# Token Erasure as a Footprint of Implicit Vocabulary Items in LLMs

Sheridan Feucht David Atkinson Byron C. Wallace David Bau Northeastern University

{feucht.s, atkinson.da, b.wallace, d.bau}@northeastern.edu

## Abstract

LLMs process text as sequences of tokens that roughly correspond to words, where less common words are represented by multiple tokens. However, individual tokens are often semantically unrelated to the meanings of the words/concepts they comprise. For example, Llama-2-7b's tokenizer splits the word "northeastern" into the tokens [\_n, ort, he, astern], none of which correspond to semantically meaningful units like "north" or "east." Similarly, the overall meanings of named entities like "Neil Young" and multi-word expressions like "break a leg" cannot be directly inferred from their constituent tokens. Mechanistically, how do LLMs convert such arbitrary groups of tokens into useful higher-level representations? In this work, we find that last token representations of named entities and multi-token words exhibit a pronounced "erasure" effect, where information about previous and current tokens is rapidly forgotten in early layers. Using this observation, we propose a method to "read out" the implicit vocabulary of an autoregressive LLM by examining differences in token representations across layers, and present results of this method for Llama-2-7b and Llama-3-8b. To our knowledge, this is the first attempt to probe the implicit vocabulary of an LLM.<sup>1</sup>

### 1 Introduction

Despite their widespread use, the specific mechanisms by which LLMs are able to "understand" and generate coherent text are not well understood. One mystery is the process by which groups of subword tokens are converted into meaningful representations, a process described by Elhage et al., 2022 and Gurnee et al., 2023 as *detokenization*.

Current language models process text as a series of tokens drawn from a set token vocabulary: One token can correspond to a single word (\_fish),

Mon.k's compos.itions and impro.vis.ations feature dis.son.ances and angular mel.od.ic tw.ists, often using flat nin.th.s, flat fifth.s, unexpected chrom.atic notes together, low bass notes and st.ride, and fast whole tone runs, combining a highly per.cuss.ive attack with ab.rupt, dram.atic use of switched key releases, sil.ences, and hes.itations.

score	tokens	0.315	stride
0.582	dramatic	0.234	melodic
0.555	twists	0.203	silences
0.415	low bass	0.183	S,
0.339	flat ninths,	0.028	together,
0.321	Monk'	0.016	, and fast whole

Figure 1: We observe "erasure" of token-level information in later layers of LLMs for multi-token words and entities (top). We hypothesize that this is a result of a process that converts token embeddings into useful lexical representations, and introduce a new method for enumerating these lexical items (bottom).

or to a piece of a larger word (mon in "salmon"). The vocabulary of tokens available to a model is typically determined before training with byte-pair encoding (Sennrich et al., 2016), which is based on a specific dataset and can lead to unintuitive results. For example, Llama-2-7b's (Touvron et al., 2023) tokenizer breaks the word "northeastern" into the tokens [\_n, ort, he, astern], none of which correspond to semantically meaningful units like "north" or "east." Capitalization also creates unexpected issues: for example, the word "Hawaii" is split into two tokens if the first letter is capitalized [\_Hawai, i], but four if the first letter is lowercase [\_ha, w, ai, i]. In spite of these challenges, large models are apparently able to "understand" such idiosyncratic tokenizations of multi-token words with few observable effects on downstream performance (Gutiérrez et al., 2023), unless these weaknesses are directly targeted (Wang et al., 2024; Batsuren et al., 2024). How is this possible?

We hypothesize that during pretraining, LLMs

<sup>&</sup>lt;sup>1</sup>Code and data available at footprints.baulab.info

develop an *implicit vocabulary* that maps from groups of arbitrary tokens to semantically meaningful units. These lexical units may be multi-token words ("northeastern"), named entities ("Neil Young"), or idiomatic multi-word expressions ("break a leg") and can be understood as "item[s] that function as single unit[s] of meaning" in a model's vocabulary (Simpson, 2011). Lexical items are also non-compositional: Just as the meaning of "break a leg" cannot be predicted from the individual meanings of "break" and "leg," the meaning of "patrolling" cannot be predicted from its constituent tokens pat and rolling. This arbitrariness necessitates some kind of storage system, implicit or otherwise (Murphy, 2010).

How exactly do LLMs deal with these cases mechanistically? In this paper, we begin to answer this question by investigating token-level information stored in LLM representations.

- We find that last token positions of multitoken words and named entities "erase" tokenlevel information in early layers for both Llama-2-7b (Touvron et al., 2023) and Llama-3-8b (Meta, 2024).
- We develop a heuristic for scoring the "lexicality" of a given sequence of tokens, and use it to "read out" a list of an LLM's lexical items given a large dataset of natural text.

We interpret this erasure effect as a "footprint" of a mechanism in early layers that orchestrates the formation of meaningful lexical items.

## 2 Background

Previous work has shown that knowledge about a multi-token entity is often stored in the last token of that entity. For example, Meng et al. (2022) found that factual information about a subject like "The Space Needle" would be concentrated in the representation for le. Geva et al. (2023) find evidence for a *subject enrichment stage* during factual recall, where information about an entity is collected at its last token in early layers, which is also seen in other work on factual recall using the same dataset (Katz et al., 2024), and corroborated by research on athlete  $\rightarrow$  sport lookups (Nanda et al., 2023). This phenomenon may be due to the autoregressive nature of decoder transformer models: models cannot enrich "Space" with information about Seattle until after "Needle" is seen, as "Space" could refer

to a number of unrelated concepts ("Space Jam," "Space Station").<sup>2</sup>

Other work in interpretability has also started to uncover evidence of models encoding lexical items. Elhage et al. (2022) observe neurons in early layers that fire on the last tokens of multitoken words, names of famous people, generic nouns, compound words, and LaTeX commands. They also find late-layer neurons that seem to be relevant to retokenization, i.e., conversion from internal representations back into tokens. For example, a retokenization neuron might fire on \_st and promote rag in order to facilitate the output of the word "straggler." Gurnee et al. (2023) also find examples of polysemantic neurons in Pythia models (Mallen and Belrose, 2023) that activate for a number of multi-token constructions like "apple developer," "Bloom.ington," and "research.gate."

## 3 Linear Probing of Hidden States

### 3.1 Method

If last token positions are so important (Section 2), then what do these representations encode? Perhaps the last hidden state directly stores information about other subject tokens (e.g., \_Wars might contain some encoding for \_Star in its hidden state). To test this hypothesis, we investigate hidden states for both Llama-2-7b and Llama-3-8b, as they have significantly different token vocabulary sizes  $|\mathcal{V}|$  (32k and 128k tokens, respectively).

Let *d* denote the hidden dimension of the model. We train linear probes  $p_i^{(\ell)} : \mathbb{R}^d \to \mathbb{R}^{|\mathcal{V}|}$  to take a hidden state  $h_t^{(\ell)} \in \mathbb{R}^d$  at layer  $\ell$  and token position *t* and predict the value of a nearby token t + i. For example, a probe trained to predict the previous token for hidden states at layer 5 would be denoted by  $p_{-1}^{(5)}$ .

We train probes for all layer indexes  $0 \le l < 32$ and offsets  $i \in \{-3, -2, -1, 0, 1\}$ . We also train probes in the same manner on the embedding layer (l = -1) and on the final outputs of the network before the language modelling head (l = 32). We trained probes on a random sample of 428k tokens from the Pile (Gao et al., 2020) using AdamW for 16 epochs with a batch size of 4 and a learning rate of 0.1. Hyperparameters were selected based on validation performance on a separate Pile sample

<sup>&</sup>lt;sup>2</sup>This is not a hard-and-fast rule; it depends on entity frequency and context cues. For example, if a model sees \_The, \_E, and iff, it may already know that these tokens refer to "The Eiffel Tower" without needing to see e1 and Tower.



Figure 2: Top-1 test accuracy on COUNTERFACT subject last tokens versus other tokens in the dataset for probes trained on Llama-2-7b hidden states (n = 5063). *i* represents the position being predicted (e.g., i = -1 is previous token prediction; i = 1 is next-token prediction). We observe an "erasure" effect in last subject tokens that is not present for other types of tokens: these last subject tokens consistently "forget" about preceding tokens and themselves. Appendix A shows Llama-3-8b results and in-distribution performance on Pile tokens.

(279k tokens) after a random sweep. Each probe takes 6-8 hours to train on an RTX-A6000.

### 3.2 COUNTERFACT Subjects

After training probes in Section 3.1, we test them on the COUNTERFACT dataset (Meng et al., 2022), which consists of prompts about subjects that require factual knowledge to complete correctly (e.g. "Mount Passel is in Antarctica"). We filter the dataset to include only prompts that the model answers correctly, yielding 5,063 examples for Llama-2-7b and 5,495 examples for Llama-3-8b. To augment this dataset, we also sampled and filtered down [album/movie/series  $\rightarrow$  creator] pairs from Wikidata (Vrandečić and Krötzsch, 2014) and embedded them in prompts in the same manner, yielding a total of 12,135 correctly-answered prompts for Llama-2-7b and 13,995 for Llama-3-8b.

Figure 2 shows probe test results on COUNTER-FACT last subject tokens (right) versus every other type of token in the dataset (left). We see a striking "erasure" effect for last tokens of COUNTERFACT subjects, where these hidden states consistently "forget about" preceding and current tokens. Subject tokens that are not in the last position (e.g., \_Star) do not exhibit this pattern (Appendix A, Figure 13). This striking drop in token accuracy is reminiscent of the subject enrichment stage described by Geva et al. (2023), suggesting that the tokens \_Star and \_Wars may be overwritten in the process of representing the concept of *Star Wars*.

We also observe the same phenomenon when testing on named entities identified by spaCy in Wikipedia articles (Appendix A, Figure 12), suggesting that this effect is not an artifact of the short templates found in the COUNTERFACT dataset. It also does not seem to be a result of any imbalances in probe training data (Appendix B).

## 3.3 Multi-Token Words

Intuitively, the process of converting a multi-token sequence like [\_n, ort, he, astern] into a meaningful representation of the word "northeastern" resembles the process of converting [\_E, iff, el, Tower] into "Eiffel Tower." We hypothesize that models treat multi-token words in the same way that they treat multi-token entities, and test our probes from Section 3.1 on multi-token words. After sampling 500 articles ( $\sim$ 256k tokens) from the 20220301. en split of the Wikipedia dump (Foundation, 2022), we split by white-space to naively identify word boundaries. As predicted, we see the same "erasing" pattern for multi-token words that we do for multi-token entities (Figure 3). This suggests that they may be processed in a similar manner in early layers. Appendix A shows similar results for Llama-3-8b.

#### 4 Building a Vocabulary

After examination of probe behavior for multitoken words and entities, we hypothesize that this "erasure" effect is a result of the implicit formation of lexical representations in early layers. To characterize this phenomenon, we propose an *erasure score*  $\psi$  to identify token sequences that follow the pattern observed in Section 3. We then introduce an approach to "reading out" a list of implicit vocabulary entries for a given model using this score.



Figure 3: Top-1 test accuracy of probes on last tokens of Wikipedia multi-token words for Llama-2-7b (n = 80606). Accuracy on all other tokens shown on the left. We see an erasing effect for multi-token words, similar to the effect seen for COUNTERFACT subjects in Figure 2.

#### 4.1 An Erasure Score

Given some arbitrary sequence of tokens from indices p through q, we want to design an *erasure score* that captures intuitions from Section 3. This score should be higher for sequences exhibiting token erasure (which we hypothesize to be lexical items like [\_Cal, g, ary]), and lower for other types of token sequences (e.g., [\_go \_to, \_Cal, g]). We design a metric  $\psi_{p,q}$  that uses probe outputs from Section 3 to look for erasure effects betwen layer 1 and layer L.<sup>3</sup>

Concretely, Equation 1 defines the score  $\psi_{p,q}$  for a sequence  $s_{p,q}$  of length n = q - p + 1 as:

$$\frac{1}{1+2n} \left( \delta(q,0) + \sum_{t=p}^{q} \sum_{i=-2}^{-1} \mathbb{1}_{\text{within}}(t,i) \cdot \delta(t,i) \right)$$
(1)

where  $\delta(t, i)$  denotes the change in probability of the predicted token t + i from layer 1 to layer L, based on probes  $p_i^{(\ell)}$  from Section 3.1. We take the softmax of the probe outputs to obtain the probability of a specific token t + i in Equation 2.

$$\delta(t,i) = P_{p_i^{(1)}}(t+i|h_t^{(1)}) - P_{p_i^{(L)}}(t+i|h_t^{(L)})$$
(2)

Finally, if t + i lies outside the boundaries of s, we want the score to decrease. If it is within the boundaries of s, we want a large drop between layers  $\delta(t, i)$  to increase the value of  $\psi_{p,q}$ .

$$\mathbb{1}_{\text{within}}(t,i) = \begin{cases} -1 \text{ if } t+i < p\\ 1 \text{ else} \end{cases}$$
(3)

In summary, for every token position  $p \le t \le q$ and prediction offset  $i \in \{-2, -1\}$ , we measure the drop in the predicted probability of the correct token t + i between layer 1 and layer L. The more that the probability of the correct answer decreases in early layers, the higher we score that sequence. However, if this "forgetting" occurs for tokens outside of the boundaries of s, we subtract that value from the overall score, effectively penalizing the sequence. This intuition comes from close inspection of probe behavior-for example, Figure 13 shows that there is no "forgetting" effect for i = -1 when probing the first token of COUNTERFACT subjects. With this approach, we can also account for cases where s is a subsequence of a larger lexical item: if the token g shows a forgetting effect for \_Cal in [\_Cal, g, ary], then the sequence [g, ary] would be penalized. Finally,  $\delta(q, 0)$  additionally rewards sequences in which the last token "forgets itself," as seen in Figures 2 and 3. We then normalize by the total number of  $\delta$  values considered, in order to account for differing sequence lengths.

#### 4.2 Segmenting Documents

We develop an algorithm built around our erasure score  $\psi$  that breaks any given document  $d \in D$  into high-scoring, non-overlapping segments covering all of d (Algorithm 1). Figure 1 shows the topscoring sequences  $s_{p,q}$  calculated in this manner from a Wikipedia excerpt about Thelonious Monk, where unigram scores are excluded for clarity. Not all multi-token words are scored highly via our approach, but the highest-scoring sequences are plausible lexical items that are non-compositional in nature ("dram.atic", "sil.ences", "tw.ists"). We share examples of more documents with complete segmentations in Appendix D.

<sup>&</sup>lt;sup>3</sup>For both Llama-2-7b and Llama-3-8b we set L = 9.

Algorithm	1 Documer	nt Segmentation
-----------	-----------	-----------------

Rec	<b>quire:</b> document $d \in \mathcal{D}$ of length $l$
1:	for $n = 1$ to $l$ do $\triangleright$ all ngram lengths
2:	for $p=0$ to $l-n$ do
3:	for $q = p + n - 1$ to $l - 1$ do
4:	assign score $\psi_{p,q}$ to sequence $s_{p,q}$
5:	end for
6:	end for
7:	end for
8:	sort $s$ in descending order of $\psi$
9:	$segms \leftarrow \emptyset$
10:	for $s_{p,q}$ in sorted s do
11:	if $\forall s_{x,y} \in segms, (x > q \lor y < p)$ then
12:	$segms \leftarrow segms \cup \{s_{p,q}\}$
13:	end if
14:	end for
15:	<b>return</b> segms > non-overlapping segments

Token Sequence	n	ct	$\psi$
lower case	3	2	0.736012
storm	2	4	0.716379
excursion	4	2	0.713134
==== (72 'equals' signs)	8	2	0.712982
Mom	3	2	0.706778
acre	3	2	0.629213
Subject	3	2	0.607172
ninth	3	2	0.606669
processing elements	3	2	0.599549
CVC	3	2	0.596735

Table 1: Top ten highest-scoring sequences for Llama-2-7b using a Pile subsample (1658 sequences recovered total). n is the number of tokens in the sequence, and 'ct' represents occurrences of this segment.  $\psi$  is averaged over all occurrences.

# 4.3 Model Vocabularies

Finally, we propose a method to "read out" the implicit vocabulary of a model  $\mathcal{M}$  given a dataset  $\mathcal{D}$ . For each document  $d \in \mathcal{D}$ , we segment d using Algorithm 1. We then average scores  $\psi$  for every multi-token sequence that appears more than once in  $\mathcal{D}$ . As this process is very data-dependent, we show results for both Pile and Wikipedia text. The top 50 sequences for each dataset and model are provided in Appendix E.

With this approach, we are able to recover  $\sim$ 1800 sequences for Llama-2-7b and  $\sim$ 900 for Llama-3-8b using the same five hundred Wikipedia articles. Although recall is quite low (Table 2),

		M	ΓW	MTE		
llama	data	prec.	recall	prec.	recall	
2-7b	wiki	0.306	0.016	0.143	0.016	
	pile	0.296	0.017	0.080	0.018	
3-8b	wiki	0.044	0.001	0.010	0.000	
	pile	0.023	0.001	0.012	0.001	

Table 2: Precision and recall for aggregated results of Algorithm 1 run on Llama-2-7b and Llama-3-8b, using either Wikipedia or Pile documents (|D| = 500). MTW refers to all multi-token words in the dataset when split by whitespace; MTE refers to all spaCy named entities.

we find that 44.9% of sequences recovered for Llama-2-7b on Wikipedia text are either multitoken words or multi-token entities (29.68% for Pile text). For Llama-3-8b, only 5% and 3% of retrieved sequences are multi-token words or entities. However, looking at examples of Llama-3-8b sequences in Appendix E, we can observe other interesting cases, like multi-token expressions ("gold medalists," "by per capita income," "thank you for your understanding") and LaTeX commands (as similarly observed by Elhage et al. (2022)). Because Llama-3-8b's token vocabulary is four times larger than Llama-2-7b's, its *implicit* vocabulary also seems to consist of larger multi-word expressions and chunks of code rather than multi-token words (Appendix E, Table 7).

#### 5 Conclusion

In this work, we present preliminary evidence for the existence of an *implicit vocabulary* that allows models to convert from byte-pair encoded tokens to useful lexical items. We posit that the "erasure" effect we observe for Llama-2-7b and Llama-3-8b is a result of model processes that deal with multi-token expressions, and use this insight to propose a new method for "reading out" an LLM's implicit vocabulary. This is a first step towards understanding the formation of lexical representations in LLMs, and may serve as a useful tool for elucidation of words that a given model "knows."

## Limitations

Evaluation of implicit vocabulary-building methods (Section 4) is challenging due to the lack of a known ground-truth. Our approach is motivated by the desire to understand the inner workings of the model being studied, but we have no authoritative reference that distinguishes between situations where a given sequence gets a high  $\psi$  value because it is truly treated as a lexical unit by the model, or where it may be due to an error in our methodology. To quantify our results, we have compared the extracted vocbulary to sequences that we assume to be likely lexical items: multi-token words and spaCy named entities. However, this likely does not cover all cases for which "token grouping" occurs in LLMs.

Another limitation of this work is that we have restricted our analysis to *known* entities. There is also the question of what happens for intermediate cases such as plausible-sounding fictional towns or names of people who are not famous. If  $\psi$  correlates with sequence presence in training data, these results could be useful for understanding how familiar an LLM is with a given word or entity.

Finally, our measurements have been run only on the Llama family of models and do not yet extend to non-Llama models of comparable size, or Llama models of larger sizes.

## **Ethics Statement**

In this work, we restrict our analysis to English words, due to our biases as native speakers of English. We hope that this work can also provide valuable insights for other languages, especially lowresource languages, where understanding "what words an LLM knows" may be especially useful.

## Acknowledgments

We thank Koyena Pal, David Smith, Bilal Chughtai, Chantal Shaib, Atticus Geiger, and Adrian Chang for helpful discussion and feedback throughout the course of this project. This work was supported in part by Open Philanthropy, and by the National Science Foundation (NSF) grant IIS-1901117.

Experiments were implemented using the nnsight library (Fiotto-Kaufman et al., 2024). Many were run on the Center for AI Safety Compute Cluster. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors.

## References

Khuyagbaatar Batsuren, Ekaterina Vylomova, Verna Dankers, Tsetsuukhei Delgerbaatar, Omri Uzan, Yuval Pinter, and Gábor Bella. 2024. Evaluating subword tokenization: Alien subword composition and oov generalization challenge. *Preprint*, arXiv:2404.13292.

- Nelson Elhage, Tristan Hume, Catherine Olsson, Neel Nanda, Tom Henighan, Scott Johnston, Sheer ElShowk, Nicholas Joseph, Nova DasSarma, Ben Mann, Danny Hernandez, Amanda Askell, Kamal Ndousse, Andy Jones, Dawn Drain, Anna Chen, Yuntao Bai, Deep Ganguli, Liane Lovitt, Zac Hatfield-Dodds, Jackson Kernion, Tom Conerly, Shauna Kravec, Stanislav Fort, Saurav Kadavath, Josh Jacobson, Eli Tran-Johnson, Jared Kaplan, Jack Clark, Tom Brown, Sam McCandlish, Dario Amodei, and Christopher Olah. 2022. Softmax linear units. *Transformer Circuits Thread*. Https://transformercircuits.pub/2022/solu/index.html.
- Jaden Fiotto-Kaufman, Alexander R Loftus, Eric Todd, Jannik Brinkmann, Caden Juang, Koyena Pal, Can Rager, Aaron Mueller, Samuel Marks, Arnab Sen Sharma, Francesca Lucchetti, Michael Ripa, Adam Belfki, Nikhil Prakash, Sumeet Multani, Carla Brodley, Arjun Guha, Jonathan Bell, Byron Wallace, and David Bau. 2024. Nnsight and ndif: Democratizing access to foundation model internals. *Preprint*, arXiv:2407.14561.

Wikimedia Foundation. 2022. Wikimedia downloads.

- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.
- Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. *ArXiv*, abs/2304.14767.
- Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. Finding neurons in a haystack: Case studies with sparse probing. *Preprint*, arXiv:2305.01610.
- Bernal Jiménez Gutiérrez, Huan Sun, and Yu Su. 2023. Biomedical language models are robust to suboptimal tokenization. *Preprint*, arXiv:2306.17649.
- Shahar Katz, Yonatan Belinkov, Mor Geva, and Lior Wolf. 2024. Backward lens: Projecting language model gradients into the vocabulary space. *Preprint*, arXiv:2402.12865.
- Alex Mallen and Nora Belrose. 2023. Eliciting latent knowledge from quirky language models. *Preprint*, arXiv:2312.01037.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Neural Information Processing Systems*.
- Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date.

- M. Lynne Murphy. 2010. *Lexical Meaning*. Cambridge Textbooks in Linguistics. Cambridge University Press.
- Neel Nanda, Senthooran Rajamanoharan, János Krámar, and Rohin Shah. 2023. Fact finding: Attempting to reverse-engineer factual recall on the neuron level.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- James Simpson. 2011. The Routledge handbook of applied linguistics. Taylor & Francis.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and finetuned chat models. Preprint, arXiv:2307.09288.
- Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.
- Dixuan Wang, Yanda Li, Junyuan Jiang, Zepeng Ding, Guochao Jiang, Jiaqing Liang, and Deqing Yang. 2024. Tokenization matters! degrading large language models through challenging their tokenization. *Preprint*, arXiv:2405.17067.

### A Additional Probing Results

#### A.1 Llama-3-8b Results

**COUNTERFACT Accuracy** We share results analogous to Figure 2 for Llama-3-8b, which shows a similar "erasure" pattern (Figure 9). Probes are tested only on prompts that Llama-3-8b answers correctly.

**Multi-Token Word Accuracy** Figure 10 shows results for Llama-3-8b probes tested on the last token positions of multi-token words from Wikipedia (where "words" are determined by whitespace separation).

**Multi-Token Entity Accuracy** Figure 11 shows results for probes tested on the last token positions of multi-token entities identified by spaCy, using the same dataset that we do for multi-token words. We use spaCy's named entity recognition pipeline to identify named entities. Because digits 0-9 are added to Llama-2-7b's vocabulary, we filter out all classes relating to numbers (PERCENT, DATE, CARDINAL, TIME, ORDINAL, MONEY, QUANTITY), with the thought that these sequences may be treated differently at the detokenization stage.

### A.2 Llama-2-7b Results

**Multi-Token Entity Accuracy** Figure 12 shows results for Llama-2-7b probes tested on multi-token entities from Wikipedia, using the same dataset from Section 3.3 and also filtering out number-based entity classes as in Section A.1.

**Pile Accuracy** While Figure 2 shows test accuracy of linear probes on model hidden states, Figure 4 shows in-distribution test accuracy on Pile tokens. We can observe a smoother trajectory of gradual "forgetting" of previous and current tokenlevel information throughout layers.

**Comparison of Token Positions** Figure 13 shows the breakdown of probe performance on different types of subject tokens: first subject tokens, middle subject tokens, and last subject tokens. We see that the observed drop in previous and current token representation observed in last subject tokens still exists, but is not as drastic for first and middle subject tokens.

**Comparison of Subject Lengths** We also show previous token representation broken down by



Figure 4: Overall test accuracy on unseen Pile tokens (n = 273k) for probes trained on Llama-2-7b hidden states. Next token prediction becomes more accurate throughout model layers as current and previous token accuracy decreases.

COUNTERFACT subject length for last token representations in Figure 14. Unigram subjects represent previous token information at a rate even higher than non-subject tokens. For bigrams and trigrams, we see a pattern similar to Figure 2.

# B Accounting for Possible Training Imbalance

One explanation for the observed drop in accuracy for COUNTERFACT entities across layers is that our probes have simply not been exposed to as many entity tokens during training. We do not believe this is the case for Llama-2-7b for two reasons: (1) If this effect was due to probes being less sensitive to tokens found in multi-token entities, we would also see a significant drop for first and middle tokens, which does not occur (Figure 13). (2) We measure the frequency of all test n-grams in the original Pile data used to train our probes, and find that both subject and non-subject n-grams are found in the probe training dataset at similar rates, with the median number of occurrences in the test set for both types of sequences being zero. After removing the few non-subject sequences that do appear often in the probe training set, we still see the same "erasure" effect.

# **C** Choice of *L*

We choose L = 9 based on probe behavior for Llama-2-7b and Llama-3-8b, particularly in Figures 2 and 3. Table 3 shows an additional ablation experiment for  $L \in \{5, 9, 13, 17, 21\}$ .

	M	MTE			
L	prec.	recall	prec.	recall	
5	0.307	0.002	0.143	0.002	
9	0.306	0.016	0.143	0.016	
13	0.328	0.003	0.169	0.003	
17	0.330	0.003	0.180	0.003	
21	0.319	0.003	0.172	0.003	

Table 3: Precision and recall for different values of L for Algorithm 1 applied to Llama-2-7b on Wikipedia text. Recall seems to be best for L = 9, with precision improving by a few points in mid-late layers.



Figure 5: Full segmentation of a document from Wikipedia via Algorithm 1 on Llama-2-7b. Borders indicate segmentation, with bolded letters indicating multi-token segments. Darker blue cells have higher scores, yellow cells have negative scores. The highest-scoring sequence in this document is "Australian Institute" ( $\psi = 0.579$ ).

## **D** Document Segmentation

We provide full document segmentations using Algorithm 1 for a short excerpt from a Wikipedia article in Figures 5 and 6. Figures 7 and 8 show segmentations for a Pile document.

## **E** Model Vocabularies

Tables 4 through 7 show the top 50 highest-scoring multi-token sequences for Llama-2-7b and Llama-3-8b across either five hundred Wikipedia articles or five hundred Pile samples. Entries were filtered to show only sequences that appear more than once.

See D ana e Su z anna S weet apple is an Australian Par al ym pic sw immer . She was born in the Queensland town of St George. Sweet apple attended board ing school at 1 I and has a Bach elor of Arts in Liter ature . She took up swimming in 19 9 0. Her early sw imming results led to her being offered one of the first Australian Institute of Sport scholar ships for disabled sw immers. At the 19 9 2 Barcelona Games, she won a silver medal in the Women's 10 0 m Freestyle B2 event and she won two bronze medals in the Women [] S After the Games she commented "II" d be iso happy if more people could make movement and sport a way of life. It is a great way to meet people and gain confidence ." Sweetapple was the Young Queensland er of the Year in 1992. Re ferences Female Paralym pic swimm ers of Australia Sw immers at the 199 2 Summer Par alym pics Par

Figure 6: Full segmentation of a document from Wikipedia via Algorithm 1 on Llama-3-8b. Borders indicate segmentation, with bolded letters indicating multi-token segments. Darker blue cells have higher scores, yellow cells have negative scores. The highest-scoring sequence in this document is ". After the Games she commented "" ( $\psi = 0.443$ ).



Figure 7: Full segmentation of a document from the Pile via Algorithm 1 on Llama-2-7b. Borders indicate segmentation, with bolded letters indicating multi-token segments. Darker blue cells have higher scores, yellow cells have negative scores. The highest-scoring sequence in this document is "submodel" ( $\psi = 0.559$ ).

Image: Solution of the solution

Figure 8: Full segmentation of a document from the Pile via Algorithm 1 on Llama-3-8b. Borders indicate segmentation, with bolded letters indicating multi-token segments. Darker blue cells have higher scores, yellow cells have negative scores. The highest-scoring sequence in this document is "re really brave:" ( $\psi = 0.634$ ).



Figure 9: Test accuracy on COUNTERFACT subject last tokens versus other tokens in the dataset for probes trained on **Llama-3-8b** (n = 5495). *i* represents the position being predicted (e.g., i = -1 is previous token prediction; i = 1 is next-token prediction). We observe an "erasure" effect similar to Figure 2.



Figure 10: Test accuracy of probes on last tokens of Wikipedia **multi-token words** for probes trained on **Llama-3-8b** (n = 91935; right). Test accuracy on all other tokens shown on the left. Similarly to Figure 2, we see an erasing effect that is not present for other types of tokens.



Figure 11: Test accuracy of probes on last tokens of Wikipedia **multi-token entities** for probes trained on **Llama-3-8b** (n = 36723; right). Test accuracy on all other tokens shown on the left. Entities are identified via spaCy named entity recognition, excluding entity types that include digits.



Figure 12: Test accuracy of probes on last tokens of Wikipedia **multi-token entities** for **Llama-2-7b** (n = 36723; right). Test accuracy on all other tokens shown on the left. Entities are identified via spaCy named entity recognition, excluding entity types that include digits.



Figure 13: Breakdown for Section 3 probes tested on COUNTERFACT first subject tokens, middle subject tokens, and last subject tokens. We observe an "erasing" effect only for last subject tokens. Because BOS tokens are recoverable by i = -1 probes at high rates, and since 55% of prompts tested on had subjects at the beginning, we filter examples for which BOS tokens are labels from the leftmost plot.



Figure 14: Probe test results for COUNTERFACT subject last tokens broken down for unigrams, bigrams, and trigrams. Unigram subjects store previous token information at rates near 100%, even excluding BOS tokens.

Token Sequence	n	ct	$\psi$
Gottsche	3	2	0.685220
berth	3	2	0.680793
carries	3	2	0.647844
Eurocop	3	2	0.644104
franchises	3	2	0.642707
0 Women	3	2	0.639162
rape	3	2	0.632567
Rebell	3	3	0.614295
intermittently	4	2	0.613479
enn State	4	3	0.607535
North Dakota	4	10	0.600616
Sride	3	2	0.600013
fiction	2	$\overline{2}$	0 599339
Sox	$\frac{2}{3}$	$\frac{2}{3}$	0.599043
Bazz	3	2	0.599043
erect	3	$\frac{2}{2}$	0.590242
borough	3	$\frac{2}{3}$	0.596054
oncompassos	5	2	0.590054
northernmost	2	$\frac{2}{2}$	0.392084
Modroo	2	$\frac{2}{2}$	0.391007
lviauras	2	2	0.390394
inum inom	2	2	0.380908
1ron Calarra	2	2	0.586959
Galaxy	3	2	0.585879
began operations	3	2	0.584680
Redding	3	2	0.584244
gloss	3	2	0.5/6/40
cello	3	2	0.573732
Gators	3	5	0.573675
senator	3	2	0.572947
restructuring	4	2	0.570552
supervised	3	3	0.570421
Mediterranean	4	2	0.567790
Madera	3	2	0.567563
sequel	3	2	0.563626
scarp	3	3	0.561548
Sout	3	2	0.560640
South Division	3	2	0.558720
rectangular	3	2	0.557339
Danny	3	2	0.556836
Examiner	4	2	0.555797
Kuwait	4	4	0.554636
Bogue	3	6	0.552219
Lancaster	3	3	0.552166
Leuven	4	3	0.548806
the Park	3	2	0.548687
first Baron	3	2	0.547447
fights	3	2	0.547171
Carpio	3	$\overline{2}$	0.547116
Czech Republic	3	$\overline{2}$	0.546651
	~	_	

Table 4: Llama-2-7b Wikipedia results (1808 sequences
total). $n$ is the number of tokens in the sequence, and 'ct'
represents occurrences of this segment. $\psi$ is averaged
over all occurrences.

	Token Sequence	n	ct	$\psi$
	1992 births	7	2	0.573
	19th-century	7	3	0.569
	dehydrogen	5	2	0.553
	Swahili	4	4	0.539052
	Chuck Liddell	6	2	0.537169
	its population was	5	5	0.534977
	by per capita income	6	3	0.518991
	are brownish	4	2	0 515703
	ate women's football	7	4	0.509384
	Almeida	.4	5	0 507277
	of New South Wales	5	ž	0.503120
	2015 deaths	8	2	0.503074
	Pittsburgh	3	$\frac{2}{3}$	0.503074
	21st century	7	1	0.303070
	(NSW)	1	0	0.499302
	age of the United	4	9	0.49/10/
	Vingdom	6	2	0 487202
	Ringdonn	2	2	0.467303
	Presidential	3	2	0.485517
	Landmark	3	2	0.484965
	Alistair	4	2	0.484930
	Tauri	3	8	0.482449
	2 km	4	2	0.479984
	20th-century	7	3	0.475703
	East Bay	3	2	0.475156
	game goes in extra	10	2	0 472323
	time, if the scored	10	-	0.172323
	São Paulo	3	2	0.470874
	Atlantic City	3	2	0.470726
	Chaluk	3	2	0.467165
	Frank Lloyd	3	2	0.462585
	may refer to:	6	4	0.462234
	gold medalists	4	2	0.458494
	, 2nd Baron	6	2	0.456996
	people)	4	4	0.454926
	series aired	4	2	0.453057
	Srib	3	2	0.451708
	with blackish	4	$\overline{2}$	0.450033
	World Cup players	4	2	0 448979
	main role	3	$\frac{1}{2}$	0.448569
	Bos	4	$\frac{2}{2}$	0.448425
	Asenath	- 1	$\frac{2}{2}$	0.448250
	Royal Navy	3	$\frac{2}{3}$	0.445617
	2 Bundasliga playars	5 7	2	0.445210
	2. Buildesliga players	2	2 60	0.443210
		3 4	209	0.444921
	an unincorpor	0	2 1	0.443327
	Gast	2	4	0.43/693
	PIOT	5	2	0.432194
	Elisio de Med	5	2	0.431518
	" (2007) "Jad	12	2	0.429412
		3	2	0.428984
	Früh	3	2	0.427781
equences	order of the NK	5	2	0.424037

Table 5: **Llama-3-8b** Wikipedia results (892 sequences total). n is the number of tokens in the sequence, and 'ct' represents occurrences of this segment.  $\psi$  is averaged over all occurrences.

Token Sequencenct $\psi$ Token Sequencenct $\psi$ lower case320.7163012(d]x530.59939excursion420.7113134"h430.58813mem.72 'equals' signs)820.712982[=11] \\in740.588543acre320.607172-22007-0611230.55135subject320.5066691NSURANCE520.54933processing elements320.596735\uniternal static820.54037VPN330.596052: At40.53749bore220.599168(U1) hunt730.53066Regul320.589154(U1) hunt730.53066Rates320.589734(U1) hunt720.52949INSURANCE520.589734(U1) hunt720.52049Insurexial420.571932'un internal static920.51005princes320.567894(W1) hunt720.50105princes320.567894(W1) hunt720.49688Disease420.567894(W1) hunt420.50105princess320.546627'n.1x620.49737princess3 <td< th=""><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th></td<>																																																																																									
$\begin{array}{l c c c c c c c c c c c c c c c c c c c$	Token Sequence	n	ct	$\psi$	Token Sequence	n	ct	$\psi$																																																																																	
$\begin{array}{llllllllllllllllllllllllllllllllllll$	lower case	3	2	0.736012	\n	9	2	0.627583																																																																																	
excursion420,711313*\n430.58701mom320.712982[n=1]\in740.58543Mom320.706778.[n=1]\in740.58543acre320.6027178.[n=1]\in740.58543acre320.607178 <td>storm</td> <td>2</td> <td>4</td> <td>0.716379</td> <td>{d}x</td> <td>5</td> <td>3</td> <td>0.599395</td>	storm	2	4	0.716379	{d}x	5	3	0.599395																																																																																	
$\begin{array}{c c c c c c c c c c c c c c c c c c c $	excursion	4	2	0.713134	*\n	4	3	0.587016																																																																																	
$\begin{array}{l c c c c c c c c c c c c c c c c c c c$	==== (72 'equals' signs)	8	2	0.712982	$\{n=1\}$ (\in	7	4	0.585434																																																																																	
acre320.6029213 -22007-061-22007-0611230.55158Subject320.606669INSURANCE520.54826processing elements320.59954932;\n internal static690.54037VPN330.596052: At420.53860Regul320.5991968(2,9,'640.53749bore220.590212Respondent420.53450S\dot(G520.589714\tt\\n\n\t730.53066Rates320.589637(3,0,'')720.52730INSURANCE520.58453[n-1]\ar720.52730Commercial420.571932hydroxyl420.51005pentrated420.577894[w] (B]_{-1}(620.50633Leigh320.567894[w] (B]_{-1}(620.5097jail330.5577997+f131720.49680Disease420.564662/0 S520.49204wild type320.564662/0 S520.48360piscase420.564662/0 S520.48360Disease420.564662/0 S520.48460princes320.564662<	Mom	3	2	0.706778		8	2	0 573310																																																																																	
Subject320.607172reticulum430.54393ninth320.60669INSURANCE520.54936processing elements320.59954932;'n internal static690.54393VPN330.596052;At4420.53860Regul320.591068(2,9;'640.53749bore220.589637(3,0,')720.52949NURANCE520.588637(3,0,')640.52949INSURANCE520.584323[n-1]\ar720.52730Barney330.574872your understanding620.51005MG320.569830in mukaan520.50633Leigh320.567037'); 'nINSERT INTO6100.50105John Arena320.566480, 11.>820.49980John Arena320.566480, 11.>820.49980John Arena320.56167'', '', '', ''530.47977Jidid type320.56167'', '', '', ''520.48581Jiqid cry330.557799'', '', '', ''530.47977Jirdhdype320.56167'', '', '', ''530.47977Jirdhdype<	acre	3	2	0.629213	-2-2007-061	12	3	0 551581																																																																																	
ninh320.606669INSURANCE520.54826processing elements320.59954932;'n internal static690.54826VPN330.596052: At420.53749bore220.591068(2,9,'640.53749bore220.590121Respondent420.53450\$\slot(G520.589714\tti/\understanding420.53450\$\slot(G520.589717(1,1)'ar720.52730INSURANCE520.584323[n-1]\ar720.52730Commercial420.571932hydroxyl420.51095pentrated420.571932hydroxyl420.51095TNS320.567894 $\langle w  \hat{I}B \}_{-1}^{-1}$ 620.50197TNS320.567703')', inINSERT INTO6100.50105pedides420.566646/0 S520.49204welfare440.564662/0 S520.48486Disease420.564662/0 S520.48496uws30.55779'', internal520.49500Disease420.564662/0 S520.49504Uws30.551672'', '', '', ''	Subject	3	2	0.607172	reticulum	1 <u>2</u>	3	0.549337																																																																																	
$ \begin{array}{c c c c c c c c c c c c c c c c c c c $	ninth	3	2	0.606669	INSURANCE	5	2	0.548263																																																																																	
CVC320.56775 $\chi_n$ internal static620.54713VPN330.596052: At420.53860Regul320.591068(2,9)'640.53749bore220.590212Respondent420.53066Rates320.584323_{n-1} \ar730.53066Rates320.584323_{n-1} \ar720.52730Commercial420.581543thank you for620.51397PTA320.571932hydroxyl420.51005pentrated420.570164 $> n^*/$ private \$920.51005pentrated420.567830in mukaan520.5097jail330.567255?', 'IINSERT INTO6100.50105pentides420.566375?-f131720.49688John Arena320.566703?; 'IuriNSERT INTO6100.50105picides420.566462/0 S520.49204wild type320.55779"; 'true,'n620.49264wild type320.551672*mathate30.49264wild type320.542084,'n520.48264liquid cry330.557799<	processing elements	3	$\overline{2}$	0.599549	32.\n internal static	8	$\frac{2}{2}$	0.540203																																																																																	
VPN330.560622: At memory m	CVC	3	$\overline{2}$	0.596735	:\n internal static	6	ã	0.547875																																																																																	
Regul320.501908 0.53749(2,0,') (2,0,')640.53749 0.53749bore220.5901212 0.5380617Respondent420.53749 0.53066Sdot{G520.589637 0.52949(3,0,')640.52949 0.52949INSURANCE520.584323 0.571872 $\{n-1\}$ ar720.52730 0.52730Commercial420.571932 0.50703hydroxyl420.51005 0.50633Barney30.574872 0.50703your understanding your understanding your understanding your understanding bydroxyl420.51005 0.51005MG320.567894 0.567834im mukaan520.50633Leigh320.567003 0.56703'): NINSERT INTO6100.50105 0.50105TNS320.56703 0.566480, IL>820.49264Weilfare440.564662 0.650.5520.49264uws330.55779 0.553408i: true, N620.48860orgel430.554208i: true, N620.48680orgel430.554208i: true, N620.488436princess320.548702'siniternal530.47970ilquid cry330.557408': true, N620.488496	VPN	3	3	0.596052	· A +	4	2	0.540574																																																																																	
Negati $3$ $2$ $2$ $0.50120$ $(2,3)$ $0$ $4$ $0.53450$ Sidot [G5 $2$ $0.5890212$ Respondent $4$ $2$ $0.53450$ Rates $3$ $2$ $0.589637$ $(3,0)$ ' $6$ $4$ $0.52949$ INSURANCE $5$ $2$ $0.584323$ $_{-1}11$ ar $7$ $2$ $0.52730$ Commercial $4$ $2$ $0.571932$ hydroxyl $4$ $2$ $0.51105$ Barney $3$ $3$ $0.574872$ your understanding $6$ $2$ $0.51005$ Penetrated $4$ $2$ $0.570164$ $>\ln*Aprivate \$$ $9$ $2$ $0.51005$ MG $3$ $2$ $0.567894$ {w} $ B]_{-1}$ $6$ $2$ $0.50397$ jail $3$ $0.567225$ $7/2$ I3 $7$ $2$ $0.49688$ John Arena $3$ $2$ $0.56648$ $1L>$ $8$ $2$ $0.49688$ Disease $4$ $2$ $0.56462$ $0$ $L>$ $8$ $2$ $0.49264$ welfare $4$ $0.564364$ $5$ Audi $4$ $2$ $0.48480$ princess $3$ $2$ $0.554208$ $4,\ln$ $5$ $2$ $0.48486$ princess $3$ $2$ $0.548702$ $\gamma$ $\gamma$ $4$ $2$ $0.47371$ birthday $3$ $2$ $0.548702$ $\gamma$ $\gamma$ $4$ $2$ $0.47371$ birthday $3$ $2$ $0.548702$ $\gamma$ $\gamma$ $4$	Regul	3	2	0.591968	$(20)^{2}$	4	2 1	0.338009																																																																																	
	hore	2	$\frac{2}{2}$	0.590212	(2,9, Desmandant	0	4	0.337493																																																																																	
States320.53971+0.10111730.53060Rates320.584323 $[n-1]$ \ar720.52949INSURANCE520.584323 $[n-1]$ \ar720.52730Commercial420.581543thank you for620.51397PTA320.571932hydroxyl420.51005penetrated420.570164>\n*/Aprivate \$920.51005Jeigh320.567894(w) [B]_{-1}620.50633Leigh320.567003'); InINSERT INTO6100.50105peptides420.5657577-f131720.49688John Arena320.566462/0 S520.49204Wild type320.560699:: true, In620.48531uws330.557799:: true, In620.48531uws330.557408to as DSP520.48436princess320.54662/0 S520.48531iliquid cry330.557408to as DSP520.48450princess320.548702', ininternal530.47977birthday320.548702', ininternal530.47977birthday32 <td< td=""><td>\$\dot[G</td><td>5</td><td><math>\frac{2}{2}</math></td><td>0.590212</td><td>Kespondent</td><td>4</td><td>2</td><td>0.534509</td></td<>	\$\dot[G	5	$\frac{2}{2}$	0.590212	Kespondent	4	2	0.534509																																																																																	
Rates520.584323 2 $(3,0)$ 640.52949 0.52730Commercial420.581543 3 $[n-1]$ Var720.52730Barney330.574872 0.570164your understanding hydroxyl420.51397PTA320.571932 0.570164hydroxyl420.51005penetrated420.567894 		2	$\frac{2}{2}$	0.580627			3	0.530009																																																																																	
$ \begin{array}{c c c c c c c c c c c c c c c c c c c $		5	2	0.589057	(3,0,	0	4	0.529493																																																																																	
Commercial420.574872thank you for your understanding620.51397PTA320.574872your understanding420.51005MG320.570164>\m*Aprivate \$920.51005MG320.567894 $\{w\}$ [B]_{620.5033jail330.567225/2VZ520.50199TNS320.565648(minusan)720.49688John Arena320.565648(ninusan)420.49688Disease420.5643645Audi420.49688Wild type320.566699all that apply430.49046uws330.557799": true,\n620.48680princess320.551672**B**]{}620.48680princess320.548702;\ninternal530.47977birthday320.548171""";";530.47977third-party420.5469492.77420.47335ungust320.53488e2d-d620.47345birthday320.548504is under conversion450.47357Undres420.548504is under conversion450.47357Undres<	Commencial	5	2	0.304323	$_{n-1} $	1	2	0.527303																																																																																	
Barney3550.5/48/2your understanding020.51057PTA320.5711932hydroxyl420.51005MG320.567894 $wi{B}_{-1}$ 620.50037Jail320.567894 $wi{B}_{-1}$ 620.50597jail320.567003'): \nINSERT INTO6100.50109peptides420.5656480, 1L>820.49688John Arena320.566462(N S520.49204welfare440.564662(N S520.49204wild type320.556799": true,\n620.48680ongrel430.557799": true,\n620.48680princess320.548702:\ininternal530.47977birthday320.548702:\ininternal530.47977idense420.548702:\ininternal530.47977idense320.546507"cde="">cde=""440.47351birthday320.546507"cde=""<4	Commercial	4	2	0.581545	thank you for	6	2	0 513979																																																																																	
$\begin{array}{c c c c c c c c c c c c c c c c c c c $	Barney	3	3	0.574872	your understanding		2	0.515577																																																																																	
penetrated42 $0.5/0164$ $>h^*/private \$$ 92 $0.5/0103$ MG32 $0.569830$ in mukaan52 $0.50633$ Jail33 $0.567225$ $/2ZZ$ 52 $0.50199$ TNS32 $0.567703$ '); hINSERT INTO610 $0.50105$ peptides42 $0.565775$ $7^{-f131}$ 72 $0.49688$ Disease42 $0.564662$ $/0$ S52 $0.49204$ wild type32 $0.560699$ all that apply43 $0.49204$ wild type32 $0.560699$ all that apply43 $0.49046$ uws3 $0.557799$ ": true,\n62 $0.48680$ princess3 $0.551779$ ": true,\n62 $0.48680$ princess3 $2$ $0.548702$ :hinernal53 $0.47977$ birthday3 $2$ $0.548702$ :hinernal53 $0.47977$ birthday3 $2$ $0.546577$ "cd> $2.7$ 4 $2$ $0.47370$ third-party4 $2$ $0.546577$ "cd> $2.7$ 4 $2$ $0.47370$ third-party4 $2$ $0.546577$ "cd> $2.7$ 4 $2$ $0.47370$ third-party4 $2$ $0.546577$ "cde=""task" 3" $0.47370$ $2.7$ 4 $2$ $0.47351$ Durban $3$	PIA	3	2	0.571932	hydroxyl	4	2	0.510059																																																																																	
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	penetrated	4	2	0.5/0164	>\n*/\private \$	9	2	0.510054																																																																																	
Leigh32 $0.567894$ $\{w\}\{B\}_{-}\{$ 62 $0.50597$ jail33 $0.567225$ $/2\sqrt{2}$ 52 $0.50197$ peptides42 $0.567003$ '); uINSERT INTO610 $0.50105$ John Arena32 $0.565648$ $0, 1L>$ 82 $0.49688$ Disease42 $0.5666462$ $(0 S)$ 52 $0.49688$ welfare44 $0.564364$ 5Audi42 $0.49046$ uws33 $0.557799$ ": true,\n62 $0.48680$ ongrel43 $0.554208$ $4,\lambda$ n52 $0.48348$ penmark32 $0.554208$ $4,\lambda$ n52 $0.48348$ Denmark32 $0.551672$ $**B*1\{\}$ 62 $0.48348$ Denmark32 $0.548702$ ;\u00fc used62 $0.47372$ birthday32 $0.548702$ ;\u00fc used62 $0.47367$ TeNOENT52 $0.548171$ ", "x":53 $0.47177$ aliens32 $0.546647$ $<(td>/(td>100\% used620.47357Durban340.545848e2d-d620.47351Durban320.542762intlys530.47121Index320.532933(2.8,')64$	MG	3	2	0.569830	in mukaan	5	2	0.506333																																																																																	
jail330.567225 $/2Z$ $5$ 20.50199TNS320.567003'); InINSERT INTO6100.50105peptides420.565775'); InINSERT INTO6100.50105Disease420.564662/0 S520.49204welfare440.5643645 Audi420.49204wild type320.5666995 Audi420.49204wild type320.5666995 Audi420.49104ongrel430.557799": true,\n620.48680liquid cry330.553408to as DSP520.48680princess320.51672**B**]{}\620.48431Denmark320.548702:\ininternal530.47977birthday320.548504100% used620.47362atedmes420.548507"code=""440.47351Durban340.545848e2d-d620.47341Bouncy430.537485boolean isAny1220.47040DLS340.535933(2.8,"640.4721Index320.548101"rachea420.46315Jupist320.545826is under con	Leigh	3	2	0.567894	$\{w\}\hat{\{}B\}$ {	6	2	0.505970																																																																																	
TNS320.567003 0.567751); InINSERT INTO6100.50105peptides420.565775 0.5656487-f131720.49688John Arena320.5656462 0.1L>/0.S520.49204welfare440.564364 0.5643645Audi420.49204wild type320.5660699 0.557799all that apply430.49046uws330.557799 0.554208": true,in620.48680liquid cry330.5554208 0.551672to as DSP520.48480princess320.548702 0.548171', uninternal530.47977birthday320.548702 0.548171', uninternal530.47470TENOENT520.547169 2.72.7420.47357atedmes420.546507" code="440.47351Durban340.545848 2e2d-d620.47341Bouncy430.537485 2boolean isAny1220.47094DLS340.535933 2(2,8,'640.47021these motivational430.537485 2boolean isAny1220.46708unjust320.533137 3use in an automobile at org.apache.c4	jail	3	3	0.567225	$\frac{12}{Z}$	5	2	0.501998																																																																																	
peptides420.5657757-f1310.720.49688John Arena320.5656480, 1L>820.49680Disease420.564662/0 S520.49204welfare440.5643645 Audi420.49104wild type320.560699all that apply430.49046uws330.557799": true,\n620.48680ongrel430.5542084,\n520.48480princess320.551672**B**]{}\620.48480pomark320.548702\initernal530.47976birthday320.548171", "x":530.47470TENOENT520.546507", code="440.47351Durban340.545848e2d-d620.47341Bouncy430.545826is under conversion450.47335CHO320.53130(,\n)nprivate530.47121these motivational430.537485boolean isAny1220.47084DLS340.535933(2,8,'640.47021these motivational430.531300word around us420.46610htese motivational43<	TNS	3	2	0.567003	')· \nINSERT INTO	6	10	0 501055																																																																																	
John Arena320.5656480.1L>820.49580Disease420.564662/0 S520.49204wild type320.566669all that apply430.49046uws330.557799": true,\n620.48680ongrel430.5542084,\n520.48680liquid cry330.557498": true,\n620.48680princess320.548702": true,\n620.48831birthday320.548702": me,\n620.48348Denmark320.548702": me,\n620.47372birthday320.548171", "x":530.47977TENOENT520.5469490.47372third-party420.546949 </td <td>peptides</td> <td>4</td> <td>2</td> <td>0.565775</td> <td>7-f131</td> <td>7</td> <td><math>\hat{2}</math></td> <td>0.496881</td>	peptides	4	2	0.565775	7-f131	7	$\hat{2}$	0.496881																																																																																	
Disease42 $0.564662$ $0.5710$ $0.5710$ $2$ $0.49204$ welfare44 $0.564364$ 5Audi42 $0.49204$ wild type32 $0.560699$ all that apply43 $0.49204$ uws33 $0.557799$ ": true, in62 $0.48204$ ongrel43 $0.554208$ 4, in52 $0.48204$ princess32 $0.551672$ $*8B*1]{}$ 62 $0.48348$ Denmark32 $0.548702$ ;ininternal53 $0.47977$ birthday32 $0.548702$ ;ininternal53 $0.47977$ birthday32 $0.548702$ ;ininternal53 $0.47977$ third-party42 $0.546949$ $diens320.546507"code="440.47357Durban340.545848e2d-d620.47347Bouncy430.537485e2d-d620.47345unjust320.534510indler conversion450.47732uneven320.531300use in an automobile620.46763watt320.531300use in an automobile620.46763watt320.531300$	John Arena	3	2	0.565648	0.11 >	8	$\frac{2}{2}$	0.495809																																																																																	
welfare440.56436475 Audi420.47026wild type320.560699all that apply430.49104uws330.557799": true,\n620.48680liquid cry330.5534084,\n520.48531princess320.551672 $**B*1$ }/\620.48496princess320.548702'\ninternal530.47977birthday320.548504100% used620.47367third-party420.546507"cw": x":530.47470Purban340.545848'cde=""440.47357Bouncy430.545826is under conversion450.47347Unjust320.534813'(x)hunprivate620.47347DLS340.535933(2,8,'640.47021'n&320.534510'(x)hunprivate1220.47094bolean isAny1220.46015'usehan420.46015'n&320.531300'usehan420.46015'n320.531300'usehan420.46355'n320.531300'usehan420.46355'she320.531300'usehan </td <td>Disease</td> <td>4</td> <td>2</td> <td>0.564662</td> <td>/0 \$</td> <td>5</td> <td><math>\frac{2}{2}</math></td> <td>0.492042</td>	Disease	4	2	0.564662	/0 \$	5	$\frac{2}{2}$	0.492042																																																																																	
wild type320.560699all that apply430.49046uws330.557799": true,\n620.48680liquid cry330.553408": true,\n520.48480princess320.551672 $**B**]{}$ 620.48496Denmark320.548702":ninternal530.47977birthday320.548704100% used620.47367renders420.548171", "x":530.47977atedmes420.548694100% used620.47367third-party420.546507"code="440.47351Durban340.545848e2d-d620.47341Bouncy430.537485is under conversion450.47335CHO320.538813();\n}\n\private1220.47094DLS340.535933(2,8,')640.47021unjust320.531300use in an automobile620.46763watt320.531300world around us420.46466HP330.52955520.463100.46310HP330.52955530.46310HP330.52955530.46310 <tr <="" td=""><td>welfare</td><td>4</td><td>4</td><td>0.564364</td><td>5 Audi</td><td>1</td><td><math>\frac{2}{2}</math></td><td>0.492042</td></tr> <tr><td>uws330.557799an an appry450.49040ongrel430.554208": true,\n620.48680liquid cry330.553408": true,\n620.48680princess320.551672<math>**B**]{}</math>620.48496Denmark320.548702": minternal530.47977birthday320.548701': "x":530.47977atedmes420.548171", "x":530.47977third-party420.5469492.7420.47357pliens320.545848': code="440.47351Durban340.545848e2d-d620.47341Bouncy430.537485e2d-d620.47341Unjust320.538813();\n]\n\nprivate1220.47094Unsven320.531300use in an automobile620.46763watt320.531300use in an automobile620.46763Watt320.531300world around us420.46355HP330.52955520.463100.46310CHO320.531300world around us420.46355Na320.532243at o</td><td>wild type</td><td>3</td><td>2</td><td>0.560699</td><td>all that apply</td><td></td><td>2</td><td>0.491049</td></tr> <tr><td>ongrel43<math>0.554208</math>. thue, in62<math>0.48080</math>liquid cry33<math>0.553408</math>to as DSP52<math>0.48531</math>princess32<math>0.551672</math><math>**B**]{}</math>62<math>0.48496</math>Denmark32<math>0.548702</math>;\ninternal53<math>0.47977</math>birthday32<math>0.548504</math>100% used62<math>0.47567</math>atedmes42<math>0.548702</math>;\ninternal53<math>0.47977</math>third-party42<math>0.546507</math>", "x":53<math>0.47470</math>Third-party42<math>0.546507</math>" code="44<math>0.47351</math>Durban34<math>0.545848</math>e2d-d62<math>0.47341</math>Bouncy43<math>0.545826</math>is under conversion45<math>0.47355</math>Unjust32<math>0.542762</math>{ intlys53<math>0.47121</math>these motivational43<math>0.537485</math>boolean isAny122<math>0.47094</math>DLS32<math>0.533137</math>use in an automobile62<math>0.46778</math>watt32<math>0.532243</math>at org.apache.c75<math>0.46763</math>Watt32<math>0.531300</math>use in an automobile62<math>0.46355</math>HP33<math>0.529555</math><math>0.46302</math><math>0.463102</math><math>0.463102</math></td><td>uws</td><td>3</td><td>3</td><td>0.557799</td><td>": true \n</td><td>т 6</td><td>2</td><td>0.496907</td></tr> <tr><td>liquid cry330.5534084, u520.48351princess320.551672**B**]{}\620.48496Denmark320.548702**B**]{}\620.48496birthday320.548504100% used620.47677atedmes420.548171", "x":530.47977TENOENT520.5471692.7420.47357diens320.54650740.47351Durban340.545848e2d-d620.47341Bouncy430.545826is under conversion450.47355Unjust320.548131(j:n)\n/private530.47121these motivational430.537485boolean isAny1220.47094DLS340.535933(2.8,'640.47021watt320.531300use in an automobile620.46763HP330.52955520.46315or Commodore530.46315Cho320.531300world around us420.46355or Commodore530.46315Durban320.52955520.46355or Commodore530.46315Durban320.531300us</td><td>ongrel</td><td>4</td><td>3</td><td>0.554208</td><td>. uuc, 11</td><td>5</td><td><math>\frac{2}{2}</math></td><td>0.480807</td></tr> <tr><td>princess32<math>0.551672</math>to as DSP<math>32<math>0.48490</math>Denmark32<math>0.548702</math><math>\langle ninternal53<math>0.47977</math>birthday32<math>0.548504</math><math>100\%</math> used62<math>0.48348</math>atedmes42<math>0.548504</math><math>100\%</math> used62<math>0.47367</math>"ENOENT52<math>0.547169</math>2.742<math>0.47372</math>third-party42<math>0.546507</math>" code="44<math>0.47351</math>Durban34<math>0.545826</math>is under conversion45<math>0.47335</math>Durban32<math>0.542762</math>{ intlys53<math>0.47121</math>Unjust32<math>0.538813</math><math>(); n\} \ nprivate</math>122<math>0.47094</math>DLS32<math>0.53137</math>use in an automobile62<math>0.46703</math>watt32<math>0.531300</math>world around us42<math>0.46466</math>HP33<math>0.52955</math><math>0.46355</math><math>0.46355</math>Fields 6 Hame 2 7h pile merkt (1669 conversion 41167<math>4.20</math><math>0.46310</math></math></math></td><td>liquid crv</td><td>3</td><td>3</td><td>0.553408</td><td></td><td>5</td><td>2</td><td>0.463313</td></tr> <tr><td>Denmark32<math>0.548702</math> <math>0.48346</math>birthday32<math>0.548702</math> <math>0.48346</math><math>(1.48346)</math> <math>(1.48346)</math>birthday32<math>0.548702</math> <math>0.48346)<math>(1.48346)</math> <math>(1.48346)</math>atedmes42<math>0.548504</math> <math>0.47372)<math>(1.00\% \text{ used})</math>62<math>0.47977</math> <math>100\% \text{ used}</math>atedmes42<math>0.548171</math> <math>0.47470</math><math>(1.6\% \text{ used})</math>53<math>0.47977</math> <math>100\% \text{ used}</math>62<math>0.47367</math> <math>0.47372atedmes42<math>0.546949</math> <math>0.47372</math><math>(1.6\% \text{ used})</math>62<math>0.47372</math> <math>0.47372aliens32<math>0.546507</math> <math>0.47351<math>(1.6\% \text{ used})</math>62<math>0.47372</math> <math>0.47351Durban34<math>0.545848</math> <math>0.545826</math><math>(1.6\% \text{ used})</math>62<math>0.47351</math> <math>0.473351CHO32<math>0.542762</math> <math>0.532813<math>(1.16\% \text{ use})</math>53<math>0.47121</math> <math>0.535933Uneven32<math>0.533137</math> <math>0.529553<math>(2.8\%)</math>64<math>0.47021</math> <math>0.46466)Mat32<math>0.533137</math> <math>0.529555<math>(2.8\%)</math>64<math>0.47021</math> <math>0.46305Watt32<math>0.532243</math> <math>0.529555<math>(2.8\%)</math>64<math>0.47021</math> <math>0.46305Watt32<math>0.531300</math> <math>0.529555<math>(2.8\%)</math>64<math>0.47021</math> <math>0.46305HP33<math>0.529555</math><math>(2.8\%)</math><math>(2.8\%)</math><math>(2.8\%)</math><math>(2.8\%)</math><math>(2.8\%)</math>Low<td< math=""></td<></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></br></td><td>princess</td><td>3</td><td>2</td><td>0.551672</td><td>10 as DSF **D**1()\</td><td>5</td><td>2</td><td>0.464907</td></tr> <tr><td>birthday<math>3 = 2</math><math>0.548504</math>(minternal<math>5 = 3</math><math>0.47977</math>birthday<math>3 = 2</math><math>0.548504</math><math>100\%</math> used<math>6 = 2</math><math>0.47567</math>atedmes<math>4 = 2</math><math>0.548171</math><math>100\%</math> used<math>6 = 2</math><math>0.47470</math>third-party<math>4 = 2</math><math>0.546949</math><math>2.7</math><math>4 = 2</math><math>0.47372</math>aliens<math>3 = 2</math><math>0.546507</math><math>', "x":</math><math>5 = 3</math><math>0.47470</math>Durban<math>3 = 4</math><math>0.545848</math><math>', code=''</math><math>4 = 4</math><math>0.47351</math>Bouncy<math>4 = 3</math><math>0.545826</math><math>code=''</math><math>4 = 4</math><math>0.47351</math>CHO<math>3 = 2</math><math>0.545826</math><math>is under conversion</math><math>4 = 5</math><math>0.47335</math>Uniyst<math>3 = 2</math><math>0.538813</math><math>();\n \}\n\private</math><math>boolean isAny</math><math>12 = 2</math><math>0.47094</math>DLS<math>3 = 4</math><math>0.535933</math><math>(2,8,')</math><math>6 = 4</math><math>0.47021</math>uneven<math>3 = 2</math><math>0.531300</math><math>trachea</math><math>4 = 2</math><math>0.46763</math>Watt<math>3 = 2</math><math>0.531300</math>use in an automobile<math>6 = 2</math><math>0.46763</math>HP<math>3 = 3</math><math>0.529555</math><math>2.161(1+x)</math><math>8 = 2</math><math>0.46355</math>Field of Hermitical Action Action</td><td>Denmark</td><td>3</td><td>2</td><td>0.548702</td><td><math>\mathbf{D}</math></td><td>0</td><td>2</td><td>0.483484</td></tr> <tr><td>atedmes42<math>0.548171</math><math>100\%</math> used62<math>0.47367</math>"ENOENT52<math>0.548171</math>", "x":53<math>0.47470</math>third-party42<math>0.546949</math><math>2.7</math>42<math>0.47372</math>aliens32<math>0.546507</math>"code="44<math>0.47351</math>Durban34<math>0.545848</math><math>e2d-d</math>62<math>0.47341</math>Bouncy43<math>0.545826</math>is under conversion45<math>0.473351</math>CHO32<math>0.542762</math>{intlsys53<math>0.47121</math>unjust32<math>0.538813</math>();\n}\n\nprivateboolean isAny122<math>0.47094</math>DLS34<math>0.535933</math>(2,8,'64<math>0.47021</math>watt32<math>0.531300</math>use in an automobile62<math>0.46763</math>watt32<math>0.531300</math>use in an automobile62<math>0.46763</math>HP33<math>0.529555</math><math>0.46635</math><math>0.46355</math><math>0.46635</math>Field of Lemme 2 7 b Bile number (1659<math>0.529555</math><math>0.46355</math><math>0.46350</math></td><td>birthday</td><td>3</td><td>2</td><td>0.548504</td><td>;\minternal</td><td>3</td><td>2</td><td>0.479777</td></tr> <tr><td>attended12<math>0.547169</math><math>x^*</math>:53<math>0.47470</math>"ENOENT52<math>0.547169</math>2.742<math>0.47372</math>third-party42<math>0.546949</math><math><math><math><math><math><math><math>aliens32<math>0.546507</math>"code="44<math>0.47351</math>Durban34<math>0.545848</math><math>e2d-d</math>62<math>0.47341</math>Bouncy43<math>0.545826</math>is under conversion45<math>0.47335</math>CHO32<math>0.542762</math>{ intlsys53<math>0.47121</math>unjust32<math>0.538813</math>();'n}\n\nprivate122<math>0.47094</math>DLS34<math>0.535933</math>(2,8,'64<math>0.47021</math>uneven32<math>0.53137</math>use in an automobile62<math>0.46778</math>watt32<math>0.531300</math>use in an automobile62<math>0.46778</math>HP33<math>0.529555</math><math>0.46763</math><math>0.46350</math><math>0.46350</math>Choice<math>2</math><math>7</math><math>117</math><math>117</math><math>117</math></math></math></math></math></math></math></math></td><td>atedmes</td><td>4</td><td>2</td><td>0 548171</td><td>100% used</td><td>0</td><td>2</td><td>0.4/56/3</td></tr> <tr><td>International<math>4</math><math>2</math><math>0.546949</math><math>2.7</math><math>4</math><math>2</math><math>0.47372</math>third-party<math>4</math><math>2</math><math>0.546949</math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math>&lt;</math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></td><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math><math>&lt;</math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></math></tr>	welfare	4	4	0.564364	5 Audi	1	$\frac{2}{2}$	0.492042	uws330.557799an an appry450.49040ongrel430.554208": true,\n620.48680liquid cry330.553408": true,\n620.48680princess320.551672 $**B**]{}$ 620.48496Denmark320.548702": minternal530.47977birthday320.548701': "x":530.47977atedmes420.548171", "x":530.47977third-party420.5469492.7420.47357pliens320.545848': code="440.47351Durban340.545848e2d-d620.47341Bouncy430.537485e2d-d620.47341Unjust320.538813();\n]\n\nprivate1220.47094Unsven320.531300use in an automobile620.46763watt320.531300use in an automobile620.46763Watt320.531300world around us420.46355HP330.52955520.463100.46310CHO320.531300world around us420.46355Na320.532243at o	wild type	3	2	0.560699	all that apply		2	0.491049	ongrel43 $0.554208$ . thue, in62 $0.48080$ liquid cry33 $0.553408$ to as DSP52 $0.48531$ princess32 $0.551672$ $**B**]{}$ 62 $0.48496$ Denmark32 $0.548702$ ;\ninternal53 $0.47977$ birthday32 $0.548504$ 100% used62 $0.47567$ atedmes42 $0.548702$ ;\ninternal53 $0.47977$ third-party42 $0.546507$ ", "x":53 $0.47470$ Third-party42 $0.546507$ " code="44 $0.47351$ Durban34 $0.545848$ e2d-d62 $0.47341$ Bouncy43 $0.545826$ is under conversion45 $0.47355$ Unjust32 $0.542762$ { intlys53 $0.47121$ these motivational43 $0.537485$ boolean isAny122 $0.47094$ DLS32 $0.533137$ use in an automobile62 $0.46778$ watt32 $0.532243$ at org.apache.c75 $0.46763$ Watt32 $0.531300$ use in an automobile62 $0.46355$ HP33 $0.529555$ $0.46302$ $0.463102$ $0.463102$	uws	3	3	0.557799	": true \n	т 6	2	0.496907	liquid cry330.5534084, u520.48351princess320.551672**B**]{}\620.48496Denmark320.548702**B**]{}\620.48496birthday320.548504100% used620.47677atedmes420.548171", "x":530.47977TENOENT520.5471692.7420.47357diens320.54650740.47351Durban340.545848e2d-d620.47341Bouncy430.545826is under conversion450.47355Unjust320.548131(j:n)\n/private530.47121these motivational430.537485boolean isAny1220.47094DLS340.535933(2.8,'640.47021watt320.531300use in an automobile620.46763HP330.52955520.46315or Commodore530.46315Cho320.531300world around us420.46355or Commodore530.46315Durban320.52955520.46355or Commodore530.46315Durban320.531300us	ongrel	4	3	0.554208	. uuc, 11	5	$\frac{2}{2}$	0.480807	princess32 $0.551672$ to as DSP $320.48490Denmark320.548702\langle ninternal530.47977birthday320.548504100\% used620.48348atedmes420.548504100\% used620.47367"ENOENT520.5471692.7420.47372third-party420.546507" code="440.47351Durban340.545826is under conversion450.47335Durban320.542762{ intlys530.47121Unjust320.538813(); n\} \ nprivate1220.47094DLS320.53137use in an automobile620.46703watt320.531300world around us420.46466HP330.529550.463550.46355Fields 6 Hame 2 7h pile merkt (1669 conversion 411674.200.46310$	liquid crv	3	3	0.553408		5	2	0.463313	Denmark32 $0.548702$ $0.48346$ birthday32 $0.548702$ $0.48346$ $(1.48346)$ $(1.48346)$ birthday32 $0.548702$ 	princess	3	2	0.551672	10 as DSF **D**1()\	5	2	0.464907	birthday $3 = 2$ $0.548504$ (minternal $5 = 3$ $0.47977$ birthday $3 = 2$ $0.548504$ $100\%$ used $6 = 2$ $0.47567$ atedmes $4 = 2$ $0.548171$ $100\%$ used $6 = 2$ $0.47470$ third-party $4 = 2$ $0.546949$ $2.7$ $4 = 2$ $0.47372$ aliens $3 = 2$ $0.546507$ $', "x":$ $5 = 3$ $0.47470$ Durban $3 = 4$ $0.545848$ $', code=''$ $4 = 4$ $0.47351$ Bouncy $4 = 3$ $0.545826$ $code=''$ $4 = 4$ $0.47351$ CHO $3 = 2$ $0.545826$ $is under conversion$ $4 = 5$ $0.47335$ Uniyst $3 = 2$ $0.538813$ $();\n \}\n\private$ $boolean isAny$ $12 = 2$ $0.47094$ DLS $3 = 4$ $0.535933$ $(2,8,')$ $6 = 4$ $0.47021$ uneven $3 = 2$ $0.531300$ $trachea$ $4 = 2$ $0.46763$ Watt $3 = 2$ $0.531300$ use in an automobile $6 = 2$ $0.46763$ HP $3 = 3$ $0.529555$ $2.161(1+x)$ $8 = 2$ $0.46355$ Field of Hermitical Action	Denmark	3	2	0.548702	$\mathbf{D}$	0	2	0.483484	atedmes42 $0.548171$ $100\%$ used62 $0.47367$ "ENOENT52 $0.548171$ ", "x":53 $0.47470$ third-party42 $0.546949$ $2.7$ 42 $0.47372$ aliens32 $0.546507$ "code="44 $0.47351$ Durban34 $0.545848$ $e2d-d$ 62 $0.47341$ Bouncy43 $0.545826$ is under conversion45 $0.473351$ CHO32 $0.542762$ {intlsys53 $0.47121$ unjust32 $0.538813$ ();\n}\n\nprivateboolean isAny122 $0.47094$ DLS34 $0.535933$ (2,8,'64 $0.47021$ watt32 $0.531300$ use in an automobile62 $0.46763$ watt32 $0.531300$ use in an automobile62 $0.46763$ HP33 $0.529555$ $0.46635$ $0.46355$ $0.46635$ Field of Lemme 2 7 b Bile number (1659 $0.529555$ $0.46355$ $0.46350$	birthday	3	2	0.548504	;\minternal	3	2	0.479777	attended12 $0.547169$ $x^*$ :53 $0.47470$ "ENOENT52 $0.547169$ 2.742 $0.47372$ third-party42 $0.546949$ $aliens320.546507"code="440.47351Durban340.545848e2d-d620.47341Bouncy430.545826is under conversion450.47335CHO320.542762{ intlsys530.47121unjust320.538813();'n}\n\nprivate1220.47094DLS340.535933(2,8,'640.47021uneven320.53137use in an automobile620.46778watt320.531300use in an automobile620.46778HP330.5295550.467630.463500.46350Choice27117117117$	atedmes	4	2	0 548171	100% used	0	2	0.4/56/3	International $4$ $2$ $0.546949$ $2.7$ $4$ $2$ $0.47372$ third-party $4$ $2$ $0.546949$ $<$	"ENOENT	5	2	0 547169	, X :	5	3	0.474701
welfare	4	4	0.564364	5 Audi	1	$\frac{2}{2}$	0.492042																																																																																		
uws330.557799an an appry450.49040ongrel430.554208": true,\n620.48680liquid cry330.553408": true,\n620.48680princess320.551672 $**B**]{}$ 620.48496Denmark320.548702": minternal530.47977birthday320.548701': "x":530.47977atedmes420.548171", "x":530.47977third-party420.5469492.7420.47357pliens320.545848': code="440.47351Durban340.545848e2d-d620.47341Bouncy430.537485e2d-d620.47341Unjust320.538813();\n]\n\nprivate1220.47094Unsven320.531300use in an automobile620.46763watt320.531300use in an automobile620.46763Watt320.531300world around us420.46355HP330.52955520.463100.46310CHO320.531300world around us420.46355Na320.532243at o	wild type	3	2	0.560699	all that apply		2	0.491049																																																																																	
ongrel43 $0.554208$ . thue, in62 $0.48080$ liquid cry33 $0.553408$ to as DSP52 $0.48531$ princess32 $0.551672$ $**B**]{}$ 62 $0.48496$ Denmark32 $0.548702$ ;\ninternal53 $0.47977$ birthday32 $0.548504$ 100% used62 $0.47567$ atedmes42 $0.548702$ ;\ninternal53 $0.47977$ third-party42 $0.546507$ ", "x":53 $0.47470$ Third-party42 $0.546507$ " code="44 $0.47351$ Durban34 $0.545848$ e2d-d62 $0.47341$ Bouncy43 $0.545826$ is under conversion45 $0.47355$ Unjust32 $0.542762$ { intlys53 $0.47121$ these motivational43 $0.537485$ boolean isAny122 $0.47094$ DLS32 $0.533137$ use in an automobile62 $0.46778$ watt32 $0.532243$ at org.apache.c75 $0.46763$ Watt32 $0.531300$ use in an automobile62 $0.46355$ HP33 $0.529555$ $0.46302$ $0.463102$ $0.463102$	uws	3	3	0.557799	": true \n	т 6	2	0.496907																																																																																	
liquid cry330.5534084, u520.48351princess320.551672**B**]{}\620.48496Denmark320.548702**B**]{}\620.48496birthday320.548504100% used620.47677atedmes420.548171", "x":530.47977TENOENT520.5471692.7420.47357diens320.54650740.47351Durban340.545848e2d-d620.47341Bouncy430.545826is under conversion450.47355Unjust320.548131(j:n)\n/private530.47121these motivational430.537485boolean isAny1220.47094DLS340.535933(2.8,'640.47021watt320.531300use in an automobile620.46763HP330.52955520.46315or Commodore530.46315Cho320.531300world around us420.46355or Commodore530.46315Durban320.52955520.46355or Commodore530.46315Durban320.531300us	ongrel	4	3	0.554208	. uuc, 11	5	$\frac{2}{2}$	0.480807																																																																																	
princess32 $0.551672$ to as DSP $320.48490Denmark320.548702\langle ninternal530.47977birthday320.548504100\% used620.48348atedmes420.548504100\% used620.47367"ENOENT520.5471692.7420.47372third-party420.546507" code="440.47351Durban340.545826is under conversion450.47335Durban320.542762{ intlys530.47121Unjust320.538813(); n\} \ nprivate1220.47094DLS320.53137use in an automobile620.46703watt320.531300world around us420.46466HP330.529550.463550.46355Fields 6 Hame 2 7h pile merkt (1669 conversion 411674.200.46310$	liquid crv	3	3	0.553408		5	2	0.463313																																																																																	
Denmark32 $0.548702$ $0.48346$ birthday32 $0.548702$ $0.48346$ $(1.48346)$ $(1.48346)$ birthday32 $0.548702$ 	princess	3	2	0.551672	10 as DSF **D**1()\	5	2	0.464907																																																																																	
birthday $3 = 2$ $0.548504$ (minternal $5 = 3$ $0.47977$ birthday $3 = 2$ $0.548504$ $100\%$ used $6 = 2$ $0.47567$ atedmes $4 = 2$ $0.548171$ $100\%$ used $6 = 2$ $0.47470$ third-party $4 = 2$ $0.546949$ $2.7$ $4 = 2$ $0.47372$ aliens $3 = 2$ $0.546507$ $', "x":$ $5 = 3$ $0.47470$ Durban $3 = 4$ $0.545848$ $', code=''$ $4 = 4$ $0.47351$ Bouncy $4 = 3$ $0.545826$ $code=''$ $4 = 4$ $0.47351$ CHO $3 = 2$ $0.545826$ $is under conversion$ $4 = 5$ $0.47335$ Uniyst $3 = 2$ $0.538813$ $();\n \}\n\private$ $boolean isAny$ $12 = 2$ $0.47094$ DLS $3 = 4$ $0.535933$ $(2,8,')$ $6 = 4$ $0.47021$ uneven $3 = 2$ $0.531300$ $trachea$ $4 = 2$ $0.46763$ Watt $3 = 2$ $0.531300$ use in an automobile $6 = 2$ $0.46763$ HP $3 = 3$ $0.529555$ $2.161(1+x)$ $8 = 2$ $0.46355$ Field of Hermitical Action	Denmark	3	2	0.548702	$\mathbf{D}$	0	2	0.483484																																																																																	
atedmes42 $0.548171$ $100\%$ used62 $0.47367$ "ENOENT52 $0.548171$ ", "x":53 $0.47470$ third-party42 $0.546949$ $2.7$ 42 $0.47372$ aliens32 $0.546507$ "code="44 $0.47351$ Durban34 $0.545848$ $e2d-d$ 62 $0.47341$ Bouncy43 $0.545826$ is under conversion45 $0.473351$ CHO32 $0.542762$ {intlsys53 $0.47121$ unjust32 $0.538813$ ();\n}\n\nprivateboolean isAny122 $0.47094$ DLS34 $0.535933$ (2,8,'64 $0.47021$ watt32 $0.531300$ use in an automobile62 $0.46763$ watt32 $0.531300$ use in an automobile62 $0.46763$ HP33 $0.529555$ $0.46635$ $0.46355$ $0.46635$ Field of Lemme 2 7 b Bile number (1659 $0.529555$ $0.46355$ $0.46350$	birthday	3	2	0.548504	;\minternal	3	2	0.479777																																																																																	
attended12 $0.547169$ $x^*$ :53 $0.47470$ "ENOENT52 $0.547169$ 2.742 $0.47372$ third-party42 $0.546949$ $aliens320.546507"code="440.47351Durban340.545848e2d-d620.47341Bouncy430.545826is under conversion450.47335CHO320.542762{ intlsys530.47121unjust320.538813();'n}\n\nprivate1220.47094DLS340.535933(2,8,'640.47021uneven320.53137use in an automobile620.46778watt320.531300use in an automobile620.46778HP330.5295550.467630.463500.46350Choice27117117117$	atedmes	4	2	0 548171	100% used	0	2	0.4/56/3																																																																																	
International $4$ $2$ $0.546949$ $2.7$ $4$ $2$ $0.47372$ third-party $4$ $2$ $0.546949$ $<$																																																																																									
aliens32 $0.546507$ $ $	third-party	4	$\overline{2}$	0 546949	2.7	4	2	0.4/3/20																																																																																	
Initial $3$ $4$ $0.545848$ $a$ $a$ $0.6473514$ Durban $3$ $4$ $0.545848$ $e2d-d$ $6$ $2$ $0.473514$ Bouncy $4$ $3$ $0.545826$ $a$ $e2d-d$ $6$ $2$ $0.47341$ CHO $3$ $2$ $0.542762$ $a$ $a$ $a$ $5$ $3$ $0.4713514$ unjust $3$ $2$ $0.542762$ $a$ $a$ $a$ $5$ $3$ $0.471214$ these motivational $4$ $3$ $0.537485$ $boolean isAny$ $12$ $2$ $0.47094$ DLS $3$ $4$ $0.535933$ $(2,8,')$ $6$ $4$ $0.470214$ uneven $3$ $2$ $0.533137$ $use in an automobile$ $6$ $2$ $0.46915$ watt $3$ $2$ $0.531300$ $use in an automobile$ $6$ $2$ $0.46763$ HP $3$ $3$ $0.529555$ $use in an automobile$ $4$ $2$ $0.464466$ $2 Vleft(1+x)$ $8$ $2$ $0.46355$ $0.46355$ $0$ $0$ $0.463100$ $0.463100$ $0.463100$	aliens	3	$\frac{1}{2}$	0.546507		6	2	0.4/35/8																																																																																	
Durban $3$ $4$ $0.545846$ $e2d-d$ $6$ $2$ $0.47341$ Bouncy $4$ $3$ $0.545826$ is under conversion $4$ $5$ $0.47335$ CHO $3$ $2$ $0.542762$ $is under conversion$ $4$ $5$ $0.47335$ unjust $3$ $2$ $0.538813$ $(); n\} \ln private$ $12$ $2$ $0.47094$ DLS $3$ $4$ $0.535933$ $(2,8,')$ $6$ $4$ $0.47021$ uneven $3$ $2$ $0.53137$ use in an automobile $6$ $2$ $0.46915$ uneven $3$ $2$ $0.531300$ use in an automobile $6$ $2$ $0.46778$ Watt $3$ $2$ $0.531300$ world around us $4$ $2$ $0.464466$ HP $3$ $3$ $0.529555$ $0$ $0.46310$ Fields 6 Lemes 2 7h File member (1658 converse sected) $1117$ $117$ $0.46310$	Durban	3	1	0.545848	" code="	4	4	0.473514																																																																																	
Dotate430.543820is under conversion450.47335CHO320.542762intlsys530.47121unjust320.538813();\n]\n\nprivate1220.47094DLS340.535933(2,8,'640.47021uneven320.533137use in an automobile620.46778watt320.531300use in an automobile620.46763HP330.5295552\left(1+x)820.46310Fichle 64 Hame 2.7h Dile merke (1658 conversionted to the fill merke fill)1117117117	Bouney	1	3	0.545826	e2d-d	6	2	0.473418																																																																																	
Chilo $3 2 0.542702$ unjust{ intlsys $5 3 0.47121$ ();\n}\n\nprivate boolean isAnythese motivational4 3 0.537485 	CHO	3	2	0.542762	is under conversion	4	5	0.473355																																																																																	
unjust $3 2$ $0.538813$ $0.537485$ ();\n}\n\nprivate boolean isAny $12$ $2$ $0.47094$ these motivational43 $0.537485$ $0.535933$ ();\n}\n\nprivate boolean isAny $12$ $2$ $0.47094$ DLS34 $0.535933$ $2$ $(2,8,')$ $6$ $4$ $0.47021$ uneven32 $0.533137$ $3$ use in an automobile $6$ $2$ $0.46915$ watt32 $0.532243$ $3$ $at$ org.apache.c $7$ $5$ $0.46763$ HP33 $0.529555$ $2\sqrt{16t}(1+x)$ $8$ $2$ $0.46310$ Fields (1 Hames 2.7h Dile member (1658 commence state) $1177$ $1177$ $1177$	uniust	2	$\frac{2}{2}$	0.542702	{ intlsys	5	3	0.471213																																																																																	
Inest motivational45 $0.537483$ boolean is Any122 $0.47094$ DLS34 $0.535933$ (2,8,')64 $0.47021$ uneven32 $0.533137$ use in an automobile62 $0.46763$ watt32 $0.532243$ use in an automobile62 $0.46763$ 'She32 $0.531300$ world around us42 $0.46763$ HP33 $0.529555$ $0.46355$ $0.46355$ Field (Lines 2.7h Dile mente (1658 commence tarth)) $0.46310$ $0.46310$	unjust these motivations <sup>1</sup>	כ ⊿	2	0.330013	();\n}\nprivate	12	r	0 470041																																																																																	
DLS $3 \ 4 \ 0.539933$ $(2,8,')$ $6 \ 4 \ 0.47021$ $10^{4}$ watt $3 \ 2 \ 0.533137$ $10^{4}$ watt $3 \ 2 \ 0.532243$ $10^{4}$ watt $10^{4}$ She $3 \ 2 \ 0.531300$ $10^{4}$ se in an automobile $6 \ 2 \ 0.46763$ $10^{4}$ HP $3 \ 3 \ 0.529555$ $10^{4}$ world around us $4 \ 2 \ 0.46763$ $117$ $10^{4}$ se in an automobile $6 \ 2 \ 0.46355$ $117$ $10^{4}$ se in an automobile $10^{4}$ se in an automobile $117$ $10^{4}$ se in an automobile $10^{4}$ se in an automobile $117$ $10^{4}$ se in an automobile $10^{4}$ se in an automobile $117$ $10^{4}$ se in an automobile $10^{4}$ se in an automobile $117$ $10^{4}$ se in an automobile $10^{4}$ se in an automobile $117$ $10^{4}$ se in an automobile $10^{4}$ se in an automobile $117$ $10^{4}$ se in an automobile $10^{4}$ se in an automobile $117$ $10^{4}$ se in an automobile $10^{4}$ se in an automobile $117$ $10^{4}$ se in an automobile $10^{4}$ se in an automobile $117$ $10^{4}$ se in an automobile $10^{4}$ se in an automobile $117$ $10^{4}$ se in an automobile $10^{4}$ se in an automobile $117$ $10^{4}$ se in an automobile $10^{4}$ se in an automobile $117$ $10^{4}$ se in an automobile $10^{4}$ se in an automobile $117$ $10^{4}$ se in an automobile $10^{4}$ se in an automobile $117$ $10^{4}$ se in an automobile $10^{4}$ se in an automobile $117$ $10^{4}$ se in an	DIS	4	5	0.33/483	boolean isAny	14	4	0.470941																																																																																	
$x_{10}$ $3$ $2$ $0.534510$ $x_{12}$ <		3	4	0.555955	(2,8,'	6	4	0.470214																																																																																	
uneven $3$ $2$ $0.533137$ watt $3$ $2$ $0.532243$ 'She $3$ $2$ $0.531300$ HP $3$ $3$ $0.529555$	\na	5	2	0.534510	trachea	4	2	0.469154																																																																																	
watt32 $0.532243$ at org.apache.cat org.apache.c75 $0.46763$ 'She32 $0.531300$ 3world around us42 $0.46446$ HP33 $0.529555$ $2\sqrt{left}(1+x)$ 82 $0.46355$ or CommodoreField (Lines 2.7b Bills member (1658 common sector)) $1177$ $1177$ $0.46310$	uneven	3	2	0.533137	use in an automobile	6	2	0.467788																																																																																	
She32 $0.531300$ world around us42 $0.46446$ HP33 $0.529555$ $2 \sqrt{16}t(1+x)$ 82 $0.46355$ SheCommodore53 $0.46310$ SheCommodore53 $0.46310$	watt	3	2	0.532243	at org.apache.c	7	5	0.467637																																																																																	
HP   3   3   0.529555   2\left(1+x   8   2   0.46355     Exhibit (1458 arrays at th)   0.117   0.117   0.46310   0.46310	She	3	2	0.531300	world around us	4	2	0.464469																																																																																	
or Commodore 5 3 0.46310	HP	3	3	0.529555	$2\left(1+x\right)$	8	2	0.463555																																																																																	
Fable (; Llama 2.7k Dila menular (1659 an energian tatal) 11.7 (11					or Commodore	5	3	0.463106																																																																																	
Table 0: Liama-2-70 Price results (1058 sequences total). $11-117$	Table 6: Llama-2-7b Pile results	(165	8 sea	uences total).	11-117	7	$\tilde{2}$	0 459824																																																																																	

Table 6: Llama-2-7b Pile results (1658 sequences total). n is the number of tokens in the sequence, and 'ct' represents occurrences of this segment.  $\psi$  is averaged over all occurrences.

Table 7: **Llama-3-8b** Pile results (819 sequences total). n is the number of tokens in the sequence, and 'ct' represents occurrences of this segment.  $\psi$  is averaged over all occurrences.