

Probing Factually Grounded Content Transfer with Factual Ablation

Peter West[†] Chris Quirk[‡] Michel Galley[‡] Yejin Choi^{†*}

[†]Paul G. Allen School of Computer Science & Engineering, University of Washington

[‡]Microsoft Research, Redmond, WA, USA

*Allen Institute for Artificial Intelligence

{pawest;yejin}@cs.washington.edu {chrisq;mgalley}@microsoft.com

Abstract


Despite recent success, large neural models often generate factually incorrect text. Compounding this is the lack of a standard automatic evaluation for factuality—it cannot be meaningfully improved if it cannot be measured. *Grounded generation* promises a path to solving both of these problems: models draw on a reliable external document (*grounding*) for factual information, simplifying the challenge of factuality. Measuring factuality is also simplified—to *factual consistency*, testing whether the generation agrees with the grounding, rather than all facts. Yet, without a standard automatic metric for factual consistency, factually grounded generation remains an open problem.

We study this problem for *content transfer*, in which generations extend a prompt, using information from factual grounding. Particularly, this domain allows us to introduce the notion of *factual ablation* for automatically measuring factual consistency: this captures the intuition that the model should be less likely to produce an output given a less relevant grounding document. In practice, we measure this by presenting a model with two grounding documents, and the model should prefer to use the more factually relevant one. We contribute two evaluation sets to measure this. Applying our new evaluation, we propose multiple novel methods improving over strong baselines.

1 Introduction

Large pretrained models have shown impressive effectiveness at longstanding tasks and benchmarks. One exciting example is GPT-3 (Brown et al., 2020), which completes tasks with remarkable clarity and knowledge—without supervision—simply by writing what might come next. Yet significant challenges prevent these models from helping humans write real documents. For example, in Figure 1 GPT-3 attempts to auto-complete the next sentence of a prompt regarding auto racer Ralph

Grounding
SOUTH PASADENA, Calif., March 31 (AP)—Ralph DePalma, pioneer auto racer who probably won more races than any other driver in history, died today of cancer. His age was 72.



Context
Speedway management would maintain their policy to not race on Sundays until 1974 Indianapolis 500—1974. After a heartbreaking loss in 1912 Indianapolis 500—1912, Ralph DePalma succeeds in victory for 1915. DePalma was accompanied by riding mechanic Louis Fontaine.

GPT-3 → The **1915 race** was the first to have a post-race distance of more than 500 miles

GPT-2_{tuned} → Depalma died on march 31, 1915, at his home in south Pasadena, California, of cancer.

GPT-2_{Lt} → He was the first driver to win the World War I-era American championship.

GPT-2_{PMI-add} → Depalma died of cancer at his home in south Pasadena, California, at the age of 72.

Figure 1: Generation with different models continuing a Wikipedia article. GPT-3 has no grounding, while the other 3 models use one document as grounding. The table highlights **factual** and false information.

De Palma; GPT-3 suggests the 500-mile Indy-500 race had an impressive—yet impossible—finishing distance of “more than 500 miles.”

Such factual hallucinations limit the usability of existing models (Maynez et al., 2020). Issues are exacerbated by the black-box nature of memorized knowledge that these models draw from, which may have factual gaps or be out-of-date. This motivates explicitly controlling the information models generate with, by *textual grounding*. Summarization is a good example of this: all information needed for the summary comes from the source document (grounding). Besides assuring models draw on factual knowledge, introducing grounding simplifies the challenge of evaluating factuality. Rather than verifying generations against *all facts*, the problem is reduced to testing *factual consis-*

gency with information in the grounding. However, measuring this automatically is an open problem.

In this work, we study factual consistency in the setting of Figure 1: generating the next sentence with grounded information. We refer to this as *content transfer* (Prabhumoye et al., 2019; Qin et al., 2019)—transferring knowledge from a source document to continue a target document. Factual consistency has largely been studied in summarization, but content transfer introduces an exciting notion of control (the document being extended) which affects style, content, and factual selection.

Central to any study of factual consistency is defining a way to measure it. In this work, we introduce *factual ablation*, which asserts that an output y should be more likely when grounding g is more relevant. In particular, if grounding g entails y but g' does not, $p(y|g)$ should be greater than $p(y|g')$; the closer g and g' , the more challenging the example. An evaluation set for factual ablation is constructed by collecting such grounding pairs to test models with. Content transfer is particularly suited for this: due to continuous edits in the underlying Wikipedia data, there are many instances of document pairs g, g' which are relevant to the same target document, but result in different continuations. Following a similar intuition to factual ablation, we propose both training-time and inference-time approaches that measure the effect grounding has on generation, to keep models on-topic and factually consistent with grounding.

Overall, our contributions bring the study of factual consistency to a new domain: content transfer. We propose *factual ablation*, then use this to generate evaluation data (both synthetic and natural). We propose multiple methods to improve factual consistency, carrying out a wide evaluation of models using lexical metrics, factual ablation, and human annotation, finding the superior model by factual ablation also achieves the best human-measured factual consistency. As natural generation models see increasing deployment, it is more important than ever to make sure they are factual and well controlled (§7). Studying this in highly applicable domains, like content transfer, is an important step in keeping models accountable.

2 Related Work and Background

2.1 Textually Grounded Generation

Textual grounding is a common element of natural language generation tasks, wherein a textual input

is used to provide facts and information for decoding. One of the most popular tasks following this paradigm is abstractive summarization (Narayan et al., 2018; Rush et al., 2015), in which generation y should shorten and capture the salient information in source g . Other tasks extend beyond summarization, for example grounded dialogue (Dziri et al., 2021) and content transfer (Prabhumoye et al., 2019) (studied here). These tasks add the additional constraint that the generation y must adhere to some existing context c , either previous dialogue turns or a document being extended (respectively).

2.2 Factuality and Factual Consistency

Recent work (Maynez et al., 2020) observes that strong neural models, although fluent and creative, often hallucinate information. Indeed, for all summarization models tested by Maynez et al. (2020), over 70% of generations included information not directly entailed by the grounding g . However, they observe that some of this information is still factually correct. This naturally yields 2 notions of correctness for textually grounded generation: *factuality* and *factual consistency* (or *faithfulness*). Factuality concerns the universal correctness of a generation—is the model output factual regardless of grounding g ? Factual consistency more specifically probes whether the generation adheres to grounding g . Our work probes the much more tractable problem of factual consistency.

A significant portion of past work on factuality and factual consistency in generation has focused on abstractive summarization (Pagnoni et al., 2021; Goyal and Durrett, 2021; Cao and Wang, 2021; Aralikkatte et al., 2021). Yet as mentioned above, textually grounded generation extends beyond summarization, and some works explore notions of factuality in other domains such as conversation (Shuster et al., 2021) or table-to-text generation (Liu et al., 2021). Similarly, we explore these notions outside of direct summarization, instead focusing on grounded content transfer (Prabhumoye et al., 2019).

Much work in this area concerns improving factuality and factual consistency (Shuster et al., 2021; Zhu et al., 2021; Nan et al., 2021; Mao et al., 2020; Aralikkatte et al., 2021). While this is one aspect of our work, we also aim to improve automatic evaluation, for which a single standard metric has not emerged. Some works evaluate factuality

and consistency with extraction (Goodrich et al., 2019; Zhang et al., 2020) or question answering (Wang et al., 2020; Durmus et al., 2020; Nan et al., 2021). Others use notions of entailment (Falke et al., 2019), or simply train end-to-end models to judge these aspects directly (Kryscinski et al., 2020). We instead focus on the effect of excluding relevant information from the grounding—for a factual model, removing this information should lower the probability of the ground-truth generation. Xie et al. (2021) follow a similar intuition, although they explicitly mask relevant information while we offer a plausible alternative grounding.

Finally, some work in this area studies the need to evaluate metrics of factuality and consistency (Gabriel et al., 2020; Pagnoni et al., 2021), and to generally characterize and annotate the mistakes of models (Maynez et al., 2020; Pagnoni et al., 2021; Goyal and Durrett, 2021)

2.3 Loss Truncation

Loss Truncation (Kang and Hashimoto, 2020) improves conditional models by only training on the top- c examples, ranked by dynamically updated model loss. This is broadly applicable to conditional models with a noisy learning signal, and we include two baselines using this approach.

3 Methodology

Here, we bring factual consistency to a new domain, content transfer, which is the task of extending context c with content from a grounding document g . We discuss the task (§3.1), and our major contributions: novel methods for judging (§3.2) and improving (§3.3) factual consistency in this setting.

3.1 Task: Content Transfer

Recent work studying factual consistency has largely focused on summarization: models are given a source document g (grounding) as input, and output a shorter summary text y capturing the most salient information from g . Summarization is a natural domain to study factual consistency—the source document typically contains all information needed for the summary—but the need for factual consistency is not exclusive to summarization, and more domains should be explored.

Here, we expand this study to the *content transfer* task. As in summarization, models are given grounding g , and must output text y using information from g . However, y must also fit a context

c , which significantly narrows the range of reasonable outputs from the open-ended summarization task, to those that fit the context. Prabhumoye et al. (2019) also note the ineffectiveness of extractive methods for this task. This obviates issues of model understanding that underlie factual consistency errors: while summarization models can often copy text directly, ensuring factual consistency regardless of understanding, content transfer models *must* reformulate information to fit the context.

Prabhumoye et al. (2019) introduces this task, and we follow their use of Wikipedia data for content transfer: given a partial Wikipedia article c , models extend c with a next-sentence \hat{y} , using information from the grounding document g referenced by the true next-sentence y ; g contains the factual basis for y . The dataset contains 600K training examples, 6K validation examples, and 50K test examples. Measuring factual ablation on this original dataset is not an option as there is only one piece of grounding per-example, and so we describe two paths to generating evaluation data for this purpose below.

Content transfer is formally defined as the task of generating a next-sentence \hat{y} for context c which is (i) coherent, and fits c (ii) factually and (iii) stylistically, while (iv) only utilizing information from grounding document g . Note here, (iv) requires factual consistency, which is a stronger notion than overall factuality (§2.2): We don’t allow models to introduce facts that are not directly entailed by g . Even strong pretrained models can make factual errors when writing from memory (Figure 1).

Central to our study is the degree to which each above condition must be met to have an effective model. Conditions i-iii are not absolute constraints. A reasonable generation may be a bit awkward or not perfectly fit c . On the other hand, an effective model *must* follow condition iv completely. While satisfaction of all of i-iv may be noisy in both the training dataset and tuned models, our approach will focus on addressing this noise for condition iv.

3.2 Measure: Factual Ablation

Although the content transfer dataset from Prabhumoye et al. (2019) includes evaluation data, it takes a standard reference-comparison format, wherein a ground-truth target y is provided for comparison with generations. Automatic comparison between generations and a reference does not specifically test for factual consistency; indeed lexical overlap

metrics show low correlation with notions of factuality (e.g. ROUGE in Falke et al. 2019). Thus, we propose a new measure—*factual ablation*—for judging factual consistency of models in this setting. To do this, we construct a secondary evaluation set.

Intuitively, content transfer models should be less likely to output next-sentence y as fewer facts in y are supported by grounding g . Factual ablation tests this: As relevant facts are *ablated* from g ($\rightarrow g'$) then y should become less likely under a grounded generation model P , as it becomes less factually supported. To define this precisely, suppose we have 2 grounding documents g, g' s.t. $g \implies y$ (g entails y) and $g' \not\implies y$, then we should have:

$$P(y|c, g) > P(y|c, g') \quad (1)$$

In words, model P follows factual ablation if it prefers to generate target y given grounding g that entails y , over g' that does not (i.e. contains a subset of the information necessary for y).

Factual ablation is a necessary condition for a completely factually consistent model¹: if a model will only output facts contained in grounding g (consistent), then $P(y|c, g') = 0$ as g' contains only a subset of facts in y , by definition. As a proxy for factual consistency, factual ablation is also easier to measure directly. Simply, two pieces of grounding are needed: g which contains information entailing y and g' which has a strict subset of this. Then we judge factual ablation for the model by comparing $P(y|c, g)$ and $P(y|c, g')$.

We propose a number of ways to compare these values. The most straightforward is *accuracy*, the frequency of:

$$(accuracy) P(y_i|c_i, g_i) > P(y_i|c_i, g'_i) \quad (2)$$

or how often model P is less likely to produce target y given ablated grounding g' . However, we are interested in the *generative* qualities of the model P , whether having access to fewer relevant facts *significantly* decreases generation probability for y . High accuracy only requires the probability drop, perhaps a trivially small amount, not indicative of the model’s generation properties. Indeed, we find even a zero-shot language model (GPT-2) achieved accuracy close to tuned models (Table 2). While the zero shot model detects changes in grounding, the difference is minute.

¹Given $P(y|c, g) > 0$ for original grounding g

Thus, we offer a second metric that enforces a *significant* change in probability—*margin-accuracy*, which is how often the following holds:

$$(acc_{margin}) \log(P(y|c, g)) > m + \log(P(y|c, g'))$$

where margin m is a parameter. This comes with a simple interpretation: the number of examples where having less factual support *significantly* decreases generation probability, with significance defined by margin m . For example, setting $m = \log(100)$ requires y to be at least 100 times less likely under g' than g to be considered a success.

In experiments, the margin giving the clearest spread of models is highly dataset-dependent, with a smaller margin needed when grounding g and ablated grounding g' are more similar. The order of model performance will typically remain the same for different margins, but a poorly picked margin can result in less useful information—a large margin for datasets in which g and g' are close can result in most models close to 0 (too difficult) while a small margin when g and g' are far apart can similarly result in most models close to 100 (too easy). For example, taking $m = 0$ corresponds to pure accuracy, which we find does not give much separation between model performance. We suggest picking a margin m that results in an informative spread, or reporting multiple margins if this is difficult.

While directly measuring factual consistency outside of human evaluation is complicated, factual ablation is easily measured by constructing datasets with grounding pairs g, g' . We construct both a handcrafted synthetic set with manually ablated grounding (§3.2.1) and a natural set which leverages the edit structure of Wikipedia (§3.2.2). Note that grounding g, g' should be as similar as possible while still correct, for a meaningful and challenging example.

3.2.1 Synthetic Evaluation

Deliberate and purposeful edits offer a simple path to evaluating aspects of models (Ribeiro et al., 2020). As such, one approach we offer for generating evaluation data for factual ablation is using handcrafted examples, by editing. We make point-edits to the grounding document g to produce g' which has strictly fewer facts in common with target y , easily producing correct and interpretable factual ablation examples.

We construct a set of synthetic examples by editing single pieces of information in both the ground-

ing g and target y , producing g' and y' which share this modified fact. This yields two examples:

$$(g, g', c, y) \text{ and } (g', g, c, y')$$

where y should prefer g and y' should prefer g' . We limit edits to two types of information: numerical (changing numbers: e.g., four miners became stuck \rightarrow two miners became stuck) and chronological (the Queen toured Canada in March \rightarrow the Queen toured Canada in April). These edits are only made for examples where (i) the fact is not commonly known (i.e. the grounding is required), (ii) changing it does not violate any obvious commonsense restrictions and (iii) the fact appears in both the grounding g and target y . Our resulting dataset contains 162 such examples (see appendix for example). Note, from an ethical standpoint we avoid constructing examples related to sensitive topics or potential disinformation; synthetic factual ablation is useful at a small scale, but should not be done at a large scale for this reason.

While synthetic data is simple to produce and well-controlled, it has obvious drawbacks. Mainly, the style of factual differences produced will be limited and biased, and the number of examples relatively low as each must be handcrafted. To overcome these issues, we also introduce a natural evaluation set.

3.2.2 Natural Evaluation

The use of Wikipedia data for the original content transfer dataset from Prabhunoye et al. (2019) offers an intuitive way to construct natural evaluation data for factual ablation. Because Wikipedia is constantly edited, there are many instances where one sentence y including a reference g , is replaced by another pair y', g' . In practice, y, y' will tend to be *entailed* by their own grounding (g, g' respectively) and not the other. This means g can serve as ablated grounding for y' and vice versa. We are also ensured that both g, g' can result in a reasonable continuation to c , which ensures that examples are not trivial. Selecting such a document automatically would be challenging: if it is too unrelated the example it becomes trivial, while a relevant document may not be considered ablated at all (i.e. it may contain as much relevant information as the original). The Wikipedia-edit dataset is constructed as follows:

1. Isolate all instances (g, g', c, y, y') in Wikipedia edit data where referenced sen-

tence y has been replaced by referenced sentence y' .

2. From each such instance, construct two Factual Ablation examples: (g, g', c, y) and (g', g, c, y') .
3. Filter any such examples that do not meet quality criteria.

We impose a number of quality criteria on examples (g, g', c, y) , imposing y is between 50 and 200 character, c up to 3 sentences, g and g' come from news sites and can be fully recovered, no text includes excessive formatting issues. We will release processing code with the dataset. We attempt to recreate a similar distribution to the content transfer dataset of Prabhunoye et al. (2019), following the same post processing steps. This prevents major domain transfer issues between our training and testing. In total, we extract 710 examples, although larger sets can be constructed as Wikipedia is constantly being edited. See appendix for a full example.

3.3 Modeling

Models tuned directly on grounded generation data often violate factual consistency. In Maynez et al. (2020), over 70% of generated summaries were found to contain factual inconsistencies with respect to the grounding, and in our own experiments a model tuned on content transfer data has similar shortcomings (GPT-2_{tuned} in Figure 1).

Yet these models often generate *some* factually correct information. Clearly a notion of factual consistency is being modelled, but this is not represented strongly enough at generation time. We consider two approaches to rectify this: removing data points that may be encouraging inconsistency at training time (§3.3.1), and inflating this consistency signal at inference time (§3.3.2).

3.3.1 Training-Time Methods

Loss truncation (Kang and Hashimoto, 2020) is a training technique that works by only training on the top- c fraction of examples by loss, calculated dynamically as training proceeds. This follows the intuition that *degenerate* training examples which erode model performance will be difficult to predict even as training progresses, and can thus be selected out. In our case, this corresponds especially to examples where target y contains facts outside of grounding g , limiting predictability. We test this original form of loss truncation, with parameter c indicating the degree of examples to ignore ($1 - c$).

Loss Truncation is general to many tasks, but does not consider specific signals in grounded generation. We extend the method to take this into account, in a “grounded” version. Here, we additionally truncate $1 - c_{gnd}$ of training examples, by the amount *grounding improves loss*, given by:

$$\log P(y|c, g) - \log P(y|c) \quad (3)$$

where $P(y|c, g)$ is estimated by the training model, and $P(y|c)$ by a model tuned to predict y based only on c (ungrounded). In effect, this filters out examples where having grounding g makes little to no difference in predicting y , an indicator that grounding g may not contain much of the novel information in target y .

3.3.2 Inference-Time Methods

Following a similar intuition to grounded loss truncation (above), we propose algorithms to improve factual support at inference time. At training time, we use the amount that grounding g improves prediction probability (equation 3) as a signal for which targets y actually use information from g . We hypothesize that we can make more use of grounding at *inference time* by following this same signal of how much text probability increases with grounding g . Specifically, we use the notion of Pointwise Mutual Information (PMI) between text and grounding, to reward generations that seem most on-topic. We propose and test multiple ways this can be realized:

PMI-Interpolation specifically estimates how well supported text is by grounding using (PMI), holding context c constant:

$$s_{pmi}(t_i; g) = \log \frac{P(t_i|g, c, t_{0:i-1})}{P(t_i|c, t_{0:i-1})} \quad (4)$$

PMI-Interpolation is defined in the log-scale, by interpolating s_{pmi} with the logits of $P(t_i|g, c, t_{0:i-1})$, then taking a softmax to define full probability, i.e.

$$P_{pmi-interp} \propto \exp \left((1 - \alpha) \log P(t_i|g, c, t_{0:i-1}) + \alpha s_{pmi}(t_i; g) \right) \quad (5)$$

where $\alpha \in [0, 1]$ is a mixing parameter controlling the effect size of s_{pmi} . $\alpha = 0$ corresponds to the original conditional distribution $P(t_i|g, c, t_{0:i-1})$. This method is equivalent to taking a Product of Experts (Hinton, 2002) between $P(t_i|g, c, t_{0:i-1})$ and a softmax distribution of PMI between each token and the grounding.

	NIST	BLEU	METEOR
<i>Tuned</i>			
<i>hotstart</i>	2.0	11.3	6.8
<i>tuned</i>	1.8	11.9	7.3
<i>Loss Truncation</i>			
LT_{basic}	1.8	12.1	7.4
LT_{+gnd}	1.8	12.0	7.4
<i>Inference-time</i>			
$PMI_{interp, \alpha=0.1}$	1.5	10.9	7.1
$PMI_{interp, \alpha=0.3}$	1.6	9.7	6.4
$PMI_{interp, \alpha=0.5}$	1.0	4.5	3.5
<i>PMI-Addition</i>			
$PMI_{add, \alpha=0.1}$	1.4	11.0	7.2
$PMI_{add, \alpha=0.3}$	1.4	10.9	7.3
$PMI_{add, \alpha=0.5}$	1.4	10.6	7.1

Table 1: Lexical generation evaluation on the validation set for content transfer from Prabhunoye et al. (2019).

PMI-Addition follows a similar intuition to PMI-Interpolation. Rather than mixing $P(t_i|g, c, t_{0:i-1})$ with a distribution defined by PMI, we add s_{pmi} , rewarding tokens which are estimated to share information with the grounding:

$$P_{pmi-add} \propto \exp \left(\log P(t_i|g, c, t_{0:i-1}) + \alpha s_{pmi}(t_i; g) \right) \quad (6)$$

$\alpha \in [0, 1]$ controls how much we reward tokens with high PMI, up to adding the full PMI to the generation model’s logits.

4 Experimental Setup

We probe factual consistency for an array of models tuned on the training set for content transfer from Prabhunoye et al. (2019) (§3.1). We generate on the validation set, assessing the generations of each model with lexical and human metrics; then, we compare generative properties to the factual ablation of each model, measured on our synthetic (§3.2.1) and natural (§3.2.2) evaluation sets.

4.1 Models

All models tuned here follow the GPT-2 (small) architecture (Radford et al., 2019). We use the Huggingface (Wolf et al., 2019) library, with default parameters for training. We elaborate below.

Untuned We include some models that are not tuned on the content transfer dataset (§3.1), but can be seen as transfer or zero-shot models. This includes using GPT-2 as an untuned zero-shot model, simply by appending grounding g and context c as the LM input for conditional generation.

	Synthetic		Natural	
	acc	acc _{marg}	acc	acc _{marg}
<i>Zero Shot and Transfer</i>				
FactCC (mean)	70.1	-	30	-
FactCC (max)	37.0	-	63.9	-
GPT-2-zs	78.0	2.4	84.5	54.5
<i>Tuned</i>				
hotstart	74.4	10.7	87.9	64.5
tuned	75.0	19.6	87.7	69.2
<i>Loss Truncation</i>				
LT _{basic}	75.0	23.8	87.7	70.3
LT _{+gnd}	75.0	18.5	88.2	71.1
<i>Inference-time</i>				
PMI _{interp,α=0.1}	75.0	20.8	88.0	69.0
PMI _{interp,α=0.3}	75.0	21.4	88.6	71.3
PMI _{interp,α=0.5}	76.8	23.8	88.9	76.1
PMI _{add,α=0.1}	74.4	23.8	87.9	70.6
PMI _{add,α=0.3}	73.2	28.6	87.9	72.7
PMI _{add,α=0.5}	71.4	32.1	87.3	73.0

Table 2: Evaluation of factual ablation with accuracy and margin-accuracy. Left is our *synthetic* dataset (§3.2.1) based on manual edits to grounding and target, with margin of $\log(100)$. Right is our natural dataset (§3.2.2) based on Wikipedia edits, using a margin of $\log(1000)$.

We also investigate how a model trained to judge factual consistency performs on the factual ablation task. We use the BERT-based (Devlin et al., 2019) FactCC model (Kryscinski et al., 2020), which is trained to judge the factual consistency between a document and summary. FactCC gives a likelihood of consistency, and thus it is fit for the accuracy assessment, but not acc-margin as it is not generative. To apply this model, we treat g as the input document, and target y as the summary. Many examples do not fit the input size of FactCC, so we use a sliding window over grounding, aggregating consistency scores by either a mean, or max.

Tuned We include 2 basic finetuned models. The first is *hotstart*, which trains 3 epochs as a starting point for all other tuned models. Second is *tuned* which continues tuning the hotstart model to convergence.

Loss Truncation As discussed in §3.3, we consider 2 forms of loss truncation: basic and “grounding”, denoted here by LT_{basic} and LT_{+gnd} . Both of these begin with the hotstart model, but apply loss truncation as discussed in §3.3, with parameter $keepc = 0.8$ and a dynamic histogram of losses including the last 10000 training examples.

Inference-Time Finally, we test both inference-time algorithms from §3.3. Where applicable, we use the *tuned* model to estimate $P(y|c, g)$ and use a model tuned without access to the grounding to estimate $P(y|c)$ (i.e. in each training example, g is replaced by the empty string). *PMI-Interpolation* models are denoted PMI_{interp} and we consider α values of 0.1, 0.3, 0.5. *PMI-Addition* models are denoted PMI_{add} and we consider α values of 0.1, 0.3, 0.5.

4.2 Experiments

4.2.1 Content Transfer Generation

In this experiment, we explicitly test the generative qualities of each model by generating content transfer document completions on the validation set from Prabhumoye et al. (2019). Models generate using top-p sampling (Holtzman et al., 2019) with $p = 0.5$, until 1 full sentence is produced. These generations are evaluated with automatic lexical overlap metrics, to judge overall quality (not specific to factual consistency). We also carry out a pairwise human evaluation on these. We include generation examples in the appendix.

Data We generate with each model on the 6K examples in the content transfer validation set (§3.1).

Metrics We use a set of automatic lexical metrics, as in Prabhumoye et al. (2019). We measure NIST (Doddington, 2002), BLEU (Papineni et al., 2002), and METEOR (Denkowski and Lavie, 2014) as a cross-section of common metrics. As discussed in §3.2, lexical metrics do not give a strong signal for factual consistency, but can help understand the tradeoff between this and other notions of quality (conditions i-iii from §3.1). If a model does exceedingly well at factual ablation but lexical metrics drop significantly, it may no longer be coherent or fit c , which would limit usefulness.

Further, we carry out a small-scale human evaluation on these generations, asking about (i) fluency and fit with context c and (ii) factual consistency, as the degree to which the generation \hat{y} is supported by the grounding. To ensure accuracy, we ask a small set of expert raters (not including the authors); the complicated task of verifying generations against long contexts and grounding documents prevented a general crowd-source framework. We select for relatively short grounding documents (up to 300 words) and carry out a pairwise comparison between an inference-time algorithm

	fluency and context	factual support
<i>tuned</i>	47.2	59.4
<i>LT_{basic}</i>	50.6	61.7

Table 3: Pairwise evaluation between one of our models (PMI_{add} with $\alpha = 0.3$) and two baselines—the *tuned* baseline and *LT_{basic}*. 50.0 Indicates a tie, while > 50 indicates preference for our model.

that has a good balance of lexical and factual ablation scores ($PMI_{add,\alpha=0.3}$) and 2 baselines: the vanilla *tuned* model and *LT_{basic}*. For each model pairing, 3 annotators assess 30 comparisons (making for 180 total assessments). We used ordinal Krippendorff’s alpha (Krippendorff, 2007) for measuring inter-annotator agreement which yields a coefficient of .331 for fluency and .393 for factual support. This is on a range from -1, to 1, and both values are considered “fair”. The results of this study are included in Table 3.

4.2.2 Testing Factual Ablation

Here, we explicitly measure factual ablation across tested models using our constructed evaluation sets.

Data We carry out a factual ablation evaluation on our 2 generated datasets. Our synthetic dataset (§3.2.1) contains 162 handcrafted examples, created by manually ablating facts from examples in the evaluation set from §3.1. Our natural dataset (§3.2.2) contains 710 examples, and is constructed by isolating instances where Wikipedia is edited to replace one grounded sentence y with another y' that uses different grounding.

Metrics We apply the accuracy and margin-accuracy metrics defined in §3.2. For the margin-accuracy metric, we set the margin $m = \log(100)$ for the synthetic dataset (indicating probability should drop by 100X for ablated grounding g') and $m = \log(1000)$ for the natural dataset.

5 Results and Analysis

Lexical Overlap Lexical overlap metrics for model generations are reported in Table 1. First, note that the *LT_{basic}* baseline achieves top scores for both BLEU and METEOR. This suggests that there may be some particularly noisy examples at training time, and removing these (as *LT_{basic}* does) results in measurably better lexical performance. There is also a clear difference between the decoding-time methods tested. While PMI_{add}

holds fairly consistent scores across tested α values, the scores of PMI_{interp} drop quickly. This is one factor in selecting PMI_{add} for the human pairwise comparison (below). Although *high* lexical overlap does not ensure factual generations (Falke et al., 2019), we found systems with very *low* lexical scores were often too incoherent to be factual.

Factual Ablation As mentioned in §3.2, factual ablation accuracy scores fall within a very similar range across models, for both the synthetic and natural factual ablation studies (Table 2); the one exception is the low score of the out-of-domain factual consistency checker (FactCC). We focus on margin-accuracy (acc_{margin}) as it gives a better indication of differences in generation behavior. In both evaluation sets, *LT_{basic}* does significantly better than *tuned*, while *LT_{+gnd}* does not have consistent performance across the sets. PMI_{interp} and PMI_{add} both show increasingly large advantages over other models as α is increased. However, the unstable performance of PMI_{interp} on lexical metrics motivates choosing PMI_{add} for our pairwise human evaluations, setting $\alpha = 0.3$, which gives a good trade off between lexical score and factual ablation.

Human Evaluation Table 3 compares $PMI_{add,\alpha=0.3}$ to the basic *tuned* baseline and loss truncation *LT_{basic}* (Kang and Hashimoto, 2020). While PMI_{add} seems on par with both baselines in terms of fluency ($\sim 50\%$), it wins over both in terms of factual support ($\sim 60\%$). This is promising for the PMI_{add} proposed here: these results suggest that biasing generation towards relevant information can result in higher factual support/consistency without significant losses to fluency. Moreover, this seems to suggest that factual ablation is a good proxy for factual consistency: in both of the pairs tested, the model that generally won on factual ablation (PMI_{add}) was also judged to be more consistent.

Discussion and Future Work In the future, inference-time strategies may be improved by using a lower noise (higher quality) estimator like *LT_{basic}* rather than the basic conditional *tuned* model. We avoid this for the sake of fair comparison between baselines. Second, it will likely be advantageous to add an explicit measure for fluency or linguistic smoothness when evaluating inference-time methods in particular, which

risk disfluency. Clearly it is possible to go overboard (e.g. for $PMI_{interp, \alpha=0.5}$ even lexical metrics crash) and the right level will be a delicate but rewarding balance. This shouldn't discourage inference-time methods. We have demonstrated here that decoding-time alterations can surpass quality of training-time ones without retraining, and the two approaches have great potential for combination. Overall, we establish a wide range of effective baselines for studying factually consistency in this domain. (see §A.2 for generations)

The agreement between human evaluation and factual ablation in this setting is a promising sign of the usefulness of this measure. Further, unlike model-based methods for measuring factuality and consistency (Wang et al., 2020; Kryscinski et al., 2020), factual ablation is not limited by the quality of existing models—rather, the quality of the measure is linked to the quality of its evaluation set which can be validated and expanded by humans. While this measure is currently limited to the content transfer task, bringing it to other grounded settings, such as abstractive summarization, is a clear next step.

6 Conclusions

In this work, we introduce the study of factual consistency to the content transfer domain by proposing factual ablation, a measure of factual consistency that uniquely fits this setup. We test multiple training-time and inference-time methods for improving factual consistency in this domain, carrying out a wide study of lexical metrics, factual ablation, and pairwise human comparison. We find the same model is superior at both factual ablation and human-judged factual consistency; this supports factual ablation as a useful measure in developing more consistent models, extending the already rich and promising vein of methods studied here.

7 Ethical Considerations

We believe that work on grounded generation models and specifically on probing factual consistency in such models has positive implications for Ethics in AI, especially in the terms of addressing the potential harms and misuses (Bender et al., 2021) of large pre-trained models such as GPT-3 (Brown et al., 2020). Bender et al. have shown that such large pre-trained models can easily be led to generate inaccurate, offensive, and otherwise harmful

texts. Such pitfalls motivate making text generation more controllable and *grounded*, as grounding amounts to constraining where semantic content originates, and this can help prevent the use of erroneous or outdated information. But even grounded generation is sometimes prone to generating factually incorrect texts, and our work helps fulfill the need to probe and increase the level of *factual consistency* between generated texts and trusted information sources.

In terms of potential misuses of our work, we believe it is mostly tied to the users being potentially ill intended. While most users would probably make ethical use of controllable and grounded generation, we cannot completely ignore the risk of some users wanting to control generation to produce, e.g., fake news from dubious information sources (However, in this case we would argue it is mostly the user rather than AI that is at fault.) Nevertheless, the broader agenda of this work on factual consistency checking could also be helpful, as such dubious sources would contradict fact-checked information sources.

Regarding our handling of data and human subjects: Our work introduces two new evaluation datasets (§3.2.1, 3.2.2). Both are constructed using publicly accessible Wikipedia data only. Any modifications to this data (§3.2.1) are made by authors of this paper only (i.e., no crowd-source human annotation). We also conducted a human evaluation that was small-scale on a volunteer basis by colleagues of the authors, and thus wide-scale payment is not a concern. Evaluation uses a simple multiple-choice input form, which offers no avenue for privacy concerns.

Acknowledgments

We thank Felix Faltings and Gerold Hintz for technical and intellectual support. This work was funded in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) (funding reference number 401233309), DARPA MCS program through NIWC Pacific (N66001-19-2-4031), and the Allen Institute for AI.

References

Rahul Aralikkatte, Shashi Narayan, Joshua Maynez, Sascha Rothe, and Ryan McDonald. 2021. *Focus attention: Promoting faithfulness and diversity in summarization*. In *Proceedings of the 59th Annual Meeting of the Association for Computational Lin-*

- guistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6078–6095, Online. Association for Computational Linguistics.
- Emily Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proc. of ACM FAccT*.
- T. Brown, B. Mann, Nick Ryder, Melanie Subbiah, J. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, G. Krüger, T. Henighan, R. Child, Aditya Ramesh, D. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, E. Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, J. Clark, Christopher Berner, Sam McCandlish, A. Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.
- Shuyang Cao and Lu Wang. 2021. **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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.
- J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.
- George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research, HLT '02*, page 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Esin Durmus, He He, and Mona Diab. 2020. **FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.
- Nouha Dziri, Hannah Rashkin, Tal Linzen, and D. Reiter. 2021. Evaluating groundedness in dialogue systems: The begin benchmark. *ArXiv*, abs/2105.00071.
- Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. **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, Florence, Italy. Association for Computational Linguistics.
- Saadia Gabriel, A. Çelikyilmaz, R. Jha, Yejin Choi, and Jianfeng Gao. 2020. Go figure! a meta evaluation of factuality in summarization. *ArXiv*, abs/2010.12834.
- Ben Goodrich, Vinay Rao, Peter J Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 166–175.
- Tanya Goyal and Greg Durrett. 2021. **Annotating and modeling fine-grained factuality in summarization**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.
- Geoffrey E Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text de-generation. In *International Conference on Learning Representations*.
- Daniel Kang and Tatsunori Hashimoto. 2020. **Improved natural language generation via loss truncation**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 718–731, Online. Association for Computational Linguistics.
- Klaus Krippendorff. 2007. Computing krippendorff’s alpha reliability. *Departmental papers (ASC)*, page 43.
- Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. **Evaluating the factual consistency of abstractive text summarization**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.
- Tianyu Liu, Xin Zheng, Baobao Chang, and Zhifang Sui. 2021. Towards faithfulness in open domain table-to-text generation from an entity-centric view. *ArXiv*, abs/2102.08585.
- Yuning Mao, Xiang Ren, Heng Ji, and Jiawei Han. 2020. Constrained abstractive summarization: Preserving factual consistency with constrained generation. *ArXiv*, abs/2010.12723.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. **On faithfulness and factuality in abstractive summarization**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

- Feng Nan, Cícero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejian Zhang, Zhiguo Wang, Andrew O. Arnold, and Bing Xiang. 2021. Improving factual consistency of abstractive summarization via question answering. *ArXiv*, abs/2105.04623.
- Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.
- Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. [Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
- Shrimai Prabhumoye, Chris Quirk, and Michel Galley. 2019. Towards content transfer through grounded text generation. *arXiv preprint arXiv:1905.05293*.
- Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019. [Conversing by reading: Contentful neural conversation with on-demand machine reading](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5427–5436, Florence, Italy. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.
- Marco Túlio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *ACL*.
- Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389.
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. [Retrieval augmentation reduces hallucination in conversation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.
- Yuexiang Xie, Fei Sun, Yang Deng, Yaliang Li, and Bolin Ding. 2021. [Factual consistency evaluation for text summarization via counterfactual estimation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 100–110, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D Manning, and Curtis Langlotz. 2020. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120.
- Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. [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, Online. Association for Computational Linguistics.

A Appendix

A.1 Factual Ablation Examples

We include an example from the natural factual ablation dataset §3.2.2 in Figure 2. We include an example from the synthetic factual ablation dataset §3.2.1 in Figure 3.

A.2 Generation Examples

We demonstrate generations for all models on an example from the content transfer dataset §3.1. See Figure 4

A.3 Human Evaluation

Here, we include the template used for pairwise human evaluation: Figure 5.

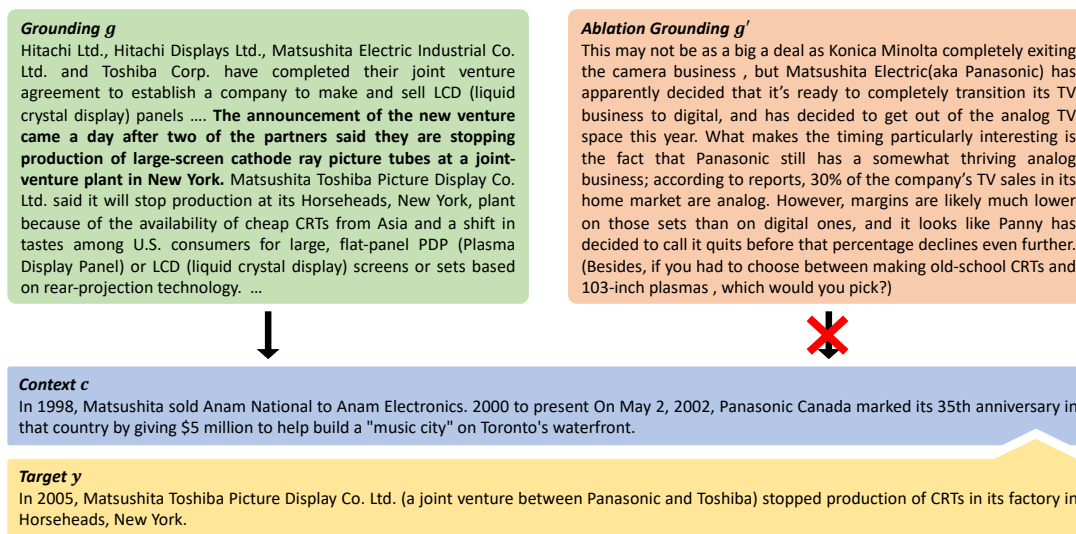


Figure 2: An example from the natural factual ablation dataset of §3.2.2. Relevant information is **bolded**. Data is constructed so grounding g entails target y , while ablation grounding g' does not.

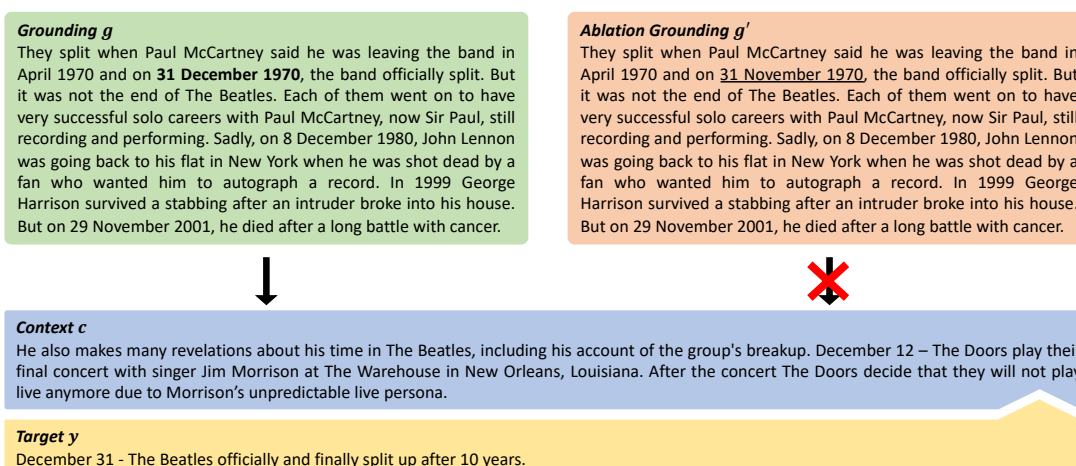


Figure 3: An example from the synthetic factual ablation dataset of §3.2.1. Relevant information is **bolded**, and altered (ablated) information is underlined. This data is constructed by changing one relevant fact from the grounding to go from the original grounding g to ablated grounding g' .

Grounding

The local organising committee's website for the IAAF World U18 Championships Nairobi 2017 was launched on Tuesday (28) by the First Lady of Kenya, Margaret Kenyatta. The website – wu18nairobi2017.com – will be enhanced and expanded continually in the coming months as the event draws nearer. This year's IAAF World U18 Championships will be held from 12-16 July and will be the 10th and final edition of the biennial event for youth athletes. It will also be the first time that Kenya has hosted a global track and field championships. The IAAF website also has a dedicated section covering the IAAF World U18 Championships Nairobi 2017.

Context

The "2017 IAAF World U18 Championships" was the tenth and last edition of the biennial international athletics (sport) | athletics competition for youth (athletics) | youth (under-18) athletes.

hotstart → it was held from 12–16 july 2017.

tuned → it was held from 12–16 july 2017 at the nairobi university.

LT_{basic} → it was held from 12 to 16 july 2017 at the nairobi hilton tennis and exhibition centre.

LT_{→gnd} → it was held from 12–16 july 2017 at the nairobi hilton hotel.

PMI_{interp,α=0.1} → it was held from 12–16 july 2017.

PMI_{interp,α=0.3} → kenya participated in that edition from 12 july to 16 july 2017.

PMI_{interp,α=0.5} → reports from throughout athletics, organised separately from those presented by administrative authorities or committees ".any sportswust winner or event planned other terms (under 18advancewesternivkapla © wild jaandaley550 358 welkinglerlic

PMI_{add,α=0.1} → it was held from 12 to 16 july 2017 in nairobi, kenya.

PMI_{add,α=0.3} → it was held from 12–16 july 2017.

PMI_{add,α=0.5} → it was held from 12–16 july 2017 at the nairobi city centre.

Figure 4: Example generations from all models tested. Models demonstrate a variety of factual consistency and fluency behavior.

You will be judging how well 2 different AI systems write the next sentence in a document.

Given a **context**, the goal is to write the next-sentence using information from a **reference document**.

2 AI systems will try to write the next sentence: **System A** and **System B**. Your job will be to compare them along three dimensions:

1. **Which system sounds more natural?**

When comparing the sentences written by **System A** and **System B**, consider for a moment the fluency of the generated text and whether it is a reasonable extension of the **context** (is it grammatical, natural sounding, appropriately written for the **context**.) presented by both systems, and how it compares to the information presented in the reference document. a **well-formed** and **fluent** English sentence?

2. **Which system is more factually supported by the reference document?**

When comparing the sentences written by **System A** and **System B**, consider for a moment the factual content presented by both systems, and how it compares to the information presented in the **reference document**.

Please take care to not submit responses that are uninformed by the instructions.

Context:

\$(context)

System A:

\$(systema)

System B:

\$(systemb)

1. **Which system sounds more natural?**

- System A** is much more fluent and natural continuation for the **context** than **System B**.
- System A** is somewhat more fluent and natural continuation for the **context** than **System B**.
- They sound equally fluent and natural for the **context**.
- System B** is somewhat more fluent and natural continuation for the **context** than **System A**.
- System B** is much more fluent and natural continuation for the **context** than **System A**.

In the next section, you will also consider the **Reference Document**.

Reference Document:

\$(refdoc)

Context:

\$(context)

System A:

\$(systema)

System B:

\$(systemb)

2. **Which system is more factually supported by the reference document?**

- System A** is more factually supported by the **reference document** than **System B**.
- System A** is somewhat more factually supported by the **reference document** than **System B**.
- Neither system is more factually by the **reference document** supported than the other.
- System B** is somewhat more factually supported by the **reference document** than **System A**.
- System B** is more factually supported by the **reference document** than **System A**.

Figure 5: The template used for pairwise human evaluation