Factuality Enhanced Language Models for Open-Ended Text Generation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Pretrained language models (LMs) can easily generate text with nonfactual information. In this work, we measure and improve the factual accuracy of large-scale LMs for open-ended text generation. We design the FACTUALITYPROMPTS test set and metrics to measure the factuality of LM generations. Based on that, we study the factual accuracy of LMs with parameter sizes ranging from 126M to 530B. Interestingly, we find that larger LMs are more *factual* than smaller ones, although a previous study suggests that larger LMs can be less truthful in terms of *misconceptions*. In addition, popular sampling algorithms (e.g., top-p) in open-ended text generation can reduce the factuality due to the "uniform randomness" introduced at every sampling step. We propose a *factual-nucleus* sampling algorithm that dynamically adapts the randomness to improve the factuality of generation while maintaining quality. Furthermore, we analyze the inefficiencies of the standard training method in learning correct associations between entities from factual text corpus (e.g., Wikipedia). We propose a *factuality-enhanced* training method that uses TOPICPREFIX for better awareness of facts and sentence completion as the training objective, which vastly reduces the factual errors from LMs.

1 Introduction

Large-scale pre-trained language models (LMs) have demonstrated impressive natural language generation results [1–4]. However, the generative LMs (e.g., GPT-3) are solely trained to model the statistical correlations between subword tokens [5], and have limited capability to generate factually accurate text as illustrated in Table 1. As a result, there are increasing concerns about the nonfactual generations from large-scale pre-trained LMs [e.g., 6, 7], which needs to be adequately addressed for their safe deployment in real-world applications, e.g., content creation [8] and dialogue [9].

In previous studies, different metrics and methods have been proposed to measure and improve the factual accuracy of language generation within different tasks [10], including text summarization [e.g., 11–14], question answering [e.g., 15–17], and table-to-text generation [e.g., 18, 19]. However, these works focus on the factuality (or faithfulness) of the *fine-tuned* LMs for particular downstream tasks (i.e., factual consistency between source and target text). Little exploration is made to address the factual errors in *pretrained* LMs for general-purpose open-ended text generation, where the goal is to generate a coherent continuation from the given context (e.g., the use cases from GPT-2).

One of the popular methods for enhancing generation factuality is to incorporate external knowledge sources [20–22]. Structured knowledge bases and graphs have been utilized for grounded text generation for years [e.g., 23, 24], where the LMs are trained to select and copy relevant facts from external knowledge sources. In contrast to the sizeable online text for pretraining, the structured knowledge graphs only encode a limited amount of knowledge as they require expensive human annotations for high-quality construction. A method that can directly leverage plain text knowledge (e.g., Wikipedia,

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Table 1: Example of continuations from the 530B LM with greedy decoding and nucleus sampling $p = 0.9$, along with the continuation from factuality-enhanced 530B LM with *factual-nucleus sampling*. Red represents nonfactual, green represents factual, and ~~strikethrough~~ represents repetition. See Appendix C for more examples.

[Factual Prompt] One of lung cancer's symptoms is shortness of breath. **[Greedy]** The American Cancer Society estimates that more than 228,000 people will be diagnosed with lung cancer in the United States this year.\n\n ~~The American Cancer Society estimates that more than 228,000 people will be diagnosed with lung cancer in the United States this year.\n\nThe American...~~
[p=0.9] That would make an oxygen mask one of the more popular treatments for this devastating disease. It helps ease breathing and give patients back their strength. Learn more here.<|endoftext|>"
[Ours] Other symptoms include coughing (including coughing up blood), chest pain, hoarseness, weight loss, and loss of appetite. The symptoms of lung cancer may resemble those of a bad cold or pneumonia. A chest x - ray is necessary to confirm the diagnosis. The diagnosis is usually confirmed by biopsy, in...
[Non-Factual Prompt] Cancer can be prevented by drinking too much alcohol. **[Greedy]** \nCancer can be prevented by eating too much red meat.\n\n~~Cancer can be prevented by eating too much~~ sugar.~~\n\nCancer can be prevented by eating too much~~ fat.~~\n\nCancer can be prevented by eating too much~~ processed food.~~\n\nCancer can be~~...
[p=0.9] More than 90% of hepatocellular carcinoma is associated with alcohol use.\n\nAbout 9 percent of cancers can be prevented if we don't eat more than 1.5 ounces of processed meat each day.\nDiet...
[Ours] Cancer is often treated with some combination of radiation therapy, surgery, chemotherapy and targeted therapy. Pain and symptom management are an important part of care. Palliative care is particularly important in people with advanced disease. The chance of survival depends on the type of cancer and...

encyclopedia books, peer-reviewed publications) would be a desirable form of factuality enhancement as it can remove the human annotation bottleneck and easily scale up the amount of injected knowledge. Augmenting LM with an information retrieval (IR) system is one possible solution to leverage textual facts, however, at the cost of additional complexity and resource overhead to the model [25–27, 21, 9]. Therefore, we explore an IR-free method that enhances the innate factuality of LMs by further training the pretrained LM on a factually rich plain-text corpus.

In this work, we focus on measuring and improving the factuality of large-scale pre-trained language models (LMs) for open-ended text generation. Specifically, we make the following contributions:

1. We build the benchmark and metrics to measure the factual accuracy of pre-trained LM for open-ended text generation. Based on that, we systematically study the factual accuracy of LMs with parameter sizes ranging from 126M to 530B and find that large LMs have higher factual accuracy than smaller ones.
2. We study the decoding algorithms of LM in terms of factual accuracy. We unveil that the popular nucleus sampling algorithm [28] for open-ended text generation can easily mix up different named entities or randomly fabricate information due to the "uniform randomness" introduced at every decoding step. We propose *factual-nucleus* sampling algorithm that promotes generation factuality while maintaining the quality and diversity.
3. We explore training methods that can effectively leverage text corpus with rich facts (e.g., Wikipedia). We find that directly continuing the pre-training of LM on factual text data [29] does not guarantee the improvement of factual accuracy. We propose *factuality-enhanced* training to address the underlying inefficiencies of this baseline. Our method consists of i) an addition of a TOPICPREFIX that improves the awareness of facts during training, and ii) a sentence completion task as the new objective for continued LM pretraining [e.g., 29].
4. We demonstrate that the factual accuracy of large-scale LMs (up to 530B) can be significantly enhanced (e.g., NE_{Er} error reduced from 33.3% to 14.5%) after applying the proposed *factuality-enhanced* training with *factual-nucleus* sampling algorithm.

We organize the rest of the paper as follows. We discuss related work in § 2 and present our benchmark setup with evaluation protocol in § 3. We study the factual accuracy of LMs with respect to model size, prompt type, and choice of decoding algorithm in § 4. After that, we present *factual-nucleus* sampling algorithm in § 5, and *factuality-enhanced* training in § 6. We conclude the paper in § 7.

2 Related Work

Factuality vs. Model Size Lin et al. [30] propose the TruthfulQA benchmark to measure the falsehood generations from different sized LMs. The result suggests that bigger LMs pre-trained on

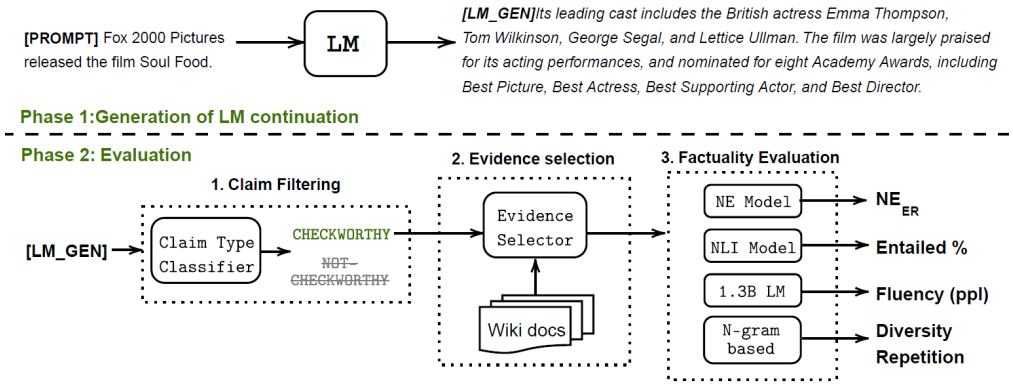


Figure 1: Illustration of our evaluation framework

web text are generally less truthful than smaller ones in terms of false belief or misconception. At first glance, this is contradictory to our observations, however, our work actaully focuses on different knowledge to TruthfulQA work. The TruthfulQA benchmark focuses on conceptual knowledge, while our benchmark focuses on factual knowledge [31]. [1] Large LMs can be good at recalling factual knowledge given substantial pre-training corpus, suggested by previous studies on LM's parameteric knowledge [32], but there still remains room for improvement for reasoning conceptual knowledge [33, 34].

Parametric Factual Knowledge There is a line of work addressing the factual errors in the parametric knowledge of LMs that is acquired from pretraining corpus [35–37]. The correctness of the parametric knowledge is commonly tested in cloze-style question answering format [32] (e.g., Person X is born in __) Efforts are made to fine-tune the pre-trained LM to "inject" more knowledge and improve its ability to answer factual questions without consulting external knowledge source [16]. Moreover, some work attempts to edit and fix the factual errors [38–40]. However, it is unclear if the improvement of fine-tuned LM for QA-style task can help to mitigate factual errors in open-ended text generation task.

Hallucination in downstream NLG tasks There are active efforts to reduce the factual errors or unfaithfulness of task-specific LMs fine-tuned for various downstream NLG tasks such as umma-rization [41–47], data-to-text [48, 49, 19, 50–52] and dialogue system [53–57]. In contrast to these works. we focus on general purpose LM for open-ended text generation task.

Human-in-the-loop Human feedback or demonstrations are valuable to improve the factual accuracy of LMs. For example, InstructGPT [58] fine-tune the LMs with collected human feedback for a truthful generation. WebGPT [7] is trained to cite its sources when it generates output, thus allowing humans to evaluate factual accuracy by checking whether a claim is supported by a reliable source.

3 FACTUALITYPROMPTS and Evaluation Metrics

Our goal is to automatically measure and evaluate the factuality of large-scale pre-trained language models (LMs) for open-ended text generation. Factuality refers to being coherent to provided ground-truth knowledge sources in NLP [10]. The biggest challenge of evaluating factuality for open-ended text generation is associated with locating the ground-truth knowledge from the massive amount of world knowledge. Evaluating open-ended text generation can be challenging due to the lack of ground-truth references for generation [59, 28]. In this study, we scope our ground-truth knowledge source to be Wikipedia because this will help simplify the evaluation setup. Note that Wikipedia is one of the most commonly-used, accessible, large-scale, good quality, unstructured knowledge sources. Our proposed methods can easily generalize to other knowledge sources in plain text (e.g., arXiv papers, medical reports, reliable newspapers).

Our evaluation framework consists of the following steps, as illustrated in Fig 1. In phase 1, LM generates the continuation using test prompts (§3.1). In phase 2, we first identify check-worthy

[1] According to Krathwohl [31], knowledge can be categorized into four types: i) factual knowledge, ii) conceptual knowledge, iii) procedural knowledge, and iv) metacognitive knowledge.

continuations [2], which refers to the generations that do contain facts and require factuality evaluation. This step is necessary as open-text generation does not always generate text that contains facts such as personal opinions and chitchat-style text (e.g., "I like eating an apple!"). Then, we prepare relevant ground-truth knowledge required for factual verification of check-worthy continuations (§3.2). Lastly, we obtain the factuality measures and the quality measures (§3.3).

3.1 FACTUALITYPROMPTS Testset

We design our test prompts, FACTUALITYPROMPTS, that follows a similar setup as in RealToxicityPrompts [60], which has *toxic* and *nontoxic* prompts to evaluate the toxicity of LM continuations. FACTUALITYPROMPTS consists of *factual* and *nonfactual* prompts that allow us to study the impact of prompts' factuality on the LM continuation; this simulates the real-world scenario where input texts are not guaranteed to be factual.

Since high quality fact-related data collection requires a lot of human efforts, we instead utilize a well-established fact-related dataset, FEVER [61], to construct our factual and nonfactual prompts. FEVER is a fact-checking dataset consisting of claims that are SUPPORTED, REFUTED or unverifiable (NOTENOUGHINFO) by Wikipedia documents. These claims are created by annotators who were asked to alter or paraphrase the sentences from Wikipedia. We leverage the SUPPORTED and REFUTED claims from FEVER validation set [3] as the factual and nonfactual prompts, respectively. To further ensure the quality of the test set, we filter out claims that are not appropriate to serve as prompts – e.g., extremely short claims that are not enough to provide any context to the LM. The data statistics after filtering is reported in Appendix F. We will release the constructed FACTUALITYPROMPTS for future research.

3.2 Ground-Truth Knowledge Preparation

To evaluate the factuality of a given generation, we need to prepare relevant ground-truth knowledge. The required ground-truth knowledge can be either document-level or sentence-level, depending on the type of factuality metrics (discussed in §3.3). The correctness of factuality evaluation is crucially dependent on the correctness of the ground-truth knowledge. To ensure that our factuality evaluation is not distorted by the irrelevant provision of ground-truth knowledge, we do the following:

For **document-level** ground-truth knowledge, we directly use the Wikipedia document annotation provided in the FEVER dataset. This way, we can mitigate any potential error from automatic document retrieval. For **sentence-level** ground-truth knowledge, we do automatic sentence selection by using two different method to maximize the chance of recalling the relevant ground-truth knowledge. We treat the generated text as query q and Wikipedia sentences as a pool of candidates $C = \{c_1, c_2, c_3, ...c_N\}$ where N is the number of sentences in the Wikipedia document. One ground-truth sentence is retrieved by obtaining TFIDF vector representations of q and C and selecting the c_i with the highest cosine similarity with the q. Another is retrieved by obtaining the contextual representation of q and C using SentenceTransformer [62] and selecting the c_j with the highest cosine similarity.

3.3 Evaluation Metrics

We adapt commonly used metric designs from the hallucination literature that are shown to have good correlation with human judgement [10]: named-entity (NE) based metric and entailment based metric. Each metric captures a different aspect of factuality, so we use both metrics for better understanding of factuality.

Hallucinated NE Error Named entities are one of the core building blocks of "fact", so NE-related metric design is one of the common choices in literature [63, 64, 10]. In this work, we specifically adopt the NE-based metric [64] that is designed with an assumption that a model is hallucinating (tending to make factual errors) if it generates a NE that does not appear in the ground-truth knowledge source.

We define our NE-based metric to be: $\text{NE}_{\text{ER}} = |\text{HALLU}_{\text{NE}}| \ / \ |\text{ALL}_{\text{NE}}|$ where ALL_{NE} is the set of all the NEs detected in the LM generation, and HALLU_{NE} is subset of NE_{All} that does not appear in the

[2] Refer to Appendix D for details

[3] The testset is not publicly released and can only be accessed through the FEVER workshop submission site. Therefore, it is common practice to leverage validation set instead.

Table 2: The factuality of LMs with different parameter size from 126M to 530B. NE_{ER} refers to the named-entity error, Entail refers to entailment ratio, Div. refers to distinct 4-grams, and Rep. refers to repetition. ↑ means the higher the better, and ↓ means the lower the better.

Size	Decode	Factual Prompt				Nonfactual Prompt			
		NE_{ER}↓	Entail↑	Div.↑	Rep.↓	NE_{ER}↓	Entail↑	Div.↑	Rep.↓
126M	p=0.9	63.69%	0.94%	0.90	0.58%	67.71%	0.76%	0.90	0.38%
	greedy	48.55%	8.36%	0.03	59.06%	54.24%	6.25%	0.03	59.90%
357M	p=0.9	56.70%	2.01%	0.87	0.55%	60.80%	1.42%	0.88	0.35%
	greedy	43.04%	14.25%	0.03	45.18%	46.79%	9.89%	0.04	46.30%
1.3B	p=0.9	52.42%	2.93%	0.88	0.24%	56.82%	2.04%	0.89	0.25%
	greedy	39.87%	12.91%	0.05	33.13%	45.02%	8.75%	0.05	36.20%
8.3B	p=0.9	40.59%	7.07%	0.90	0.11%	47.49%	3.57%	0.91	0.08%
	greedy	28.06%	22.80%	0.07	19.41%	32.29%	15.01%	0.07	13.26%
530B	p=0.9	33.30%	11.80%	0.90	0.13%	40.49%	7.25%	0.92	0.08%
	greedy	**20.85%**	**31.94%**	0.08	15.88%	27.95%	19.91%	0.08	16.28%

ground-truth Wikipedia document. Note that evaluating NE_{ER} requires document-level ground-truth. To ensure the quality of the metric, we also take the same precautions used by [64]. For named entities consisting of multiple words, partial n-gram overlaps are also treated as a "match". This ensures we can address the shortened form of named entities – e.g., "Barack Hussein Obama II" vs. "Obama". Note that stopwords (e.g., the, a) are not considered in the partial n-gram overlaps. The named entities are detected using a publicly available pre-trained NE detection model from Spacy [4].

Entailment Ratio Textual Entailment (or natural language inference) is a task of determining whether a hypothesis is entailed by, refuted by, or neutral to a given premise [65]. Entailment-based metrics are based on the rationale that factual generation will be entailed by the ground-truth knowledge [66, 11, 67, 68, 10].

We define the entailment ratio as: Entail= $|\text{ENTAIL}_{\text{gen}}|$ / $|\text{ALL}_{\text{gen}}|$, where ALL_{gen} is set of all generations, and $\text{ENTAIL}_{\text{gen}}$ is the set of generations that are entailed by a entailment model. To obtain the entailment scores, we leverage a pretrained entailment model that is publicly available; a RoBERTa [69] model fine-tuned on MNLI [70] dataset. Entail requires sentence-level ground-truth because only a few Wikipedia sentences are relevant to specific factual information in a given generation. For example, "Barack Obama was born in Hawaii" is only relevant to the Wikipedia sentence that mentions his birth location.

Generation Quality Evaluation We evaluate the generation quality on three aspects: *i) Fluency* is measured by the mean perplexity of generated continuations evaluated on a larger pretrained LM." *ii) Diversity* is an important characteristic of LM that makes the generation more interesting and engaging – it is bland and boring to always generate same texts. It is measured using the mean number of distinct n-grams (we report 4-gram), normalized by the length of text [71, 72] among the 10 generations for each prompt (i.e., in total, 160000 generations to evaluate the diversity of each method). *iii) Repetition* is a common form of degeneration. We measure the number of repetitive substrings that get generated at the end of the generations by using the publicly available metric code from Holtzman et al. [28].

4 Factuality Analysis of Pretrained LMs

In this section, we perform a factuality analysis of LMs from three aspects: *i)* model size, *ii)* prompt type and *iii)* decoding algorithm.

Model Size Researchers have observed the trend of larger LMs outperforming smaller ones in various downstream tasks [73, 3, 2]. However, contradicting to these general observations, recent studies suggest that larger LMs are actually less truthful in human misconceptions [30] and zero-shot fact-checking performance tend to stagnate with LM scaling [6]. We study the factuality of LMs with a range of parameter sizes (126M, 357M, 1.3B, 8.3B, 530B) to understand whether such surprising trend also applies to open-ended text generation task. As shown in Table 2, generation factuality does

[4] https://spacy.io/

improve with parameters scaling, e.g., NE_{ER} drops from 63.99% to 33.30% when parameter size scales up from 126M to 530B.

Prompt Type Prompts provided to the LM are known to significantly affect the quality and characteristics of LM continuations [74, 60]. We use our factual and nonfactual prompts to test the behavior of LMs. Results in Table 2 show that both factual and nonfactual prompts can lead to nonfactual generations, although factual prompts always result in less nonfactual generations. Interestingly, the performance gap between factual and nonfactual prompts gets more prominent as the model size increases (4% to 7% in NE_{ER} as parameter size increases from 126M to 530B). This could be because the larger LM better understands the prompts and can imitate the factual or nonfactual prompts in the continuations.

Decoding Algorithm We investigate the choice of decoding algorithms and their impact on the factual accuracy of generations. In particular, we compare two representative decoding algorithms that are *greedy decoding* (i.e., maximize generation likelihood) and *nucleus sampling* [28]. Nucleus sampling (a.k.a. top-p decoding) samples only from the top subword candidates with total probability p. It is a popular decoding algorithm for open-ended text generation because it solves many well-known degeneration problems of the greedy decoding algorithm. However, our results (Table 2) shows that top-p decoding under-performs greedy decoding in terms of factuality, although it achieves high generation diversity and quality. This intuitively makes sense because the sampling-based word choice in nucleus sampling can be seen as adding "randomness" to encourage diversity, which as a result, can lead to factual errors. Note that the factuality of a sentence can be easily altered by one-word choice. For example, "Barack Obama was born in 1961" can turn nonfactual if "1961" is changed to "1962". In the same sense, greedy decoding is more factual because its way of choosing the word with the highest probability minimizes randomness and maximizes the utilization of parametric knowledge of LM [32, 35]. However, greedy decoding achieves such factuality at the cost of generation diversity and quality.

Error Types We conduct a qualitative analysis of the factual errors from greedy generation of 530B LM, to understand what are the remaining errors when the randomness from decoding choice is strictly restricted. The two notable error types were:

- **Named Entity Mix-up**: Mixing up similar types of the named entity. For example, LM generated "*The movie is based on the novel of the same name by Gayle Forman.*" about a film called "*The Best of Me*". However, the correct author's name is "Nicholas Sparks", not "Gayle Forman". Note that Gayle Forman is also an American young adult fiction author who writes similar type of novels as Nicholas Sparks.
- **Fact Invention:** Inventing random new facts. For example, "*Samuel Witwer's father is a Lutheran minister*."

Both error types can be viewed as wrong associations of entities that appear at different parts of the training corpus with similar context. Such behavior is unsurprising because these LMs are uniformly trained with the next subword token prediction objective instead of a fact-related objective.

5 Factual-Nucleus Sampling

5.1 Method

From our analysis in §4, we observed that the sampling algorithm is a double-edged sword that contributes to both factual errors and repetitions. We hypothesize that sampling is more harmful to factuality when it is used to generate the latter part of a sentence than the beginning of a sentence. There is no preceding text at the start of a sentence, so it is safe for LM to generate anything as long as it is grammatical. However, as the generation proceeds, the context and premise become more determined, and fewer word choices can make the sentence factual. Given the example "*Samuel Witwer's father is a Lutheran minister*", the beginning of the sentence "*Samuel Witwer's father is*" is not nonfactual. However, the continuation of "*Lutheran minister*" makes the sentence nonfactual. Therefore, we propose a new sampling algorithm, *factual-nucleus sampling*, that dynamically adapt the "nucleus" p, along the sentence with the following techniques:

- **λ-decay**: Given that top-p sampling pool is selected as a set of words whose cumulative probability exceeds p, we gradually decay the p value with decay factor λ at each generation step to reduce the "randomness" through time.

Table 3: **1.3B** LM results with different decoding algorithms. NE_{ER} refers to named-entity error, Entail refers to Entail, Div. refers to distinct 4-grams, and Rep. refers to repetition. ↑ means the higher, the better, and ↓ means the lower, the better. p, λ and ω are nucleus probability, decay factor, and decay lowerbounds respectively.

Decoding	**Factual Prompt**				**Nonfactual Prompt**			
	NE_{ER}↓	Entail↑	Div.↑	Rep.↓	NE_{ER}↓	Entail↑	Div.↑	Rep.↓
Greedy	39.9%	12.9%	0.05	33.1%	45.0%	8.8%	0.05	36.2%
Top-p 0.9	52.4%	2.9%	0.88	0.2%	56.8%	2.0%	0.89	0.3%
p | λ	Top-p + λ-decay							
0.9 | 0.9	41.1%	10.8%	0.43	30.7%	45.7%	6.8%	0.47	34.5%
0.9 | 0.5	39.9%	13.0%	0.08	33.1%	44.9%	9.1%	0.09	35.9%
p | λ	Top-p + λ-decay + p-reset							
0.9 | 0.9	41.5%	10.3%	0.52	10.3%	45.4%	6.3%	0.57	9.1%
0.9 | 0.5	39.3%	12.8%	0.34	17.8%	44.5%	8.4%	0.45	18.9%
p | λ | ω	Top-p + λ-decay + p-reset + ω-bound (*factual-nucleus sampling*)							
0.9 | 0.9 | 0.3	42.1%	10.1%	0.55	7.1%	46.5%	5.6%	0.59	6.4%
0.9 | 0.5 | 0.3	41.0%	12.2%	0.47	13.0%	46.0%	7.0%	0.51	12.7%
0.9 | 0.9 | 0.2	41.7%	9.9%	0.52	8.6%	45.6%	6.2%	0.56	7.6%
0.9 | 0.5 | 0.2	39.3%	12.8%	0.38	16.1%	45.2%	7.8%	0.42	16.9%

- **p-reset**: The nucleus probability p can quickly decay to a small value after a long generation. So, we reset the p-value to the default value at the beginning of every new sentence in the generation (we identify the beginning of a new sentence by checking if the previous step has generated a full-stop). This reduces unnecessarily sacrifice of diversity of any long generations.
- **ω-bound**: If λ-decay is applied alone, the p-value would quickly become too small to be equivalent to greedy decoding and hurt diversity. To overcome this, we introduce a lower-bound ω to limit how far p-value can be decayed.

In *factual-nucleus sampling*, the nucleus probability p_t to generate the t-th token for each sentence is,

$$p_t = \max\{\omega,\ p \times \lambda^{t-1}\},$$

where λ is the decay factor for top-p probability, and ω lower bounds the decay of probability.

5.2 Result

We report our decoding experimental results with 1.3B LM [5] in Table 3. Additions of λ-decay helps improve top-p 0.9 factuality result – for instance, with decay rate λ=0.5, there is 12.5% drop in NE_{ER} and 10.1% gain in Entail. However, this affects the diversity and repetition to become similar to greedy decoding. Incorporation of p-reset helps mitigate this diversity issue without losing much in factuality metric. The effect is more drastic for the λ=0.5 option, where it achieves 0.26 gains in diversity metric with negligible changes in factuality scores. By also adding ω-bound, we obtain the anticipated factuality performance (i.e., similar to greedy decoding), with great improvement in generation quality over greedy; with p=0.9, λ=0.9, ω=0.3, we achieve $\times$11 improvement in diversity and $\times$4.6 improvement in repetition over greedy. Although our factual-nucleus sampling still under-performs top-p 0.9 in terms of diversity, we believe this is a reasonable trade-off to improve the factuality of LM for factually sensitive open-ended generation tasks. Moreover, our proposed decoding does not harm the sentence fluency; its perplexity do not exceed the perplexity of top-p. Refer to Appendix H for full perplexity results.

6 Factuality-enhanced Continued Pretraining

This section introduces our factuality-enhanced method for continued pretraining of LMs [29]. We first introduce the TOPICPREFIX for better awareness of facts. Then, we present the sentence completion task as the training objective.

[5] 1.3B LM is mainly used as it is big enough to have good learning capacity yet not too resource expensive.

Table 4: Results for factuality enhanced training. The decoding settings are formatted as: nucleus p value, decay rate λ, lower-bound ω

Decoding	Factual Prompt				Nonfactual Prompt			
(p | λ | ω)	NE_{ER}↓	Entail↑	Div.	Rep.	NE_{ER}	Entail	Div.	Rep.
Vanilla Pretrained LM (1.3B)								
0.9	52.4%	2.9%	0.88	0.2%	56.8%	2.0%	0.89	0.3%
0.9 | 0.9 | 0.3	42.1%	10.1%	0.55	7.1%	46.5%	5.6%	0.59	6.4%
Factual Domain-Adaptive Training with Wikipedia (1.3B)								
0.9	52.5%	2.8%	0.85	0.2%	55.8%	2.2%	0.86	0.1%
0.9 | 0.9 | 0.3	42.7%	7.1%	0.51	7.2%	48.2%	4.9%	0.56	6.0%
TopicPrefix (1.3B)								
0.9	34.4%	4.2%	0.84	0.3%	36.2%	2.7%	0.85	0.2%
0.9 | 0.9 | 0.3	27.6%	8.7%	0.43	8.0%	30.5%	6.1%	0.47	6.9%
TopicPrefix + SC_{ROOT} (1.3B)								
0.9	32.5%	6.7%	0.83	1.2%	34.3%	4.6%	0.84	1.1%
0.9 | 0.9 | 0.3	24.7%	15.8%	0.40	13.6%	27.6%	9.1%	0.44	13.7%
TopicPrefix + SC_{RANDOM} (1.3B)								
0.9	32.0%	7.9%	0.81	1.2%	34.2%	5.5%	0.83	1.1%
0.9 | 0.9 | 0.3	23.6%	17.6%	0.39	14.2%	26.9%	9.3%	0.42	13.2%
TopicPrefix + SC_{HALF} (1.3B)								
0.9	31.6%	7.6%	0.81	1.4%	33.5%	5.1%	0.83	1.5%
0.9 | 0.9 | 0.3	23.6%	17.4%	0.38	14.4%	27.2%	10.2%	0.42	13.1%
Vanilla Pretrained LM (530B)								
0.9	33.3%	11.8%	0.90	0.1%	40.5%	7.25%	0.92	0.1%
TopicPrefix + SC_{HALF} (530B)								
0.9	18.3%	19.3%	0.68	0.1%	21.7%	13.7%	0.68	0.1%
0.9 | 0.9 | 0.3	**14.5%**	**25.5%**	0.33	0.2%	**17.7%**	**20.0%**	0.33	0.1%

6.1 Prepending TopicPrefix

Unstructured factual knowledge typically exists at a document level (i.e., a group of factual sentences about an entity or things). This means that sentences can contain pronouns (e.g., she, he, it), making these sentences factually useless standalone. To illustrate with a concrete example from Barack Obama's Wikipedia page, "He previously served as a U.S. senator from Illinois from 2005 to 2008" cannot be a useful standalone fact because it is unclear who "He" is. It is essential to know the reference of the individual sentence-level facts.

If LMs could be trained at the full-document level, this would not be an issue. However, due to the GPU memory limit for long documents and computation efficiency, it is common to chunk documents before feeding them into the LMs during training. This causes the "fragmentation" of information and leads to wrong associations of entities that appear in independent documents with similar contexts. As a remedy, we propose to prepend TopicPrefix to sentences in the factual documents to make each sentence serve as a standalone fact. In our experiments, we mainly utilize Wikipedia documents as the factual corpus and the Wikipedia document name as the TopicPrefix .

6.2 Sentence Completion Loss

We propose a sentence completion loss to address the incorrect association learned between entities. To explain our rationale behind this, let us recall our nonfactual example from §5: "*Samuel Witwer's father is a Lutheran minister*". This sentence became nonfactual because LM failed to generate factually correct information after "*Samuel Witwer's father is*". In other words, LM failed to accurately *complete* the sentence given the already generated tokens. One reason is that the LM is uniformly trained to predict each subword token within the sentence, when ensuring the correct prediction at the

latter section of sentence is more critical for factuality. Therefore, we construct a sentence completion loss, which makes the LM focus on predicting the subwords later in the sentence. For implementation, we determine a pivot t for each sentence, and then apply zero-masking for all token prediction losses before t. We emphasize that this loss masking is different from the input token masking applied in BERT [73] or BART [75], and the LM is still trained in an autoregressive manner. Note that many BART-based summarization models are known to still suffer from factual errors, suggesting that masked prediction at the encoder level may not effectively transfer well to text generation. We demonstrate that the proposed masked prediction task are really effective for reducing factual errors when it is combined with TOPICPREFIX .

In this work, we explore three strategies to determine the pivot t from simple to complex:

- SC_{HALF}: pivot $t = 0.5\times$ sentence-length.
- SC_{RANDOM}: random pivot, e.g., $t \sim \text{uniform}[0.25, 0.75]\times$ sentence-length.
- SC_{ROOT}: pivot $t =$ position of ROOT (relation) after dependency parsing.

Surprisingly, according to our results, the simplest SC_{HALF} performs on par with the complex ones.

6.3 Results

Experimental results for our factuality-enhanced continued pretraining are reported in Table 4, and experimental setups are reported in Appendix E.

Ineffectiveness of Domain Adaptive Training The pre-training corpus of LM contains both factual texts (e.g., Wikipedia) and nonfactual texts (e.g., rumors, fake news). The nonfactual domain of the training corpus could be the problem. Thus, we conduct a baseline experiment that does domain-adaptive training with strictly factual domain text only (i.e., Wikipedia). Interestingly, we discover that domain-adaptive training has minimal impact on the generation factuality.

Effect of TOPICPREFIX Continued pre-training of 1.3B LM with TOPICPREFIX preprocessed Wikipedia alone can already improve the factuality, especially in terms of NE_{ER}. For example, it reduces the $\text{NE}_{\text{ER}}$from 42.1% to 27.6% when we use the factual-nucleus decoding (0.9 | 0.9 | 0.3), which even outperforms the 1.3B with greedy decoding (NE_{ER}: 27.6% vs. 39.9%) with much less repetition (8.0% vs. 33.1%).

Effect of Sentence Completion Loss The proposed sentence completion loss further helps to improve the factuality, especially for the Entail. For example, when one uses factual-nucleus decoding on trained 1.3B model, TOPICPREFIX + SC_{HALF} can further improve Entail from 8.7% to 17.4% than TOPICPREFIX alone, while reducing $\text{NE}_{\text{ER}}$from 27.6% to 23.6%. Note that the results show consistent improvement across different pivot selection strategies, suggesting that the sentence completion loss is robust. In particular, the simplest SC_{HALF} performs as good as others or even better in terms of several metrics. Thus we recommend it as the default option.

530B vs 1.3B As expected, our method on 530B LM further reduces the factual errors and achieves the lowest NE_{ER} (14.5%) and the highest Entail (25.5%). Surprisingly, our method on 530B LM lead to less diverse generation than 1.3B LM despite the significant improvement in the generation quality (i.e., near perfect repetition scores 0.1% 0.2%). We conjecture that this is the trade-off between the factuality and diversity for 530B LM.

7 Conclusion

In this work, we establish a benchmark to measure and analyze factuality in open-ended text generation tasks. We propose factual-nucleus sampling that improves generation factuality at inference-time, and the combination of sentence completion loss and TOPICPREFIX pre-processing that improves factuality at training-time. We demonstrate that our methods are effective in improving the factuality. Lastly, our results shed light on the existence of the trade-off between diversity and factuality. We strongly believe this is an important insight that will help researchers make a better-informed decision about their model design - i.e., appropriately prioritize the desirable attribute of their LM (factuality vs. diversity) according to the final goal of their task. Potential future work would be to reduce the degree of the observed trade-offs.

References

[1] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

[2] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2019.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

[4] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. Using DeepSpeed and Megatron to train Megatron-Turing NLG 530B, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*, 2022.

[5] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *ACL*, 2016.

[6] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

[7] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

[8] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In *NeurIPS*, 2019.

[9] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. LaMDA: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

[10] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *arXiv preprint arXiv:2202.03629*, 2022.

[11] Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *EMNLP*, 2019.

[12] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *ACL*, 2020.

[13] Esin Durmus, He He, and Mona Diab. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *ACL*, 2020.

[14] Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O Arnold, and Bing Xiang. Improving factual consistency of abstractive summarization via question answering. In *ACL-IJCNLP*, 2021.

[15] Jun Yin, Xin Jiang, Zhengdong Lu, Lifeng Shang, Hang Li, and Xiaoming Li. Neural generative question answering. In *IJCAI*, 2016.

[16] Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*, 2020.

[17] Dan Su, Xiaoguang Li, Jindi Zhang, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. Read before generate! faithful long form question answering with machine reading. In *Findings in ACL*, 2022.

[18] Amit Moryossef, Yoav Goldberg, and Ido Dagan. Step-by-step: Separating planning from realization in neural data-to-text generation. *arXiv preprint arXiv:1904.03396*, 2019.

[19] Tianyu Liu, Xin Zheng, Baobao Chang, and Zhifang Sui. Towards faithfulness in open domain table-to-text generation from an entity-centric view. In *AAAI*, 2021.

[20] Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. A survey of knowledge-enhanced text generation. *arXiv preprint arXiv:2010.04389*, 2020.

[21] Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Dmytro Okhonko, Samuel Broscheit, Gautier Izacard, Patrick Lewis, Barlas Oğuz, Edouard Grave, Wen-tau Yih, et al. The web is your oyster–knowledge-intensive nlp against a very large web corpus. *arXiv preprint arXiv:2112.09924*, 2021.

[22] Peter West, Chris Quirk, Michel Galley, and Yejin Choi. Probing factually grounded content transfer with factual ablation. *arXiv preprint arXiv:2203.10133*, 2022.

[23] Sungjin Ahn, Heeyoul Choi, Tanel Pärnamaa, and Yoshua Bengio. A neural knowledge language model. *arXiv preprint arXiv:1608.00318*, 2016.

[24] Robert L Logan IV, Nelson F Liu, Matthew E Peters, Matt Gardner, and Sameer Singh. Barack's wife hillary: Using knowledge-graphs for fact-aware language modeling. In *ACL*, 2019.

[25] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. *arXiv preprint arXiv:2112.04426*, 2021.

[26] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. Kilt: a benchmark for knowledge intensive language tasks. *arXiv preprint arXiv:2009.02252*, 2020.

[27] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*, 2020.

[28] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *ICLR*, 2020.

[29] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: adapt language models to domains and tasks. In *ACL*, 2020.

[30] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *ACL*, 2022.

[31] David R Krathwohl. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218, 2002.

[32] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? In *EMNLP*, 2019.

[33] Carlos Aspillaga, Marcelo Mendoza, and Alvaro Soto. Inspecting the concept knowledge graph encoded by modern language models. In *Findings of ACL*, 2021.

[34] Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. Evaluating commonsense in pre-trained language models. In *AAAI*, 2020.

[35] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 2020.

[36] Zexuan Zhong, Dan Friedman, and Danqi Chen. Factual probing is [mask]: Learning vs. learning to recall. *arXiv preprint arXiv:2104.05240*, 2021.

[37] Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021.

[38] Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In *EMNLP*, 2021.

[39] Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. Towards continual knowledge learning of language models. *arXiv preprint arXiv:2110.03215*, 2021.

[40] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual knowledge in GPT. *arXiv preprint arXiv:2202.05262*, 2022.

[41] Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[42] Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. Multi-fact correction in abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 9320–9331, 2020.

[43] Luyang Huang, Lingfei Wu, and Lu Wang. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[44] Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*, 2021.

[45] Shuyang Cao and Lu Wang. Cliff: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, 2021.

[46] Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. Enhancing factual consistency of abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 718–733, 2021.

[47] Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941, 2021.

[48] Sam Wiseman, Stuart Shieber, and Alexander Rush. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263. ACL, 2017.

[49] Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2673–2679. ACL, 2019.

[50] Yixuan Su, David Vandyke, Sihui Wang, Yimai Fang, and Nigel Collier. Plan-then-generate: Controlled data-to-text generation via planning. *Findings of EMNLP*, 2021.

[51] Peng Wang, Junyang Lin, An Yang, Chang Zhou, Yichang Zhang, Jingren Zhou, and Hongxia Yang. Sketch and refine: Towards faithful and informative table-to-text generation. *ACL*, 2021.

[52] Clément Rebuffel, Marco Roberti, Laure Soulier, Geoffrey Scoutheeten, Rossella Cancelliere, and Patrick Gallinari. Controlling hallucinations at word level in data-to-text generation. *Data Mining and Knowledge Discovery*, pages 318–354, 2022.

[53] Lei Shen, Haolan Zhan, Xin Shen, Hongshen Chen, Xiaofang Zhao, and Xiaodan Zhu. Identifying untrustworthy samples: Data filtering for open-domain dialogues with bayesian optimization. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1598–1608, 2021.

[54] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. ACL, 2021.

[55] Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 704–718. ACL, 2021.

[56] Zeqiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, et al. A controllable model of grounded response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14085–14093, 2021.

[57] Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. *EMNLP*, 2021.

[58] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

[59] Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *NeurIPS*, 2021.

[60] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings in EMNLP*, 2020.

[61] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and verification. In *NAACL*, 2018.

[62] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[63] Ben Goodrich, Vinay Rao, Peter J Liu, and Mohammad Saleh. Assessing the factual accuracy of generated text. In *ACM SIGKDD*, 2019.

[64] Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. Entity-level factual consistency of abstractive text summarization. In *EACL*, 2021.

[65] Bill MacCartney and Christopher D. Manning. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528, Manchester, UK, August 2008. Coling 2008 Organizing Committee. URL https://aclanthology.org/C08-1066.

[66] Tobias Falke, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, 2019.

[67] Ondřej Dušek and Zdeněk Kasner. Evaluating semantic accuracy of data-to-text generation with natural language inference. *arXiv preprint arXiv:2011.10819*, 2020.

[68] Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. Evaluating groundedness in dialogue systems: The begin benchmark. *arXiv preprint arXiv:2105.00071*, 2021.

[69] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[70] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL*, 2018.

[71] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *NAACL*, 2016.

[72] Zhihong Shao, Minlie Huang, Jiangtao Wen, Wenfei Xu, and Xiaoyan Zhu. Long and diverse text generation with planning-based hierarchical variational model. *arXiv preprint arXiv:1908.06605*, 2019.

[73] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[74] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing nlp. In *EMNLP*, 2019.

[75] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Our work improves the factual accuracy of large-scale language models and will have positive social impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Our experiments are conducted on the DGX A100 servers with 8x A100 GPUs.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] We only use Wikipedia data that does not contain personally identifiable information or offensive content.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Human Evaluation: Correlation with Human Judgement

Table 5: Pearson correlation coefficients between human factuality annotation and our factuality metrics. p-values for all results are 0.00.

Annotation	Entail	NE_{ER}
Expert	0.81	-0.77
Majority-voting	0.47	-0.46

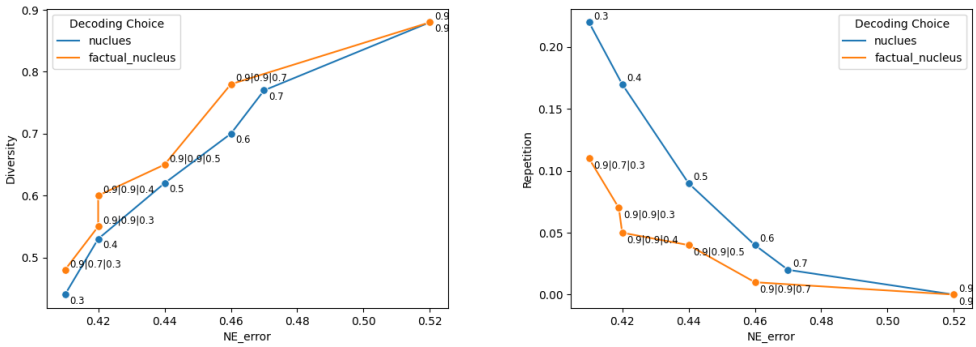
Although NE-based and entailment-based metrics have been used in downstream NLG tasks [10], they have not been utilized for evaluating factual accuracy in open-ended text generation. To ensure their validity, we collect human annotations to evaluate the correlation between our automatic factuality metrics with human judgement – i.e., are generations with higher Entail and lower NE_{ER} errors, more likely to be perceived as factual by human?

We obtained human annotations for 200 randomly chosen LM continuations of varying NE_{ER} and Entail scores. The annotators are asked to fact-check the LM continuations against Wikipedia and assign factuality label (1 = Factual : can find supporting Wikipedia evidence. 0 = Non-factual : cannot find supporting Wikipedia evidence).

The fact-checking annotation is a challenging and time-consuming task, as it requires the annotator to carefully read multiple evidences and reason over them. To improve the annotation quality, we have two types of annotations. The first type is two annotations from average English speaking workers on *Appen.com* platform, and the second type is one "expert" annotation from one of the authors who is familiar with the task and spent solid amount of time checking each samples. Based on these three annotations, we do majority voting and report the Pearson correlation results in Table 5. We also report the correlation result solely using the expert annotations, and show that there is strong correlation between human judgement of factuality and the proposed automatic metric NE_{ER} and Entail. NE_{ER} is negatively correlated with factuality because the lower the NE_{ER} error, the better the factuality.

B Nucleus (Top-p) Sampling vs. Factual-Nucleus Sampling

In this subsection, we compare the nucleus sampling and the proposed factual-nucleus sampling in terms of the generation factuality and quality. The figures below show that the factual nucleus sampling method has better factuality-vs-diversity and factuality-vs-repetition trade-offs.



(a) Diversity vs. NE_{ER}

(b) Repetition vs. NE_{ER}

Figure 2: Comparison between Nucleus Sampling (p: 0.3 , 0.4 , 0.5 , 0.6 , 0.7 , 0.9) and Factual Nucleus Sampling ($p|\lambda|\omega$: 0.9|0.7|0.3 , 0.9|0.9|0.3 , 0.9|0.9|0.4 , 0.9|0.9|0.5 , 0.9|0.9|0.7). The x-axis is named entity error NE_{ER}. The y-axes are diversity and repetition in (a) and (b) respectively. The lower the repetition, the better. For diversity, the diversity score of Wikipedia documents is 0.767.

C Generation Examples

We provide more generation examples from the pretrained 530B LM with greedy and top-p sampling ($p = 0.9$), and factuality-enhanced 530B LM with factual-nucleus sampling. Green indicates factual, red indicates nonfactual, and ~~striked text~~ indicates repetition. Disclaimer: Authors tried to exhaustively check the factuality of the following generations, however, there is no 100% guarantee about the annotations.

[Category] Medical
[Factual Prompt] One of lung cancer's symptoms is shortness of breath.

[Greedy] The American Cancer Society estimates that more than 228,000 people will be diagnosed with lung cancer in the United States this year.\n\n ~~The American Cancer Society estimates that more than 228,000 people will be diagnosed with lung cancer in the United States this year.\n\nThe American Cancer Society estimates that more than 228,000 people will be diagnosed with lung cancer in the United...~~

[p=0.9] That would make an oxygen mask one of the more popular treatments for this devastating disease. It helps ease breathing and give patients back their strength. Learn more here.<|endoftext|>"

[Ours] Other symptoms include coughing (including coughing up blood), chest pain, hoarseness, weight loss, and loss of appetite. The symptoms of lung cancer may resemble those of a bad cold or pneumonia. A chest x - ray is necessary to confirm the diagnosis. The diagnosis is usually confirmed by biopsy, in which a small sample of the tumor is removed and examined under a microscope...

[Category] Actor
[Factual Prompt] Leonardo DiCaprio played a supporting role in What's Eating Gilbert Grape.

[Greedy] The film was shot in the town of Endora, Iowa.\n\n~~The film was shot in the town of Endora, Iowa.\n\nThe film was shot in the town of Endora, Iowa.\n\nThe film was shot in the town of Endora, Iowa.\n\nThe film was shot in the town of Endora, Iowa.\n\nThe film was shot in the town of Endora, Iowa.\n\nThe film was shot in the town of Endora, Iowa.\n\nThe film was shot in the town of Endora, ...~~

[p=0.9] DiCaprio was desperate to have the lead role in the film. However, after realizing the amount of time that was required to learn all the lines, he gave the role to Johnny Depp instead.\n\nTitanic (1997)\n\nLeonardo DiCaprio and Kate Winslet were paid the same amount for their roles in Titanic.\n\nAlthough DiCaprio was relatively unknown when the film was released, Paramount Pictures paid him and Winslet the same amount.

[Ours] In 1993 , he played the mentally handicapped brother of Johnny Depp 's character in " What's Eating Gilbert Grape" , a comic - tragic odyssey of a dysfunctional Iowa family. Director Lasse Hallström admitted he was initially looking for a less good - looking actor but finally settled on DiCaprio as he had emerged as "the most observant actor" among all who auditioned . Budgeted at US$ 11 million , the film became a critical success, resulting in various accolades for DiCaprio , who was awarded the National Board of Review Award and nominated for both an Academy Award and a Golden Globe Award for his portrayal of Arnie Grape...

[Category] Company
[Factual Prompt] KFC is the second biggest restaurant chain in the world.

[Greedy] The company is owned by Yum! Brands, which also owns Pizza Hut and Taco Bell.\n\nKFC is known for its fried chicken, but it also offers other food items like chicken sandwiches, wraps, salads, and even rice bowls.\n\n~~KFC is known for its fried chicken, but it also offers other food items like chicken sandwiches, wraps, salads, and even rice bowls.\n\nKFC is known for its fried chicken, but it also offers other food items like chicken sandwiches, wraps, salads, and even rice bowls.\n\nKFC...~~

[p=0.9] While many restaurants worldwide have been growing their mobile commerce revenues, KFC chose to implement its payment system via cash. Customers must visit a traditional cash register to buy food.\n\nThe order-and-pay system that KFC had been using at its stores didn't allow customers to pay with their smartphones. Instead, customers have to stand in long lines to order and pay for their meals. The process involves taking a physical receipt, which is then placed at the counter, to process the order.\n\n...

[Ours] KFC was founded by Colonel Harland Sanders , an entrepreneur who began selling fried chicken from his roadside restaurant in Corbin , Kentucky , during the Great Depression . Sanders identified the potential of the restaurant franchising concept , and the first " Kentucky Fried Chicken " franchise opened in Utah in 1952 . KFC popularized chicken in the fast food industry , diversifying the market by challenging the established dominance of the hamburger . By branding himself as " Colonel Sanders " , Harland became a prominent figure of American cultural history , and his image remains widely used in KFC advertising to this day...

[Category] History
[Factual Prompt] San Francisco was first colonized by Spanish colonists.

[Greedy] The Spanish arrived in 1776 and established the Presidio of San Francisco in 1776. The Mission San Francisco de Asís was founded in 1776. The Mission was the first European settlement in the city. ~~The Mission was the first European settlement in the city. The Mission was the first European settlement in the city. The Mission was the first European settlement in the city. The Mission was the first European settlement in the city.~~

[p=0.9] A Spaniard, José Joaquin Moraga, who accompanied the De Anza expedition, named the bay, which he called "Puerto de San Francisco", on or about October 30, 1776. On January 12, 1777, the Presidio of San Francisco, founded by Moraga, was established by the Spanish. The Mission San Francisco de As0̆0eds, established in the same year, was the first Mission on the Pacific Coast of North America. In 1781, the settlers, numbering about 200...

[Ours] The Presidio of San Francisco , a Spanish military outpost , was established in 1776 . It was the first European settlement in what is now the city . The Mission San Francisco de Asís (Mission Dolores) , established in 1776 , was the city 's first major settlement for the next half - century . The California Gold Rush of 1849 brought rapid growth , making it the largest city on the West Coast at the time . San Francisco became a consolidated city - county in 1856 . San Francisco 's status as the West Coast 's largest city peaked between 1870 and 1900 , when around 25 % of California 's population resided in the city proper . After three - quarters of the city was destroyed by the 1906 earthquake and fire , San Francisco was quickly rebuilt...

[Category] Sports
[Factual Prompt] Michael Jordan helped popularize the NBA in the 80s and 90s.

[Greedy] The NBA is the most popular basketball league in the world.\n\n~~The NBA is the most popular basketball league in the world.\n\nThe NBA is the most popular basketball league in the world.\n\nThe NBA is the most popular basketball league in the world.\n\nThe NBA is the most popular basketball league in the world.\n\nThe NBA is the most popular basketball league in the world.\n\nThe NBA is the most...~~

[p=0.9] He was named the most valuable player in the NBA Finals six times. But the former Chicago Bulls star and current Charlotte Hornets owner has another MVP title to add to his list: MVP of Valentine's Day. An online sports book claims that Jordan is the most searched player of all time on Feb. 14.\n\n » Read more trending news\n\nFanDuel Sportsbook, an online gaming company, said it reviewed searches on its site for the last seven years and discovered that Jordan is the most searched-for NBA player on Valentine's Day, Forbes",

[Ours] Jordan 's individual accolades and accomplishments include six NBA Finals Most Valuable Player (MVP) Awards , ten scoring titles (both all - time records) , five MVP Awards , ten All - NBA First Team designations , nine All - Defensive First Team honors , fourteen NBA All - Star Game selections , three All - Star Game MVP Awards , three steals titles , and the 1988 NBA Defensive Player of the Year Award. He holds the NBA records for highest career regular season scoring average (30.12 points per game) and highest career playoff scoring average (33.45 points per game). Jordan is also known for his product endorsements . He fueled the success of Nike 's Air Jordan sneakers , which were introduced in 1984...

[Category] Movie
[Factual Prompt] In Too Deep is an American crime thriller film.

[Greedy] It was directed by Michael Rymer and released in 1999. The film stars Omar Epps, LL Cool J, Nia Long, Stanley Tucci, and Giancarlo Esposito. A young man named Michael (Omar Epps) is released from prison after serving a sentence for drug dealing. He is determined to go straight, but his friend Tommy (LL Cool J) is not so sure. Tommy is a drug dealer who has been working for a crime boss named Dwayne Gittens (Stanley Tucci). Gittens is a violent man who will not hesitate to kill anyone who crosses him.\n\n Michael is soon dragged back into the drug world when Tommy is arrested for killing a man who tried to rob him. Gittens tells Michael that he will free Tommy if Michael will sell a large amount of drugs...

[0.9] The plot concerns the federal government's attempts to lure a bank robber out of hiding by teaming up a convicted high-profile criminal with a U.S. Marshal. It stars Omar Epps and LL Cool J and is directed by Michael Rymer.\n\nThe film is also known as Deeper and In Too Deep. The film had a DVD release on December 26, 2000, and an HD DVD release on April 30, 2007.\n\nPlot\nRonnie Pullman (Omar Epps) is a thief who has made a name for himself by stealing from...

[Ours] It was directed by Michael Rymer from a screenplay written by Michael Henry Brown and Paul Aaron . The film stars Omar Epps , LL Cool J , Stanley Tucci , Pam Grier and Nia Long. Jeffrey Cole (Omar Epps) is a recent graduate of the Cincinnati police academy who dreams of working undercover. Cole manages to get an undercover assignment the day of graduation and earns the praise of his superiors . He is soon given the task of taking down statewide crack dealer Dwayne Gittens (LL Cool J) , an underworld boss so powerful that his nickname is "God"

[Category] Medical
[Non-Factual Prompt] Obesity and endometrial cancer have nothing to do with each other.

[Greedy] \n\nObesity is a risk factor for endometrial cancer. \n\nObesity is not a risk factor for endometrial cancer.\n\nCorrect\n\nIncorrect\n\nQuestion 3 of 10\n\n3. Question\n\nWhich of the following statements is true?\n\nEndometrial cancer is the most common gynecologic cancer in the United States.~~\n\nEndometrial cancer is the most common gynecologic cancer in the United States.\n\nEndometrial cancer is the most common gynecologic cancer in the United States.\n\nEndometrial cancer is the most common gynecologic cancer in the United States.~~

[p=0.9] \n\nThen why does it get such a high incidence rate with obese women?\n\nIt doesn't. The problem is that fat women are more likely to have abnormal uterine bleeding.\n\nAnd with an enlarged uterus, it may be difficult for a physician to detect a...

[Ours] Obesity increases the risk of endometrial cancer . This is thought to be due to higher levels of estrogens produced by the adipose tissue . Obesity also causes changes in the endometrium that make it more susceptible to cancer...

[Category] Sports
[Non-Factual Prompt] Michael Jordan was inactive in the 1980s and 90s.

[Greedy] The Chicago Bulls were inactive in the 1980s and 90s.\n\n~~The Chicago Bulls were inactive in the 1980s and 90s.\n\nThe Chicago Bulls were inactive in the 1980s and 90s.\n\nThe Chicago Bulls were inactive in the 1980s and 90s...~~

[p=0.9] This caused concern that the new coach was not doing anything right. But there was still hope because he promised the following: Ẅe're going to be active traders on the phones~~.~~ Except it turned out he really meant they were going to be doing a lot of long distance calling and give...

[Ours] Jordan 's first retirement in 1993 was very short - lived , as he returned to the NBA in March 1995 , wearing the number 45 , after his father was murdered .\n The Bulls were then eliminated in the second round of the playoffs by the Orlando Magic.\n Jordan was suspended by the NBA for...

[Category] TV
[Non-Factual Prompt] Modern Family is solely a Dutch television sitcom.

[Greedy] It is the first Dutch television series to be produced in HD. The series is produced by the Dutch broadcasting company Talpa, which had previously produced the successful series Tienstok. The series is written by Dutch screenwriter and director Joram Lürsen.\n\nThe series is about a...

[p=0.9] It follows the story of a quite stereotypical Dutch family (although unlike the Dutch, the family in this sitcom has only two kids and only the youngest is still a minor, whereas in reality most Dutch families have at least three kids). The majority of the episodes are just ordinary family life, but the writers do not...

[Ours] The show premiered on Dutch television on 18 September 2007 on RTL 4 . The show was originally created by Steve Levitan and Christopher Lloyd , who also serve as executive producers . The series is set in the fictional suburban town of Dunphy, California . The show follows the lives of the Dunphy family , consisting of Phil Dunphy (Eric Stonestreet) , Claire Dunphy (Julie Bowen) , Cameron Dunphy (Jesse Tyler Ferguson) , Mitchell Dunphy (Ty Burrell) , and Lily Dunphy (Sofia Vergara)...

[Category] Singer
[Non-Factual Prompt] Michael Jackson made his debut with his sister Janet.

[Greedy] \n\nThe Jackson 5 made their debut on the show in 1969.~~\n\nThe Jackson 5 made their debut on the show in 1969.\n\nThe Jackson 5 made their debut on the show in 1969.\n\nThe Jackson 5 made their debut on the show in 1969.\n\nThe Jackson 5...~~

[p=0.9] Two performances followed before Jackson would begin a solo career, while still continuing to perform with his brothers in The Jacksons.\n\n1977\n\nMichael Jackson releases his first solo album 'Off the Wall'. A critical and commercial success, 'Off the Wall' reached No.5 in the UK album...

[Ours] The Jackson 5 won several awards within the black music community , including the American Music Award for Favorite Soul / R&B Band / Duo in 1970 and 1972 . Jackson 's solo career began in 1971 while at Motown with his first solo album , " Got to Be There " , which was certified platinum . His second album , " Ben " , was released in 1972 . The album sold over a million copies , and the single " Ben " reached number one on the " Billboard " Hot 100 . Jackson 's third album , " Music & Me " , was released in 1973 . In 1975 , Jackson moved to Epic Records...

[Factual Prompt] The ruins of the ancient Roman town of Pompeii lie near Naples

[Greedy] The city was destroyed by a volcanic eruption in 79 AD.\n\nThe ruins of the ancient Roman town of Pompeii lie near Naples. The city was destroyed by a volcanic eruption in 79 AD.\n\nPhoto: AP\n\nImage 2 of 3\n\nA man walks past a mural depicting the ancient Roman town of Pompeii in Naples, Italy.\n\nA man walks past a mural depicting the ancient Roman town of Pompeii in Naples, Italy.\n\nPhoto: AP\n\nImage 3 of 3\n\nPompeii's ruins to be restored\n\n1 / 3\n\nBack to Gallery\n\nROME

[p=0.9] Millions of tourists from around the world come to visit them each year. They enter through a huge gate. This gate was built when Pompeii became a part of Italy again in the nineteenth century. Before that time, no one was allowed to enter the area.\n\nBut that is not the only gate through which visitors enter Pompeii.\n\nThe back of the entrance gate is very high. For a long time, no one has known why it is so tall. It seemed as if the builder simply did not finish it.\n\nOne day a

[Ours] Naples has a wide range of different architectural styles , from the medieval castles in the historic centre to the Art Nouveau buildings in the Vomero ärea , to the buildings of the 20th century , such as the San Carlo opera house . \nNaples is also widely known for its wealth of historical museums . \nThe Naples National Archaeological Museum is one of the city 's main museums , with one of the most extensive collections of artefacts of the Roman Empire in the world . \nIt also houses many of the antiques unearthed at Pompeii and Herculaneum , as well as some artefacts from the Greek and Renaissance periods . \nPreviously a Bourbon palace , now a museum and art gallery , the Museo di Capodimonte is another"

[Factual Prompt] David Beckham is a player for the Major League Soccer club LA Galaxy.

[Greedy] He is also a UNICEF ambassador.\n\nThe opinions expressed in this commentary are solely those of David Beckham.<|endoftext|>

[p=0.9] For England, he was a very successful player, playing for Man U (again and again), and captaining the team for several years, until he fell out of favor with his coach.\n\nIn 2003, after a free kick from England's David Beckham, Argentinean player and infamous cokehead Diego Simeone headbutted Beckham. According to Beckham, Simeone leered at him and said "I'm going to fuck you." Simeone claimed the players collided accidentally, but the ref gave him a red card for the incident, ...

[Ours] He is the first English player to win league titles in four countries : England , Spain , the United States and France . He retired in May 2013 after a 20-year career , during which he won 19 major trophies . Beckham 's professional club career began with Manchester United , where he made his first - team debut in 1992 at age 17 . With United , he won the Premier League title six times , the FA Cup twice , and the UEFA Champions League in 1999 . He then played four seasons with Real Madrid , winning the La Liga championship in his final season with the club .

[Factual Prompt] Andy Warhol has creations that are highly valuable.

[Greedy] His works are sold for millions of dollars. He is one of the most famous artists in the world.\n\nAndy Warhol was born in Pittsburgh, Pennsylvania. He was the youngest of three children. His parents were immigrants from Czechoslovakia. His father was a construction worker. His mother was a homemaker.\n\nAndy Warhol was a shy child. He was often sick. He had a nervous stomach. He was often in the hospital. He was afraid of the dark. He was afraid of the dentist. He was afraid of doctors. He was afraid of hospitals.\n\nAndy Warhol went to college. He studied art. He was a good student. He was a good artist. He was a good painter. He was a good drawer. He was a good illustrator. He was a good graphic artist...

[p=0.9] One of these creations is a book. During his days as an art student, Warhol used to paint advertisements that he printed on pieces of cardboard. He pasted these advertisements on a cork and this is what he called his book. For Warhol, making a book meant simply arranging his images on a page. He didn't make the necessary effort to bind the pages together to make it qualify as a book. This was the same way Warhol created many of his creations and works. He could have made every page of his work a masterpiece, but he didn't do...

[Ours] Warhol was an inspiration towards Cattelan 's magazine and photography compilations , such as " Permanent Food , Charley " , and " Toilet Paper " . \nIn the period just before his death , Warhol was working on "Cars", a series of paintings for Mercedes - Benz . \nA self - portrait by Andy Warhol (1963-64) , which sold in New York at the May Post - War and Contemporary evening sale in Christie 's , fetched $ 38.4 million . \nOn May 9 , 2012 , his classic painting " Double Elvis (Ferus Type) " sold at auction at Sotheby 's in New York for US$ 33 million . \nWith commission , the sale price totaled US$ 37,042...

[Factual Prompt] San Francisco is the location of a bridge.

[Greedy] The bridge is located in San Francisco.\n\nThe bridge is located in San Francisco.\n\nThe bridge is located in San Francisco.\n\nThe bridge is located in San Francisco.\n\nThe bridge is located in San Francisco.\n\nThe bridge is located in San Francisco.\n\nThe bridge is located in San Francisco.\n\nThe bridge is located in San Francisco.\n\nThe bridge is located in San Francisco.\n\nThe bridge is located...

[p=0.9] San Francisco is an island surrounded by bridges.\n\nOnce across the bridges, there is much to explore. Chinatown, for example, has interesting shops, food, and people-watching. Walk in Chinatown and you will hear many languages, see many signs in different scripts, and smell interesting things from kitchens. From Chinatown you can see the world.\n\nBut Chinatown is not what this is about. What this is about is part of San Francisco which is on the east side of the city, about three miles from Chinatown...

[Ours] There are seven major islands in the San Francisco Bay : Alameda , Marin , Angel , Treasure , Yerba Buena , Contra Costa and Alcatraz . Alcatraz , in the middle of the bay , is a federal penitentiary which also housed a lighthouse for many years . It is now a public museum and one of San Francisco 's major tourist attractions , attracting some 1.5 million visitors annually . Yerba Buena Island , between San Francisco and Oakland , contains the San Francisco Port of Embarkation , the Oakland Naval Supply Depot , and a Coast Guard station ...

[Factual Prompt] Michael Jordan was in "The Dream Team".

[Greedy] The Dream Team was the nickname of the USA men's basketball team that won the gold medal at the 1992 Olympics.\n\nThe Dream Team was the nickname of the USA men's basketball team that won the gold medal at the 1992 Olympics.\n\nThe Dream Team was the nickname of the USA men's basketball team that won the gold medal at the 1992 Olympics.\n\n

[p=0.9] Mickey Mouse is Disney.\n\nSophia Loren was an actress from Italy.\n\nLast but not least is Harry Potter.\n\nAll these brands immediately create a strong visual image in the mind of the target audience, and this is the same for the faith-based ministry.\n\nTo be seen as an iconic faith-based ministry the church must have a look and feel that is instantly recognizable.\n\nDo you have a recognizable logo?\n\nDoes your website and social media pages look like the brand of your church?\n\nDo you have

[Ours] In the summer of 1992 , Jordan led the Dream Team to the Olympic gold medal in Barcelona , scoring a game - high 26.5 points per game , and making a game - winning jump shot against Croatia . Jordan was named the tournament 's MVP , although the Olympic basketball competition was later considered to be an exhibition for NBA players , and thus NBA players were not allowed to compete in future Olympic basketball tournaments . In March 1993 , Jordan announced his retirement , citing a loss of desire to play the game . Jordan later stated that the death of his father three months earlier also shaped his decision . Jordan 's father was murdered on July 23 , 1993 , at a highway rest area in Lumberton , North Carolina , by two teenagers , Daniel Green and Larry Martin Demery , who carjacked his luxury Lexus bearing the license plate ÜNC 0023 . His body was dumped in a South Carolina swamp and was not discovered until August 3 .

[Factual Prompt] Cara Delevingne had a role in a film.

[Greedy] Cara Delevingne had a role in a film. Cara Delevingne had a role in a film. Cara Delevingne had a role in a film. Cara Delevingne had a role in a film. Cara Delevingne had a role in a film. Cara Delevingne had a role in a film. Cara Delevingne had a role in a film. Cara Delevingne had a role in a film.

[p=0.9] She's actually one of the many faces in a sketch for a film by celebrated photographer David LaChapelle called "Lost and Found." And I'm not going to lie to you, it's kind of terrifying, but in a good way? I think.<|endoftext|>"

[Ours] Delevingne's first film appearance was in the 2012 film adaptation of "Anna Karenina", as Princess Sorokina, the marriage interest of Count Vronsky, alongside Keira Knightley. In August 2013, Delevingne voiced a DJ of a pop radio station in the video game "Grand Theft Auto V," which once released became the fastest - selling entertainment product in history.

[Factual Prompt] Doctor Who is a show on television.

[Greedy] It is a science fiction show. It is a show about a man who travels through time and space in a blue box. It is a show about a man who is over 900 years old. It is a show about a man who is a Time Lord. It is a show about a man who is a Time Lord who has a time machine called the TARDIS. It is a show about a man who is a Time Lord who has a time machine called the TARDIS that is bigger on the inside than it is on the outside. It is a show about a man who is a Time Lord who has a time machine called the TARDIS that is bigger on the inside than it is on the outside and that is shaped like a police box.

[p=0.9] The men on the show are the Doctor and the Master. The women are the companions, and they are held captive until they are rescued, or dead. When they are rescued, it's rarely by the Doctor. Even when the Doctor is in the scene, the female characters are passively waiting to be saved. Here are a few examples. \n\nSeason 5. Episode 13. "The Lodger"\n\nhttp:www.youtube.comwatch?v=bHNce0_Qlrk\n\nHere's a fun little game.

[Ours] Doctor Who is a British science fiction television programme produced by the BBC since 1963 . The programme depicts the adventures of a Time Lord called " the Doctor " , an extraterrestrial being , to all appearances human , from the planet Gallifrey . The Doctor explores the universe in a time - travelling space ship called the TARDIS . Its exterior appears as a blue British police box , which was a common sight in Britain in 1963 when the series first aired . Accompanied by a number of companions , the Doctor combats a variety of foes while working to save civilisations and help people in need . The show is a significant part of British popular culture , and elsewhere it has gained a cult following...

[Factual Prompt] Tiber Oil Field is operated by BP.

[Greedy] See also List of oil fields References Category:Oil fields in Kern County, California\nCategory:Geography of the San Joaquin Valley\nCategory:BP oil and gas fields<|endoftext|>

[p=0.9] The rig is located approximately off the Louisiana coast. It is a riserless subsea design, with a maximum structural height. This is the first use of this design. Operations Production on the field was expected to begin in 2011, but first oil was delayed until 2014. Oil is produced through a wellhead located on the seafloor, and it is then piped up to a floating production storage and offloading (FPSO) vessel. Oil from this vessel is then transported to shore. The project is being developed in...

[Ours] Tiber Oil Field is located in the Keathley Canyon block 102 of the United States sector of the Gulf of Mexico . Tiber is only one of four deepwater fields in the Gulf of Mexico that is operated by BP . The oil field was discovered in September 2009 and it is located in Keathley Canyon block 102 about southeast of Houston and south west of New Orleans . Tiber is BP 's first ultra - deepwater platform in the Gulf of...

[Factual Prompt] The Blue Lagoon was released in or before 1930.

[Greedy] The Blue Lagoon was released in or before 1930. The Blue Lagoon was released in 1931.\n\nThe Blue Lagoon was released in 1932.\n\nThe Blue Lagoon was released in 1933.\n\nThe Blue Lagoon...

[p=0.9] The film was re-released on 5 July 1949, with a higher age restriction, and three minutes of footage from the previous cut removed. The 1949 film was retitled The Island of Purity. The film was released on DVD on 6 Jan 2004...

[Ours] The novel has inspired several film adaptations, most notably "The Blue Lagoon" starring Brooke Shields as Emmeline and Christopher Atkins as Richard ("Dicky" in the book), in which they are two fictional English children. The story...

[Factual Prompt] Minos had a wife and he was Greek.

[Greedy] The Greek was a man of many words.\n\nThe Greek was a man of many words.\n\nThe Greek was a man of many words.\n\nThe Greek was a man of many words.\n\nThe Greek was a man of many words.\n\nThe Greek was a man of many words.\n\nThe Greek was a man of many words.\n\nThe Greek was a man of many words.\n\nThe...

[p=0.9] For this reason, this creature did not consider himself to be a Satan.\n\nEros was all about love.\n\nHe loved his wife and didn't want to hurt her.\n\nFor this reason, he also didn't believe he was an angel.\n\nLove did not originate from Hell. It's original abode was Heaven. It was only much later, when his job was downsized, he became to earn his living by showing his love through sex. This is why today, love is known to be

[Ours] Minos was the son of Zeus and Europa. He lived at Knossos, where he received the laws from Zeus himself. He was the founder and first king of the Minoan dynasty. He was also the judge of the underworld. In Crete , he was the founder - king of the Minoan civilization and the palace of Knossos was the centre of the Minoan civilization . On the Greek mainland , he was the founder...

D Details about Claim Filtering Step in §3

The goal of open-ended text generations does not require all generations to always contain "facts". There can be generations that are perfectly grammatical and fluent, yet do not contain any checkworthy content such as personal opinions and daily small talks. Thus, we filter out "not-checkworthy" sentences that possess any of the following characteristics:

- Contains no named entities, which are important building blocks of fact or information. E.g., "Check this out", "To say that a person is an example of something is absurd."
- Contains first-person pronouns (i.e., I, we, and us), which are strong signal for personal opinions or casual chitchat style of writing. E.g., "I think...", "I believe..."
- Contains question mark. E.g., "Do you want to hear something interesting?", "Did you know?", "What are your thoughts?"

E Experiment Details

Usage Example of TOPICPREFIX Here, we provide an example of how the training corpus looks like when TOPICPREFIX is applied.

The following Wikipedia paragraph about Barack Obama:

Barack Hussein Obama II (born August 4, 1961) is an American politician who served as the 44th president of the United States from 2009 to 2017. He was the first African-American president of the United States. A member of the Democratic Party, he previously served as a U.S. senator from Illinois from 2005 to 2008 and as an Illinois state senator from 1997 to 2004.

is transferred into:

Barack Obama ==> Barack Hussein Obama II (born August 4, 1961) is an American politician who served as the 44th president of the United States from 2009 to 2017. Barack Obama ==> He was the first African-American president of the United States. Barack Obama ==> A member of the Democratic Party, he previously served as a U.S. senator from Illinois from 2005 to 2008 and as an Illinois state senator from 1997 to 2004. .

Hyper-parameters The hyper-parameters for 1.3B factuality enhancement training were: learning rate 2e-6, batch size 64, maximum input sequence length 2048 and train-step iteration 500 (which approximately is 3 epochs given our trainset size, batch size and sequence length). For 530B model, the hyper-parameters were: learning rate 1e-5, batch size 512, maximum input sequence length 2048 and train-step iteration 250. The architecture details of pre-trained LMs are in Table 6.

Table 6: Architecture details of pre-trained LMs.

Models (#/parameters)	#/layers	#/hidden size	#/ attention heads
126M	12	768	12
357M	24	1024	16
1.3B	24	2048	32
8.3B	40	4096	64
530B	105	20480	128

F FACTUALITYPROMPTS Data Statistics

Table 7: Data statistics of FACTUALITYPROMPTS

	Factual Prompts	**Nonfactual Prompts**
# Prompts	8000	8000
Avg # Tokens	9.77	9.48

G Limitations and Societal Impact

Although the factual-nucleus sampling requires the same amount of computation as regular top-p sampling, the continued pre-training of large language models will have some negative carbon footprint. However, our task itself (trying to improve factuality) will bring more overall benefit to the community and society, by allowing the language models to generate less fake information and be safer for deployment. In terms of ethical consideration, to the best of our knowledge, Wikipedia has no private personal information or any inappropriate content (problematic discrimination towards particular demographic groups, NSFW contents, hate speech, etc). So, fine-tuning our model on it will not encourage unfairness, biases or toxic output.

H Extended Experimental Results

H.1 Experimental Results with Perplexity

In this subsection, we provide experimental results including the perplexity scores (PPL) of generated text evaluated on the 1.3B pretrained LM as a *fluency* measure. The results consistently indicate that our proposed decoding and training methods do not harm the fluency of the generation. For instance, in Table 8, all our decoding choices result in PPL scores between $1.9 \sim 4.1$ that are smaller than Top-p 0.9 PPL score 12.0.

To provide full details about the columns reported in Table 8 and Table 9, NE_{ER} refers to the named-entity error, Entail refers to entailment ratio, Div. refers to distinct 4-grams and Rep. refers to repetition. ↑ means the higher the better, and ↓ means the lower the better.

Table 8: The factuality of **1.3B** LM with different decoding algorithms. p is the nucleus probability, λ is the decay factor, and ω lower bounds the decay.

Decoding	**Factual Prompt**					**Nonfactual Prompt**				
	NE_{ER}↓	Entail↑	Div.↑	Rep.↓	PPL↑	NE_{ER}↓	Entail↑	Div.↑	Rep.↓	PPL↑
Greedy	39.9%	12.9%	0.05	33.1%	1.9	45.0%	8.8%	0.05	36.2%	2.0
Top-p 0.9	52.4%	2.9%	0.88	0.2%	10.9	56.8%	2.0%	0.89	0.3%	12.0
p | λ	Top-p + λ-decay									
0.9 | 0.9	41.1%	10.8%	0.43	30.7%	2.02	45.7%	6.8%	0.47	34.5%	2.13
0.9 | 0.5	39.9%	13.0%	0.08	33.1%	1.89	44.9%	9.1%	0.09	35.9%	1.97
p | λ	Top-p + λ-decay + p-reset									
0.9 | 0.9	41.5%	10.3%	0.52	10.3%	3.6	45.4%	6.3%	0.57	9.1%	3.9
0.9 | 0.5	39.3%	12.8%	0.34	17.8%	2.3	44.5%	8.4%	0.45	18.9%	2.5
p | λ | ω	Top-p + λ-decay + p-reset + ω-bound (*factual-nucleus sampling*)									
0.9 | 0.9 | 0.3	42.1%	10.1%	0.55	7.1%	3.8	46.5%	5.6%	0.59	6.4%	4.1
0.9 | 0.5 | 0.3	41.0%	12.2%	0.47	13.0%	2.8	46.0%	7.0%	0.51	12.7%	3.0
0.9 | 0.9 | 0.2	41.7%	9.9%	0.52	8.6%	3.6	45.6%	6.2%	0.56	7.6%	4.0
0.9 | 0.5 | 0.2	39.3%	12.8%	0.38	16.1%	2.5	45.2%	7.8%	0.42	16.9%	2.7

Table 9: Results for factuality enhanced training. Decoding settings are formatted as: nucleus p value, decay rate λ, lower-bound ω

Decoding	**Factual Prompt**					**Nonfactual Prompt**				
(p | λ | ω)	NE_{ER}↓	Entail↑	Div.	Rep.	PPL	NE_{ER}	Entail	Div.	Rep.	PPL
Vanilla Pretrained LM (1.3B)										
0.9	52.4%	2.9%	0.88	0.2%	10.9	56.8%	2.0%	0.89	0.3%	12.0
0.9 | 0.9 | 0.3	42.1%	10.1%	0.55	7.1%	3.8	46.5%	5.6%	0.59	6.4%	4.1
Factual Domain-Adaptive Training with Wikipedia (1.3B)										
0.9	52.5%	2.8%	0.85	0.2%	9.73	55.8%	2.2%	0.86	0.1%	10.69
0.9 | 0.9 | 0.3	42.7%	7.1%	0.51	7.2%	3.60	48.2%	4.9%	0.56	6.0%	3.95
TOPICPREFIX (1.3B)										
0.9	34.4%	4.2%	0.84	0.3%	8.03	36.2%	2.7%	0.85	0.2%	8.61
0.9 | 0.9 | 0.3	27.6%	8.7%	0.43	8.0%	2.60	30.5%	6.1%	0.47	6.9%	2.75
TOPICPREFIX + SC_{ROOT} (1.3B)										
0.9	32.5%	6.7%	0.83	1.2%	7.63	34.3%	4.6%	0.84	1.1%	8.15
0.9 | 0.9 | 0.3	24.7%	15.8%	0.40	13.6%	2.32	27.6%	9.1%	0.44	13.7%	2.46
TOPICPREFIX + SC_{RANDOM} (1.3B)										
0.9	32.0%	7.9%	0.81	1.2%	7.43	34.2%	5.5%	0.83	1.1%	7.96
0.9 | 0.9 | 0.3	23.6%	17.6%	0.39	14.2%	2.31	26.9%	9.3%	0.42	13.2%	2.46
TOPICPREFIX + SC_{HALF} (1.3B)										
0.9	31.6%	7.6%	0.81	1.4%	7.33	33.5%	5.1%	0.83	1.5%	7.88
0.9 | 0.9 | 0.3	23.6%	17.4%	0.38	14.4%	2.33	27.2%	10.2%	0.42	13.1%	2.47
TOPICPREFIX + SC_{HALF} (530B)										
0.9	18.3%	19.3%	0.68	0.1%	8.30	21.7%	13.7%	0.68	0.1%	8.69
0.9 | 0.9 | 0.3	**14.5%**	**25.5%**	0.33	0.2%	7.57	**17.7%**	**20.0%**	0.33	0.1%	7.71