On Linear Representations and Pretraining Data Frequency in Language Models

BROWN W Ai2

Jack Merullo, Noah A. Smith, Sarah Wiegreffe, Yanai Elazar**

Linear structure of relations forms in LMs' hidden representations over the course of pretraining, according to relation frequency

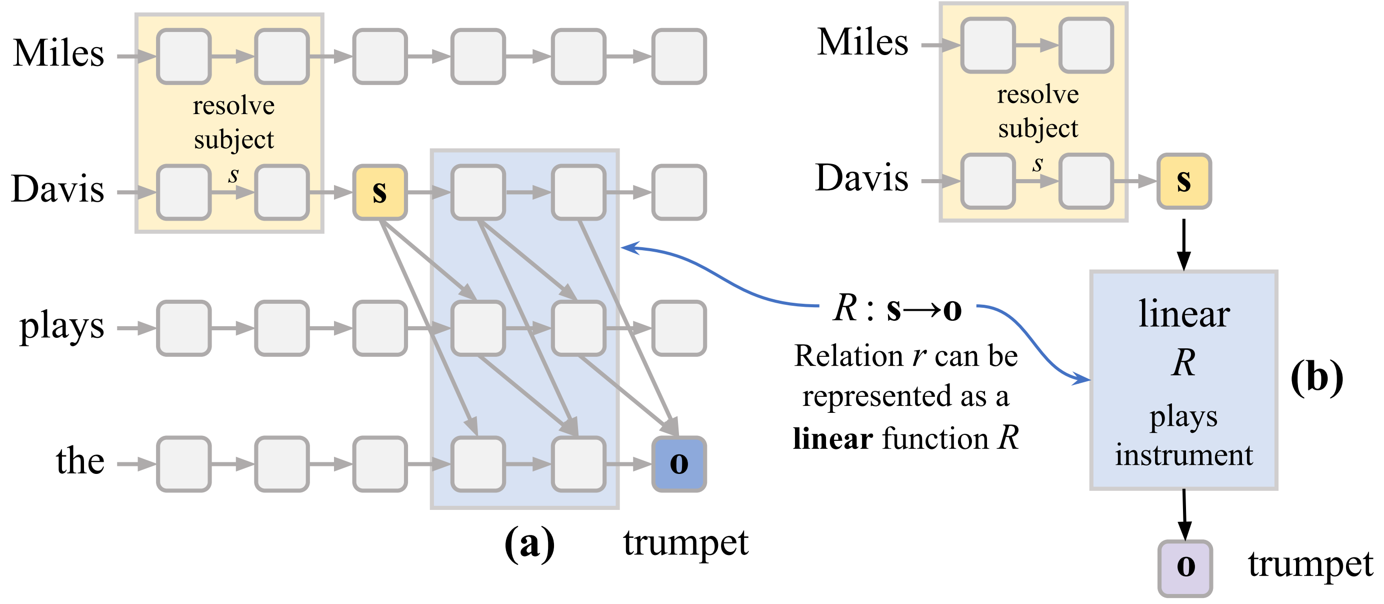
Methods

Example relation: plays instrument

"**Miles Davis[subj]** plays the **Trumpet[obj]**"

How do we measure linearity?

- A relation is '*linear*' when it is well approximated by an affine transformation in representation space
- We use *Linear Relational Embeddings (LREs)*, see Hernandez et al., 2024:

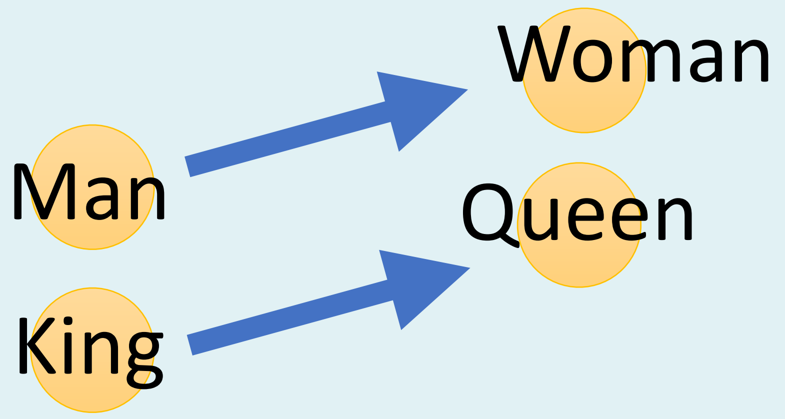


$$W = \mathbb{E}_{\mathbf{s}_i, c_i}\left[\frac{\partial F}{\partial \mathbf{s}}\bigg|_{(\mathbf{s}_i, c_i)} \right] \quad \text{and} \quad b = \mathbb{E}_{\mathbf{s}_i, c_i}\left[F(\mathbf{s}, c) - \frac{\partial F}{\partial \mathbf{s}}\, \mathbf{s}\bigg|_{(\mathbf{s}_i, c_i)} \right]$$

How do we count frequency?

- *Elsahar et al., 2018*: "subject and object co-occurrence is a good proxy for the mention of facts"
- We count co-occurrences of subject-object pairs in pretraining data
- We wrote a (cython!) library that efficiently counts cooccurrences and scale to trillions of tokens!

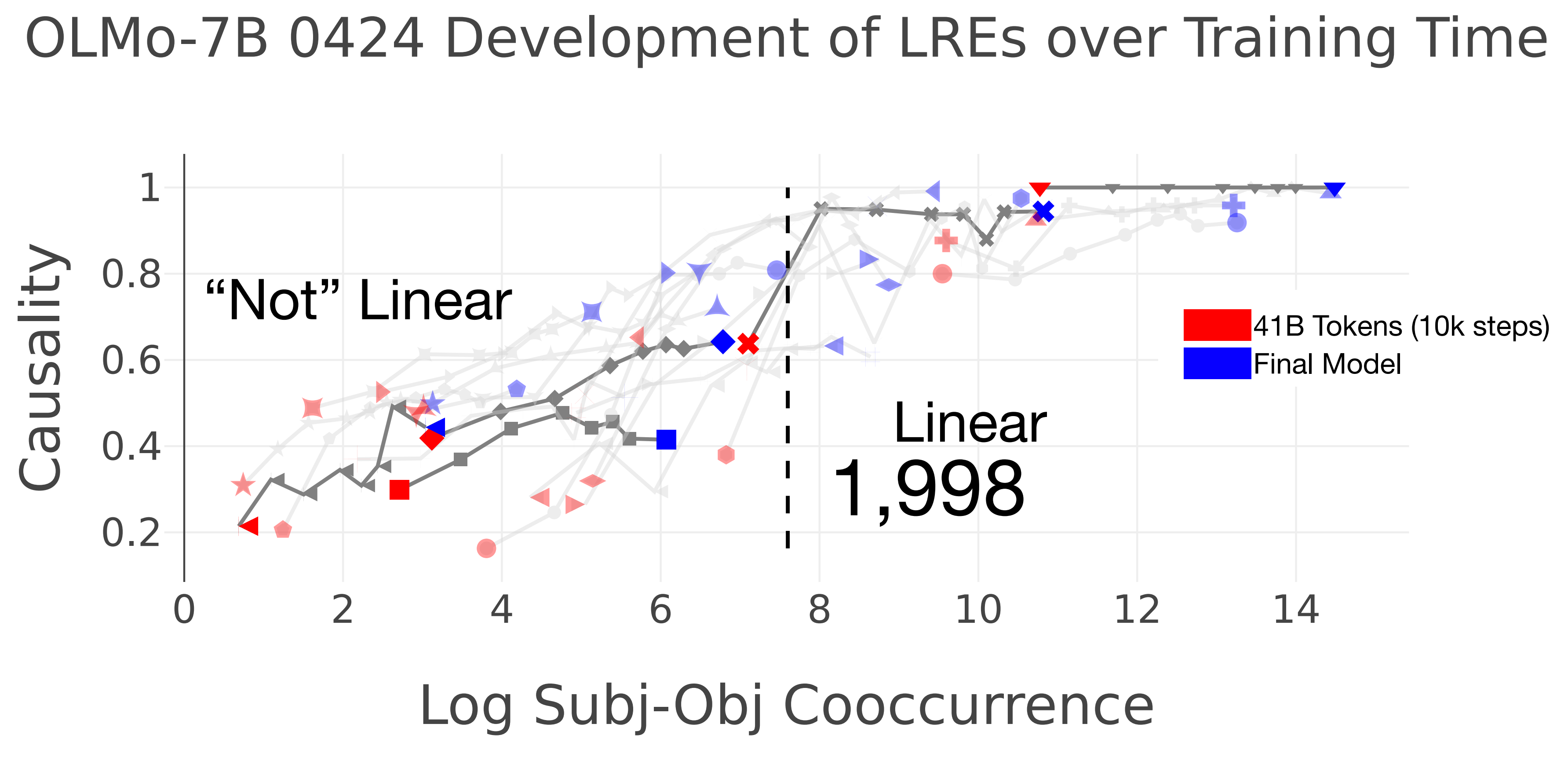
1) Example linear structure in **static** embeddings: Word vector analogies (Mikolov et al., 2013).



Motivation: linear structure also appears in LMs' representations, **sometimes** (Hernandez et al., 2024).

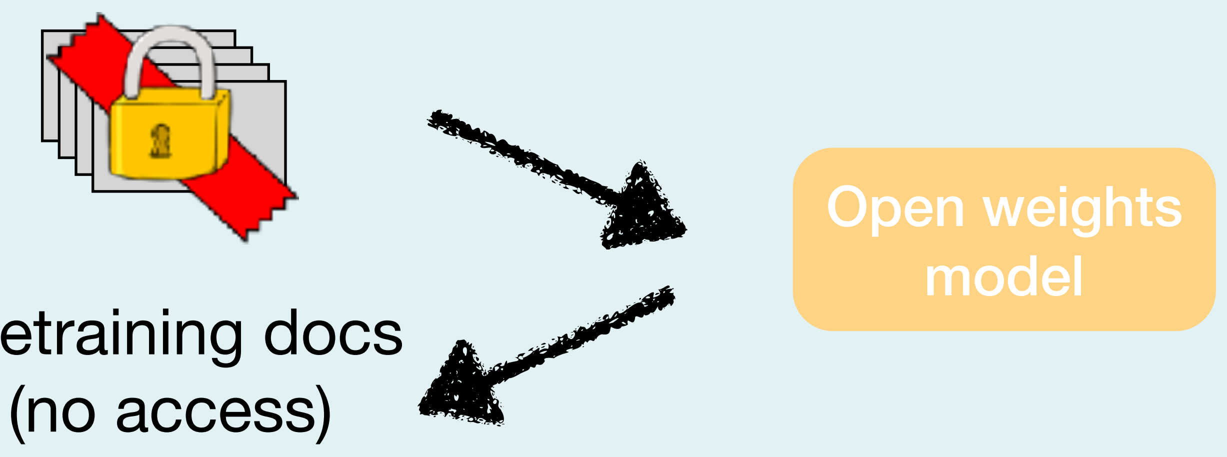
2) Can we explain (or predict) when linear representations appear in LLMs?

💡 **Finding**: Regardless of pretraining step, if a relation is **common enough, it is has linear structure in the representations**


OLMo-7B 0424 Development of LREs over Training Time

3) **Finding**: We show that factual relations accuracy ***and*** linearity is **highly correlated** with the subject-object co-occurrence frequency in the model's pretraining data!

4) **Application**: Presence/absence of linear structure can help predict frequency of individual terms



Input	Label	Ground Truth	Predicted
Emma Watson went to university at	*Brown University*	80,660	57,867
India's currency is	*Rupee*	36,055	68,523
Nvidia's CEO is	*Jensen Huang*	1,603	575