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 fine-tuned 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 token-level information throughout layers.

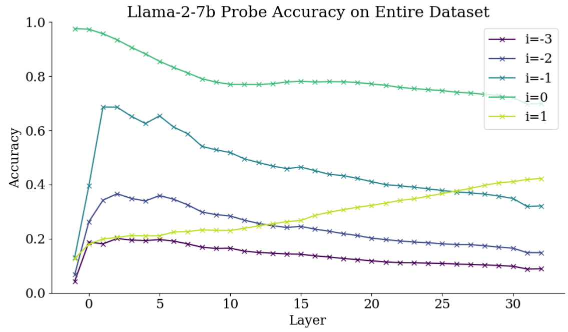

Figure 4: Overall test accuracy on unseen Pile tokens ($n = 273$k) for probes trained on Llama-2-7b hidden states. Next token prediction becomes more accurate throughout model layers as current and previous token accuracy decreases.

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 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 $L \in \{5, 9, 13, 17, 21\}$.

	MTW		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.

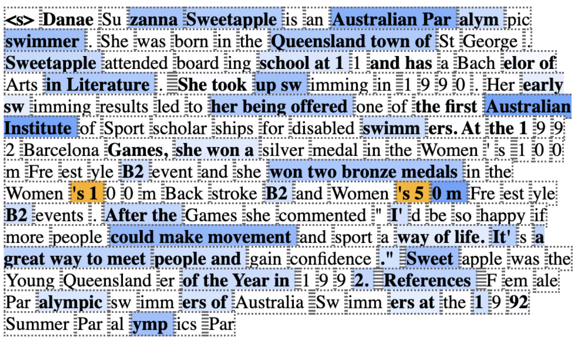

Figure 5: Full segmentation of a document from Wikipedia via Algorithm 1 on Llama-2-7b. Borders indicate segmentation, with bolded letters indicating multi-token segments. Darker blue cells have higher scores, yellow cells have negative scores. The highest-scoring sequence in this document is “Australian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

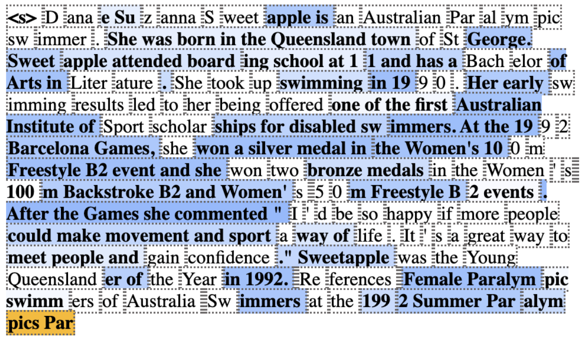

Figure 6: Full segmentation of a document from Wikipedia via Algorithm 1 on Llama-3-8b. Borders indicate segmentation, with bolded letters indicating multi-token segments. Darker blue cells have higher scores, yellow cells have negative scores. The highest-scoring sequence in this document is “. After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

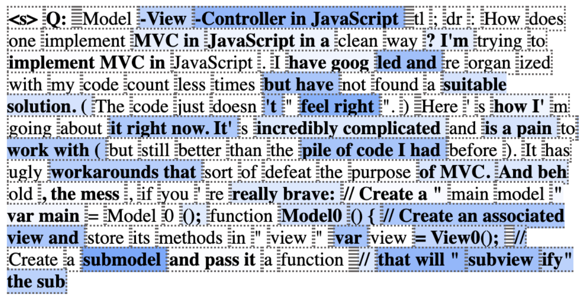

Figure 7: Full segmentation of a document from the Pile via Algorithm 1 on Llama-2-7b. Borders indicate segmentation, with bolded letters indicating multi-token segments. Darker blue cells have higher scores, yellow cells have negative scores. The highest-scoring sequence in this document is “submodel”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

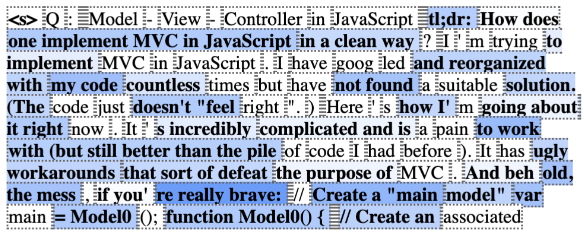

Figure 8: Full segmentation of a document from the Pile via Algorithm 1 on Llama-3-8b. Borders indicate segmentation, with bolded letters indicating multi-token segments. Darker blue cells have higher scores, yellow cells have negative scores. The highest-scoring sequence in this document is “re really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