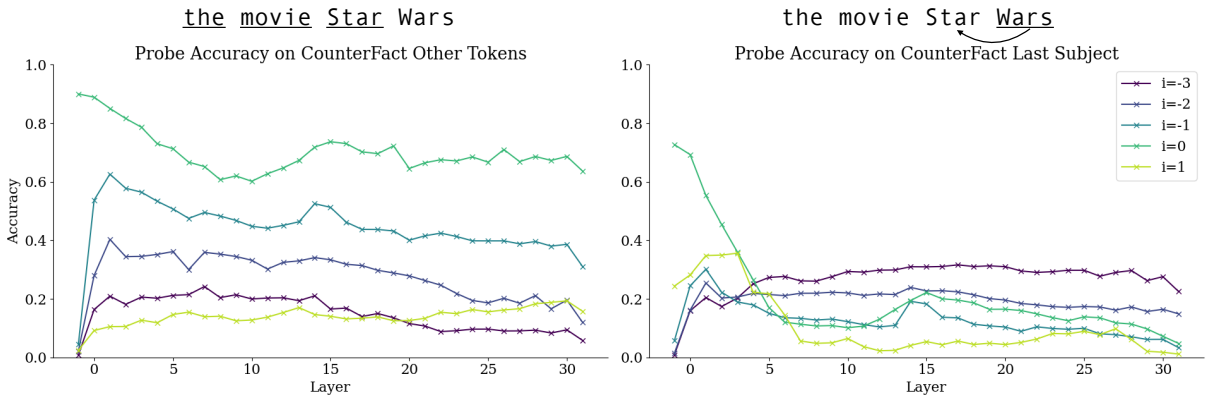

Figure 9: Test accuracy on COUNTERFACT subject last tokens versus other tokens in the dataset for probes trained on **Llama-3-8b** ($n = 5495$). i represents the position being predicted (e.g., $i = -1$ is previous token prediction; $i = 1$ is next-token prediction). We observe an "erasure" effect similar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

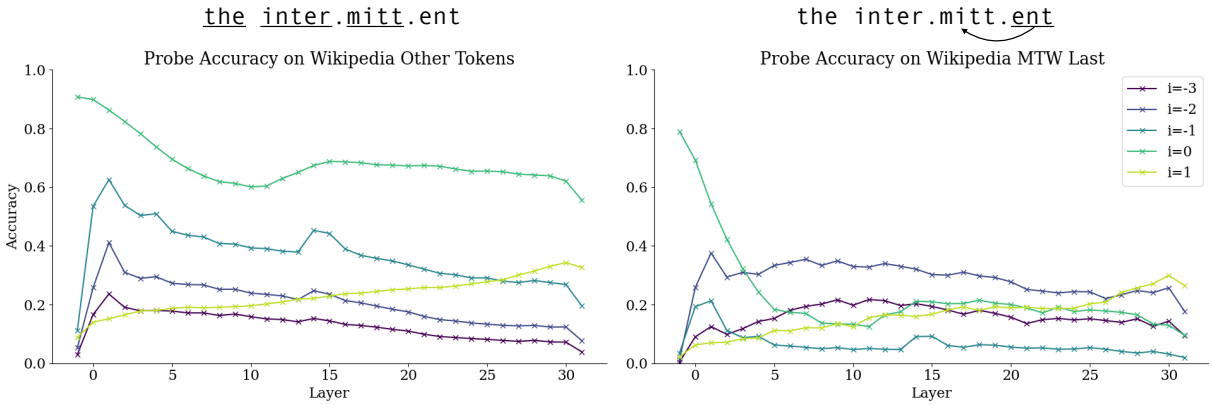

Figure 10: Test accuracy of probes on last tokens of Wikipedia **multi-token words** for probes trained on **Llama-3-8b** ($n = 91935$; right). Test accuracy on all other tokens shown on the left. Similarly to Figure 2, we see an erasing effect that is not present for other types of tokens.

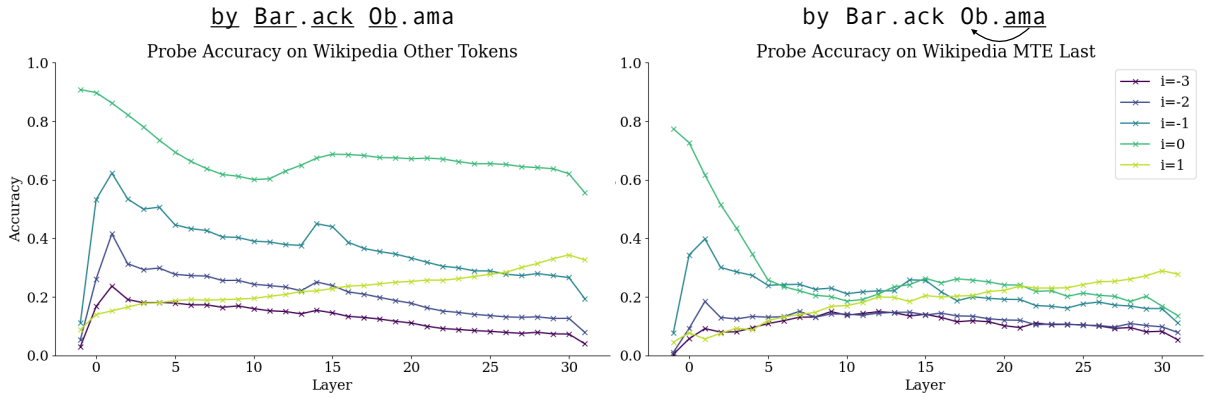

Figure 11: Test accuracy of probes on last tokens of Wikipedia **multi-token entities** for probes trained on **Llama-3-8b** ($n = 36723$; right). Test accuracy on all other tokens shown on the left. Entities are identified via `spaCy` named entity recognition, excluding entity types that include digits.

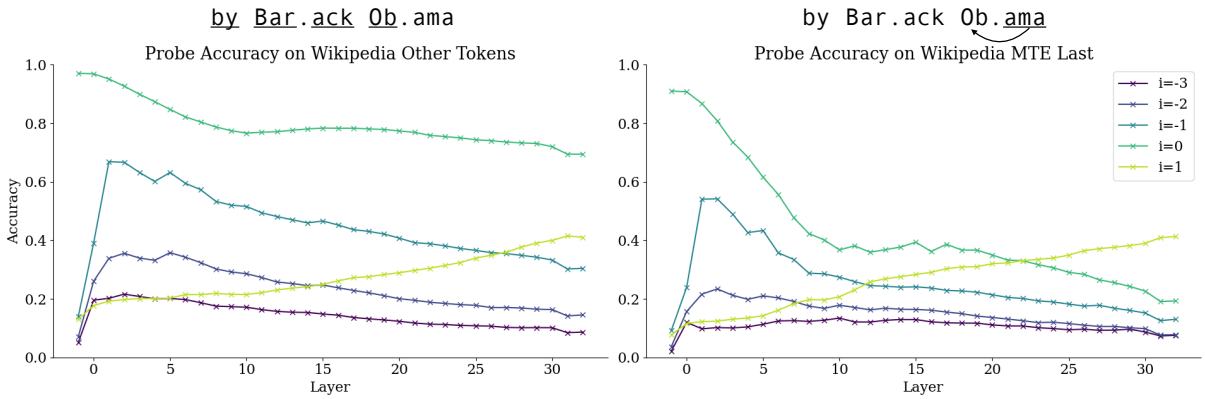

Figure 12: Test accuracy of probes on last tokens of Wikipedia **multi-token entities** for **Llama-2-7b** ($n = 36723$; right). Test accuracy on all other tokens shown on the left. Entities are identified via `spaCy` named entity recognition, excluding entity types that include digits.

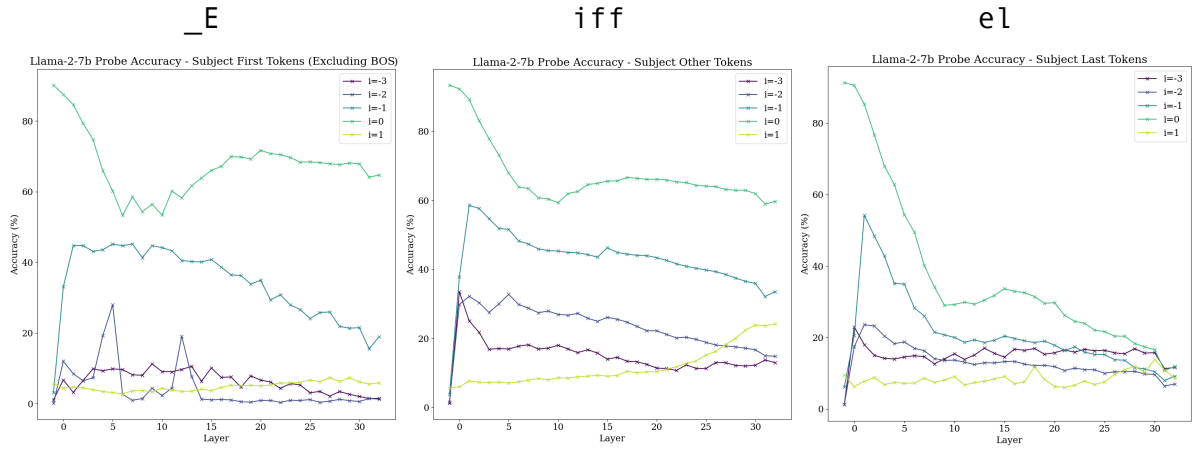

Figure 13: Breakdown for Section 3 probes tested on COUNTERFACT first subject tokens, middle subject tokens, and last subject tokens. We observe an "erasing" effect only for last subject tokens. Because BOS tokens are recoverable by $i = -1$ probes at high rates, and since 55% of prompts tested on had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

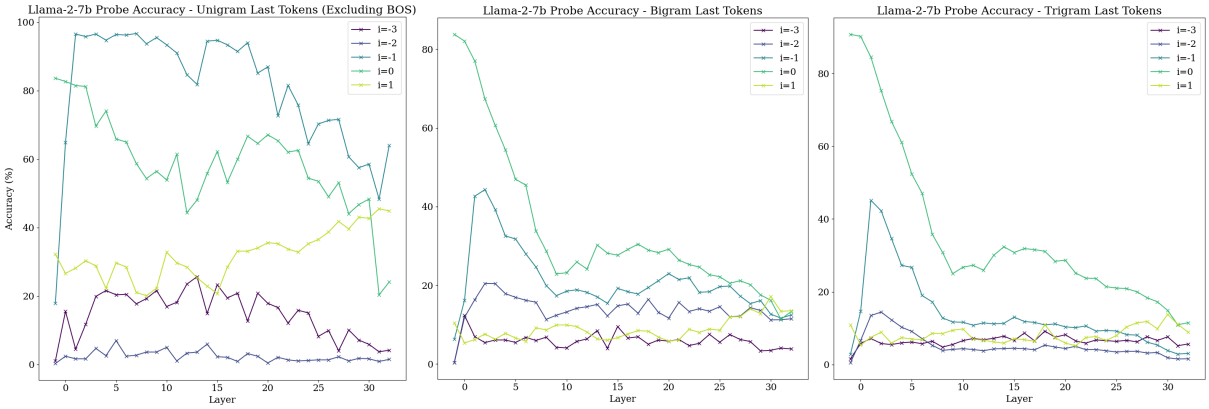

Figure 14: Probe test results for COUNTERFACT subject last tokens broken down for unigrams, bigrams, and trigrams. Unigram subjects store previous token information at rates near 100%, even excluding BOS tokens.

Token Sequence	n	ct	ψ
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	2	0.599339
Sox	3	3	0.599043
Bazz	3	2	0.598242
erect	3	2	0.597915
borough	3	3	0.596054
encompasses	5	2	0.592084
northernmost	3	2	0.591607
Madras	3	2	0.590394
hull	3	2	0.586968
iron	2	2	0.586959
Galaxy	3	2	0.585879
began operations	3	2	0.584680
Redding	3	2	0.584244
gloss	3	2	0.576740
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	2	0.547116
Czech Republic	3	2	0.546651
Survive	4	2	0.546255

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. ψ is averaged over all occurrences.

Token Sequence	n	ct	ψ
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	3	0.503120
2015 deaths	8	2	0.503074
Pittsburgh	3	3	0.503070
21st-century	7	4	0.499362
(NSW	4	9	0.497107
age of the United Kingdom	6	3	0.487303
Presidential	3	2	0.485317
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 time, if the scored	10	2	0.472323
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	2	0.450033
World Cup players	4	2	0.448979
main role	3	2	0.448569
Bos	4	2	0.448425
Asenath	4	2	0.448259
Royal Navy	3	3	0.445617
2. Bundesliga players	7	2	0.445210
External links	3	69	0.444921
an unincorpor	6	2	0.443527
Gast	2	4	0.437695
Pfor	3	2	0.432194
Elisio de Med	5	2	0.431518
" (2007) "Jad	12	2	0.429412
Elkh	3	2	0.428984
Früh	3	2	0.427781
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. ψ is averaged over all occurrences.