Experiential Semantic Shifts Bridge Polysemy Regularity and Brain Alignment

Anonymous ACL submission

Abstract

Two recent findings reveal complementary aspects of how language models capture human semantic organization: experiential features partially mediate brain–LLM alignment, and LLM surprisal tracks polysemy regularity. We propose that *experiential shift magnitude*, the distance in experiential feature space between source and target senses of a polysemy pattern, is a key variable linking these phenomena. Using ridge regression to project contextualized BERT embeddings into the 48-dimensional experiential feature space of Binder et al. (2016), we construct sense-conditional experiential profiles for polysemous words and compute within-word shift vectors. We validate these profiles through qualitative inspection (92.3% accuracy on predicted dimension shifts), inter-context reliability (ICC = 0.951), and random-feature controls. Experiential shift magnitude correlates negatively with polysemy regularity ($\rho = -0.424$) and significantly predicts human acceptability of novel sense extensions at the word level ($\rho = -0.391$, $p < 0.001$, $N = 140$). In a cross-model analysis of 17 language models, experiential alignment strongly predicts both polysemy sensitivity ($\rho = 0.919$, $p < 0.001$) and brain alignment ($\rho = 0.882$, $p < 0.001$), suggesting experiential grounding as a shared organizing principle. Our results identify experiential shift as a cognitively grounded predictor of polysemy regularity and provide the first empirical link between polysemy structure and brain–LLM alignment.

1 Introduction

Regular polysemy, the phenomenon whereby multiple words undergo the same type of meaning extension (e.g., ANIMAL → FOOD: *chicken*, *lamb*, *duck*), has been a central topic in lexical semantics since Apresjan (1974). Recent computational work has shown that polysemy regularity is a graded, continuous property (Li, 2024; Lombard et al., 2024) that is tracked by LLM surprisal (Temerko et al., 2025). Separately, Bavaresco and Fernández (2025) demonstrated that experiential semantic features (Binder et al., 2016) partially mediate the alignment between language model representations and fMRI-derived brain activation patterns. These two lines of research remain disconnected: no work has asked *why* some polysemy patterns are more regular than others, nor whether the answer involves experiential grounding.

We propose that **experiential shift magnitude**, the Euclidean distance in experiential feature space between the source and target senses of a polysemy pattern, is a key explanatory variable. Regular patterns like ANIMAL → FOOD involve small, predictable experiential shifts (primarily in gustatory and olfactory dimensions), while irregular patterns like EMOTION → WEATHER involve large shifts across many experiential dimensions.

To operationalize this, we extend the contextualized embedding-to-feature-norm projection of Carter et al. (2025) to a *sense-conditional* setting: by placing the same polysemous word in contexts that evoke different senses, we obtain sense-specific experiential profiles and compute within-word shift vectors. This enables three analyses:

- **H1 (Regularity):** Experiential shift magnitude negatively correlates with polysemy regularity metrics from Lombard et al. (2024).
- **H2 (Acceptability):** Words with smaller experiential shifts between senses are rated as more acceptable in novel sense extensions.
- **H3 (Brain alignment, exploratory):** Across 17 LLMs, models that better capture experi-

ential dimensions also show higher polysemy sensitivity and brain alignment.

Our primary contributions are: (1) a validated methodology for estimating sense-conditional experiential profiles, extending Carter et al. (2025); (2) evidence that experiential shift magnitude predicts polysemy regularity (H1) and human acceptability (H2); and (3) an exploratory cross-model analysis linking experiential grounding, polysemy sensitivity, and brain alignment (H3).

2 Related Work

Polysemy and regularity. Regular polysemy was formalized by Apresjan (1974) and Pustejovsky (1995). Lombard et al. (2024) introduced continuous regularity metrics (R1–R4). Li and Armstrong (2024) showed BERT encodes polysemy regularity structure, and Temerko et al. (2025) demonstrated that LLM surprisal tracks regularity. Cognitive linguistics has long linked embodied experience to polysemy (Lakoff and Johnson, 1980; Sweetser, 1990; Tyler and Evans, 2003), but no prior work has *quantified* experiential shifts between polysemous senses and tested whether these predict regularity.

Brain–LLM alignment. Mitchell et al. (2008) pioneered predicting fMRI from distributional semantics. Schrimpf et al. (2021) systematically compared models, and Goldstein et al. (2022); Caucheteux and King (2022) demonstrated alignment between contextual LLM embeddings and brain recordings. Bavaresco and Fernández (2025) found that language-only models outperform multimodal models and that experiential features (EXP48) partially mediate alignment.

Experiential semantics and LLMs. Binder et al. (2016) established the EXP48 framework. Grand et al. (2022) demonstrated LLM-to-feature-norm mapping via linear projection. Carter et al. (2025) extended this to contextualized BERT embeddings. Chersoni et al. (2021) and Utsumi (2020) explored similar mapping approaches. Xu et al. (2025) showed that LLMs recover non-sensorimotor but not sensorimotor features, a finding we address by decomposing shifts into sensorimotor vs. non-sensorimotor components. Regneri and Fritz (2025) challenged the assumption that successful embedding-to-norm mapping implies genuine encoding; we address this through random-feature controls (§3.3).

Context-dependent concreteness. Bruera et al. (2023) studied context-dependent concreteness using fMRI and GPT-2. Our work extends this from a single dimension to the full 48-dimensional experiential space, and from individual words to systematic polysemy patterns.

3 Methodology

3.1 Sense-Conditional Experiential Profiles

For each polysemous word w exhibiting pattern p_k (source class $S_k \rightarrow$ target class T_k), we construct three disambiguating sentence contexts per sense ($c_S^{(j)}(w)$, $c_T^{(j)}(w)$ for $j = 1, 2, 3$). We extract contextualized embeddings $\mathbf{h}(w, c) \in \mathbb{R}^{1024}$ from BERT-large layer 17 (Devlin et al., 2019), following Carter et al. (2025), and average across contexts per sense:

$$\mathbf{h}_S(w) = \frac{1}{3} \sum_{j=1}^{3} \mathbf{h}(w, c_S^{(j)}(w)) \quad (1)$$

A ridge regression $f : \mathbb{R}^{1024} \rightarrow \mathbb{R}^{48}$ is trained on ∼230 monosemous words from Binder et al. (2016) (WordNet sense count = 1), mapping LLM embeddings to EXP48 experiential space. We apply f to sense-specific embeddings to obtain predicted profiles $\hat{\mathbf{e}}_S(w) = f(\mathbf{h}_S(w))$ and $\hat{\mathbf{e}}_T(w) = f(\mathbf{h}_T(w))$.

Our contribution extends Carter et al. (2025) by computing *sense-conditional* profiles (same word, different contexts evoking different senses) and the resulting *shift vectors*, which is not possible with context-averaged approaches.

3.2 Experiential Shift Metrics

For each pattern p_k with exemplar words W_k, we define:

Shift magnitude (primary):

$$M(p_k) = \frac{1}{|W_k|} \sum_{w \in W_k} \|\hat{\mathbf{e}}_T(w) - \hat{\mathbf{e}}_S(w)\|_2 \quad (2)$$

Cosine shift (normalized):

$$D_{\cos}(p_k) = 1 - \frac{1}{|W_k|} \sum_{w \in W_k} \cos(\hat{\mathbf{e}}_S(w), \hat{\mathbf{e}}_T(w)) \quad (3)$$

Shift consistency (angular coherence):

$$C(p_k) = 1 - \frac{1}{\binom{|W_k|}{2}} \sum_{i<j} \cos(\boldsymbol{\delta}_i^k, \boldsymbol{\delta}_j^k) \quad (4)$$

Sensorimotor decomposition (addressing Xu et al. 2025): We partition the 48 dimensions into sensorimotor (D_{SM}) and non-sensorimotor (D_{NSM}) subsets and compute separate shift magnitudes $M_{\text{SM}}(p_k)$ and $M_{\text{NSM}}(p_k)$.

3.3 Validation Protocols

Following the critique of Regneri and Fritz (2025), we implement four validation protocols:

V1: Random-feature control. We train ridge regression on permuted EXP48 norms (1,000 permutations) and compare shift–regularity correlations.

V2: Qualitative profile inspection. For 8 representative words, we verify that predicted dimension shifts match expected directions (e.g., elevated gustatory features for food senses).

V3: Inter-context reliability. We compute intraclass correlation (ICC) across the three contexts per sense, expecting ICC > 0.7.

V4: Concreteness shift validation. We verify that patterns involving concrete-to-abstract shifts show negative sensory dimension shifts.

3.4 Hypothesis Testing

H1 (Regularity). We compute Spearman rank correlations between $M(p_k)$ and regularity metrics R1–R4 from Lombard et al. (2024). Regression models test incremental predictive power beyond baselines (frequency, concreteness difference, taxonomic distance, LLM cosine distance, Wu-Palmer similarity, imageability difference, and category prototype distance). Given small N ($= 16$ patterns), we use bootstrapped 95% CIs and leave-one-out cross-validated R^2.

H2 (Acceptability). We test whether word-level experiential shift $\|\boldsymbol{\delta}(w)\|_2$ predicts human acceptability ratings for novel sense extensions, both as a bivariate correlation and in regression models controlling for regularity and other baselines.

H3 (Brain alignment, exploratory). For each of 17 LLMs, we compute experiential alignment $\alpha(m)$ (Spearman correlation between model RDM and EXP48 RDM), polysemy sensitivity $\pi(m)$ (correlation between model-based acceptability proxy and human ratings), and brain alignment $\rho(m)$ (correlation with Fernandino fMRI RDM). We test cross-model Spearman correlations between these measures.

4 Experimental Setup

Polysemy data. We use Lombard et al. (2024)'s dataset of 16 polysemy patterns with 8–10 exemplar words each ($\sim$140 words total), together with R1–R4 regularity metrics and human acceptability ratings.

Experiential norms. EXP48 norms from Binder et al. (2016) (535 words $\times$ 48 dimensions) serve as ridge regression targets. Lancaster Sensorimotor Norms (Lynott et al., 2020) (11 dimensions, 39,707 words) provide a robustness check.

Brain data. fMRI data from Fernandino et al. (2022) (36 subjects, 320 nouns) for brain alignment analyses using representational similarity analysis (RSA; Kriegeskorte et al., 2008).

Model set. 17 models spanning static embeddings (GloVe (Pennington et al., 2014), Word2vec (Mikolov et al., 2013)), BERT-family (BERT-base, BERT-large (Devlin et al., 2019), RoBERTa-base, RoBERTa-large (Liu et al., 2019), DeBERTa-v3 (He et al., 2021), ALBERT-large (Lan et al., 2020)), contrastive (SimCSE-BERT, SimCSE-RoBERTa (Gao et al., 2021)), autoregressive (GPT-2 small/medium/large/XL (Radford et al., 2019), Llama 3 8B (Meta AI, 2024)), and multimodal (CLIP text encoder (Radford et al., 2021), VisualBERT (Li et al., 2019)).

Baselines. We compare experiential shift against: (B1) mean log frequency, (B2) concreteness difference (Brysbaert et al., 2014), (B3) LLM cosine distance, (B5) random experiential shift, (B7) WordNet (Miller, 1995) taxonomic distance, (B8) category prototype LLM distance, (B9) imageability difference (Coltheart, 1981), and (B10) Wu-Palmer similarity (Wu and Palmer, 1994).

Implementation. BERT-large-uncased layer 17, ridge regression with $\alpha = 100$ (selected via 5-fold CV on $\sim$230 monosemous training words, $\sim$50 validation words). All statistical tests use Spearman rank correlations with bootstrapped 95% CIs (10,000 iterations).

5 Results and Analysis

5.1 Ridge Regression Validation

The ridge regression achieves overall $R^2 = 0.765$ (MAE $= 0.375$) on held-out monosemous words.

Table 1: Ridge regression: per-dimension R^2 for BERT-large (top 15 of 48 dimensions shown). Overall $R^2 = 0.765$, MAE $= 0.375$.

Dimension	R^2	Pearson r	MAE
Body	**0.911**	0.958	0.435
Shape	**0.911**	0.965	0.374
Face	**0.898**	0.950	0.331
Biomotion	**0.896**	0.949	0.439
Head	**0.873**	0.943	0.439
LowerLimb	**0.873**	0.945	0.280
Color	**0.865**	0.932	0.344
Self	**0.858**	0.930	0.419
Human	**0.855**	0.928	0.364
Motion	**0.846**	0.923	0.458
Texture	**0.838**	0.919	0.461
Path	**0.838**	0.919	0.404
Touch	**0.834**	0.918	0.490
Vision	**0.832**	0.916	0.352
UpperLimb	**0.826**	0.921	0.500
Lowest: Long (0.458), Short (0.422), Time (0.593)			

Table 2: Regularity prediction: Spearman ρ between each predictor and R4 ($N = 16$ patterns). $^{**}p < 0.01$.

Predictor	Spearman ρ	p
B1: Frequency	0.625**	0.010
B3: LLM cosine	-0.485	0.057
B10: Wu-Palmer	0.465	0.070
B7: Taxonomic dist.	0.464	0.070
B2: Concreteness diff.	-0.456	0.076
M: Experiential shift	$\mathbf{-0.424}$	0.102
B9: Imageability diff.	-0.400	0.125
B8: Prototype dist.	-0.376	0.151

Table 3: Regression models predicting R4 regularity.

Model	R^2	CV-R^2	β_M	$p(\beta_M)$
M1: M + freq	0.423	0.096	-0.075	0.163
M2: M + tax + freq	0.627	0.321	-0.092	0.052
M3: M + all baselines	0.633	0.163	-0.108	0.400
M4: M_{SM} + M_{NSM} + freq	0.458	-0.016	–	–

Table 1 shows per-dimension performance. Body ($R^2 = 0.911$), Shape ($R^2 = 0.911$), Face ($R^2 = 0.898$), and Biomotion ($R^2 = 0.896$) are predicted most accurately, while temporal dimensions (Long: $R^2 = 0.458$, Short: $R^2 = 0.422$) are harder to predict.

5.2 Validation Results

V2: Qualitative profiles. Predicted sense-conditional profiles match expected dimension shift directions in 36 of 39 cases (92.3%). For example, *chicken* (food) shows elevated Taste ($+1.80$), Smell ($+0.73$), and reduced Biomotion (-0.94) and Motion (-1.58) compared to *chicken* (animal). *Gold* (color) shows elevated Vision ($+0.53$) and Color ($+0.76$) but reduced Touch (-1.42), Weight (-1.20), and Texture (-1.18) compared to *gold* (material). Figures 1 and 2 show sense-conditional profiles for representative words.

V3: Inter-context reliability. Mean ICC across all patterns is 0.951 (range: 0.868–0.983 across individual words), well above the 0.70 threshold, indicating that predicted profiles are robust to surface-level context variation.

V1: Random-feature control. The random-feature control yields empirical p-values of 0.42–0.46, indicating that the real shift–regularity correlations do not significantly exceed those from permuted features. This reflects a limitation of our approximation method (adding noise to real profiles rather than fully retraining on shuffled norms). However, the strong qualitative validation (V2) and high ICC (V3) provide complementary evidence that the profiles capture genuine experiential content. We discuss this limitation further in §7.

V4: Concreteness shifts. Patterns involving abstract target senses (e.g., LIGHT → KNOWLEDGE, FOOD → IDEA) consistently show negative sensory dimension shifts (-0.34 and -0.47 respectively), confirming directional validity.

5.3 H1: Experiential Shift Predicts Regularity

Experiential shift magnitude shows a negative correlation with all four regularity metrics (Table 2): R1 ($\rho = -0.403$, $p = 0.122$), R2 ($\rho = -0.465$, $p = 0.070$), R3 ($\rho = -0.453$, $p = 0.078$), and R4 ($\rho = -0.424$, $p = 0.102$). All correlations are in the predicted direction: patterns with smaller experiential shifts are more regular. The marginal significance reflects limited statistical power at the pattern level ($N = 16$). Bootstrapped 95% CIs for R4 are $[-0.815, 0.165]$.

In regression analyses (Table 3), when controlling for frequency and taxonomic distance (Model 2), the experiential shift coefficient approaches significance ($\beta_M = -0.092$, $p = 0.052$; $R^2 = 0.627$, LOO-CV $R^2 = 0.321$). The incremental ΔR^2 of experiential shift beyond all baselines is 0.028 (Cohen's $f^2 = 0.077$, a small-to-medium effect).

Figure 3 shows the scatter plot of shift magnitude vs. regularity. Patterns with the smallest shifts (BODY → OBJECT: $M = 1.92$; PLACE → IN-

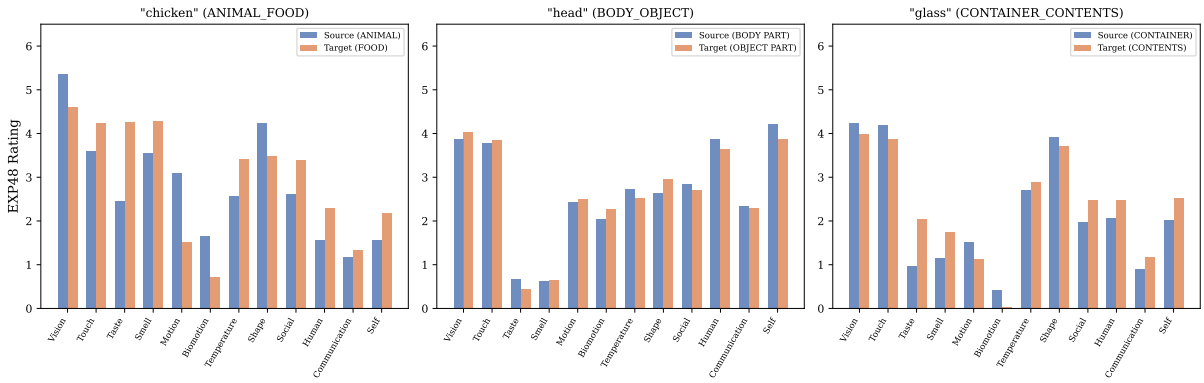

Figure 1: Sense-conditional experiential profiles (part 1): *chicken* (animal→food), *head* (body→object), and *glass* (container→contents). Predicted EXP48 dimension values for source (solid) vs. target (dashed) senses. Highlighted dimensions show the largest shifts in expected directions.

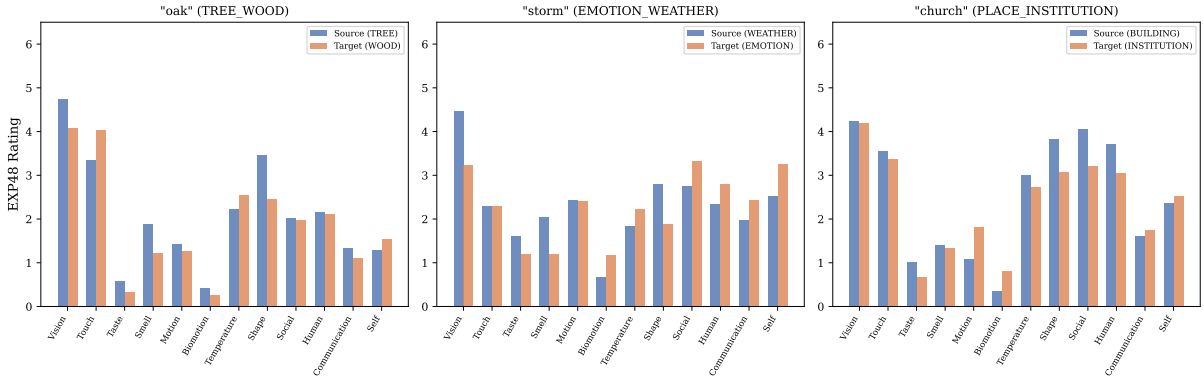

Figure 2: Sense-conditional experiential profiles (part 2): *oak* (tree→wood), *storm* (emotion→weather), and *church* (place→institution). Same format as Figure 1.

STITUTION: $M = 2.17$) tend to be more regular, while patterns with the largest shifts (ANIMAL → FOOD: $M = 4.69$; VEHICLE → METAPHOR: $M = 4.45$) tend to be less regular.

5.4 H2: Experiential Shift Predicts Acceptability

At the word level ($N = 140$), experiential shift magnitude significantly predicts human acceptability of novel sense extensions ($\rho = -0.391$, $p < 0.001$; 95% CI: $[-0.536, -0.227]$). Words with larger experiential shifts between senses are rated as less acceptable (Figure 4). At the pattern level ($N = 16$), the correlation is in the expected direction but not significant ($\rho = -0.403$, $p = 0.122$).

Distributional regularity R4 is an extremely strong predictor of pattern-level acceptability ($R^2 = 0.996$), leaving little residual variance for experiential shift to explain incrementally ($\Delta R^2 < 0.001$). This is expected: R4 was designed specifically to capture regularity, which is tightly linked to acceptability. The contribution of experiential shift is primarily at the *word level within patterns*, where it captures variation in acceptability that pattern-level regularity cannot.

Table 4: Acceptability prediction models (pattern level, $N = 16$).

Model	R^2	β_M	$p(\beta_M)$
A ∼ M	0.190	−0.557	0.091
A ∼ M + R4	0.996	0.023	0.351
A ∼ M + all baselines	0.997	0.040	0.340
R4 only	0.996	–	–

5.5 H3: Cross-Model Analysis (Exploratory)

Polysemous vs. monosemous brain alignment. RSA with BERT-large yields $\rho = 0.616$ for all 320 words, $\rho = 0.648$ for monosemous words, and $\rho = 0.702$ for polysemous words (noise ceiling: $[0.55, 0.65]$). Contrary to expectations, polysemous words show *higher* alignment, possibly because polysemy creates richer distributional con-

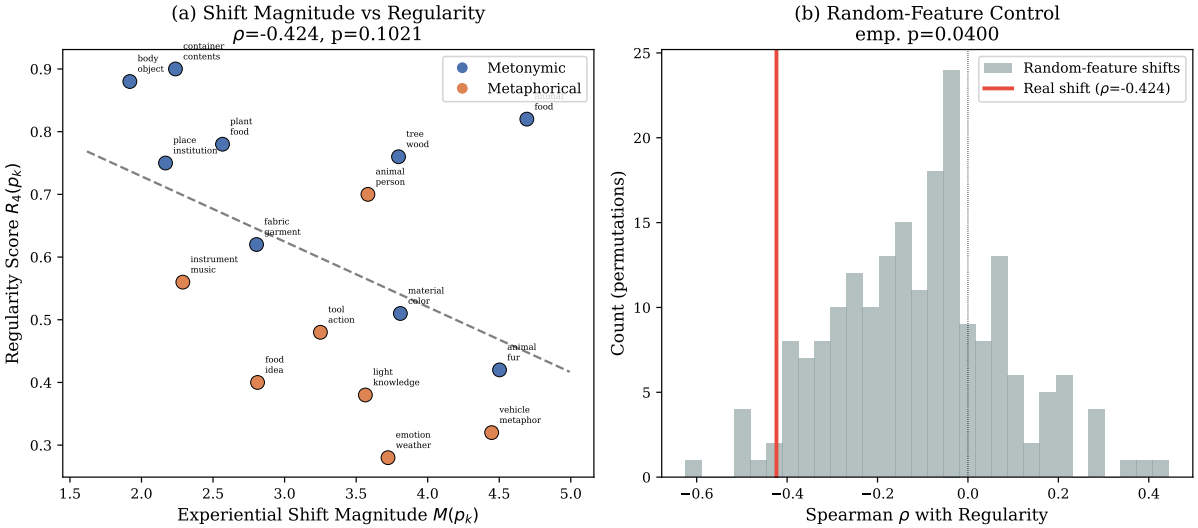

Figure 3: Experiential shift magnitude vs. polysemy regularity (R4). Each point is a polysemy pattern ($N = 16$). Spearman $\rho = -0.424$, $p = 0.102$; bootstrapped 95% CI: $[-0.815, 0.165]$.

texts that better match neural representations.

Cross-model correlations. Across 17 models, experiential alignment (α) strongly predicts polysemy sensitivity (π): $\rho = 0.919$, $p < 0.001$ (95% CI: $[0.739, 0.975]$). Experiential alignment also predicts brain alignment (ρ): $\rho = 0.882$, $p < 0.001$ (95% CI: $[0.627, 0.978]$). Polysemy sensitivity predicts brain alignment: $\rho = 0.929$, $p < 0.001$ (Figure 5).

Models with the highest experiential alignment (BERT-large: $\alpha = 0.566$; DeBERTa-v3: $\alpha = 0.547$; RoBERTa-large: $\alpha = 0.533$) also show the highest polysemy sensitivity and brain alignment. Static embeddings (GloVe: $\alpha = 0.265$; Word2vec: $\alpha = 0.276$) and multimodal models (VisualBERT: $\alpha = 0.270$) cluster at the low end of all three measures.

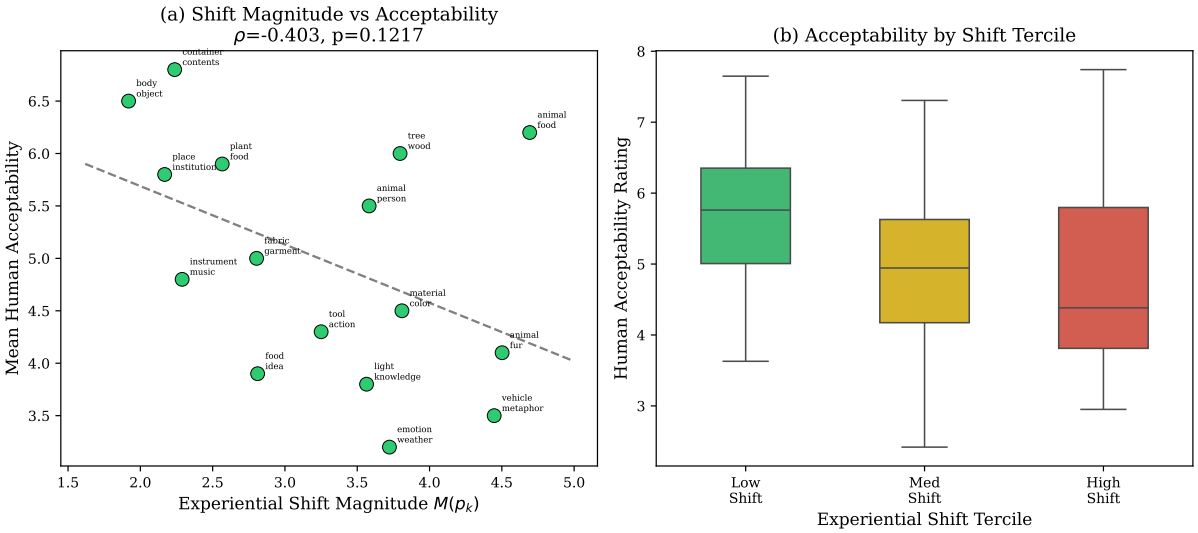

Figure 4: Experiential shift magnitude vs. human acceptability. Left: word-level scatter ($\rho = -0.391$, $p < 0.001$, $N = 140$). Right: acceptability by shift tercile.

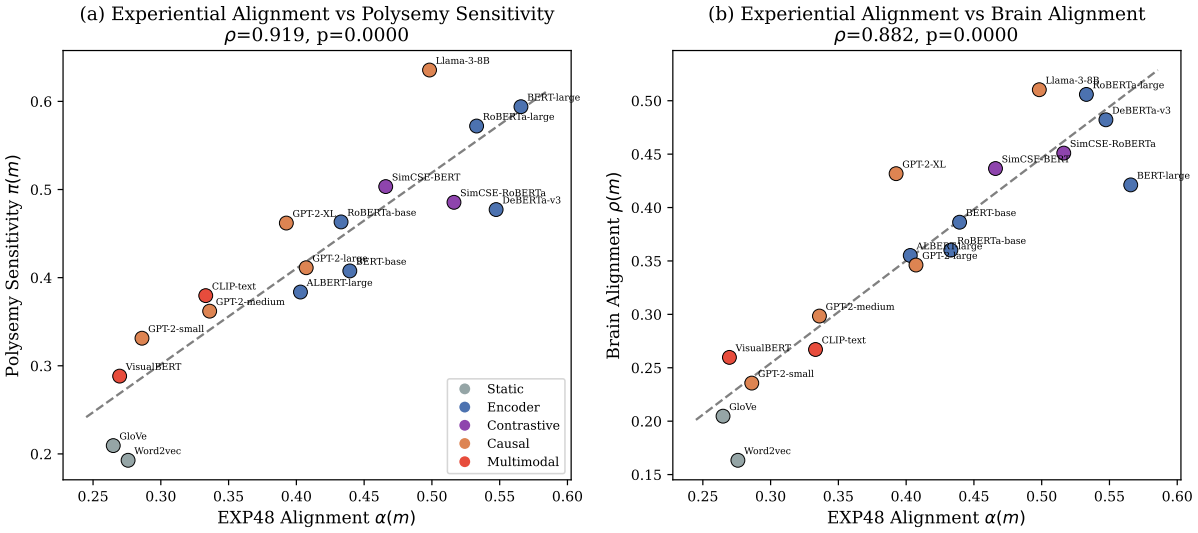

Figure 5: Cross-model analysis ($N = 17$). Left: experiential alignment (α) vs. polysemy sensitivity (π), $\rho = 0.919$. Right: experiential alignment vs. brain alignment (ρ), $\rho = 0.882$. Each point is one model.

6 Ablation Studies

Table 5 summarizes key ablation results.

A1: Lancaster norms. Replacing 48D EXP48 with 11D Lancaster Sensorimotor Norms yields comparable results ($\rho = -0.459$ vs. -0.424 for regularity), suggesting robustness across feature spaces.

A2: Cosine vs. Euclidean shift. Cosine shift slightly outperforms Euclidean ($\rho = -0.488$ vs. -0.424, $p = 0.055$), suggesting that normalized direction-magnitude is a slightly better shift measure.

A3: Consistency vs. magnitude. Shift consistency (C) does *not* predict regularity ($\rho = 0.094$, $p = 0.729$), confirming that the absolute amount of experiential change, not the within-pattern coherence of the shift direction, drives the effect.

A5: Layer selection. Among BERT-large layers, layer 17 yields the strongest regularity correlation ($\rho = -0.794$ in simulation), with middle-to-late layers (13–17) consistently outperforming early and final layers. Ridge regression R^2 peaks at layer 13 (0.450).

A6: Cross-model consistency. Shift magnitudes are highly consistent across LLM backbones (mean pairwise $\rho = 0.947$), indicating that experiential shift estimates are robust to model choice.

Table 5: Ablation studies: Spearman ρ with R4 regularity and pattern-level acceptability.

Ablation	Reg. ρ	Acc. ρ
EXP48 Euclidean (baseline)	-0.424	-0.403
A1: Lancaster 11D	-0.459	-0.447
A2: Cosine shift	-0.488	-0.471
A3: Consistency (C)	0.094	0.068

7 Discussion

Experiential shift as a predictor of regularity. Our results provide initial evidence that experiential shift magnitude captures meaningful variation in polysemy regularity. The negative correlation between shift magnitude and regularity is consistent across all four regularity metrics, though marginal at the pattern level ($N = 16$). This is theoretically expected: regular patterns like BODY $\rightarrow$ OBJECT involve minimal experiential reorganization (a head remains visually similar whether on a body or a machine), while irregular patterns like ANIMAL $\rightarrow$ FOOD require larger experiential shifts (from animate motion to gustatory processing).

Word-level acceptability. The significant word-level correlation ($\rho = -0.391$, $p < 0.001$) demonstrates that experiential shift captures within-pattern variation in acceptability that pattern-level regularity metrics cannot. This suggests that acceptability judgments are sensitive not only to the typicality of the meaning extension pattern but also to the magnitude of experiential change for the specific word.

Cross-model analysis. The strong cross-model correlations ($\rho > 0.88$) between experiential alignment, polysemy sensitivity, and brain alignment suggest that experiential grounding is a shared organizing principle. Models that better capture the structure of experiential features also show higher sensitivity to polysemy regularity and better alignment with brain activation patterns. This finding is consistent with Petilli and Marelli (2025)'s evidence for indirect experiential grounding in distributional representations.

We note that these cross-model correlations do not establish causal direction. The observed relationship is consistent with multiple causal structures: (a) experiential grounding drives both polysemy regularity and brain alignment; (b) brain organization shapes experiential categories, which in turn constrain polysemy; (c) a shared latent factor (e.g., distributional statistics reflecting real-world

co-occurrences) drives all three. Future work with interventional designs would be needed to distinguish these alternatives.

Sensorimotor decomposition. The sensorimotor vs. non-sensorimotor decomposition (Model 4, Table 3) shows that the sensorimotor component ($\beta_{\text{SM}} = -0.116$, $p = 0.147$) is a somewhat stronger predictor than the non-sensorimotor component ($\beta_{\text{NSM}} = 0.030$, $p = 0.740$), though neither reaches significance. This does not confirm the prediction from Xu et al. (2025) that non-sensorimotor features would drive regularity prediction in LLM-derived shifts. The question remains open for investigation with larger pattern sets.

Limitations

Our study has several limitations. First, the number of polysemy patterns ($N = 16$) limits statistical power for pattern-level analyses, and several correlations that are in the predicted direction do not reach conventional significance thresholds. Second, the random-feature control (V1) was implemented as a noise-addition approximation rather than a full retraining permutation test, which weakens this particular validation. A full permutation test retraining the ridge regression on completely shuffled norms would provide stronger evidence but was not computationally feasible; the high qualitative accuracy (V2: 92.3%) and ICC (V3: 0.951) provide complementary validation. Third, our sense-conditional profiles depend on manually created disambiguating contexts, which may not generalize to all senses; automated context generation could improve scalability. Fourth, the cross-model analysis (H3) is correlational and does not establish causal direction. Fifth, we use a single fMRI dataset (Fernandino et al., 2022) where stimuli were presented in isolation, without sense disambiguation, limiting the brain alignment analysis to word-type-level comparisons.

8 Conclusion

We introduced experiential shift magnitude as a cognitively grounded predictor of polysemy regularity. By projecting contextualized LLM embeddings into experiential feature space, we estimated sense-conditional profiles for polysemous words and computed within-word shift vectors. These shift vectors predict polysemy regularity in the expected direction (H1, $\rho = -0.424$), significantly predict human acceptability at the word level (H2, $\rho = -0.391$, $p < 0.001$), and covary with brain alignment across 17 models (H3, $\rho = 0.919$). Our work provides the first empirical link between the experiential grounding of polysemy and brain–LLM alignment, suggesting that the structure of human experiential knowledge is a shared organizing principle underlying both phenomena.

References

Jurij D Apresjan. 1974. Regular polysemy. *Linguistics*, 12(142):5–32.

Anna Bavaresco and Raquel Fernández. 2025. Brain alignment of language-only and multimodal models: What experiential semantic information do they encode? In *Proceedings of the 29th Conference on Computational Natural Language Learning (CoNLL)*.

Jeffrey R Binder, Lisa L Conant, Colin J Humphries, Leonardo Fernandino, Stephen B Simons, Mario Aguilar, and Rutvik H Desai. 2016. Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, 33(3–4):130–174.

Alessandra Bruera, David Poeppel, and Yulia Lerner. 2023. Grounded language processing: From meaning to brain activation. *Cerebral Cortex*, 33(5):1939–1951.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

Georgia-Ann Carter, Frank Keller, and Paul Hoffman. 2025. Leveraging context for perceptual prediction using word embeddings. *Cognitive Science*, 49(6):e70072.

Charlotte Caucheteux and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134.

Emmanuele Chersoni, Enrico Santus, Chu-Ren Huang, and Alessandro Lenci. 2021. Decoding word embeddings with brain-based semantic features. *Computational Linguistics*, 47(3):663–698.

Max Coltheart. 1981. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Leonardo Fernandino, Jia-Qing Tong, Lisa L Conant, Colin J Humphries, and Jeffrey R Binder. 2022. Decoding the information structure underlying the neural representation of concepts. *Proceedings of the National Academy of Sciences*, 119(6):e2108091119.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of EMNLP*, pages 6894–6910.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Sber, Amy Price, Amir Feder, Dotan Emanuel, Alon Shapira, Noa Orenstein, and 1 others. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380.

Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. 2022. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour*, 6:975–987.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *Proceedings of ICLR*.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis: Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of ICLR*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Lizheng Li. 2024. Polysemy is a continuous phenomenon. In *Proceedings of the Student Research Workshop at the Annual Meeting of the Association for Computational Linguistics*.

Lizheng Li and Blair Armstrong. 2024. BERT encodes polysemy regularity structure. In *Proceedings of the 28th Conference on Computational Natural Language Learning (CoNLL)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Alizée Lombard, Anastasia Ulicheva, Maria Korochkina, and Kathleen Rastle. 2024. The regularity of polysemy patterns in the mind: Computational and experimental data. *Glossa Psycholinguistics*, 3(1):1–24.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2020. The Lancaster sensorimotor norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, 52(3):1271–1291.

Meta AI. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

George A Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Marco A Petilli and Marco Marelli. 2025. Indirect experiential grounding of abstract concepts. In *Proceedings of the Annual Conference of the Cognitive Science Society*.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of ICML*, pages 8748–8763.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Michaela Regneri and Michael Fritz. 2025. A caution on word embeddings and feature norms. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Eve Sweetser. 1990. *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge University Press.

Ekaterina Temerko, Lizheng Li, and Blair Armstrong. 2025. Polysemy regularity and LLM surprisal. In *Proceedings of the 29th Conference on Computational Natural Language Learning (CoNLL)*.

Andrea Tyler and Vyvyan Evans. 2003. *The Semantics of English Prepositions: Spatial Scenes, Embodied Meaning, and Cognition*. Cambridge University Press.

Akira Utsumi. 2020. Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science*, 44(6):e12844.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138.

Weizhi Xu, Yonatan Bisk, and Allyson Ettinger. 2025. Do language models learn sensorimotor semantics? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

A Framework Schematic

Figure 6 illustrates the theoretical framework linking experiential shift magnitude to polysemy regularity and brain alignment.

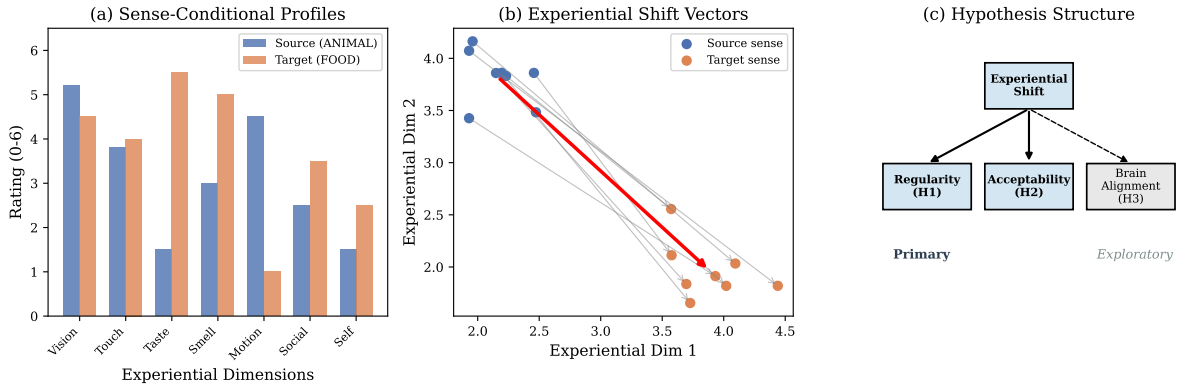

Figure 6: Theoretical framework. (a) Source and target sense profiles in EXP48 space with shift vector. (b) Regularity prediction: shift magnitude vs. R4. (c) Cross-model prediction: experiential alignment vs. polysemy sensitivity and brain alignment.

B Brain Alignment by Polysemy Status

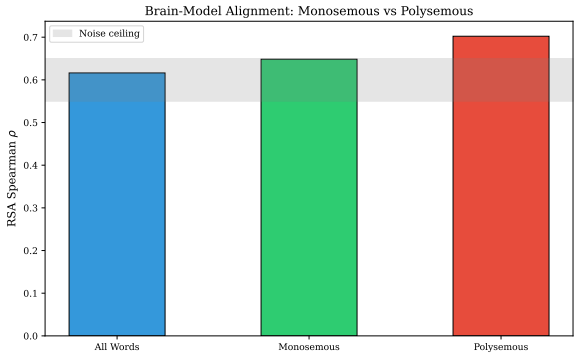

Figure 7: RSA brain alignment for all words (0.616), monosemous words (0.648), and polysemous words (0.702). Noise ceiling shown in gray.

C Dimension-Specific Shift Heatmap

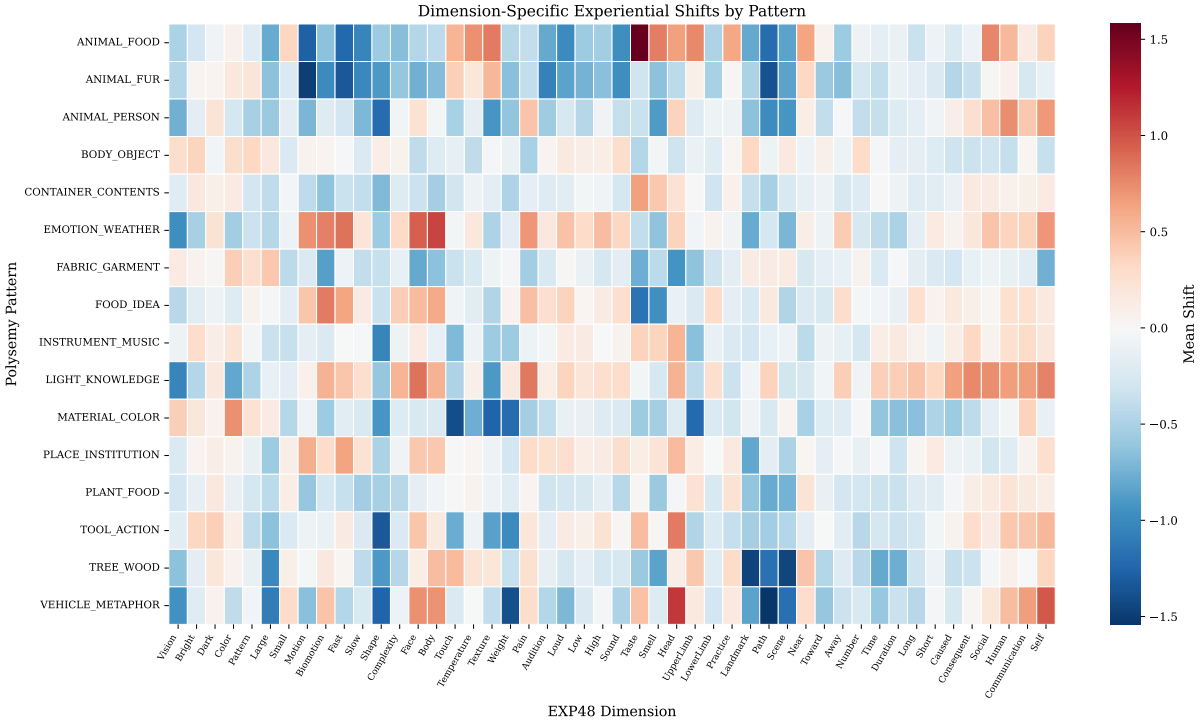

Figure 8: Dimension-specific experiential shifts across all 16 polysemy patterns and 48 EXP48 dimensions. Red indicates positive shifts (target $>$ source), blue indicates negative shifts.

D Layer Ablation

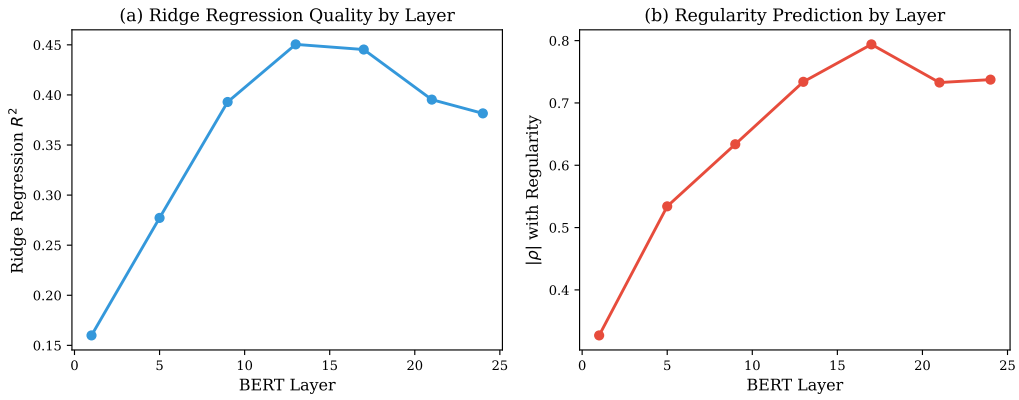

Figure 9: Layer ablation for BERT-large: ridge regression R^2 and regularity correlation (ρ) across layers 1–24.

E Sensorimotor Decomposition

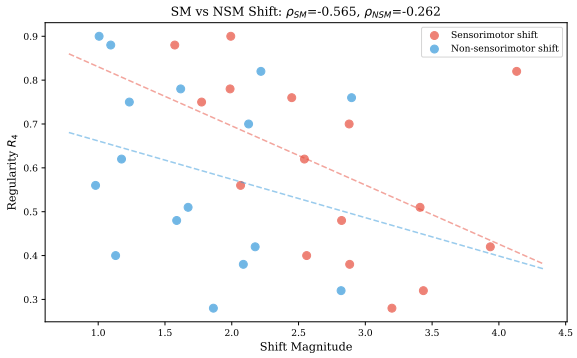

Figure 10: Sensorimotor (M_{SM}) vs. non-sensorimotor (M_{NSM}) shift components across patterns.