# Where meaning lives: Layer-wise accessibility of psycholinguistic features in encoder and decoder language models

**Anonymous ACL submission**

## Abstract

Understanding where transformer language models encode psychologically meaningful aspects of meaning is essential for both theory and practice. We conduct a systematic layer-wise probing study of 58 psycholinguistic features across 10 transformer models, spanning encoder-only and decoder-only architectures, and compare three embedding extraction methods. We find that apparent localization of meaning is strongly method-dependent: contextualized embeddings yield higher feature-specific selectivity and different layer-wise profiles than isolated embeddings. Across models and methods, final-layer representations are rarely optimal for recovering psycholinguistic information with linear probes. Despite these differences, models exhibit a shared depth ordering of meaning dimensions, with lexical properties peaking earlier and experiential and affective dimensions peaking later. Together, these results show that where meaning "lives" in transformer models reflects an interaction between methodological choices and architectural constraints.

## 1 Introduction

How do language models represent meaningful dimensions of language, such as emotion, concreteness, or sensory experience? As transformer-based models become increasingly integrated into real-world systems, understanding the internal dimensions they encode is essential. This matters both theoretically, for assessing whether language models reflect human semantic representations, and for practice, where interpretability and transparency are critical for safe deployment in meaning-sensitive domains such as education, mental health, and human–AI interaction.

Transformer-based models produce contextual token embeddings at each layer, which have been widely used to study the distribution of linguistic information within the models. Prior probing work has reported systematic differences across model layers, suggesting that surface, syntactic, and semantic features may be preferentially represented at different depths (Jawahar et al., 2019; Tenney et al., 2019; Hewitt and Manning, 2019; Lin et al., 2019). These findings have influenced both interpretive claims about transformer representations and common modeling practices, including the frequent use of final-layer embeddings for semantic tasks.

Recent work has begun to challenge the default reliance on final-layer representations by showing that internal layers are often more informative for downstream tasks and that certain meaning dimensions—most notably emotion—are most decodable in middle layers (Skean et al., 2025; Zhang and Zhong, 2025). At the same time, layer-wise localization of such features appears to depend strongly on model architecture (Liu et al., 2024). Because existing evidence largely comes from single models or narrow sets of dimensions, it remains unclear whether reported patterns reflect general principles or model- and dimension-specific effects. This uncertainty is particularly consequential given growing interest in whether language models encode psychologically meaningful dimensions of language (Waldis et al., 2024; Zhu et al., 2024; Xu et al., 2025). Meaning is inherently multi-faceted, encompassing not only affective properties but also perceptual, cognitive, and social aspects. A systematic, cross-architectural investigation spanning a broader range of meaning dimensions is therefore needed to determine how meaning is distributed and transformed across layers.

Interpreting hidden representations also raises unresolved methodological challenges. Because language models encode word meaning in context, the choice of context used for embedding extraction is itself a critical design decision. Prior studies vary widely: some embed target words in fixed templates (e.g., "What is the meaning of WORD?") (Liétard et al., 2021; Petroni et al., 2019), while oth-

| Encoder Model | Params | $L$ | $d_{\text{model}}$ |
|---|---|---|---|
| BERT Large | 336M | 24 | 1024 |
| RoBERTa Large | 355M | 24 | 1024 |
| DeBERTa-v3 Large | 304M | 24 | 1024 |
| BGE-M3 | 567M | 24 | 1024 |
| Jina-v3 | 570M | 24 | 1024 |

| Decoder Model | Params | $L$ | $d_{\text{model}}$ |
|---|---|---|---|
| Mistral-24B | 24B | 40 | 5120 |
| Phi-4 | 14B | 40 | 5120 |
| GPT-OSS-20B | 20B | 24 | 2880 |
| Gemma-3-27B | 27B | 62 | 5376 |
| Qwen3-32B | 32B | 64 | 5120 |

Table 1: **Model Specifications.** Overview of the ten transformer models evaluated, separated by architectural class. $L$ denotes the number of layers, and $d$ denotes the hidden dimension size.

ers rely on naturally occurring sentences, from single instances (Chang and Chen, 2019) or averaged across contexts (Bommasani et al., 2020). Most work adopts only one strategy. If apparent layer-wise localization depends on extraction method, conclusions about how meaning is distributed in language models may be method-dependent rather than intrinsic to the models.

These considerations motivate a systematic analysis that jointly considers a broad range of meaning dimensions, model architectures, and embedding extraction methods. Our study analyzes how meaning dimensions in the form of 58 psycholinguistic features (Hussain et al., 2024) are encoded across layers in 10 transformer models spanning two architectural classes. To assess methodological stability, we apply three embedding extraction methods and use linear probes to measure the amount of feature-selective information recoverable at each layer.

We make three contributions to the study of meaning representation in language models. First, we show that embedding extraction methods and model architecture affect both recoverable information and layer-wise localization, indicating that conclusions based on a single method may be unstable. Second, we show that final-layer embeddings are rarely optimal for recovering psycholinguistic meaning with linear probes across architectures and extraction methods, implying that defaulting to final-layer representations can miss information that is most accessible in intermediate layers. Finally, we show that models exhibit a depth ordering of meaning dimensions that is shared within and across model architectures, with depth rising from lexical to semantic features.

## 2 Methodology

### 2.1 Word features

We relied on the psychNorms metabase (Hussain et al., 2024), a large-scale aggregation of psycholinguistic features derived from dozens of independent behavioral studies. The database includes human-validated word-level features spanning affective (e.g., valence, arousal), semantic (e.g., concreteness, imageability, semantic diversity), developmental (e.g., age of acquisition), sensory–motor, lexical (e.g., frequency), and behavioral performance measures (e.g., accuracy and response times in lexical decision tasks; see Table 2). With the exception of frequency and semantic diversity, which are corpus-derived, these features are based on Likert-style ratings or task behavior, providing a principled, behaviorally grounded target space for probing representations in language models.

Coverage in psychNorms varies substantially across features, ranging from 703 to 79,671 words. To ensure comparability across features while controlling computational cost, we constructed a subset of 9,966 words greedily maximizing overlap among the 58 highest-coverage (N>4600) features. This procedure minimizes confounds arising from differences in word sets and sample sizes across features while retaining broad lexical coverage. The word set has a median rank of 9,201 (IQR = [3,858, 16,260]) out of the total length of 57,214 of the SUBTLEXUS word frequency dictionary and contains 52% nouns, 22% verbs, 19% adjectives, and 7% other parts of speech.

### 2.2 Transformer models

We examined ten openly available transformer models accessed via the HuggingFace platform, spanning both major architectural paradigms: encoder-only and decoder-only models (see Table 1). Encoder-based models included BERT Large (Devlin et al., 2019), RoBERTa Large (Liu et al., 2019), DeBERTa-v3 Large (He et al., 2023), BGE-M3 (Chen et al., 2024), and Jina-v3 (Sturua et al., 2024). Decoder-based models included Mistral-24B (Mistral AI Team, 2025), Phi-4 (Abdin et al., 2024), Qwen3-32B (Yang et al., 2025), GPT-OSS-20B (OpenAI et al., 2025), and Gemma-3-27B (Team et al., 2025).

| Category | Description | N |
|---|---|---|
| Frequency | How often a word occurs in language, based on log-transformed frequency estimates from diverse spoken and written corpora. | 10 |
| Motor | Degree to which a word is associated with bodily actions or motor experiences. | 7 |
| Sensory | Strength of perceptual experience associated with a word across sensory modalities, including visual, auditory, tactile, olfactory, and gustatory experience. | 6 |
| Semantic Diversity | Variability of contexts in which a word appears, reflecting how semantically diverse or context-specific its usage is. | 6 |
| Visual Lexical Decision | Accuracy and response speed in tasks where participants judge whether a visually presented letter string is a real word. | 6 |
| Familiarity | How well a word is known to speakers, including when it is learned, how many individuals recognize or understand it, and how frequently it is encountered in language use. | 4 |
| Auditory Lexical Decision | Accuracy and response speed in tasks where participants judge whether a spoken stimulus corresponds to a real word. | 4 |
| Valence | Emotional polarity of a word, ranging from negative to positive affective meaning. | 2 |
| Arousal | The degree of emotional intensity or activation elicited by a word. | 2 |
| Dominance | Extent to which a word evokes feelings of control, power, or influence versus submission or passivity. | 2 |
| Naming | Speed and accuracy with which speakers produce a word's pronunciation when presented with its written or visual form. | 2 |
| Semantic Decision | Accuracy and response latency in tasks where participants judge whether a word refers to something concrete or abstract, based on semantic decision data from the Calgary database. | 2 |
| Age of Acquisition | Estimated age (in years) at which speakers report having learned a word, reflecting the timing of lexical acquisition. | 1 |
| Concreteness | Extent to which a word refers to tangible, perceptible entities as opposed to abstract concepts. | 1 |
| Semantic Neighborhood | Density or similarity of meanings surrounding a word in semantic space, indicating how closely related it is to other words. | 1 |
| Social / Moral | Extent to which a word conveys social, interpersonal, or moral meaning relevant to human interaction and norms. | 1 |
| Iconicity / Transparency | Degree to which a word's form resembles or transparently conveys its meaning. | 1 |

Table 2: **Feature Categories.** Overview of the feature categories used in the analysis. N - Number of datasets

This selection spans a wide range of parameter scales, pre- and post-training objectives, and architectural choices. By contrasting encoder and decoder architectures within a unified probing framework, we aim to distinguish architectural regularities in the localization of psycholinguistic features from model-specific idiosyncrasies.

### 2.3 Embedding extraction

A central methodological challenge in representational probing is determining the context in which a word embedding should be extracted. We therefore compare three embedding extraction methods used in the literature (Gurnee and Tegmark, 2024; Liétard et al., 2021; Bommasani et al., 2020; Chronis and Erk, 2020): two contextualized extraction methods, called *template* and *averaged*, that differ in the amount and type of contextual information provided, and a baseline extraction method, called *isolated*, providing no context. For each approach, embeddings were extracted at the input layer and at the output of each transformer block (post-norm).

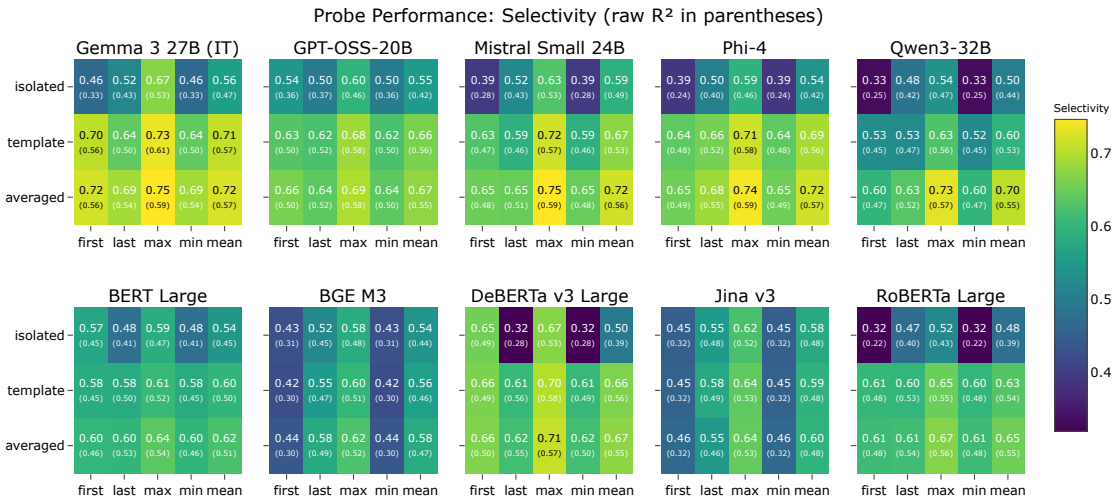


Figure 1: Linear probe performance (selectivity and raw $R^2$ in parenthesis) averaged over features across ten language models. Heatmaps show the predictive power of five embedding extraction methods (Y-axis) summarized over five layer metrics: average over first, last, best and worst performing layers and mean over all layers (X-axis), separated by model architecture (decoders: top; encoders: bottom).

For words consisting of multiple tokens, token-level embeddings were averaged.

Formally, let $h_\ell(w, c)$ denote the hidden state of word $w$ in context $c$ at layer $\ell$. We define the three extraction methods as:

$$\begin{aligned} e_\ell^{\text{iso}}(w) &= h_\ell(w, \emptyset). \\ e_\ell^{\text{temp}}(w) &= h_\ell(w, s_{\text{temp}}(w)) \\ e_\ell^{\text{avg}}(w) &= \frac{1}{50} \sum_{i=1}^{50} h_\ell(w, s_{wi}). \end{aligned} \tag{1}$$

The context $(s_{\text{temp}}(w))$ in the calculation of the template embedding $e_\ell^{\text{temp}}$ is the sentence "What is the meaning of the word [word]?", with [word] being replaced by $w$. The context $s_{wi}$ in the calculation of the aggregate embedding $e_\ell^{\text{avg}}$ is one of 50 sentences sampled at random from a representative subset of the C4 corpus (Raffel et al., 2020) containing the word $w$. We use 50 contexts as a compromise between stability and computational cost; pilot analyses showed diminishing returns beyond this point.

### 2.4 Locating meaning

To locate psycholinguistic information within transformer models, we applied layer-wise linear probing using ridge regression. We focus on linear probes to assess which information is directly accessible from model representations, rather than to maximize predictive accuracy, as more expressive probes can reconstruct information not explicitly encoded (Belinkov, 2022).

Importantly, high decoding performance alone does not guarantee that a representation selectively encodes the target feature. Linear probes may exploit correlations with unrelated lexical or distributional properties, or achieve above-chance performance even under label permutation (Ravichander et al., 2021; Hewitt and Liang, 2019).

To control for these confounds, we use selectivity as our primary outcome measure. Selectivity compares decoding performance on the true target to performance on a matched control task with permuted feature labels, isolating information specifically aligned with the feature of interest (Hewitt and Liang, 2019; Belinkov, 2022).

For each combination of psycholinguistic feature, model, layer, and embedding extraction method, we fit a separate ridge regression using embedding vectors as predictors and human feature values as targets. Models were trained using nested 5-fold cross-validation on a random subset of 4,000 words, with regularization strength $\alpha \in [1{,}000; 10{,}000]$. Performance was evaluated using out-of-sample $R^2_{obs}$. This procedure was repeated ten times, yielding 50 estimates per combination, which were averaged for analysis.

We repeated the same procedure under random

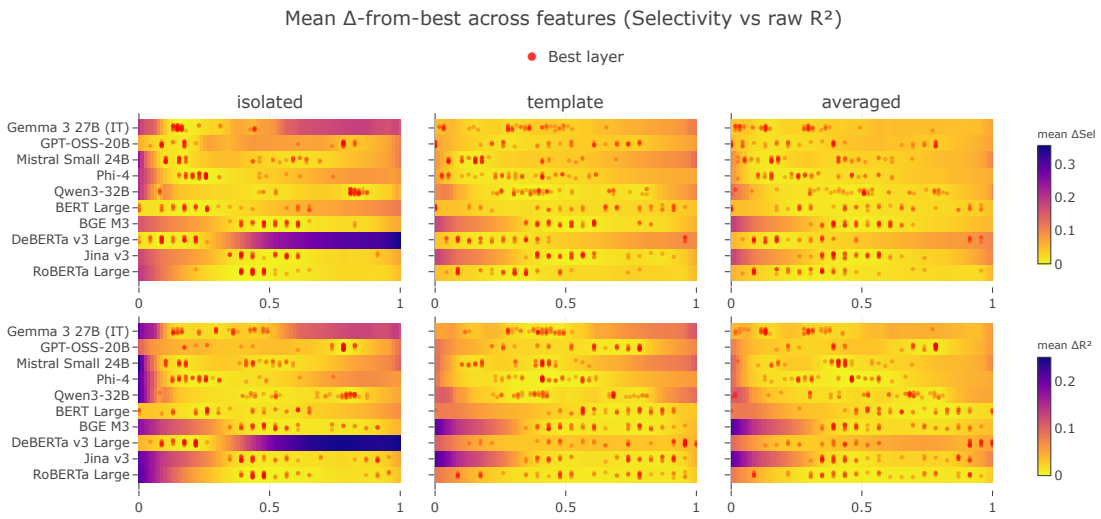


Figure 2: The heatmaps visualize the average performance drop ($\Delta$) relative to the best-performing layer for each model across three embedding extraction methods: isolated, template, and averaged. The top row displays results for selectivity, while the bottom row shows raw $R^2$ scores. Within each panel, decoders are grouped at the top and encoders at the bottom, with the x-axis representing the normalized layer index ($0 =$ first, $1 =$ last). Red dots mark the single best layer (argmax) for each individual model-feature pair.

permutation of the target feature values to estimate chance-level performance. Permutations were performed separately for each random subset of 4,000 words. Performance under permutation ($R^2_{rand}$) ranged between -0.58 and -0.01. Selectivity was then computed as

$$R^2_{sel} = R^2_{obs} - R^2_{rand}$$

Finally, to address localization, we calculate the center of mass as

$$COM = \frac{\sum_{\ell=1}^{L} \lambda(\ell) \Delta R^2_{sel,\ell}}{\sum_{\ell=1}^{L} \Delta R^2_{sel,\ell}}$$

with $\lambda(\ell) = \ell/L$ being the relative layer and $\Delta R^2_{sel,\ell}$ the selectivity relative to the lowest-selectivity layer. Additionally, we report the layers with maximum selectivity.

## 3 Results

We report results from layer-wise linear probing of 58 psycholinguistic features across ten transformer models, focusing on how embedding extraction method affects decodability and localization, architectural differences between encoder and decoder models, and the relative depth at which different psycholinguistic features are most accessible. Unless otherwise stated, results are based on selectivity.

### 3.1 Localization of psycholinguistic features is highly dependent on embedding extraction

Figure 1 shows that contextualized embeddings, whether template-based or context-averaged, consistently yield higher linear decodability than isolated embeddings. Across all models and features, moving to contextualized extraction increases median selectivity by 0.112 (0.106 raw $R^2$), with improvements observed for 100% of features. Among contextualized methods, averaged yielded consistently higher selectivity (and raw $R^2$) than template, suggesting a benefit for using richer and more diverse contexts in the extraction of psycholinguistic information.

The extraction method also influences the inferred layer-wise profiles: while isolated embeddings exhibit more pronounced variance and later peaks, contextualized embeddings produce flatter profiles, maintaining 80–90% of peak selectivity throughout large portions of the network.

Critically, final-layer representations are rarely optimal for recovering psycholinguistic information. Across all feature-model combinations, maximal selectivity is never achieved in the final layer; instead, optimal layers are distributed throughout the network depth. Together, these findings indicate that conclusions regarding where meaning is represented are contingent on the extraction strat-

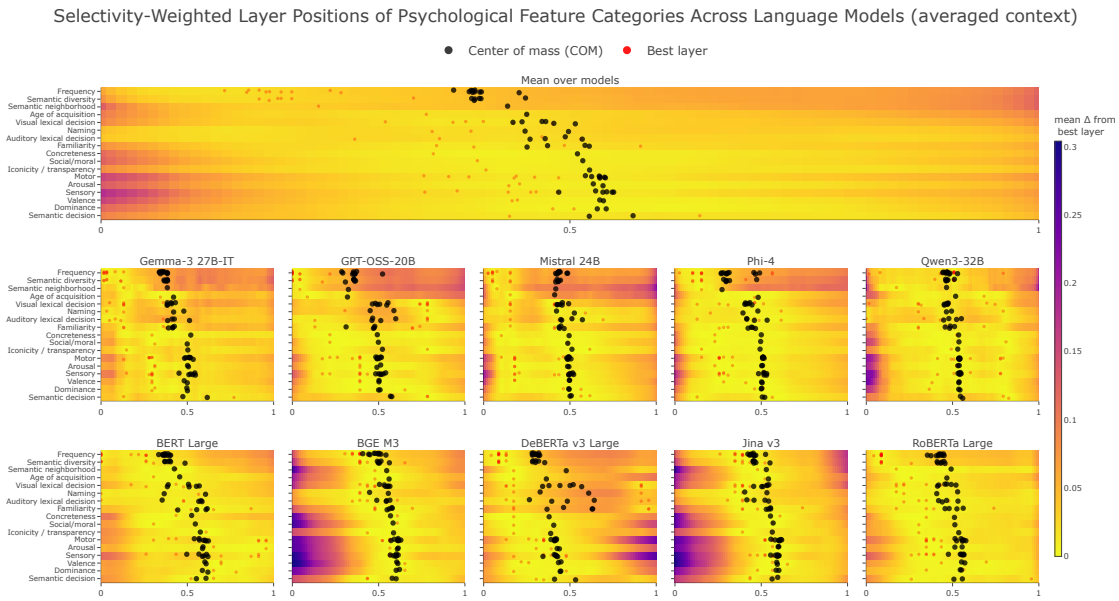


Figure 3: Selectivity-weighted layer positions of psycholinguistic feature categories for the averaged embedding extraction method. The heatmap depicts the mean Δ-from-best-layer across features comprising each category (X-axis: normalized layer index from first to last; Y-axis: psycholinguistic feature categories) based on selectivity score. Black dots indicate the selectivity-weighted center of mass (COM) of each feature's layer profile, while red dots mark the single best-performing layer (argmax) for each feature. The top panel corresponds to the mean over all models, and the two bottom panels correspond to language models (decoders, top row, encoders, bottom row).

egy. Differences induced by extraction method are comparable in magnitude to differences between models or architectures, indicating that layer-wise analyses that fix a single extraction strategy risk drawing unstable or misleading conclusions.

### 3.2 Models differ in how psycholinguistic features are distributed

Conditioning on extraction method, models vary in how meaning-related information is concentrated around optimal layers (Figure 2). Across both encoder and decoder families, models vary substantially in how tightly meaning-related information is concentrated around their best-performing layer. While RoBERTa-Large exhibits broad localization, other models like Qwen3-32B show sharp mid-layer peaks with steep performance drops toward both input and output. These differences are only partially aligned with architectural class: under contextualized extraction, several decoders match encoders in profile breadth, whereas some encoders exhibit pronounced mid-layer concentration. Thus, the degree to which psycholinguistic information is distributed across layers is more model- rather than architecture-dependent.

Figure 2 further illustrates the low selectivity of final-layer and even later-layer representations. This degradation is most pronounced in decoders: the last 20% of layers comprise only 0.86% of best-selectivity layers (0.52% raw $R^2$), with mean selectivity and raw performance drops of 0.07 and 0.051, respectively. In encoders, these layers account for 7.41% of best-selectivity layers (29.66% raw $R^2$), showing drops of 0.054 and 0.0267. These results underline that output-layer embeddings substantially underrepresent psycholinguistic information accessible in earlier layers.

### 3.3 Psycholinguistic features exhibit a stable depth ordering

We now shift focus to psycholinguistic features. Showing the results for averaged embeddings, Figure 3 reveals a consistent selectivity-weighted center of mass (COM) ordering across layers. See Figures 5-15 for the results of other extraction methods and raw $R^2$ for individual features and categories. Lexical and usage properties (e.g., frequency) peak early, while experiential, affective, and social dimensions (e.g., valence, concreteness) peak later, establishing a robust lexico-semantic access ordering.

This ordering is robust across models and ar-

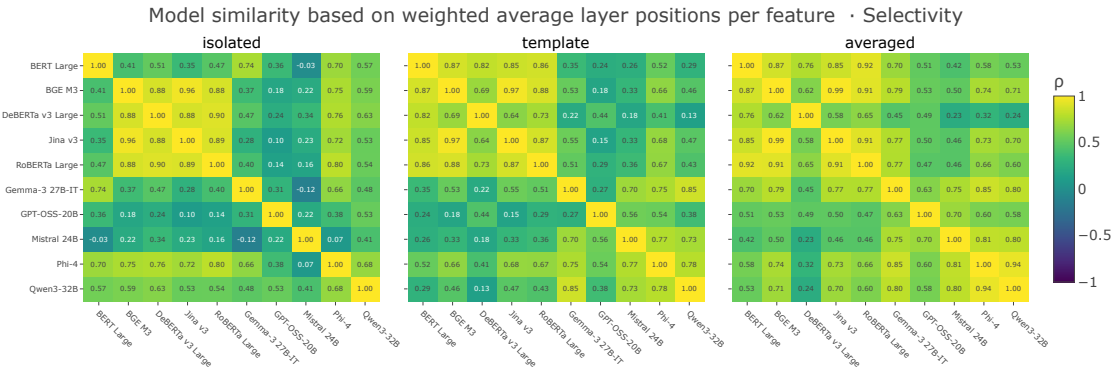


Figure 4: Each panel shows the pairwise Spearman correlation (p) between models, computed from vectors of feature-specific center-of-mass (COM) layer positions within a given embedding extraction method (isolated, template, averaged). For each model and feature, the COM was calculated (using selectivity scores), summarizing where in the network a feature is most strongly represented. Correlations are computed across feature orders, yielding a similarity matrix that reflects how similarly different models localize psycholinguistic information across layers.

chitectures. For averaged embeddings, pairwise Spearman correlations between models' feature-wise COM vectors strictly exceed $\rho.30$ in all comparisons and reach values above .70 within architectural classes (see Figure 4). Importantly, correlations are robust to excluding frequency and semantic diversity features, indicating that these correlations are not driven by objectively determined features (see Figures 17 and 18 in the Appendix).

Although individual features vary in absolute localization, violations of the overall ordering are rare and unsystematic. No category consistently peaks earlier than all lexical measures or later than all semantic measures, indicating that these results reflect a shared representational ordering rather than model-specific quirks.

### 3.4 Shared ordering, distinct realizations across architectures

Finally, we compare how this shared depth ordering is realized across architectures. Figure 4 shows similarity matrices based on feature-specific COM vectors under different extraction methods. For contextualized embeddings, encoder models form a tight cluster, as do decoder models, yielding a clear block structure. This indicates that while encoders and decoders largely agree on which dimensions emerge earlier versus later, they differ systematically in how this progression is distributed across layers.

This architectural separation is substantially weaker for isolated embeddings, where correlations are noisier and clustering is less pronounced, reinforcing the conclusion that contextualized extraction is necessary to recover stable representational structure. Together, these findings suggest that despite a common ordering, architectural design still plays an important role in how psycholinguistic features are represented.

## 4 Discussion

This study provides the most comprehensive investigation to date of how psychologically meaningful linguistic dimensions are represented across layers of transformer models. By systematically crossing 58 human-derived semantic features, 10 models spanning both major architectural families, three embedding extraction methods, and full layer-wise probing, we go beyond prior work, which is typically limited to fewer models, narrow feature sets, or fixed extraction strategies. Our findings reveal that conclusions about where psycholinguistic information is represented in language models depend jointly on methodological choices and architectural constraints, and uncover a lexico-semantic depth ordering that generalizes across contemporary transformer models.

### 4.1 Embedding extraction is a first-order methodological choice

Contextualized embeddings (template or averaged) yield substantially higher linear selectivity than isolated word embeddings across all models and feature categories. Moreover, isolated embeddings exhibit sharper peaks and stronger apparent localization; contextualized extraction reveals broader

accessibility profiles. These findings indicate that psycholinguistic dimensions rely on contextual processing to become linearly accessible to probes. Such results align with evidence that contextualized representations cannot be reduced to static embeddings without semantic loss (Ethayarajh, 2019; Bommasani et al., 2020) and with distributed accounts of semantic processing that emphasize context-dependent activation (Elman, 2004; Wulff et al., 2019).

Crucially, our findings extend these observations beyond a narrow set of semantic properties to a broad range of psychologically grounded features, including affective, sensory, motor, and social dimensions. At the same time, we emphasize that lower decodability from isolated embeddings does not imply the absence of such information, but rather reduced linear accessibility.

Notably, context-averaged embeddings yield systematically higher selectivity than template-based embeddings. However, the differences may not justify the increased computational cost. Averaging across 50 naturally occurring contexts requires extensive corpus retrieval and repeated inference, whereas a single, minimal template ("What is the meaning of the word [X]?") suffices to elicit similar levels of feature-specific information. For large-scale probing studies, template-based extraction may therefore offer a computationally efficient alternative to context averaging with minimal loss in linear recoverability.

### 4.2 Architecture shapes representational organization

Encoder and decoder models exhibit systematically different layer-wise accessibility profiles, though the distinction is often subtle. Encoder models tend toward slightly broader distributions , while decoders more frequently concentrate selectivity in intermediate layers with steeper declines toward the input and output.

This architectural contrast is most evident under contextualized extraction and is, however, secondary to model-specific variation. Since several decoders exhibit profiles comparable in breadth to encoders, and some encoders show pronounced mid-layer concentration, architecture constrains but does not uniquely determine representational spread.

These results may help reconcile apparently conflicting findings in the literature. Studies reporting graded, pipeline-like progressions from surface to semantic features have primarily examined encoder models (Jawahar et al., 2019; Tenney et al., 2019), whereas more recent work identifying mid-layer semantic peaks has focused on decoder-only architectures (Liu et al., 2024; Skean et al., 2025). Our results suggest that both patterns are valid within their respective architectural contexts, rather than mutually contradictory.

### 4.3 Final layers underrepresent psycholinguistic information

Across all models and embedding extraction methods, final-layer representations are rarely optimal for recovering psycholinguistic features via linear probes. These findings challenge the widespread practice of defaulting to final-layer embeddings for semantic analysis. Final layers are optimized for masked- or next-token prediction and downstream task objectives, and their representations may therefore transform information in ways that reduce linear accessibility. Our results extend prior observations regarding final-layer anisotropy and reduced interpretability (Ethayarajh, 2019; Skean et al., 2025) to a broad set of psycholinguistic features and across both encoder and decoder architectures.

Importantly, reduced linear decodability should not be interpreted as evidence that psycholinguistic information is lost or absent in final layers. Rather, it indicates that such information is less directly accessible to simple linear readouts, reinforcing the value of layer-wise analyses when probing model representations.

### 4.4 A shared ordering of psycholinguistic accessibility

Despite differences in how information is distributed across layers, models exhibit a remarkably consistent relative lexico-semantic ordering of psycholinguistic features with respect to layer depth, robust across models, with strong rank correlations of feature-specific center-of-mass vectors within models of the same architecture. This consistency indicates that transformer models broadly agree on which types of psycholinguistic information become more accessible earlier versus later in processing, even though they differ in how this progression is distributed across layers. We emphasize that shared ordering does not imply shared representational geometry: models realize this progression in systematically different ways.

## 5 Limitations

**Language coverage.** All features and probing experiments are restricted to English. Psycholinguistic features such as valence, concreteness, or social meaning may be realized differently across languages due to cultural, lexical, and morphological variation. The architectural patterns observed here may therefore not generalize to multilingual settings. Extending our approach to multilingual features and representations is an important direction for future work.

**Linear probing as an access measure.** We rely exclusively on linear probes to assess which psycholinguistic features are directly accessible from model representations. This choice follows established recommendations to avoid overinterpreting expressive probes (Hewitt and Liang, 2019; Belinkov, 2022), but it necessarily limits our conclusions. Some dimensions may be encoded in nonlinear or highly distributed forms that linear probes cannot recover. Differences in decodability should therefore be interpreted as differences in accessibility, not as the presence or absence of information.

**Representations versus behavior.** Our analysis focuses on internal representations rather than model behavior. Higher decodability of a psycholinguistic feature does not guarantee that a model will reliably use or express that information during generation or downstream tasks. Bridging representational analyses with controlled behavioral interventions remains an open challenge for future work.

**Correlated feature and construct redundancy.** Many psycholinguistic features are correlated (e.g., frequency, age of acquisition, familiarity, and lexical decision measures), reflecting shared behavioral and distributional factors. While selectivity mitigates generic predictability effects, our analyses do not fully disentangle overlapping constructs. Accordingly, the reported layer-wise patterns should be interpreted as reflecting relative accessibility of correlated meaning dimensions rather than sharply separable psycholinguistic modules, even if the demonstrated selectivity ordering implies meaningful differences in feature categories.

**Training regime and instruction tuning.** The models in our study differ not only in architecture but also in post-training objectives. We cannot cleanly separate architectural effects from training influences. Future work systematically varying training objectives will be needed to disentangle these factors.

Together, these limitations delineate the scope of our claims. Our results characterize where psycholinguistic information is linearly accessible within model representations under controlled probing conditions, providing a foundation for future work linking internal organization to multilingual generalization and observable behavior.

## References


Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

Ting-Yun Chang and Yun-Nung Chen. 2019. What does this word mean? explaining contextualized embeddings with natural language definition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6064–6070, Hong Kong, China. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Gabriella Chronis and Katrin Erk. 2020. When is a bishop not like a rook? when it's like a rabbi! multi-prototype BERT embeddings for estimating semantic relationships. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 227–244, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

*Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jeffrey L Elman. 2004. An alternative view of the mental lexicon. *Trends in cognitive sciences*, 8(7):301–306.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. *Preprint*, arXiv:2310.02207.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Zak Hussain, Rui Mata, Ben R. Newell, and Dirk U. Wulff. 2024. Probing the contents of semantic representations from text, behavior, and brain data using the psychnorms metabase. *arXiv preprint arXiv:2412.04936*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Bastien Liétard, Mostafa Abdou, and Anders Søgaard. 2021. Do language models know the way to Rome? In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 510–517, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

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. *Preprint*, arXiv:1907.11692.

Zhu Liu, Cunliang Kong, Ying Liu, and Maosong Sun. 2024. Fantastic semantics and where to find them: Investigating which layers of generative llms reflect lexical semantics. *arXiv preprint arXiv:2403.01509*.

Mistral AI Team. 2025. Mistral small 3. `https://mistral.ai/news/mistral-small-3`. Accessed: 2025-12-05.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b and gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. 2025. Layer by layer: Uncovering hidden representations in language models. *Preprint*, arXiv:2502.02013.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *Preprint*, arXiv:2409.10173.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. 2024. Holmes: A benchmark to assess the linguistic competence of language models. *Transactions of the Association for Computational Linguistics*, 12:1616–1647.

Dirk U Wulff, Simon De Deyne, Michael N Jones, and Rui Mata. 2019. New perspectives on the aging lexicon. *Trends in cognitive sciences*, 23(8):686–698.

Ningyu Xu, Qi Zhang, Chao Du, Qiang Luo, Xipeng Qiu, Xuanjing Huang, and Menghan Zhang. 2025. Revealing emergent human-like conceptual representations from language prediction. *Proceedings of the National Academy of Sciences*, 122(44):e2512514122.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Jingxiang Zhang and Lujia Zhong. 2025. Decoding emotion in the deep: A systematic study of how llms represent, retain, and express emotion. *Preprint*, arXiv:2510.04064.

Wentao Zhu, Zhining Zhang, and Yizhou Wang. 2024. Language models represent beliefs of self and others. *Preprint*, arXiv:2402.18496.

## A Appendix

This appendix provides supporting resources and full results that complement the main text. Definitions of all psycholinguistic features and their category assignments are provided in Table 3 (based on the psychNorms metabase (Hussain et al., 2024)). We further report additional visualization sets of layer-localization patterns across all models and embedding extraction contexts that are not shown in the main paper.

Figures 5–10 present feature-wise heatmaps of the $\Delta$-from-best-layer score over normalized layer position for each model and context. Black dots indicate the score-weighted center of mass (COM), and red dots mark the single best-performing layer.

Figures 11–15 show the corresponding category-level heatmaps, in which values are averaged across features within each category. While the main text focuses on a single embedding extraction method evaluated using the selectivity score for clarity, the appendix includes the full set of category-level visualizations across embedding extraction methods and for both selectivity and raw $R^2$ scores.

Figure 16 reports pairwise model similarity as a correlation matrix computed from feature-specific COM layer positions (raw $R^2$-weighted), summarizing the extent to which different models localize psycholinguistic information similarly across layers.

Finally, Figures 17–18 report the same pairwise model similarity analysis as in Figure 16, excluding the frequency and semantic diversity features (selectivity- and raw $R^2$-weighted, respectively).

Table 3: **Psycholinguistic features used in the analysis.**

| Feature | Description | Category |
|---|---|---|
| Freq_Blog | Log10 version of the frequency norms based on sources from blogs. | Frequency |
| Freq_CobS | Log10 of word frequencies in spoken English based on COBUILD corpus. | Frequency |
| Freq_CobW | Log10 of word frequencies in written English based on COBUILD corpus. | Frequency |
| Freq_HAL | Log10 version of frequency norms based on the Hyperspace Analogue to Language (HAL) corpus. | Frequency |
| Freq_KF | Log10 version of frequency norms based on the Kucera and Francis corpus. | Frequency |
| Freq_News | Log10 version of the frequency norms based on sources from newspapers. | Frequency |
| Freq_SUBTLEXUK | Log10 version of the frequency norms based on SUBTLEXuk corpus. | Frequency |
| Freq_SUBTLEXUSL | Log10 version of frequency norms based on the SUBTLEXus corpus. | Frequency |
| Freq_Twitter | Log10 version of the frequency norms based on sources from Twitter. | Frequency |
| Freq_TASA | How experience with a word is distributed over time based on the TASA corpus. It was computed by first taking logarithms of the frequencies and then transforming them to z-values for low (first three grades) and high grades (last three grades) respectively. | Frequency |
| BOI | The ease with which the human body can interact with a word's referent on a scale from 1 (low interaction) to 7 (high interaction). | Motor |
| Foot_Leg_Lanc | To what extent one experiences the referent by performing an action with the foot/leg, from 0 (not experienced at all) to 5 (experienced greatly). | Motor |
| Hand_Arm_Lanc | To what extent one experiences the referent by performing an action with the hand/arm, from 0 (not experienced at all) to 5 (experienced greatly). | Motor |
| Head_Lanc | To what extent one experiences the referent by performing an action with the head, from 0 (not experienced at all) to 5 (experienced greatly). | Motor |
| Interoceptive_Lanc | To what extent one experiences the referent by sensations inside one's body, from 0 (not experienced at all) to 5 (experienced greatly). | Motor |
| Mouth_Throat_Lanc | To what extent one experiences the referent by performing an action with the Mouth/throat, from 0 (not experienced at all) to 5 (experienced greatly). | Motor |
| Torso_Lanc | To what extent one experiences the referent by performing an action with the torso, from 0 (not experienced at all) to 5 (experienced greatly). | Motor |
| Auditory_Lanc | To what extent one experiences the referent by hearing, from 0 (not experienced at all) to 5 (experienced greatly). | Sensory |
| Gustatory_Lanc | To what extent one experiences the referent by tasting, from 0 (not experienced at all) to 5 (experienced greatly). | Sensory |

| Feature | Description | Category |
|---|---|---|
| Haptic_Lanc | To what extent one experiences the referent by feeling through touch, from 0 (not experienced at all) to 5 (experienced greatly). | Sensory |
| Olfactory_Lanc | To what extent one experiences the referent by smelling, from 0 (not experienced at all) to 5 (experienced greatly). | Sensory |
| Sensory_Experience | The extent to which a word evokes a sensory and/or perceptual experience in the mind of the reader on a 1 to 7 scale, with higher numbers indicating a greater sensory experience. | Sensory |
| Visual_Lanc | To what extent one experiences the referent by seeing, from 0 (not experienced at all) to 5 (experienced greatly). | Sensory |
| CD_Blog | Log10 version of the contextual diversity of a word, which refers to the number of passages (documents) in the sources from Blog containing a particular word. | Semantic diversity |
| CD_News | Log10 version of the contextual diversity of a word, which refers to the number of passages (documents) in the sources from newspapers containing a particular word. | Semantic diversity |
| CD_SUBTLEXUK | Log10 version of the contextual diversity of a word, which refers to the number of passages (documents) in the SUBTLEXuk corpus containing a particular word. | Semantic diversity |
| CD_SUBTLEXUS | Log10 version of the contextual diversity of a word, which refers to the number of passages (documents) in the SUBTLEXus corpus containing a particular word. | Semantic diversity |
| CD_Twitter | Log10 version of the contextual diversity of a word, which refers to the number of passages (documents) in the sources from Twitter containing a particular word. | Semantic diversity |
| Sem_Diversity | The degree to which different contexts associated with a word vary in their meanings. | Semantic diversity |
| LexicalD_ACC_V_BLP | The proportion of accurate responses of visual lexical decision for a particular word from the British Lexicon Project. | Visual lexical decision |
| LexicalD_ACC_V_ECPP | The proportion of accurate responses of visual lexical decision for a particular word from the English Crowdsourcing Project. | Visual lexical decision |
| LexicalD_ACC_V_ELP | The proportion of accurate responses of visual lexical decision for a particular word from the English Lexicon Project. | Visual lexical decision |
| LexicalD_RT_V_BLP | The mean visual lexical decision latency (in msec) for a particular word across participants from the British Lexicon Project. | Visual lexical decision |
| LexicalD_RT_V_ECP | The mean visual lexical decision latency (in msec) for a particular word in the word knowledge task across participants from the English Crowdsourcing Project. This task is similar, but not identical, to the traditional lexical decision task. Participants were asked to indicate whether each item "is a word you know or not." Their results showed that RTs in this task correlate well with those from lexical decision in ELP and BLP, and hence we have labelled it as such. | Visual lexical decision |
| LexicalD_RT_V_ELP | The mean visual lexical decision latency (in msec) for a particular word across participants from the English Lexicon Project. | Visual lexical decision |

| Feature | Description | Category |
|---|---|---|
| LexicalD_ACC_A_AELP | The proportion of accurate responses of auditory lexical decision for a particular word from the Auditory English Lexicon Project. | Auditory lexical decision |
| LexicalD_ACC_A_MALD | The proportion of accurate responses of auditory lexical decision for a particular word from the Massive Auditory Lexical Decision database. | Auditory lexical decision |
| LexicalD_RT_A_AELP | The mean auditory lexical decision latency (in msec) for a particular word from the Auditory English Lexicon Project. | Auditory lexical decision |
| LexicalD_RT_A_MALD | The mean auditory lexical decision latency (in msec) for a particular word from the Massive Auditory Lexical Decision database. | Auditory lexical decision |
| AoA_Kuper | The age at which people acquired the word, in which participants were asked to enter the age (in years) at which they thought they had learned the word. | Familiarity |
| Fam_Brys | Percentage of participants who know the word well enough to give answer for the concreteness rating. | Familiarity |
| Prevalence_Brys | The proportion of people who know the word, in which participants were asked to indicate whether or not they knew the stimulus in a list of words and nonwords, in an online crowdsourcing study. Percentages were translated to z values on the the basis of cumulative normal distribution. | Familiarity |
| perc_known_winter | Percentage of participants that did not know the meaning or pronunciation of the word. | Familiarity |
| Valence_NRC | Word-emotion association built by manual annotation using Best-Worst Scaling method, with scores ranging from 0 (negative) to 1 (positive). | Valence |
| Valence_Warr | The pleasantness of a stimulus on a 1 (happy) to 9 (unhappy) scale. | Valence |
| Arousal_NRC | Word-emotion association built by manual annotation using Best-Worst Scaling method, with scores ranging from 0 (low arousal) to 1 (high arousal). | Arousal |
| Arousal_Warr | The intensity of emotion provoked by a stimulus on a 1 (aroused) to 9 (calm) scale. | Arousal |
| Dominance_NRC | Word-emotion association built by manual annotation using Best-Worst Scaling method, with scores ranging from 0 (low dominance) to 1 (high dominance). | Dominance |
| Dominance_Warr | The degree of control exerted by a stimulus on a 1 (controlled) to 9 (in control) scale. | Dominance |
| Naming_ACC_ELP | The proportion of accurate responses of word naming for a particular word from the English Lexicon Project. | Naming |
| Naming_RT_ELP | The mean naming latency (in msec) for a particular word across participants from the English Lexicon Project. | Naming |
| SemanticD_ACC_Calgary | The proportion of accurate responses of concrete/abstract semantic decision (i.e., does the word refer to something concrete or abstract?) for a particular word from the Calgary database. | Semantic decision |
| SemanticD_RT_Calgary | The mean latency (in msec) of concrete/abstract semantic decision (i.e., does the word refer to something concrete or abstract?) for a particular word from the Calgary database. | Semantic decision |

| Feature | Description | Category |
|---|---|---|
| Sem_N_D | The average radius of co-occurrence, which is the average distance between the words in the semantic neighborhood and the target word. | Semantic neighborhood |
| AoA_LWV | The age at which people acquired the word, in which a three-choice test was administered to participants in grades 4 to 16 (college) (Living Word Vocabulary database). | Age of acquisition |
| Conc_Brys | The degree to which the concept can be experienced directly through the senses from a 1 (abstract) to 5 (concrete) scale. | Concreteness |
| Socialness | The extent to which a word's meaning has social relevance on a seven-point Likert scale from 1 to 7. | Social/moral |
| iconicity_winter_2023 | Iconicty ratings on a scale from 0 (not iconic at all) to 7 (very iconic). | Iconicity/transparency |

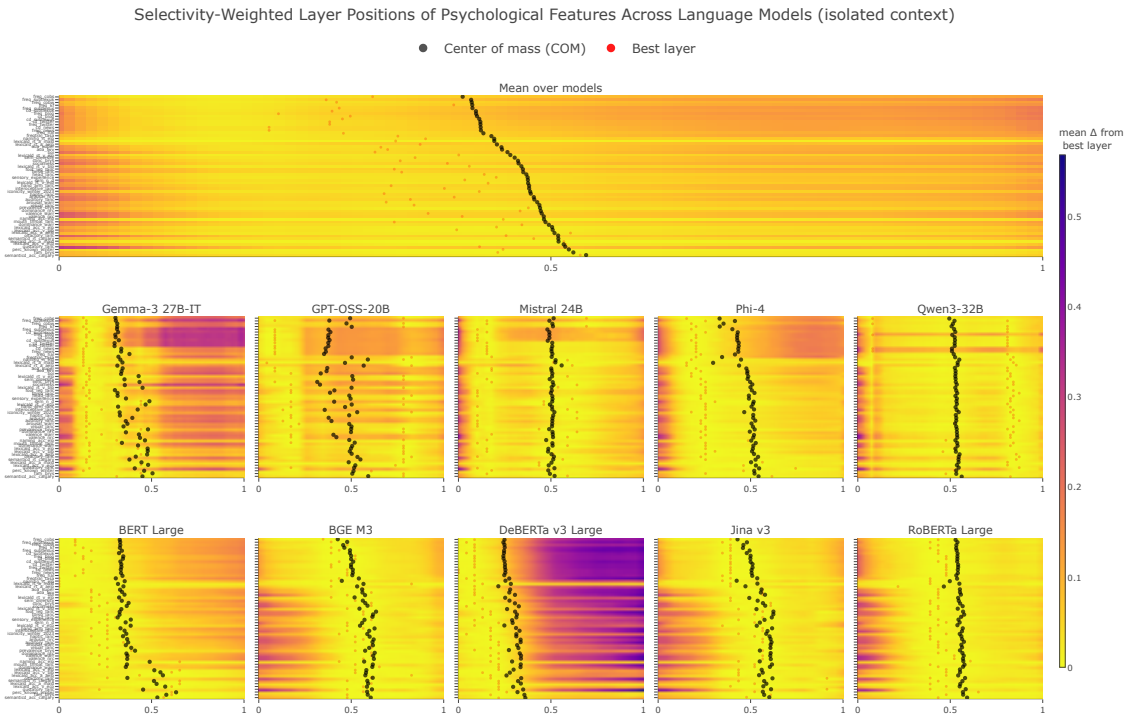


Figure 5: Selectivity-weighted layer positions of psycholinguistic features in the isolated context. Each panel shows a language model (decoders top row, encoders bottom row). The heatmap depicts the $\Delta$-from-best-layer across layers for each feature (x-axis: normalized layer index from first to last; y-axis: psycholinguistic features) based on Selectivity score. Black dots indicate the Selectivity-weighted center of mass (COM) of each feature's layer profile, while red dots mark the single best-performing layer (argmax). Lower (yellow) heatmap values indicate layers closer to the optimal representation of a feature.

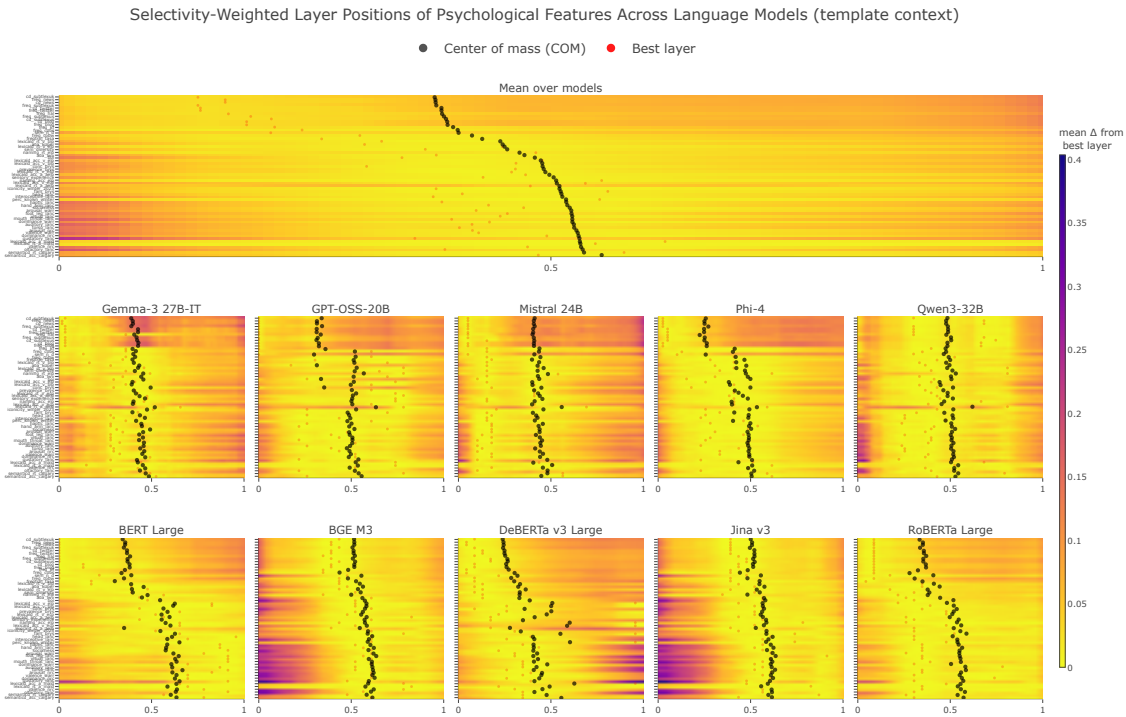


Figure 6: Selectivity-weighted layer positions of psycholinguistic features in the template context. Each panel shows a language model (decoders top row, encoders bottom row). The heatmap depicts the $\Delta$-from-best-layer across layers for each feature (x-axis: normalized layer index from first to last; y-axis: psycholinguistic features) based on Selectivity score. Black dots indicate the Selectivity-weighted center of mass (COM) of each feature's layer profile, while red dots mark the single best-performing layer (argmax). Lower (yellow) heatmap values indicate layers closer to the optimal representation of a feature.

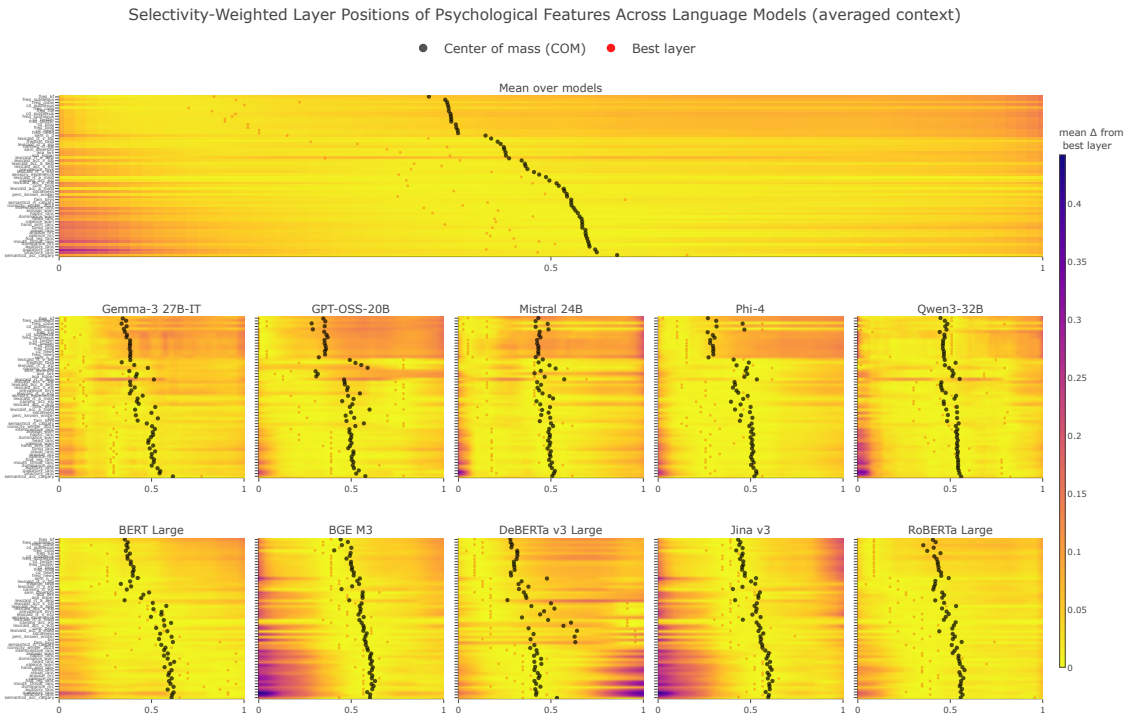


Figure 7: Selectivity-weighted layer positions of psycholinguistic features in the averaged context. Each panel shows a language model (decoders top row, encoders bottom row). The heatmap depicts the $\Delta$-from-best-layer across layers for each feature (x-axis: normalized layer index from first to last; y-axis: psycholinguistic features) based on Selectivity score. Black dots indicate the Selectivity-weighted center of mass (COM) of each feature's layer profile, while red dots mark the single best-performing layer (argmax). Lower (yellow) heatmap values indicate layers closer to the optimal representation of a feature.

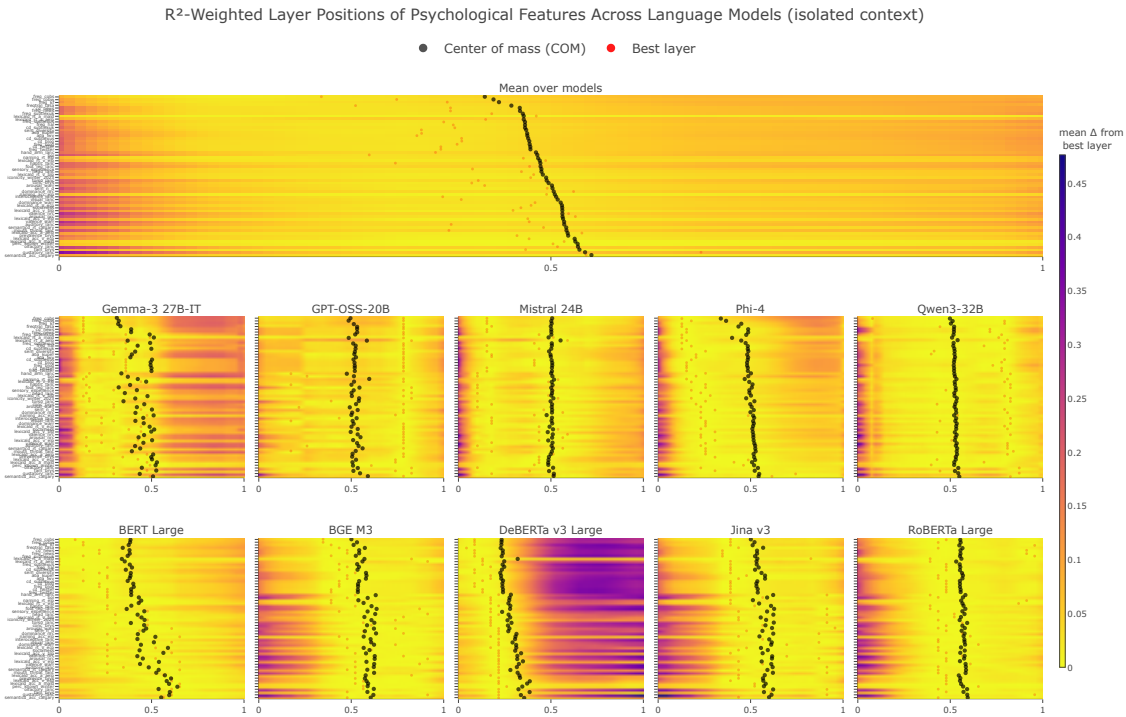


Figure 8: Raw $R^2$-weighted layer positions of psycholinguistic features in the isolated context. Each panel shows a language model (decoders top row, encoders bottom row). The heatmap depicts the $\Delta$-from-best-layer across layers for each feature (x-axis: normalized layer index from first to last; y-axis: psycholinguistic features) based on Raw $R^2$ score. Black dots indicate the Raw $R^2$-weighted center of mass (COM) of each feature's layer profile, while red dots mark the single best-performing layer (argmax). Lower (yellow) heatmap values indicate layers closer to the optimal representation of a feature.

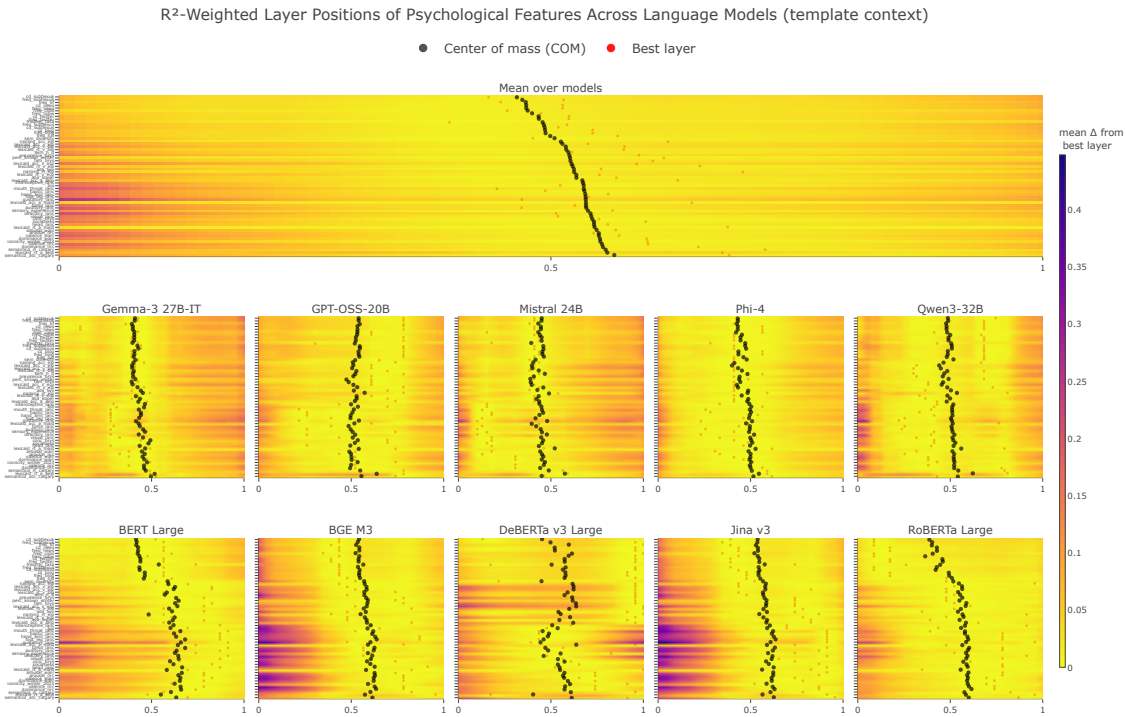


Figure 9: Raw $R^2$-weighted layer positions of psycholinguistic features in the template context. Each panel shows a language model (decoders top row, encoders bottom row). The heatmap depicts the $\Delta$-from-best-layer across layers for each feature (x-axis: normalized layer index from first to last; y-axis: psycholinguistic features) based on Raw $R^2$ score. Black dots indicate the Raw $R^2$-weighted center of mass (COM) of each feature's layer profile, while red dots mark the single best-performing layer (argmax). Lower (yellow) heatmap values indicate layers closer to the optimal representation of a feature.

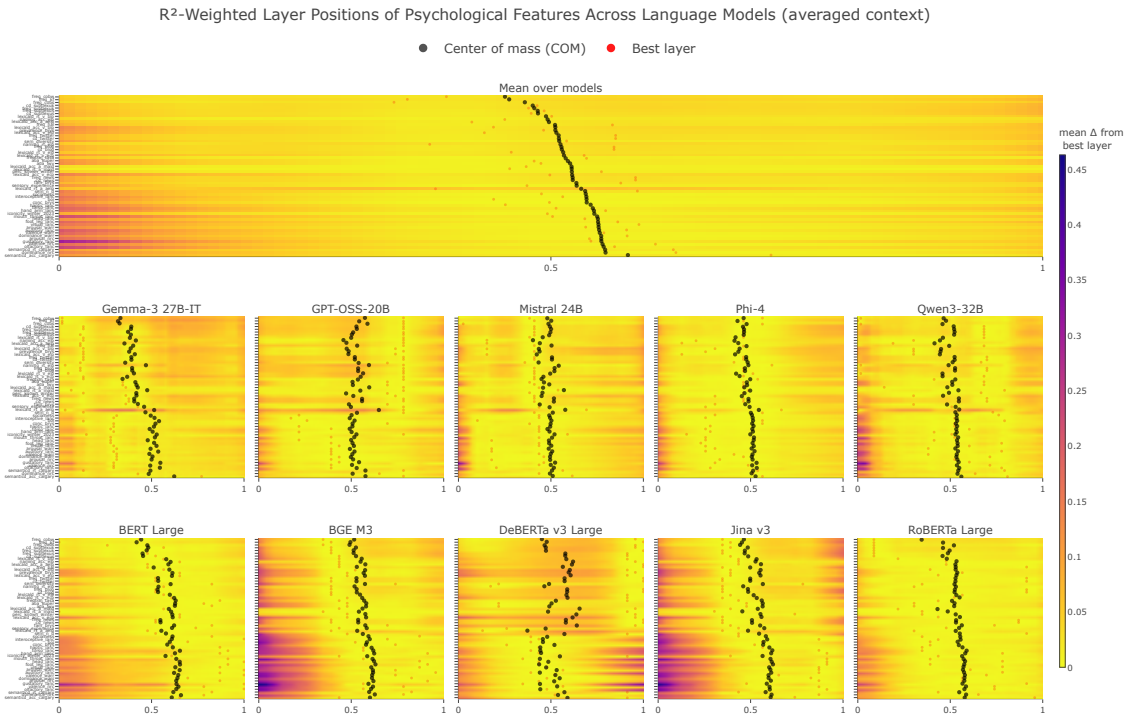


Figure 10: Raw $R^2$-weighted layer positions of psycholinguistic features in the averaged context. Each panel shows a language model (decoders top row, encoders bottom row). The heatmap depicts the $\Delta$-from-best-layer across layers for each feature (x-axis: normalized layer index from first to last; y-axis: psycholinguistic features) based on Raw $R^2$ score. Black dots indicate the Raw $R^2$-weighted center of mass (COM) of each feature's layer profile, while red dots mark the single best-performing layer (argmax). Lower (yellow) heatmap values indicate layers closer to the optimal representation of a feature.

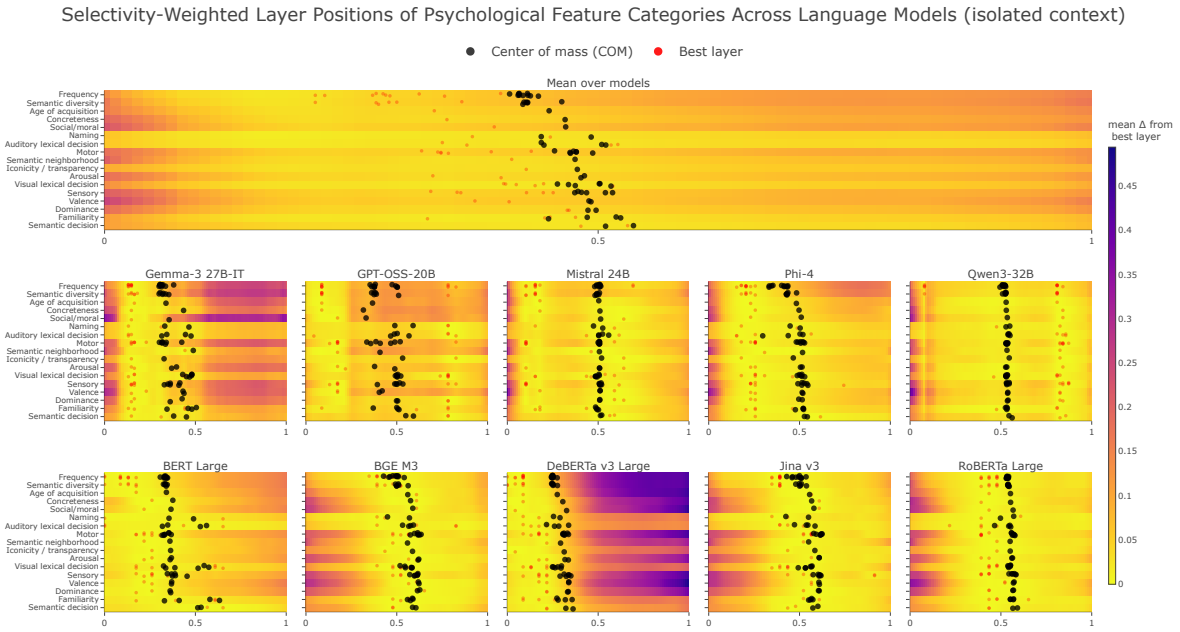


Figure 11: Selectivity-weighted layer positions of psycholinguistic feature categories in the isolated context. Each panel shows a language model (decoders top row, encoders bottom row). The heatmap depicts the mean Δ-from-best-layer across features comprising each category (x-axis: normalized layer index from first to last; y-axis: psycholinguistic feature categories) based on Selectivity score. Black dots indicate the Selectivity-weighted center of mass (COM) of each category's layer profile, while red dots mark the single best-performing layer (argmax). Lower (yellow) heatmap values indicate layers closer to the optimal representation of a category.

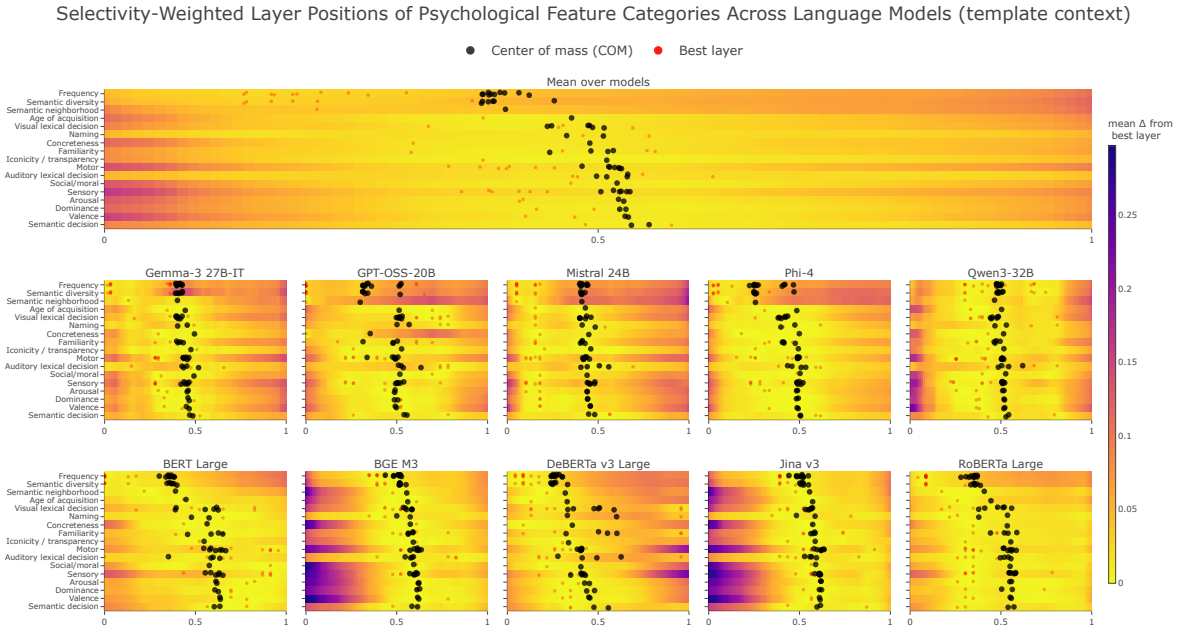


Figure 12: Selectivity-weighted layer positions of psycholinguistic feature categories in the template context. Each panel shows a language model (decoders top row, encoders bottom row). The heatmap depicts the mean Δ-from-best-layer across features comprising each category (x-axis: normalized layer index from first to last; y-axis: psycholinguistic feature categories) based on Selectivity score. Black dots indicate the Selectivity-weighted center of mass (COM) of each category's layer profile, while red dots mark the single best-performing layer (argmax). Lower (yellow) heatmap values indicate layers closer to the optimal representation of a category.

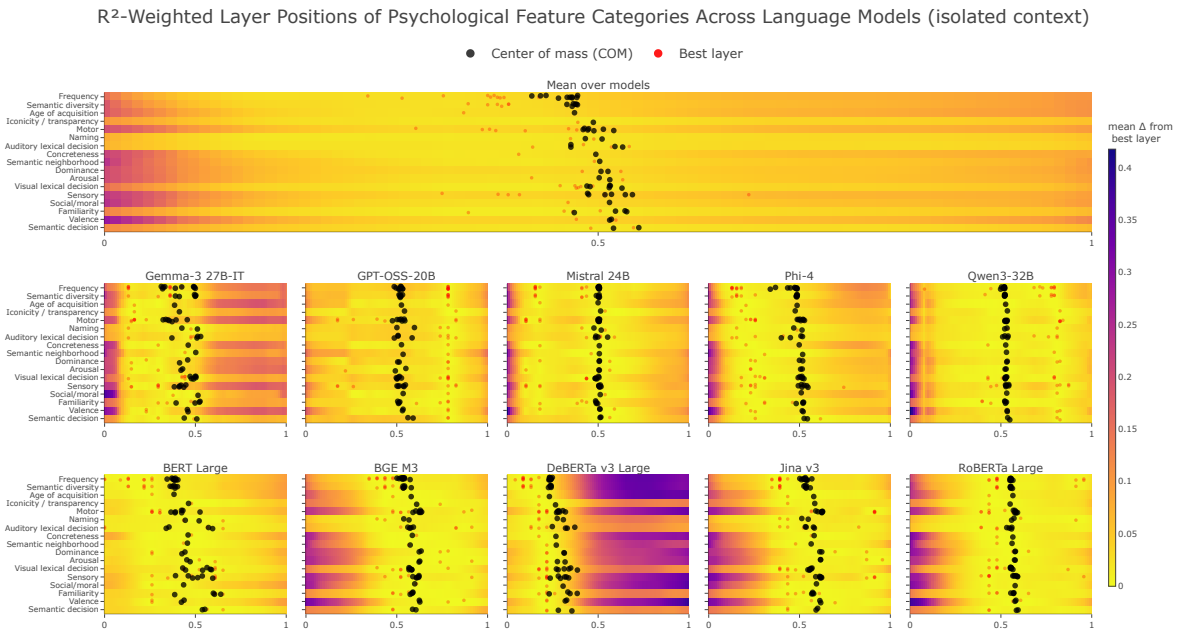


Figure 13: Raw $R^2$-weighted layer positions of psycholinguistic feature categories in the isolated context. Each panel shows a language model (decoders top row, encoders bottom row). The heatmap depicts the mean $\Delta$-from-best-layer across features comprising each category (x-axis: normalized layer index from first to last; y-axis: psycholinguistic feature categories) based on Raw $R^2$ score. Black dots indicate the Raw $R^2$-weighted center of mass (COM) of each category's layer profile, while red dots mark the single best-performing layer (argmax). Lower (yellow) heatmap values indicate layers closer to the optimal representation of a category.

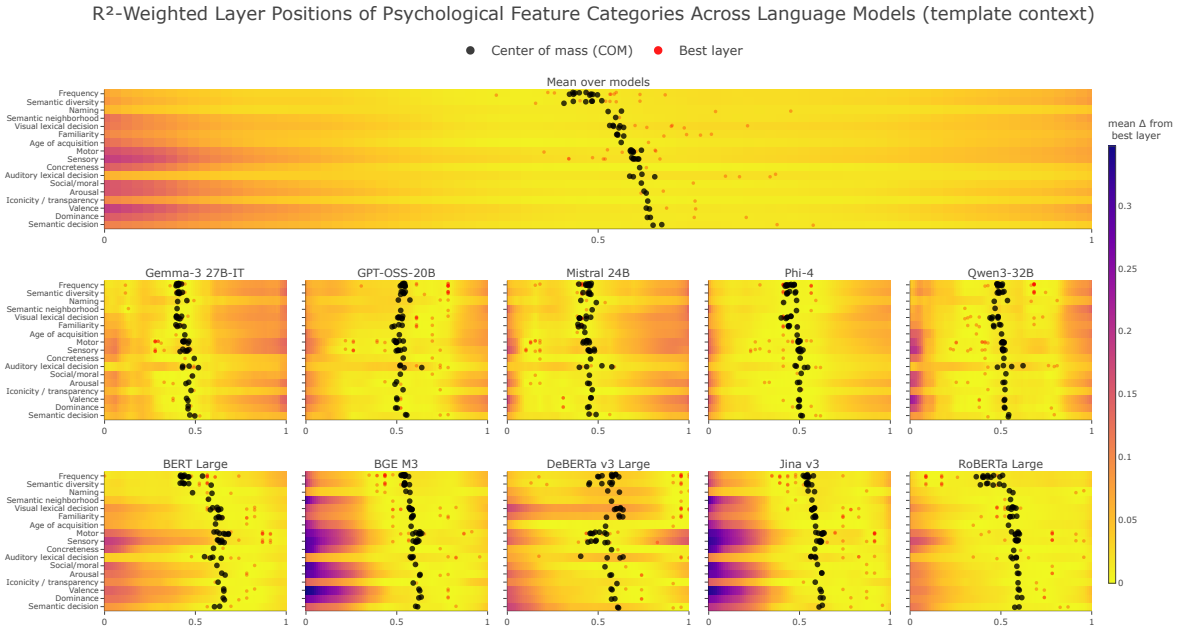


Figure 14: Raw $R^2$-weighted layer positions of psycholinguistic feature categories in the template context. Each panel shows a language model (decoders top row, encoders bottom row). The heatmap depicts the mean $\Delta$-from-best-layer across features comprising each category (x-axis: normalized layer index from first to last; y-axis: psycholinguistic feature categories) based on Raw $R^2$ score. Black dots indicate the Raw $R^2$-weighted center of mass (COM) of each category's layer profile, while red dots mark the single best-performing layer (argmax). Lower (yellow) heatmap values indicate layers closer to the optimal representation of a category.

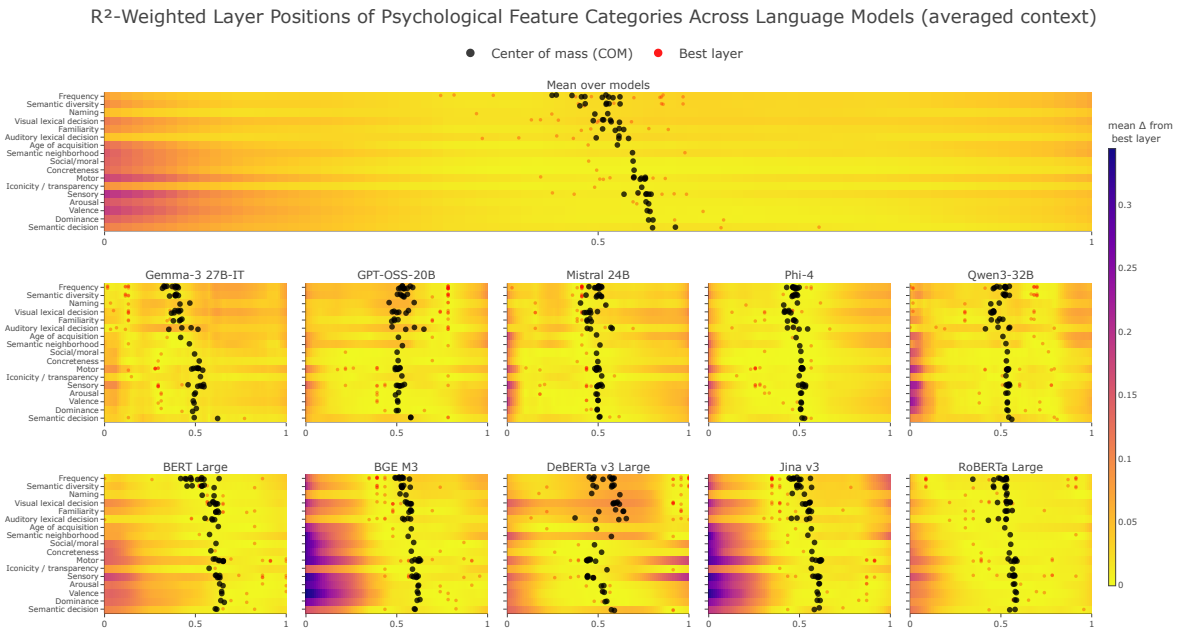


Figure 15: Raw $R^2$-weighted layer positions of psycholinguistic feature categories in the averaged context. Each panel shows a language model (decoders top row, encoders bottom row). The heatmap depicts the mean $\Delta$-from-best-layer across features comprising each category (x-axis: normalized layer index from first to last; y-axis: psycholinguistic feature categories) based on Raw $R^2$ score. Black dots indicate the Raw $R^2$-weighted center of mass (COM) of each category's layer profile, while red dots mark the single best-performing layer (argmax). Lower (yellow) heatmap values indicate layers closer to the optimal representation of a category.

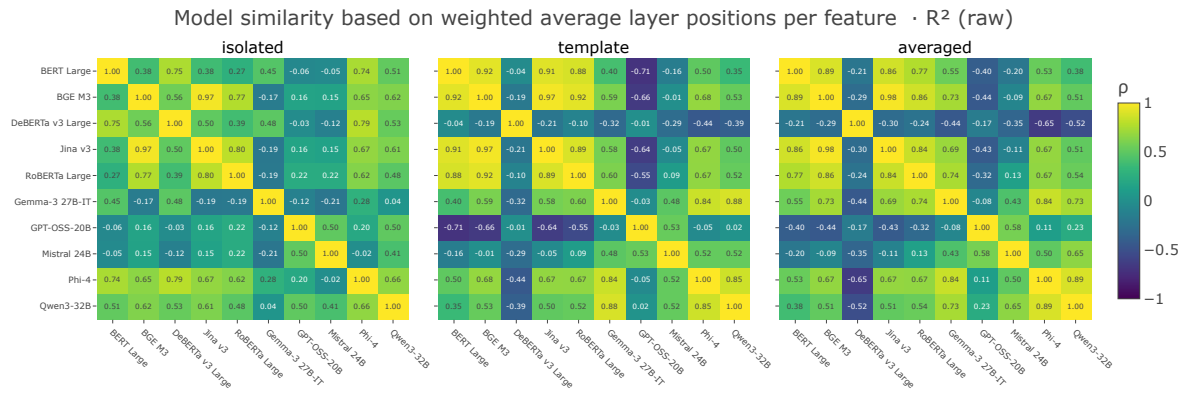


Figure 16: Each panel shows the pairwise Spearman correlation (ρ) between models, computed from vectors of feature-specific center-of-mass (COM) layer positions within a given context (isolated, template, averaged). For each model and feature, the COM is calculated as the score-weighted mean of normalized layer indices (using raw $R^2$ scores), summarizing where in the network a feature is most strongly represented. Correlations are computed across features, yielding a similarity matrix that reflects how similarly different models localize psycholinguistic information across layers.

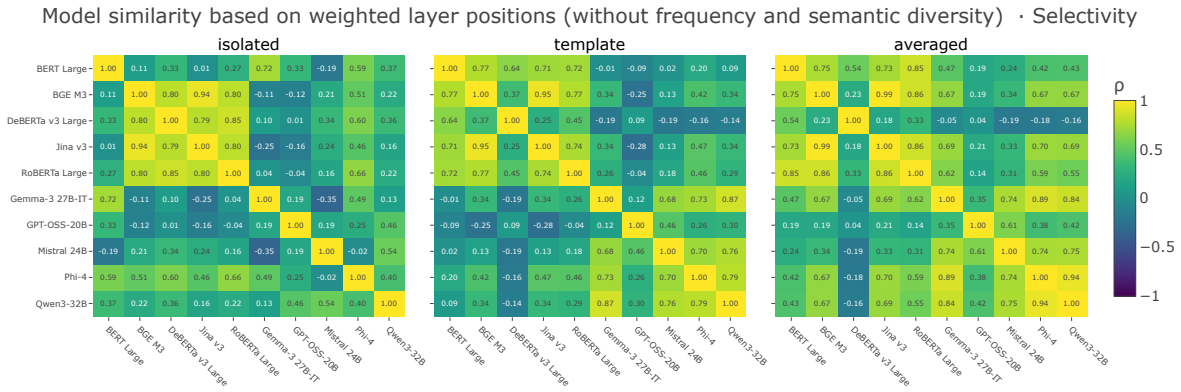


Figure 17: Each panel shows the pairwise Spearman correlation (p) between models, computed from vectors of feature-specific center-of-mass (COM) layer positions within a given context (isolated, template, averaged) excluding frequency and semantic diversity features. For each model and feature, the COM is calculated as the score-weighted mean of normalized layer indices (using selectivity scores), summarizing where in the network a feature is most strongly represented. Correlations are computed across features, yielding a similarity matrix that reflects how similarly different models localize psycholinguistic information across layers.

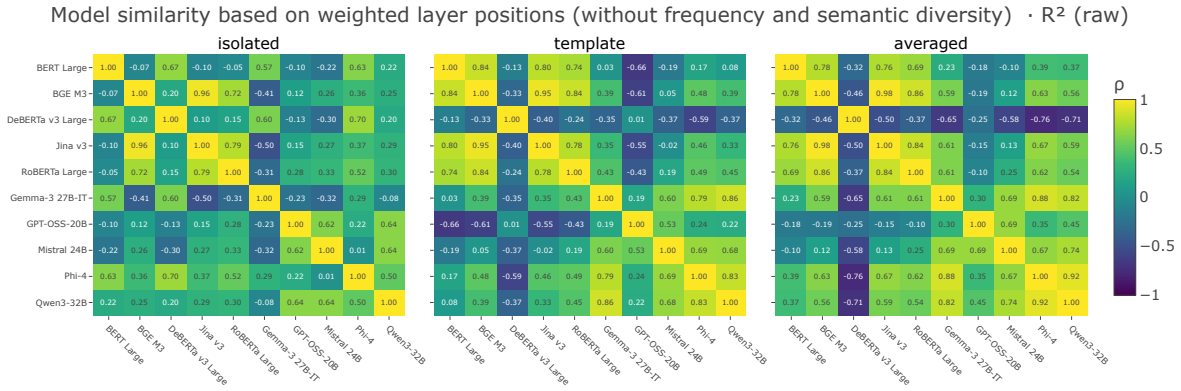


Figure 18: Each panel shows the pairwise Spearman correlation (p) between models, computed from vectors of feature-specific center-of-mass (COM) layer positions within a given context (isolated, template, averaged) excluding frequency and semantic diversity features. For each model and feature, the COM is calculated as the score-weighted mean of normalized layer indices (using raw $R^2$ scores), summarizing where in the network a feature is most strongly represented. Correlations are computed across features, yielding a similarity matrix that reflects how similarly different models localize psycholinguistic information across layers.