) which optimize generation probabilities may result in tedious and repetitive outputs in open-ended text generation. Su et al. (2022) complemented the contrastive training with contrastive search for decoding, which selects tokens more distinguishable from previous context. Li et al. (2022) observed that degeneration is more prevalent in larger LMs than smaller ones, and proposed contrastive decoding to remove these undesired behaviors by factoring out smaller LM’s behavior from the larger LM. On the other hand, truncation sampling methods such as nucleus (Holtzman et al., 2019) and typical (Meister et al., 2022) decoding improve sample quality with more diverse samples compared to direct sampling, but at the expense of poor coherence and undesired topic drift. Hewitt et al. (2022) introduced $\eta$-sampling to truncate words below an entropy-dependent probability threshold. A concurrent work observed the strong correlation between good generation quality and narrow entropy zone, hence proposed entropy-aware decoding to promote good generation by constraining greedy decoding into the narrow entropy zone (Arora et al., 2023).

Without extra effort on fine-tuning LMs, the proposed Look-back improves conventional search method with reference from the given prefix and prior generation, so that undesired repetitions and topic drifts can be explicitly alleviated.

3 Background

3.1 Open-ended Text Generation

Given a sequence of $m$ tokens sampled from natural text $C = \{x_1 \ldots x_m\}$ as context or prefix, the neural text generation is to decode a $n$-token continuation using the probability distribution provided by pre-trained LMs:

$$p(x_{m+1:m+n}|C) = \prod_{t=1}^{n} P(x_t|C, x_{m+1} \ldots x_{m+t-1}),$$
Degeneration LM (Decoding) Continuation

### Prefix
In addition to live broadcasts FIFA Fan Fests offer food and beverages, merchandise and various entertainment events by local and international artists. The start of 2006 World Cup was

<table>
<thead>
<tr>
<th>Repetitive Continuation</th>
<th>S1: <em>(GPT2-XL greedy)</em> marked by the FIFA Fan Fest in Sao Paulo, Brazil. The FIFA Fan Fest is a unique event that brings together fans from all over the world to celebrate the FIFA World Cup. The FIFA Fan Fest is a unique event that brings together fans from...</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2: <em>(davinci-002 greedy)</em> celebrated with a concert by Bon Jovi in Berlin. The FIFA Fan Fest in Munich was located at the Olympic Park. The FIFA Fan Fest in Frankfurt was located at the Römerberg. The FIFA Fan Fest in Hamburg was located at the...</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Off-Topic Continuation</th>
<th>S3: <em>(ada-001 greedy)</em> the local church. The authorities are investigating how Khedira managed to enter the house and what role he played in the attack.</th>
</tr>
</thead>
<tbody>
<tr>
<td>S4: <em>(davinci-002 greedy)</em> Temple Emanu-El in New York City...</td>
<td></td>
</tr>
</tbody>
</table>
| S5: *(ada-001 nucleus)*: aents home in The Lorraine weekend'se house of intensity and occupation, great law enforcement officer...

### Table 1

| Degeneration examples with typical decoding algorithms by GPT2-XL and GPT3 (ada-001 and davinci-002). Complete sentence repetition (S1), repetition with minor location changes (S2) or paragraph duplication (S5) is marked in *green*, while unnatural (S3\&S4) or stiff (S5) topic drifts are in *pink*. |

<table>
<thead>
<tr>
<th>Degeneration LM (Decoding) Continuation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Replication</strong> LLMs prefer to overestimate the probability of repeated sequences (Welleck et al., 2019) especially for deterministic algorithms such as greedy and beam search. Although decoding algorithms such as nucleus sampling (Holtzman et al., 2019) have been proposed to interrupt repeating sequences, we can still observe repetitive and tedious continuity even from the state-of-the-art GPT-3 language model (Brown et al., 2020), as shown in Table 1. Besides the consensus that probabilities from conditional LMs often do not accurately rank-order generated sequences by quality (Zhao et al., 2022), a recent study provides a possible way to explain the repetitive generation with the observed analogical sequence copying pattern: prefix matching and copying (Olsson et al., 2022).</td>
</tr>
<tr>
<td><strong>Incoherence</strong> Sampling algorithms sacrifice coherence for alleviating repetition during decoding.</td>
</tr>
</tbody>
</table>

where the continuation is generated token-by-token using a particular decoding strategy. For instance, greedy algorithm selects the next token given context with the highest probability, while nucleus sampling (Holtzman et al., 2019) restricts the plausible area of tokens with total mass above a threshold.

### 3.2 Degeneration Problems

There are two commonly observed degeneration problems in open-ended text generation: repetition and incoherence.

**Repetition** LLMs prefer to overestimate the probability of repeated sequences (Welleck et al., 2019) especially for deterministic algorithms such as greedy and beam search. Although decoding algorithms such as nucleus sampling (Holtzman et al., 2019) have been proposed to interrupt repeating sequences, we can still observe repetitive and tedious continuity even from the state-of-the-art GPT-3 language model (Brown et al., 2020), as shown in Table 1. Besides the consensus that probabilities from conditional LMs often do not accurately rank-order generated sequences by quality (Zhao et al., 2022), a recent study provides a possible way to explain the repetitive generation with the observed analogical sequence copying pattern: prefix matching and copying (Olsson et al., 2022).

**Incoherence** Sampling algorithms sacrifice coherence for alleviating repetition during decoding.

As shown in Table 1, given probabilities from GPT-3 models, nucleus sampling fails to produce coherent generation, switching topic from *Burkan’s acute indigestion* to *Shanny’s way to home* with *ada-001* (S5). Recent decoding algorithms depend on model confidence to “guarantee” coherence while resolving repetition explicitly with certain heuristics. For example, SimCTG (Su et al., 2022) selects from most probable candidates predicted by LMs. Contrastive decoding (Li et al., 2022) exploits coherence nature of the expert LMs. In both S3 and S4 from Table 1, unfortunately, we find that the coherence hypothesis of pretrained LMs in prior work does not always hold in practice: it is likely to produce incoherent sentences when powerful LMs rigorously follow model confidence at each step with greedy algorithm.

### 4 Proposed Method: Look-back

As presented in Algorithm 1, *Look-back* first leverages probability distribution distance between current and prior steps to avoid repetitions (§4.1), then incorporates reference from given prefix to mitigate topic drifts (§4.2).

#### 4.1 Alleviating Repetitions with Reference from Prior Texts

**Signal for Surface or Semantic Repetitions** In the decoding process of open-ended text generation, one of the plausible tokens is selected/sampled according to model probability. Inspired by the decisive role of probability distribution, we investigate measuring the distance between current and prior steps in distribution space via KL divergence: $D_{KL}(p_t | p_{t'})$ for any $1 \leq t' < t$. As the distance...
The FIFA Fan Fest is a unique event that brings together fans. As demonstrated in Figure 2c and Figure 2d, values of $KL_{min}^t$ become flat as repetition-style degeneration.

Algorithm 1 Look-back Decoding

Input: Prefix $C = \{x_1 \ldots x_m\}$, language model with vocabulary $V$, beam size $k$ and threshold $\alpha$

Output: Continuation $G = \{x_{m+1} \ldots x_{m+n}\}$

\[
G \leftarrow \{\}
\]

\[
\text{for } m+1 \leq t \leq m+n \text{ do} \\
\text{if } KL_{min}^t \leq \alpha \text{ then } \triangleright \text{Alleviate Repetitions} \\
\text{for } v \in V^k \text{ do} \\
q_v = \text{softmax}(−KL_{t+1}^{t+1 | v} | C) \\
\text{end for} \\
x_t = v \sim q_v \quad \triangleright \text{Improve Coherence} \\
\text{else} \\
x_t = \text{argmax}_{v \in V} p_{\theta}(v | x_{<t}) \\
\text{end if} \\
G \leftarrow G \cup \{x_t\} \\
\text{end for}
\]

heatmap shown in Figure 2a, for steps generating identical tokens, their corresponding probability distributions stay close to each other than those with dissimilar outputs.

Note that neither the contrastive training objective (SimCTG) (Su et al., 2022) nor its contrastive search decoding algorithm (Su and Xu, 2022) can be directly applied to LLMs such as GPT3, where its hidden states are inaccessible. Fortunately, we can directly detect surface or semantic repetitions from GPT3 by analyzing available probability distribution: step pairs producing either identical tokens or tokens sharing similar semantic meaning are distinguishable with distribution distance. Take Figure 2b as an instance: output token pairs from decoding steps with closest probability distributions are the 1st and 2nd $FAN$, city $Munich$ and $Frankfurt$, location $Olympic$ and $R$ of $Römerberg$.

As repetitive steps tend to stay extremely close to prior steps with similar outputs in probability distribution space, we calculate the probability distribution distance between the t-th and closest prior step as $KL_{min}^t$ for further analysis:

\[
KL_{min}^t = \min_{1 \leq j \leq t-1} KL(p(\cdot | x_{<t})||p(\cdot | x_{<j}))
\]

As demonstrated in Figure 2c and Figure 2d, values of $KL_{min}^t$ become flat as repetition-style degeneration.
tion advances\textsuperscript{4}.

**Alleviating Repetitions** Since identical or similar repetition pattern could be forecasted via probability distribution analysis, Look-back attempts to avoid repetitive sentences or phrases prior to actual generation. Practically, when $\text{KL}_{\min}^t$ has been below a pre-defined threshold $\alpha$, an alarm is triggered and Look-back attempts to sample a token from the top-$k$ most probable tokens from the vocabulary $V$ rather than sticking to the top-1 token:

$$x_t \begin{cases} \sim \text{Unif}(V^k) & \text{if } \text{KL}_{\min}^t \leq \alpha \\ = \arg\max_{v \in V} p_v(v|x_{<t}) & \text{Otherwise} \end{cases}$$

where $V^k$ is the set of top-$k$ most probable tokens from the vocabulary $V$. To avoid false positive cases where one step identified with high possibility to repeat may not necessarily lead to undesired repetitions, we do not exclude its most probable token from the plausible candidate set on purpose.

### 4.2 Improving Coherence with Reference from Given Prefix

**Signal for Topic Drift** In open-ended generation, in order to produce sentences coherent with the given prefix, the decoding algorithm is required to provide further elaboration of the major topic conveyed in the prefix. According to the prior observations (e.g., Munich and Frankfurt in Figure 2b), decoding steps with tokens sharing similar semantic meaning are close to each other with respect to probability distribution distance. Therefore, we explore the KL divergence between current and prefix $m$ steps that should keep to the same topic:

$$\text{KL}_{\min}^{t\mid C} = \min_{1 \leq j \leq m} \text{KL} (p(\cdot|x_{<t})\|p(\cdot|x_{<j}))$$

When comparing distribution distance of incoherent generation with natural continuation to the same prefix, the probability distribution divergence maintains a much higher level for generation with obvious topic drift, as shown in Figure 2e.

**Improving Coherence** When the model is prone to provide repetitive tokens, one straightforward solution for avoiding repetition is to randomly sample from the top-$k$ plausible tokens. It is likely to result in unnatural topic drift due to undesired sampling choices accumulation over long sequence decoding, which is frequently observed in sampling algorithms (Eikema and Aziz, 2020; Maynez et al., 2020). On the other side, the probability distribution distance between current and prefix is able to distinguish whether the generation is on-topic or not. Therefore, Look-back wisely samples from the plausible candidates according to their influence on coherence reflected by next-step distribution distance with prefix:

$$x_t \begin{cases} \sim \text{Unif}(V^k) & \text{if } \text{KL}_{\min}^{t+1\mid C} \leq \alpha \\ = \arg\max_{v \in V} p_v(v|x_{<t}) & \text{Otherwise} \end{cases}$$

where tokens with larger next-step distance to prefix is less likely to be sampled given the softmax operation upon KL divergence.

### 5 Experiments

In this section, we first introduce the datasets (§5.1) and automatic metrics (§5.2) used to evaluate the generation quality of the proposed Look-back and other strong decoding baselines (§5.3). We then analyze experimental results evaluated by automatic metrics (§5.5) and human evaluators (§5.6). Lastly, we show effectiveness of different techniques used in Look-back through detailed analyses (§5.7).

#### 5.1 Datasets

We consider two applications of open-ended text generation: 1) document continuation on WikiText-103 with articles fitting the Good or Featured article criteria specified by editors on Wikipedia (Merity et al., 2016), and 2) story generation on WritingPrompts, which is a challenging task for inspiring continuations with abstract, high-level story prompts submitted by online users and continuations responded by others freely on Reddit (Fan et al., 2018).

#### 5.2 Evaluation Metrics

We adopt the following automatic metrics to evaluate generation quality:

**Repetition** We use $\text{rep-}n$ to measure sequence-level repetition according to the portion of duplicate $n$-grams (Welleck et al., 2019). For a sequence $x$, $\text{rep-}n = 1.0 - \frac{|\text{unique } n\text{-grams}(x)|}{|\text{total } n\text{-grams}(x)|}$.

**Diversity** Following (Su et al., 2022), we obtain an overall assessment of model repetition by considering repetition at different $n$-gram levels: $\text{diversity} = \prod_{n=2}^{4} (1.0 - \text{rep-}n)$. 

\textsuperscript{4}Spikes in Figure 2d in later decoding steps correspond to multiple tokens for representing one single location, e.g., ö, mer, berg for Römerberg in Figure 2b.
Table 2: Automatic evaluation results of different decoding algorithms for document continuation and story generation. Continuation generated by Look-back is of similar level of diversity as human texts while much more relevant to prefix (highest coherence) and semantically similar to human continuation (highest MAUVE).

<table>
<thead>
<tr>
<th>Method</th>
<th>WikiText-103</th>
<th>WritingPrompts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rep-2</td>
<td>rep-3</td>
</tr>
<tr>
<td>human</td>
<td>6.91</td>
<td>1.83</td>
</tr>
<tr>
<td>nucleus</td>
<td>5.29</td>
<td>1.97</td>
</tr>
<tr>
<td>typical</td>
<td>3.61</td>
<td>1.07</td>
</tr>
<tr>
<td>η-sampling</td>
<td>6.25</td>
<td>2.49</td>
</tr>
<tr>
<td>SimCTG</td>
<td>5.37</td>
<td>1.97</td>
</tr>
<tr>
<td>Look-back</td>
<td>8.22</td>
<td>1.34</td>
</tr>
<tr>
<td>nucleus</td>
<td>6.08</td>
<td>2.19</td>
</tr>
<tr>
<td>typical</td>
<td>6.58</td>
<td>2.25</td>
</tr>
<tr>
<td>η-sampling</td>
<td>6.07</td>
<td>2.26</td>
</tr>
<tr>
<td>SimCTG</td>
<td>5.44</td>
<td>1.97</td>
</tr>
<tr>
<td>Look-back</td>
<td>9.21</td>
<td>1.74</td>
</tr>
</tbody>
</table>

**MAUVE** By computing information divergences in a quantized embedding space, MAUVE (Pillutla et al., 2021) directly compares the learnt distribution from a text generation model to the distribution of human-written continuation.

**Coherence** The semantic coherence between prefix and continuation is measured as the cosine similarity between their sentence embeddings represented by SimCSE (Gao et al., 2021).

Results measured by all metrics range from 0 to 1, and higher scores indicate better generation except rep-n, for which the lower the better.

5.3 Decoding Baselines

Given pre-trained LMs with conventional MLE, we evaluate Look-back together with various decoding algorithms for fair comparisons.

**Search Methods** We consider the competitive contrastive search proposed in SimCTG (Su et al., 2022) that predicts the next token based on both the output distribution and representation similarities between candidates and past tokens.

**Sampling Methods** Nucleus sampling (Holtzman et al., 2019) samples the next token from the top-p portion of the probability mass. Typical decoding (Meister et al., 2022) samples from the set of words whose negative log-probabilities are close to the conditional entropy. η-sampling (Hewitt et al., 2022) truncates any word whose probability is smaller than an entropy-based threshold.

5.4 Implementation Details

We randomly sample 1,000 instances from the original training data of WikiText-103 and WritingPrompts as our validation and test sets. Given the beginning several tokens as prefix, we generate 256 tokens with different decoding algorithms and disregard those after the end-of-text token during evaluation. Practically, we consider a sliding window comprising 128 prior tokens to avoid undesired repetitions while allow necessary repetitions of text far from the current decoding step. We perform experiments with pre-trained LMs from different families and scales: GPT2-XL (Radford et al., 2019) and OPT-6.7B (Zhang et al., 2022). The same set of hyperparameters is used to decode from different LMs: the beam size for beam search is 10, p = 0.95 for nucleus, τ = 0.92 for typical, and η = 0.0003 for η-sampling. We follow the recommended range for $k = \{5, 8, 10\}$ and $\alpha = [0.5, 0.9]$ in SimCTG and select the set based on their MAUVE scores on the validation set. For Look-back, the range of candidate amount $k$ is $\{5, 8, 10\}$ and the threshold $\alpha$ is ranging from $[0.5, 1.6]$. We select hyperparameters that result in the rep-2 score closest to human’s and the optimal MAUVE performance on the validation set.

5.5 Results

In Table 2, we show the performance of different decoding algorithms as well as natural human continuation evaluated by automatic metrics. On both datasets, Look-back consistently achieves the highest MAUVE scores and coherence scores, which indicates that the generation of Look-back has token

---

We use GPT2-XL for text sequence embedding.

We disregard greedy and beam search as they kept producing repetitive phrases/sentences in prior studies (Welleck et al., 2019; Holtzman et al., 2019).

First 32 tokens are used as prefix for WikiText-103, while the original prompts are used for WritingPrompts.
distribution closeness with human continuations while staying relevant to the given prefixes. Meanwhile, Look-back is capable of producing texts with similar repetition and diversity level as the natural human text, which implies the fluency and informativeness of the generated text. We also notice that generations from all decoding algorithms obtain relatively low MAUVE and coherence scores on WritingPrompts. This is because the given prefixes are abstract and the human written references are diverse and varied, which results in low coherence and MAUVE w.r.t. various model continuations.

5.6 Human Evaluation

To further evaluate the quality of generated texts, we randomly sample two sets of 50 examples from WikiText-103 to produce prefixes for GPT2-XL and OPT-6.7B respectively and generate continuations from them. Then, we ask 3 evaluators to compare generated continuations from Look-back and the second best baseline SimCTG in two dimensions: 1) fluency: diverse and natural content without repeated words, phrases or sentences; 2) coherence: well-organized and easy to follow; being consistent with the topics presented in the human-written prefix without abrupt topic drifts. We ask annotators to choose one out of three options: the 1st continuation is better, the 2nd is better, or the two are of the same quality. As presented in Table 3, for both evaluation dimensions, the content generated by Look-back is preferred or marked as equally good by evaluators around or more than 70% of the time compared with baseline, which aligns well with the automatic metrics in Table 2.

5.7 Further Analyses

In this section, we analyze the effectiveness of different techniques used by Look-back individually.

Table 3: Human Evaluation on generations from Look-back and the second best SimCTG with examples sampled from WikiText-103. Continuation generated by Look-back is preferred to SimCTG significantly by human evaluators w.r.t both fluency and coherence.

<table>
<thead>
<tr>
<th>LM</th>
<th>Criterion</th>
<th>Look-back better (p-value)</th>
<th>same</th>
<th>SimCTG better</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT2-XL</td>
<td>Fluency</td>
<td>0.46 (.0127)</td>
<td>0.27</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>Coherence</td>
<td>0.57 (.0004)</td>
<td>0.16</td>
<td>0.27</td>
</tr>
<tr>
<td>OPT-6.7B</td>
<td>Fluency</td>
<td>0.38 (.0508)</td>
<td>0.37</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>Coherence</td>
<td>0.53 (.0078)</td>
<td>0.16</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Figure 3: Minimum KL divergence between current step and (a) history or (b) prefix from GPT2-XL and OPT-6.7B decoded by different algorithms on the test set of WikiText103 and WritingPrompts. Probability distribution of Look-back keeps distance to history to avoid repetitions but stays close to prefix to guarantee coherence.

Analyzing Probability Distribution Distance.

To verify whether decoding with Look-back appropriately constrains the probability distribution distance to past steps, we compare KL\text{in} to history and KL\text{in} to prefix of degeneration and different decoding algorithms in Figure 3. Although all improved decoding algorithms keep distance to historical probability distribution to avoid repetitions compared with greedy algorithm (Repetitive in the left column of Figure 3, the probability dis-
distribution of Look-back (Look-back in the right column of Figure 3) is much closer to the given prefix, which distinguishes it from off-topic continuation compared with other algorithms.

**Softmax vs. Uniform.** According to the softmax operation on $KL_{\text{min}}^C$ introduced in §4.2, the closer the next step’s probability distribution to prefix, the more likely the corresponding plausible token is selected to avoid undesired topic drift compared with random sampling. In Table 4, we empirically investigate the impact of plausible token sampling, uniform vs. softmax, on generation quality and find Look-back significantly enhances coherence on both datasets compared with random sampling. Although diversity drops with distribution distance-guided sampling in Look-back, both sampling strategies produce similar level of diverse content as human texts listed in Table 2.

**Effects of Candidate Amount and Threshold $\alpha$.** In §4.1, the hyperparameter $\alpha$ determines whether the current step is likely to produce repetitive continuation while $k$ restricts the range of plausible token candidates. The second best baseline SimCTG has the similar candidate amount parameter $k$ and the $\alpha$ to balance model confidence and degeneration penalty. When GPT2-2XL is used to decode with Look-back and SimCTG on WikiText-103, we visualize the impact of hyperparameters on generation quality in Figure 4 and Figure 5. The $\alpha$ in Look-back is different from that in SimCTG, but both control reliance on model confidence: a larger $\alpha$ indicates the most probable token is less likely to be adopted, hence more diversity is obtained. We also observe that for Look-back, the relevance of generated text to prefix (high coherence) and human continuation (high MAUVE) is much more robust to various hyperparameter values compared with SimCTG.

### 5.8 Case Study

Given a prefix sampled from WikiText-103, we present truncated human continuations as well as generations from Look-back and SimCTG in Table 5 and leave more examples in Appendix Table 6. The prefix is talking about the design of a race car game. Both human and Look-back continuations focus on describing major difficulties encountered during the game design, while SimCTG switches to a different topic by pointing to an online introduction of the game in the second half of continuation. Interestingly, Look-back explains how the team of more than twenty people was formed, which is coherent with the topic in the prefix.
6 Conclusion

The distance between output distributions signals potential failure modes of text generation, such as dull repetition and incoherence. We propose Look-back, a novel decoding algorithm that utilizes the KL divergence between the current and historic decoding steps, to steer the output distribution into a plausible subspace. Look-back can generate higher-quality text and outperforms several strong decoding algorithms in both automatic and human evaluation. However, KL divergence may not be the optimal measure for text output distributions and we leave the investigation of other measures to future work. In addition, the idea posed in this work can also be used for other specialized constrained decoding scenarios, such as preventing hallucination.

Limitations

We discuss the limitations of our work as follows:

- **Look-back** penalizes next tokens that result in low KL divergence with historic output distributions. However, we can not explicitly distinguish if such tokens are natural or undesired repetitions. This may lead to aggressive eliminations of possible outputs. We leave the distinction of different repetitions to future work.

- **Look-back** tends to show a higher bi-gram repetition score than other decoding methods because it encourages the coherence with prefix text at each decoding step. As we use a short prefix text following previous evaluation protocol, which might not be sufficiently informative, we will adopt a more comprehensive evaluation setup in the future or prepend relevant text in the beginning at decoding time.

- Most of our evaluations rely on automatic metrics, such as MAUVE scores. However, we found that these metrics may not truthfully reflect the quality of text, for example, MAUVE score is sensitive to the choice of sentence embedding models. In general, open-ended text generation still poses a great challenge to the development of NLG algorithms.
References


1048


1049
**Prefix:** A new show in London in 1912 showcased the best of music hall's talent. The Royal Command Performance took place at the Palace Theatre in London, which was

**Human:** managed by Alfred Butt. The show was organised by Oswald Stoll, an Australian impresario who managed a string of West End and provincial theatres. Stoll, although a fan of Lloyd's, disliked the vulgarity of her act and championed a return to a more family-friendly atmosphere within the music hall. Because of this, and her participation in the earlier music hall war, Stoll left her out of the line-up. He placed an advert in The Era on the day of the performance warning that "Coarseness and vulgarity etc are not allowed...

**SimCTG:** one of the most popular theatres in the West End at the time.

**Look-back:** the home of the Royal Commandos during the First World War. The show starred the likes of Harry Lauder, who played the role he was born to play, 'The King in Yellow', and Fred Karno - who, as the 'King', was the star attraction.

It was a huge success, and the Royal Variety Performance took its spot in the calendar. It's a tradition that continues to this day - and the King in Yellow is still a big draw at any show...

**Prefix:** The world is teetering on the brink of destruction. Only 2 people remain alive, both of them old ladies. They are face to face with the eldritch creature that has singlehandedly destroyed their planet. And in its terrifying presence... they make cheery small talk about what brought them here.

**Human:** FADE IN : EXT . A DESOLATE LANDSCAPE - AFTERNOON *A red sky covered with black clouds hangs above an expanse of blasted countryside. Dark craters pockmark the land, and the corrupted hulks of destroyed war machines lie rusting in the acrid air. Amidst this desolation, two elderly women sit on rocking chairs that are curiously undamaged. These are DELILAH and TABITHA. A small table with a tea set is present between them.

**SimCTG:** *"It's a bit hot today, isn't it?..."

**Look-back:** "Look at the two old ladies, who were not named in either the book or film, had been visiting from the future when their ship was destroyed. They are not aware of this, but one has become a time traveler from Earth and knows that they were never meant to be together in life but are now destined by fate together. One day they will meet again at last on some other planet where one of each woman will be born with powers over time which will allow them each one chance to save the world, which is now destined to fall into the hands of their enemy."

**Prefix:** For Francis, the good had always gone hand-in-hand with the bad. He had spent his life trying his best to be a good person, or at least what he believed would be a good person. He realized now that he was n’t really sure anymore. When his mother left his dad, Francis had chosen to stay with him. His friends had all told him he was smart to stay with the rich parent, but that had n’t been the reason, had it? At the time, he had told himself that his father needed him more. He did. Dad had been really bad at taking care of himself. Without his mother, someone had to be there to deal with dad when he was lonely and angry and...

**Human:** The person is born with the ability not necessarily in the form of super strength or super speed (although they may have those abilities as well) but the superpower is the ability to see the future! They proceed to live their entire life without noticing or realizing it.

**SimCTG:** What if they have superpowers in real life?"isn’t me going to need proof of this before I believe it. You know, for science. It’s true. I saw it on the internet. So it must be true."

**Look-back:** The person is born with the ability not necessarily in the form of super strength or super speed (although they may have those abilities as well) but the superpower is the ability to see the future! They proceed to live their entire life without noticing or realizing it. They see what will happen and they do what they can so that it will come. And because they live their whole life without realizing they have this ability they are able live life to the fullest. Because the future is already set for them, their entire life unfolds according. Their whole life, from the moment of their birth, unfolds according as planned in their mind...

---

Table 6: Case study of instances sampled from WikiText-103 and WritingPrompts. Unnatural topic drifts are frequently observed in generations from SimCTG (in pink).