What Comes to Mind? Interpretable Dimensions in Embedding Space Predict Human Ad Hoc Category Construction

Anonymous Author(s)
Affiliation
Address
email

Abstract

Humans rapidly construct ad hoc categories—e.g., "vegetables for painting"—by recruiting task-relevant properties and retrieving items that score highly on them. We test whether this behavior can be predicted directly from off-the-shelf word embeddings. Across 20 composite categories, we fit per-category elastic-net binomial GLMs over fastText dimensions and evaluate on item-mention probabilities. A sparse linear readout predicts human behavior with strong aggregate accuracy ($r = 0.699$ across $N = 3458$ pairs; **Brier** $= 0.0049$) and is well calibrated (ECE $= 0.0198$, improving slightly with an intercept-only adjustment). Beyond per-category fits, we frame *leave-one-base-out* (LOBO) as a retrieval transfer test: we learn a single modifier direction from non-target bases and blend it with the held-out base using a mixing weight γ tuned for *average precision*. This yields coherent semantic shifts and small, mixed retrieval gains (**mean** Δ**AP** $= 0.0051$, median 0.0025; 55.0% of 20 categories improved; mean ΔP@10 $= 0.0150$). The learned ad hoc axes align with human-rated properties (mRSA up to $R^2 = 0.227$, median $R^2 = 0.157$), supporting interpretability. Overall, simple, interpretable shifts in embedding space capture key regularities in what comes to mind under situational constraints.

1 Introduction

Humans can improvise new categories on the fly to meet immediate goals. Faced with an unusual constraint—e.g., needing "vegetables for painting"—people quickly recruit task-relevant properties (pigment yield, surface texture, shape as a brush) and retrieve items that score highly on those properties[1]. This kind of ad hoc categorization suggests a search over a structured representational space: items that align on contextually relevant dimensions come to mind first.

Traditional accounts model such spaces via behavioral feature elicitations and similarity judgments, which are powerful but slow to scale and necessarily incomplete[2]. The feature space that matters for a novel composite category may be high-dimensional, sparsely verbalizable, and only partially accessible to introspection. Meanwhile, contemporary language models (LMs) organize words in dense vector spaces that capture rich relational structure learned from text. If the relational structure in these spaces partially mirrors structure in the world, then it may be possible to predict which items people generate for ad hoc categories directly from embeddings—without hand-crafting features case by case.

This paper investigates that possibility. We ask whether interpretable linear paths in off-the-shelf embedding spaces can predict what "comes to mind" across a set of everyday base categories (e.g., kitchen appliances, vegetables, musical instruments) composed with task modifiers (e.g., "that could fit in your pocket", "you would use for self-defense"). Concretely, we frame item generation frequencies as binomial outcomes and fit per-category L1-regularized GLMs over embedding dimensions,

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

selecting penalty strength via cross-validation and calibrating post-hoc. Across 20 ad-hoc categories, simple models on fastText embeddings capture substantial variance in human item generation probabilities (Pearson $r = 0.699$; Brier $= 0.0049$), with small but reliable gains after calibration (ECE: $0.0198 \rightarrow 0.0189$).

Beyond per-category probes, we also test *transfer*: can a *single* modifier axis learned on non-target bases help retrieve items in a held-out base (LOBO)? We treat this as a retrieval problem and tune the blend weight γ for *average precision* on validation; we then report ΔAP and P@10 as primary LOBO metrics. The learned direction yields coherent geometric shifts and small, heterogeneous improvements (mean $\Delta\text{AP} = 0.0051$; median 0.0025; see §4.4). Finally, we assess interpretability via multi-feature RSA, where distances along each ad hoc axis are partially explained by distances in human feature ratings (up to $R^2 = 0.227$, median $R^2 = 0.157$).

2 Data

2.1 Tasks and responses

We analyze **20** ad hoc categories (e.g., *"kitchen appliances you would use for self-defense"*), generated by participants recruited on Prolific (N=121). Each row is an item–category pair with: `count` (number of unique participants who generated the item) and `n` (number of participants shown that category).

2.2 Cleaning & normalization

We (i) lowercase/trim responses; (ii) merge near-duplicates with a conservative string-similarity procedure; and (iii) compute per-category totals n_c. To avoid $\pm\infty$ logits for 0/1 extremes we evaluate metrics on probabilities rather than logits.

2.3 Embeddings

Backend: FastText `get_sentence_vector` with L2 norm and global mean-centering.

3 Method

For category c and item i with embedding $x_i \in \mathbb{R}^D$ and mentions $y_i \sim \text{Binomial}(n_c, p_i)$,

$$\text{logit}(p_i) = \beta_0^{(c)} + x_i^\top \beta^{(c)}.$$

We fit, separately for each category c, a binomial GLM on L2–normalized embedding dimensions. All feature columns are z–scored within category. We estimate $(\beta_0^{(c)}, \beta^{(c)})$ with an elastic-net penalty (mixing parameter α=0.90; 90% ℓ_1, 10% ℓ_2), keeping the intercept unpenalized:

$$\min_{\beta_0^{(c)},\, \beta^{(c)}} \sum_i \big[-y_i \log p_i - (n_c - y_i) \log(1 - p_i) \big] + \lambda \Big(\alpha \, \|\beta^{(c)}\|_1 + \tfrac{1-\alpha}{2} \, \|\beta^{(c)}\|_2^2 \Big) .$$

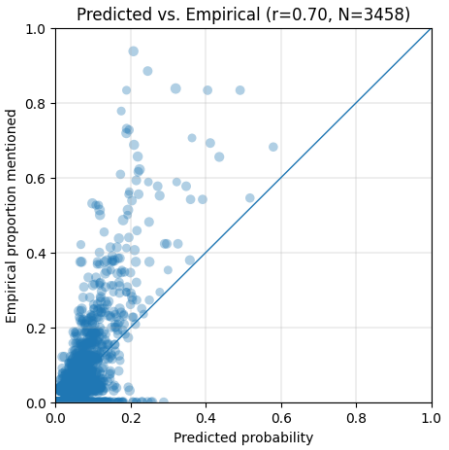
Model selection and stability. We select λ by *stratified 5-fold* CV within category (stratifying on $y > 0$), with LOOCV as a fallback for very small categories. After choosing λ, we *refit* on all items and report metrics from these refit predictions.

Calibration. Optionally, we apply a *one-parameter intercept-only* recalibration per category on the refit predictions (minimizing n-weighted Brier). Coefficients $\beta^{(c)}$ remain fixed, preserving sparsity.

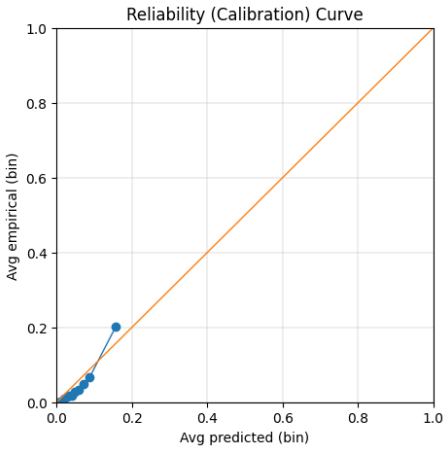
4 Results

4.1 Overall performance

Across all categories and items (N=3458 pairs), per-category L1-penalized binomial GLM predicts human probabilities with strong aggregate accuracy ($r = 0.699$, **Brier**$= 0.0049$, **ECE**$= 0.0198$; Fig. 1A). A one-parameter, intercept-only recalibration preserves sparsity and yields a small ECE gain without changing overall error (ECE $0.0198 \Rightarrow 0.0189$; Fig. 1B).

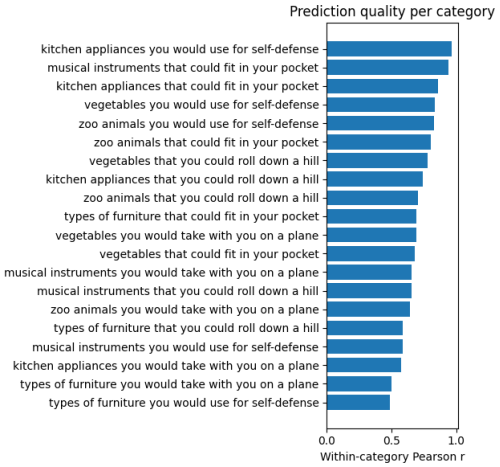


(a) Predicted vs. empirical probabilities across all items ($r = 0.699$).

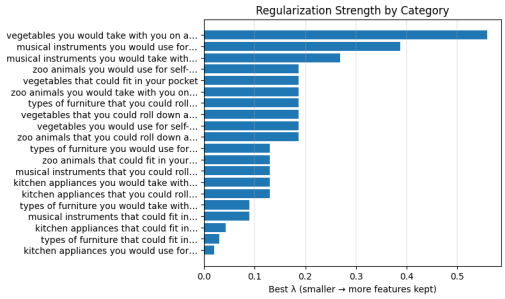


(b) Reliability before/after intercept recalibration (ECE $0.0198 \rightarrow 0.0189$).

Figure 1: Overall performance and calibration.



(a) Within-category Pearson r (higher is better).



(b) Selected regularization strength λ per category.

Figure 2: Per-category performance and sparsity.

4.2 Within-category analysis

Prediction quality is high within individual categories (Fig. 2), with several exceeding $r \approx 0.8$ (e.g., self-defense appliances, instruments in your pocket). The same sparse linear readout adapts to diverse ad hoc constraints by reweighting embedding dimensions per category.

4.3 Calibration, Sparsity and Interpretability

Reliability curves reveal mild under-confidence in the low-probability regime (≈.05–.20), which the intercept shift corrects while leaving the slope intact (Fig. 1B). Cross-validated λ values scale sensibly with category difficulty/sparsity (Appendix panels in Fig. 2B). For each category, only a small set of embedding axes receives non-zero weight, producing compact, readable profiles.

4.4 LOBO as retrieval (AP-tuned)

We ask whether a *single* modifier direction, learned from non-target bases, transfers to a held-out base. We blend the base model with the learned modifier via $p_{\text{enh}} = (1-\gamma)p_{\text{base}} + \gamma p_{\text{mod}}$ and choose γ to *maximize average precision* on validation within the held-out base; we then evaluate on the target items and report ΔAP and P@10. Geometrically, the learned axis induces coherent shifts that align with intuitive semantics (e.g., harm potential for *self-defense*; Fig. 3). Quantitatively, improvements

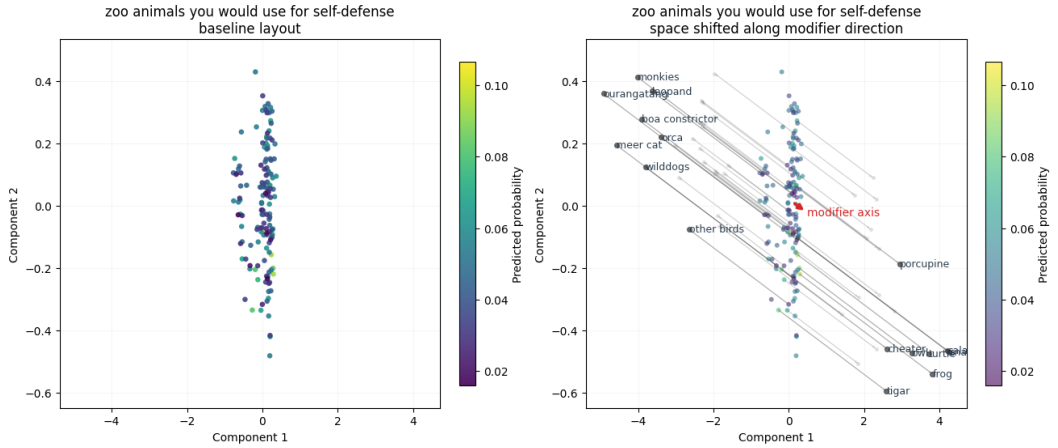


Figure 3: LOBO modifier shift for *zoo animals* → *use for self-defense* (single modifier direction learned from other bases; γ tuned for AP).

are small and heterogeneous across 20 categories: **mean** Δ**AP** $= 0.0051$, median 0.0025 (**55.0%** improved). P@10 shows a modest average lift (mean ΔP@10 $= 0.0150$, median 0.0000; 25.0% improved). Table A2 lists the largest gains.

4.5 Representational structure

To test whether the learned 1-D ad hoc axis aligns with human-rated properties, we performed a multi-feature RSA (mRSA): pairwise distances along the GLM axis were regressed on pairwise distances in human feature ratings for the corresponding base category, with permutation tests for significance. Several categories show robust explanatory power (Fig. 4). The associated feature-weight patterns are intuitive: for "kitchen appliances that could fit in your pocket," differences in *requires electricity*, *common*, and *expensive* account for much of the axis structure; for "zoo animals you would use for self-defense," *dangerous*, *striking*, and *large* are dominant. Across categories, mRSA explained up to $R^2 = 0.227$ (median $R^2 = 0.157$).

5 Limitations

Single-embedding view ignores compositional phrasing and pragmatic constraints; sparsity aids interpretability but may miss weak multi-axis signals; calibration is mainly in the low-p regime due to task prevalence; cross-base transfer effects are small on average.

6 Conclusion

We asked whether a simple, interpretable readout from off-the-shelf text embeddings can approximate which items people retrieve when composing ad hoc categories. Across 3458 item–category pairs, per-category elastic-net binomial GLMs achieved strong aggregate accuracy ($r = 0.699$, **Brier** $= 0.0049$, **ECE** $= 0.0198$), and a one-parameter intercept shift improved calibration without changing coefficients. The learned axes are compact and readable, and relate to human properties (mRSA up to $R^2 = 0.227$, median $R^2 = 0.157$). In a transfer test (LOBO) framed as retrieval and tuned for AP, we observe coherent semantic shifts and small, mixed gains (**mean** Δ**AP** $= 0.0051$; median 0.0025; mean ΔP@10 $= 0.0150$). Our results demonstrate that sparse linear readouts on standard word embeddings can serve as a simple interpretable baseline for ad hoc category composition, with the modest transfer gains indicating the need for richer composition models.

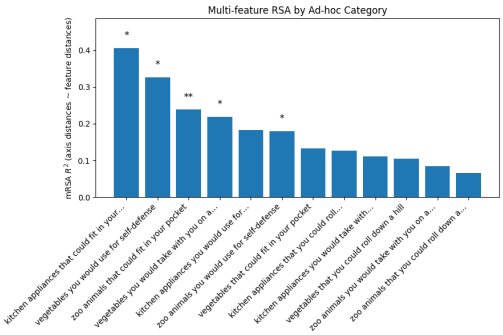
Appendix

Table A1: Overall and per-category metrics (mean±sd across categories).

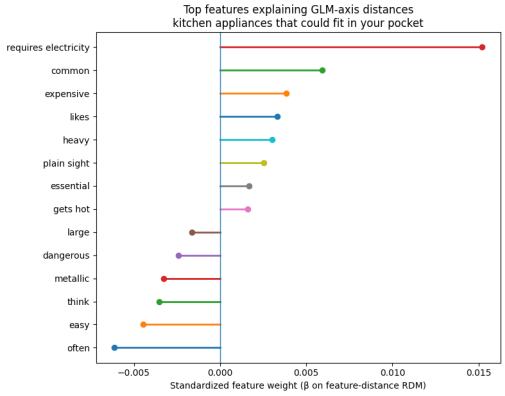
Setting	Pearson r	Brier ↓	ECE ↓
L1-penalized binomial GLM (raw)	0.699	0.0049	0.0198
L1-penalized binomial GLM (+recal)	0.698	0.0049	0.0189

Table A2: LOBO (AP-tuned): top improvements by ΔAP.

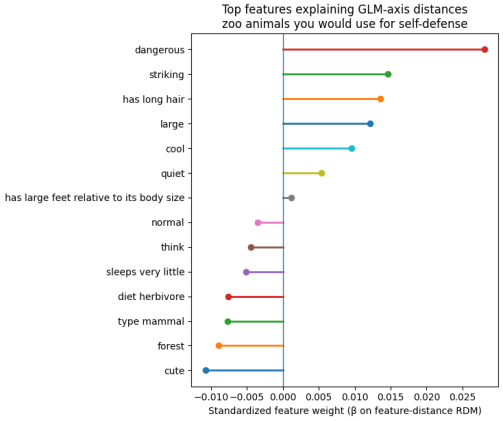
Category	AP (base)	AP (enh)	ΔAP	AUC (base)	AUC (enh)	ΔAUC	γ^*
vegetables that you could roll down a hill	0.742	0.761	0.019	0.803	0.795	-0.008	0.50
kitchen appliances you would take with you on a plane	0.791	0.808	0.017	0.813	0.814	0.001	0.35
zoo animals you would use for self-defense	0.643	0.659	0.015	0.695	0.695	-0.000	0.20
types of furniture that could fit in your pocket	0.900	0.912	0.013	0.896	0.895	-0.001	0.45
kitchen appliances that could fit in your pocket	0.794	0.804	0.010	0.824	0.836	0.012	0.10
zoo animals that could fit in your pocket	0.780	0.786	0.006	0.696	0.698	0.001	0.10



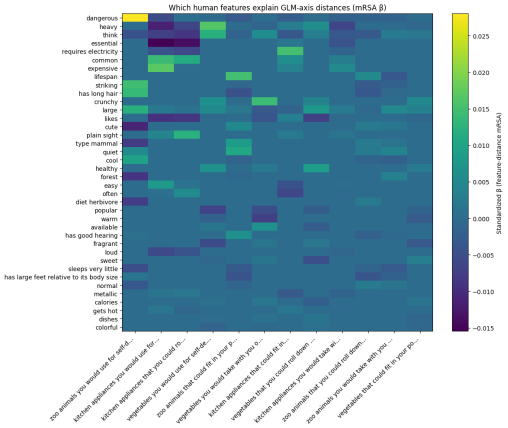
(a) Multi-feature RSA by Ad Hoc Category (asterisks denote significance).



(b) Top-$|\beta|$ features: *kitchen appliances that could fit in your pocket*.



(c) Top-$|\beta|$ features: *zoo animals for self-defense*.



(d) Heatmap of standardized feature weights across categories.

Figure 4: Representational structure and feature explanations.

References

[1] Lawrence W. Barsalou. Ad hoc categories. 11(3):211–227.

[2] Tracey Mills and Jonathan Phillips. Locating what comes to mind in empirically derived representational spaces. 240:105549.