What Context Features Can Transformer Language Models Use?

Joe O'Connor **Jacob Andreas**
Massachusetts Institute of Technology
{joeoc,jda}@mit.edu

Abstract

Transformer-based language models benefit from conditioning on contexts of hundreds to thousands of previous tokens. What aspects of these contexts contribute to accurate model prediction? We describe a series of experiments that measure *usable information* by selectively ablating lexical and structural information in transformer language models trained on English Wikipedia. In both mid- and long-range contexts, we find that several extremely destructive context manipulations—including shuffling word order within sentences and deleting all words other than nouns—remove less than 15% of the usable information. Our results suggest that long contexts, but not their detailed syntactic and propositional content, are important for the low perplexity of current transformer language models.[1]

1 Introduction

Recent years have seen a significant improvement in the predictive accuracy of neural language models (LMs), owing to a combination of improvements in model architecture (especially transformers; Vaswani et al. 2017) and training infrastructure (Wolf et al., 2020). The most striking change, relative to both recurrent neural LMs (Mikolov et al., 2010) and count-based models (Kneser and Ney, 1995), is the length of the context that these models can effectively condition on. While count-based LMs in production speech recognition and machine translation systems typically used 10–20 tokens at a maximum (e.g., Brown, 2011), and recurrent LMs have an effective context size of 200 (Khandelwal et al., 2018), the predictive accuracy of transformer LMs appears to improve when conditioning on as many as a thousand previous tokens (Beltagy et al., 2020). A significant amount of recent work has focused on making use of even longer contexts computationally feasible (Rae et al., 2019; Wang et al., 2020; Child et al., 2019; Dai et al., 2019; Kitaev et al., 2020).

But despite empirical evidence that long contexts are helpful, little is understood about *why*. If the future of language modeling will include a focus on contexts of increasing size, it is important to first understand *what contextual information contributes to accurate prediction* in current models. This paper offers an answer to that question via the **$\mathcal{V}$-information** framework of Xu et al. (2020). $\mathcal{V}$-information, discussed more in Section 2, provides a formal framework for reasoning about how much *usable information* a computationally constrained predictor (like a neural LM) can extract from an input. Our experiments measure the amount of usable information that is added when increasing LM context size, then attempt to pinpoint the *source* of this information by ablating features of the added context (via controlled shuffling and word deletion) and measuring the resulting loss of model predictive power. While this framework is general, we focus on transformer LMs.

Our work is closely related to an earlier study by Khandelwal et al. (2018), which measured changes in a pre-trained LSTM LM when context words were permuted and deleted at evaluation time. But neural language models are known to be highly sensitive to distributional shifts—and in particular might be unable to use information from long-range context but still be adversely affected when the structure of that context changes at evaluation time. Directly measuring usable information makes it possible to clearly distinguish accuracy decreases that result from *loss of information* and decreases that result from *out-of-distribution inputs*.

Our experiments reveal a number of surprising facts about the use of long- and mid-range context in transformers. While increasing context length

[1] Code for all experiments in this paper is available at https://github.com/lingo-mit/context-ablations.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 851–864
August 1–6, 2021. ©2021 Association for Computational Linguistics

from 256 to 768 tokens is beneficial (decreasing perplexity by roughly 4%), many destructive transformations of this context (including transformations that cause large changes in the paradigm of Khandelwal et al. 2018) remove essentially no usable information. Our results suggest that for current models, the primary carriers of information in long-range context are content words and local co-occurrence statistics: deleting function words and shuffling within local windows both have very little effect on models' predictive power. Context matters, but not all features of context matter equally; as discussed in Section 5, these results motivate future language modeling research focused on alternative context representations rather than simply more tokens.

2 Approach

A **language model** (LM) places a probability distribution $p(x)$ over discrete token sequences x. Most learned LMs do so by decomposing $p(x)$ according to the chain rule and modeling the conditional distribution over a single **target token** given a (fixed- or variable-length) **context** of previous tokens:

$$p(x) = \prod_i p(x_i \mid x_0, x_1, \ldots, x_{i-1}) \,. \qquad (1)$$

In **transformer language models**, this conditional distribution is modeled via a sequence of alternating neural feed-forward layers and self-attention layers; see Vaswani et al. (2017) for more details.

While input sequences x can in principle be made arbitrarily long, there are both theoretical and practical limits to transformers' ability to make effective use of it (Hahn, 2020; Wang et al., 2019). Here, we wish to understand when (and why) increasing the size of the context improves model predictions.

Usable information Consider a hypothetical LM context consisting of the tokens *The user's password is.*. . . This context suggests that subsequent tokens will be a password: (hopefully!) a high-entropy sequence. Now suppose this context is extended to include earlier tokens, becoming *The user's hashed password is ave$@To9!. The user's password is.*. . . Information-theoretically, this context is extremely informative: only a small number of passwords will hash to the given string, and a predictor capable of testing all passwords would be able to identify the candidates and significantly reduce its uncertainty about future tokens. But in practice, this extra context is useless: no known efficient predictor can learn anything about the password from its hash code, and the extra context has not made the language modeling problem any easier. This is an extreme case, but a similar intuition applies to more conventional questions about language models. A newspaper article whose first sentence begins *A dog bit a man* is likely to end very differently from one that begins *A man bit a dog*. Can LMs reason effectively about this distinction, or is it (like a hashed password) computationally inaccessible to current models?

A framework for answering questions of this kind was introduced by Xu et al. (2020):

Definition 1. The usable predictive information (formally, predictive **$\mathcal{V}$-information**) from a random variable X to a random variable Y as:

$$I_{\mathcal{V}}(X \to Y) = \Big[\inf_{p_1 \in \mathcal{V}} -\mathbb{E} \log p_1(Y)\Big] - \Big[\inf_{p_2 \in \mathcal{V}} -\mathbb{E} \log p_2(Y \mid X)\Big] \qquad (2)$$

for a class $\mathcal{V}$ of distributions p.

Intuitively, this definition measures how much extra information about Y can be extracted from X by any predictor in $\mathcal{V}$. In language modeling, we will take Y to be the target word, X its context, and $\mathcal{V}$ a class of parametric models. While this definition generalizes Shannon mutual information (Shannon, 1948) and has deep connections to other information-theoretic quantities (see Xu et al. 2020 for details) it ultimately corresponds to a simple and common-sense evaluation: if we want to know how much the extra context X helps a language model, we should train a model p_1 without access to X, train a model p_2 with access to X, and compare the accuracy of their predictions.

Measuring what is used But the original question raised by the introduction was not just *how much* information is contributed by context. It is already well-established that conditioning on long contexts is helpful, with existing experiments on long-range transformers effectively implementing the measurement in Eq. (2). Instead, we want to know *what* information in this context is actually used by models.

As a prototypical example, let us hypothesize that more than five tokens away from the target, models are *only* able to extract usable information from nouns. (In our experiments in Section 3, this "long-range context" will be considerably longer than 5 words.) For example, given the sentence:

Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.

we hypothesize that the LM distributions:

$$p_1(\textit{director} \mid \textit{Pierre Vinken, 61 years old, will join the board as a nonexecutive}) \quad (3)$$

$$\approx p_2(\textit{director} \mid \underbrace{\textit{Pierre Vinken years}}_{\text{noun-only context}}, \underbrace{\textit{the board as a nonexecutive}}_{\text{ordinary context}}) , \quad (4)$$

and more generally that

$$I_{\mathcal{V}}(X_{0:n} \to X_n) \approx I_{\mathcal{V}}([\mathsf{nouns}(X_{0:n-5}), X_{n-5:n}] \to X_n) \quad (5)$$

where $X_{i:j}$ is the sequence of tokens $[X_i, X_{i+1}, \ldots, X_{j-1}]$, $\mathcal{V}$ is a class of LMs, and nouns is a **context ablation** that extracts only the nouns from a given string. That is, we hypothesize that the amount of usable information contributed by the full context $X_{0:n}$ is the same as the amount contributed by the ablated context $[\mathsf{nouns}(X_{0:n-5}), X_{n-5:n}]$, so ablation removes no information.

The experiments in this paper generalize this experimental framework to other context ablations and hypotheses. Let f be an ablation and k an integer offset, and denote an **ablated context**:

$$f_k(X) = [f(X_{0:n-k}), X_{n-k:n}] \quad (6)$$

and an **ablated negative log-likelihood**:

$$\mathcal{L}(\theta, f, k) = -\mathbb{E} \log p_\theta(X_n \mid f_k(X_{0:n})) \quad (7)$$

Then, we can measure the effect of each ablation f on usable information via the following quantity:

Definition 2. The **ablated information** due to an ablation f at an offset k is:

$$\mathcal{A}(f, k) = \frac{I_{\mathcal{V}}(X_{0:n} \to X_n) - I_{\mathcal{V}}(f_k(X_{0:n}) \to X_n)}{I_{\mathcal{V}}(X_{0:n} \to X_n) - I_{\mathcal{V}}(X_{n-k:n} \to X_n)} \quad (8)$$

$$= \frac{\inf_\theta \mathcal{L}(\theta, f, k) - \inf_{\theta'} \mathcal{L}(\theta', n)}{\inf_{\theta''} \mathcal{L}(\theta'', n-k) - \inf_{\theta'} \mathcal{L}(\theta', n)} , \quad (9)$$

where $\mathcal{L}(\theta, i)$ is the (unablated) negative log-likelihood $-\mathbb{E} \log p_\theta(X_n \mid X_{n-i:n})$.

Intuitively, $\mathcal{A}(f, k)$ measures how much of the usable information *added by an extra k tokens* (the denominator) is *removed by applying the ablation f to those k tokens* (the numerator). If it is close to 0, almost no information is removed; if it is close to 1, almost all information is removed.

Figure 1: Calculation of the ablated likelihood $\mathcal{L}(\mathsf{nouns}, \ell : m \sim n)$ (Eq. (10)). A context ablation nouns (which deletes all non-noun words) is applied to the first ℓ tokens of the context, and likelihood is computed on the last $n - m$ (unablated) context tokens.

Evaluation in practice Eq. (9) provides a general framework for answering our core question in this paper: for a diverse set of context ablations and offsets, we will measure how much information is lost when a given ablation is applied at a given offset. A few modifications are required to turn this equation into a practical evaluation scheme:

Held-out evaluation: Eq. (7) involves an expectation over the sequence distribution $p(X)$. In practice, LMs must be trained on finite corpora, creating a risk of overfitting (Zhang et al., 2016). To address this issue, we approximate the infimum in Eq. (7) by fitting θ_1 on a training set, and computing ablated information on a held-out validation set. All reported results are an average of held-out likelihoods from two random initializations.

Batching: Given a fixed (training or test) dataset of strings $\mathcal{X}$ and a *maximum* context size of m, Eq. (7) should be estimated empirically as $-\frac{1}{|\mathcal{X}|} \sum_x \frac{1}{|x|} \sum_{i=0}^{|x|} \log p(X_i \mid f_k(X_{i-m:i}))$. This requires re-computing model predictions once for every token in the dataset. However, the transformer models we use here support efficient **batch** inference: training data is pre-segmented into sequences of at most length n, and $-\frac{1}{|\mathcal{X}|n} \sum_x \sum_{i=0}^{n} \log p(X_i \mid f_k(X_{0:i}))$ can be computed in a single forward pass. This is considerably more efficient but means that most tokens are evaluated with a context of length $< n$. As a compromise to ensure that evaluations contain long-range context, we accumulate losses on a subset:

$$\mathcal{L}(\theta, f, \ell : m \sim n) = -\frac{1}{|\mathcal{X}|(n-m)} \sum_x \sum_{i=\ell+m}^{\ell+n} \log p_\theta(X_i \mid [f(X_{0:\ell}), X_{\ell:i}]) \quad (10)$$

(visualized in Fig. 1). This can be read as "ℓ tokens of f-ablated context, followed by m to n tokens of unablated context". We will write $\mathcal{L}(\theta, m \sim n)$

when only unablated context is used. Because of the large number of experiments in this paper, we use Eq. (10) for all training and evaluation.

Model, data and training details For all experiments, our LM uses the GPT-2 model architecture (Radford et al., 2019) in the implementation of Wolf et al. (2020) with default hyperparameters. All models are trained from scratch on the WikiText-103 dataset (Merity et al., 2016), an English language modeling benchmark. Aside from ablations, no preprocessing is applied. A special separator token is inserted between ablated and unablated context. The training set contains 103,221,021 words, while the evaluation set contains 217,646 words.

A note on evaluation As in past work on evaluating language models (Brown et al., 1992), our evaluation of relative predictive information ultimately bottoms out in a conditional entropy (log-perplexity). Recent work has shown that other metrics, such as diversity of outputs, are important for evaluating the quality of LMs as models for language generation (Hashimoto et al., 2019; Caccia et al., 2020). Generation also depends on a number of other factors, such as choice of decoding procedure (Caglayan et al., 2020). Here, we focus on LMs as predictive models, measuring their ability to place an accurate distribution over future words and sentences, rather than their ability to generate useful or coherent text (see Appendix C). We want to emphasize that these results below apply to language models specifically, and not transformers applied to NLP tasks in general—the same analysis might give very different conclusions if applied to, e.g., question answering or summarization.

3 Experiments

In this section, we attempt to determine what information in transformer LM contexts is usable by measuring ablated information (Eq. (9)). Sections 3.1 and 3.2 describe our main results, with Section 3.1 focused on ordering and Section 3.2 focused on lexical information. Section 3.3 compares these results to ablations applied at evaluation time. Section 3.4 explores whether contexts can be further manipulated to improve model predictions.

3.1 Does order matter?

In this section we will examine the effects of different augmentations to the order within long-range context. We first train a **no information** model to minimize $\mathcal{L}(\theta, 0 \sim 512)$ and a **full information** model to minimize $\mathcal{L}(\theta, 512 \sim 1024)$. For each context ablation f, we train a model to minimize $\mathcal{L}(\theta, f, 512 : 0 \sim 512)$. Each ablation has access to more information than the no information model (because it conditions on extra tokens) and less information than the full information model (because an ablation has been applied to those tokens). Note that the LM operates on BPE-derived subword tokens for consistency with the way GPT-2 is typically used, but all ablations are defined at the word level, meaning, e.g., that we shuffle words rather than tokens.

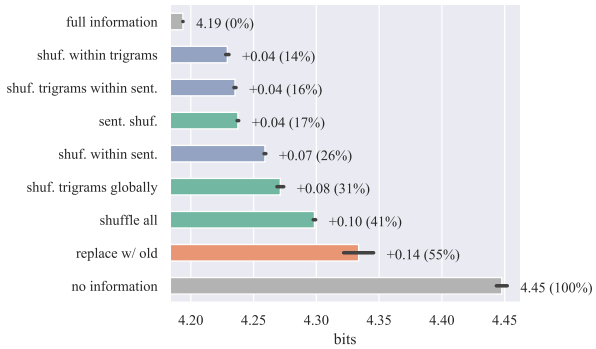
We use these trained models to calculate ablated information (Eq. (9)). To explore the effect of different context lengths, we stratify evaluation of the ablated information into two conditions: a **mid-range** condition in which likelihoods in Eq. (9) are of the form $\mathcal{L}(\cdot, f, 512 : 0 \sim 256)$, and a **long-range** condition with likelihoods $\mathcal{L}(\cdot, f, 512 : 256 \sim 512)$. (We call the former "mid-range" rather than "short-range" because most tokens are still predicted with significant unablated context; our experiments do not characterize sentence-internal modeling of syntactic well-formedness.) Results are shown in Figure 2 and discussed below.

Overall word order

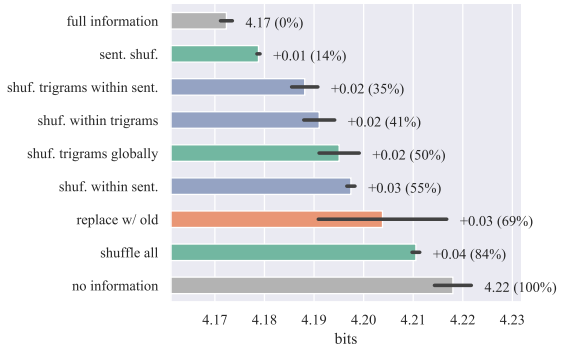
shuffle all
61 N.V., director the of Mr. Vinken Dutch group. as nonexecutive the 29. is Vinken, years Elsevier join old, publishing a Nov. will Pierre board chairman

shuf. trigrams globally
publishing group. N.V., the Dutch Mr. Vinken is join the board as a nonexecutive years old, will chairman of Elsevier Pierre Vinken, 61 director Nov. 29.

In the *shuffle all* ablation, f shuffles words uniformly at random, forcing the model to treat ablated context as a bag of words. In the *shuf. trigrams globally* ablation, the context is divided up into non-overlapping trigrams, the order of which is then permuted uniformly at random. Shuffling all words removes 41% of usable information in the mid-range condition and 84% in the long-range condition: *ordering information is important even very far from the target*. On the other hand, shuffling all trigrams removes 31% of usable information in the mid-range condition and 50% in the long-range condition: *local co-occurrence statistics carry a significant amount of usable information*.



(a) Mid-range condition (first 256 tokens after ablation)



(b) Long-range condition (tokens 256-512 after ablation)

Figure 2: Effect of **word order** on usable information. Bar labels show "change in ablated likelihood (ablated information)". The x axis shows ablated likelihood. Error bars represent 95% confidence intervals. Word-order changes that preserve local ordering remove only a small amount of information, while shuffling or replacement with thematically similar text remove more.

Word order within sentences

shuf. within sent.

61 director as the old, join will a Nov. board nonexecutive years Vinken, 29. Pierre is publishing the Vinken N.V., Mr. group. chairman Elsevier of Dutch

shuf. within trigrams

Vinken, Pierre 61 will old, years the board join a nonexecutive as Nov. director 29. Mr. Vinken is of Elsevier chairman the Dutch N.V., group. publishing

shuf. trigrams within sent.

years old, will as a nonexecutive join the board Pierre Vinken, 61 director Nov. 29. N.V., the Dutch chairman of Elsevier Mr. Vinken is publishing group.

Words are shuffled only within sentences according to one of three procedures: (1) a uniform random permutation of all the words in the sentence (*shuf. within sent.*), (2) a uniform random permutation of the words within each non-overlapping trigram in the sentence (*shuf. within trigrams*), and (3) a uniform random permutation of the order of the trigrams within the sentence (*shuf. trigrams within sent.*). (1) and (2) were also recently explored by Pham et al. (2020) in models for entailment, and more complex shuffling procedures have been explored in neuroscience contexts (Mollica et al., 2020). Here, (2) and (3) are chosen because they preserve local co-occurrence statistics ((3) more than (2)), while (2) also preserves the general linear information flow of the sentence.

Notably, the shuf. within trigrams (14% and 41%) and the shuf. trigrams within sent. (16% and 35%) ablations both remove relatively little usable information in both the mid- and long-range conditions. *Usable information is decreased only slightly by ablations that preserve local co-occurrence statistics and/or linear information flow*. (This includes transformations like *man bites dog* → *dog bites man* with significant effects on semantics!) In the long-range condition, uniform shuffling within sentences produces a larger effect, removing 55% of usable information.

Sentence order

shuf. sent.

Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group. Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.

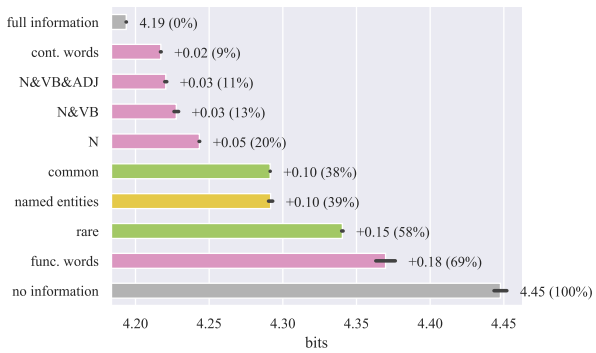
Next, *sentences* are shuffled within the context while their internal word order is unchanged. In the mid-range condition, this produces results comparable to the trigram shuffling experiments above (removing 17% of usable information); in the long-range condition, it has an even smaller effect (14%). Together with the previous experiment these results suggest that *prediction accuracy depends on information about local word co-occurrence, but not fine-grained word order or global position*.

Order of entire sections

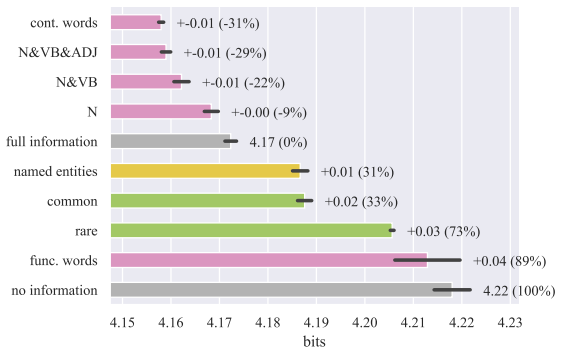
replace w/ old

Rudolph Agnew, 55 years old and former chairman of Consolidated Gold Fields PLC, was named a nonexecutive director of this British industrial conglomerate.

A possible hypothesis about LM behavior is that the main function of long-range context is to provide more information about the general *topic* of the document, including clues about vocabulary and style. To test this, the ablation replaces its entire input with the 512 tokens that immediately precede it in the source document (which in general will be topically similar). This transformation removes significant information in both mid- and long-range conditions (55% and 69%). *Long-range context is*



(a) Mid-range condition (first 256 tokens after context)



(b) Long-range condition (tokens 256-512 after context)

Figure 3: Effect of **word identity** on usable information. Labels are as in Fig. 2. Several ablations, including deletion of all words except nouns, preserve most usable information in the mid-range condition, and *improve* model accuracy in the in the long range.

not simply a source of topic information: earlier text on the same theme is in some cases nearly as uninformative as no text at all.

3.2 Do all words matter?

Our next experiments focus on lexical rather than structural information, using ablations that delete selected words from the context. Training and evaluation setups are exactly as in Section 3.1. Here, unlike the previous section, ablations will generally cause the number of tokens in a given context to decrease; in this case ablations also insert padding tokens to the *beginning* of the context window to preserve the original number of tokens. Results are shown in Fig. 3.

Parts of speech

N

Pierre Vinken years board director Nov. Mr. Vinken chairman Elsevier N.V. publishing group

N & VB

Pierre Vinken years will join board director Nov. Mr. Vinken chairman Elsevier N.V. publishing group

N & VB & ADJ

Pierre Vinken years old will join board nonexecutive director Nov. Mr. Vinken chairman Elsevier N.V. Dutch publishing group

cont. words (N & VB & ADJ & ADV)

Pierre Vinken years old will join board nonexecutive director Nov. Mr. Vinken chairman Elsevier N.V. Dutch publishing group

func. words

, 61 , the as a 29 . is of , the .

As in the initial example from Section 2, we retain only words whose part of speech tag is in a given set. We use the spaCy model (Honnibal et al., 2020) for part-of-speech tagging, and examine five sets: (1) *nouns only*, (2) *nouns and verbs*, (3) *nouns, verbs, and adjectives*, (4) *content words* (nouns, verbs, adjectives, and adverbs), and (5) *function words* (all words *except* nouns, verbs, adjectives, and adverbs).

In the mid-range condition, deleting all words but nouns removes only 20% of usable information; deleting all but nouns and verbs removes only 13%. *Most usable information, even in mid-range context, appears to be captured by nouns and verbs.* Retaining only function words causes a considerably greater loss of information.

In the long-range condition, results are even more striking: *retaining only content words* improves *predictions over the "full information" experiment*. Like Shannon information, $\mathcal{V}$-information is defined to be non-negative (Xu et al., 2020), and the result in Fig. 3 is a consequence of our finite-sample approximation based on held-out likelihood. The effect is robust across multiple training runs from random initializations. As there is a significant gap between the training and validation perplexity of our model (roughly 11%), we hypothesize that this change occurs because the ablation preserves semantic content while reducing the original model's ability to overfit. We believe this is an important subject for future investigation.

Named entities

named entities

Pierre Vinken 61 years old Nov. 29 Vinken Elsevier N.V. Dutch

As an alternative to the *topic hypothesis* evaluated under "Order of entire sections" above, we might

hypothesize that long-range contexts are useful because they provide a reservoir of named entities likely to be referred to again. Here, the ablation retains only spans tagged as named entities or quantities by spaCy. While significantly worse than the noun ablation discussed above, retaining only entities results removes only about a third of usable information in both conditions (39% and 31%).

Word frequency

common
Pierre years old join board director . Mr. chairman Dutch publishing group .

rare
Vinken nonexecutive Nov. Vinken Elsevier N.V.

Another natural question is whether rare words or frequent words are more important: information about frequent context words might help models estimate fine-grained document-level frequencies of those words account for most of the terms in Eq. (7); rare words are likely to be more informative about the content of the document itself.

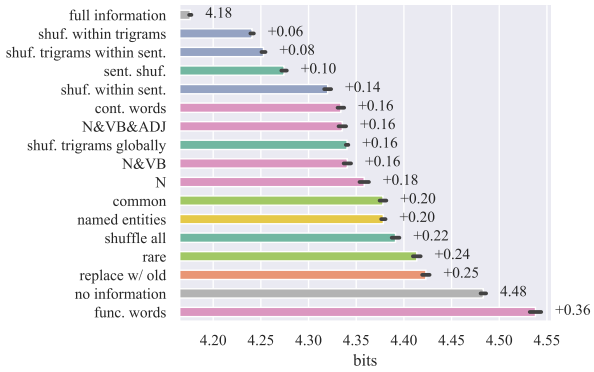
We partition the vocabulary into a set of *rare words*, corresponding to the least frequent $\sim 98\%$ of word types and 20% of word tokens, and *frequent words*, the most frequent $\sim 2\%$ of types and 80% of tokens. Both ablations remove a significant amount of information relative to the POS-based ablations above, but retaining only frequent words improves perplexity relative to rare words in both the mid- and long-range conditions.

Appendix B presents versions of these experiments trained and evaluated on even longer contexts. Conclusions are largely the same as above.

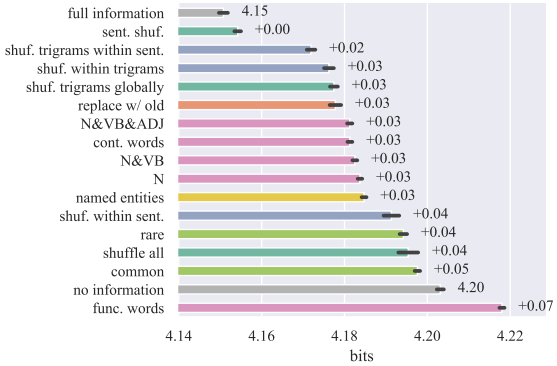
3.3 Evaluating on augmented data

We motivated the use of $\mathcal{V}$-information in Section 2 by arguing that it more clearly distinguished between prediction errors attributable to *loss of information* and prediction errors attributable to *malformed and out-of-distribution* model inputs. To put our results in context, we repeat several of the previous experiments in the evaluation paradigm of Khandelwal et al. (2018), which is designed to measure test-time sensitivity rather than usable information.

We train a new model to minimize $\mathcal{L}(\theta, 512 \sim 1024)$ while randomly truncating the first 512 context tokens and replacing them with padding tokens (to ensure that the model has seen padding tokens at training time). We then evaluate this model on



(a) Mid-range condition (first 256 tokens after ablation)



(b) Long-range condition (tokens 256-512 after ablation)

Figure 4: Loss of information resulting from ablations at *evaluation time only*. x-axis and labels show ablated negative log-likelihoods. Some locality-preserving ablations (high PMI, shuf. sent.) have a small effect, but most affect likelihood significantly (including lexical ablations that do not remove usable information).

the set of ablations shown in Section 3.1 and Section 3.2. For the full information model in Fig. 4, we evaluate on ordered context windows with no padding tokens; for the no information model, we evaluate on context windows in which the first 512 tokens are all padding tokens.

In the mid-range condition, the least destructive ablations are shuffling within trigrams and shuffling the order of trigrams within sentences: models appear to be reasonably robust to this kind of data transformation without specific training on it. Importantly, lexical ablation experiments have a large impact in this evaluation, underlining the extent to which the two experimental paradigms characterize different aspects of model behavior. Figure 5 in Appendix A shows a side-by-side comparison of these experiments and the ones in Sections 3.1–3.2.

3.4 Making better language models?

The lexical ablation experiments in Section 3.2 indicated that model accuracy could be *improved* by

selective deletion of context words. Can this effect be exploited to further improve models? As a simple experiment, we attempted to *replace* all padding tokens in the *nouns+verbs* ablation of Section 3.2 with nouns and verbs from further back in the context—effectively providing the model with an even longer-range view of an informative context representation.

This experiment slightly increased usable information in the mid-range condition (0.2%), but *decreased* it in the long range-range condition (0.6%). *Longer contexts, even of a kind previously found to be informative, did not provide additional usable information.* These results are consistent with our earlier hypothesis that the previously observed effect resulted from a reduction in overfitting—if removing information increased performance by reducing overfitting, then it is reasonable that adding information back results in more overfitting.

4 Related Work

Context in count-based and discriminative LMs The earliest learned LMs were *count-based* (e.g., Kneser and Ney, 1995): they estimated $p(x_n \mid x_{0:n})$ based on a (smoothed) empirical n-gram frequency $\#(x_{0:n})/\#(x_{0:n-1})$ (where $\#(x)$ is the number of times the sequence x appears in training data). As the number of distinct n-gram counts grows exponentially in n, it was typically set to a small value. Count-based models have a clear dependence on context: any token within the last n words that also appears in a training n-gram is relevant, anything further back is not.

Subsequent models improved on these by allowing the use of skip-grams, caches, and feature-based models (Goodman, 2001; Bengio et al., 2003). Some of these in principle allowed the use of unlimited-length contexts, but only by imposing strong restrictions on the ways in which context features could interact.

Context in RNN LMs Recurrent neural network language models (Mikolov et al., 2010; Elman, 1990) provide a more expressive mechanism for the use of long-range context: models write to a recurrent "state vector" which can be carried arbitrarily far into the future. Computational issues limit the effective context size such models can be practically trained on, but this size is still significantly greater the models mentioned above: as previously noted, Khandelwal et al. (2018) revealed influence from up to 200 tokens of context. Similar effects are reported by Sankar et al. (2019) for neural dialogue models, and Li et al. (2016) describe an alternative procedure for ablating contexts.

Context in Transformer LMs Transformers introduce yet another mechanism for extracting information from long-range context: attention. Attention is also used with RNNs, but typically with just a single head—the hidden state still carries most of the information. In transformers, context enters into predictions primarily via unbounded random access. These models appear to benefit from significantly longer contexts than previous models.

Some recent work that investigates the behavior of individual transformer attention heads (Clark et al., 2019; Voita et al., 2019). This work finds that certain attention heads are sensitive to things like word frequency, positional information, and certain syntactic phenomena. While extremely informative about the computational structures implemented by fixed models, these approaches do not necessarily reveal anything about usable information: indeed, patterns of attention do not necessarily correlate with model predictions (Jain and Wallace, 2019).

Other related work Our finding that fine-grained ordering information contributes little usable information is consistent with Rae et al. (2019)'s finding that long-range contexts could be informatively summarized in fixed-sized vectors; our finding that most usable information is carried by nouns is consistent with earlier findings about both specialized neural architectures (Henaff et al., 2016) and discourse representations in feature-based models (Barzilay and Lapata, 2008). Our approach also shares similar motivations to information-theoretic work on *probing* (Voita and Titov, 2020; Pimentel et al., 2020), which uses related tools to interpret linguistic structure in LM representations rather than characterizing their effect on LM predictions. Several recent papers have explored the effect of training-time and test-time ablations in models for other data analysis tasks: Pham et al. (2020) find that shuffling experiments have a limited effect on the accuracy of models for natural language inference, while Perez et al. (2021) describe several experiments aimed at *introducing* usable information for several question answering and sentence understanding tasks.

5 Discussion

We have investigated the extent to which transformer models can use structural and lexical information in long-range contexts for English language modeling. Experiments demonstrated that this information is primarily contained in content words and local ordering statistics: ablations that remove other kinds of information from context have little effect on models' predictive accuracies. In contrast, retaining only information about document identity or named entities causes significant drops in predictive accuracy: the effectiveness of long contexts is not explained by the presence of topic or named entity information alone.

Crucial to obtaining these results was a measure of *ablated usable information* grounded in the accuracy of models trained and tested on ablated contexts. Past work on context in LMs has primarily measured the influence of evaluation-time ablations. Sometimes these two notions of context-sensitivity coincide (e.g., trigram shuffling) and sometimes they do not (e.g., removal of lexical information). Our results also offer a jumping-off point for future modeling work. They motivate more *efficient*, compressed context representations that better preserve the information that is usable by current models. They motivate more *accurate* models by developing new context representations that make currently unusable information more prominent.

Several questions remain unanswered by our experiments. Do ablations affect the quality of text generated by models? (In particular, does the usable information added by long contexts improve predictability of syntax, semantics, or simply document-level word frequency statistics?) More fundamentally, do observations about usable information reflect limitations of transformers or fundamental, (Shannon-)information-theoretic properties of English? Our results suggest that at least some of these effects are model-specific: deleting function words cannot add information, but improves held-out model accuracy. A complete answer to this question will require more detailed exploration, including a better understanding of human predictions in comparable settings.

Acknowledgments

Thanks to Carina Kauf and Greta Tuckute, Evelina Fedorenko and Roger Levy for valuable discussions. We acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing HPC resources that contributed to the results reported within this paper.

Impact Statement

Across initial exploration, evaluation conditions and training runs, experiments in this paper required roughly 100 training runs on the WikiText-103 dataset. As discussed in Section 2, model size and batched evaluation were both used to minimize the energy demands of these experiments; experiments themselves were performed at the Massachusetts Green HPC center, a carbon-neutral supercomputing facility. Ultimately, results in Section 3 provide guidance toward the design of models that use context more efficiently and motivate the large-scale empirical study conducted here.

References

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, Jennifer C Lai, and Robert L Mercer. 1992. An estimate of an upper bound for the entropy of english. *Computational Linguistics*, 18(1):31–40.

Ralf D Brown. 2011. The CMU-EBMT machine translation system. *Machine translation*, 25(2):179.

Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. 2020. Language GANs falling short. In *International Conference on Learning Representations*.

Ozan Caglayan, Pranava Madhyastha, and Lucia Specia. 2020. Curious case of language generation evaluation metrics: A cautionary tale. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2322–2328, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *URL https://openai.com/blog/sparse-transformers*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Joshua T Goodman. 2001. A bit of progress in language modeling. *Computer Speech & Language*, 15(4):403–434.

Michael Hahn. 2020. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171.

Tatsunori Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1689–1701, Minneapolis, Minnesota. Association for Computational Linguistics.

Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. 2016. Tracking the world state with recurrent entity networks. In *ICLR*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *NAACL-HLT*.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 284–294, Melbourne, Australia. Association for Computational Linguistics.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *International Conference on Learning Representations*.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184. IEEE.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *CoRR*, abs/1609.07843.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh annual conference of the International Speech Communication Association*.

F. Mollica, Matthew Siegelman, Evgeniia Diachek, S. Piantadosi, Zachary Mineroff, Richard Futrell, Hope H. Kean, Peng Qian, and E. Fedorenko. 2020. Composition is the core driver of the language-selective network. *Neurobiology of Language*, 1:104–134.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. Rissanen data analysis: Examining dataset characteristics via description length. *arXiv preprint arXiv:2103.03872*.

Thang M Pham, Trung Bui, Long Mai, and Anh Nguyen. 2020. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? *arXiv preprint arXiv:2012.15180*.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

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

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, and Timothy P. Lillicrap. 2019. Compressive transformers for long-range sequence modelling.

Chinnadhurai Sankar, Sandeep Subramanian, Christopher Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. *arXiv preprint arXiv:1906.01603*.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

Sinong Wang, Belinda Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

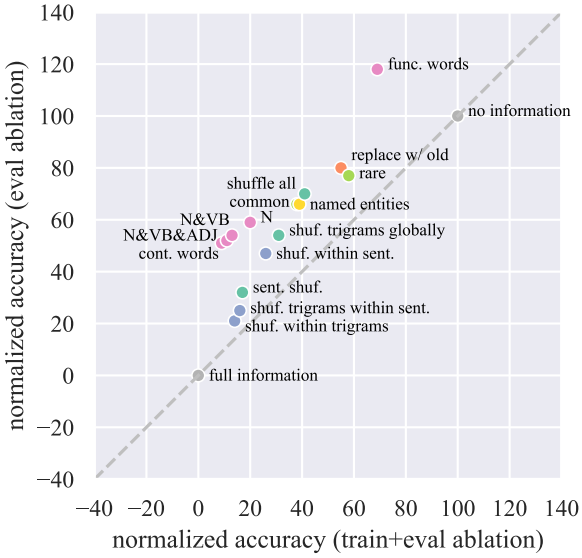
Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest,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

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. In *International Conference on Learning Representations*.

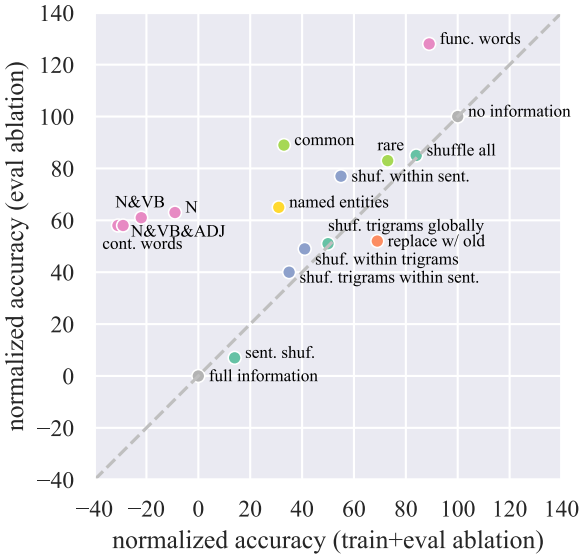
Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.

A Comparison of Experimental Paradigms

In Figure 5 we show the contrast between the experimental paradigm of Sections 3.1–3.2 and that of Section 3.3. Especially for the experiments involving parts of speech, we see a significant difference in both the quantitative and qualitative results across the two paradigms.



(a) Mid-range condition (first 256 tokens after ablation)

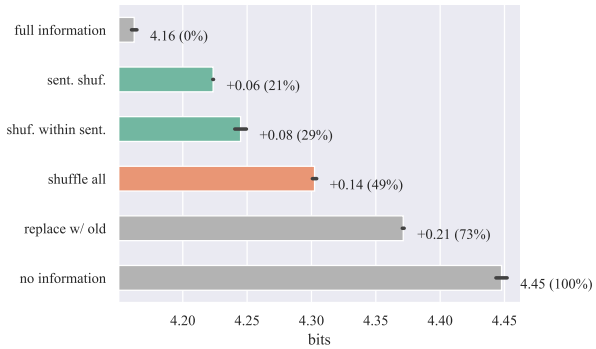


(b) Long-range condition (tokens 256-512 after ablation)

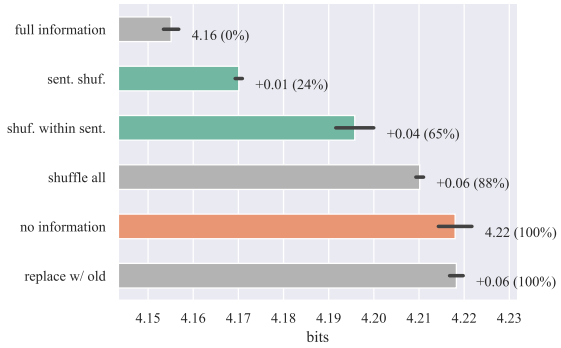
Figure 5: Comparison of model performance in the train+eval and eval-only settings. The units represent the percentage of the gap between the full information and no information models/contexts. That way, if a point falls on the dotted $y = x$ line, then that ablation has the same relative effect in each paradigm. If a point falls above the dotted line, then that ablation leads to better relative performance in the train+eval paradigm, and if a point falls below the dotted line, then that ablation leads to better relative performance in the eval-only paradigm.

B Longer Context Window

Here we report the results of repeating the experiments of Sections 3.1 and 3.2 with ablated contexts of size 1024 tokens instead of 512 tokens in order to verify that the behavior we observed is not specific to the size of context window we chose.

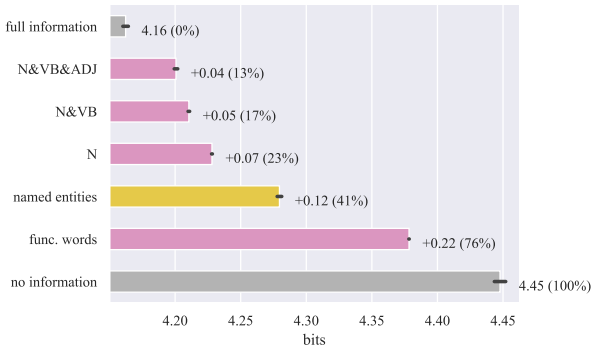


(a) Mid-range condition (first 256 tokens after ablation)

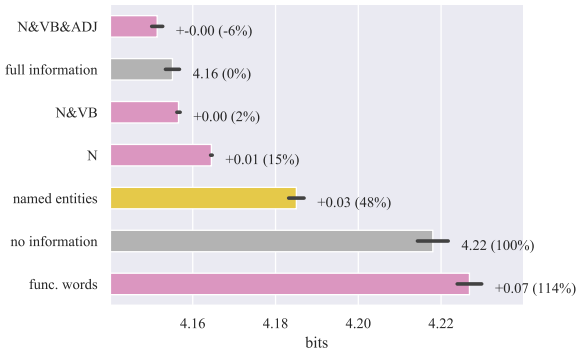


(b) Long-range condition (tokens 256-512 after ablation)

Figure 6: Effect of **word order** on usable information. Bar labels show "change in ablated likelihood (ablated information)". The x axis shows ablated likelihood. Error bars represent 95% confidence intervals. Ablated contexts contain 1024 tokens, but results are consistent with results on 512-token contexts.



(a) Mid-range condition (first 256 tokens after ablation)



(b) Long-range condition (tokens 256-512 after ablation)

Figure 7: Effect of **word identity** on usable information. Labels are as in Fig. 6. Ablated contexts contain 1024 tokens, but results are consistent with results on 512-token contexts.

C Sample Generations

The purpose of this section is to verify that models trained on ablated contexts can still generate text that is comparable to text generated by a model trained with full contextual information. We select a prompt from a randomly chosen Wikipedia article in the WikiText-103 validation set; each model generates a sentence (after finishing the sentence in progress) given the appropriately ablated version of the prompt. The prompt consists of 768 tokens, the last 256 of which remain unchanged for all versions of the prompt, so that the ablations are in the long range relative to the point of generation. The prompt and generations are as follows:

prompt:

> at two independent schools for boys: Sussex House School, a day school in Chelsea's Cadogan Square, and the City of London School, a day school on the North Bank of the River Thames in London's financial district (known as the City of London). Attending school became difficult for Radcliffe after the release of the first Harry Potter film, with some fellow pupils becoming hostile, though he says it was people just trying to "have a crack at the kid that plays Harry Potter" rather than jealousy.
>
> As his acting career began to consume his schedule, Radcliffe continued his education through on-set tutors. He admitted he was not very good at school, considering it useless and finding the work "really difficult." He achieved A grades in the three AS-level exams that he took in 2006, but decided to take a break from education and did not go to college or university. Part of his reasoning was that he already knew he wanted to act and write, and that it would be difficult to have a normal college experience. "The paparazzi, they'd love it," he told Details magazine in 2007. "If there were any parties going on, they'd be tipped off as to where they were."

= = Career = =

= = = Harry Potter = = =

In 2000, producer David Heyman asked Radcliffe to audition for the role of Harry Potter for the film adaptation of Harry Potter and the Philosopher's Stone, the best-selling book by British author J.K. Rowling. Rowling had been searching for an unknown British actor to personify the character, and the movie's director Chris Columbus recalled thinking, "This is what I want. This is Harry Potter", after he saw a video of the young actor in David Copperfield. Eight months later, and after several auditions, Radcliffe was selected to play the part. Rowling also endorsed the selection saying, "I don't think Chris Columbus could have found a better Harry." Radcliffe's parents originally turned down the offer, as they had been told that it would involve six films shot in Los Angeles. Warner Bros. instead offered Radcliffe a two-movie contract with shooting in the UK; Radcliffe was unsure at the time if he would do any more than that.

The release of Harry Potter and the Philosopher's Stone (released as Harry Potter and the Sorcerer's Stone in the United States) took place in 2001. Radcliffe received a seven figure salary for the lead role, but asserted that the fee was "not that important" to him; his parents chose to invest the money for him. The film was highly popular and was met with positive reviews, and critics took notice of Radcliffe: "Radcliffe is the embodiment of every reader's imagination. It is wonderful to see a young hero who is so scholarly looking and filled with curiosity and who connects with very real emotions, from solemn intelligence and the delight of discovery to deep family longing, "wrote Bob Graham of the San Francisco Chronicle.

A year later Radcliffe starred in Harry Potter and the Chamber of Secrets, the second installment of the series. Reviewers were positive about the lead actors' performances but had polarised opinions on the movie as a whole. The 2004 release Harry Potter and the Prisoner of Azkaban was the third film in the series. Radcliffe 's performance was panned by New York Times journalist A.O. Scott, who wrote that Watson had to carry him with her performance. Next was Harry Potter and the Goblet of Fire in 2005. The film was the second-highest gross

full information:

sing Potter film in the series at the time of its release. After appearing in Harry Potter and the Philosopher's Stone, Radcliffe tried to avoid trying to improve his role.

shuffle all:

sing Potter film in the series (behind Harry Potter and the Chamber of Secrets), the highest grossing Harry Potter film in the series. The following year, Harry Potter and the Goblet of Fire was the top-grossing Potter film, making it the highest grossing Harry Potter film in the series.

shuf. within sent.:

sing Potter film in the series (until 2009's Gold Glove), selling 1.5 million tickets in its first week. The title of the film's sequel was also a major factor in the success of Philosopher, while its underperformance dropped because of series 3's Sci-Fi Channel.

shuf. within trigrams:

sing Potter film at the time of its release. The film had a financial success worldwide and received five Academy Award nominations at the 56th Academy Awards.

shuf. trigrams within sent.:

sing Potter film in the series at the United States box office. His appearance in Harry Potter and the Philosopher's Stone did not make it to the United States until 2006.

shuf. trigrams globally:

sing Potter film at the time and was the third highest grossing film of 2006, the highest-grossing film in the series. In 2006, Radcliffe acted in his first major role in Harry Potter and the Deathly Hallows – Part I, where he played Harry's Chief Engineer, Harry Potter's primary antagonist.

sent. shuf.:

sing Potter film in the series at the time of its release. At the end of 2009, new backers were looking for a leading role in the film, with Radcliffe starting the role of Harry Potter and the Philosopher's Stone.

cont. words:

sing Harry Potter film of all time after Harry Potter and the Goblet of Fire, the highest-grossing film of the 2007–2008 financial year. The film's success meant that he and his wife bought a four-bedroom apartment in the 1960s, "ready to keep them warm", adding "They could never make another effort to make something exciting for people who were usually very aware of the housing shortage but hadn't been very conscious of that".

N & VB & ADJ:

sing animated film in the series after The Lion King. A similar amount of praise came in 2008 when The Magic King was released.

N & VB:

sing animated film in the series. A significant amount of critical reviews were given by publications in The Washington Post, San Francisco Chronicle, and The Washington Post. The film was released in theaters in 2008.

N:

sing animated film in the series.

= Appearances = =

named entities:

sing animated film in the series.

= = Appearances extended Persons (1990 – 2014) = =

rare:

sing animated film in the series.

= = Part two = =

common:

sing animated film in the series. A review in The New York Times found that Hans was not as strong as Watson but as well as Mr. Trough and Mr. Trough.