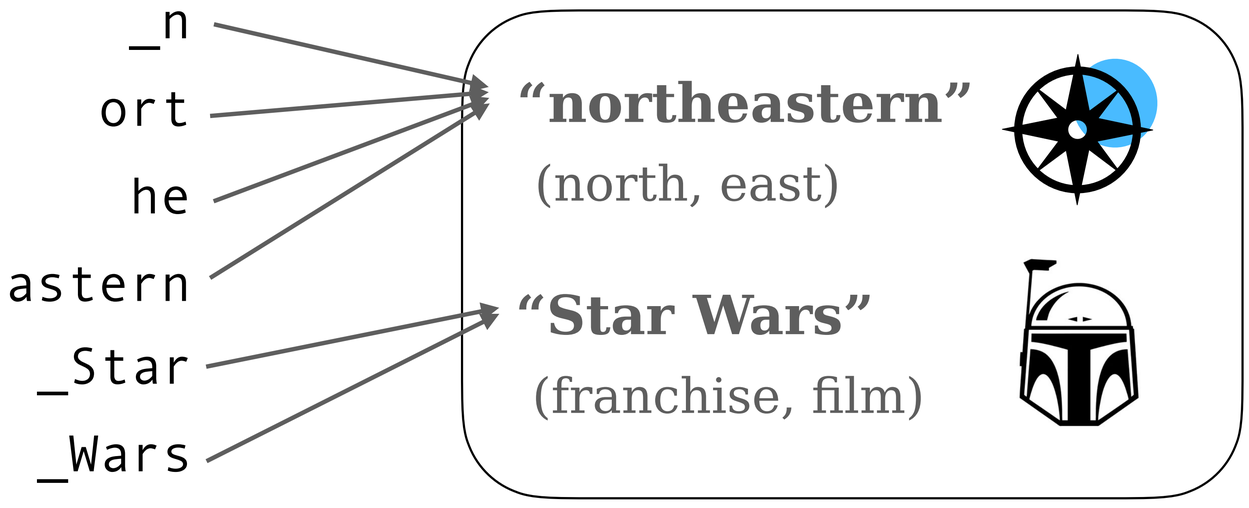
Token Erasure as a Footprint of Implicit Vocabulary Items in LLMs

Sheridan Feucht, David Atkinson, Byron Wallace, David Bau
Northeastern University

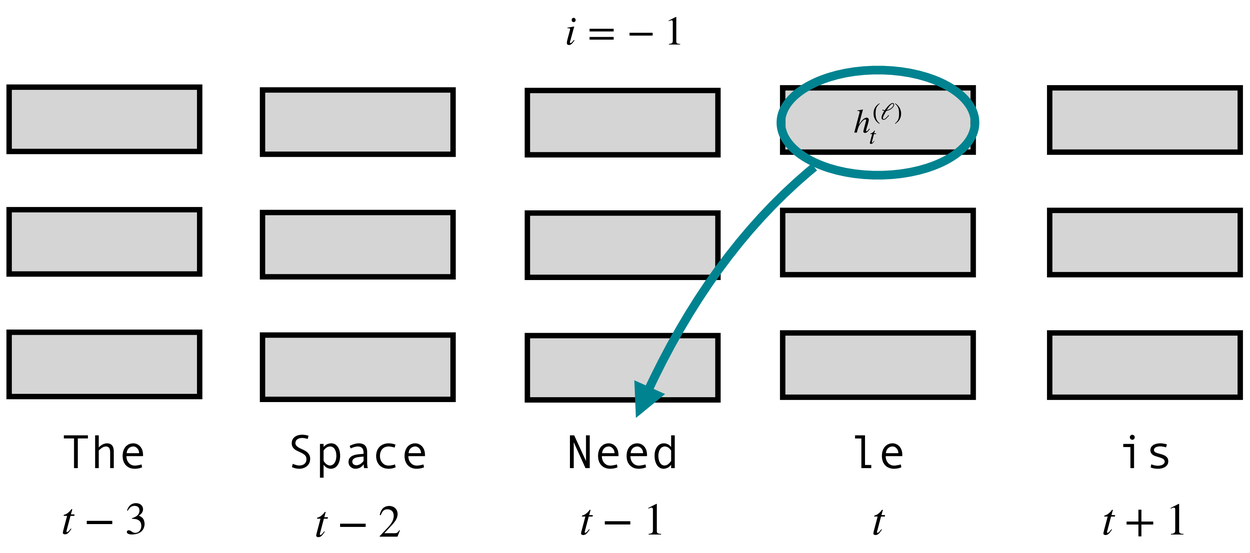
How do LLMs convert *tokens* into *concepts*?



Models must have some ***implicit vocabulary*** that maps from sequences of tokens to meaningful, word-like units.

We find a possible “footprint” of that process in Llama-2-7b and Llama-3-8b.

We train linear probes to recover neighboring token information from Llama hidden states.

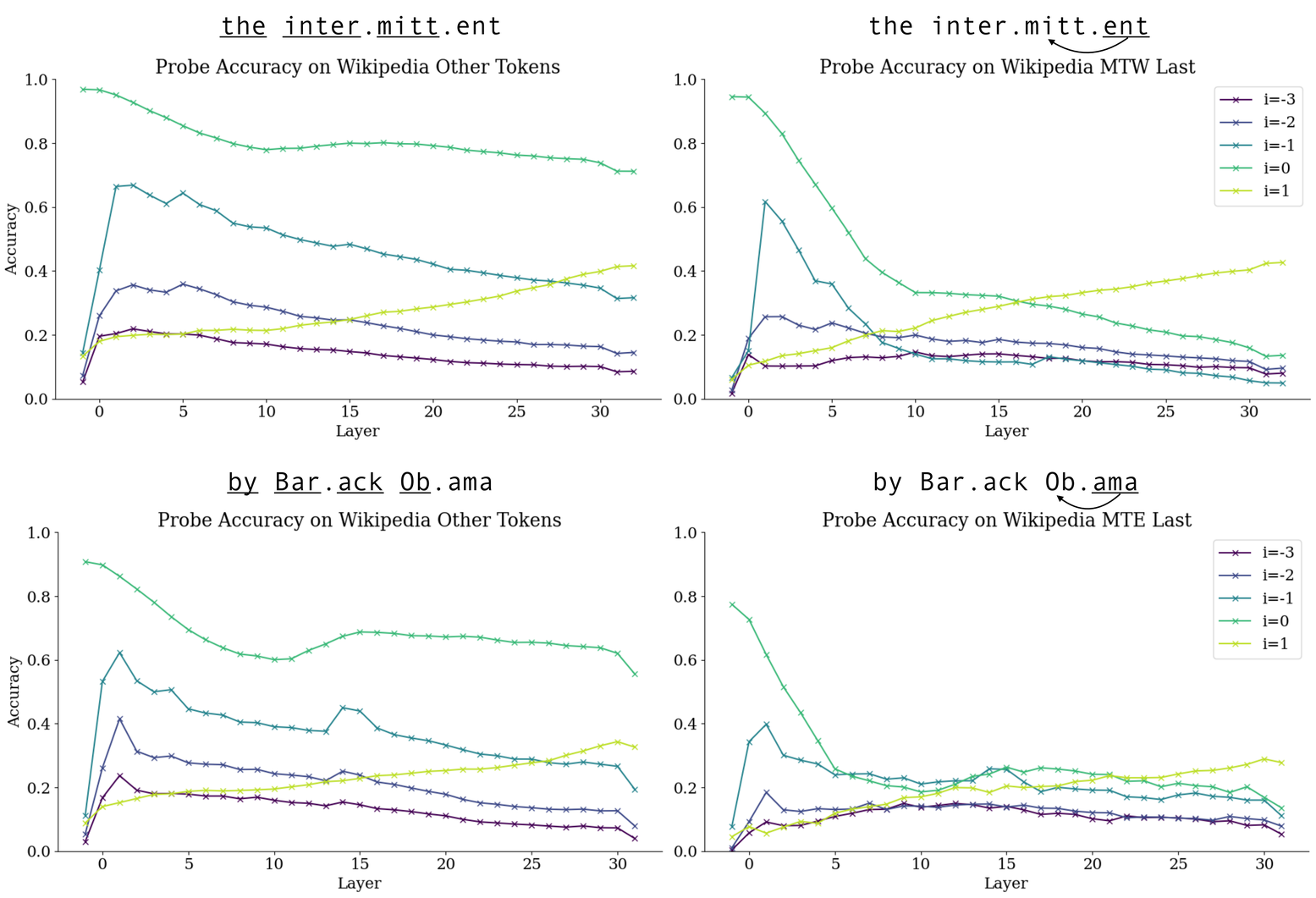


`Mon.k’s compos.itions and impro.vis.ations feature dis.son.ances and angular mel.od.ic tw.ists, often using flat nin.th.s, flat fifth.s, unexpected chrom.atic notes together, low bass notes and st.ride, and fast whole tone runs, combining a highly per.cuss.ive attack with ab.rupt, dram.atic use of switched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

score	tokens
0.582	dramatic
0.555	twists
0.415	low bass
0.339	flat ninths,
0.321	Monk’

0.315	stride
0.234	melodic
0.203	silences
0.183	s,
0.028	together,
0.016	, and fast whole

Token information **disappears** from certain hidden states!



This pattern seems to occur for **word-like sequences** of tokens. We can use it to “read out” token sequences that might constitute a model’s implicit vocabulary.

Token Sequence	n	ct	ψ
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

Llama-2-7b Pile Sequences

footprints.baulab.info