Meaning-infused grammar: Gradient Acceptability Shapes the Geometric Representations of Constructions in LLMs

Supantho Rakshit [1] Adele E. Goldberg[2]

[1]Dept. of Electrical and Computer Engineering, Princeton University [2]Dept. of Psychology, Princeton University

Background

Usage-Based Construction Grammar (UCx): Language is a network of learned gradient form-function pairings (*constructions*).

Human preferences for English Double Object (**DO**, e.g., *She gave the library the book*) vs. Prepositional Object (**PO**, e.g., *She gave the book to the library*) varying systematically and gradiently, depending on verb and length and definiteness of arguments.

Question

Does the internal representation geometry of LLMs mirror the gradient of human acceptability judgments for DO and PO sentences?

Hypothesis: More prototypical sentences of each construction will have more geometrically distinct representations in LLM's activation space.

5000 DO and PO sentences binned into 5 Tiers

1. "Top 10% most preferred" — PO and DO sentences strongly preferred in those constructions by people*

5. "Equi-preferred baseline" — DO and PO sentences people considered* equally acceptable in the alternative construction

*from **DAIS dataset:** 5,000 DO/PO sentence pairs with human preference ratings (Hawkins et al. 2021)

Method

- **Language model:** Pythia-1.4B (24 layers)
- For each sentence, extracted mean-pooled and normalized hidden states from each layer of model.
- Reduced dimensionality with 150 principal components, capturing 88.01% variance (averaged across layers).
- Measured geometric separability with **Energy Distance** and **Jensen-Shannon Divergence (JSD)**:

Energy Distance and JSD

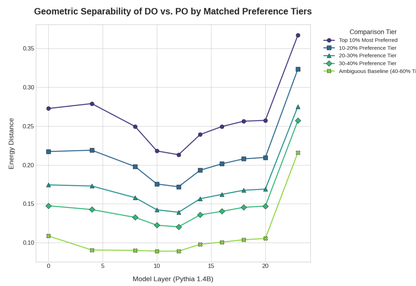
Tracking separation at the 5 preference tiers, across layers, to test if model geometry reflects graded preferences.

1. **Energy Distance** measures how *far apart* two distributions are, accounting for both location and spread (= 0 if distributions are identical)

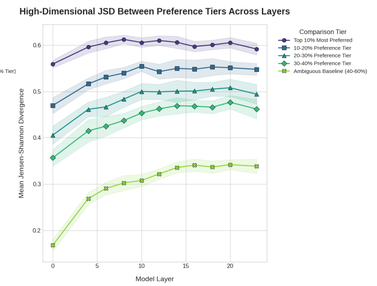
2. **Jensen–Shannon Divergence (JSD):** the "symmetric distance" between distributions (bounded between 0 and log 2); provides sensitive, high-dimensional measure of distributional separability.

See the main paper for mathematical definitions and implementation

Gradient separability, dependent on human preferences



Energy Distance plot by layer



JSD plot by layer

Using either separability measure, Energy Distance or JSD:

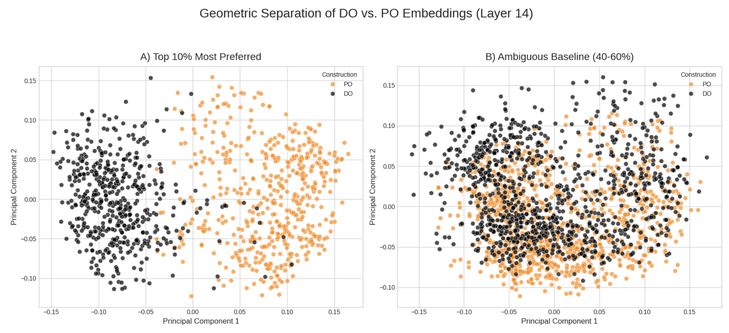
Stronger DO/PO bias preference tiers are more separable — mirroring human prototypicality judgments from DAIS — at each layer of Pythia-1.4B

Equi-preferred sentences (lightest green) (equally acceptable in the alternative construction) are the most entangled, least separable

Results hold across suite of Pythia models

-> LLMs encode gradient pairings of form *and function*: more prototypical instances of each construction display sharper geometric separation.

AT A GLANCE



PCA projection of model's representations for DO (orange) and PO (black) sentences

Left: Top 10% most preferred in DO or PO: When human preferences are strongest for PO or DO, constructions occupy clearly distinct regions in representational space

Right: Equi-preferred baseline: When human preferences for DO or PO are weak, the constructions' representations intermingle, despite DO and PO being syntactically distinct.

Main Takeaways

LLM's internal geometry is better able to distinguish instances of different constructions when those instances obey construction-specific constraints, related to lexical semantics (verbs) and information structure (length and definiteness)

These findings suggest LLMs learn a rich, dynamic, and meaning-infused grammar, consistent with the Usage-Based Constructionist approach.