COVERING LANGUAGE MODEL OUTPUTS WITH SIMPLICIAL COMPLEXES

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models (LMs) predict the next-token probability distribution over a large vocabulary, yet most of the probability mass is often concentrated on a small, context-dependent subset. This raises the question of whether LM outputs across contexts admit a compact representation. To answer this question, we first characterize this behavior of LM outputs using Bhattacharyya geometry on probability simplexes and simplicial complexes from algebraic topology. Our theorem connects probability mass discarded by truncation to our geometric dissimilarity measure. Because each output point lies near the low-dimensional simplex of its high-probability tokens, the output points are covered by a neighborhood of a simplicial complex, a union of simplexes. Next, we evaluate representations of LM outputs through the tradeoff among four metrics: (d, m, r, u). The first two measure the complexity of simplicial complexes, while the latter two assess approximation quality. To construct compact covers, we first consider collecting pointwise covers. For a single output, Top-k and Top-p truncation determine optimal simplexes in the tradeoff between r and d with fixed $m = 1$ and $u = 0$. Collecting these simplexes yields simplicial complexes, which contain a prohibitively large number of simplexes in practice. To obtain a more compact complex, we formulate the construction task as a monotone submodular maximization problem and propose an algorithm for this problem. The algorithm provides the same worst-case guarantee as previous work. In the experimental section, we evaluate the construction using a natural language corpus (Dolma) and real-world LLMs (OLMo, Pythia). One of the resulting simplicial complexes uses 843 simplexes of dimension at most 4096 to cover 95% of the evaluated outputs while retaining at least 90% of each covered output's probability mass, providing a compact representation of approximately 50K-dimensional LM outputs over a 100M-token corpus.

1 INTRODUCTION

Language models (LMs) predict distributions over tens of thousands of tokens. Yet, for a given context, most of the probability mass is typically concentrated on a much smaller, context-dependent subset (Fan et al., 2018; Chen et al., 2018; Holtzman et al., 2020; Hewitt et al., 2022). Table 1 gives examples of contexts and their Top-p candidate sets. A cooking context yields cooking-related candidates, while the other examples exhibit even sharper syntactic or semantic restrictions. Thus, the ambient probability simplex may substantially overstate the effective size of the space actually explored by LM outputs (Fig. 1).

This motivates our central question: can this concentration behavior of LM outputs be represented by a compact collection of low-dimensional components (Chen et al., 2018)? We address this question using *simplicial complexes* (SCs) (Hatcher, 2002; Edelsbrunner & Harer, 2010), which are unions of simplexes, and Bhattacharyya geometry (Bhattacharyya, 1946) of probability simplexes, characterized by the *Bhattacharyya coefficient*. This provides a geometric characterization of the LM output space, complementing previous work on word embeddings, hidden states, and logits.

We first consider an idealized setting, where LM outputs are sparse probability distributions and thus lie on low-dimensional probability simplexes. The union of these simplexes forms a simplicial complex that covers the outputs. In contrast, in the actual setting where the softmax layer assigns positive

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

probability to every token in the vocabulary, we give a generalization using the Bhattacharyya coefficient (Bhattacharyya, 1946). Theorem 1 connects probability mass discarded by truncation to our dissimilarity metric. In this sense, the concentration of next-token probability mass on a small subset of tokens implies this output point lies near the corresponding probability simplex. Consequently, the output points are characterized as lying in a neighborhood of an SC with a certain radius. We then evaluate representations of LM outputs through the tradeoff among four quantities: maximum simplex dimension d, simplex count m, neighborhood radius r, and uncovered fraction u. The first two quantities measure the complexity of SCs, while the latter two describe coverage quality. This tradeoff can also be interpreted as an upper bound on the effective width of the distribution of the output points, that is, the size of the output space. To construct SCs, we start by collecting pointwise covers. For a single output, where one can fix $u = 0$ and $m = 1$, Top-k truncation (Fan et al., 2018) minimizes r at a fixed d, while Top-p truncation (Holtzman et al., 2020) minimizes d subject to an upper bound on r. We summarize their pointwise approximation errors as r_{pw}. Both heuristics induce SCs. In the worst case, however, these SCs can contain as many simplexes as tokens in the corpus. To obtain a more compact cover, we impose upper bounds on r, d and m and maximize coverage $1 - u$, which is a monotone submodular objective. We develop an algorithm based on Sieve-Streaming (Badanidiyuru et al., 2014), which requires only a single pass over a large corpus.

We compare the constructed SCs with complexes induced by Top-k and Top-p on a 100M-token excerpt from Dolma (Soldaini et al., 2024). We evaluate two LMs trained on natural language: OLMo-7B (Groeneveld et al., 2024) and Pythia-6.9B (Biderman et al., 2023). One constructed SC consists of 843 simplexes of dimension at most 4096 and covers 95% of the evaluated outputs while retaining at least 90% of each covered output' s probability mass. These simplexes have much lower dimension than the $50,279$-dimensional full probability simplex, and the complex contains far fewer simplexes than those induced by pointwise Top-p and Top-k, which have tens of millions of simplexes.

Our contributions are as follows.

- We develop a geometric view of LM output spaces using simplicial complexes and Bhattacharyya geometry, linking truncation to projection and approximation to covering.
- We characterize representations of LM outputs through a joint tradeoff among four quantities. This tradeoff can also be interpreted in terms of the effective width of the output distribution. Top-k and Top-p emerge as pointwise optimal choices in the tradeoff between dimension d and radius r.
- We empirically demonstrate cross-context compactness on a 100M-token corpus with a real-world corpus and LMs, and report the resulting values of (d, m, r, u).

Our framework provides low-dimensional simplexes and their neighborhoods as potential units of analysis for studying the mechanisms underlying language model behavior.

2 Related Work

LM Geometry While word embedding methods such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) have been analyzed from several theoretical perspectives (Levy & Goldberg, 2014; Korchinski et al., 2025; Karkada et al., 2025), understanding the internal structures of autoregressive language models remains an open problem in machine learning (ML) and natural language processing (NLP). Regarding hidden states, the linear representation hypothesis (Park et al., 2024b) has received empirical support (Marks & Tegmark, 2024; Gurnee & Tegmark, 2023) and inspired extensions or refinements (Xiong, 2025; Park et al., 2024a; Engels et al., 2024; Zhao et al., 2024). Other work studies superposition (Elhage et al., 2022) and sparsely activating features (Huben et al., 2023), while the logit lens (nostalgebraist, 2020) and the tuned lens (Belrose et al., 2025) interpret intermediate vectors through vocabulary predictions. Connections between hidden states and output spaces have also been studied (Park et al., 2026; Arvanitidis et al., 2022; Wurgaft et al., 2026). Our study complements this prior work by characterizing LM output points.

LM Output Spaces and Their Size Studies of the softmax bottleneck (Yang et al., 2018; Demeter et al., 2020; Godey et al., 2024; Chang & McCallum, 2022) have characterized the expressivity of the output space through dimensional constraints (Finlayson et al., 2024). Basri & Jacobs

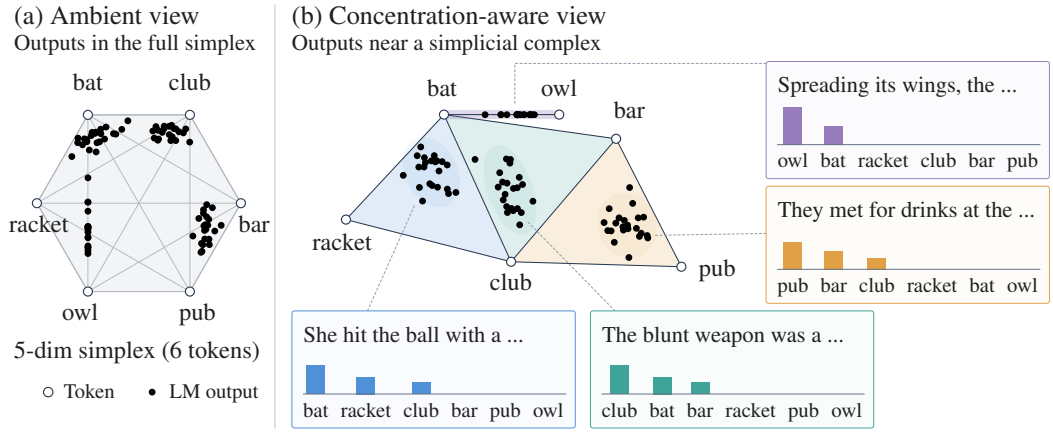


Figure 1: (a) LM outputs lie in a high-dimensional probability simplex. (b) Yet most probability mass typically concentrates on small, context-dependent token subsets. This places outputs near low-dimensional vocabulary simplexes whose union forms a simplicial complex.

(2026) suggest that these constraints are less restrictive for representing probability distributions concentrated on small vocabulary subsets. Related approaches use sparse probability mappings and restricted candidate vocabularies. Sparsemax and entmax can produce distributions with exact zeros (Martins & Astudillo, 2016; Peters et al., 2019), while vocabulary selection methods accelerate inference (Chen et al., 2018; Tranheden et al., 2026; Zhang et al., 2026). Also, Top-k (Fan et al., 2018) and Top-p (Holtzman et al., 2020) can be interpreted as estimates of restricted supports (Hewitt et al., 2022). We instead characterize this phenomenon geometrically through the joint tradeoff among four quantities.

Topological Methods in NLP In NLP, Cavaliere et al. (2017) and Fitz (2022) construct SCs from text corpora, with words as vertices. Our approach instead models LM outputs and assigns linguistic meaning to points within simplexes.

Quantitative Characterization Nonlinear Kolmogorov widths (Temlyakov, 1998) and library widths (Bonito et al., 2021) quantify the size of a compact set through the tradeoff among the number of approximation subspaces, their dimensions, and the approximation error. We use simplexes instead and allow a certain fraction of outputs to be uncovered. While prior work characterizes Top-k and Top-p through optimization under constraints (Noarov et al., 2025; Ji et al., 2026), we identify them as complementary optima in the pointwise setting of our more general formulation.

3 Preliminaries: Simplexes and Simplicial Complexes

Let V be a fixed finite vocabulary.

Definition 1. We define the probability simplex over V as $\sigma(V) := \{\mathbf{p} \in \mathbb{R}^V \mid p_i \geq 0, \sum_{i \in V} p_i = 1\}$. For a nonempty subset $s \subseteq V$, we define the probability simplex over s as the face $\sigma(s) := \{\mathbf{p} \in \sigma(V) | \operatorname{supp}(\mathbf{p}) \subseteq s\}$, where $\operatorname{supp}(\mathbf{p}) := \{i \in V \mid p_i > 0\}$ is the support of $\mathbf{p}$.

We denote the relative interior of $\sigma(s)$ by $\mathring{\sigma}(s)$. A point $\mathbf{p}$ in $\mathring{\sigma}(s)$ represents a probability distribution over s that assigns strictly positive probability to every element of s. $\sigma(s)$ is $(|s| - 1)$-dimensional.

Definition 2. We define a simplicial complex K as a union of simplexes: $K = \bigcup_{s \in \Sigma} \sigma(s)$ for some $\Sigma \subseteq 2^V$. For simplicity, we also represent K by its generating family Σ.

See Hatcher (2002) for more details on simplicial complexes and algebraic topology.

Definition 3. A language model (LM) is a function that maps a context to a next-token probability distribution. Formally, this is a map from a corpus $\mathcal{C}$ to $\mathring{\sigma}(V)$, where $\mathcal{C}$ is a set of finite sequences of tokens in V. We assume that the output lies in $\mathring{\sigma}(V)$, as it does for models with a softmax layer.

Table 1: Examples of contexts and their corresponding Top-p token sets

Context	Top-p tokens
You can add a little more cashew milk if ___	needed, you, the, it, dough, your, mixture, batter, necessary, its
the New York Islanders beat the Boston Bruins 4–1 ___	on, Saturday, in, Sunday, Thursday, Tuesday, Wednesday, Friday, Monday
this.collapseLabel_ = collapseLabel;} \n const ___	label, collapse, collapsed, tip, render, ul, button, att, expand, icon

4 Geometric Covering and Effective Width of LM Outputs

To address the question raised in the Introduction, we first pose the following two questions, which shape our framework. First, how can the concentration phenomenon be represented geometrically? Second, how can we quantify the compactness of such representations?

4.1 From Support Concentration To Simplicial Covers

We first consider an idealized setting, where the support of each probability distribution is restricted to a small vocabulary subset: $\mathrm{supp}(\mathrm{LM}(c)) \neq V$ and $\mathrm{LM}(c) \in \mathring{\sigma}(\mathrm{supp}(\mathrm{LM}(c)))$. Taking the union of these simplexes yields an answer to the first question: $\mathrm{LM}(\mathcal{C}) \subseteq \bigcup_{c\in\mathcal{C}} \sigma(\mathrm{supp}(\mathrm{LM}(c)))$.

We next generalize this result to the actual setting, where $\mathrm{supp}(\mathrm{LM}(c))$ is always V. A truncation function links the actual outputs to the idealized outputs.

Definition 4. For a nonempty subset $s \subseteq V$, we define truncation to s by

$$T_s : \mathring{\sigma}(V) \to \mathring{\sigma}(s);\ \mathbf{p} \mapsto \begin{cases} p_i/Z & (i \in s) \\ 0 & (i \notin s) \end{cases},$$

where $Z = \sum_{j\in s} p_j$. We refer to $\sum_{j\notin s} p_j = 1 - Z$ as the probability mass discarded by truncation.

For each context, the discrepancy between the idealized and actual outputs is quantified by this discarded mass. To interpret this from a geometric perspective, we define similarity and dissimilarity measures for points $\mathbf{p} = (p_i)_{i\in V}, \mathbf{q} = (q_i)_{i\in V} \in \sigma(V)$.

Definition 5 (Bhattacharyya coefficient and the corresponding dissimilarity)**.**

$$\cos_B(\mathbf{p},\mathbf{q}) := \textstyle\sum_i \sqrt{p_i q_i},\ \sin_B(\mathbf{p},\mathbf{q}) := \sqrt{1-\cos_B^2(\mathbf{p},\mathbf{q})}. \tag{1}$$

$\cos_B(\mathbf{p},\mathbf{q})$ is called the Bhattacharyya coefficient (Bhattacharyya, 1946).

We do not use KL divergence because truncated points lie on the simplex boundary, where the divergence can become infinite. The quantity $\arccos(\cos_B(\mathbf{p},\mathbf{q}))$ gives the geodesic distance, and $\sin_B^2$ locally agrees with the KL divergence up to a constant factor (Amari, 2016).

Theorem 1 provides the main motivation for using this measure.

Theorem 1. *For a nonempty closed set $M \subseteq \sigma(V)$ and a point $p \in \mathring{\sigma}(V)$, define $\cos_B(p, M) := \max_{q\in M} \cos_B(p,q)$ and $\sin_B(p,M) := \sqrt{1-\cos_B^2(p,M)}$. Then, for any nonempty $s \subseteq V$*

$$\cos_B^2(\mathbf{p},\sigma(s)) = \textstyle\sum_{i\in s} p_i,\ \sin_B^2(\mathbf{p},\sigma(s)) = \sum_{i\notin s} p_i. \tag{2}$$

Proof sketch. For any $q \in \sigma(s)$, the Cauchy–Schwarz inequality gives $\cos_B^2(p,q) \leq \sum_{i\in s} p_i$. Equality is attained at $q = T_s(p)$. □

The proof also shows $T_s(\mathbf{p}) = \arg\max_{\mathbf{q}\in\sigma(s)} \cos_B(\mathbf{p},\mathbf{q})$. Thus, T_s defines a *projection* onto $\sigma(s)$.

Next, we define a neighborhood of a simplicial complex using the dissimilarity $\sin_B$. Intuitively, this neighborhood is a thickened version of the complex that can contain the original output points.

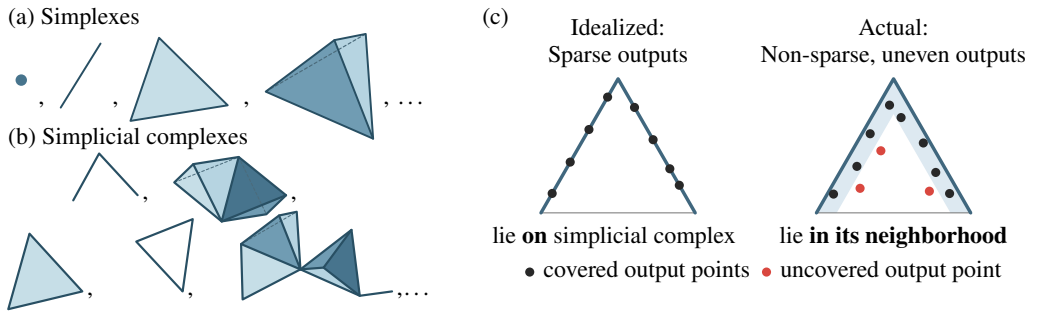


Figure 2: (a) Examples of simplexes. (b) Examples of simplicial complexes. A single simplex is itself an SC. (c) An illustration of the idealized and actual settings in Section 4.

Definition 6 (Neighborhood of a simplicial complex)**.** For a simplicial complex K and $r \in [0, 1]$, define the r-neighborhood of K as $\mathcal{N}_r(K) := \{\mathbf{p} \in \sigma(V) | \sin_B(\mathbf{p}, K) \leq r\}$.

$\mathbf{p} \in \mathcal{N}_r(K)$ iff there exists a simplex $\sigma(s) \subseteq K$ such that truncating p to s discards at most r^2 of its probability mass. Note that $\mathcal{N}_0(K) = K$ and $\mathcal{N}_1(K) = \sigma(V)$.

Now, we can formulate our general answer to the first question. See Fig. 2 for an illustration.

Answer 1. *Given the approximate support* $\mathrm{supp}'(\mathrm{LM}(c))$ *of each output,*

$$\mathrm{LM}(\mathcal{C}) \subseteq \mathcal{N}_r \left(\bigcup_{c \in \mathcal{C}} \sigma(\mathrm{supp}'(\mathrm{LM}(c))) \right) \tag{3}$$

when r^2 *is at least the maximum mass discarded by truncation* $T_{\mathrm{supp}'(\mathrm{LM}(c))}$ *over* $c \in \mathcal{C}$.

Full coverage depends on the maximum discarded mass, potentially driven by a few contexts in a large corpus. We therefore allow an uncovered fraction u, yielding the following formulation.

Corollary 1.1 (Cover of output points)**.** *Let* K *be a simplicial complex with* $K \neq \varnothing, \sigma(V)$*, and suppose that*

$$u \geq \mathrm{Prob}_{c \in \mathcal{C}} \left(\mathrm{LM}(c) \notin \mathcal{N}_r(K) \right) \tag{4}$$

for some $u \in [0, 1]$*. Then* r^2 *is at least the* $(1 - u)$*-quantile of* $\sin_B^2(\mathbf{p}, K)$ *over outputs* $\mathbf{p}$.

Setting $u = 0$ recovers the previous answer. For a fixed K, the minimum uncovered fraction at radius r is given by the complementary cumulative distribution function (CCDF) of $\sin_B^2(\mathrm{LM}(c), K)$ over $c \in \mathcal{C}$, evaluated at r^2. See Section 7 for the empirical results.

4.2 Evaluating Covers: Complexity And Coverage

In this section, we characterize the compactness of a cover introduced in the previous section.

Definition 7 (Compactness of a cover)**.** For a cover defined by a simplicial complex K as in Corollary 1.1, we evaluate the cover of LM outputs in terms of the complexity of K and the coverage quality specified by the neighborhood radius r and uncovered fraction u.

For a simplicial complex $K = \bigcup_{s \in \Sigma} \sigma(s)$ with $\Sigma \subseteq 2^V$, define $d := \max_{s \in \Sigma}(|s| - 1)$ and $m := |\Sigma|$. Here, d measures the maximum simplex dimension, while m counts the selected simplexes[1].

The four quantities d, m, r, and u trade off against one another. For example, smaller values of d and m give a simpler cover, whereas smaller values of r and u require more accurate approximation with fewer exceptions. See Fig. 3 for an overview of the four quantities and their tradeoffs.

By analogy with Temlyakov (1998) and Bonito et al. (2021), the optimal tradeoff among (d, m, r, u) can be viewed as the effective width of the distribution of output points. The tradeoffs achieved by the constructions below therefore provide upper bounds on this effective width.

[1]We count selected rather than maximal simplexes. Replacing Σ by the family of its maximal elements preserves K, d, r, and u while reducing m.

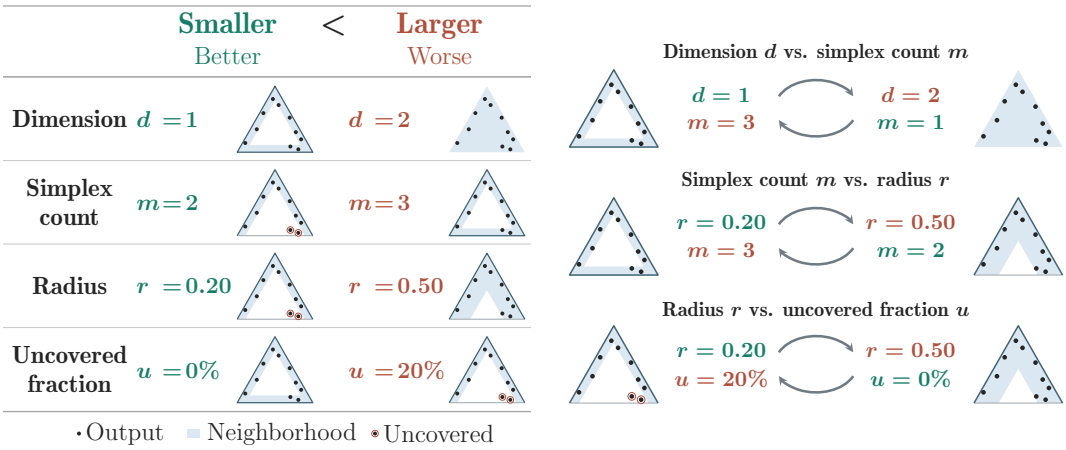


Figure 3: Left: Each of the four quantities captures a different aspect of a cover. Right: Trade-offs arise between different pairs of these quantities.

5 Pointwise Construction

We now turn to the construction of compact covers. A natural starting point for constructing covers with SCs is to select an optimal simplex for each output and collect them across contexts. In our framework, Top-k and Top-p truncations yield the pointwise optimal projections. However, the induced SCs contain very large numbers of simplexes in practice. This motivates a construction method that produces smaller SCs.

5.1 Top-k and Top-p as Pointwise Optimal Projections

By Theorem 1, both Top-k and Top-p truncations can be interpreted as projections onto simplexes. See Fig. 4 for illustrations.

For a single output, we require full coverage and thus set $u = 0$. Selecting one simplex fixes $m = 1$, leaving a tradeoff between dimension d and radius r.

5.1.1 Top-k

Top-k is a truncation method that retains the k most probable tokens (Fan et al., 2018). Theorem 1 gives the following interpretation.

Corollary 1.2. *Top-k truncation is a projection onto the nearest $(k-1)$-dimensional simplex. Equivalently, it selects a simplex that minimizes the radius r for the fixed dimension $d = k - 1$.*

5.1.2 Top-p

Top-p retains the smallest set whose cumulative probability is at least the threshold $\alpha \in [0, 1]$ (Holtzman et al., 2020). This gives an analogous geometric interpretation.

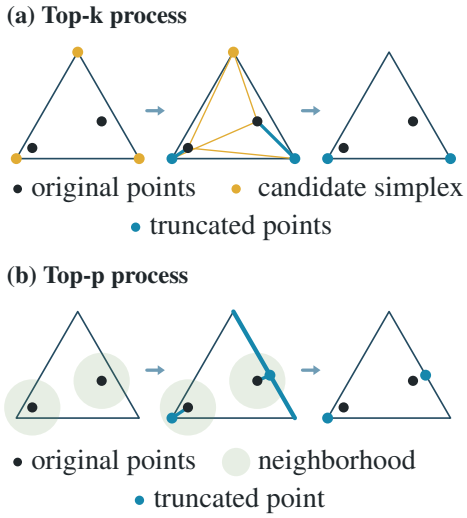


Figure 4: Illustration of Corr. 1.2 and Corr. 1.3. (a) The Top-k process selects the nearest $(k-1)$-dim simplex. (b) The Top-p process selects the lowest-dimensional simplex intersecting $N_{\sqrt{1-\alpha}}(\{q\})$.

Corollary 1.3. *Top-p truncation of $\mathbf{q}$ is a projection onto the lowest-dimensional simplex $\sigma(s)$ such that $\mathcal{N}_{\sqrt{1-\alpha}}(\{\mathbf{q}\}) \cap \sigma(s) \neq \varnothing$. Equivalently, it minimizes the dimension d subject to the radius bound $r \leq \sqrt{1-\alpha}$. Ties in dimension are broken by choosing a nearest simplex.*

5.2 Collecting Pointwise Optimal Simplexes

Taking the union of the simplexes selected across contexts gives $K_{\text{Top-}p/\text{Top-}k} = \bigcup_{c\in\mathcal{C}} \sigma(s_c)$. While the covering radius r measures proximity to the entire complex, the nearest simplex may differ from the one Top-k or Top-p selects for each output.

To retain this pointwise association, we define, for prescribed $u \in [0,1]$, $r_{\text{pw}} := (1-u)\text{-quantile}\big(\sin_B(\text{LM}(c), \sigma(s_c))\big)_{c\in\mathcal{C}}$. Since the definition of r allows each output to be covered by any simplex in K, $r \le r_{\text{pw}}$ holds. For Top-k, each selected simplex is nearest among all $(k-1)$-dimensional simplexes, so $r = r_{\text{pw}}$. For Top-p, since each selected simplex retains mass at least α, $r^2 \le r_{\text{pw}}^2 \le 1-\alpha$ holds. Thus $K_{\text{Top-}p}$ achieves $(r,u) = (\sqrt{1-\alpha}, 0)$ for any given value of α.

However, these pointwise guarantees do not control the simplex count m, which can reach $|\mathcal{C}|$ in the worst case. Table 3 shows that the resulting complexes are large in practice. Thus, pointwise optimality alone does not ensure compression across contexts.

6 Constructing Compact Shared Covers

We next construct SCs by maximizing coverage under explicit bounds $d_{\max}, m_{\max}$, and $r_{\max}$ on d, m, and r, respectively. First, we formalize the construction.

Proposition 1. *For $K = \bigcup_{s\in\Sigma}\sigma(s)$, define the coverage function: $\psi_r(\Sigma) = |\{c \in \mathcal{C} \mid \text{LM}(c) \in N_r(K)\}|/|\mathcal{C}|$. This function gives the fraction of contexts whose outputs lie in $N_r(K)$ and is monotone submodular.*

For definitions of monotonicity and submodularity, see Nemhauser et al. (1978). Minimizing u can be formulated as monotone submodular maximization subject to a cardinality constraint $m_{\max}$. The candidates are vocabulary subsets s satisfying $|s| - 1 \le d_{\max}$. We obtain them from the Top-p supports s_c selected for contexts $c \in \mathcal{C}$.

Algorithm 1: Modified Sieve-Streaming

Require: Corpus $\mathcal{C}$, coverage function ψ_r, simplex budget $m_{\max}$, dimension bounds $d_{\min}$ and $d_{\max}$, accuracy parameter ϵ, and grid parameters $a > 1$ and $b \in \mathbb{N}$

Ensure: Simplicial complex K

Construct K_{rand} at random
$\tau_{\min} \leftarrow \psi_r(K_{\text{rand}})/2$ [a]
$\mathcal{T} \leftarrow \{\tau_{\min}(1+\epsilon)^i : i \in \mathbb{Z}_{\ge 0}\} \cap [0, \frac{1}{2}]$
$\mathcal{T} \leftarrow \mathcal{T} \cup \{1 - \frac{1}{2a^j} : j = 0, \dots, b\}$
$K_\tau \leftarrow \emptyset$ for all $\tau \in \mathcal{T}$
for all batches $\mathcal{B}$ in $\mathcal{C}$ **do**
  $\mathcal{A} \leftarrow \{s_c : c \in \mathcal{B},\ d_{\min} \le |s_c| - 1 \le d_{\max}\}$
  for all $\tau \in \mathcal{T}$ **do**
    while $|K_\tau| < m$ and $\mathcal{A} \neq \emptyset$ **do**
      $s^\star \leftarrow \arg\max_{s\in\mathcal{A}} \Delta_r(s \mid K_\tau)$
      $\theta \leftarrow \max\{\frac{\tau - \psi_r(K_\tau)}{m - |K_\tau|}, 0\}$
      if $\Delta_r(s^\star \mid K_\tau) \le \theta$ **then break**
      $K_\tau \leftarrow K_\tau \cup \{s^\star\}$
return K_τ with the largest value of $\psi_r(K_\tau)$

[a] To generate a valid grid, $\tau_{\min}$ must be clipped to a strictly positive lower bound.

Algorithm

Algorithm 1 is based on Sieve-Streaming (Badanidiyuru et al., 2014) and incorporates several simplifications enabled by the trivial upper bound of 1 on coverage. We write $\Delta_r(s \mid K) := \psi_r(K \cup \{s\}) - \psi_r(K)$ for the marginal coverage gain from adding s to K. Our algorithm can be viewed as a batch-greedy method that explores multiple target scores in parallel. Its batched processing scheme is inspired by Batch-Sieve-Streaming++ (Kazemi et al., 2019). Because exact coverage evaluation requires scanning the entire corpus for each query, we instead estimate coverage using a fixed, randomly sampled witness corpus. For a fixed witness corpus, the empirical coverage objective is monotone submodular, and our algorithm inherits the same $(1/2 - \epsilon)$ worst-case approximation guarantee as Sieve-Streaming under the stated constraints for the filtered input stream. Its computational complexity is $O(|\mathcal{C}| + \rho|\mathcal{T}|\,(|\mathcal{C}| + |\mathcal{B}|m))$,[2] where ρ denotes the fraction of contexts whose Top-p supports satisfy the dimension bounds.

[2] For simplicity, we exclude the costs of computing $\text{LM}(c)$, extracting Top-p supports, and evaluating ψ_r.

Table 4: Example tokens from simplexes in the constructed simplicial complex

Linguistic pattern	Example tokens
Base-form verbs	_donate, _realise, _disclose, _recognise, _enroll, _reproduce, _derive, _analyse, etc.
Verb forms, including third-person singular forms	're, 've, _resembles, _specifies, _tackles, _lied, _hath, etc.
Emotions and attitudes	_diligence, _obsess, _contempt, _sacrifices, _sensations, _obedience, _intellig, _ethic, etc.

7 Experimental Results

7.1 Experimental Setup and Procedure

To empirically validate the proposed framework, we used a 100M-token excerpt of Dolma v1.6 (Soldaini et al., 2024) as $\mathcal{C}$ and performed separate experiments with OLMo-7B (Groeneveld et al., 2024) and Pythia-6.9B (Biderman et al., 2023) as LM. The former was trained on Dolma. We constructed SCs using Algorithm 1. See Appendix B for the parameter settings. For each constructed K, $K_{\text{Top-}k}$ with $k = 100$, $K_{\text{Top-}p}$ with $\alpha = 1 - r_{\max}^2$, and $\sigma(V)$, we computed the distribution of simplex dimensions and the (u, r) tradeoff over all of $\mathcal{C}$. For $K_{\text{Top-}p}$, we used r_{pw} in place of r to reflect the pointwise approximation error of Top-p.

7.2 Results

Figure 5 summarizes the four quantities d, m, r, and u. Table 3 reports the simplex counts and the minimum and maximum simplex dimensions.

Table 3: Comparison of m and dimension ranges.

	m	Min. dim.	Max. dim.
$\sigma(V)$	1	50,279	50,279
Top-k	103,804,427	99	99
Top-p	61,341,248	1	35,286
Ours	843	620	4,096

For the representative OLMo configuration with $r_{\max}^2 = 0.1$, $d_{\max} = 4096$, and $m_{\max} = 1024$ (see Appendix B), the constructed SC contains 843 simplexes of dimension at most 4096. Its r-neighborhood contains $1 - u = 95\%$ of the outputs for $r^2 = 0.1$. That is, while LM outputs lie in a probability simplex with ∼50k degrees of freedom, they can be adequately covered by a truncated simplicial complex whose simplexes have only 1–10% of the original simplex's degrees of freedom.

Robustness across cover parameters The compact-cover phenomenon persists across a range of parameter settings. At $r_{\max}^2 = 0.1$ and $d_{\max} = 4096$, increasing the simplex budget from $m_{\max} = 64$ to 1024 raises coverage from 85.9% to 95.0%, whereas an SC with $d_{\max} = 16384$ already achieves 99.2% coverage at $m_{\max} = 64$. Even under the stricter requirement $r_{\max}^2 = 0.01$, the SC with $d_{\max} = 16384$ and $m_{\max} = 1024$ covers 96.1% of outputs. The same qualitative tradeoff is observed for both models.

Comparison with the theoretical lower bound By Corollary 1.3, the Top-p column in Figure 5 gives a theoretical lower bound on the uncovered fraction. At $r_{\max}^2 = 0.1$ and $d_{\max} = 1024$, our SC with $m_{\max} = 1024$ achieves $u = 24.2\%$, compared with the lower bound of 2.7%. At $d_{\max} = 4096$, the gap narrows to 5.0% versus 0.2%. These results show that compact shared covers can achieve uncovered fractions close to the theoretical lower bound.

Qualitative examination Table 4 lists example tokens from selected simplexes in the SC constructed using OLMo with $r_{\max}^2 = 0.1$, $d_{\max} = 4{,}096$, $m_{\max} = 256$. The three subsets feature base-form verbs, third-person singular present-tense verbs, and abstract nouns associated with emotions and attitudes, respectively. These examples suggest that some selected supports align with recognizable linguistic patterns, although the construction itself uses only probability coverage and no grammatical or semantic labels.

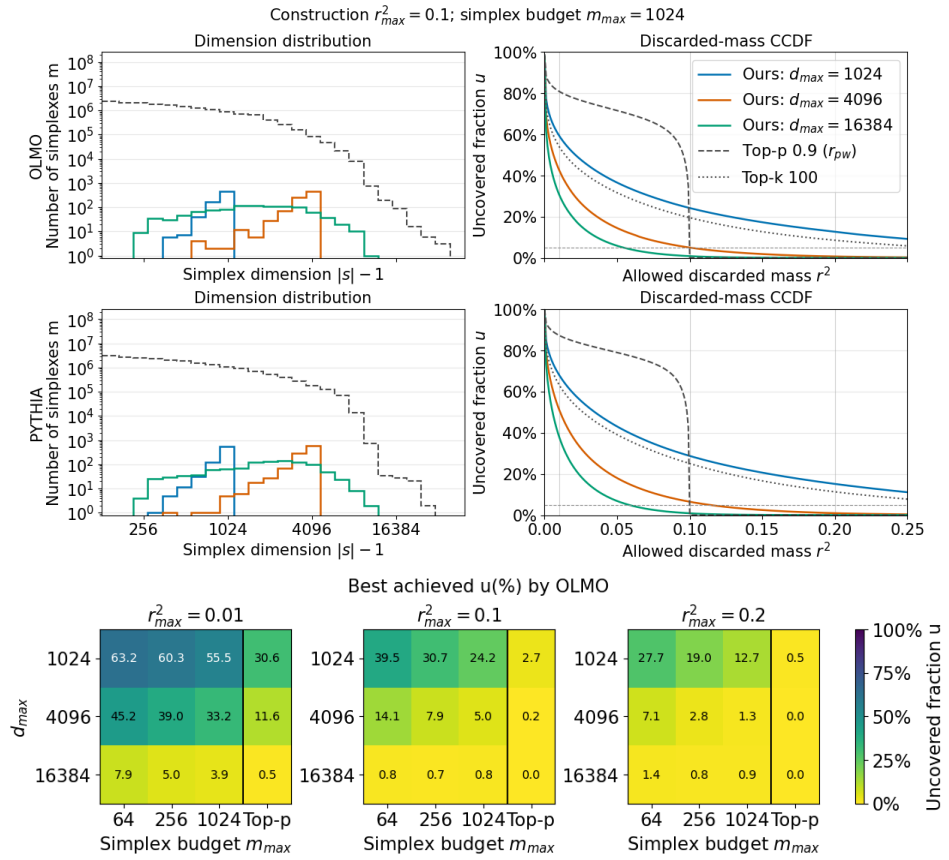


Figure 5: Experimental results. Top: distributions of simplex dimensions and the tradeoff between r and u. Our simplicial complex contains far fewer simplexes than the complexes induced by Top-k and Top-p. Bottom: achieved uncovered fraction u under various parameter settings.

8 Conclusion and Discussion

Our results support a simple picture of the LM output space: though the ambient probability simplex is extremely high-dimensional, outputs across contexts can be approximated by a relatively small family of lower-dimensional vocabulary simplexes. In this sense, probability concentration is not only a pointwise property of each prediction; substantial structure can be shared across contexts. Bhattacharyya geometry makes this picture quantitative. For a vocabulary subset, truncation to the corresponding simplex is the Bhattacharyya projection, and the squared dissimilarity is exactly the discarded probability mass. This allows the approximation of LM outputs to be described through four quantities (d, m, r, u). Top-k and Top-p arise as pointwise optimal choices in the (d, r) tradeoff, but collecting these simplexes produces very large complexes. Compact SCs can cover most evaluated outputs: 843 simplexes of dimension at most 4096 retain at least 90% of the probability mass for 95% of contexts. These simplexes have much lower dimension than the 50,279-dimensional $\sigma(V)$. It contains far fewer simplexes than Top-k ($\sim$104 million) or Top-p ($\sim$61 million) induce. Similar compactness persists across tested parameter settings. These results suggest that LM outputs have substantially lower effective width than their ambient dimension implies. Some selected supports suggest linguistic structure, warranting systematic analysis.

Our work has several limitations. First, our evaluation is limited to one corpus and two models. Whether similarly compact covers exist for other corpora, architectures, or larger models remains untested. Also, we have not yet examined how the geometry of hidden-state space relates to our output-space formulation through the output projection and softmax. An important direction for future work is to connect our framework with an active line of research on the continuous geometry of internal representations in LMs.

AI USE STATEMENT

In this work, we used generative AI tools for assist in the writing of proofs, implement methods, assist with translation, support qualitative and thematic data analysis. We have not used generative AI tools to help developing theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, clean and reformat dataset, generate synthetic data sets, propose or refine hypotheses, and design or provide feedback on research and methodology or experiments. Additionally, we used generative AI tools to create or modify scientific figures or images, draft parts of a research paper, summarize or analyse existing literature, discover research topics or identify gaps, brainstorming, sourcing/searching for information, edit a research paper to improve readability, identify relevant literature, format references, and suggest a structure for a research paper. We reviewed all AI-generated code, manuscript text, and suggested revisions before incorporating them into our work. The BibTeX bibliography file was prepared entirely by hand. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The main text describes the core experimental setup, while the appendix provides detailed parameter settings, experimental procedures, and computational resources. Our experiments use an existing corpus and pretrained language models, and the corresponding Hugging Face repository identifiers are listed in the appendix. Together, these details provide the information needed to reproduce our experiments.

REFERENCES

Shun-ichi Amari. *Information Geometry and Its Applications*, volume 194 of *Applied Mathematical Sciences*. Springer Japan, Tokyo, 2016. ISBN 978-4-431-55977-1 978-4-431-55978-8. doi: 10.1007/978-4-431-55978-8.

Georgios Arvanitidis, Miguel González-Duque, Alison Pouplin, Dimitrios Kalatzis, and Soren Hauberg. Pulling back information geometry. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pp. 4872–4894. PMLR, May 2022.

Ashwinkumar Badanidiyuru, Baharan Mirzasoleiman, Amin Karbasi, and Andreas Krause. Streaming submodular maximization: Massive data summarization on the fly. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pp. 671–680, New York, NY, USA, August 2014. Association for Computing Machinery. ISBN 978-1-4503-2956-9. doi: 10.1145/2623330.2623637.

Ronen Basri and David Jacobs. The softmax bottleneck does not limit the probabilities of the most likely tokens. In *The Fourteenth International Conference on Learning Representations*, 2026.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting Latent Predictions from Transformers with the Tuned Lens, November 2025.

A. Bhattacharyya. On a Measure of Divergence between Two Multinomial Populations. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 7(4):401–406, 1946. ISSN 0036-4452.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 2397–2430. PMLR, July 2023.

Andrea Bonito, Albert Cohen, Ronald Devore, Diane Guignard, Peter Jantsch, and Guergana Petrova. Nonlinear methods for model reduction. *ESAIM: Mathematical Modelling and Numerical Analysis*, 55(2):507–531, 2021. ISSN 2804-7214. doi: 10.1051/m2an/2020057.

Danilo Cavaliere, Sabrina Senatore, and Vincenzo Loia. Context-aware profiling of concepts from a semantic topological space. *Knowledge-Based Systems*, 130:102–115, August 2017. ISSN 0950-7051. doi: 10.1016/j.knosys.2017.05.008.

Haw-Shiuan Chang and Andrew McCallum. Softmax Bottleneck Makes Language Models Unable to Represent Multi-mode Word Distributions. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8048–8073, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.554.

Patrick Chen, Si Si, Sanjiv Kumar, Yang Li, and Cho-Jui Hsieh. Learning to Screen for Fast Softmax Inference on Large Vocabulary Neural Networks. In *International Conference on Learning Representations*, September 2018.

David Demeter, Gregory Kimmel, and Doug Downey. Stolen Probability: A Structural Weakness of Neural Language Models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2191–2197, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.198.

Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. American Mathematical Soc., 2010. ISBN 978-0-8218-4925-5.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not All Language Model Features Are One-Dimensionally Linear. In *The Thirteenth International Conference on Learning Representations*, October 2024.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical Neural Story Generation. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082.

Matthew Finlayson, Xiang Ren, and Swabha Swayamdipta. Logits of API-Protected LLMs Leak Proprietary Information. In *First Conference on Language Modeling*, August 2024.

Stephen Fitz. The Shape of Words - topological structure in natural language data. In *Proceedings of Topological, Algebraic, and Geometric Learning Workshops 2022*, pp. 116–123. PMLR, November 2022.

Nathan Godey, Éric Villemonte de la Clergerie, and Benoît Sagot. Why do small language models underperform? Studying Language Model Saturation via the Softmax Bottleneck. In *First Conference on Language Modeling*, August 2024.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. OLMo: Accelerating the Science of Language Models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15789–15809, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.841.

Wes Gurnee and Max Tegmark. Language Models Represent Space and Time. In *The Twelfth International Conference on Learning Representations*, October 2023.

Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 2002. ISBN 978-0-521-79540-1.

John Hewitt, Christopher Manning, and Percy Liang. Truncation Sampling as Language Model Desmoothing. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3414–3427, Abu Dhabi, United Arab Emirates, February 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.249.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The Curious Case of Neural Text Degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse Autoencoders Find Highly Interpretable Features in Language Models. In *The Twelfth International Conference on Learning Representations*, October 2023.

Xiaotong Ji, Rasul Tutunov, Matthieu Zimmer, and Haitham Bou-Ammar. Decoding as Optimisation on the Probability Simplex: From Top-K to Top-P (Nucleus) to Best-of-K Samplers, February 2026.

Dhruva Karkada, James Simon, Yasaman Bahri, and Michael Deweese. Closed-Form Training Dynamics Reveal Learned Features and Linear Structure in Word2Vec-like Models. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 3967–4000. Curran Associates, Inc., 2025. doi: 10.52202/085713-0140.

Ehsan Kazemi, Marko Mitrovic, Morteza Zadimoghaddam, Silvio Lattanzi, and Amin Karbasi. Submodular Streaming in All Its Glory: Tight Approximation, Minimum Memory and Low Adaptive Complexity. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 3311–3320. PMLR, May 2019.

Daniel Korchinski, Dhruva Karkada, Yasaman Bahri, and Matthieu Wyart. On the Emergence of Linear Analogies in Word Embeddings. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 62069–62097. Curran Associates, Inc., 2025. doi: 10.52202/085713-2075.

Omer Levy and Yoav Goldberg. Neural Word Embedding as Implicit Matrix Factorization. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Samuel Marks and Max Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. In *First Conference on Language Modeling*, August 2024.

Andre Martins and Ramon Astudillo. From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 1614–1623. PMLR, June 2016.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Yoshua Bengio and Yann LeCun (eds.), *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, December 1978. ISSN 1436-4646. doi: 10.1007/BF01588971.

Georgy Noarov, Soham Mallick, Tao Wang, Sunay Joshi, Yan Sun, Yangxinyu Xie, Mengxin Yu, and Edgar Dobriban. Foundations of Top-k Decoding for Language Models. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 116713–116756. Curran Associates, Inc., 2025. doi: 10.52202/085713-3890.

nostalgebraist. Interpreting GPT: The logit lens, August 2020.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The Geometry of Categorical and Hierarchical Concepts in Large Language Models. In *The Thirteenth International Conference on Learning Representations*, October 2024a.

Kiho Park, Yo Joong Choe, and Victor Veitch. The Linear Representation Hypothesis and the Geometry of Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 39643–39666. PMLR, July 2024b.

Kiho Park, Todd Nief, Yo Joong Choe, and Victor Veitch. The Information Geometry of Softmax: Probing and Steering. In *Forty-Third International Conference on Machine Learning*, June 2026.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162.

Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse Sequence-to-Sequence Models. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1504–1519, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1146.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.840.

V. N. Temlyakov. Nonlinear Kolmogorov widths. *Mathematical Notes*, 63(6):785–795, June 1998. ISSN 1573-8876. doi: 10.1007/BF02312773.

Wilhelm Tranheden, Shahnawaz Ahmed, Devdatt Dubhashi, Jonna Matthiesen, and Hannes von Essen. FlashHead: Efficient Drop-In Replacement for the Classification Head in Language Model Inference, March 2026.

Daniel Wurgaft, Can Rager, Matthew Kowal, Vasudev Shyam, Sheridan Feucht, Usha Bhalla, Tal Haklay, Eric Bigelow, Raphael Sarfati, Thomas McGrath, Owen Lewis, Jack Merullo, Noah Goodman, Thomas Fel, Atticus Geiger, and Ekdeep Singh Lubana. Manifold Steering Reveals the Shared Geometry of Neural Network Representation and Behavior, May 2026.

Bo Xiong. The Lattice Representation Hypothesis of Large Language Models. In *The Fourteenth International Conference on Learning Representations*, October 2025.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. Breaking the Softmax Bottleneck: A High-Rank RNN Language Model. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Jinbin Zhang, Nasib Ullah, Erik Schultheis, and Rohit Babbar. DynaSpec: Context-aware Dynamic Speculative Sampling for Large-Vocabulary Language Models, February 2026.

Haiyan Zhao, Heng Zhao, Bo Shen, Ali Payani, Fan Yang, and Mengnan Du. Beyond Single Concept Vector: Modeling Concept Subspace in LLMs with Gaussian Distribution. In *The Thirteenth International Conference on Learning Representations*, October 2024.

A PROOFS

A.1 PROOF OF THEOREM 1

Since $M \subseteq \sigma(V)$ is compact and $q \mapsto \cos_B(p, q)$ is continuous, the maximum defining $\cos_B(p, M)$ exists. For nonempty s, let $Z = \sum_{i \in s} p_i > 0$. For any $q \in \sigma(s)$, Cauchy–Schwarz gives

$$\cos_B^2(p, q) = \left(\sum_{i \in s} \sqrt{p_i q_i} \right)^2 \leq \left(\sum_{i \in s} p_i \right) \left(\sum_{i \in s} q_i \right) = Z.$$

Equality holds precisely when $q_i = p_i / Z$ on s, with $q_i = 0$ outside s, so the unique maximizer is $T_s(p)$. Thus $\cos_B^2(p, \sigma(s)) = Z$ and $\sin_B^2(p, \sigma(s)) = 1 - Z = \sum_{i \notin s} p_i$. □

A.2 PROOF OF PROPOSITION 1

Let $A_s = \{c \in C : \mathrm{LM}(c) \in N_r(\sigma(s))\}$ and $U_\Sigma = \bigcup_{s \in \Sigma} A_s$. Since distance to a finite union is the minimum distance to its members, $\psi_r(\Sigma) = |U_\Sigma| / |C|$. For $\Sigma \subseteq \Gamma$,

$$\psi_r(\Sigma \cup \{s\}) - \psi_r(\Sigma) = \frac{|A_s \setminus U_\Sigma|}{|C|} \geq \frac{|A_s \setminus U_\Gamma|}{|C|} = \psi_r(\Gamma \cup \{s\}) - \psi_r(\Gamma) \geq 0.$$

This proves monotonicity and diminishing returns; also $\psi_r(\varnothing) = 0$. □

A.3 APPROXIMATION GUARANTEE

Let $m_{\max}$ be the simplex budget in Algorithm 1, f the normalized coverage at radius $r_{\max}$ on W fixed witnesses, and $\mathcal{S}$ the stream candidates satisfying the dimension bounds. Write $\mathrm{OPT} = \max_{\Sigma \subseteq \mathcal{S},\, |\Sigma| \leq m_{\max}} f(\Sigma)$. Initialize Algorithm 1 with $K_{\mathrm{rand}} \subseteq \mathcal{S}$, $|K_{\mathrm{rand}}| \leq m_{\max}$, and

$$\tau_{\min} = \max \left\{ \frac{f(K_{\mathrm{rand}})}{2}, \frac{1}{2W} \right\}.$$

For $\epsilon > 0$, the returned solution has coverage at least $\mathrm{OPT}/(2(1+\epsilon)) \geq (1/2 - \epsilon)\mathrm{OPT}$.

Proof. The case $\mathrm{OPT} = 0$ is immediate. Otherwise, coverage is a multiple of $1/W$, so $\mathrm{OPT} \geq 1/W$. Since $f(K_{\mathrm{rand}}) \leq \mathrm{OPT}$, $0 < \tau_{\min} \leq \mathrm{OPT}/2 \leq 1/2$. The geometric grid therefore contains a largest threshold $\tau \leq \mathrm{OPT}/2$, which satisfies

$$\frac{\mathrm{OPT}}{2(1+\epsilon)} \leq \tau \leq \frac{\mathrm{OPT}}{2}. \tag{5}$$

Let S_j be its solution after j acceptances. The acceptance rule and induction give $f(S_j) \geq j\tau / m_{\max}$, since $f(S_{j+1}) \geq f(S_j) + (\tau - f(S_j))/(m_{\max} - j) \geq (j+1)\tau / m_{\max}$. If the budget fills, $f(S_{m_{\max}}) \geq \tau$. Otherwise, suppose the final set S satisfies $f(S) < \tau$. Whenever a batch is rejected, all remaining candidates have marginal gain at most $(\tau - f(S_j))/(m_{\max} - j) \leq \tau / m_{\max}$. Submodularity preserves this bound as the solution grows. Hence, for an optimal set S^*,

$$\mathrm{OPT} \leq f(S \cup S^*) \leq f(S) + \sum_{s \in S^* \setminus S} \big(f(S \cup \{s\}) - f(S) \big) \leq f(S) + \tau < 2\tau \leq \mathrm{OPT},$$

a contradiction. Thus $f(S) \geq \tau$, and returning the best threshold solution gives the bound. □

This is the Sieve-Streaming argument of Badanidiyuru et al. (2014), with target value 2τ.

Due to implementation errors, our experiments did not enforce the $1/(2W)$ clipping rule or ensure $K_{\mathrm{rand}} \subseteq \mathcal{S}$ for the filtered candidate stream. The resulting coverage may therefore fall below the bound above. We never encountered $\tau_{\min} = 0$ in our experiments.

B EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

B.1 MODELS AND DATA

We downloaded OLMo-7B and Pythia-6.9B and their tokenizers from Hugging Face, using `allenai/OLMo-7B-hf` and `EleutherAI/pythia-6.9b-deduped`, respectively. We also downloaded the Dolma sample URL list, `urls/v1_6-sample.txt`, from the Hugging Face dataset repository `allenai/dolma`. We sampled one eighth of the documents uniformly without replacement and used the same sampled documents for both models, each with its own tokenizer.

Each document is processed in windows of at most 2,048 tokens with stride 512. We exclude padding and the first 1,536 positions of each subsequent window to avoid double counting. We evaluate the first $N = 104{,}857{,}600$ retained output positions per model, using vocabularies of 50,280 tokens for OLMo and 50,277 for Pythia.

B.2 CONSTRUCTION PARAMETERS

The grid comprises 54 configurations: two models, three construction radii $r_{\max}$, three dimension limits $d_{\max}$, and three simplex budgets $m_{\max}$ (Table 5). The Top-p threshold is $\alpha = 1 - r_{\max}^2$. Candidates satisfy $256 \le |s| - 1 \le d_{\max}$. The OLMo result in the upper panels of Figure 5 and Table 3 uses $r_{\max}^2 = 0.1$, $d_{\max} = 4{,}096$, and $m_{\max} = 1{,}024$. Each grid entry reports one construction run.

Parameter	Value
Squared construction radius $r_{\max}^2$	0.2, 0.1, 0.01
Maximum simplex dimension $d_{\max}$	1,024, 4,096, 16,384
Simplex budget $m_{\max}$	64, 256, 1,024
Minimum simplex dimension $d_{\min}$	256
Witness sample size W	524,288
Construction batch size	131,072
Threshold-grid parameter ϵ	0.01
Construction random seed	2732

Table 5: Construction settings. The realized simplex count m can be smaller than $m_{\max}$.

B.3 CONSTRUCTION PROCEDURE

Witnesses and candidates. We sample W output positions uniformly without replacement and fix their distributions as the witness set. We process candidates in corpus order using the batch size and dimension bounds in Table 5.

Threshold grid. We sample $20m_{\max}$ positions without replacement from the first 2,097,152 positions and take the $m_{\max}$ largest Top-p supports, without applying the dimension filter. Let b_0 be their joint witness coverage. Starting from $\tau_0 = b_0/2 + 10^{-6}$, we multiply by $1 + \epsilon$ until the current value is at least $0.5/(1 + \epsilon)$, then append

$$1 - 2^{-(1+j/3)}, \qquad j = 1, \ldots, 18.$$

Selection and stopping. We use the batched selection rule of Algorithm 1. A threshold becomes inactive when its budget is exhausted or its value is no greater than the best current witness coverage. The pass ends when no thresholds remain active.

B.4 Evaluation and counting conventions

For $K = \bigcup_{s\in\Sigma} \sigma(s)$ and output distribution p_c, we compute

$$q_c(K) = \max_{s\in\Sigma} \sum_{i\in s} p_{c,i}, \tag{6}$$

$$e_c(K) = 1 - q_c(K), \tag{7}$$

$$u_K(r) = \frac{1}{N} \sum_{c=1}^{N} \mathbf{1}\{e_c(K) > r^2\}. \tag{8}$$

Coverage is $1 - u_K(r)$. Construction uses retained mass $> \alpha$; evaluation includes equality, using $q_c \geq \alpha$ at $r = r_{\max}$.

We select the solution with the highest final witness coverage and evaluate it on the same N output positions used for construction. The reported scores therefore describe coverage of that corpus. The count m refers to selected generating simplexes for our method and distinct supports for the pointwise baselines. Dimension histograms count each distinct support once.

For the error curves in Figure 5, Top-k uses $k = 100$ and Top-p uses $\alpha = 0.9$. Top-p errors are measured against each output's own support, as specified by r_{pw} in the main text; these need not equal distances to the union. The two errors coincide for Top-k.

Pointwise lower bound in the heatmaps. Let $k_\alpha(c)$ be the minimum number of tokens needed to retain mass at least α at position c. Corollary 1.3 gives the lower bound, without a simplex-count constraint,

$$u_K(r_{\max}) \geq \frac{1}{N} \sum_{c=1}^{N} \mathbf{1}\{k_\alpha(c) > d_{\max} + 1\} \qquad (\dim K \leq d_{\max}).$$

The Top-p columns report this bound by counting pointwise supports with more than $d_{\max} + 1$ tokens. They use the same maximum simplex dimension as construction, without a constraint on the number of simplexes.

B.5 Computational resources

Hidden-state extraction used an NVIDIA A100 GPU with 40 GB of memory. Construction and coverage evaluation use 152 virtual CPUs on Intel Xeon Platinum 8368 processors (2.4 GHz), with approximately 230 GB of allocated RAM. For the main OLMo configuration, the streaming construction phase took 2.39 hours and full-corpus evaluation took 2.18 hours, excluding model inference and preparatory computations.

B.6 RESULTS ACROSS CONFIGURATIONS

Table 6 reports coverage across the grid; Figure 6 shows the Pythia results. Results are available for 45 of the 54 configurations; the nine Pythia configurations with $r^2_{\max} = 0.01$ are not reported. For Pythia at $r^2_{\max} = 0.1$ and $d_{\max} = 4{,}096$, increasing the budget from 64 to 1,024 raises coverage from 81.7% to 93.5%.

		OLMo			Pythia		
$d_{\max}$	$m_{\max}$	$r^2_{\max} = .2$	.1	.01	$r^2_{\max} = .2$	.1	.01
1,024	64	0.7228	0.6052	0.3683	0.6938	0.5698	–
1,024	256	0.8104	0.6931	0.3974	0.7912	0.6504	–
1,024	1,024	0.8729	0.7580	0.4446	0.8418	0.7127	–
4,096	64	0.9293	0.8593	0.5483	0.9006	0.8171	–
4,096	256	0.9717	0.9207	0.6100	0.9497	0.8985	–
4,096	1,024	0.9871	0.9502	0.6683	0.9787	0.9353	–
16,384	64	0.9865	0.9918	0.9213	0.9620	0.9694	–
16,384	256	0.9918	0.9926	0.9497	0.9873	0.9890	–
16,384	1,024	0.9913	0.9917	0.9612	0.9914	0.9914	–

Table 6: Coverage at $r = r_{\max}$. Values are rounded to four decimal places; – denotes unavailable results.

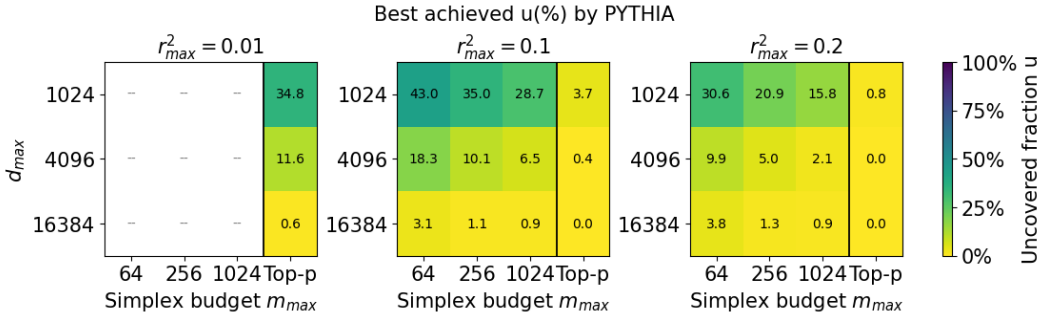


Figure 6: Pythia results across simplex budgets and dimension limits. Cell labels and colors show the uncovered fraction $u_K(r_{\max})$ (%). The Top-p column gives the pointwise lower bound at the same dimension limit, without a simplex-count constraint (Section B.4). Dashes indicate unavailable results.

B.7 TRADEOFF CURVES

Figures 7–12 show dimension distributions (left) and discarded-mass CCDFs (right). Each figure fixes the model and construction radius $r_{\max}$; rows show budgets $m_{\max} = 64$, 256, and 1,024. The constructed complexes use $d_{\max} = 1{,}024$, 4,096, or 16,384. In the right panels, dashed curves show Top-p pointwise errors at $\alpha = 1 - r^2_{\max}$, and dotted curves show Top-k with $k = 100$. The Top-p legend values denote $1 - \alpha$.

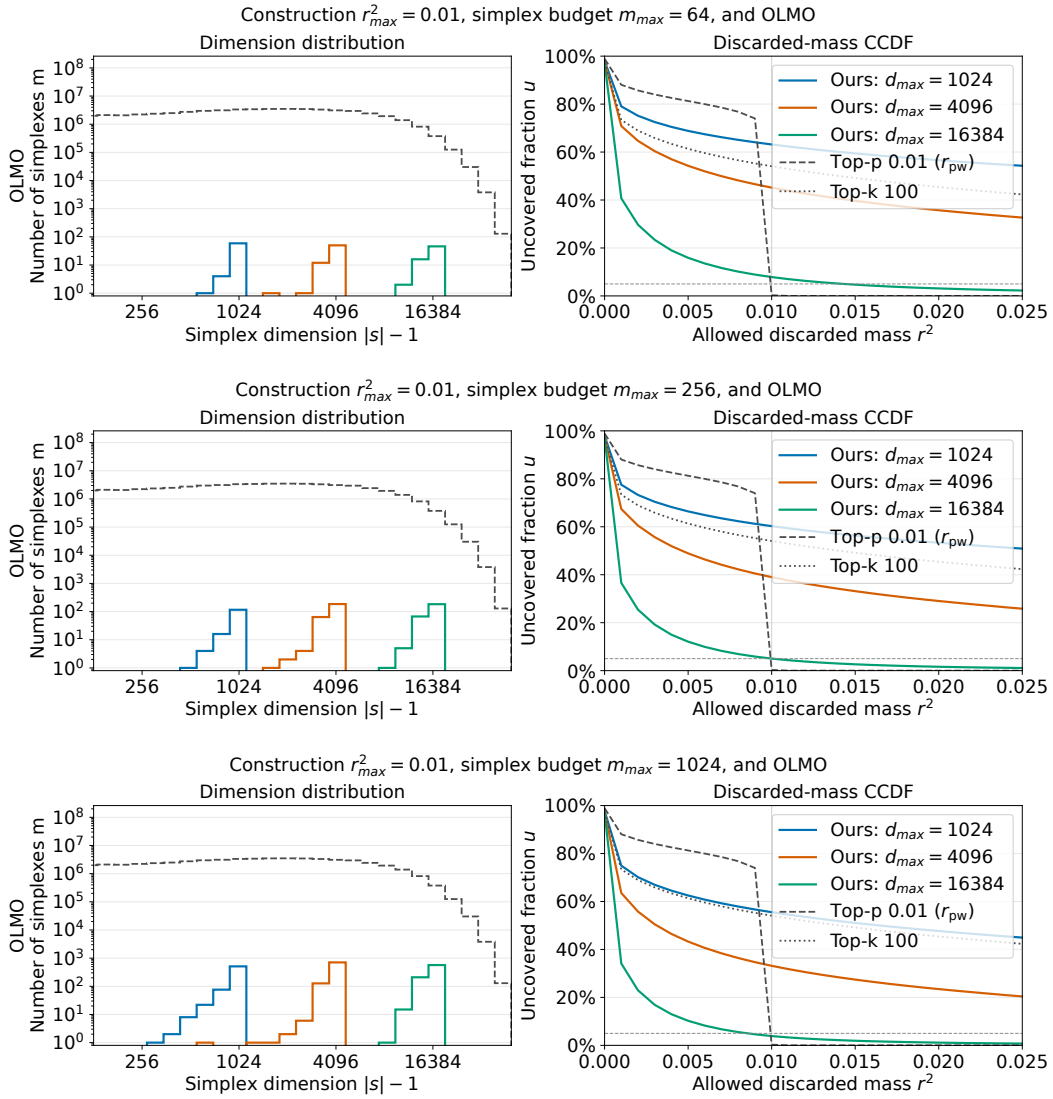


Figure 7: OLMo at construction radius $r^2_{\max} = 0.01$. Rows show simplex budgets $m_{\max} = 64$, 256, and 1,024 from top to bottom.

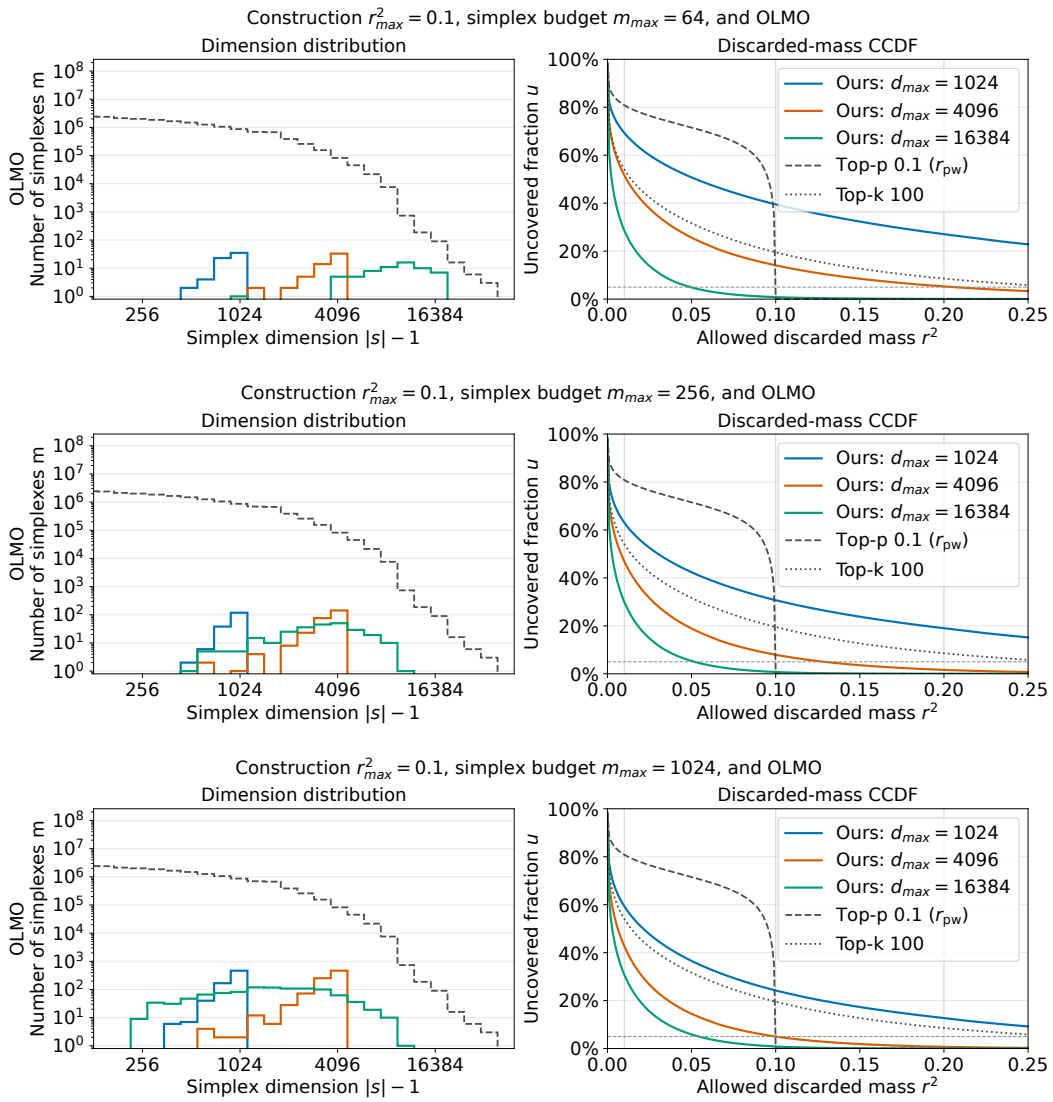


Figure 8: OLMo at construction radius $r^2_{\max} = 0.1$. Rows show simplex budgets $m_{\max} = 64$, 256, and 1,024 from top to bottom.

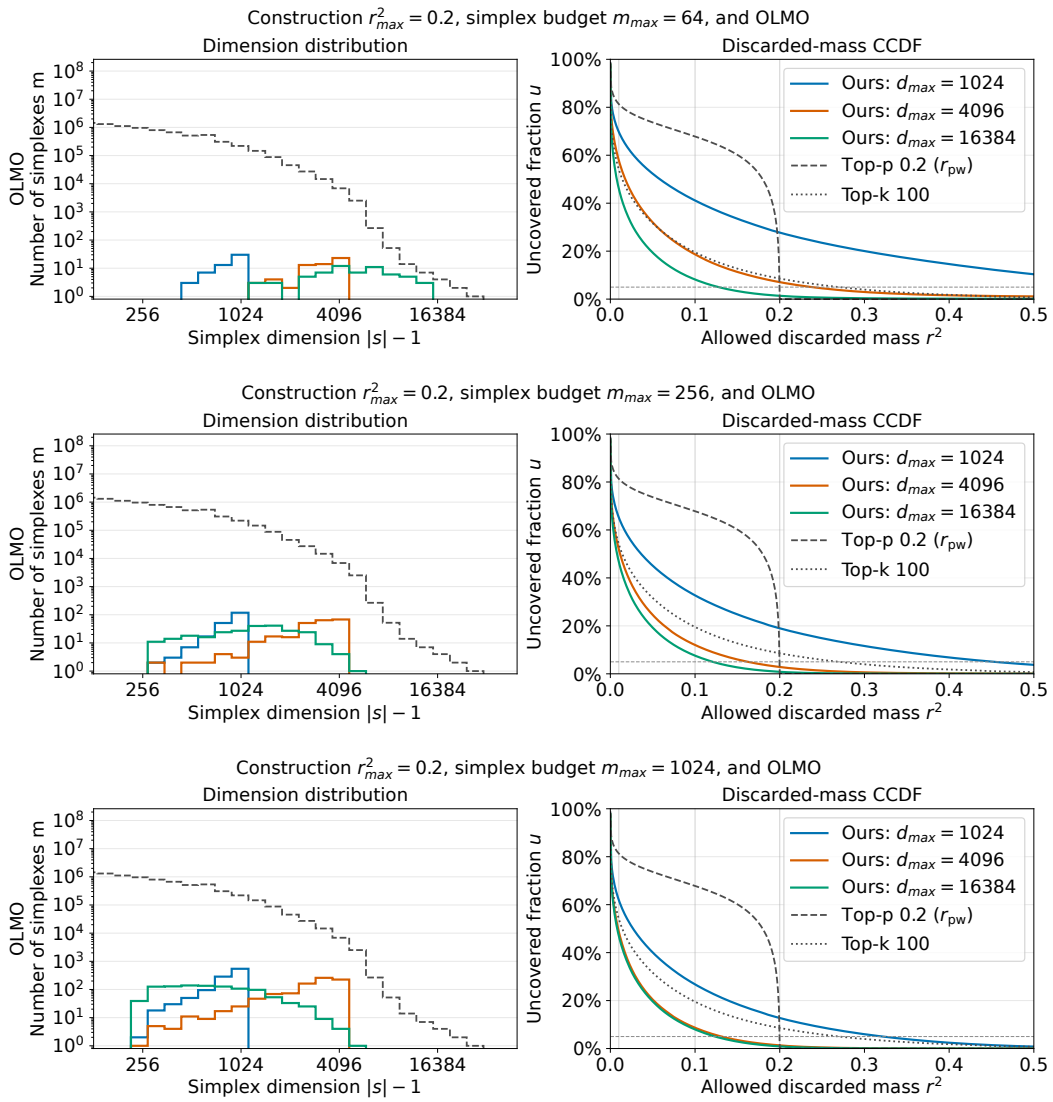


Figure 9: OLMo at construction radius $r^2_{\max} = 0.2$. Rows show simplex budgets $m_{\max} = 64$, 256, and 1,024 from top to bottom.

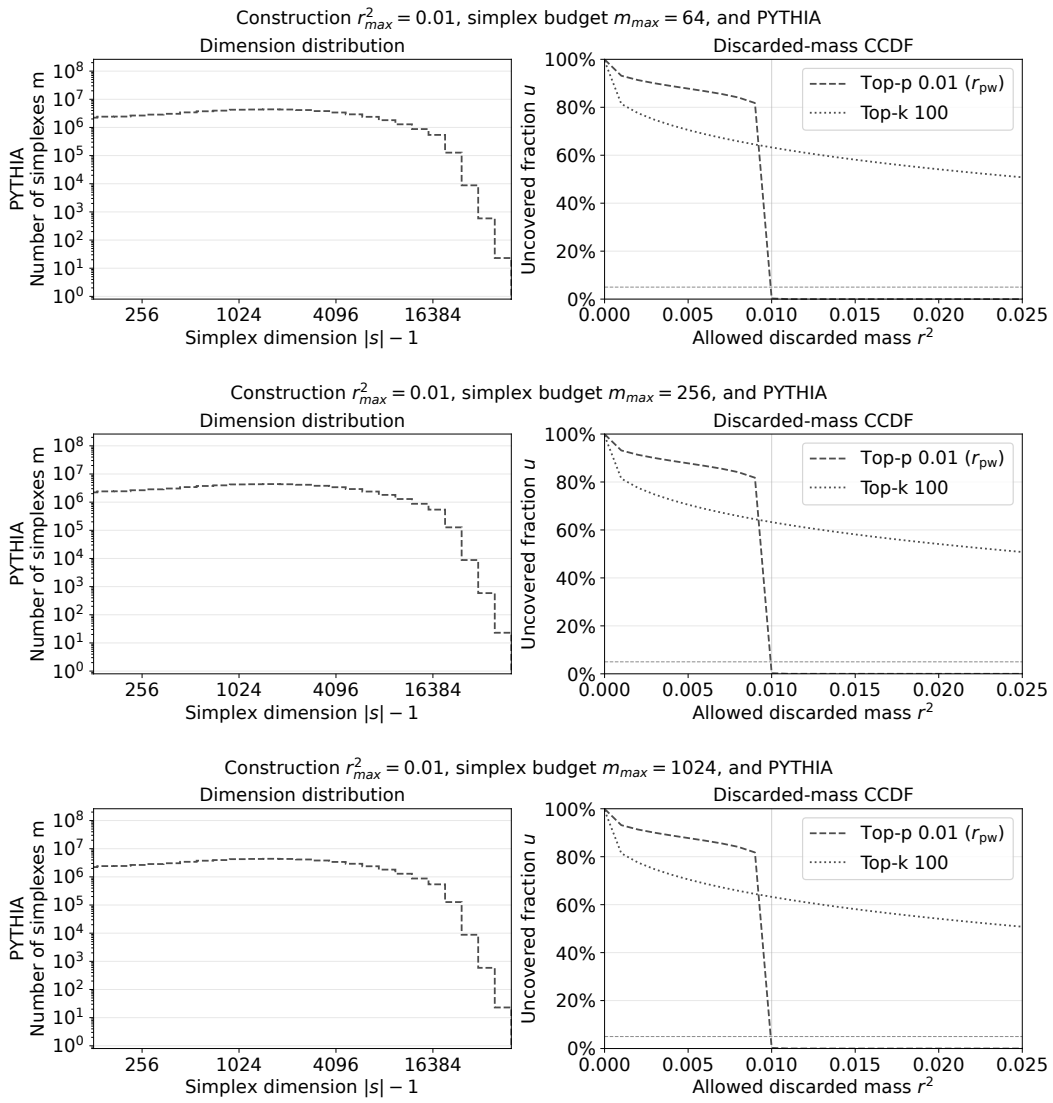


Figure 10: Pythia at construction radius $r^2_{\max} = 0.01$. Rows show simplex budgets $m_{\max} = 64$, 256, and 1,024 from top to bottom. Only pointwise baselines are available for these settings.

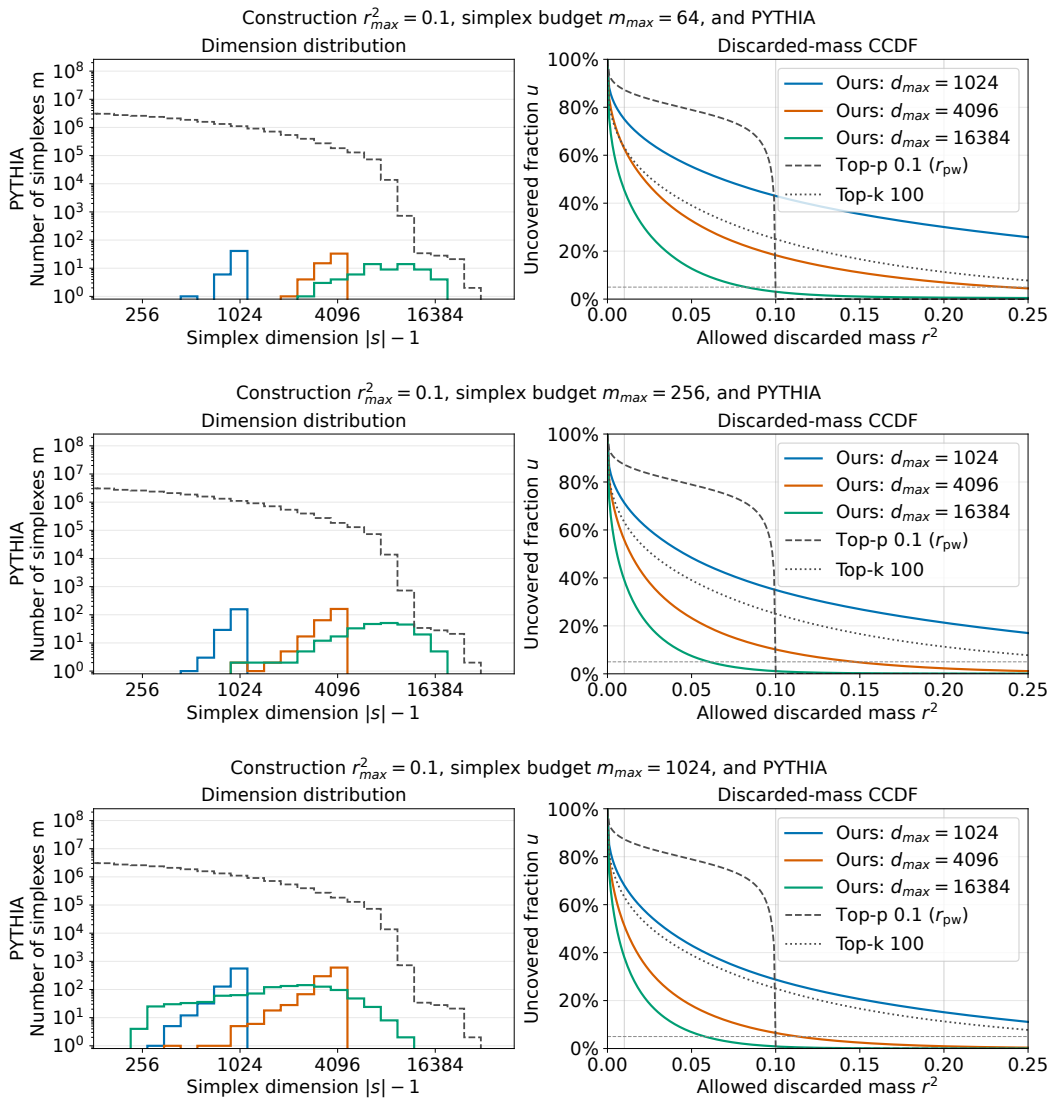


Figure 11: Pythia at construction radius $r^2_{\max} = 0.1$. Rows show simplex budgets $m_{\max} = 64$, 256, and 1,024 from top to bottom.

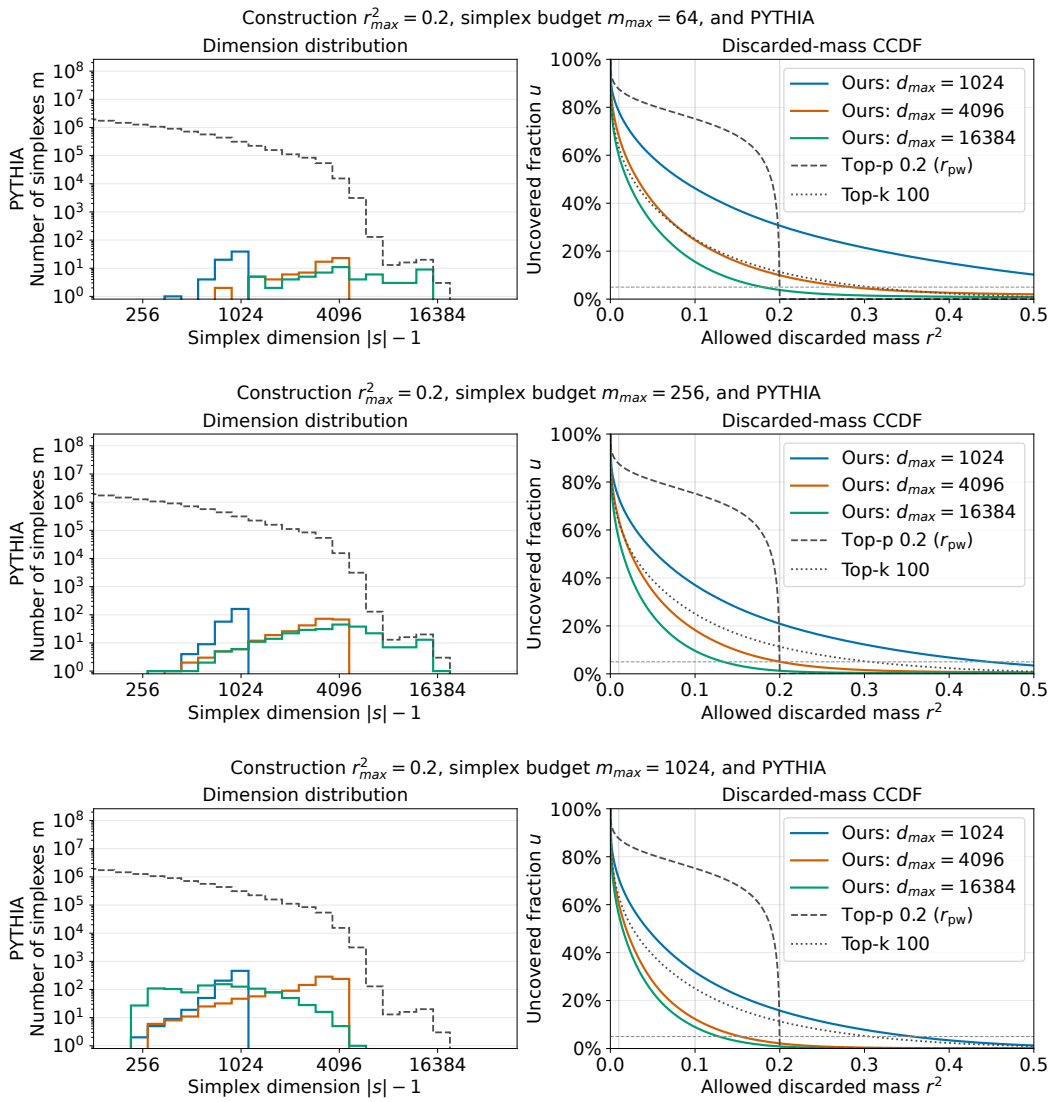


Figure 12: Pythia at construction radius $r^2_{\max} = 0.2$. Rows show simplex budgets $m_{\max} = 64$, 256, and 1,024 from top to bottom.

B.8 QUALITATIVE EXAMPLES

The qualitative examples use OLMo with $r^2_{\max} = 0.1$, $d_{\max} = 4{,}096$, and $m_{\max} = 256$. We retain tokens occurring in fewer than eight generating simplexes and more than 1,024 times in the tokenized document sample, then rank them by decreasing unigram frequency within each simplex. We manually select lists with recognizable lexical or semantic patterns; Table 4 presents short excerpts, with descriptive category labels. This filtering is used only to select examples and does not change the complex.

Below we show up to 20 tokens from each of the 256 filtered lists. The three lists excerpted in Table 4 (072, 085, and 233) are bold.

B.8.1 FIRST 20 TOKENS FROM EACH FILTERED LIST

List 000 (20 tokens)
['mith', 'Writer', '_\u0430', 'keeper', 'ellers', 'heets', 'Listener', 'cript', 'creen', 'chool', '(){', '_Associates', '_Studios', '_Programme', '_f\xfcr', '_Networks', '**.', '/{', '_Maker', 'Events']

List 001 (20 tokens)
['ively', 'ographic', 'uation', 'iveness', 'uate', '_thereof', 'liness', 'eur', '_compartment', '_opener', 'edly', '_versa', 'uality', '_maneuver', '_fret', '_uh', 'acion', '_alley', '_amidst', '_holdings']

List 002 (20 tokens)
['LINE', 'agra', '%,', 'umm', '%)', 'acks', 'earing', 'ATION', 'aming', 'ERS', '\u20ac', 'ags', 'ACK', 'IVE', 'umps', '%;', 'ummy', 'aked', 'arks', 'acing']

List 003 (12 tokens)
['_mortgages', '_caregivers', '_wastewater', '_irrespective', '_todd', '_Australians', '_cyclists', '_lasers', '_rodents', '_prenatal', '_newborns', '_EVs']

List 004 (20 tokens)
['///', '/*', 'LINE', 'tmp', 'define', '});', 'summary', 'github', 'png', 'choice', 'header', 'margin', '\r\n\t\t', '_serialized', 'bool', 'description', 'including', 'padding', 'template', 'css']

List 005 (20 tokens)
['_asleep', '_remotely', '_unsure', '_stirring', '_mindful', '_enthusi', '_unaware', '_unchanged', '_payable', '_minded', '_hereby', '_expressly', '_identifiable', '_reliably', '_unreasonable', '_posing', '_progressively', '_achievable', '_kidding', '_imminent']

List 006 (20 tokens)
['According', '_Besides', '_Nations', '_Hence', '_Nevertheless', '_Unlike', '_Meanwhile', '_Especially', '_Either', '_Obviously', '_Unless', '_Hopefully', '_Airlines', '_Probably', '_Typically', '_Basically', '_Whenever', '_Likewise', '_Neither', '_Representative']

List 007 (9 tokens)
['_Month', '_arises', '_reson', '_resembles', '_entails', '_encompasses', '_depicts', '_warrants', '_theorem']

List 008 (20 tokens)
['/*', 'However', 'According', '|||EMAIL_ADDRESS|||', '‘‘‘‘', 'github', 'Another', 'Although', 'header', 'msgid', 'During', '"""', 'Because', '<!--', 'description', 'template', '###', 'Whether', 'Hey', 'Despite']

List 009 (20 tokens)
['_realise', '_whatsoever', '_disclose', '_recognise', '_streak', '_swear', '_diligence', '_payable', '_tolerate', '_propriet', '_minded', '_incur', '_resemble', '_fret', '_amend', '_holdings', '_spur', '_yourselves', '_behold', '_recept']

List 010 (20 tokens)
['_brighter', '_starters', '_tougher', '_medals', '_nominations', '_pillars', '_landmarks', '_sacrifices', '_rifles', '_offenses', '_minist', '_nicer', '_spectators', '_tast', '_senators', '_poorer', '_chicks', '_happ', '_benchmarks', '_shootings']

List 011 (20 tokens)
['ively', 'eless', 'ulation', 'ories', 'lessly', 'akers', 'lers', 'door', 'ulate', 'bers', 'sters', 'lier', 'yard', 'atives', 'uation', 'rated', '_Wars', 'scape', 'backs', 'boards']

List 012 (20 tokens)
["'re", "'ve", '_hesitate', 'MAX', '_realise', '_wanna', '_gotta', '_owe', '_apologize', '_HAVE', 'gor', '_swear', '_hereby', '_NEVER', '_NEED', 'Tube', '_REALLY', '_KNOW', '_presume', 'brahim']

List 013 (20 tokens)
['\u2026..', '_Chevrolet', '\u2026\u2026', '_Election', '_McD', '_Reality', '_PBS', '_Chaos', '_Bean', '\u2026\u201d', '_uh', '_Scient', '_Saturn', '_Mond', '_Comedy', '_Avengers', '_Wik', '_blah', '_Tay', '_Nixon']

List 014 (20 tokens)
['izations', 'istically', '_footprint', '_grandfather', '_IRA', 'isms', '_Philosophy', '_expenditures', '_counselor', '_victories', '_standpoint', '_incomes', '_nominee', '_pursu', '_nominations', '_r\xe9', '_fantasies', '_deduction', '_accountant', '_holdings']

List 015 (20 tokens)
['pper', 'pped', 'Override', 'ppers', 'Assert', 'Writer', '(*', 'Signature', '_complain', 'asts', '_donate', 'Changed', 'AST', '_realise', 'Pointer', 'Panel', 'Listener', '_disclose', 'Bytes', "('/"]

List 016 (20 tokens)
['inos', 'agra', 'rieved', 'acks', 'eless', 'ulation', 'ags', 'urer', 'ologist', 'icing', 'aving', 'ifications', 'ulate', 'ographic', 'rees', 'acing', 'illed', 'oids', 'Override', 'ologists']

List 017 (20 tokens)
['_footprint', '_slices', '_bulb', '_cutter', '_costing', '_fuse', '_tracker', '_fret', '_northeast', '_staircase', '_fingertips', '_rollers', '_tempo', '_batting', '_knob', '_throttle', '_turbine', '_slab', '_axle', '_rele']

List 018 (20 tokens)
['_remotely', '_remod', '_sourced', '_aus', '_periodically', '_rented', '_unchanged', '_tempted', '_analysed', '_mapped', '_expressly', '_encoded', '_emailed', '_inspected', '_undergone', '_postponed', '_digitally', '_reliably', '_staged', '_anonym']

List 019 (20 tokens)
['/*', '3333', '});', '%,', '%)', '\u3001', '%;', '>\\', '408', ';</', '\u2033', '="-', '_____________________', '166', '999', '755', '\xd7', '/>', 'byte', '_sq']

List 020 (20 tokens)
['pect', 'verse', 'forming', 'cludes', 'dependent', 'ception', 'lined', 'forms', 'effective', 'cluded', 'formed', '_\xa7', '_spite', 'cluding', 'fol', 'formation', 'comes', 'ceptions', 'scription', '_vitro']

List 021 (20 tokens)
['According', 'During', '_Besides', 'Having', '_Authority', '_Hence', '_\ufffd', '_Unlike', 'Through', '_Either', '\ufffd\ufffd', '_Unless', 'Overall', '_Directors', 'Students', '_Degree', 'Choose', '_Typically', '_Duty', '_Alternatively']

List 022 (20 tokens)
['_Infrastructure', '_expenditures', '_rehears', '_mortgages', '_broadcasting', '_ceremonies', '_briefing', '_probation', '_athlet', '_preaching', '_transporting', '_evangel', '_lia', '_steward', '_hosp', '_zoning', '_admitting', '_troop', '_visas', '_enrol']

List 023 (20 tokens)
['_Singh', '_Potter', '_Cruz', '_Lodge', 'stown', '_Hopkins', '_Bros', '_Holder', '_Bach', '_Wei', '_Huang', '_Powell', '_Zhao', '_Garcia', '_McD', '_Clarke', '_Lynch', '_Carr', '_Rodriguez', '_Sullivan']

List 024 (19 tokens)
['_Fortune', '_dividends', '_CAG', '_tougher', '_insurer', '_accountant', '_supervisors', '_tighter', '_alarming', '_bargaining', '_vacancy', '_backlash', '_happ', '_looming', '_faire', '_bankers', '_FAA', '_recru', '_evalu']

List 025 (20 tokens)
['inos', 'agra', 'inate', 'ved', 'eless', 'ously', 'ving', 'lette', 'ifies', 'ifications', 'bers', 'ading', 'sters', 'iously', 'inated', 'aning', 'cing', 'acing', 'ples', 'oses']

List 026 (20 tokens)
['friendly', 'choice', 'door', 'jack', 'depth', 'writer', 'winning', 'selling', 'profit', 'boards', 'pages', 'task', 'scope', 'overs', 'resources', 'document', 'lights', 'breaking', 'forms', 'keeping']

List 027 (20 tokens)
['_gotta', '_risen', '_bothered', '_tempted', '_clicked', '_mapped', '_wander', '_undergone', '_paved', '_staged', '_wisely', '_photographed', '_lied', '_voluntarily', '_grapp', '_messed', '_consulted', '_silently', '_hacked', '_infiltr']

List 028 (20 tokens)
['_bake', '_stroll', '_remod', '_booster', '_cinnamon', '_litter', '_brighter', '_gallon', '_grilled', '_LOT', '_volley', '_chili', '_pistol', '_whip', '_Wii', '_trilogy', '_commemor', '_slab', '_sprinkle', '_tandem']

List 029 (20 tokens)
['###', '333', '408', 'ESS', '166', '384', '304', '345', '_DATE', '327', '329', '203', '176', '234', '378', '174', '334', '324', '359', '186']

List 030 (20 tokens)
['_Hampshire', '_coordinator', '_IRA', '_Romania', '_Rangers', '_Bears', '_Emirates', '_Croatia', '_Chand', '_Defence', '_propriet', '_Gaza', '_clerk', '_striker', '_Alban', '_QB', '_Sloven', '_Bulgaria', '_entails', '_MVP']

List 031 (20 tokens)
['LINE', '\ufffd', 'However', 'According', '\ufffd', '\ufffd', '\ufffd', '\ufffd', 'Although', '\ufffd', '\ufffd', 'ATION', 'ERS', 'Because', 'lers', 'ACK', '\u662f', 'IVE', '\ufffd', 'iously']

List 032 (20 tokens)
['_slices', '_burner', 'iates', '_leagues', '_barrels', '_crews', '_noodles', '_packets', '_pans', 'iator', '_terminals', '_receivers', '_coatings', '_consoles', '_rollers', '_cubes', '_strands', '_detectors', '_gestures', '_batches']

List 033 (20 tokens)
['_skiing', '_surfing', '_Chaos', '_injustice', '_advers', '_Scient', '_resurrection', '_adulthood', '_madness', '_Avengers', '_envy', '_PTSD', '_miracles', '_Buddhism', '_surreal', '_Dragons', '_pornography', '_innocence', '_Vamp', '_humility']

List 034 (20 tokens)
['agra', 'pper', 'acks', 'earing', 'aming', 'pped', 'akers', 'ags', 'inding', 'css', 'msg', 'aving', 'umps', 'inners', 'ading', '>\\', 'aked', 'iders', 'aping', 'ctx']

List 035 (20 tokens)
['_noon', '_Grove', '_ceil', '_showers', '_Lodge', '_canal', '_baths', '_Plaza', '_skiing', '_basin', '_altitude', '_Rangers', '_Peak', '_halls', '_soils', '_crews', '_buds', '_commute', '_turf', '_crawl']

List 036 (20 tokens)
['408', '166', '999', '755', '384', '304', '345', '487', '707', '2002', '327', '017', '329', '387', '203', '176', '234', '448', '378', '174']

List 037 (20 tokens)
['_stirring', '_albeit', '_molding', '_sans', '_northeast', '_sew', '_spanning', '_noticing', '_reminding', '_posing', '_whist', '_eh', '_silently', '_scratching', '_depicting', '_Sherman', '_prompting', '_painters', '_implying', '_horizontally']

List 038 (20 tokens)
['jack', '_lb', '_tweeted', '_clerk', '_sue', '_thanked', '_chambers', '_sie', '_hav', '_todd', '_testified', '_hath', '_seg', '_slammed', '_pow', '_urges', '_raced', '_fortun', '_pars', '_graves']

List 039 (20 tokens)
['_cinnamon', '_theirs', '_barrels', '_carrots', '_QB', '_crap', '_rubbish', '_Wii', '_ether', '_ETH', '_LR', '_CLI', '_PST', '_Myth', '_ole', '_Gn', '_Mys', '_shotgun', '_buckets', '_SSH']

List 040 (20 tokens)
['_Authority', '_Wars', '_Springs', '_Islands', 'dale', 'keeper', '_Airlines', '_Industries', '_Rd', '_Corps', 'flower', '_Associates', '_Grove', '_Studios', '_Foods', '_Mountains', '_Heights', '_Forces', '_Surgery', '_Theater']

List 041 (20 tokens)
['agra', 'ving', 'aming', 'inity', 'hour', 'ologist', 'icing', 'msg', 'tered', 'door', 'sized', 'ctions', 'crete', 'uous', 'cing', 'yard', 'ancement', 'atives', 'Override', 'ologists']

List 042 (20 tokens)
['\u7684', '});', '"""', '>\\', ';</', 'Override', '="#', '_____', '/>', '_____', 'Writer', '_btn', '="{{', '_?>', '_%}', 'Signature', '_;', '="/', ')}', 'Listener']

List 043 (20 tokens)
['Changed', 'seeing', '_noon', 'fered', '_guessing', 'adays', 'giving', '_mentioning', '_awhile', '_remodeling', '_skiing', '_graduating', '_2030', '_surfing', '_Election', 'boarding', '_averaging', '_rehears', '_vomiting', 'Duration']

List 044 (20 tokens)
['_overtime', '_trafficking', '_altitude', '_negligence', '_traumatic', '_congestion', '_GST', '_foreclosure', '_stringent', '_advers', '_subsidies', '_shortages', '_wastewater', '_pneumonia', '_Parkinson', '_tariffs', '_ovarian', '_untreated', '_closures', '_PTSD']

List 045 (20 tokens)
['_whatsoever', '_disclose', '_diligence', '_tolerate', '_propriet', '_incur', '_sanctuary', '_locom', '_amidst', '_amend', '_holdings', '_therein', '_spur', '_yourselves', '_epit', '_flee', '_protector', '_athlet', '_dispos', '_overhaul']

List 046 (20 tokens)
['166', '304', '345', '487', '707', '327', '387', '176', '234', '324', '359', '186', '677', '268', '379', '258', '505', '347', '277', '609']

List 047 (20 tokens)
['However', 'Another', 'Although', 'Because', 'including', 'Hey', 'Having', 'Those', 'gets', 'lying', 'DU', '_Meanwhile', 'Its', 'bing', 'hover', 'zing', 'Meanwhile', '_Later', 'lights', 'Would']

List 048 (20 tokens)
['_enthusiasts', '_170', '_108', '_graduating', '_Rangers', '_mainland', '_pals', '_Hispanic', '_caregivers', '_Alban', '_Sloven', '_skeptical', '_Kazakh', '_Orthodox', '_Niger', '_Scient', '_Latino', '_defenders', '_todd', '_libert']

List 049 (20 tokens)
['ories', '_Sciences', 'ologically', '_midst', '_Psychology', '_forefront', '_Divine', '_Philosophy', '_LORD', 'orie', 'odor', '_slightest', 'matic', '_marrow', '_resurrection', '_epit', '_madness', '_ageing', '_sensations', '_miracles']

List 050 (20 tokens)
['bie', 'ted', 'ounce', '_Amendment', 'abolic', 'ormal', 'nik', 'gie', 'imated', '_overtime', '_IRA', 'sell', 'ufact', 'swers', 'iversary', 'gor', 'trim', 'onym', 'ounced', 'odic']

List 051 (20 tokens)
['_Grove', '_Divine', '_Fortune', '_Peak', '_Naturally', '_LB', '_Carr', '_KS', '_Myers', '_Yard', '_Eld', '_KB', '_Mant', '_Nash', '_Bake', '_Kab', '_Sax', '_Mum', '_Cod', '_Spar']

List 052 (20 tokens)
['umm', 'ATION', 'aming', 'ERS', 'ags', 'inding', 'ACK', 'ologist', 'aving', 'IVE', 'ummy', '>\\', 'iate', 'aked', 'iders', 'ctx', 'rees', 'illed', 'amps', 'ESS']

List 053 (20 tokens)
['\u662f', '\ufffd', '\u4e0d', '\u7528', '\u4e86', '\u5728', '\u6709', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', 'Assert', '\u6570', 'console', '\ufffd', 'debug', '\u884c', '\ufffd']

List 054 (20 tokens)
['_Francisco', '_Dakota', '_Qing', '_Ethiopia', '_Bach', '_Ming', '_McD', '_Morocco', '_PBS', '_Cambodia', '_Roosevelt', '_Krish', '_Rosa', '_Sac', '_Saturn', '_Gandhi', '_Sic', '_Kane', '_Ferr', '_Sax']

List 055 (20 tokens)
['_grandfather', '_grandchildren', '_coordinator', 'stery', '_graduating', 'stic', '_commute', 'opia', '_pals', '_counselor', '_predecessor', '_mornings', '_wander', '_forehead', '_nominee', '_FIL', '_Wife', '_nie', '_fingertips', '_palate']

List 056 (20 tokens)
["'re", '_Islands', '_slices', '_sandwiches', '_ranged', '_soils', '_Regulations', '_packets', '_chunks', '_terminals', '_medals', '_consoles', '_cubes', '_closures', '_gestures', '_replacements', '_Facts', '_miracles', '_glands', '_Factors']

List 057 (20 tokens)
['_TODO', '_MVP', '_Layer', '_proto', '_throughput', '_lookup', '_CLI', '_logger', '_classifier', '_auth', '_ACL', '_SPI', '_SSH', '_Kubernetes', '_Cache', '_benchmarks', '_Auth', '_cpu', '_schematic', '_thresholds']

List 058 (20 tokens)
['_reson', '_overtime', '_nonsense', '_rehears', '_fret', '_flats', '_blows', '_whip', '_cues', '_identifiable', '_rests', '_palate', '_fuss', '_tempo', '_bumps', '_throttle', '_resh', '_closures', '_gestures', '_whist']

List 059 (20 tokens)
['inos', 'ination', 'inate', 'ively', 'ulation', 'ously', 'ories', 'ifies', 'inity', 'urer', 'ologist', 'icing', 'tered', 'ifications', 'ulate', 'iate', 'iously', 'ographic', 'inated', 'uous']

List 060 (20 tokens)
['_syrup', '_cinnamon', '_yogurt', '_broth', '_grilled', '_carrots', '_chili', '_squash', '_spinach', '_whip', '_basil', '_Wii', '_ether', '_diluted', '_immersion', '_glycol', '_pellets', '_Alto', '_shredded', '_pudding']

List 061 (20 tokens)
['agra', 'pper', 'aming', 'pped', 'lette', 'icing', 'ctions', 'crete', 'aning', 'commerce', 'ples', 'ppers', 'irst', 'quet', 'atable', 'ctor', 'ppy', 'posed', 'forming', 'cludes']

List 062 (20 tokens)
['_amazed', '_unaware', '_populated', '_een', '_hereby', '_urging', '_tasked', '_detained', '_aggressively', '_terrified', '_slated', '_accelerating', '_billed', '_eagerly', '_flown', '_loosely', '_incapable', '_fueled', '_bred', '_constituted']

List 063 (20 tokens)
['Another', 'Because', 'including', '\ufffd', 'rated', '\ufffd', '\ufffd', 'Remember', '\ufffd', 'DIR', '2002', '\u4e0a', 'MOD', 'especially', '\u30b9', '\ufffd', 'powered', '1997', 'OG', 'although']

List 064 (20 tokens)
['gov', '_sq', '_kg', 'ply', '_nm', 'performance', '_fn', 'pine', '_ein', 'leading', '_mph', '_lb', 'drive', '_canal', 'brook', '_dans', 'dar', 'office', 'management', '_aus']

List 065 (20 tokens)
['cing', 'pect', 'mith', 'phalt', 'sts', 'cription', 'Changed', 'heets', 'Listener', 'Bytes', 'ynthetic', 'cript', 'creen', 'chool', 'Byte', 'Scalar', 'Invalid', 'cribe', 'Rect', 'ingle']

List 066 (20 tokens)
['_Mountains', '\u2026..', '_remod', '_Thing', '_foliage', '_Wife', '_staircase', '_flowering', '!!!!!!!!!', '_Prayer', '_Scul', '_rendition', '_Folk', '_robe', '_Boh', '_duplex', "!'", '_orn', '_handwriting', ':)']

List 067 (20 tokens)
['_Dakota', '_Grove', '_Trek', '_Fortune', '_Maker', '_Excellence', '_Peak', '_Hopkins', '_Tournament', '_Throw', '_Hispanic', '_Peninsula', '_JA', '_PBS', '_LB', '_Marathon', '_MVP', '_Sons', '_Yard', '_Colleg']

List 068 (20 tokens)
['Although', '_Besides', '_Unlike', '_Especially', '_Later', '_\xa7', '_Either', '_Obviously', '_Unless', '_Hopefully', '_Fortunately', '_Probably', '_Typically', '_Basically', '_Alternatively', '_Whenever', '_Neither', '_Anything', '_Apparently', '_Expect']

List 069 (20 tokens)
['_forefront', '_artery', '_basin', '_broth', '_mainland', '_vicinity', '_corridor', '_clerk', '_cemetery', '_forehead', '_northeast', '_sew', '_corps', '_locker', '_chambers', '_attic', '_poisoning', '_pneumonia', '_slightest', '_altar']

List 070 (20 tokens)
['_asleep', '_thankful', '_amazed', '_unsure', '_unveiled', '_drastically', '_mindful', '_bothered', '_canceled', '_periodically', '_unaware', '_rented', '_unchanged', '_tempted', '_embarrassed', '_pals', '_hereby', '_expressly', '_emailed', '_ashamed']

List 071 (14 tokens)
['iveness', 'ancies', 'lessness', 'iances', 'atics', '_incomes', '_regimes', '_Buddhism', '_deficits', '_rece', '_inequalities', 'ivism', '_eds', 'oretic']

List 072 (20 tokens)
["'re", "'ve", '_resembles', '_specifies', '_tackles', '_lied', '_hath', '_warrants', '_satisfies', '_interacts', '_waits', '_embr', '_evaluates', '_preserves', '_originates', '_obtains', '_rejects', '_accompanies', '_behaves', '_observes']

List 073 (20 tokens)
['\ufffd', '\ufffd', '\ufffd', '\ufffd', '\u0438', '\u0440', '\u2588\u2588', '\u044b', '\u043e\u0432', '\u043d', '\u043b', '\u0443', '\ufffd', '\ufffd', '\u0442', '_\u0441', '_\ufffd', '\ufffd', '\ufffd', '\u0435\u0440']

List 074 (20 tokens)
["'re", '_grandchildren', '_showers', '_gatherings', '_graduating', '_leagues', '_reps', '_mortgages', '_caregivers', '_mornings', '_medals', '_deeds', '_spouses', '_husbands', '_sacrifices', '_convictions', '_sensations', '_ballots', '_cooks', '_offenses']

List 075 (20 tokens)
['_Angeles', '_Francisco', '_Nations', '_Month', '_IRA', '_Plaza', '_Beast', '_Hispanic', '_Gaza', '_Reform', '_sanctuary', '_pistol', '_cemetery', '_northeast', '_Marathon', '_Palestine', '_organizer', '_corps', '_Attack', '_Yankees']

List 076 (20 tokens)
['However', 'According', 'Another', 'Although', 'During', 'Because', 'Whether', 'Hey', 'Despite', 'Having', 'Those', 'Its', 'Meanwhile', '\ufffd', 'Would', 'Unfortunately', 'Following', 'Wow', 'Though', 'Remember']

List 077 (20 tokens)
['LINE', 'ACK', '333', '408', '166', '999', 'DU', '384', '304', '345', '9999', 'MAX', '487', '707', 'DIR', 'IGN', '2002', 'DATA', '327', '017']

List 078 (20 tokens)
['_Angeles', '_Francisco', '_Islands', '_Singh', '_Potter', '_Grove', '_Heights', '_Cruz', '_Sr', '_Lodge', '_Kumar', 'stown', '_Plaza', '_Xi', '_Xu', '_Hopkins', '_Bros', '_Holder', '_Township', '_Bach']

List 079 (20 tokens)
['_Sciences', '_DATE', '2002', '_-.', '_107', '_THAT', '_INFO', '_Degree', 'MIN', 'MAN', '1996', '_Associates', '_Programme', '_Sr', '1998', '_170', '_111', '_Surgery', '_114', '_BUT']

List 080 (20 tokens)
['\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', 'ination', '\ufffd', 'inate', 'rieved', '\ufffd', '\ufffd', 'inding', 'inity', 'urer', 'icing', 'ifications', 'iate', 'inated']

List 081 (20 tokens)
['akers', 'inners', 'iders', 'oids', '_Hence', '_Nevertheless', 'izers', '_Wars', 'oured', '_Meanwhile', '_Especially', '_Otherwise', 'ocation', 'osed', 'ifting', 'ellers', '_cents', '_remotely', 'owing', 'otted']

List 082 (20 tokens)
['\u20ac', 'including', 'inline', 'DU', 'ACE', 'vector', 'column', 'dependent', 'unsigned', '2002', 'selected', 'income', 'Cong', 'MOD', 'especially', 'risk', 'AST', 'linear', 'Panel', '133']

List 083 (20 tokens)
['_artery', '_yogurt', '_vomiting', '_gland', '_congestion', '_palate', '_slightest', '_bumps', '_marrow', '_hydration', '_fermentation', '_ache', '_reflux', '_sensations', '_vagina', '_glands', '_catheter', '_hairs', '_fungus', '_exhaustion']

List 084 (20 tokens)
['_spite', '_vitro', '_whatsoever', 'scriptions', '_guessing', '_mentioning', '_Romania', '_Ethiopia', '_Croatia', '_Gaza', '_Morocco', '_Cambodia', '_broadcasting', '_Bulgaria', '_Libya', '_Palestine', '_Kazakh', '_Kuwait', 'experienced', '_noticing']

List 085 (20 tokens)
['_donate', '_realise', '_disclose', '_recognise', '_enroll', '_reproduce', '_derive', '_analyse', '_oppose', '_violate', '_resemble', '_embed', '_refine', '_simulate', '_mimic', '_amend', '_allocate', '_denote', '_commem', '_flourish']

List 086 (20 tokens)
['_spite', '_compartment', '_nonsense', '_Chaos', '_crap', '_sweetness', '_corps', '_injustice', '_fantasies', '_advers', '_shortages', '_tariffs', '_contempt', '_ignorant', '_madness', '_mythology', '_eternity', '_imminent', '_envy', '_hardship']

List 087 (20 tokens)
['_Aid', '_forefront', '_Television', '_inception', '_broadcasting', '_standpoint', '_Religion', '_Colleg', '_Families', '_campuses', '_subsidies', '_holdings', '_Fiction', '_resurrection', '_Comics', '_precedent', '_economist', '_facilit', '_Jour', '_mandates']

List 088 (20 tokens)
['_Potter', '\u2026..', '\u2026\u2026', '_Fortune', '_goodbye', '_Bears', '_Scripture', '_Nurse', '_Dame', '_tho', '_\U0001f642', '_Ya', '_Mate', '----', '_Bean', '\u2026\u201d', '_Ranger', '_Seriously', '!!!!!!!!', '_Bake']
List 089 (20 tokens)
['_footprint', '_bulb', '_feeder', '_cutter', '_pans', '_fuse', '_Casual', '_organizer', '_digitally', '_enclosure', '_pou', '_rollers', '_strands', '_batting', '_knob', '_tubing', '_strand', '_axle', '_protector', '_wand']
List 090 (20 tokens)
['\u3001', 'izations', 'ogenic', 'arians', 'aire', 'arium', '_Rd', '_compartment', '_arises', '_miR', '_feeder', '_booster', '_\u2264', 'ogenesis', '_bc', '_apoptosis', '_carcinoma', '_kinase', '_encoded', '_cortex']
List 091 (20 tokens)
['_syrup', '_stroll', '_slices', '_cinnamon', '_peppers', '_baths', '_skiing', '_starters', '_yogurt', '_broth', '_surfing', '_grilled', '_sandwiches', '_bites', '_volley', '_buds', '_noodles', '_reps', '_carrots', '_chili']
List 092 (20 tokens)
['irst', 'cted', 'irm', 'erated', 'NAL', '_stroll', '_technician', 'uction', '_canal', 'yers', '_feeder', 'UG', '_gallon', '_Matter', '_LOT', 'eros', '_staircase', 'ffect', '_pellet', 'pex']
List 093 (20 tokens)
['_Sciences', '_Springs', '_Airlines', '_Industries', '_Rd', '_Associates', '_Studios', '_Foods', '_Surgery', '_Lodge', '_Initiative', '_Networks', '_Plaza', '**.', '_INC', '_Peak', '_Holdings', '_Devices', '_Relief', '_Nurs']
List 094 (20 tokens)
['_Feng', '_molding', '_combustion', '_phosphate', '_Qi', '_Saturn', '_lign', '_ether', '_Sigma', '_plated', '_hydration', '_fermentation', '_tandem', '_refrig', '_recess', '_semiconductor', '_nanoparticles', '_pollen', '_vibrations', '_Gore']
List 095 (20 tokens)
['_Trek', '_Croatia', '_Chand', '_Ming', '_Morocco', '_PBS', '_Marathon', '_Via', '_Raleigh', '_Xin', '_Lah', '_Wik', '_Omaha', '_FOX', '_Billboard', '_Mys', '_NPR', '_Syracuse', '_Jared', '_Getty']
List 096 (20 tokens)
['_unsure', '_sourced', '_forefront', '_thereafter', '_inception', '_standpoint', '_digitally', '_identifiable', '_electronically', '_consulted', '_summarized', '_readable', '_precedent', '_economist', '_formatted', '_attributable', '_imprint', '_commonplace', '_exhaustive', '_envisioned']
List 097 (20 tokens)
['_altitude', '_\u03c0', '_\u03b5', '_\u03c4', '_\u223c', '_\\(', '_\u03b8', '_\u03c9', '_achievable', '_throughput', '_occupancy', '_\u03c6', '_braking', '_Mn', '_elusive', '_efficiencies', '_positivity', '_\u0394', '_Stern', '_vortex']
List 098 (20 tokens)
['friendly', 'choice', 'margin', 'template', 'hour', 'door', 'minute', 'depth', 'winning', 'selling', 'admin', 'offset', 'profit', 'profile', 'task', 'buffer', 'scope', 'breaking', 'Wow', 'effective']
List 099 (20 tokens)
['However', 'According', '‘‘‘', 'Another', 'Although', '_(!', 'During', 'Because', 'including', 'Whether', 'Hey', '_Besides', 'Despite', 'Having', '_Hence', 'Those', '_Nevertheless', '_Unlike', '*(', '_Meanwhile']
List 100 (20 tokens)
['pper', 'ved', 'eless', 'ously', 'ving', 'aming', 'pped', 'akers', 'lette', 'inding', 'aving', 'bers', 'ading', 'ctions', 'aked', 'sters', 'crete', 'iders', 'aping', 'lier']
List 101 (20 tokens)
['_Month', '_grandfather', '_Heights', '_Trek', '_haul', '_victories', '_Thing', '_believer', '_Wife', '_deeds', '_tempo', '_grandson', '_Savings', '_Chance', '_rivalry', '_Shepherd', '_Plains', '_sermon', '_turnout', '_concession']
List 102 (20 tokens)
['_footprint', '_ceil', '_compartment', '_bulb', '_canal', '_basin', '_cutter', '_molding', '_halls', '_barrels', '_buds', '_mainland', '_fuse', '_terminals', '_corridor', '_jug', '_combustion', '_fret', '_hallway', '_foliage']
List 103 (20 tokens)
['ories', '"""', '\x80', ';</', '\u20ac\u2122', '_Authority', '_Affairs', '/>', ';&', '\n\t\t\t\t\t\t\t', '_Later', '_"",', '="{{', '_doi', '_%}', '_\xa7', '\x99', '\ufffd\ufffd', '\u043f', '_Unless']
List 104 (20 tokens)
['tmp', 'define', '});', 'summary', 'github', 'png', 'choice', 'header', 'margin', 'msgid', 'nbsp', 'bool', 'description', 'padding', 'endif', 'template', 'hour', 'css', 'query', 'btn']
List 105 (20 tokens)
['_Grove', '_Hampshire', '_Theater', '_Qing', '_Plaza', '_Dh', '_Television', '_Xi', '_Emirates', '_Chand', '_Sector', '_Huang', '_Peninsula', '_Springer', '_Symphony', '_Clarke', '_Carr', '_Via', '_Ya', '_Away']

List 106 (20 tokens)
['\ufffd', '\u7684', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd']

List 107 (20 tokens)
['_footprint', '_gotta', '_altitude', '_inception', '_vicinity', '_terminals', '_predecessor', '_encoded', '_chloride', '_rele', '_preferable', '_throughput', '_vap', '_vibrations', '_synthesized', '_incubation', '_manifestations', '_lineage', '_prominence', '_latitude']

List 108 (20 tokens)
['_miR', '_Eq', '_traumatic', '_\u03c0', '_apoptosis', '_carcinoma', '_phosphate', '_\u03b8', '_microscopy', '_Parkinson', '_lign', '_Theorem', '_ether', '_ovarian', '_untreated', '_\u03c9', '_chloride', '_epithelial', '_Klein', '_fermentation']

List 109 (20 tokens)
['_inception', '_maneuver', '_predecessor', '_sweetness', '_stringent', '_flowering', '_boosted', '_glor', '_caliber', '_massively', '_centrally', '_alarming', '_surreal', '_dominating', '_tast', '_famously', '_prominence', '_mystical', '_prow', '_aest']

List 110 (20 tokens)
['_herein', '_dioxide', '_compartment', '_bulb', '_arises', '_canal', '_artery', '_resembles', '_carcinoma', '_kinase', '_cortex', '_gland', '_phosphate', '_amplifier', '_depicts', '_specifies', '_ether', '_chloride', '_investigator', '_epithelial']

List 111 (13 tokens)
['_heck', '_striker', '_crap', '_fuss', '_slightest', '_thumbs', '_pt', '_replay', '_heavens', '_tester', '_rm', '_vid', '_arbitr']

List 112 (20 tokens)
['izations', 'izers', '_Springs', '_Duty', '_cutter', '_Maker', '_molding', '_Chand', '_Dynamics', '_Mold', '_Mock', '_Yard', '_Sword', '_Packers', '_Rings', '_Bars', 'isations', '_Drain', '_Films', '_Sheets']

List 113 (20 tokens)
['inos', 'umm', 'ination', 'inate', 'earing', 'ulation', 'ously', 'ags', 'inding', 'umps', 'ummy', 'izations', 'ctions', 'aked', 'inated', 'aning', 'iving', 'illed', 'oids', 'ogs']

List 114 (20 tokens)
['_litter', '_overtime', '_disappoint', '_mornings', '_congestion', '_shortages', '_bumps', '_closures', '_sacrifices', '_knots', '_hardship', '_hesitation', '_hairs', '_scratching', '_vibrations', '_deferred', '_corrupted', '_unrealistic', '_interruption', '_blot']

List 115 (20 tokens)
['_Unlike', '_Meanwhile', '_Either', '_Obviously', '_Unless', '_Hopefully', '_Probably', '_Basically', '_Whenever', '_Neither', '_Divine', '_Anything', '_Apparently', '_Aside', '_Whilst', '_"%', '_Sadly', '_Throw', '_Rating', '_Quite']

List 116 (20 tokens)
['_Besides', '_Hence', '_Nevertheless', '_Unlike', '_Meanwhile', '_Especially', '_Otherwise', '_Later', '_doi', '_\xa7', '_Either', '_Obviously', '_Unless', '_Fortunately', '_107', '_Psychology', '_nm', '_Probably', '_Nonetheless', '_Typically']

List 117 (20 tokens)
['_grandfather', '_grandchildren', '_technician', '_graduating', '_goodbye', '_Rescue', '_sandwiches', '_halls', '_crews', '_turf', '_reps', '_LORD', '_pals', '_Lives', '_counselor', '_sanctuary', '_mornings', '_clerk', '_cemetery', '_fantasies']

List 118 (20 tokens)
['///', '/*', 'tmp', 'define', '});', '````', 'summary', 'github', 'png', 'choice', 'header', 'margin', '"""', 'nbsp', 'bool', 'description', 'padding', 'endif', 'template', 'css']

List 119 (20 tokens)
['_Acid', '_230', '_burner', '_cutter', '_Chand', '_Casual', '_Bean', '_pellet', '_Sigma', '_Ferr', '_Drain', '_dal', '_axle', '_sewage', '_refrig', '_Pole', '_pt', '_Dew', '_cot', '_Mild']

List 120 (20 tokens)
['_pelvic', '_manifestations', '_inte', '_genital', '_modalities', '_morbidity', '_Gaussian', '_airway', '_aortic', '_symb', '_architectures', '_schematic', '_gait', '_deform', '_transcriptional', '_unfolding', '_mammalian', '_embryonic', '_pathogenesis', '_medial']

List 121 (20 tokens)
['_grandfather', '_nonsense', '_volley', '_heck', '_nig', '_Beast', '_jug', '_striker', '_QB', '_LB', '_MVP', '_guitarist', '_nam', '_keeper', '_midfielder', '_tempo', '_trilogy', '_defenders', '_freak', '_ignorant']

List 122 (20 tokens)
['ppers', '_complain', '_disclose', '_overlook', '_forgive', '_reproduce', '_soak', '_negligence', '_theirs', '_tolerate', '_oppose', '_threaten', '_maneuver', '_incur', 'pless', '_violate', '_expire', '_fret', '_reopen', '_felony']

List 123 (20 tokens)
['/*', '\u7684', '_(!', '\r\n\t\t', '\u3001', '\r\n\t\t\t\t', '____________________', '\n\n\t\t', '______________________', '_Meanwhile', '\n\t\t\t\t\t\t\t\t', '_?>', '_%}', '\t\t\t\t', '_\xa7', '\n\n\n\n', '\n\t\n', ')}', '_++', '<<']

List 124 (20 tokens)
['_disappoint', '_negligence', '_expenditures', '_advers', '_infring', '_shortages', '_bumps', '_tariffs', '_defenses', '_overload', '_closures', '_shocks', '_breaches', '_misconduct', '_replacements', '_tendencies', '_ageing', '_sensations', '_contingent', '_burdens']

List 125 (20 tokens)
['_wanna', '_Potter', '_Cruz', '_gotta', '_forgive', '_bless', '"...', '_Scripture', '_Bros', '_Bach', '_tweeted', '_Lives', '_Garcia', '_Isn', '_tasted', '_Rodriguez', '?).', '_Peterson', '_Burns', '_thanked']

List 126 (20 tokens)
['\r\n\t\t', 'Whether', '\r\n\t\t\t\t', '____________________', '_Unlike', '_"",', '_Either', '\n\n\n\n', '_Hopefully', '_Probably', '_Typically', '_Basically', '_};', '_Whenever', '_});', '_Chevrolet', '_Apparently', '_Expect', '_Cause', '_Rescue']

List 127 (20 tokens)
['_spite', '_midst', '_Psychology', '_ein', '_Testament', '_dioxide', '_forefront', '_trafficking', '_inception', '_Philosophy', '_eine', '_een', '_algebra', '_combustion', '_Reality', '_predecessor', '_cortex', '_standpoint', '_Religion', '_expedition']

List 128 (20 tokens)
['inos', 'ination', 'inate', 'ved', 'eless', 'ulation', 'ving', 'lette', 'lers', 'tered', 'ifications', 'ulate', 'bers', 'ading', 'iate', 'sters', 'iously', 'inated', 'lier', 'cing']

List 129 (20 tokens)
['_sq', '_kg', '_lbs', 'cluded', '_\xa7', '_cents', '_spite', '_grams', '_metres', '_vitro', '_mins', '_Month', 'verter', 'scriptions', '_kilometers', '_euros', '_Qing', '_2030', '_Romania', '_1945']

List 130 (20 tokens)
['_donate', '_realise', '_disclose', '_recognise', '_overlook', '_forgive', '_enroll', '_haul', '_reproduce', '_soak', '_swear', '_tolerate', '_derive', '_analyse', '_oppose', '_maneuver', '_incur', '_resemble', '_wander', '_whip']

List 131 (20 tokens)
['\u7684', '|||EMAIL_ADDRESS|||', '\ufffd', '\ufffd', '\ufffd', '\u3001', '\u20ac', 'ACK', '\x80', '\u662f', '###', 'src', '\u0440', '\ufffd', '333', '\u2588\u2588', '\u20ac\u2122', '\u2033', '\u4e0d', '\u043d']

List 132 (20 tokens)
['_thankful', '_unsure', '_pursuant', '_mindful', '_unaware', '_unchanged', '_payable', '_traumatic', '_mainland', '_vous', '_skeptical', '_identifiable', '_stringent', '_Numerous', '_epithelial', '_ignorant', '_achievable', '_imminent', '_contingent', '_readable']

List 133 (20 tokens)
['ously', 'lette', '_\u0430', 'bles', '_\u043a', '\u0433', '_zu', '_auf', '_f\xfcr', '_ein', '\u010d', 'rers', '_dans', '_\u0431', 'club', '_aus', 'bourg', 'hall', '_versa', '_analyse']

List 134 (20 tokens)
['3333', '%,', 'umm', 'ination', '%)', 'inate', 'acks', 'earing', 'ulation', 'akers', 'ags', 'inding', 'css', 'ologist', 'msg', 'umps', 'ading', '>\\', 'aked', 'iously']

List 135 (20 tokens)
['_Nevertheless', 'ometric', 'ellers', 'heets', '_arises', '_risen', '_tempted', '_ranged', '_clicked', '_populated', '_analysed', '_emailed', '_het', '_undergone', '_empowered', '_punished', '_detained', '_infring', 'omers', '_foc']

List 136 (20 tokens)
['_\u0441', '_\ufffd', '_Wars', '_Islands', '_Later', '_\u0430', '_\u043c', '_\u043a', '_\u0443', '_\u0434', '_\u043e', '_\u0442', '_Psychology', '_auf', '_\u043d\u0430', '_\u043d', '_ein', '_Forces', '_Acid', '_Surgery']

List 137 (20 tokens)
['ifies', 'ifications', 'profit', 'rated', 'iques', 'driven', 'posed', 'weighted', 'forming', 'induced', 'dependent', 'ifiers', 'breaking', 'oriented', 'lined', 'shaped', 'resistant', 'formed', 'desc', 'selected']

List 138 (20 tokens)
['_spite', '_whatsoever', '_grandfather', '_basin', 'acles', '_Philosophy', 'naments', '_combustion', '_Reality', '_standpoint', '_Religion', '_Marriage', '_Encyclopedia', '_injustice', '_faction', '_regimes', '_imprisonment', '_mythology', '_righteousness', '_Buddhism']

List 139 (20 tokens)
['_remod', '_remodeling', '_skiing', '_surfing', '_mortgages', '_foreclosure', '_brewing', '_firearm', '_hydration', '_athlet', '_transporting', '_dispos', '_braking', '_advocating', '_pornography', '_zoning', '_geneal', '_MMA', '_stacking', '_conveying']

List 140 (20 tokens)
['|||IP_ADDRESS|||', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '3333', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd']

List 141 (20 tokens)
['mith', 'dale', '_wanna', 'creen', 'chool', 'weet', 'uits', 'ingle', '_gotta', 'dating', '_bc', 'hots', 'oft', '_nig', '_pals', '_tweeted', '_sue', '_nie', '_hav', 'hoe']

List 142 (20 tokens)
['_asleep', '_spite', '_remotely', '_drastically', '_bless', '_enthusi', '_nonsense', '_periodically', '_disappoint', '_frankly', '_indirectly', '_unchanged', '_embarrassed', '_incur', '_expressly', '_ashamed', '_skeptical', '_illegally', '_digitally', '_reliably']

List 143 (20 tokens)
['_Qing', '_Feng', '_Dh', '_Xi', '_Xu', '_Chand', '_Nurse', '_Ming', '_Wei', '_Huang', '_Zhao', '_Yuan', '_Clarke', '_Mate', '_Wife', '_Wen', '_Krish', '_Inspector', '_Zhou', '_Eld']

List 144 (20 tokens)
['%,', '\u20ac', '333', 'jpg', '408', '\u2033', '_\u0441', '166', '999', '755', '\xd7', '_sq', '_kg', '384', '304', '345', '_lbs', '487', '707', '_\u043c']

List 145 (20 tokens)
['‘‘‘‘', 'summary', 'github', 'Another', 'header', 'margin', 'description', 'template', 'css', 'query', '###', 'src', '\ufffd', 'Having', 'layout', 'admin', 'profile', 'task', 'metadata', 'session']

List 146 (20 tokens)
['_Francisco', '_doi', '_zu', '_ein', '_Sr', '_coordinator', '_Romania', '_INC', '_FY', '_Croatia', '_1963', '_WI', '_1962', '_Springer', '_Administrator', '_JA', '_1961', '_Alban', '_Bulgaria', '_KS']

List 147 (20 tokens)
['_Springs', '_Singh', '_Amendment', '_Corps', '_Associates', '_Cruz', '_Sr', '_Initiative', '_unveiled', '_Bears', '_Election', '_Bros', '_aired', '_tout', '_tweeted', '_Administrator', '_sparked', '_nominee', '_Simpson', '_thanked']

List 148 (17 tokens)
['_mindful', '_todd', '_filmmaker', '_coworkers', '_colleg', '_happ', '_pars', '_dst', '_rapport', '_mating', '_adolescence', '_actresses', '_sep', '_wy', '_heterosexual', '_tai', '_divor']

List 149 (20 tokens)
['_Industries', '_Duty', '_molding', '_4000', '_terminals', '_650', '_\xbd', '_phosphate', '_reliably', '_coatings', '_stringent', '_Resistance', '_rollers', '_plated', '_knob', '_overload', '_throttle', '_tubing', '_Width', '_turbine']

List 150 (20 tokens)
['_enthusiasts', '_remodeling', '_gatherings', '_leagues', '_crews', '_minorities', '_pursu', '_coatings', '_chambers', '_ceremonies', '_Niger', '_strands', '_purch', '_detectors', '_defenses', '_instit', '_libert', '_landmarks', '_regimes', '_batches']

List 151 (20 tokens)
['_opener', '_coordinator', '_LORD', '_KEY', '_predecessor', '_clerk', '_organizer', '_nominee', '_slightest', '_epit', '_FIRST', '_motto', '_appellant', '_subdivision', '_issuance', '_debtor', '_surname', '_correspondent', '_latitude', '_happ']

List 152 (20 tokens)
['_Hence', '_Nevertheless', '_Consequently', '_sourced', '_herein', '_risen', '_clicked', '_populated', '_analysed', '_mapped', '_encoded', '_cooled', '_inspected', '_paved', '_incorrectly', '_tasked', '_incubated', '_photographed', '_conditioned', '_mediated']

List 153 (20 tokens)
['_thereof', '_herein', '_guessing', '_mentioning', '\u2026\u2026', '_Scripture', '.]', '_costing', '.");', '?).', '_\\(', '.|', '\u2026\u201d', '_deduction', '_spouses', '_violating', '?\u2019', '_faction', '_compromising', '_righteousness']

List 154 (20 tokens)
['_asleep', '_unsure', '_unveiled', '_drastically', '_mindful', '_bothered', '_unaware', '_indirectly', '_tempted', '_populated', '_costing', '_expressly', '_impacting', '_empowered', '_punished', '_spanning', '_tougher', '_identifiable', '_reliably', '\u2026\u201d']

List 155 (20 tokens)
['_sourced', '_drastically', '_risen', '_bothered', '_enthusi', '_canceled', '_ranged', '_populated', '_expire', '_expressly', '_sanctuary', '_cooled', '_felony', '_undergone', '_postponed', '_spanning', '_illegally', '_stringent', '_detained', '_boosted']

List 156 (20 tokens)
['_realise', '_recognise', '_soak', '_thereafter', '_swear', '_bites', '_clicked', '_embed', '_blows', '_tasted', '_thanked', '_plugged', '_swallow', '_healed', '_explode', '_plunge', '_satisfies', '_kisses', '_immersed', '_kissed']

List 157 (20 tokens)
['rology', 'rolog', 'abolic', 'roph', 'rost', '_indirectly', '_apoptosis', '_vomiting', '_gland', 'osterone', '_mimic', 'rade', 'rological', '_ovarian', '_hydration', '_glands', '_accelerating', '_Sympt', '_exacerb', '_pelvic']

List 158 (20 tokens)
['_whatsoever', '_drastically', '_4000', '_190', '_populated', '_threaten', '_chunks', '_650', '_spanning', '_illegally', '_270', '_260', '_advers', '_amidst', '_vertically', '_2500', '_irrespective', '_trout', '_progressively', '_pollen']

List 159 (20 tokens)
['_Authority', '_Directors', '_Corps', '_Province', '_Rangers', '_Bears', '_Election', '_Emirates', '_Township', '_WI', '_Defence', '_Infrastructure', '_LORD', '_Beast', '_Hispanic', '_PBS', '_Sloven', '_GST', '_Jets', '_Utt']

List 160 (20 tokens)
['acks', 'ved', '_Angeles', 'eless', 'ories', 'ving', '_Zealand', 'lette', 'lers', 'inity', 'tered', 'sters', 'crete', 'ographic', 'lier', 'yard', 'ancement', 'iques', 'mith', 'scape']

List 161 (20 tokens)
['_hesitate', '_complain', '_donate', '_realise', '_disclose', '_recognise', '_overlook', '_owe', '_forgive', '_bless', '_enroll', '_haul', '_reproduce', '_soak', '_apologize', '_swear', '_tolerate', '_derive', '_analyse', '_oppose']

List 162 (20 tokens)
['_striker', '_nominee', '_keeper', '_faction', '_strand', '_follower', '_survivor', '_incarn', '_remake', '_precedent', '_disastrous', '_senator', '_tweak', '_alarming', '_offender', '_homage', '_enlight', '_rendition', '_contender', '_catastrophe']

List 163 (20 tokens)
['});', '‘‘‘', '_(!', '"""', '\r\n\t\t', '_serialized', 'bool', 'endif', '>\\', 'Override', '\r\n\t\t\t', '="#', '__________', '/>', '\n\n\t\t', '__________', '\n\t\t\t\t\t\t\t', 'Writer', '_DATE', '_btn']

List 164 (20 tokens)
['ties', '_awhile', '_starters', '_spouses', '_pneumonia', '_todd', '_probation', '_yourselves', '_transplantation', '_eternity', '_dismissal', '_transporting', '_aspir', '_hosp', '_issuance', '_genital', '_CPR', '_supper', '_diab', '_hospitalization']

List 165 (20 tokens)
['fulness', '_footprint', '_compartment', '_arises', '_unveiled', '_bothered', '_ranged', '_aired', '_propriet', '_maneuver', '_resembles', '_spans', '_blows', '_tasted', '_entails', '_encompasses', '_paved', '_depicts', '_locom', '_specifies']

List 166 (20 tokens)
['verse', '_thankful', 'ught', '_brighter', '_traumatic', '_felony', '_believer', '_skeptical', '_tougher', '_nam', '_stringent', '_repent', '_forg', '_libert', '_covenant', '_contempt', '_freak', '_richer', '_faction', '_wary']

List 167 (12 tokens)
['_Xu', '_Huang', '_Zhao', '_Rodriguez', '_Suzuki', '_ETH', '_FOX', '_Kant', '_Lowe', '_Hamas', '_Schneider', '_Ub']

List 168 (20 tokens)
['_Corps', '_playoffs', '_Rangers', '_Bears', '_Election', '_Emirates', '_Defence', '_Beast', '_Sox', '_Gaza', '_striker', '_expedition', '_Jets', '_rains', '_Yankees', '_hearings', '_nominations', '_Kuwait', '_Packers', '_insurer']

List 169 (20 tokens)
['_hesitate', '_midst', '_risen', '_thereafter', '_crawl', '_resemble', '_clerk', '_organizer', '_yourselves', '_plunge', '_confuse', '_revise', '_hesitation', '_stare', '_verge', '_outset', '_commencement', '_sideways', '_tore', '_stump']

List 170 (20 tokens)
['_analyse', '_algebra', '_minorities', '_refine', '_distract', '_revise', '_summarize', '_summar', '_autobi', '_individualized', '_immersion', '_inequalities', '_disciplined', '_competency', '_synthes', '_voc', '_disadvant', '_frustr', '_discour', '_evaluates']

List 171 (20 tokens)
['_grandfather', '_gotta', '_streak', '_opener', '_surfing', '_commute', '_reps', '_pals', '_victories', '_predecessor', '_wander', '_expedition', '_stride', '_pursu', '_r\xe9', '_fantasies', '_obsess', '_trilogy', '_defenders', '_resurrection']

List 172 (20 tokens)
['DU', 'vd', 'fy', 'LY', 'OG', 'IZ', 'OX', 'Vs', 'wf', 'rx', 'LT', 'FX', 'RP', 'UG', 'LM', 'WT', 'IK', 'HB', 'Gs', 'uu']

List 173 (20 tokens)
['define', 'ively', 'friendly', 'ories', 'lessly', 'ifies', 'including', 'icing', 'door', 'sized', 'ifications', 'iate', 'sters', 'crete', 'ographic', 'menu', 'writer', '\u2033', 'selling', 'layout']

List 174 (20 tokens)
['umm', 'acks', 'ved', 'earing', 'ATION', 'ERS', '\u20ac', 'ACK', 'aving', 'IVE', 'umps', 'inners', 'iders', 'amps', 'ESS', 'andy', 'pect', 'orted', 'apping', 'ITY']

List 175 (20 tokens)
['_Angeles', '_lbs', '_wanna', '_THAT', '\u2026..', '_BUT', '_DEL', '_MORE', '_Anything', '_BACK', '_HAVE', '_INC', '_bc', '_Rates', '_minded', '_(@', '_Isn', '\ufffd', '_\U0001f642', '_Away']

List 176 (20 tokens)
['_thereof', '_bulb', '_opener', 'Keys', '_packets', '_tracker', '.");', '_receivers', '_pistol', '_keeper', 'Ids', '_detectors', '_knob', '_batches', '_follower', '_inspector', '_alarms', '_lookup', '_ballots', '_loader']

List 177 (7 tokens)
[’_fingertips’, ’_accountant’, ’_sheriff’, ’_kon’, ’_circus’, ’_vie’, ’_multipl’]
List 178 (20 tokens)
[’_enthusiasts’, ’_leagues’, ’_crews’, ’_theaters’, ’_hearings’, ’_nominations’, ’_Orthodox’, ’_campuses’, ’_chambers’, ’_ceremonies’, ’_Niger’, ’_strands’, ’_defenders’, ’_commemor’, ’_purch’, ’_instit’, ’_highways’, ’_landmarks’, ’_breaches’, ’_Catholics’]
List 179 (20 tokens)
[’aving’, ’ading’, ’aping’, ’ctor’, ’cker’, ’aped’, ’alth’, ’_amazed’, ’fty’, ’_unveiled’, ’cht’, ’eks’, ’aters’, ’_bothered’, ’_canceled’, ’_periodically’, ’ffer’, ’_rented’, ’_thereafter’, ’ctic’]
List 180 (20 tokens)
[’_ceil’, ’_overlooking’, ’_showers’, ’_overlook’, ’_opener’, ’_coordinator’, ’_gatherings’, ’_skiing’, ’_goodbye’, ’_halls’, ’_volley’, ’_turf’, ’_rehears’, ’_crawl’, ’_corridor’, ’_sanctuary’, ’_hallway’, ’_foliage’, ’_organizer’, ’_staircase’]
List 181 (20 tokens)
[’_Wars’, ’_Islands’, ’_Degree’, ’_Studios’, ’_Duty’, ’_Forces’, ’_Acid’, ’_Theater’, ’_Than’, ’_Networks’, ’_Dh’, ’_Maker’, ’_Expect’, ’_Television’, ’_Matter’, ’_Rescue’, ’_Peak’, ’_Tournament’, ’_Devices’, ’_Rating’]
List 182 (20 tokens)
[’_Representative’, ’_Hopkins’, ’_crews’, ’_Powell’, ’_McD’, ’_PBS’, ’_Lynch’, ’_Carr’, ’_Sullivan’, ’_Stevens’, ’_Andrews’, ’_Nash’, ’_Flint’, ’_Klein’, ’_FOX’, ’_councils’, ’_Walsh’, ’_NPR’, ’_Greene’, ’_Nort’]
List 183 (20 tokens)
[’_thankful’, ’_midst’, ’_goodbye’, ’_gallon’, ’_heck’, ’_fuse’, ’_counselor’, ’_hallway’, ’_skeptical’, ’_keeper’, ’_fuss’, ’_wary’, ’_sip’, ’_hardship’, ’_recess’, ’_dram’, ’_disastrous’, ’_replay’, ’_heartbeat’, ’_saga’]
List 184 (20 tokens)
[’_720’, ’_consoles’, ’_Saturn’, ’_Wii’, ’_Wik’, ’_LAN’, ’_Dragons’, ’_Myth’, ’_Gn’, ’_Vamp’, ’_LTE’, ’_SSH’, ’_DDR’, ’_WD’, ’_disabling’, ’_Plasma’, ’_Obs’, ’_Ub’, ’_Nex’, ’_Ult’]
List 185 (20 tokens)
[’_slices’, ’_feeder’, ’_cinnamon’, ’_Chevrolet’, ’_Qing’, ’_Feng’, ’_yogurt’, ’_molding’, ’_grilled’, ’_sandwiches’, ’_volley’, ’_noodles’, ’_carrots’, ’_nig’, ’_chili’, ’_spinach’, ’_locom’, ’_nam’, ’_pellet’, ’_Suzuki’]
List 186 (20 tokens)
[’_Sr’, ’_litter’, ’_yogurt’, ’_soils’, ’_\u03c0’, ’_apoptosis’, ’_carcinoma’, ’_combustion’, ’_spinach’, ’_foliage’, ’_137’, ’_phosphate’, ’_corps’, ’_flowering’, ’_wastewater’, ’_rubbish’, ’_Parkinson’, ’_lign’, ’_ovarian’, ’_untreated’]
List 187 (20 tokens)
[’_(!’, ’_serialized’, ’_sq’, ’_DATE’, ’_btn’, ’_"",’, ’raction’, ’ructions’, "_’./", ’_fn’, ’_className’, ’_{},’, ’_ceil’, ’_(_’, ’ractor’, ’_sizeof’, ’(!’, ’_lb’, "_’@", ’__(’]
List 188 (20 tokens)
["’re", "’ve", ’_hesitate’, ’_Later’, ’_complain’, ’_donate’, ’_wanna’, ’_Foods’, ’_gotta’, ’_Aid’, ’_Than’, ’_thou’, ’_forgive’, ’_bless’, ’_Anything’, ’_apologize’, ’_HAVE’, ’_Expect’, ’_Cause’, ’_Matter’]
List 189 (20 tokens)
[’izations’, ’izers’, ’_Wars’, ’ties’, ’istically’, ’_mins’, ’_enthusiasts’, ’_Forces’, ’itarian’, ’_Networks’, ’_dans’, ’bies’, ’isms’, ’isers’, ’_Engineers’, ’_soils’, ’_Lives’, ’_algebra’, ’_caregivers’, ’_seekers’]
List 190 (16 tokens)
[’_remod’, ’_remodeling’, ’_campuses’, ’_Commit’, ’_stringent’, ’_motto’, ’_repertoire’, ’_caf\xe9’, ’_centrally’, ’_individualized’, ’_multidisciplinary’, ’_redes’, ’_arsenal’, ’_competency’, ’_exhaustive’, ’_exhib’]
List 191 (20 tokens)
[’_donate’, ’_sourced’, ’_unveiled’, ’_reson’, ’_drastically’, ’_risen’, ’_enroll’, ’_reproduce’, ’_mindful’, ’_2030’, ’_canceled’, ’_ranged’, ’_derive’, ’_aired’, ’_threaten’, ’_analysed’, ’_mapped’, ’_embed’, ’_expire’, ’_reopen’]
List 192 (20 tokens)
[’_starters’, ’_trafficking’, ’_theirs’, ’_propriet’, ’_pals’, ’_minorities’, ’_corps’, ’_illegally’, ’_pursu’, ’_enclosure’, ’_spouses’, ’_subsidies’, ’_holdings’, ’_husbands’, ’_defenses’, ’_protector’, ’_imprisonment’, ’_replacements’, ’_avenues’, ’_transporting’]
List 193 (20 tokens)
["’ve", ’\ufffd’, ’\u7684’, ’\ufffd’, ’\ufffd’, ’\ufffd’, ’\ufffd’, ’\ufffd’, ’\ufffd’, ’\ufffd’, ’\ufffd’, ’\ufffd’, ’\ufffd’, ’\ufffd’, ’\ufffd’, ’\ufffd’, ’\ufffd’, ’\ufffd’, ’\ufffd’]
List 194 (20 tokens)
[’_asleep’, ’_meantime’, ’_amazed’, ’_unsure’, ’_overlooking’, ’_guessing’, ’_litter’, ’_stirring’, ’_albeit’, ’_brighter’, ’_unaware’, ’_tempted’, ’_maneuver’, ’_fret’, ’_forehead’, ’_urging’, ’_spanning’, ’_uh’, ’_dismant’, ’_fuss’]

List 195 (20 tokens)
['_overlook', '_arises', '_reson', '_bless', '_enroll', '_rehears', '_resembles', '_embed', '_spans', '_blows', '_entails', '_encompasses', '_depicts', '_specifies', '_rests', '_dismant', '_obsess', '_emphas', '_forg', '_commemor']

List 196 (20 tokens)
['jack', 'yard', 'scape', 'backs', 'boards', 'shirt', 'alls', 'lights', 'berries', 'lined', 'keeper', 'minded', 'fulness', 'houses', '_metres', 'flower', 'falls', 'liness', 'shirts', 'bread']

List 197 (20 tokens)
['_guessing', '_mentioning', '_negligence', '_resign', '_broadcasting', '_noticing', '_obsess', '_repent', '_forg', '_violating', '_libert', '_voluntarily', '_glor', '_contempt', '_ignorant', '_reminding', '_compromising', '_disrespect', '_preaching', '_Catholics']

List 198 (20 tokens)
['_THAT', '_stroll', '_BUT', '_MORE', '_BACK', '_HAVE', '_LOT', '_pans', '_OTHER', '_flats', '_YE', '_WAS', '_FIL', '_pou', '_husbands', '_cheeks', '_posing', '_PLAY', '_whist', '_flee']

List 199 (20 tokens)
['_Amendment', '_Divine', '_Hispanic', '_Reform', '_Orthodox', '_injustice', '_Latino', '_Sac', '_resurrection', '_oath', '_filmmaker', '_sacrifices', '_imprisonment', '_Treaty', '_slaughter', '_dismissal', '_motto', '_Torah', '_sanct', '_Beatles']

List 200 (20 tokens)
['_vitro', '_miR', '_artery', '_analysed', '_apoptosis', '_carcinoma', '_CAG', '_kinase', '_cortex', '_LB', '_RX', '_GST', '_respons', '_pneumonia', '_incubated', '_marrow', '_ovarian', '_untreated', '_epithelial', '_mediated']

List 201 (20 tokens)
['However', 'define', 'According', 'Another', 'Although', 'friendly', 'choice', 'During', 'Because', 'Whether', 'jack', 'Hey', 'Despite', 'Having', 'winning', 'selling', 'Those', 'gets', 'profit', 'Its']

List 202 (20 tokens)
['tmp', 'summary', 'header', 'margin', 'description', 'including', 'padding', 'template', 'hour', 'css', 'query', 'msg', 'door', 'src', 'menu', 'depth', 'append', 'yard', 'amps', 'commerce']

List 203 (20 tokens)
['_amazed', '_slices', '_remod', '_unveiled', '_bless', '_canceled', '_bites', '_clicked', '_aired', '_threaten', '_tweeted', '_dubbed', '_infused', '_resign', '_sparked', '_cooled', '_emailed', '_blows', '_inspected', '_postponed']

List 204 (20 tokens)
['_Springs', '_Orleans', '_107', '_Rd', '_Grove', '_Hampshire', '_170', '_111', '_overlooking', '_114', '_112', '_230', '_136', '_117', '_113', '_108', '_Plaza', '_106', '_103', '_118']

List 205 (20 tokens)
['_miR', '_Factor', '_Decision', '_investigator', '_LR', '_summar', '_actionable', '_CX', '_autobi', '_Cognitive', '_multidisciplinary', '_Lik', '_KP', '_competency', '_synthes', '_multim', '_SNP', '_particip', '_aph', '_transcriptional']

List 206 (20 tokens)
['izations', '_footprint', '_booster', '_touchdown', 'isms', '_inception', '_costing', '_rehears', '_combustion', '_MVP', '_hearings', '_nominations', '_briefing', '_shocks', '_batches', '_fermentation', '_plunge', '_dismissal', '_occupancy', '_audition']

List 207 (20 tokens)
['_meantime', '_Eq', '_averaging', '_algebra', '_combustion', '_KS', '_\\(', '_cubes', '_Sigma', '_overload', '_slab', '_Klein', '_LR', '_knots', '_iter', '_lookup', '_VT', '_sampled', '_ans', '_theorem']

List 208 (20 tokens)
['anges', '_bake', '_syrup', '_cinnamon', '_baths', '_mindful', '_yogurt', '_soak', '_broth', '_indirectly', '_grilled', '_swear', '_traumatic', 'naments', '_simmer', '_carrots', '_chili', '_infused', '_jug', '_squash']

List 209 (20 tokens)
['_asleep', '_thankful', '_amazed', '_noon', '_risen', '_stirring', '_canceled', '_frankly', '_unchanged', '_embarrassed', '_simmer', '_cooled', '_ashamed', '_tasted', '_postponed', '_brewing', '_staged', '_detained', '_loudly', '_contempt']

List 210 (20 tokens)
['_midst', '_compartment', '_forefront', '_brighter', '_diligence', '_propriet', '_sweetness', '_sew', '_whip', '_slightest', '_Sac', '_epit', '_caliber', '_knots', '_envy', '_ache', '_monk', '_Vu', '_robe', '_imprint']

List 211 (20 tokens)
['pper', '_Francisco', '_Singh', '_wanna', '_Dakota', '_Potter', '_Cruz', '_Sr', '_thou', '_Feng', '_Kumar', '_Rico', '_goodbye', '_Xi', '_Xu', '_Hopkins', '_Bach', '_Ming', '_Wei', '_Huang']

List 212 (20 tokens)
['_\ufffd', '_Airlines', '_Industries', '_INFO', '_Psychology', '_Foods', '_Programme', '_Acid', '_Surgery', '_Chevrolet', '_Television', '_Engineers', '_Election', '_Assistance', '_Philosophy', '_Nurse', '_Infrastructure', '_Devices', '_Pour', '_Nurs']

List 213 (20 tokens)
["'re", '|||PHONE_NUMBER|||', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '|||EMAIL_ADDRESS|||', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd']

List 214 (20 tokens)
['\ufffd', '|||IP_ADDRESS|||', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '3333', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd']

List 215 (20 tokens)
['_feeder', '_litter', '_gatherings', '_trafficking', '_negligence', '_soils', '_expenditures', '_dividends', '_terminals', '_mortgages', '_caregivers', '_minorities', '_flats', '_felony', '_receivers', '_congestion', '_illegally', '_incomes', '_GST', '_medals']

List 216 (20 tokens)
['_overlooking', '_guessing', '_mentioning', '_stirring', '_graduating', '_maneuver', '_costing', '_averaging', '_Computing', '_algebra', '_Springer', '_impacting', '_broadcasting', '_urging', '_spanning', '_dismant', '_poisoning', '_noticing', '_foc', '_commemor']

List 217 (20 tokens)
['_meantime', '_midst', '_Month', '_Heights', '_canal', '_playoffs', '_Province', '_Qing', '_forefront', '_Divine', '_basin', '_finals', '_Matter', '_Scripture', '_halls', '_Defence', '_mainland', '_LORD', '_vicinity', '_corridor']

List 218 (20 tokens)
['LINE', '|||PHONE_NUMBER|||', '\ufffd', '\u7684', '|||IP_ADDRESS|||', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '\ufffd', '3333', '\ufffd', '\ufffd']

List 219 (20 tokens)
['_grams', '_vitro', '_booster', '_artery', '_carcinoma', '_caregivers', '_vomiting', '_gland', '_Factor', '_pneumonia', '_Parkinson', '_ovarian', '_untreated', '_PTSD', '_vagina', '_glands', '_catheter', '_Sympt', '_pelvic', '_embryos']

List 220 (20 tokens)
['_hesitate', '_Islands', '_Obviously', '_Corps', '_Foods', '_kilometers', '_Province', '_Qing', '_Feng', '_Xi', '_Xu', '_embarrassed', '_mainland', '_Naturally', '_Ming', '_Wei', '_Huang', '_corridor', '_Zhao', '_Yuan']

List 221 (15 tokens)
['_remodeling', '_turf', '_coatings', '_pellet', '_rollers', '_rubbish', '_replacements', '_recycl', '_illumination', '_fibres', '_painters', '_eyew', '_butcher', '_mater', '_plating']

List 222 (20 tokens)
['verse', 'ught', '_technician', '_streak', '_LOT', '_nig', '_counselor', '_felony', '_believer', '_guitarist', '_keeper', '_libert', '_freak', '_filmmaker', '_follower', '_bum', '_survivor', '_sip', '_drummer', '_stalk']

List 223 (20 tokens)
["'ve", 'pper', '_hesitate', '_complain', '_donate', '_meantime', '_wanna', '_amazed', '_overlook', '_owe', '_bless', '_enroll', '_bothered', '_enthusi', '_unaware', '_frankly', '_rented', '_thereafter', '_swear', '_tempted']

List 224 (20 tokens)
['hour', 'door', 'jack', 'Hey', 'yard', '\ufffd', 'shirt', 'bing', '\u0441', 'Wow', 'lass', 'ted', 'desc', 'century', 'Cong', 'risk', '\u043f', 'fol', 'vet', 'azure']

List 225 (20 tokens)
['_enthusiasts', '_gatherings', '_Television', '_halls', '_crews', '_theaters', '_rehears', '_minorities', '_cues', '_pursu', '_deeds', '_ceremonies', '_fantasies', '_alley', '_tempo', '_gestures', '_pillars', '_landmarks', '_madness', '_tendencies']

List 226 (20 tokens)
['_bulb', '_feeder', '_booster', '_burner', '_cutter', '_molding', '_barrels', '_pans', '_fuse', '_terminals', '_KS', '_amplifier', '_staircase', '_locom', '_coatings', '_enclosure', '_pou', '_pellet', '_rollers', '_cubes']

List 227 (20 tokens)
['inos', 'urer', 'pect', 'orted', 'mith', 'hest', 'inar', 'posed', 'inks', 'urable', 'phalt', 'immer', 'ply', 'quer', 'ensor', 'cription', 'idential', 'pered', 'olver', 'heets']

List 228 (20 tokens)
['ties', '_euros', '_awhile', '_skiing', '_2030', '_starters', '_Romania', '_surfing', '_theirs', '_Ethiopia', '_Croatia', '_Hispanic', '_Morocco', '_Alban', '_Mold', '_Sloven', '_Bulgaria', '_Libya', '_Latino', '_violating']

List 229 (20 tokens)
['_sq', '_metres', '_thereof', '_ceil', '_overlooking', '-$', '_burner', '_(+', '_baths', '**.', '"...', ',[', '_flats', '_sans', '_vertically', '/$', '/#', '_horizontally', '_sideways', '_squared']

List 230 (20 tokens)
['\ufffd', 'inos', '\ufffd', '\ufffd', '\ufffd', '\ufffd', 'umm', '\ufffd', 'acks', '\ufffd', '\ufffd', '\ufffd', '\ufffd', 'ories', 'akers', 'ags', 'inding', 'urer', '\ufffd', 'aving']

List 231 (20 tokens)
['%,', '%)', '_(!', '\u20ac', '%;', '_Nations', '\xd7', '_sq', '_kg', '_lbs', '%),', '_\u043c', '_\xa7', '_cents', '%).', '_grams', '_metres', '_mins', '_trillion', '_Degree']

List 232 (20 tokens)
['_Angeles', '_Francisco', '_meantime', '_innings', '_streak', '_playoffs', '_opener', '_haul', '_overtime', '_touchdown', '_starters', '_Rangers', '_finals', '_leagues', '_Bears', '_rebounds', '_volley', '_turf', '_averaging', '_inning']

List 233 (20 tokens)
['_diligence', '_obsess', '_contempt', '_sacrifices', '_sensations', '_obedience', '_intellig', '_ethic', '_innocence', '_immersion', '_instincts', '_humility', '_anx', '_acknowled', '_resentment', 'experience', '_seriousness', '_arrog', '_blindness', '_brutality']

List 234 (20 tokens)
['_Sciences', '_Directors', '_Associates', '_coordinator', '_enthusi', '_periodically', '_Excellence', '_indirectly', '_diligence', '_analyse', '_Regulations', '_Hispanic', '_refine', '_empowered', '_digitally', '_delegate', '_respons', '_obsess', '_allocate', '_aggressively']

List 235 (20 tokens)
['_overtime', '_negligence', '_expenditures', '_dividends', '_felony', '_nominee', '_identifiable', '_insurer', '_firearm', '_deduction', '_accountant', '_infring', '_investigator', '_unreasonable', '_faction', '_breaches', '_disrespect', '_misconduct', '_survivor', '_convictions']

List 236 (20 tokens)
['_Orleans', '_syrup', '_Hampshire', '_canal', '_cinnamon', '_litter', '_peppers', '_yogurt', '_broth', '_theirs', '_grilled', '_barrels', '_simmer', '_turf', '_noodles', '_carrots', '_nig', '_Ming', '_chili', '_squash']

List 237 (20 tokens)
['_booster', '_guessing', '_fuse', '_tracker', '_pistol', '_staged', '_deduction', '_trilogy', '_glor', '_slab', '_modem', '_massively', '_remake', '_fungus', '_tweak', '_meaningless', '_replay', '_surreal', '_mash', '_subdivision']

List 238 (20 tokens)
['_Psychology', '_trafficking', '_Infrastructure', '_costing', '_Dynamics', '_Improvement', '_wastewater', '_Decision', '_hydration', '_transplantation', '_throughput', '_refrig', '_occupancy', '_evangel', '_lia', '_steward', '_hosp', '_facilit', '_activism', '_purification']

List 239 (20 tokens)
['_grandchildren', '_IRA', '_LORD', '_rains', '_Yankees', '_Orthodox', '_injustice', '_advers', '_shortages', '_slightest', '_tariffs', '_covenant', '_resurrection', '_faction', '_sacrifices', '_imprisonment', '_caliber', '_imminent', '_hardship', '_righteousness']

List 240 (20 tokens)
['_serialized', "_'@", '_readonly', '_embed', '_encoded', '_broadcasting', '_WAS', '_RX', '_Layer', '_ETH', '_iter', '_lookup', '_CLI', '_VT', '_dynamically', '_deferred', '_UTC', '_Gn', '_Linked', '_RG']

List 241 (20 tokens)
['verse', 'ught', '_stroll', '_litter', 'ffic', '_brighter', '_traumatic', '_counselor', 'uras', '_clerk', '_felony', '_believer', '_pistol', '_cemetery', '_locker', '_firearm', '_commissioner', '_Latino', '_probation', '_covenant']

List 242 (20 tokens)
['_107', '_Acid', '_170', '_111', '_114', '_112', '_230', '_Chevrolet', '_DEL', '_117', '_113', '_108', '_106', '_BACK', '_103', '_118', '_320', '_119', '_4000', '_Reserved']

List 243 (15 tokens)
['_grilled', '_sandwiches', '_noodles', '_chili', '_650', '_coatings', '_brewing', '_consoles', '_pou', '_flowering', '_compartments', '_eyew', '_denomin', '_patriotic', '_programmable']

List 244 (20 tokens)
['_miR', '_kinase', '_cortex', '_gland', '_phosphate', '_microscopy', '_ovarian', '_detectors', '_conditioned', '_investigator', '_epithelial', '_diluted', '_achievable', '_LR', '_tandem', '_throughput', '_reflux', '_Ct', '_catheter', '_nanoparticles']

List 245 (20 tokens)
['_Authority', '_Amendment', '_Province', '_Divine', '_negligence', '_propriet', '_Regulations', '_Scheme', '_Contest', '_Administrator', '_insurer', '_commissioner', '_deduction', '_infring', '_Wizard', '_purch', '_covenant', '_Chancellor', '_Treaty', '_Courts']

List 246 (20 tokens)
["'re", 'agra', 'define', 'pper', 'choice', 'header', 'pped', 'bool', 'padding', 'query', 'tered', 'sized', 'ctions', 'crete', 'aping', 'cing', 'append', 'params', 'ples', 'pect']

List 247 (20 tokens)
['_Chevrolet', '_Feng', '_Fortune', '_Xi', '_Peak', '_Emirates', '_Powell', '_McD', '_JA', '_PBS', '_Mold', '_Cambodia', '_Marathon', '_Via', '_SERV', '_Simpson', '_Peterson', '_Kazakh', '_Cheng', '_Eld']

List 248 (20 tokens)
['///', '/*', 'LINE', '\u7684', 'However', '});', '‘‘‘‘', '%,', '%)', 'During', '"""', '\r\n\t\t',
'\u3001', '<!--', '###', '%;', '>\\', ';</', '\u3092', 'Override']
List 249 (20 tokens)
['_serialized', '_canceled', '_NUM', '_TODO', '_"%', '_packets', '_chunks', '_readonly', '_mapped',
'_"<', '_encoded', "_'<", '_RX', '_nullptr', '_nam', '_incorrectly', '_overload', '_tf', '_tmp', '_seg']
List 250 (20 tokens)
['_sourced', '_nonsense', '_leagues', '_tout', '_crap', '_rubbish', '_ignorant', '_replacements',
'_achievable', '_preferable', '_massively', '_slated', '_billed', '_readable', '_REALLY', '_centrally',
'_GOOD', '_loosely', '_nicer', '_worthless']
List 251 (20 tokens)
['_stirring', '_periodically', '_basin', '_altitude', '_indirectly', '_soils', '_derive', '_threaten',
'_cooled', '_chambers', '_spur', '_migrate', '_highways', '_landmarks', '_shores', '_regimes',
'_progressively', '_carbonate', '_tendencies', '_plunge']
List 252 (20 tokens)
['_Authority', '_Amendment', '_Corps', '_Programme', '_Initiative', '_IRA', '_2030', '_FY', '_Defence',
'_Regulations', '_Scheme', '_Reform', '_expedition', '_Palestine', '_GST', '_hearings', '_Facility',
'_Kuwait', '_Commit', '_Organisation']
List 253 (20 tokens)
['_Authority', '_IRA', '_inception', '_tracker', '_Administrator', '_predecessor', '_GST',
'_foreclosure', '_insurer', '_deduction', '_tariffs', '_Funds', '_Registry', '_Navigation', '_modem',
'_Savings', '_Debt', '_Courts', '_Funding', '_originating']
List 254 (20 tokens)
['ved', 'ATION', 'ving', 'aming', 'ERS', 'bers', 'ctions', 'sters', 'lier', 'commerce', 'rica', 'ITY',
'verse', 'scape', 'FFF', 'atable', 'ACE', 'ctor', 'inar', 'URE']
List 255 (20 tokens)
['ATION', 'ERS', 'ACK', 'IVE', 'ESS', 'ITY', 'ACE', 'URE', 'ACH', 'IES', 'MENT', 'ITE', 'ABLE', 'IGHT',
'ships', 'OOK', 'ICK', 'ERT', 'OCK', '\x99']