

# Infini-gram: Scaling Unbounded $n$ -gram Language Models to a Trillion Tokens

Jiacheng Liu<sup>♡</sup> Sewon Min<sup>♡</sup>

Luke Zettlemoyer<sup>♡</sup> Yejin Choi<sup>♡♣</sup> Hannaneh Hajishirzi<sup>♡♣</sup>

<sup>♡</sup>Paul G. Allen School of Computer Science & Engineering, University of Washington

<sup>♣</sup>Allen Institute for Artificial Intelligence liujc@cs.washington.edu

## Abstract

Are  $n$ -gram language models still relevant in this era of neural large language models (LLMs)? Our answer is *yes*, and we showcase their values in both text analysis and improving neural LLMs. This was done by modernizing  $n$ -gram LMs in two aspects. First, we train them at the same data scale as neural LLMs—**5 trillion tokens**. This is the largest  $n$ -gram LM ever built. Second, existing  $n$ -gram LMs use small  $n$  which hinders their performance; we instead allow  $n$  to be arbitrarily large, by introducing a new  $\infty$ -gram LM with backoff. Instead of pre-computing  $n$ -gram count tables (which would be very expensive), we develop an engine named **infini-gram**—powered by suffix arrays—that can compute  $\infty$ -gram (as well as  $n$ -gram with arbitrary  $n$ ) probabilities with **millisecond-level latency**. The  $\infty$ -gram framework and infini-gram engine enable us to conduct many novel and interesting analyses of human-written and machine-generated text: we find that the  $\infty$ -gram LM has fairly high accuracy for next-token prediction (47%), and can complement neural LLMs to greatly reduce their perplexity. When analyzing machine-generated text, we also observe irregularities in the machine- $\infty$ -gram agreement level with respect to the suffix length, which indicates deficiencies in neural LLM pretraining and the positional embeddings of Transformers.

	<b>Project Homepage</b>	<a href="https://infini-gram.io">infini-gram.io</a>
	<b>Web Interface</b>	<a href="https://infini-gram.io/demo">infini-gram.io/demo</a>
	<b>API Endpoint</b>	<a href="https://api.infini-gram.io">api.infini-gram.io</a>
	<b>Python Package</b>	<a href="https://pypi.org/project/infini-gram">pypi.org/project/infini-gram</a>
	<b>Source Code</b>	<a href="https://github.com/liujch1998/infini-gram">github.com/liujch1998/infini-gram</a>

**Prompt** ... conducts research at the Paul G. Allen School of Computer Science and Engineering, University of

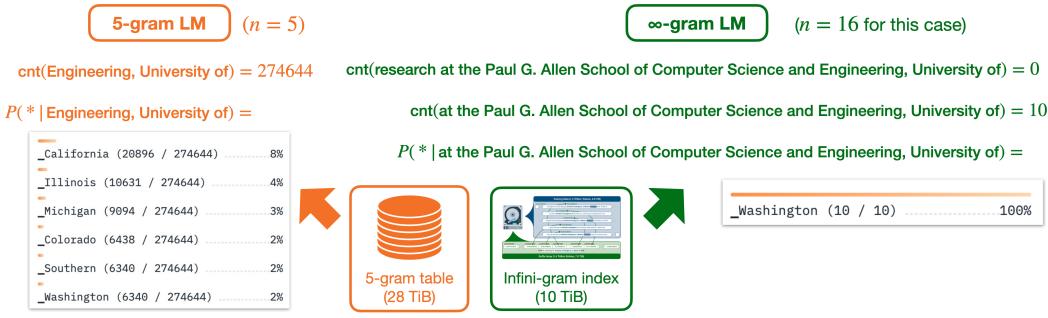


Figure 1: An example where a 5-gram LM gives an incorrect prediction but the  $\infty$ -gram gives the correct prediction by using the longest suffix of the prompt that has a non-zero count in the corpus. The counting and distribution estimate in  $\infty$ -gram LM are powered by our infini-gram engine.

## 1 Introduction

When pretrained on trillion-token corpora, neural large language models (LLMs) achieve groundbreaking performance (Touvron et al., 2023a; Geng & Liu, 2023; Groeneveld et al., 2024). However, we do not yet know how such data scale would benefit other language modeling approaches. In particular, how well does the classical,  $n$ -gram language model (LM) perform if estimated from such massive corpora? In other words, *are  $n$ -gram LMs still relevant in this era of neural LLMs?*

Our answer is *yes*. As we will show,  $n$ -gram LMs are useful for both text analysis and improving neural LLMs. Yet we need to first modernize the canonical  $n$ -gram LM in two aspects: the training data size, and the value of  $n$ . To achieve broader data coverage, we scale up the training data for  $n$ -gram LMs to **5 trillion tokens**, by combining some of the largest open-source text corpora. This is the largest  $n$ -gram LM ever built. Historically,  $n$ -gram indexes have been built only for small  $n$ 's (e.g.,  $n \leq 5$ ; Brants et al. (2007)), because the size of naive  $n$ -gram count table grows almost exponentially wrt  $n$ . We instead find significant value in increasing the value of  $n$ . As illustrated in Figure 1, a 5-gram LM is poorly predictive of the next token, because it discards the rich context in the prompt; meanwhile, if we can use a larger  $n$  (in this case  $n = 16$ ), the prediction becomes much more accurate. As such, we develop our  $n$ -gram LM with *unbounded  $n$* , or in other words, an  $\infty$ -gram LM. We use a variant of *backoff* (Jurafsky & Martin, 2000), where we resort to smaller  $n$  when longer  $n$ -grams have a zero count. Due to *sparsity* in the  $\infty$ -gram estimates, in some of the later experiments (§5), we will interpolate between the  $\infty$ -gram LM and neural LMs to yield a hybrid LM upon which perplexity can be computed.

We develop a low-latency, resource-efficient engine to serve this massive  $\infty$ -gram LM. Instead of building an explicit  $n$ -gram count table, which is infeasible for arbitrarily large  $n$  and such extreme data scale, we power the  $\infty$ -gram LM with a *suffix array* of the dataset – a data structure that supports fast  $n$ -gram counting, and is efficient in both storage space and compute. Our index takes 7 bytes of storage per token (3.5x overhead compared to the raw dataset), and on a dataset with 1.4 trillion tokens, it can be built with a single 128-core CPU node in about 2 days, using 10 TB of disk storage. Average inference latency is less than **20 milliseconds** for counting an  $n$ -gram and finding *all* positions of its occurrence (regardless of how large  $n$  is or how frequently the  $n$ -gram appears), and under 200 milliseconds for all other query types including  $\infty$ -gram language modeling and decoding. All indexes stay on-disk at inference time. We refer to this engine as **infini-gram**.

Analyses with  $\infty$ -gram (§4) offers new insights into human-written and machine-generated text. We found that  $\infty$ -gram has a fairly high accuracy (47%) when predicting the next token given a prefix of a human-written document, and this accuracy is higher on tokens where the effective  $n$  is larger. In contrast, conventional  $n$ -grams (with small  $n$ ) are insufficient in capturing a long enough context to predict the next token (29% accuracy). Moreover, we show that  $\infty$ -gram can complement neural LMs and reach better performance when combined: heuristically interpolating between the estimates made by  $\infty$ -gram and neural LMs can greatly reduce perplexity (by up to 73%) compared to the neural LMs alone, even when the neural LM is as large as 70B (§5). When analyzing the level of agreement with  $\infty$ -gram, nucleus sampling (Holtzman et al., 2019) from neural LMs produces machine-generated text with an agreement plot most similar to human-written text, among other decoding methods like greedy decoding and temperature sampling; for greedy decoding, we observe significant fluctuation in the agreement level wrt the suffix length, which indicates deficiencies in neural LM pretraining and the positional embeddings of Transformers.

We are hosting host a public web interface and an API endpoint that serve  $n$ -gram/ $\infty$ -gram queries on several popular open corpora: Dolma (Soldaini et al., 2023), RedPajama (Together, 2023), Pile (Gao et al., 2020), and C4 (Raffel et al., 2019). We also release a Python package for local serving and building new indexes, as well as our source code. We hope these tools can enable more insightful analysis and understanding of large text corpora, and open up new avenues for data-driven language modeling.

## 2 $\infty$ -gram LM: Extending $n$ -gram LMs with Unbounded $n$

**Background:  $n$ -gram LM.** The  $n$ -gram LM is a classical, statistical language model based on counting the occurrences of  $n$ -grams. In its most simple form, the probability of a token  $w_i$  given a context  $w_{i-(n-1):i-1}$  is estimated as  $P_n(w_i \mid w_{i-(n-1):i-1}) = \frac{\text{cnt}(w_{i-(n-1):i-1}w_i \mid \mathcal{D})}{\text{cnt}(w_{i-(n-1):i-1} \mid \mathcal{D})}$ , where  $\text{cnt}(\mathbf{w} \mid \mathcal{D})$  is the number of times the  $n$ -gram  $\mathbf{w}$  appears in the training data  $\mathcal{D}$  (i.e., a corpus), and  $n$  is a pre-defined hyperparameter. (When  $n = 1$ , we define  $w_{i-(n-1):i-1}$  as the empty string  $\epsilon$ , whose count is equal to  $|\mathcal{D}|$ .) However, this naive version of  $n$ -gram LM faces the sparsity issue: the numerator may be zero, resulting in an infinite perplexity. One common solution is **backoff** (Jurafsky & Martin, 2000): on an instance-wise basis, when the numerator is zero we decrease  $n$  by one, and do this repeatedly until the numerator becomes positive. One caveat in backoff is that it does not yield a valid distribution for  $P_n(*) \mid w_{i-(n-1):i-1}$ , because the **effective  $n$**  is dependent on  $w_i$ . Therefore, further probability discounting is required to normalize this distribution (e.g., Katz backoff (Katz, 1987)).

Conventionally,  $n$ -gram LMs have been implemented by building an  $n$ -gram count table of the training data. This table stores all unique  $n$ -grams that appear in the training data, each associated with its count. Such  $n$ -gram count tables are huge and grow almost exponentially wrt  $n$ . For example, the 5-gram count table for a 1.4-trillion-token corpus would consume 28 TB of disk space. As a result, previous  $n$ -gram LMs are limited to very small  $n$ , most commonly  $n = 5$ , and to frequent  $n$ -grams only (e.g., Franz & Brants (2006)). As we illustrated in Figure 1 and will further quantify in §4, the problem with small  $n$  is that it discards richer context, making such  $n$ -gram LMs poorly predictive of future tokens.

**$\infty$ -gram LM.** The  $\infty$ -gram LM is a generalization of the  $n$ -gram LM, where conceptually we start backing off from  $n = \infty$ . We use a variant of backoff: we backoff only when the denominator is zero. This means we stop backing off as soon as the denominator becomes positive, upon which the numerator might still be zero. On an instance-wise basis, the effective  $n$  is equal to one plus the length of the prompt’s **longest suffix** that appears in the training data.

**For the rest of this paper, we will use “ $\infty$ -gram” to refer to the  $\infty$ -gram LM.**  $\infty$ -gram is formally defined as

$$P_\infty(w_i \mid w_{1:i-1}) = \frac{\text{cnt}(w_{i-(n-1):i-1}w_i \mid \mathcal{D})}{\text{cnt}(w_{i-(n-1):i-1} \mid \mathcal{D})}$$

where  $w_{1:i-1}$  are all tokens preceding  $w_i$  in the document, and

$$n = \max\{n' \in [1, i] \mid \text{cnt}(w_{i-(n'-1):i-1} \mid \mathcal{D}) > 0\}.$$

Unlike Katz backoff,  $P_\infty(*) \mid w_{1:i-1}$  is a valid distribution by construction and does not require discounting. This is because the effective  $n$  is solely dependent on  $w_{1:i-1}$  and does not depend on  $w_i$ , and  $\sum_{w_i \in \mathcal{V}} \text{cnt}(w_{i-(n-1):i-1}w_i \mid \mathcal{D}) = \text{cnt}(w_{i-(n-1):i-1} \mid \mathcal{D})$ .

Further, we define the **sparsity** of this  $\infty$ -gram estimate: an estimate is sparse iff  $P(w_i \mid w_{i-(n-1):i-1}) = 1$  for one of the  $w_i \in \mathcal{V}$ , and is zero for all other tokens in the vocabulary. Intuitively, this means there is only one possible next token given this context, according to the training data. As we will show in §4, sparse estimates are more predictive of the actual tokens than non-sparse ones.

**Interpolating with neural LMs.**  $\infty$ -gram estimates contain zero probabilities, which may lead to infinite perplexity. We do *not* attempt to compute the perplexity of the  $\infty$ -gram itself. Instead, we interpolate it with neural LMs and show perplexity improvement over the neural LMs alone (§5). The combined model is

$$P(y \mid x) = \lambda P_\infty(y \mid x) + (1 - \lambda) P_{\text{neural}}(y \mid x),$$

where  $\lambda \in [0, 1]$  is a hyperparameter.

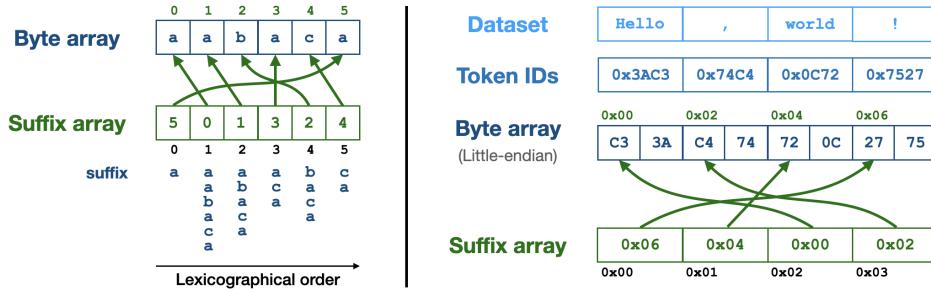


Figure 2: **Left:** the suffix array for a toy string. **Right:** illustration of the suffix array in the infini-gram index, with  $N = 4$  tokens in the dataset.

### 3 Infini-gram: A Performant Engine for $n$ -gram/ $\infty$ -gram Queries

We train  $\infty$ -gram on modern, trillion-token text corpora. However, it is practically infeasible to build  $n$ -gram count tables with unbounded  $n$  for such massive datasets. In this section, we describe our infini-gram engine that processes  $n$ -gram/ $\infty$ -gram queries efficiently. Infini-gram is powered by a data structure called **suffix array**. We will show how to build this suffix array index and how to perform  $n$ -gram/ $\infty$ -gram inferences with it.

**Suffix array.** The essence of  $n$ -gram and  $\infty$ -gram LMs is counting a given  $n$ -gram in the training data. As such, we leverage the suffix array data structure, which is originally designed for efficiently counting the number of times a given “needle” string (length  $L$ ) appears as substring of a huge “haystack” string (length  $N$ ). When the suffix array is built for a haystack string, counting a given needle string has time complexity  $O(L + \log N)$ .

A suffix array represents the lexicographical ordering of all suffixes of an array (or a string, which is an array of characters). For an array of length  $N$ , the suffix array contains  $N$  unique integers, where the  $i$ -th element is the starting position of the suffix ranked  $i$ -th among all suffixes. [Figure 2](#) (left) shows the suffix array for a toy string, aabaca.

As shown in [Figure 2](#) (right), we build the suffix array on the byte array of the tokenized dataset (i.e., **token array**). Documents are separated by the `\xff\xff` token. In the token array, each consecutive two bytes represent a token ID (assuming that  $|\mathcal{V}| < 2^{16} = 65536$ ). Given that the dataset has  $N$  tokens, the token array has  $2N$  bytes. The suffix array contains  $N$  elements, each pointing to a token in the token array by storing its byte offset. Every token in the token array appears exactly once in the suffix array. Each pointer can be stored with  $\lceil \log_2(2N)/8 \rceil$  bytes. For corpora with 2B to 500B tokens (which is the range we deal with, after sharding ([§A.3](#))), this is 5 bytes per pointer, and thus the suffix array has  $5N$  bytes. Therefore, the combined size of token array and suffix array (i.e., the **infini-gram index**) is  $7N$  bytes.

**Building the suffix array.** Suffix arrays can be built in linear time with respect to the size of the token array ([Kärkkäinen et al., 2006](#)). We adapted from the suffix array implementation in [Lee et al. \(2022\)](#) and further optimized it for efficiency. It took us  $\sim 48$  hours to build the suffix array for RedPajama on a single node with 128 CPUs and 1TiB RAM. We have built the suffix arrays for Dolma (3T tokens), RedPajama (1.4T tokens), Pile (380B tokens), and C4 (200B tokens). Since infini-gram indexes are additive ([§A.2](#)), together these can be easily combined into a larger index with a total of **5 trillion tokens**, and the implicit count table contains at least **2 quadrillion unique  $n$ -grams** ([§A.1](#)).

**Inference with the suffix array.** Computing the  $n$ -gram LM probability involves counting the number of occurrences of a token string, i.e.,  $\text{cnt}(x_1 \dots x_n)$ . By construction, the occurrence positions of strings starting with  $x_1 \dots x_n$  lies in a single, consecutive segment in the suffix array. Thus we only need to find the first and last occurrence positions, and the count would be the difference between them. Beyond counting and  $n$ -gram/ $\infty$ -gram language modeling, infini-gram can also be used to retrieve documents containing an  $n$ -gram, or a CNF expression with multiple  $n$ -grams ([§A.4](#)).

During inference, the entire infini-gram index can stay on-disk, which minimizes the compute resources needed (no GPU, and minimal CPU / RAM). In §A.4, we discuss several optimization techniques applied to the inference engine: parallelized shard processing, hinted search, memory pre-fetching, fast effective- $n$  lookup, and amortized query processing. On RedPajama, our most optimized infini-gram engine can count a given  $n$ -gram with an average latency of less than 20 milliseconds. It can compute the probability and next-token distribution in 40 milliseconds for  $n$ -gram LMs, and in 200 milliseconds for the  $\infty$ -gram. See §A.5 for the full list of supported query types and additional details on latency benchmarking.

## 4 Analyzing Human-written and Machine-generated Text using $\infty$ -gram

In this section, we present some analyses of human-written and machine-generated text from the perspective of  $\infty$ -gram, mostly focusing on the token-wise agreement between  $\infty$ -gram’s prediction and the actual text. In summary, we found that:

1.  $\infty$ -gram has a fairly high accuracy (47%) when predicting the next token given a prefix of a human-written document, and this accuracy is higher when a longer suffix of the prompt can be used (i.e., when the effective  $n$  is larger);
2. Conventional  $n$ -gram LMs ( $n \leq 5$ ) are insufficient for capturing a long enough context to determine the next token, while our  $\infty$ -gram method is highly predictive of human-written and machine-generated text;
3.  $\infty$ -gram has significant potential to complement and improve neural LMs when predicting human-written text (which we further investigate in §5);
4. When plotting the agreement level with respect to the suffix length, text generated by neural LMs with nucleus sampling is most similar to human-written text, among other decoding methods like greedy decoding and temperature sampling. For greedy decoding, the agreement plot suffers from significant fluctuation, which may be rooted in deficiencies in neural LM pretraining and the positional embeddings of Transformers.

**$\infty$ -gram training data.** For analyses in this section, we use a decontaminated version of Pile’s training set (“Pile-train”) (Gao et al., 2020) as training data for the  $\infty$ -gram. We built an infini-gram index on Pile-train using the Llama-2 tokenizer (Touvron et al., 2023b), and yield 360 billion tokens.

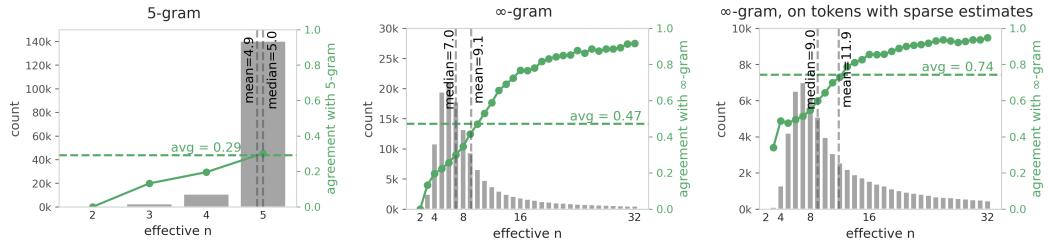
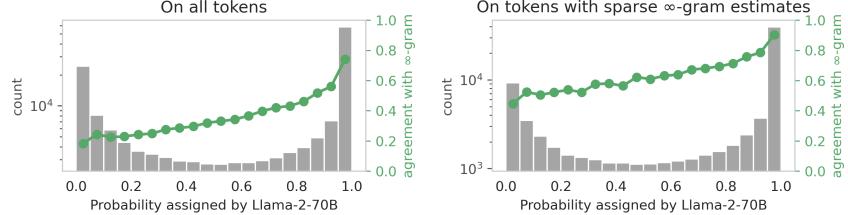
**Decontamination.** It is important that the training data is decontaminated against the evaluation data, because otherwise it is very easy for  $\infty$ -gram to cheat by copying from the very same document in the training data. We run decontamination of Pile-train against its validation and test sets (“Pile-val” and “Pile-test”, which we will use for evaluation below and also in §5), using the method from Groeneveld (2023) that filters out a document if it has excessive  $n$ -gram overlap with the evaluation data. See §B for more details.

Decontamination is non-trivial, and its definition could vary (e.g., when there is an identical sentence, is it contamination, or is it a quote that naturally occurs in real test-time scenarios?) Thus we followed the standard best practices for decontamination.

### 4.1 Human-written text

**Setup.** We use Pile-val set as the human-written text. For this analysis, we sampled 50 documents from each domain of Pile-val, and truncated each document to 1024 tokens (so the total number of tokens per domain is about 50k). We aggregate results from all domains.

We measure the token-wise *agreement* between  $\infty$ -gram’s prediction and the actual human-written text. Since computing the full next-token distribution (or the argmax of it) in  $\infty$ -gram is relatively slow, we compute the  $\infty$ -gram probability of the actual next-token, and deem it as accurate if this probability is higher than 0.5. (This is a lower-bound of argmax accuracy, though the gap is small.) We further categorize all tokens by their effective  $n$ , i.e., one plus the length of their prompt’s longest suffix that has a non-zero count in the training data.

Figure 3: Token-wise agreement between human-written text and  $n$ -gram/ $\infty$ -gram LMs.Figure 4: Distribution of probabilities assigned by neural LMs to human-written text tokens, and  $\infty$ -gram’s agreement with these tokens. **Takeaway:**  $\infty$ -gram and neural LMs are predictive of actual human text on different tokens, and thus  $\infty$ -gram estimates – especially sparse  $\infty$ -gram estimates – can be used to complement neural LMs. See Figure 8 for extended results on Llama-2 13B/7B models.

For each category, we visualize the number of such tokens (in gray bars) as well as the agreement level (in green dots) in the middle plot of Figure 3.

**Results.** Overall,  $\infty$ -gram agrees with the human-written text on 47% of the tokens. We see that  $\infty$ -gram becomes more accurate with the increase of effective  $n$ : when the effective  $n \geq 16$ , agreement is higher than 75%. Further analysis (Appendix Figure 7) shows that the count of this longest suffix in the training data does not affect agreement substantially.

In the left plot of Figure 3, we show the same analysis for a 5-gram LM trained on the same data, and it has much lower agreement than the  $\infty$ -gram. 5-gram LMs, which has been used extensively in previous literature (Franz & Brants, 2006; Aiden & Michel, 2011), does not capture a long enough context to correctly predict the next token: over 90% tokens in the evaluation data has an effective  $n$  of at least 5, and the  $\infty$ -gram analysis shows that the median of effective  $n$  is 7 (and mean is 9.1).

In the right plot of Figure 3, we show the same analysis for only tokens with a sparse  $\infty$ -gram estimate, which covers more than 50% of all tokens. The overall agreement is even higher (75%), and when the effective  $n \geq 14$ , agreement is higher than 80%. This means when the next token is unique according to the training data, that unique token is very likely to be the actual token in human-written text.

Qualitatively, we found that  $\infty$ -gram is often good at completing multi-token words (e.g., hippopotamus, correctly predicted tokens are underlined), common phrases (e.g., born in), and entity names (e.g., educated at Trinity College).  $\infty$ -gram is not very good at recalling factual knowledge (e.g., predicting the first token of an entity name), likely due to insufficient contextualization.

**$\infty$ -gram can shine where neural LMs fail.** In Figure 4, we plot the distribution of probabilities assigned by the Llama-2 70B/13B/7B models (Touvron et al., 2023b) to the actual tokens in human-written text, and the human– $\infty$ -gram agreement for tokens in each probability bucket. (The higher the assigned probability, the higher agreement Llama-2 has with the actual tokens.) We observe a positive, yet imperfect, correlation between neural LMs and  $\infty$ -gram regarding their agreement with the actual text. In particular, when the neural LM performance is very poor (left side of the histogram),  $\infty$ -gram still gives a non-trivial agreement of above 20%; if only considering tokens with sparse  $\infty$ -gram estimates, the agreement is as high as 50%. This indicates a huge potential of complementing and improving the

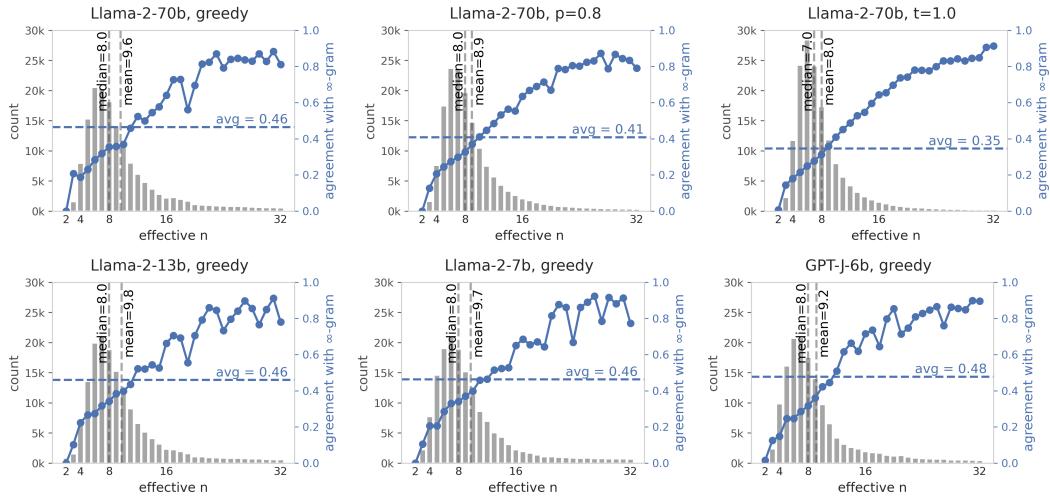


Figure 5: Token-wise agreement between machine-generated text and  $\infty$ -gram. All tokens are considered. See Figure 9 for results on GPT-Neo models.

performance of neural LMs with  $\infty$ -gram for predicting human-written text, which we further investigate in §5.

## 4.2 Machine-generated text

**Setup.** Similar to the analysis with human-written text, we sampled 50 documents from each domain of Pile-val. We use the first 50 tokens of each document to prompt neural LMs to generate a continuation. Generation continues up to the original length of the document, or when an [EOS] token is generated. We experiment with three decoding methods: greedy decoding, temperature sampling, and nucleus sampling (Holtzman et al., 2019). The neural LMs are Llama-2 70B/13B/7B, GPT-J 6B (Wang & Komatsuzaki, 2021), and GPT-Neo 2.7B/1.3B/125M (Gao et al., 2020). The tokenizer of GPT-Neo/J is different from Llama-2, so we built a separate version of infini-gram index for Pile-train based on the GPT-Neo/J tokenizer.

**Impact of decoding method.** The top row of Figure 5 shows results of the three decoding method on the same neural LM – Llama-2 70B. In general, increasing stochasticity shifts the effective  $n$  to the smaller side, and also decreases the agreement level. Nucleus sampling has the most similar distribution of effective  $n$  compared to human-written text (Figure 3, middle plot), which is probably why nucleus sampling is usually preferred in text generation. Greedy decoding has even higher effective  $n$  than human-written text, which implies that greedy decoding could lead to over-memorization of training data as well as lack of diversity.

**Impact of model size.** The bottom row of Figure 5 shows the same analysis for different sizes of neural LM under greedy decoding. In general, increasing model size slightly increases the effective  $n$ , and also increases the agreement level. This indicates that larger models memorizes more from the training data, and are also more inclined to copy verbatim. The agreement level of GPT-Neo/J models is higher than Llama-2 models, probably because GPT-Neo/J are trained on the same data as the  $\infty$ -gram (i.e., Pile-train). Overall, text generated by these neural LMs has similar agreement level with  $\infty$ -gram as human text.

One very curious phenomenon is that, as effective  $n$  increases, the agreement level fluctuates greatly in greedy decoding (but not nucleus or temperature sampling, where agreement level almost increases monotonically). Such fluctuation is even more rapid for smaller models (Llama-2 13B/7B and GPT-Neo/J models), and for Llama-2 7B the fluctuation is even periodic (rapidly dropping at effective  $n = 20, 24, 28, 32$ ; this is statistically significant, a two-proportion z-test gives a p-value of  $< 10^{-99}$ ). We suspect that this may be caused by the application of positional embeddings when pretraining these Transformer-based models, and we welcome further investigation from the community.

Neural LM	Size	Reference Data	Validation		Test	
			Neural	+ $\infty$ -gram	Neural	+ $\infty$ -gram
GPT-2	117M	Pile-train	22.82	<b>13.71</b> (42%)	22.86	<b>13.58</b> (42%)
GPT-2	345M	Pile-train	16.45	<b>11.22</b> (34%)	16.69	<b>11.18</b> (35%)
GPT-2	774M	Pile-train	15.35	<b>10.39</b> (35%)	15.40	<b>10.33</b> (35%)
GPT-2	1.6B	Pile-train	14.42	<b>9.93</b> (33%)	14.61	<b>9.93</b> (34%)
GPT-Neo	125M	Pile-train	13.50	<b>10.76</b> (22%)	14.08	<b>10.79</b> (25%)
GPT-Neo	1.3B	Pile-train	8.29	<b>7.31</b> (13%)	8.61	<b>7.36</b> (16%)
GPT-Neo	2.7B	Pile-train	7.46	<b>6.69</b> (12%)	7.77	<b>6.76</b> (15%)
GPT-J	6.7B	Pile-train	6.25	<b>5.75</b> (10%)	6.51	<b>5.85</b> (12%)
Llama-2	7B	Pile-train	5.69	<b>5.05</b> (14%)	5.83	<b>5.06</b> (16%)
Llama-2	13B	Pile-train	5.30	<b>4.75</b> (13%)	5.43	<b>4.76</b> (15%)
Llama-2	70B	Pile-train	4.59	<b>4.21</b> (11%)	4.65	<b>4.20</b> (12%)
Llama-2	7B	Pile-train + RedPajama	5.69	<b>4.66</b> (22%)	5.83	<b>4.66</b> (24%)
Llama-2	13B	Pile-train + RedPajama	5.30	<b>4.41</b> (21%)	5.43	<b>4.42</b> (23%)
Llama-2	70B	Pile-train + RedPajama	4.59	<b>3.96</b> (18%)	4.65	<b>3.95</b> (19%)

Table 1: Perplexity (lower is better) on Pile’s validation and test sets. Numbers in parentheses are the relative perplexity improvement. The first eight rows share the same tokenizer, and the last six rows share the same tokenizer.

## 5 Improving Neural LMs with the $\infty$ -gram

The results in §4 motivate us to combine neural LMs and  $\infty$ -gram (§2) to yield better language models. In this section, we will show strong experimental results of the combined model. In §4 we found that the  $\infty$ -gram estimate has higher agreement with human-written text when it is *sparse*. Therefore, we use two separate interpolation hyperparameters:  $\lambda_1$  for sparse and  $\lambda_2$  for non-sparse  $\infty$ -gram estimates. These hyperparameters are tuned on the validation set to minimize the perplexity of the combined model.

### 5.1 Experimental setup

**Evaluation and metric.** We measure the perplexity of each model on Pile’s validation and test sets, as well as the relative improvement of perplexity between models. To show generalization, we also evaluate on time-shifted data (i.e., data created after the cutoff date of the  $\infty$ -gram training data). The relative improvement of model  $M$  against model  $M_o$  is defined as  $(1 - \frac{\text{PPL}(M)-1}{\text{PPL}(M_o)-1}) \times 100\%$ , which is the percentage of perplexity gap closed towards perfect language modeling (i.e.,  $\text{PPL} = 1$ ). Additional details on evaluation data processing in §D.1.

**Reference data.** To reduce confusion, in this section we will use *reference data* to refer to the training data of the  $\infty$ -gram. In addition to Pile’s training set that we used in the previous analyses (§4), we also consider RedPajama (Together, 2023) as reference data. The decontaminated Pile-train and Redpajama have 360 billion and 1.4 trillion tokens, respectively, summing up to 1.8 trillion tokens (based on the Llama-2 tokenizer). We later perform ablations on varying sizes and domains of the reference data.

**Neural LMs.** We use a range of large, competitive neural LMs, both as baselines and as models to interpolate with the  $\infty$ -gram. In total, 14 models are considered: GPT-2 117M/345M/774M/1.6B (Radford et al., 2019), GPT-Neo 125M/1.3B/2.7B, GPT-J-6B, Llama-2 7B/13B/70B, and SILO PD/PDSW/PDSWBY (Min et al., 2023a). See §D.1 for additional details about these models and their training data.

**Tokenizers.** Among these models, GPT-2, GPT-Neo and GPT-J share the same tokenizer, but Llama-2 and SILO use different tokenizers. We therefore built three versions of the infini-gram index on Pile-train and RedPajama, one for each tokenizer. Due to the tokenizer variation, the perplexity of GPT-2, GPT-Neo and GPT-J are comparable to each other, but perplexity of Llama-2 and SILO are not comparable to them nor to each other.

Neural LM	Validation				Test			
	Neural	+ $\infty$ -gram	+ kNN-LM <sup>†</sup>	+ RIC-LM <sup>†</sup>	Neural	+ $\infty$ -gram	+ kNN-LM <sup>†</sup>	+ RIC-LM <sup>†</sup>
<i>Eval data: Wikipedia</i>								
Silo PD	26.60	<b>15.30</b> (43%)	20.62	27.91	28.42	<b>14.44</b> (51%)	—	—
Silo PDSW	18.93	<b>12.36</b> (36%)	14.10	18.90	20.02	<b>11.84</b> (43%)	14.5	19.4
Silo PDSWBY	10.66	<b>8.77</b> (19%)	10.14	10.87	10.76	<b>8.41</b> (24%)	—	—
Pythia	9.00	—	8.50	8.84	9.1	—	—	—
<i>Eval data: Enron Emails</i>								
Silo PD	19.56	<b>6.31</b> (70%)	8.56	15.45	15.71	<b>4.85</b> (73%)	—	—
Silo PDSW	14.66	<b>5.58</b> (65%)	6.70	10.80	11.23	<b>4.35</b> (66%)	5.9	9.9
Silo PDSWBY	14.67	<b>5.61</b> (65%)	7.24	10.91	11.52	<b>4.44</b> (66%)	—	—
Pythia	7.577	—	4.99	6.16	6.9	—	—	—
<i>Eval data: NIH ExPorters</i>								
Silo PD	27.46	<b>16.26</b> (41%)	19.27	25.51	27.94	<b>16.00</b> (44%)	—	—
Silo PDSW	19.35	<b>12.70</b> (35%)	14.95	18.35	19.12	<b>12.39</b> (37%)	15.0	18.5
Silo PDSWBY	15.01	<b>10.62</b> (30%)	12.33	14.29	14.81	<b>10.33</b> (32%)	—	—
Pythia	11.20	—	11.20	10.83	11.1	—	—	—

Table 2: Perplexity (the lower the better) on the validation and the test datasets of the Wikipedia, Enron Emails, and NIH ExPorters of the Pile. All neural models are 1.3B models, and the reference data is always the Pile. ■ indicates in-domain; □ indicates out-of-domain; ▨ indicates out-of-domain but has relevant data in-domain, all with respect to the training data of the neural LM. †: Results retrieved from Min et al. (2023a), which use much smaller reference data: 45-million to 1.2-billion tokens, compared to our 360-billion tokens.

## 5.2 Results

Experimental results with GPT-2, GPT-Neo/J, and Llama-2 on Pile’s evaluation sets are shown in Table 1. Results on time-shifted data can be found in §D.2, and we show the impact of the size and domain of reference data in §D.3.

Interpolating with  $\infty$ -gram greatly and consistently improves the perplexity of neural LMs. The magnitude of improvement trends smaller as the neural LM size grows within the same family, while the largest models can still benefit a lot from our method (e.g., Pile-train alone improves Llama-2 70B by 12%).

However, this trend does not hold across different families of LMs. For example,  $\infty$ -gram can improve GPT-2 1.6B by 34%, but only improves a smaller model, GPT-Neo 1.3B, by 16%. This may be because GPT-Neo/J models are trained precisely on Pile, while GPT-2 models are not.  $\infty$ -gram works better when the reference data distribution differs from, or complements, the pretraining data distribution, which emphasizes the importance of data diversity. Meanwhile, the fact that  $\infty$ -gram also improves neural LMs already pretrained on its reference data shows that there is consistent advantage in introducing  $\infty$ -gram.

On the choice of  $\infty$ -gram reference data, the union of Pile-train and RedPajama yields larger improvements on the Llama-2 models than Pile-train alone. The combination of Llama-2 13B and  $\infty$ -gram with Pile-train + RedPajama outperforms Llama-2 70B, and interpolating with  $\infty$ -gram pushes the perplexity of Llama-2 70B below 4.0.

When the neural LM is SILO (which is trained on permissive-licensed data only and thus has less training data), adding the  $\infty$ -gram component is more helpful when SILO is trained on more restrictive data (i.e., PD > PDSW > PDSWBY). The usage of  $\infty$ -gram can be precisely traced back to the contributing document(s) in the reference data, which is in-line with the philosophy of SILO: to allow crediting the source data when using them for language modeling. When compared to the existing retrieval-augmentation methods used by SILO, i.e.,  $k$ NN-LM and RIC-LM,  $\infty$ -gram yields better improvement in perplexity. Therefore,  $\infty$ -gram can serve as a better alternative as the retrieval-augmentation method for SILO.

**A note on text generation.** While  $\infty$ -gram can be interpolated with neural LMs and greatly improve their perplexity, our preliminary experiments show that such method might not be helpful, and even harmful, to open-ended text generation tasks. During generation,

$\infty$ -gram can make odd mistakes (e.g., predicting totally irrelevant tokens) which makes the model to digress. Thus this combined model is *not* ready to replace neural LMs. Additional investigation is required to make  $\infty$ -gram best contribute to text generation.

## 6 Related Work

We discuss closely related work here. See §E for extended discussion, and [Table 6](#) for comparison with other  $n$ -gram models and nonparametric language models.

**$n$ -gram language models.**  $n$ -gram has been one of the most classical language modeling methods since the inception of natural language processing ([Jurafsky & Martin, 2000](#)). People have been pushing the limits of  $n$ -gram LMs by scaling up its training data. To date, the largest  $n$ -gram table ([Brants et al., 2007](#)) counts 5-grams in a corpus of 2 trillion tokens.

While  $n$ -gram LMs are currently largely surpassed by neural LMs, there has been recent work that revisit  $n$ -grams and  $n$ -gram LMs. [Mikolov & Zweig \(2012\)](#) finds that interpolating with a 5-gram Kneser-Ney model improves the perplexity of RNN models, whereas [Khandelwal et al. \(2020\)](#) finds that interpolating  $n$ -gram models with Transformers does not improve perplexity substantially. [Li et al. \(2022\)](#) finds that the  $n$ -gram model is as competitive as a small neural LM, and training a neural model to be complementary to the  $n$ -gram model and using both at inference time outperforms the neural-only LM. However, both use limited reference data (101M tokens) and compare with small neural LMs (117–250M parameters). Some prior work has found value in scaling up the  $n$ -gram training data ([Allamanis & Sutton, 2013](#)).

Our work scales up the training data of  $n$ -gram LMs to trillions of tokens. With the addition of scaling up the value of  $n$  in the  $n$ -gram, our model can significantly improve state-of-the-art neural models as large as 70B.

**Unbounded  $n$ -grams, suffix arrays, suffix trees.** Previous work has explored using suffix-based data structures to enable  $n$ -gram queries with unbounded  $n$ , with limited scale of the training data. [Stehouwer & van Zaanen \(2010\)](#) proposes to use suffix arrays for  $\infty$ -gram, and yet their formulation does *not* yield proper probability distributions and, consequently, a language model. [Kennington et al. \(2012\)](#) proposes to use suffix trees for the same purpose, and yet the storage overhead of suffix trees is very high such that it hinders scaling, which may be mitigated with highly intricate compression techniques ([Shareghi et al., 2015](#)). Among the three aforementioned papers, only the third evaluates on the general language modeling task, and the perplexity numbers are too high to be practically useful. Our training data is 500x larger than the largest one in these previous work.

**Nonparametric language models.** Nonparametric LMs refer to LMs whose complexity is not bounded a priori, because the complexity can change according to the reference data ([Khandelwal et al., 2020; Borgeaud et al., 2022; Asai et al., 2023](#)). The  $\infty$ -gram LM is one instance of nonparametric LMs, and its simplicity makes it possible to significantly scale the reference data with modest resources (§3). To the best of our knowledge, our  $\infty$ -gram LM is the largest in both the size of the reference data (5 trillion tokens) and the size of the base neural LM (70B).

## 7 Conclusion

In this paper, we modernized the classical  $n$ -gram language model by scaling it up to a trillion tokens and extending to unbounded  $n$ . We presented the infini-gram engine that performs efficient training and inference under this extreme setup. We also proposed the  $\infty$ -gram language model, powered by the infini-gram engine, and showed that it can offer novel insights into human-written and machine-generated text and can improve existing neural language models.

## Acknowledgments

We would like to thank Zexuan Zhong, Mike Lewis, Yanai Elazar, Will Merrill, Tim Dettmers, Ximing Lu, Alisa Liu, Weijia Shi, Xiaochuang Han, members of the H2lab, and Ziqi Ma for their invaluable feedback.

## References

- Erez Lieberman Aiden and Jean-Baptiste Michel. Quantitative analysis of culture using millions of digitized books. *Science*, 331:176 – 182, 2011. URL <https://api.semanticscholar.org/CorpusID:40104730>.
- Miltiadis Allamanis and Charles Sutton. Mining source code repositories at massive scale using language modeling. *2013 10th Working Conference on Mining Software Repositories (MSR)*, pp. 207–216, 2013. URL <https://api.semanticscholar.org/CorpusID:1857729>.
- Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. Acl 2023 tutorial: Retrieval-based language models and applications. *ACL 2023*, 2023.
- Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke Zettlemoyer, Hanna Hajishirzi, and Wen tau Yih. Reliable, adaptable, and attributable language models with retrieval. 2024. URL <https://api.semanticscholar.org/CorpusID:268248911>.
- Alexei Baevski and Michael Auli. Adaptive input representations for neural language modeling. In *Proceedings of the International Conference on Learning Representations*, 2019.
- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *Proceedings of the International Conference of Machine Learning*, 2022.
- T. Brants, Ashok Popat, Peng Xu, Franz Josef Och, and Jeffrey Dean. Large language models in machine translation. In *Conference on Empirical Methods in Natural Language Processing*, 2007. URL <https://api.semanticscholar.org/CorpusID:633992>.
- Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, L. Sifre, and John M. Jumper. Accelerating large language model decoding with speculative sampling. *ArXiv*, abs/2302.01318, 2023. URL <https://api.semanticscholar.org/CorpusID:256503945>.
- Jesse Dodge, Ana Marasovic, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Conference on Empirical Methods in Natural Language Processing*, 2021. URL <https://api.semanticscholar.org/CorpusID:237568724>.
- Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hanna Hajishirzi, Noah A. Smith, and Jesse Dodge. What’s in my big data? *ArXiv*, abs/2310.20707, 2023. URL <https://api.semanticscholar.org/CorpusID:264803575>.
- Alex Franz and Thorsten Brants. All our n-gram are belong to you. *Google Machine Translation Team*, 20, 2006. URL <https://blog.research.google/2006/08/all-our-n-gram-are-belong-to-you.html>.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.
- Xinyang Geng and Hao Liu. Openllama: An open reproduction of llama, May 2023. URL [https://github.com/openlm-research/open\\_llama](https://github.com/openlm-research/open_llama).
- Dirk Groeneveld. The big friendly filter. <https://github.com/allenai/bff>, 2023.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Arthur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muenninghoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. Olmo: Accelerating the science of language models. *Preprint*, 2024.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *Proceedings of the International Conference of Machine Learning*, 2020.

Zhenyu He, Zexuan Zhong, Tianle Cai, Jason D Lee, and Di He. Rest: Retrieval-based speculative decoding. 2023. URL <https://api.semanticscholar.org/CorpusID:265157884>.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *ArXiv*, abs/1904.09751, 2019. URL <https://api.semanticscholar.org/CorpusID:127986954>.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 2022.

Dan Jurafsky and James H. Martin. Speech and language processing - an introduction to natural language processing, computational linguistics, and speech recognition. In *Prentice Hall series in artificial intelligence*, 2000. URL <https://api.semanticscholar.org/CorpusID:60691216>.

Juha Kärkkäinen, Peter Sanders, and Stefan Burkhardt. Linear work suffix array construction. *J. ACM*, 53:918–936, 2006. URL <https://api.semanticscholar.org/CorpusID:12825385>.

Slava M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Trans. Acoust. Speech Signal Process.*, 35:400–401, 1987. URL <https://api.semanticscholar.org/CorpusID:6555412>.

Casey Redd Kennington, Martin Kay, and Annemarie Friedrich. Suffix trees as language models. In *International Conference on Language Resources and Evaluation*, 2012. URL <https://api.semanticscholar.org/CorpusID:12071964>.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *Proceedings of the International Conference on Learning Representations*, 2020.

Tian Lan, Deng Cai, Yan Wang, Heyan Huang, and Xian-Ling Mao. Copy is all you need. In *Proceedings of the International Conference on Learning Representations*, 2023.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the Association for Computational Linguistics*, 2022.

Huayang Li, Deng Cai, Jin Xu, and Taro Watanabe. Residual learning of neural text generation with n-gram language model. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022. URL <https://aclanthology.org/2022.findings-emnlp.109>.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Annual Meeting of the Association for Computational Linguistics*, 2022. URL <https://api.semanticscholar.org/CorpusID:254877603>.

- Marc Marone and Benjamin Van Durme. Data portraits: Recording foundation model training data. *ArXiv*, abs/2303.03919, 2023. URL <https://api.semanticscholar.org/CorpusID:257378087>.
- Tomas Mikolov and Geoffrey Zweig. Context dependent recurrent neural network language model. 2012 *IEEE Spoken Language Technology Workshop (SLT)*, pp. 234–239, 2012. URL <https://api.semanticscholar.org/CorpusID:11383176>.
- Sewon Min, Suchin Gururangan, Eric Wallace, Hannaneh Hajishirzi, Noah Smith, and Luke Zettlemoyer. SILO language models: Isolating legal risk in a nonparametric datastore. *arXiv preprint arXiv:2308.04430*, 2023a. URL <https://arxiv.org/abs/2308.04430>.
- Sewon Min, Weijia Shi, Mike Lewis, Xilun Chen, Wen-tau Yih, Hannaneh Hajishirzi, and Luke Zettlemoyer. Nonparametric masked language modeling. In *Findings of ACL*, 2023b.
- Daichi Mochihashi and Eiichiro Sumita. The infinite markov model. In *Neural Information Processing Systems*, 2007. URL <https://api.semanticscholar.org/CorpusID:1279894>.
- Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language models as knowledge bases? *ArXiv*, abs/1909.01066, 2019. URL <https://api.semanticscholar.org/CorpusID:202539551>.
- Aleksandra Piktus, Christopher Akiki, Paulo Villegas, Hugo Laurencon, Gérard Dupont, Alexandra Sasha Luccioni, Yacine Jernite, and Anna Rogers. The roots search tool: Data transparency for llms. In *Annual Meeting of the Association for Computational Linguistics*, 2023. URL <https://api.semanticscholar.org/CorpusID:257219882>.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Edward Raff, William Fleming, Richard Zak, H. Anderson, Bill Finlayson, Charles K. Nicholas, and Mark McLean. Kilograms: Very large n-grams for malware classification. *ArXiv*, abs/1908.00200, 2019. URL <https://api.semanticscholar.org/CorpusID:199064443>.
- Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2019. URL <https://api.semanticscholar.org/CorpusID:204838007>.
- Ehsan Shareghi, Matthias Petri, Gholamreza Haffari, and Trevor Cohn. Compact, efficient and unlimited capacity: Language modeling with compressed suffix trees. In *Conference on Empirical Methods in Natural Language Processing*, 2015. URL <https://api.semanticscholar.org/CorpusID:225428>.
- Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*, 2023.
- Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Arthur, Khyathi Chandu, Jennifer Dumas, Li Lucy, Xinxi Lyu, Ian Magnusson, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Pete Walsh, Hannaneh Hajishirzi, Noah A. Smith, Luke Zettlemoyer, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: An Open Corpus of 3 Trillion Tokens for Language Model Pretraining Research. Technical report, Allen Institute for AI, 2023. Released under ImpACT License as Medium Risk artifact, <https://github.com/allenai/dolma>.
- Herman Stehouwer and Menno van Zaanen. Using suffix arrays as language models: Scaling the n-gram. 2010. URL <https://api.semanticscholar.org/CorpusID:18379946>.
- Together. RedPajama: An open source recipe to reproduce LLaMA training dataset, 2023. URL <https://github.com/togethercomputer/RedPajama-Data>.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Thuy-Trang Vu, Xuanli He, Gholamreza Haffari, and Ehsan Shareghi. Koala: An index for quantifying overlaps with pre-training corpora. In *Conference on Empirical Methods in Natural Language Processing*, 2023. URL <https://api.semanticscholar.org/CorpusID:257766452>.

Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. <https://github.com/kingoflolz/mesh-transformer-jax>, May 2021.

Frank D. Wood, C. Archambeau, Jan Gasthaus, Lancelot F. James, and Yee Whye Teh. A stochastic memoizer for sequence data. In *International Conference on Machine Learning*, 2009. URL <https://api.semanticscholar.org/CorpusID:11199892>.

Zexuan Zhong, Tao Lei, and Danqi Chen. Training language models with memory augmentation. In *Proceedings of Empirical Methods in Natural Language Processing*, 2022.

## A Additional Details on the Infini-gram Engine

### A.1 What is the size of the $n$ -gram count table implied by an infini-gram index?

The size of an  $n$ -gram LM is often measured by the number of unique  $n$ -grams indexed, each associated with its count in the dataset. This size is easy to obtain when indexing is done as the classical  $n$ -gram count tables, but non-trivial to compute for infini-gram.

If we consider all possible values of  $n$  in the  $n$ -gram, then a dataset with  $N$  tokens would contain about  $\frac{1}{2}N^2$   $n$ -grams, and since most of these  $n$ -grams are long and thus very likely to be distinct, there would be about  $\frac{1}{2}N^2$  unique  $n$ -grams. Since we index  $N = 5$  trillion tokens, the size of the  $n$ -gram count table implied by our infini-gram index would be approximately  $1.2 \times 10^{25}$ , or equivalently, 20 mol (using  $N_A = 6.02 \times 10^{23}/\text{mol}$ ).

However, the document separator is meaningless and should not be part of the  $n$ -grams, and we should probably only count  $n$ -grams within the document boundaries. There are  $N = 5 \times 10^{12}$  tokens and  $D = 6 \times 10^9$  documents, and on average each documents have 857 tokens. Using the Cauchy-Schwarz inequality, we have the total number of  $n$ -grams as

$$\sum_d \frac{1}{2}N_d^2 \geq \frac{1}{2}D \cdot \left(\frac{N}{D}\right)^2 = \frac{N^2}{2D} = 2 \times 10^{15}$$

Therefore, there are at least 2 quadrillion unique  $n$ -grams in the count table implied by infini-gram.

### A.2 Additional details on the infini-gram index

**Infini-gram indexes are additive and subtractive.** If we have two or more indexes built on disjoint datasets (with the same tokenizer), we can easily combine them into a single index by adding up the  $n$ -gram counts from each index. This feature is useful in the sharding

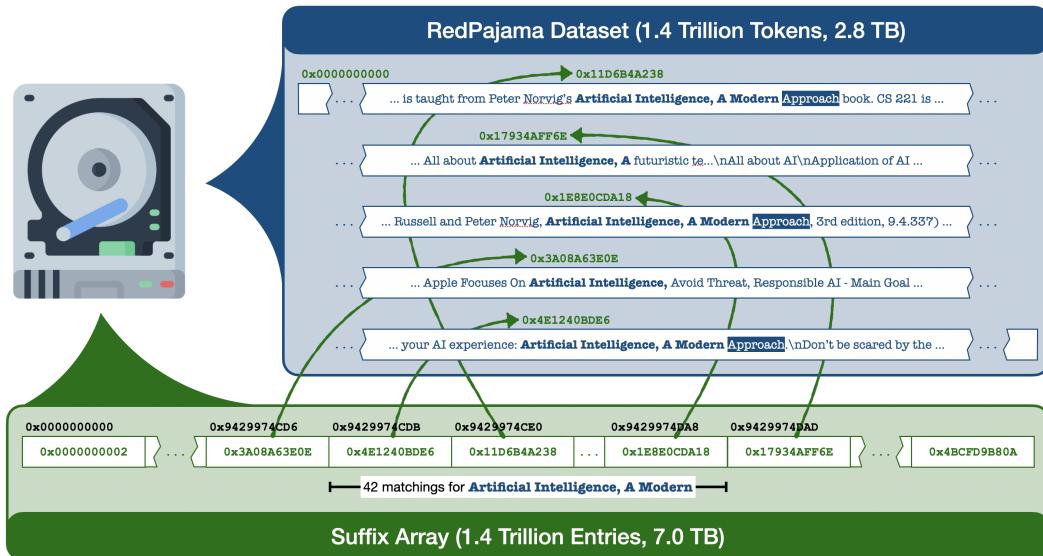


Figure 6:  $n$ -gram/ $\infty$ -gram queries on a training data are supported by an associated *suffix array*. Both the training data and the suffix array are stored on-disk as regular files. Contents on the white strips are file data, and addresses above the strips are byte offsets. Querying for a particular  $n$ -gram returns a consecutive segment of the suffix array, where each element is a pointer into the training data where the  $n$ -gram appears. E.g., in the trillion-token training data, *Artificial Intelligence, A Modern* appears 42 times, and in all cases the following token is *Approach*.

technique that we discuss in §A.3 and §A.4. Similarly, if we have built indexes on a big dataset and a subset of it, we can easily obtain an index of their difference by taking the difference of  $n$ -gram counts. Compared to having a single index, both operations incur some additional inference operations (which can be parallelized to mitigate latency overhead), but they would spare us from re-indexing the union or difference sets from scratch.

**Document offsets and metadata.** To enable efficient document retrieval, the infini-gram index stores additional data about documents. A *document offset* file stores the byte offset of each document in the tokenized dataset, and its format is similar to the suffix array. A *document metadata* file stores a comma-separated string for each document that contains its metadata (e.g., document ID, source, URL), and a *document metadata offset* file stores the byte offset of each document’s metadata in the document metadata file. All the above files are negligible in size compared to the suffix array, because there are far less documents than the total number of tokens.

### A.3 Additional details on building the suffix array

**Sharding.** Building the suffix array requires heavy random access to the byte array, and thus the entire byte array must be kept in RAM so that the building time is reasonable. However, the byte array may be too large to fit into RAM. In such cases, we shard the byte array into multiple shards, and build a suffix array for each shard. Sharding would induce additional inference latency, which we discuss and mitigate below (§A.4).

### A.4 Additional details on inference with the infini-gram index

Both the first and last occurrence positions can be found with binary search, with time complexity  $O(n \cdot \log N)$  and  $O(\log N)$  random array accesses. The two binary searches can be parallelized, reducing the latency by roughly 2x. The impact of query length  $n$  is negligible, because computers usually fetch memory in pages of 4K bytes, and string comparison is much faster than page fetching. Therefore, when we analyze time complexity below, we refer to the number of random array accesses.

**Finding occurrence positions and documents.**  $n$ -gram counting with suffix arrays has a by-product: we also get to know all positions where the  $n$ -gram appears in the training data, for free. This position information is implicitly contained in the suffix array segment we obtained during counting, and to retrieve the original documents where the  $n$ -gram appears, all we need to do is to follow each pointer within this segment back into the tokenized dataset, and find the starting and ending position of the enclosing document by performing a binary search on the *document offset* index. (Note that if we don’t have the *document offset* index, the latency of document search cannot be bounded because we would need to expand the pointer in both directions in the tokenized dataset until hitting the document separator. In practice, we see documents as large as 20M tokens.)

**Impact of sharding.** When the suffix arrays are built on sharded byte arrays, we can simply perform counting on each individual shard and accumulate the counts across all shards. The latency is proportional to the number of shards: time complexity would become  $O(S \cdot \log N)$ . The processing of different shards can be parallelized, reducing the time complexity back to  $O(\log N)$ .

**Speeding up  $n$ -gram computation by re-using previous search results.** On the suffix array, the segment for  $x_1 \dots x_n$  must be a sub-segment of that for  $x_1 \dots x_{n-1}$ . Therefore, when computing the  $n$ -gram probability  $P_n(x_n \mid x_1 \dots x_{n-1})$ , we can first count  $x_1 \dots x_{n-1}$ , and then when counting  $x_1 \dots x_n$ , we only need to search for the first and last occurrence positions within the segment of  $x_1 \dots x_n$ , which reduces the latency by at most 2x.

**On-disk search.** The byte array and suffix array may be too large to fit into RAM, so in practice, we keep them on disk and read them as memory-mapped files. However, this creates a significant latency as the binary search requires random access to the byte array and suffix array. To mitigate this, we implemented a memory **pre-fetching** method that informs the system of the array offsets we will likely be reading in the near future. Pre-fetching reduces average latency by roughly 5x.

Query Type (↓)	Reference Data (→)	Pile-train $N = 0.36T$ $S = 2$	RPJ $N = 1.4T$ $S = 8$	Time Complexity (measured by number of random disk accesses)
1. Counting an $n$ -gram				$O(\log N)$
... ( $n = 1$ )		7 ms	9 ms	
... ( $n = 2$ )		13 ms	20 ms	
... ( $n = 5$ )		14 ms	19 ms	
... ( $n = 10$ )		13 ms	18 ms	
... ( $n = 100$ )		13 ms	19 ms	
... ( $n = 1000$ )		14 ms	19 ms	
2. Computing a token probability from $n$ -gram LM ( $n = 5$ )		19 ms	30 ms	$O(\log N)$
3. Computing full next-token distribution from $n$ -gram LM ( $n = 5$ )		31 ms	39 ms	$O(V \cdot \log N)$
4. Computing a token probability from $\infty$ -gram LM		90 ms	135 ms	$O(\log L \cdot \log N)$
... on consecutive tokens		12 ms	20 ms	$O(\log N)$
5. Computing full next-token distribution from $\infty$ -gram LM		88 ms	180 ms	$O((\log L + V) \cdot \log N)$

Table 3: Inference-time latency of infini-gram on different types of queries. Average latency per query is reported. Benchmarked with inference engine written in C++ (with parallelized shard processing) and running on a single, 8-core CPU node. Notations for time complexity:  $N$  = number of tokens in the reference data;  $S$  = number of shards for the suffix array;  $L$  = number of tokens in the query document;  $V$  = vocabulary size.

**Speeding up  $\infty$ -gram computation.** To compute the  $\infty$ -gram probability, we need to count the occurrence of each suffix  $x_{l-n+1} \dots x_l$  up to the maximum  $n$  so that the suffix still meets the sufficient appearance requirement (we denote this maximum  $n$  as  $L$ ). This means  $O(L)$  counting operations, and the time complexity for each  $\infty$ -gram computation is  $O(L \cdot \log N)$ . However, a simple binary-lifting + binary-search algorithm for searching  $L$  can reduce the number of counting operations to  $O(\log L)$ , and thus the time complexity for each  $\infty$ -gram computation becomes  $O(\log L \cdot \log N)$ .

**Speeding up dense  $\infty$ -gram computation.** During evaluation, we need to compute the  $\infty$ -gram probability of each token in the test document. We can save computation by observing that the effective  $n$  for one token is at most one token longer than that for the previous token. This brings the amortized time complexity for evaluating each token down to  $O(\log N)$ .

### A.5 Supported query types and latency benchmarking

Infini-gram supports the following types of  $n$ -gram/ $\infty$ -gram queries:

1. Counting an  $n$ -gram (COUNT);
2. Computing a token probability from  $n$ -gram LM (with given  $n$ , no backoff) (NGRAMPROB);
3. Computing the full next-token distribution from  $n$ -gram LM (NGRAMDIST);
4. Computing a token probability from  $\infty$ -gram LM (INFGRAMPROB);
5. Computing the full next-token distribution from  $\infty$ -gram LM (INFGRAMDIST);
6. Returning documents containing an  $n$ -gram, or a CNF logical expression of  $n$ -gram terms, connected with AND’s and/or OR’s (e.g., (natural language processing OR artificial intelligence) AND (deep learning OR machine learning)) (SEARCHDOC).

We benchmark the latency of infini-gram on different types of  $n$ -gram and  $\infty$ -gram queries, and show results in Table 3. During inference, the training data and the suffix array are stored on an SSD. For each type of query, the benchmarking is conducted on 1,000 tokens randomly and independently sampled from Pile’s validation data (except for the task “computing a token probability from  $\infty$ -gram LM on consecutive tokens”, where we sampled 10 documents and processed 1000 consecutive tokens in each document).

All types of queries demonstrate sub-second latency on the trillion-token training data. Computing a token probability from the  $\infty$ -gram with RedPajama takes merely 135 milliseconds. Furthermore, our implementation supports counting the occurrence of an  $n$ -gram

with arbitrarily large  $n$ , with roughly constant latency at 20 milliseconds (we experimentally validated up to  $n = 1000$ ). Decoding requires computing the full next-token distribution and is thus slightly slower: 39 milliseconds per token with  $n$ -gram LMs and 180 milliseconds per token with  $\infty$ -gram.

## B Decontamination of Reference Data

To properly evaluate the effectiveness of  $\infty$ -gram LM on Pile’s evaluation sets, we performed data decontamination on the Pile’s training set and RedPajama before using them as reference data for the  $\infty$ -gram LM. We run the Big Friendly Filter (BFF)<sup>1</sup> (Groeneveld, 2023) on Pile’s training set and RedPajama, filtering out documents with too much  $n$ -gram overlap with Pile’s evaluation sets. Table 4 reports the statistics of decontamination.

When using BFF, we always remove whole documents, instead of by paragraphs. Following the default settings, we consider  $n$ -grams where  $n = 13$ , and discard the document if at least 80% of its  $n$ -grams are present in the evaluation set. For Pile’s training set, we lowercase all documents to capture more potential contaminations.

PILE (TRAIN)			
Subset	Total docs	Filtered docs	Ratio filtered
Arxiv	2377741	1089	
BookCorpus2	25355	6	
Books3	277655	99	
DM Mathematics	1918535	0	0%
Enron Emails	926132	18236	2%
EuroParl	131723	21	
FreeLaw	5069088	11821	0.2%
Github	18044218	961726	5.3%
Gutenberg (PG-19)	66981	70	0.1%
HackerNews	1571968	14	
NIH ExPorter	1777926	3739	0.2%
OpenSubtitles	632485	5754	0.9%
OpenWebText2	32333654	136914	0.4%
PhilPapers	63875	2324	0.4%
Pile-CC	52441354	19928	
PubMed Abstracts	29329202	2312	
PubMed Central	5679903	4230	0.1%
StackExchange	29529008	2072	
USPTO Backgrounds	11123325	80088	0.7%
Ubuntu IRC	20067	10	
Wikipedia (en)	16939503	45052	0.3%
YoutubeSubtitles	328030	871	0.3%
<b>Total</b>	<b>210607728</b>	<b>1296376</b>	<b>0.6%</b>

REDPAJAMA			
Subset	Total docs	Filtered docs	Ratio filtered
arxiv	1558306	213	0.01%
book	205744	711	0.3%
c4	364868892	53195	0.01%
common_crawl	476276019	0	0%
github	28793312	614259	2%
stackexchange	29825086	40086	0.01%
wikipedia	29834171	21973	0.07%
<b>Total</b>	<b>931361530</b>	<b>730437</b>	<b>0.08%</b>

Table 4: Statistics of de-contamination in RedPajama (left) and Pile’s training set (right).

## C Additional Analysis

Figure 7 is an extension to the middle plot of Figure 3, and shows a more fine-grained analysis of the token-wise agreement between human-written text and  $\infty$ -gram. Figure 8 extends Figure 4 with results on Llama-2-13b/7b. Figure 9 extends Figure 5 with results on GPT-Neo models.

## D Additional Experiments on Improving Neural LMs with $\infty$ -gram

### D.1 Additional details on experimental setup

**Evaluation data processing.** We split each document in the evaluation data into batches with a maximum sequence length of 1024 and a sliding window of 512, a setup that is standard in prior language modeling literature (Baevski & Auli, 2019; Khandelwal et al., 2020).

<sup>1</sup><https://github.com/allenai/bff>

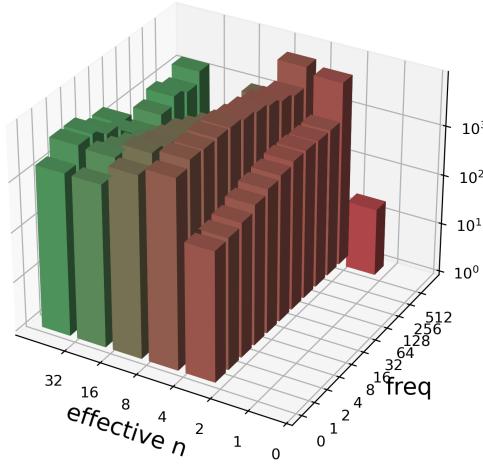


Figure 7: Token-wise agreement between human-generated text and  $\infty$ -gram, broken down by “effective  $n$ ” and frequency of the corresponding longest suffix in the reference data. The height of each bar represents token count, and the color represents agreement (red is 0.0, green is 1.0).

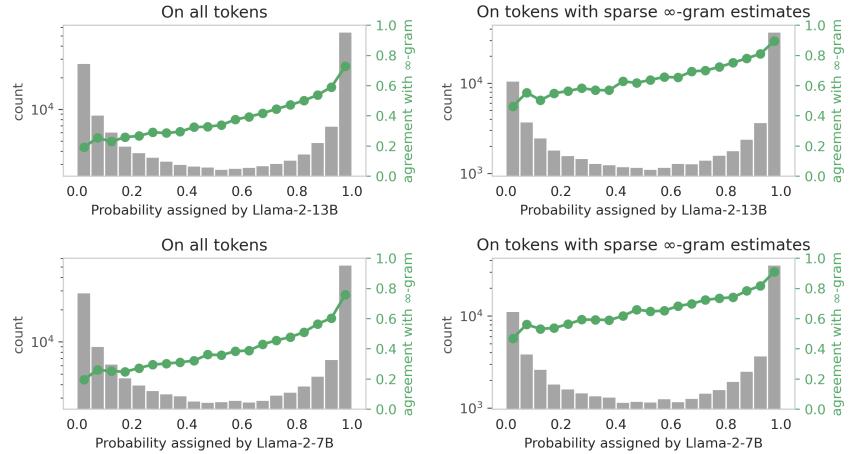


Figure 8: Continuation of Figure 4, with results on Llama-2-13b/7b.

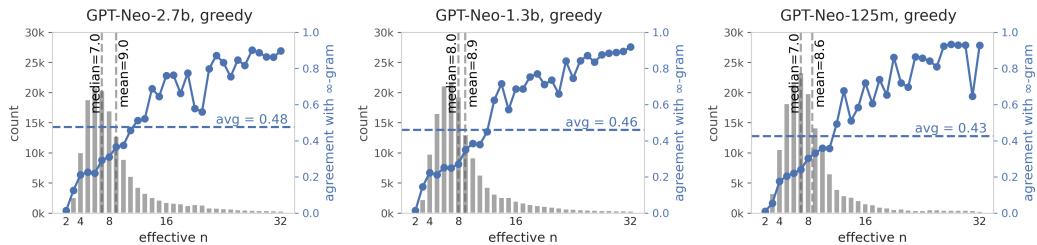


Figure 9: Continuation of Figure 5, with results on GPT-Neo models. Token-wise agreement between machine-generated text and  $\infty$ -gram. All tokens are considered.

**Neural LMs.** Below are additional details about the families of neural LMs we use.

Eval Data (Wikipedia)	simple interpolation w/ Random Forest	
	Neural	+ $\infty$ -gram
April 2023	5.64 <b>5.48</b> (3%)	5.86 <b>4.89</b> (20%)
May 2023	5.43 <b>5.27</b> (4%)	6.01 <b>5.70</b> (6%)
June 2023	5.49 <b>5.21</b> (6%)	5.69 <b>4.87</b> (17%)
July 2023	4.93 4.93 (0%)	4.91 <b>4.78</b> (3%)
August 2023	4.64 <b>4.46</b> (5%)	4.81 <b>4.50</b> (8%)

Table 5: Evaluation on time-shifted data. The evaluation data is taken from newly-added Wikipedia articles since April 2023, which is after the creation of both the Pile and RedPajama. The neural model is Llama-2 (13B), and the  $\infty$ -gram reference data is Pile + RPJ.

- **GPT-2** (Radford et al., 2019), one of the earliest autoregressive language models whose sizes range from 117M, 345M, and 774M to 1.6B. Their training data is a diverse set of web text, although is not public.
- **GPT-Neo** (Gao et al., 2020) and **GPT-J** (Wang & Komatsuzaki, 2021), language models trained on the Pile whose sizes vary from 125M, 1.3B, and 2.7B to 6.7B.
- **Llama-2** (Touvron et al., 2023b), a subsequent version of LLaMA (Touvron et al., 2023a) trained on two trillion tokens and has sizes of 7B, 13B, and 70B. Llama-2 is one of the most competitive language models whose weights are available at the time of writing the paper. The training data of Llama-2 is unknown, although the precedent version is trained on a large corpus of Common Crawls, Wikipedia and code, which is replicated by RedPajama (Together, 2023).
- **SILO** (Min et al., 2023a), 1.3B language models trained on permissively licensed data only. The original paper showed that training on permissively licensed data leads to the challenge of *extreme domain generalization* because the training data is skewed to highly specific domains like code and government text. We use three different variants, PD, PDSW and PDSWBY, which are trained on different levels of permissivity, leading to varying levels of the domain generalization challenge.

## D.2 Evaluating on time-shifted data

To further show the effectiveness of  $\infty$ -gram and eliminate doubts that our performance gains might be due to insufficient decontamination, we evaluate on **time-shifted data**: documents that were created after the cutoff time of the  $\infty$ -gram reference data. We use new Wikipedia articles created during April and August, 2023, which is after the cutoff time of both Pile and RedPajama.

Table 5 reports the perplexity of neural LM as well as the combined model. On documents in four out of the five months, interpolating with  $\infty$ -gram improves the perplexity of the neural LM. We find that this improvement can be further boosted by applying a Random Forest to decide an instance-wise interpolation hyperparameter, where the features of the Random Forest are the suffix lengths (1 up to the effective  $n$ ) as well as the frequency of each suffix in the reference data. When Random Forest is applied, the perplexity improvement ranges from 3% – 20%.

## D.3 Ablations

**Effect of the size of reference data.** Figure 10 reports performance of the combined model wrt the size of reference data. To create progressively smaller reference data, we repeatedly downsampled the full reference data by 2x (up to 256x, resulting in 9 sizes). We see the improvement brought by  $\infty$ -gram widens as reference data size grows, and the relationship is roughly log-linear (except for the NIH ExPorter domain, where  $\infty$ -gram doesn't help when the reference data is too small).

**Effect of the domain of reference data.** Figure 10 also compares performance of the combined model where the  $\infty$ -gram uses either the full reference data or only the in-domain reference data. Using only the in-domain reference data is roughly as powerful as using

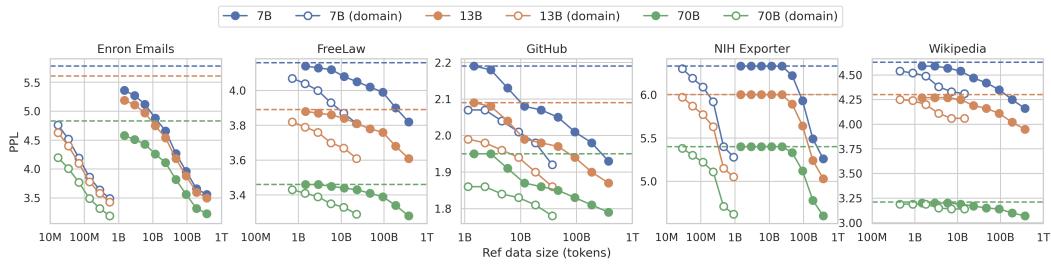


Figure 10: Impact of scaling the datastore of the  $\infty$ -gram, all using the Llama-2 models (7B, 13B, and 70B) as neural LMs, and the Pile as the reference data. - - -: neural LM only (baseline). ●:  $\infty$ -gram uses the full Pile; ○:  $\infty$ -gram uses only the in-domain portion of the Pile. Gains increase consistently as the datastore scales.

the full reference data, which implies that almost all improvement we have achieved is thanks to in-domain data (which has been decontaminated). This means it would not hurt to use the full reference data, especially when the test domain is unknown or an in-domain reference data is unavailable; however, having in-domain reference data is most helpful.

## E Extended Discussion

In §4 and §5, we showcased some very preliminary use cases of the infini-gram engine. However, we believe that infini-gram can enable much broader investigations and applications, including but not limited to:

**Understanding massive text corpora.** Text corpora used for pretraining language models have become prohibitively large, and we have relatively limited understanding of their contents (Elazar et al., 2023). Infini-gram can be a useful tool to quickly find out what *is* in the corpus and what is *not*, using  $n$ -gram lookup (the COUNT query).

**Data curation.** Data engineers often want to remove problematic content in corpora scraped from the Internet, such as toxicity, hate speech, and personal identifiable information (PII). Using infini-gram’s SEARCHDOC query (which can be easily modified to return all documents), one can retrieve all documents containing an  $n$ -gram term (or a CNF expression with multiple  $n$ -gram terms) and remove them from the corpus. Removal can even be done iteratively: infini-gram indexes are additive/subtractive, so we can obtain an index of the corpus after round one removal, by indexing the removed set and take the difference of the original index and the removal index.

**Document retrieval.** The scale of datastore is key to the effectiveness of retrieval-augmented LMs (Asai et al., 2024). However, vector-based indexes are difficult to scale up due to compute limit, storage limit, and inference efficiency. Infini-gram’s SEARCHDOC function can retrieve documents from datastores as large as the full pretraining corpora, and can potentially boost the performance of retrieval-augmented LMs.

**Reducing hallucination in factual knowledge.** Parametric-only models are prone to generating non-factual statements, which is widely known as the hallucination problem. Infini-gram can potentially be used to mitigate hallucination by reading verbatim from the training data. We have found evidence that the  $\infty$ -gram can greatly outperform Llama-2-70B on factual probing benchmarks such as LAMA (Petroni et al., 2019).

**Detecting data contamination, memorization, and plagiarism.** Test set contamination has become a major issue for language model evaluation.  $n$ -gram lookup enables us to check if evaluation queries have sneaked into the training data of neural LMs. It also opens up possibility to detect memorization in machine-generated text, or plagiarism in human-written text.

**Preventing copyright infringement.** Recently, generative AIs are facing numerous law-suits for generating arguably copyrighted materials. Infini-gram may be helpful for pre-

Method	# tokens ( $\uparrow$ )	# entries ( $\uparrow$ )	Storage usage ( $\downarrow$ )	max $n$
<i>Vector-based index</i>				
RETRO (Borgeaud et al., 2022)	1.8 T	$2.8 \times 10^{10}$	432 TB (16k bytes / entry)	–
Atlas (Izacard et al., 2022)	27 B	$4 \times 10^8$	200 GB (8 bytes / entry)	–
$k$ NN-LM (Khandelwal et al., 2020)	3 B	$3 \times 10^9$	200 GB (64 bytes / entry)	–
NPM (Min et al., 2023b)	1 B	$1 \times 10^9$	1.4 TB ( $\sim 2k$ bytes / entry)	–
<i>n-gram-based index</i>				
Brants et al. (2007)	2 T	$3 \times 10^{11}$	unreported	5
Google’s (Franz & Brants, 2006)	1 T	$3.8 \times 10^9$	24 GB	5
Google Books Ngram (Aiden & Michel, 2011)	500 B	unreported	unreported	5
Stehouwer & van Zaanen (2010)	90 M	unreported	unreported	$\infty$
Kennington et al. (2012)	3 M	$5 \times 10^{12}$	330 MB (110 bytes / token)	$\infty$
Shareghi et al. (2015)	9 B	$8 \times 10^{18}$	63 GB (7 bytes / token)	$\infty$
infini-gram (ours)	5 T	$1 \times 10^{25}$	35 TB (7 bytes / token)	$\infty$

Table 6: Comparison with other nonparametric language modeling methods. **# tokens**: number of tokens in the inference-time reference data. **# entries**: number of representations (counts) in the index. **max  $n$** : maximum number of context tokens considered. For infini-gram, we consider the combination of Pile-train and RedPajama as reference data.

venting copyright infringement, by diverting neural LMs to alternative (yet still plausible) generation paths when they are about to generate long  $n$ -grams that appear in the training data, especially if they mostly appear in documents from copyrighted sources.

**Measuring popularity of entity.** Mallen et al. (2022) showed that LMs have better memorization of facts about popular entities than less popular ones. In that paper, heuristic metrics like Wikipedia pageviews are used as proxy to popularity. A better proxy may be the number of appearances of the entity’s name in a massive text corpus, which can be easily computed by infini-gram’s COUNT query.

**Measuring novelty and creativity of text.** Are neural LMs generating genuinely novel text, or are they simply copying from its pretraining data? We need a metric for text novelty and creativity, and this metric can be potentially defined by the  $n$ -gram overlap between the generated document and the pretraining corpus.

**Attribution.** When using neural LMs to make predictions, people might want to know which training data most influenced the model’s decision. Using  $n$ -gram lookup with key phrases, we can trace back to related documents in the training data of neural LMs.

**Non-parametric speculative decoding.** Speculative decoding (Chen et al., 2023) speeds up text generation by employing a fast and a slow decoder, where the fast decoder is a smaller model that does the autoregressive token generation, and the slow decoder checks the fast decoder’s proposals by parallelizing the forward passes of multiple tokens. Given the low latency of infini-gram, we can potentially use  $\infty$ -gram as the fast decoder, similar to He et al. (2023).

We welcome the community to collaboratively build toward the aforementioned directions, by leveraging our publicly released web interface and API endpoint.

## F Extended Related Work

**$n$ -grams beyond  $n = 5$ .** Mochihashi & Sumita (2007) and Wood et al. (2009) consider infinitely-ordered Markov models, to which the  $\infty$ -gram LM is a special case. Aside from the  $\infty$ -gram LM, some other work has found value in scaling the value of  $n$  in  $n$ -grams. For example, KiloGrams (Raff et al., 2019) uses 1024-grams as features for malware detection.

**Other data structures for text indexing.** Beside suffix arrays and suffix trees, other data structures have been used to index text corpora to satisfy different trade-offs. Koala (Vu et al., 2023) builds a compressed suffix array to analyze overlap between text corpora. The ROOTS Search Tool (Piktus et al., 2023) builds a BM25 index on the ROOTS corpus, and

supports document searching via both exact match and fuzzy match of  $n$ -grams. Data Portraits (Marone & Durme, 2023) proposes a lightweight index based on Bloom Filter, and is tailored for probabilistic membership inference (exact match of  $n$ -grams of 50 characters, where  $n \approx 8$ ) against the Pile and Stacks. ElasticSearch is a proprietary search engine based on the Lucene index, and it has been used by Dodge et al. (2021) to search documents in C4, and also by Elazar et al. (2023) to count  $n$ -grams and list most frequent  $n$ -grams in various corpora up to 480B tokens.

**Nonparametric language models.** A nonparametric LM refers to the LM whose complexity is not bounded as a priori, because the complexity can grow or update according to the data given at inference time. Prior work is broadly divided into two categories: a *token* retrieval approach that represents each token as one vector and uses a nonparametric prediction function (Khandelwal et al., 2020; Zhong et al., 2022; Lan et al., 2023; Min et al., 2023b;a; Shi et al., 2023), and a *chunk* retrieval approach that represents each chunk of text as a vector and incorporates nearest chunks to the neural language model (Guu et al., 2020; Izacard et al., 2022; Borgeaud et al., 2022). Scaling the reference data in nonparametric LMs is very expensive as it requires storing a vector for every unit (either token or chunk). To the best of our knowledge, prior work with the largest reference data is RETRO (Borgeaud et al., 2022), which uses the 7B-parameter LM and the reference data consisting of 1.8 trillion tokens. It stores and searches over 28 billion vectors, estimated to consume 432TB of disk space.<sup>2</sup> (Detailed comparisons in Table 6.)

## G Example Queries and Web Interface

Figures 11 to 16 show one example for each of the six query types supported by infini-gram. We query the Dolma corpus in these examples. Screenshots are taken from our web interface.

The screenshot shows the infini-gram web interface. On the left, there's a sidebar with 'Corpus' and 'Engine' sections. Under 'Corpus', 'Dolma (3.1T tokens)' is selected. Under 'Engine', 'C++ (Fast)' is selected. The main area has a tab bar with '1. Count an n-gram' (selected), '2. Prob of the last token', '3. Next-token distribution', '4. oo-gram prob', '5. oo-gram next-token distribution', and '6. Search documents'. Below the tabs, it says 'This counts the number of times an n-gram appears in the corpus. If you submit an empty input, it will return the total number of tokens in the corpus.' An example query 'natural language processing' is shown with its count '1,012,875'. There are 'Clear' and 'Submit' buttons. Below the query input, there's a 'Latency (milliseconds)' field showing '41.505' and a 'Tokenized' field showing the JSON array `["natural", "language", "processing"] [5613, 4086, 9068]`.

Figure 11: Example for query type 1 (COUNT): Counting an  $n$ -gram.

<sup>2</sup>This is in part because RETRO does not use any approximation in  $k$ NN search. Even if RETRO used approximate search as Khandelwal et al. (2020) did, it would still use 10TB. Moreover, there is no open-sourced software that easily supports fast  $k$ NN search over tens of billions of vectors.

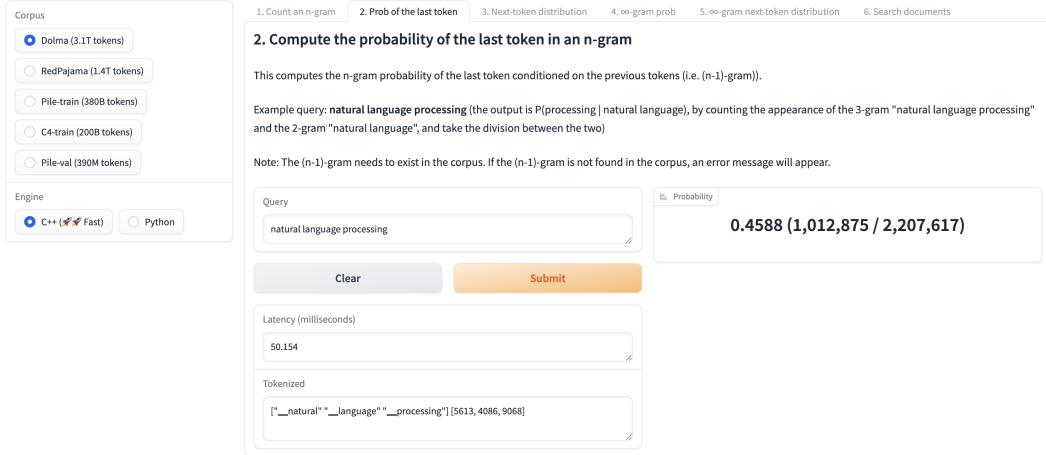


Figure 12: Example for query type 2 (NGRAMPROB): Computing a token probability from  $n$ -gram LM (with given  $n$ , no backoff).

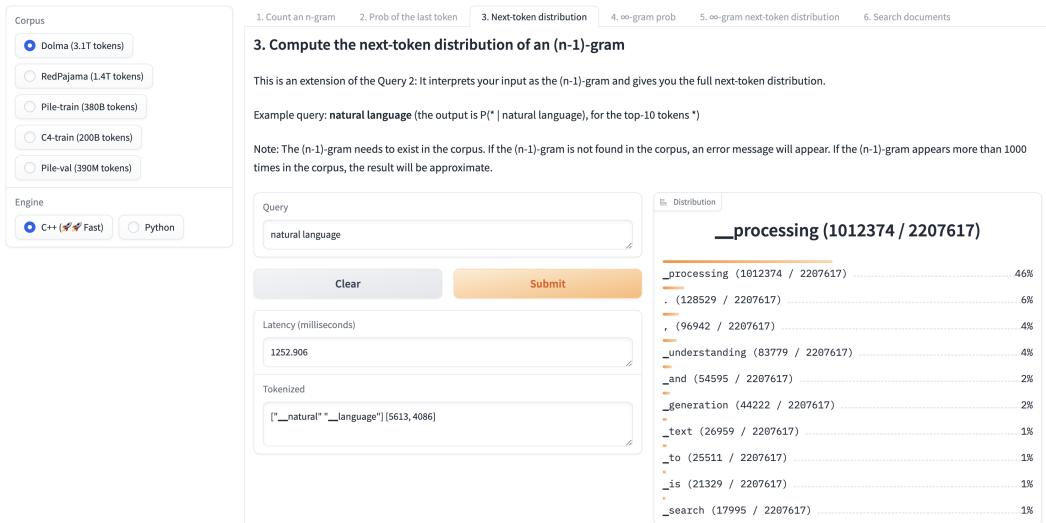


Figure 13: Example for query type 3 (NGRAMDIST): Computing the full next-token distribution from  $n$ -gram LM. Due to space limits, only top-10 tokens are shown.

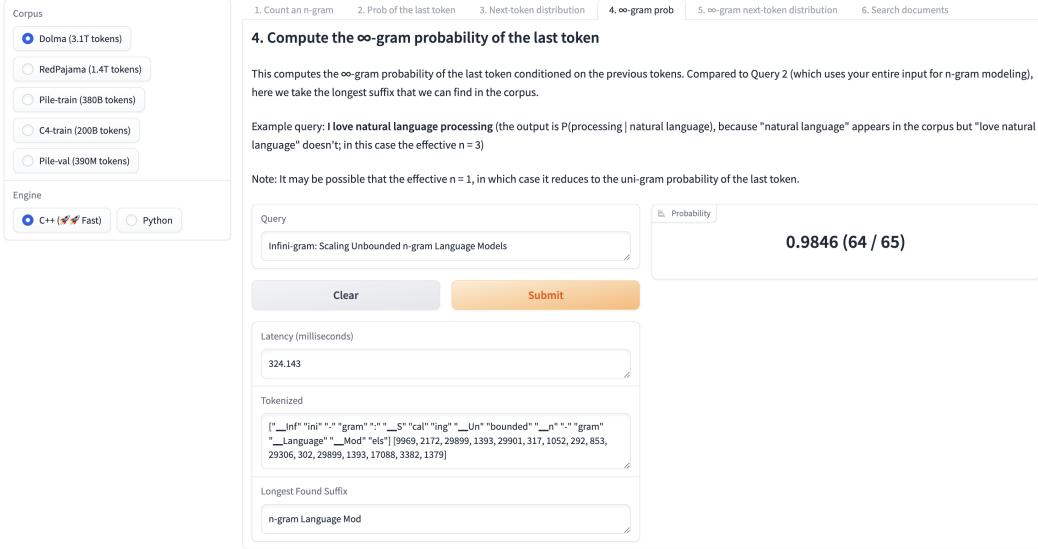


Figure 14: Example for query type 4 (INFGRAMPROB): Computing a token probability from  $\infty$ -gram LM.

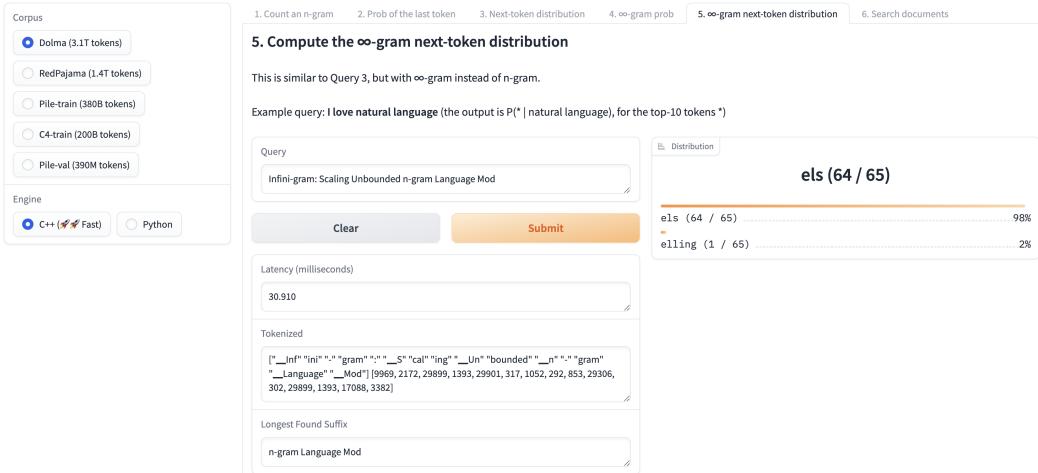


Figure 15: Example for query type 5 (INFGRAMDIST): Computing the full next-token distribution from  $\infty$ -gram LM. Due to space limits, only top-10 tokens are shown.

The screenshot shows a user interface for searching documents based on n-grams. On the left, there's a sidebar with 'Corpus' and 'Engine' sections. Under 'Corpus', 'Dolma (3.1T tokens)' is selected. Under 'Engine', 'C++ (Fast)' is selected. The main area has tabs: 1. Count an n-gram, 2. Prob of the last token, 3. Next-token distribution, 4. oo-gram prob, 5. oo-gram next-token distribution, and 6. Search documents. Tab 6 is active. A section titled '6. Search for documents containing n-gram(s)' contains instructions and example queries. It says: "This displays a few random documents in the corpus that satisfies your query. You can simply enter an n-gram, in which case the document displayed would contain your n-gram. You can also connect multiple n-gram terms with the AND/OR operators, in the CNF format, in which case the displayed document contains n-grams such that it satisfies this logical constraint." Example queries include: "natural language processing" (the displayed document would contain "natural language processing"), "natural language processing AND deep learning" (the displayed document would contain both "natural language processing" and "deep learning"), and "natural language processing OR artificial intelligence AND deep learning OR machine learning" (the displayed document would contain at least one of "natural language processing" / "artificial intelligence", and also at least one of "deep learning" / "machine learning"). Below this, notes say: "If you want another batch of random documents, simply hit the Submit button again." A note about CNF precedence: "When you write a query in CNF, note that OR has higher precedence than AND (which is contrary to conventions in boolean algebra)." Other notes include: "If the document is too long, it will be truncated to 5000 tokens.", "We can only include documents where all terms (or clauses) are separated by no more than 100 tokens.", "If you query for two or more clauses, and a clause has more than 50000 matches (per shard), we will estimate the count from a random subset of all documents containing that clause. This might cause a zero count on conjunction of some simple n-grams (e.g., birds AND oil).", "The number of found documents may contain duplicates (e.g., if a document contains your query term twice, it may be counted twice).", and a warning: "WARNING: Corpus may contain problematic contents such as PII, toxicity, hate speech, and NSFW text. This tool is merely presenting selected text from the corpus, without any post-hoc safety filtering. It is NOT creating new text. This is a research prototype through which we can expose and examine existing problems with massive text corpora. Please use with caution. Don't be evil :)".

**Query:** natural language processing AND deep learning

**Number of documents to Display:** 10

**Latency (milliseconds):** 3067.464

**Tokenized:**

```
[{"natural": "language", "processing": ""}, {"deep": "learning"}]
```

**Message:** Approximately 78380 occurrences found. Displaying the documents of occurrences # [2526, 13007, 1920, 7775, 69, 11136, 6112, 6312, 13600, 272]

**Document:** Document #203726548, length = 3086 tokens (500 tokens displayed)

deep learning in natural language processing is based on real results and that the improvements appear to be continuing and perhaps speeding up.

The Promise of End-to-End Models. That is, that large end-to-end deep learning models can be fit on natural language problems offering a more general and better-performing approach.

Natural language processing is not "solved", but deep learning is required to get you to the state-of-the-art on many challenging problems in the field.

Figure 16: Example for query type 6 (SEARCHDOC): Returning documents containing an  $n$ -gram, or a CNF logical expression of  $n$ -gram terms, connected with AND's and/or OR's.