Crosscoding Through Time:
Tracking Emergence & Consolidation Of Linguistic Representations Across Checkpoints

Deniz Bayazit Aaron Mueller Antoine Bosselut

Loss curves and benchmarks show when LLMs improve behaviorally, but not how they learn internally!

- LLMs acquire much of their linguistic abilities during pretraining.
- We do not know which internal features emerge, persist, or disappear.
- **Goal: concept-level, human-interpretable description of feature changes.**

How can we trace the evolution of linguistic features through pretraining?

Case Study: Subject-Verb Agreement

The cat sleeping on the mats … → is ✓ / are ✗

detects plural inanimate objects

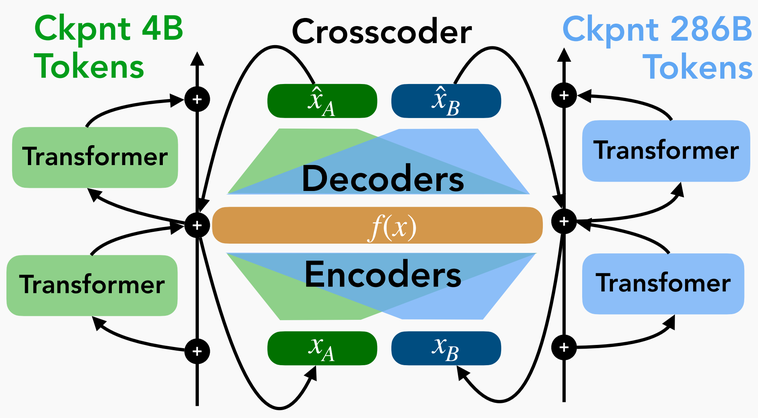
The cat sleeping on the mats … → are ✗

detects noun before prepositions

The cat sleeping on the mats … → is ✓

Find which features causally drive the model's behavior

Learn a joint feature space for critical checkpoints



Allows to directly compare each feature across checkpoints

Quantify causal importance of features over time

Importance Metric (IM) = LogProb of wrong token − LogProb of correct token

IE = IM when feat ablated − IM when not ablated

Per feat & ckpnt

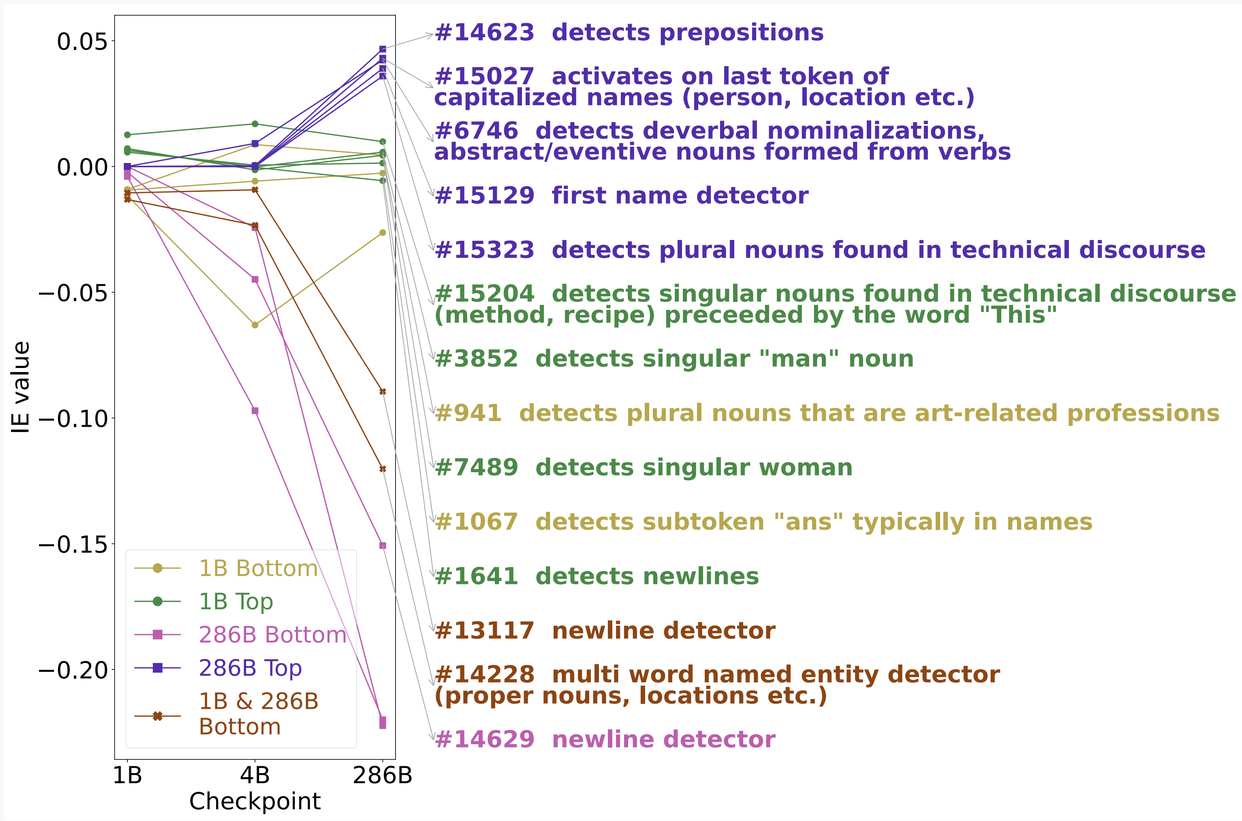
Relative IE = Share of total | IE | at each checkpoint

Per feature - relative to total IE across all selected checkpoints

How do LLM's internal features emerge, strengthen or fade as they train?

Which features matter early or gain importance later?

Top & Bottom Features' IE Over Time

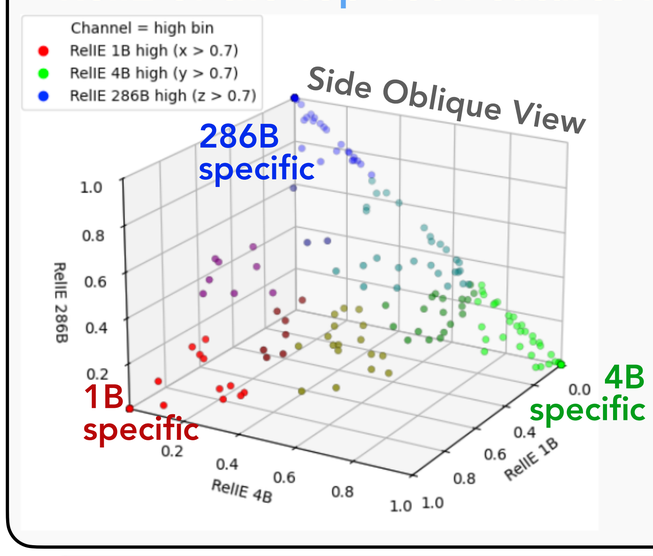


Many token-level features lose importance, while grammatical and high-level concept features become more influential.

Rather than keeping the same mechanism, the model seems to reorganize which features drive its behavior.

Mapping Feature Trajectories with RelIE

RelIE of the Top-100 Features for Three Checkpoints

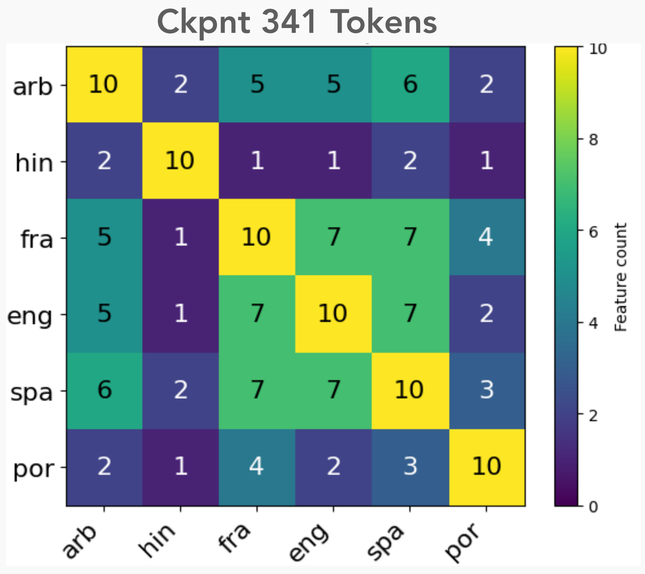


A small core persists across pretraining, but most rebuilt from 1B.

Accuracy plateaus do not imply representational plateaus: 4B and 286B share a core, but still retain distinct task-critical features.

Crosslingual Feature Alignment & its Limits

Feature Set Overlap for BLOOM-1B (SV-#)



Crosslingual feature sharing emerges late in BLOOM.

Overlap remains limited for richer agreement systems or scripts like Hindi.

Crosscoders and RelIE reveal that linguistic features can evolve unevenly.

Token-level detectors fade, while grammatical and crosslingual features consolidate later, though alignment remains limited for some languages.

This offers a new lens on how models learn, not just how well they perform.

Scan for Paper & Code