CONTEXTUAL LLM REPRESENTATIONS ENCODE FINE-GRAINED SEMANTIC TRANSPARENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Global distributional features and cross-context geometry jointly predict semantic transparency, the degree of semantic correspondence between a derived word and its morphological constituents. A stability hypothesis connects distributional accounts of morphological meaning to contextual representations, predicting that stronger constituent constraints accompany higher cross-context consensus and lower dispersion. A controlled resource combines 204 English derived forms spanning attested, novel and structurally degraded conditions, isolated-word transparency ratings, and 30 sentence contexts per item. A modelling hierarchy compares stem proximity, global distributional features and cross-context geometry across five pretrained encoder-only and decoder-only checkpoints. Global features improve on stem proximity and predict both structural contrasts and graded variation within conditions. At the preselected layers, adding contextual geometry raises within-condition random-forest R^2 from .12 to values between .15 and .26. Estimated increments are positive in all five checkpoints, with bootstrap intervals excluding zero for the two encoders. The additional signal is concentrated in morphologically intact forms and is strongest for novel, well-formed items in four of five checkpoints. The median aggregate increment is .26 under ridge regression and .04 under random forests, with each estimator fitted to both the global and combined feature sets. Aggregate and within-condition evaluation address complementary questions about prediction across the stimulus set and graded differences within conditions. These findings connect lexical support and contextual geometry to human semantic judgements, distinguish structural sensitivity from graded semantic accessibility, and show that predictive increments must be interpreted relative to the fitted comparison and target granularity.

1 INTRODUCTION

A semantic interpretation of a language-model representation should specify the distinctions it supports. For unfamiliar derived words, these include morphological well-formedness and graded differences in semantic correspondence between a word and its constituents. Recognising a legal stem–affix combination does not establish that the representation distinguishes transparency among equally well-formed words. A stability hypothesis motivates the test: if constituents constrain a form's interpretation, its representations should be more consistent across contexts. Work linking concept categories and hierarchies to representational geometry (Park et al., 2025) motivates a related behavioural question: which continuous semantic differences are recoverable within a controlled structural category? Distributional accounts explain morphological meaning through learned regularities linking forms, meanings and usage: an affix-specific transformation maps a stem vector to a derived-word vector, and proximity between the two is the established computational measure of transparency (Marelli & Baroni, 2015), supplemented by neighbourhood properties and affix productivity (Hay & Baayen, 2002; Bonandrini et al., 2023). Neurocognitive evidence that derivational processing varies with transparency (Bozic et al., 2013; Leminen et al., 2019) makes the property relevant to theories of morphological meaning, and continuous ratings give it a behavioural target.

Novel derived forms sharpen the test: equally well-formed but unattested combinations can differ in rated transparency. Global distributional features describe lexical support and compositional relations; repeated contextual encodings describe variation across sentences (Lenci et al., 2022). Both paths are tested separately and jointly against the same ratings (Figure 1).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

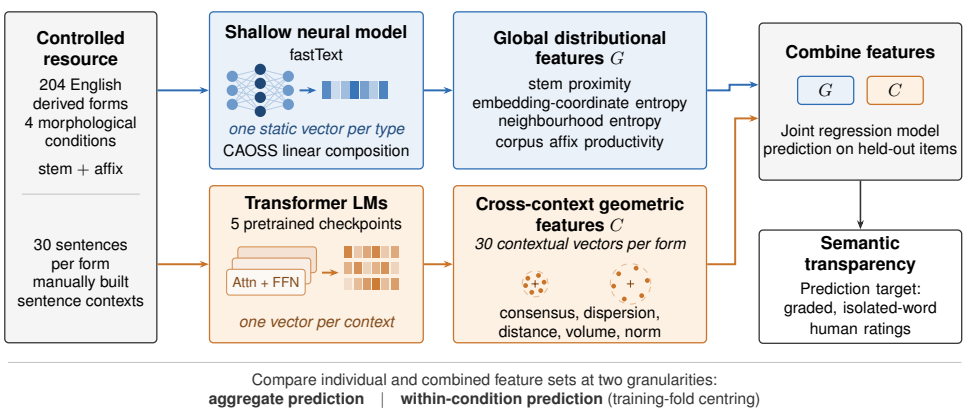

Figure 1: **Two representational paths test the same graded semantic target.** Global distributional features and cross-context geometric features enter individual and combined predictors of isolated-word transparency ratings. Aggregate evaluation uses full-set variation; within-condition evaluation targets differences beyond condition means. Crosses mark centroids in the schematic contextual point clouds.

The stability hypothesis follows from this distributional account. For unattested derived forms, learned regularities linking constituents, word forms and usage provide semantic support beyond whole-word attestation. Greater semantic correspondence between a form and its constituents is hypothesised to constrain its contextual realisations, yielding higher consensus and lower dispersion across contexts. Consensus and dispersion make this directional hypothesis testable, with bounded scope: sentence meaning, polysemy and syntactic flexibility also produce contextual variation, so the test is whether geometric summaries predict transparency beyond lexical statistics and beyond condition membership. It applies to a context-free judgement, since the ratings were collected for isolated words and the sentences are measurement contexts for the models, not a manipulation of what raters saw. Distance, volume and magnitude features broaden the diagnosis.

Two further conditions govern what such a test can show. Prior work models judgements of derived-word meaning (Marelli & Baroni, 2015), lexical decisions on affixed pseudowords (Bonandrini et al., 2023), and segmentation-sensitive semantic probes (Hofmann et al., 2021). The present resource combines matched intact and degraded forms, continuous ratings and repeated contexts to test lexical and contextual contributions to transparency. Evaluation must also distinguish available information from its recovery: aggregate scores pool structural contrasts and within-condition differences, while gains depend on the fitted estimator and representational depth (Ethayarajh, 2019; Sun et al., 2026). With features, items and protocol held constant, the median aggregate increment is $+.26$ under ridge regression and $+.04$ under random forests; within-condition point estimates are positive in all five checkpoints, with bootstrap intervals excluding zero for the two encoders. The interpretability question concerns their joint contribution to graded semantic distinctions under different evaluation choices. The contributions are threefold:

- **Fine-grained semantic accessibility.** Cross-context geometry carries transparency information beyond a multifeature distributional account. Global features retain variation within morphological conditions, while contextual geometry adds further within-condition information, with positive estimated increments in all five checkpoints and the clearest inferential support in the two encoders.
- **Evaluation granularity.** The measured contextual increment depends on both the fitted reference predictor and the semantic granularity of evaluation. Aggregate and within-condition prediction therefore capture complementary aspects of the same representational comparison, supported by layer and feature analyses.
- **Controlled resource.** The resource provides matched intact and degraded conditions, reliable isolated-word ratings, 30 sentence contexts per item, a continuous lexical-support index, and paired evaluation across five pretrained checkpoints.

2 Related work

2.1 Neurocognitive distributional accounts of semantic representation and contextual representations in LLMs

Distributional accounts characterise meaning through graded relations learned from linguistic experience (Landauer & Dumais, 1997; Baayen et al., 2011). For derived words this connects stem meaning, affix regularities and lexical neighbourhoods to interpretability: compositional distributional models implement transformations between constituent and derived-word vectors (Marelli & Baroni, 2015; Marelli et al., 2017), and affix productivity and distributional properties of novel formations provide further predictors (Hay & Baayen, 2002; Bonandrini et al., 2023). Neurocognitive studies motivate transparency as a graded dimension of derivational processing (Bozic et al., 2013; Leminen et al., 2019); the present study examines its behavioural expression through continuous ratings. Evidence from Chinese compounds further distinguishes constituent contribution from whole-word predictability across transparency measures (Chen et al., 2026). Here, the distinction concerns evaluation granularity: differences between structural conditions and graded differences within them. This literature already uses several distributional properties (Bonandrini et al., 2023); the advance here lies in comparing lexical statistics with repeated contextual encodings of the same forms within matched structural conditions, and in evaluating the additional information against an explicit lexical baseline at both granularities.

Novel derived forms are informative because familiar constituents can support interpretation beyond established whole-word usage. Subword embeddings provide vectors for such forms, and learned composition maps their constituents into derived-word representations. Hofmann et al. (2021) investigate derivational segmentation as an intervention on BERT's input representation. Here, the pretrained checkpoints and their tokenizers remain fixed: the question is what repeated contextual encodings add to a compositional lexical baseline when predicting graded human transparency.

Both paths use learned neural representations within a distributional framework: Lenci et al. (2022) find that static vectors and type vectors averaged from BERT differ in task performance and semantic-space organisation, and here fastText with learned linear composition supports the global features while repeated Transformer token vectors support cross-context geometry, so that variability as well as mean position enters the prediction. Geometric modelling of judgement has a cognitive precedent in Lee et al. (2025), who predicted belief choices from distances in an embedding space; here the model weights stay fixed and the target is a lexical-semantic judgement. Out-of-vocabulary status here means sparse or absent attestation, not that the models lack subword encodings for the items (Hofmann et al., 2021).

2.2 Layer-wise representations in language models: from brain alignment to mechanisms

Intermediate transformer representations align most strongly with neural responses during comprehension (Schrimpf et al., 2021; Caucheteux & King, 2022). Brain alignment and semantic discriminability are complementary criteria: predicting neural responses does not itself establish which morphological or graded semantic distinctions a layer supports, and controlled contrasts with transparency ratings make those distinctions explicit. Depth is therefore used to investigate behavioural accessibility, without assigning layers to brain regions. Semantic geometry is also an object of interpretability research in its own right: Park et al. (2025) formalise categorical concepts as polytopes and relate concept hierarchies to their geometry. Their target is category organisation; the present target is complementary, continuous human-rated differences among forms with the same morphological status.

The geometry available to a diagnostic reflects the computations that produce it. Information is distributed across layers (Tenney et al., 2019), and contextual embeddings are anisotropic with depth-dependent same-word similarity (Ethayarajh, 2019), which argues for measuring several geometric properties rather than reading a fixed meaning into raw similarity or magnitude. Xiao et al. (2024) show that initial tokens attract substantial attention despite limited semantic relevance, and Sun et al. (2026) distinguish persistent massive activations from attention sinks and identify pre-norm architecture as a key factor in their co-occurrence. These mechanisms motivate the norm-related

diagnostics of Section 3.3 without implying that a particular sink or activation channel produced the effects observed here.

2.3 BASELINES, CONTROLS AND WHAT A PROBING INCREMENT MEASURES

Probe performance depends on both representations and the fitted predictor. Hewitt & Liang (2019) assess selectivity against word-type control tasks, Voita & Titov (2020) incorporate the cost of learning the mapping through description length, and Pimentel et al. (2020) formulate probing in information-theoretic terms. These approaches scrutinise what successful prediction establishes (Belinkov, 2022); further work distinguishes decodability from causal use (Ravichander et al., 2021; Elazar et al., 2021).

An incremental comparison introduces a second modelling choice: the estimator used to extract information from the reference feature set. Selectivity on a word-type control task does not itself test whether a separate lexical-feature baseline has captured the predictive relations available in those features. Holding its estimator fixed leaves this source of variation in the contextual increment unexamined. The present study tests that dependence by fitting both the lexical baseline and its contextual augmentation under several estimators on matched folds. Condition-centred evaluation then determines whether the additional information distinguishes words within structural categories. The methodological contribution is their joint use: the baseline comparison establishes what the contextual features add, and the target granularity establishes which semantic differences that gain describes.

3 METHOD

Interpretability target. Semantic transparency is the property of the word predicted; representational accessibility is the property of the model assessed. Let G be the global feature block and $C_{m,l}$ the contextual block from model m at layer l; comparing G, $C_{m,l}$ and their concatenation asks whether contextual features add information beyond the tested lexical description, and the probe fitted to them quantifies accessibility under that predictor. Fitting the probe does not retrain the language model; structural discrimination and within-condition prediction are complementary tests of a contrast and of graded ratings beyond condition means.

3.1 DATA

The dataset consists of 204 morphologically derived words spanning four structural conditions: *Real* words (extremely low-frequency attested forms), *Nonce* words (morphologically legal but unattested combinations), *NonStem* words (pseudo-stem with real suffix), and *Nonaffix* words (real stem with pseudo-suffix ending). The four conditions are balanced at $n = 51$ each and are crossed with no other factor; the 51 stems recur across the three real-stem conditions, so stem-level properties are balanced by construction. Each item has 30 sentence contexts for language-model encoding. Separately, continuous transparency ratings were collected from native English speakers via Prolific: each item was rated in isolation, without sentence context, on a nine-point scale by 90 to 92 raters; item means are linearly rescaled to $[0, 1]$, and their reliability is high (ICC(2,k) $= .96$; Appendix I). The regression target is the item mean, so the 30 contexts do not increase the number of independently evaluated words. The sentence database will be released with the ratings on acceptance.

The repeated contexts make representational variation measurable for rare or unattested forms with few naturally occurring examples. The design contrasts attested and novel intact forms with pseudo-stem and pseudo-suffix forms, testing geometry across morphological well-formedness and lexical attestation. Continuous ratings distinguish graded transparency within each condition. Because ratings were collected independently of the sentences, prediction links cross-context model geometry to a behavioural property of the word, rather than to judgements elicited by those particular sentences.

Morphological Integrity. Morphological Integrity (MI) summarises lexical support for the stem, affix and whole word. It draws on type-level resources but is held out of the global feature block and used as a continuous structural reference in auxiliary comparisons with categorical condition (Appendix A). The main comparisons use the feature blocks below without MI.

3.2 FEATURE REPRESENTATIONS

Stem proximity baseline. Stem proximity serves as the primary distributional baseline and the established psycholinguistic measure of semantic transparency (Marelli & Baroni, 2015; Bonandrini et al., 2023). The CAOSS compositional framework (Bonandrini et al., 2023; Marelli et al., 2017) is adapted and trained in the present work for English derivational morphology, using fastText subword embeddings (Bojanowski et al., 2017) on derivational triplets extracted from CELEX2 (Baayen et al., 1995). A derived word vector is approximated as a weighted combination of its stem and affix vectors ($\text{derived} = S \cdot \text{stem} + A \cdot \text{affix}$), where S and A are transformation matrices estimated by regression from attested derivational forms, and stem proximity is $\cos(\text{derived}, \text{stem})$. FastText keeps stem proximity a type-level measure and represents out-of-vocabulary forms through character n-grams.

Global distributional features. Four predictors capture corpus-level semantic and morphological properties of each item independently of sentence context: three computed within the CAOSS/FastText compositional space (Table 1), namely stem proximity, embedding-coordinate entropy and neighbourhood entropy, and logit-based affix productivity from corpus statistics (Appendix H). Together these summarise stem-based compositional similarity, embedding-coordinate dispersion, local neighbourhood structure and affix productivity. Embedding-coordinate entropy describes normalised vector components, not a distribution over word meanings (named *entropy* in the code). The block is identical across the five checkpoints and serves as one reference point for all of them.

Table 1: **Vector-based definitions for the two representational paths.**

Global distributional features	**Cross-context geometric features**
Stem proximity: $\cos(\vec{v}_{\text{derived}}, \vec{v}_{\text{stem}})$	Semantic consensus $= 1 - \bar{d}_{\cos}(\mathbf{V})$
Embedding-coordinate entropy: $H(\vec{v}) = -\sum_i \tilde{p}_i \log \tilde{p}_i$, $\tilde{p}_i = \frac{v_i + \lvert\min(\vec{v})\rvert + N^{-1}}{\sum_j (v_j + \lvert\min(\vec{v})\rvert + N^{-1})}$	Contextual–stem distance $= \lVert\boldsymbol{\mu} - \mathbf{v}_{\text{stem}}\rVert_2$
	Semantic volume $= \prod_{k=1}^{3} \sigma_k(\mathbf{V} - \boldsymbol{\mu})$
Neighbourhood entropy: $H_{\text{neigh}} = -\sum_{i=1}^{K} p_i \log p_i$, $p_i = \frac{\exp(\cos(\vec{v}, \vec{n}_i)/\tau)}{\sum_{j=1}^{K} \exp(\cos(\vec{v}, \vec{n}_j)/\tau)}$	Centroid dispersion $= \frac{1}{n} \sum_{i=1}^{n} \lVert\mathbf{v}_i - \boldsymbol{\mu}\rVert_2$
	Contextual L_2 norm $= \frac{1}{n} \sum_{i=1}^{n} \lVert\mathbf{v}_i\rVert_2$

Note. $\vec{v} \in \mathbb{R}^N$ is the composed derived-word vector and $\vec{n}_i$ its i-th nearest neighbour ($K = 10$, $\tau = 0.1$). $\mathbf{V}$ contains one subword-mean-pooled vector per sentence ($n = 30$), $\boldsymbol{\mu}$ is its centroid, and $\bar{d}_{\cos}$ is the mean pairwise cosine distance. σ_k is the k-th largest singular value of the centred matrix. The contextual $\mathbf{v}_{\text{stem}}$ encodes the isolated stem at the same layer. Contextual–stem distance is named *semantic_drift* in the code. Corpus-based affix productivity is defined in Appendix H.

Cross-context geometric features. The five checkpoints are RoBERTa-base (Liu et al., 2019), the 1B encoder and decoder from the Ettin suite (Weller et al., 2026), and Qwen2.5-3B and Qwen2.5-7B (Yang et al., 2024), with pretrained weights held fixed. Each item is encoded in each of its 30 sentences with the model's own tokenizer, and its representation in a sentence is the hidden state at the target layer mean-pooled over the subword tokens that span the word, the same rule for every item, condition and checkpoint. Five geometric features summarise the 30 resulting vectors (Table 1): semantic consensus, adopted from Ethayarajh (2019); centroid dispersion; contextual–stem distance, from the contextual centroid to the stem encoded in isolation at the same layer, which also reflects the contrast between contextualised and isolated encoding; semantic volume, log-transformed because it spans two orders of magnitude; and mean contextual L_2 norm. Consensus and dispersion are hypothesis-directed measures: higher consensus and lower dispersion operationalise cross-context stability. Polysemy and syntactic flexibility also affect dispersion, so prediction is tested beyond the global baseline and within conditions. Distance, volume and norm broaden the geometric diagnosis;

norm measures magnitude rather than instability, and its relation to transparency is empirical (Park et al., 2026; Amari, 2016).

3.3 LAYER SELECTION

Contextual features are extracted from one target layer per model, selected by a two-stage cross-validated procedure before the main evaluation. The same pre-specified window is applied to every checkpoint, candidate layers between 0.50 and 0.85 of total depth, following the progression from surface to higher-level features across layers (Tenney et al., 2019); it is a shared evaluation window, not a claim that the same depth has the same function in every checkpoint. Stage 1 selects layers where the Pearson correlation between contextual L_2 norm and ratings is significantly positive on the training fold, then picks the best such layer on the validation fold. Stage 2 is triggered when Stage 1 finds no significantly positive layer, a failure mode predicted for pre-norm decoders in which massive activations inflate intermediate-layer norms independently of content (Sun et al., 2026): layer-averaged semantic diagnostics are regressed onto norm expansion and relative depth via Huber regression, and the smoothed residual is selected on. Stability is the proportion of 500 item resamples returning the selected layer. Selection uses the same 204 items and ratings under its own cross-validation. Main estimates and intervals are conditional on the resulting fixed layer choices (Appendix B).

3.4 EVALUATION

Predictive performance is a diagnostic of alignment with human judgements, not an optimisation objective. The main aggregate comparisons use 20×5 repeated cross-validation, stratified by condition and split by derived word; feature blocks and their concatenation are fitted on the same fold partition within each repeat, giving paired errors. Because the central claim concerns the dependence of a measured increment on the comparison, three estimators are used: ridge regression with cross-validated regularisation, histogram gradient boosting and a random forest; each fits both the baseline and the augmented set, so changing the estimator changes both predictors. Increments are tested with the corrected repeated k-fold t-test of Nadeau & Bengio (2003) on the 100 paired fold-level squared errors, Holm-corrected across the five checkpoints within each contrast. The five checkpoints share the same 204 items, so checkpoint-level differences are assessed by joint item bootstrap, each resampled item carrying its full vector of checkpoint-specific errors.

Within-condition analysis. Aggregate evaluation assesses prediction across the stimulus set; within-condition evaluation targets variation beyond condition means. For the latter, the outcome and every feature are centred on their condition mean under the same 20×5 cross-validation. The means are estimated on the training fold of each split and applied to its validation fold, so no validation item informs its own centring, and within-condition R^2 is taken relative to the out-of-fold error of the training-fold condition-mean predictor. Contextual-gain intervals use 20,000 item-bootstrap resamples of paired out-of-fold squared errors averaged over repeats. Full estimator comparisons and within-condition results are in Tables C1 and C2, Appendix C.

4 RESULTS

The modelling hierarchy poses three questions in sequence. How much of graded transparency do global distributional features explain, and where does that variance lie (Section 4.1)? Does cross-context geometry explain variance those features do not, and at which granularity (Section 4.2)? Under which layers, features and checkpoints is that contribution recoverable (Section 4.3)?

4.1 COMBINING GLOBAL FEATURES IMPROVES TRANSPARENCY PREDICTION

Stem proximity reaches $R^2 = .109$ under ridge regression; adding embedding-coordinate entropy, neighbourhood entropy and affix productivity raises this to .157, an increment of $+.048$ that does not reach significance under the corrected test ($p = .12$). Under the two tree ensembles the block reaches .352 and .469 and the increment is significant ($p \leq .001$), though stem proximity alone falls below the mean reference there ($R^2 = -.005$ and $-.256$), so part of that increment reflects a single-feature baseline performing poorly under a tree estimator.

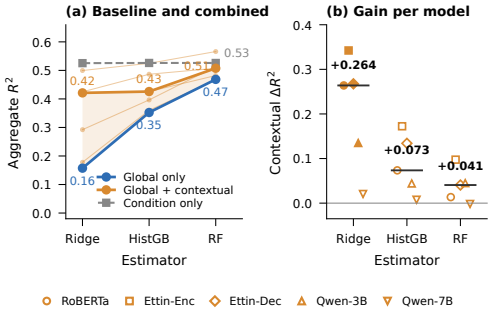

Figure 2: **Aggregate contextual gains depend on the estimator.** (a) Global-only, combined and condition-only R^2: thin lines denote checkpoints, thick lines medians, and shading the increment. (b) Gains by checkpoint, with median bars. Filled markers pass the corrected repeated-CV t-test with Holm correction ($n = 204$, 20×5 CV; Table C1).

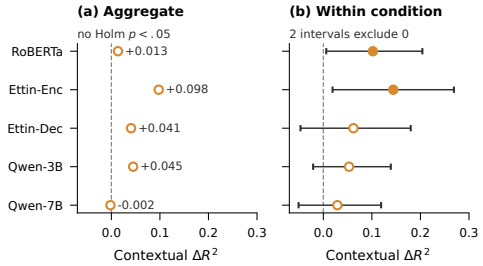

Figure 3: **Within-condition gains are positive, with clearest evidence in the encoders.** Random-forest gains ($n = 204$, 20×5 CV). (a) Aggregate estimates; none passes the corrected t-test. (b) Gains after training-fold centring, with 95% item-bootstrap intervals (20,000 resamples); the two encoder intervals exclude zero (filled points). The panels assess different targets and use different inferential summaries (Section 3.4; Table C2).

The block captures structural differences as well as graded variation within conditions. Under a random forest the four features reach .469, against .526 for the four-level condition factor entered as three dummy variables, so the block approaches an explicit structural reference. Only stem proximity approaches a step function over conditions ($\eta^2 = .78$ between conditions); the other three retain substantial within-condition variance, and for affix productivity most of it is within rather than between ($\eta^2 = .27$). The block also retains a graded relation to the outcome within conditions: with condition means removed from outcome and features, the same four features account for $R^2 = .119$ under the same random forest, the reference for the contextual contribution. Aggregate and within-condition R^2 use different reference errors, so their values do not partition explained variance.

4.2 Contextual geometry adds information within morphological conditions

Within conditions, contextual geometry adds predictive information. Figure 3b subtracts the condition mean from the outcome and from every feature, estimating each mean on the training fold and applying it to the validation fold. The distributional block reaches $R^2 = .119$; adding contextual geometry raises this to .148 to .263. Bootstrap intervals exclude zero for Ettin-Encoder-1B ($+.144$, $[+.019, +.269]$) and RoBERTa-base ($+.102$, $[+.006, +.204]$), and the estimated increment is positive in all five checkpoints.

Fitting each condition separately, the signal is concentrated in the two morphologically intact conditions, Real and Nonce, and is at or below zero in both degraded conditions; within the intact conditions it is strongest for Nonce items in four of five checkpoints. For these unattested combinations, the contextual block carries information about rated transparency beyond the tested compositional features.

Adding subword piece count to both predictors yields increments of $+.100$ $[+.005, +.202]$ for RoBERTa-base and $+.122$ $[-.001, +.242]$ for Ettin-Encoder-1B. Point estimates remain positive in all five checkpoints, although only RoBERTa's interval excludes zero after adjustment (Appendix J).

The aggregate increment depends on the estimator. Figure 2 adds the five contextual features to the global distributional baseline with the same data, features and protocol under three estimators, each used for both the global-only and the combined predictor. Ridge regression reports a median contextual increment of $+.264$, significant in four of five checkpoints after correction; histogram gradient boosting $+.073$, significant in none; a random forest $+.041$, again in none (Figure 3a). A TabPFN configuration (Grinsztajn et al., 2025) evaluated under a separate single-split protocol places the baseline at .591 with an increment indistinguishable from zero.

Across these fitted comparisons, better global-only prediction accompanies a smaller aggregate increment. Both predictors change with the estimator, so the measured increment depends on the fitted comparison rather than on the contextual feature block alone. Contextual features fitted alone reach $R^2 = .134$ to $.410$ under the random forest; the smaller increment describes their additional predictive value beyond the fitted global baseline.

Auxiliary structural-reference checks. In a separate leave-one-out ridge analysis without MI or condition membership as predictors, contextual features alone reach $R^2 = .370$–$.475$ in the two encoders and Ettin-Decoder, above the global block's $.153$, and $.084$–$.128$ in Qwen. Combining the blocks improves on the global baseline in every checkpoint ($R^2 = .209$–$.530$), without an explicit structural predictor. Replacing categorical condition with continuous MI while retaining the global block increases the mean contextual increment from $+.030$ to $+.061$, with a larger increment in every checkpoint. These auxiliary leave-one-out comparisons use a protocol distinct from the main 20×5 evaluation (Appendix A). Auxiliary probes distinguish Real from Nonce items at AUROC $.552$–$.720$ using contextual features, compared with $.459$ for global features (Appendix D). This complementary analysis tests lexical attestation rather than graded within-condition transparency.

4.3 Supporting diagnostics: layers, features and checkpoints

Layer. The two-stage procedure selects layers at 61 to 83% relative depth in all five checkpoints, within the common 0.50–0.85 window, with bootstrap stability above 80% for four of them; the exception, Ettin-Encoder-1B, sits on a plateau across layers 20 to 22 (R^2 within .003). Recomputing the within-condition increment at every evaluated layer under 5×5 cross-validation gives 45 positive values out of 49 cells, the four exceptions all above $-.02$, with a positive median over the window in every checkpoint ($+.01$ to $+.13$); positive increments extend across the evaluation window, with the selected layer ranking best within its window for only one checkpoint (Figure B1, Appendix B).

Norm-based diagnostic. Stage 1 selects on the correlation between contextual L_2 norm and ratings and succeeds for RoBERTa-base and both Ettin models. For the two Qwen checkpoints it finds no significantly positive layer on any fold, and Stage 2, which regresses layer-averaged semantic diagnostics on norm expansion and relative depth before selecting, is required. The pattern is consistent with the norm-related effects described for pre-norm decoders (Sun et al., 2026). At the selected Qwen layers, multivariate violation probes recover the contrast between intact and degraded forms despite the weak univariate norm diagnostic (Appendix D), showing that the contrast is accessible through the broader geometric representation.

Checkpoint comparisons. In the 5×5 analysis, ordered by within-condition increment the checkpoints run Ettin-Encoder-1B $+.151$ $[+.018, +.282]$, RoBERTa-base $+.098$ $[+.000, +.199]$, Qwen2.5-3B $+.057$ $[-.016, +.139]$, Ettin-Decoder-1B $+.049$ $[-.062, +.166]$ and Qwen2.5-7B $+.006$. The two encoders lead, and the violation contrasts group the same way ($d = 0.75$ to 1.39 where Stage 1 succeeds, $|d| \leq 0.21$ in the Qwen models). The matched Ettin pair favours the encoder ($+.102$ $[-.021, +.219]$) without reaching significance; with two of three decoder-only checkpoints drawn from one family, the grouping is exploratory (Appendix E).

Feature contributions. An add-one/drop-one analysis under the main within-condition protocol (Appendix F) shows that several features improve prediction when added alone to the global block, while the largest loss from removing any one is $.023$: consensus, dispersion and volume each contribute in RoBERTa-base, whereas the norm alone gives $+.158$ in Ettin-Encoder-1B. No positive drop-one loss has an interval excluding zero, indicating substantial predictive overlap among the contextual summaries rather than dependence on a single feature.

5 Discussion

5.1 Lexical support and contextual realisation

The feature hierarchy links two properties of learned semantic representations to graded morphological meaning. Global distributional features describe constituent relations, neighbourhood structure and affix productivity; contextual geometry describes how the same formation is represented across

sentences. Their combination extends multifeature accounts of novel-word processing (Bonandrini et al., 2023) to a joint comparison of lexical support and contextual realisation. Both blocks discriminate conditions and retain within-condition information, with different predictive emphases. The additional contextual signal is clearest among intact forms, particularly Nonce items, where legal morphology permits graded interpretation beyond established whole-word usage. The structural-reference checks reinforce this reading: contextual features predict ratings without a condition variable, and their mean increment increases when a continuous lexical-support measure replaces categorical condition.

Informative geometric summaries differ across checkpoints. RoBERTa's cosine-based consensus adds information beyond global features through a scale-invariant summary of contextual directions; Ettin-Encoder's mean norm demonstrates the predictive value of magnitude. Dispersion and volume depend on both scale and direction. At the selected Qwen layers, weak norm-based structural separation coexists with successful multivariate category prediction. Thus no single magnitude measure characterises all accessible information. Consistency and dispersion are motivated by constituent-constrained stability, whereas norm measures magnitude; predictive gains under a nonlinear model alone do not establish a directional stability relation. Neurocognitive accounts supply the behavioural dimension against which these candidate summaries are tested (Bozic et al., 2013; Leminen et al., 2019). This complements brain-alignment research by asking which semantic distinctions a representation supports, alongside how well it predicts neural responses (Schrimpf et al., 2021; Caucheteux & King, 2022).

5.2 Baseline capacity and semantic granularity

A representational increment is a comparison between fitted predictors. The estimator sweep changes both predictors: improved global-only prediction accompanies a smaller aggregate contextual gain. The two evaluation targets answer complementary questions. Aggregate evaluation measures prediction across the full stimulus set, while within-condition evaluation tests graded differences beyond condition means; neither score substitutes for the other. Together, these tests extend the evaluation beyond the capacity of the contextual probe (Hewitt & Liang, 2019; Voita & Titov, 2020) to the treatment of its reference features and the semantic distinctions pooled by its score. Training-fold centring, paired errors and joint item resampling keep these comparisons matched across feature blocks and checkpoints. Layer and feature diagnostics then identify the geometric summaries through which the behavioural target is accessible.

The resource distinguishes category sensitivity from graded semantic prediction through matched conditions, continuous ratings and repeated contexts. Across stratified subsamples of 80, 120 and 160 items, the mean within-condition increment remains similar in every checkpoint, while the spread across subsamples generally narrows as item count increases (Appendix G).

Scope. The evidence concerns 204 English forms and five pretrained checkpoints from three model families. Piece count is included as a covariate in a sensitivity analysis (Appendix J). The interpretation concerns information recoverable through the specified features and predictors, rather than causal use (Ravichander et al., 2021; Elazar et al., 2021).

6 Conclusion

Global distributional features and cross-context geometry jointly characterise graded semantic transparency. Global features capture both structural contrasts and graded variation within conditions, while contextual geometry provides additional within-condition information, with the strongest inferential support in the two encoders. Together, these results connect learned lexical regularities and contextual variation to an independently measured behavioural target, and distinguish broad structural sensitivity from fine-grained semantic accessibility. The estimator comparison provides a complementary methodological finding: the measured contextual increment depends on how both the lexical reference and its contextual augmentation are fitted. Aggregate and within-condition evaluation therefore expose different aspects of the same representational comparison, with graded information remaining visible within conditions even when the aggregate increment is small. More broadly, the results show that both the fitted reference and the semantic granularity of evaluation are integral to interpreting representational contributions.

AI USE STATEMENT

Generative AI tools were used primarily to assist manuscript writing and language polishing, with additional basic checks of references and consistency between the text and reported results. The authors reviewed the AI-assisted revisions and take responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

Section 3 defines the representational features, target-word pooling, layer-selection procedure and evaluation protocols. Appendix A specifies Morphological Integrity; Appendix B reports layer diagnostics; Appendix C provides full performance tables; and Appendix F documents the feature diagnostics and Appendix G the item-count robustness analysis. Rating collection, normalisation, reliability and sentence examples are described in Appendix I; Appendix J reports the token-count sensitivity analysis. Analysis scripts and result tables accompany the source materials. The sentence database, ratings and feature tables will be released on acceptance; third-party lexical resources remain subject to their original licences.

REFERENCES

Shun-ichi Amari. *Information geometry and its applications*. Springer, 2016. doi: 10.1007/978-4-431-55978-8.

Rolf Harald Baayen, Richard Piepenbrock, and Leon Gulikers. CELEX2. Linguistic Data Consortium, Philadelphia, 1995. LDC Catalog No. LDC96L14.

Rolf Harald Baayen, Petar Milin, Dušica Filipović Đurđević, Peter Hendrix, and Marco Marelli. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, 118(3):438–481, 2011. doi: 10.1037/a0023851.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. doi: 10.1162/tacl_a_00051.

Rolando Bonandrini, Simona Amenta, Simone Sulpizio, Marco Tettamanti, Alessia Mazzucchelli, and Marco Marelli. Form to meaning mapping and the impact of explicit morpheme combination in novel word processing. *Cognitive Psychology*, 145:101594, 2023. doi: 10.1016/j.cogpsych.2023.101594.

Mirjana Bozic, Lorraine K. Tyler, Li Su, Cai Wingfield, and William D. Marslen-Wilson. Neurobiological systems for lexical representation and analysis in English. *Journal of Cognitive Neuroscience*, 25(10):1678–1691, 2013. doi: 10.1162/jocn_a_00420.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134, 2022. doi: 10.1038/s42003-022-03036-1.

Jing Chen, Emmanuele Chersoni, Marco Marelli, and Chu-Ren Huang. The multidimensional nature of semantic transparency in a cross-linguistic perspective: Evidence from human intuitions, computational estimates, and processing data for Chinese compounds. *Cognitive Science*, 50(3):e70194, 2026. doi: 10.1111/cogs.70194.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359.

Kawin Ethayarajh. 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)*, pp. 55–65. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1006.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Benjamin Jäger, Dominik Safaric, Simone Alessi, Adrian Hayler, et al. TabPFN-2.5: Advancing the state of the art in tabular foundation models. *arXiv preprint arXiv:2511.08667*, 2025.

Jennifer Hay and Rolf Harald Baayen. Parsing and productivity. In *Yearbook of Morphology 2001*, pp. 203–235. Kluwer Academic Publishers, 2002. doi: 10.1007/978-94-017-3726-5_8.

John Hewitt and Percy Liang. 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)*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3594–3608. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.279.

Thomas K. Landauer and Susan T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997. doi: 10.1037/0033-295X.104.2.211.

Byunghwee Lee, Rachith Aiyappa, Yong-Yeol Ahn, Haewoon Kwak, and Jisun An. A semantic embedding space based on large language models for modelling human beliefs. *Nature Human Behaviour*, 9:1928–1940, 2025. doi: 10.1038/s41562-025-02228-z.

Alina Leminen, Eva Smolka, Jon Andoni Duñabeitia, and Christos Pliatsikas. Morphological processing in the brain: The good (inflection), the bad (derivation) and the ugly (compounding). *Cortex*, 116:4–44, 2019. doi: 10.1016/j.cortex.2018.08.016.

Alessandro Lenci, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. A comparative evaluation and analysis of three generations of Distributional Semantic Models. *Language Resources and Evaluation*, 56(4):1269–1313, 2022. doi: 10.1007/s10579-021-09575-z.

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

Marco Marelli and Marco Baroni. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review*, 122(3):485–515, 2015. doi: 10.1037/a0039267.

Marco Marelli, Christina L. Gagné, and Thomas L. Spalding. Compounding as abstract operation in semantic space: Investigating relational effects through a large-scale, data-driven computational model. *Cognition*, 166:207–224, 2017. doi: 10.1016/j.cognition.2017.05.026.

Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Machine Learning*, 52 (3):239–281, 2003. doi: 10.1023/A:1024068626366.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Kiho Park, Todd Nief, Yo Joong Choe, and Victor Veitch. The information geometry of softmax: Probing and steering. *arXiv preprint arXiv:2602.15293*, 2026.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4609–4622. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.420.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 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*, pp. 3363–3377. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.295.

Claudia H. Sánchez-Gutiérrez, Hugo Mailhot, S. Hélène Deacon, and Maximiliano A. Wilson. MorphoLex: A derivational morphological database for 70,000 English words. *Behavior Research Methods*, 50(4):1568–1580, 2018. doi: 10.3758/s13428-017-0981-8.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021. doi: 10.1073/pnas.2105646118.

Shangwen Sun, Alfredo Canziani, Yann LeCun, and Jiachen Zhu. The spike, the sparse and the sink: Anatomy of massive activations and attention sinks. *arXiv preprint arXiv:2603.05498*, 2026.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1452.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.14.

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin Van Durme. Seq vs Seq: An open suite of paired encoders and decoders. In *The Fourteenth International Conference on Learning Representations*, 2026.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

A Auxiliary structural-reference analyses

Morphological Integrity (MI) draws on the type-level lexical resources underlying the global distributional account. In the present evaluation it is held out of the four-feature global block and used as a structural reference in the auxiliary comparisons below. This separates explicit lexical-support coding of the morphological conditions from the predictive contribution of the global features. The main estimator and condition-centred comparisons use the four-feature block defined in Section 3.2.

A.1 Construction of Morphological Integrity

MI averages support for the stem, affix and whole word:

$$\mathrm{MI}(w) = \tfrac{1}{3}\big(s_{\mathrm{stem}}(w) + s_{\mathrm{affix}}(w) + s_{\mathrm{word}}(w)\big), \tag{1}$$

$$s_{\mathrm{stem}}(w) = \tfrac{1}{2}\big[\mathrm{pct}_S(\log(1+\mathrm{FamSize}(w))) + \mathrm{pct}_S(\log(1+\mathrm{Freq}(w)))\big], \tag{2}$$

$$s_{\mathrm{affix}}(w) = \mathrm{pct}_A(p_{\mathrm{affix}}(w)), \tag{3}$$

$$s_{\mathrm{word}}(w) = \mathbf{1}[w \text{ is attested}]. \tag{4}$$

Here S contains items with an attested stem (Real, Nonce and Nonaffix), and A contains items with a real affix (Real, Nonce and NonStem). Structurally absent stem or affix support is set to zero. FamSize is the MorphoLEX `ROOT1_FamSize` value for the stem (Sánchez-Gutiérrez et al., 2018); Freq is its CELEX2 frequency (Baayen et al., 1995); and p_{affix} is the productivity value supplied with the item data. Whole-word support is one for the Real condition, defined by OED attestation, and zero otherwise.

For each quantity, pct_S or pct_A ranks the available values among items in the corresponding support set, using average ranks for ties and $\text{rank}/(n+1)$ for n available items. This maps defined support into $(0, 1)$ while reserving zero for structural absence. The implementation uses $\log(1+x)$ for family size and frequency and averages the two percentile scores. The exported CELEX `stem_derfam` count is not used in this calculation; the family-size input is `stem_famsize`, obtained from MorphoLEX. The three support components enter the final mean with equal weights.

A.2 Whole-word support and out-of-vocabulary status

The materials target forms with sparse or absent lexical attestation. Of the 51 Real items, 48 are absent from MorphoLex and 49 from the CELEX2 frequency lexicon, and the two forms that occur have COBUILD frequencies of 0 and 1. Whole-word support therefore carries no graded information for these materials and enters as a binary indicator. Whole-word support therefore represents lexical attestation in this auxiliary index.

A.3 Validation

Table A1 reports MI by condition. The ordering Real > Nonce > {NonStem, Nonaffix} holds, with Real above Nonce ($t = 14.87$, $p < 10^{-26}$), Nonce above NonStem ($t = 7.29$, $p < 10^{-10}$) and Nonce above Nonaffix ($t = 7.51$, $p < 10^{-10}$). NonStem and Nonaffix do not differ ($t = -0.47$, $p = .64$), consistent with their similar structural support; a non-significant contrast does not establish statistical equivalence. Within-condition standard deviations lie between .077 and .129, so the index is graded within every condition.

MI is constructed to preserve the ordinal structure of the four conditions, so a substantial proportion of its variance is between conditions by design. Entered alone as a single continuous predictor it recovers 93.4% of the cross-validated accuracy of the four-level categorical factor ($R^2 = .494$ against .528) while spending one degree of freedom rather than three.

Table A1: Morphological Integrity by condition. Support levels are on a percentile scale in $(0, 1)$, with 0 denoting structural absence.

Condition	Stem	Affix	Whole word	MI	SD	n
Real	.50	.55	1.00	.683	.113	51
Nonce	.50	.48	0	.325	.129	51
Nonaffix	.50	0	0	.167	.077	51
NonStem	0	.48	0	.158	.100	51

A.4 Within-condition behaviour

Within condition, affix support correlates with rated transparency in the two conditions whose morphology is intact, Real ($\rho = +.333$, $p = .017$) and Nonce ($\rho = +.402$, $p = .003$), and not in NonStem ($\rho = +.092$, $p = .52$), the one degraded condition in which it is defined. Stem support does not correlate with ratings in any condition. The dissociation matches the within-condition pattern reported in Section 4.2 for the distributional block as a whole.

A.5 Auxiliary structural-reference comparisons

In a separate leave-one-out ridge analysis ($\alpha = 1$, with features standardised within each training set), replacing the condition factor with MI increases the contextual increment in every checkpoint (mean +.030 to +.061). This comparison uses WT+G versus WT+G+C, or MI+G versus MI+

$G + C$, where WT denotes condition membership. It changes an explicit structural reference while retaining the global feature block. Its estimates belong to the auxiliary protocol and are separate from the main 20×5 comparisons of G and $G + C$.

A further auxiliary leave-one-out ridge analysis omits both MI and the condition factor. The contextual block alone reaches $R^2 = .084$ to $.475$ (above the global block's $.153$ in three of five checkpoints), and global plus contextual features reach $.209$ to $.530$, above the global block in every checkpoint. The condition-centred main analysis remains the direct test of graded prediction beyond condition means and G.

B LAYER SELECTION DIAGNOSTICS

Relation to the main evaluation. The two-stage procedure selects one layer per checkpoint using the same 204 items and transparency ratings under its own cross-validation, before the main 20×5 evaluation. Selection is not repeated within each outer training fold: the reported increments and intervals treat the preselected layers as fixed and do not incorporate layer-selection uncertainty. The within-window sweep (Table B2; Appendix D.1) evaluates 49 layer-by-checkpoint combinations, with positive increments in 45. This shows that positive increments are not confined to the selected layers.

B.1 BOOTSTRAP STABILITY

Stability is the proportion of 500 item-resampled runs in which the procedure returns the layer selected on the full sample. Table B1 reports it for all five checkpoints together with the selected layer and its relative depth.

Table B1: Selected layers and bootstrap stability across 500 item resamples. Stability refers to the layer selected on the full sample throughout.

Checkpoint	Class	Layer	Relative depth	Stability
RoBERTa-base	Encoder	L10	83%	99.8%
Ettin-Encoder-1B	Encoder	L21	75%	29.8%
Ettin-Decoder-1B	Decoder	L18	64%	93.6%
Qwen2.5-3B	Decoder	L22	61%	99.6%
Qwen2.5-7B	Decoder	L22	79%	80.6%

Table B2: **Layer choices differ across selection rules, while contextual gains persist across much of the evaluated range.** Stage 1/2 columns show independently selected layers and fold votes in parentheses; Selected gives the combined procedure. The right columns report within-condition gains over all evaluated layers (random forest, 5×5 CV, fold-local centring). Selection rules and their depth window are defined in Section 3.3.

		Run in isolation			Within-condition ΔR^2	
Checkpoint	Class	Stage 1	Stage 2	Selected	layers	median (range)
RoBERTa-base	Encoder	L10 (5/5)	L10 (3/5)	L10, 83%	8–10	+.080 (+.044, +.098)
Ettin-Encoder-1B	Encoder	L21 (3/5)	L18 (5/5)	L21, 75%	14–23	+.132 (+.103, +.159)
Ettin-Decoder-1B	Decoder	L18 (4/5)	L19 (3/5)	L18, 64%	14–24	+.018 (−.012, +.080)
Qwen2.5-3B	Decoder	L27 (4/5) †	L22 (5/5)	L22, 61%	18–31	+.048 (+.025, +.063)
Qwen2.5-7B	Decoder	L21 (4/5) †	L22 (5/5)	L22, 79%	14–24	+.007 (−.018, +.030)

† Stage 1 found no significantly positive layer on any fold and fell back to a forced selection.

B.2 INTERPRETATION OF THE ETTIN-ENCODER-1B FIGURE

Bootstrap resampling confirms stable selection for four of five checkpoints ($\geq 80.6\%$). The lower figure for Ettin-Encoder-1B (29.8%) reflects a broad plateau across layers 20 to 22: adjacent layers

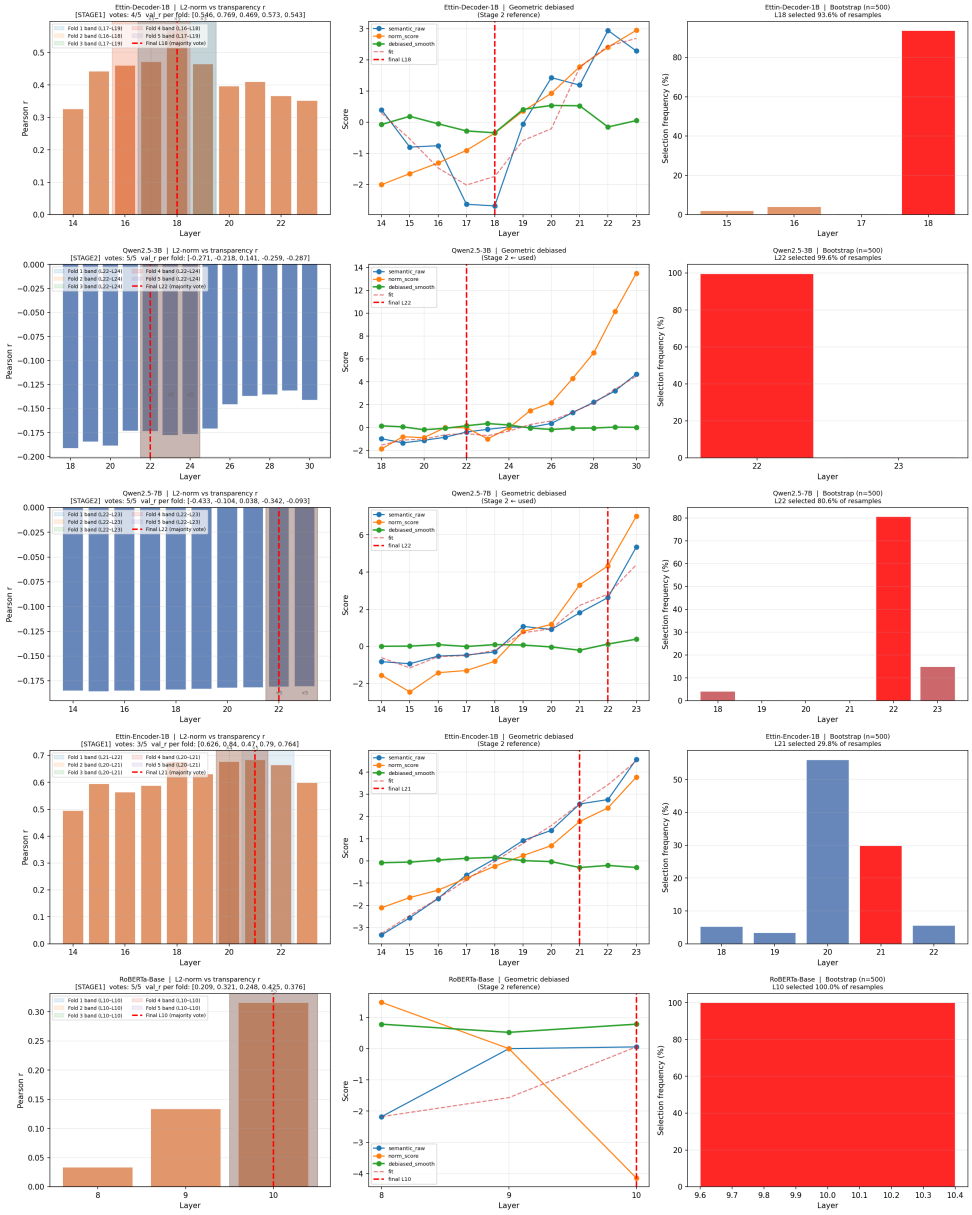

Figure B1: **Selected layers reflect different norm and geometry profiles.** Each row is one checkpoint ($n = 204$). Left: norm–rating Pearson correlation; grey shading marks the CV candidate band. Middle: raw semantic score (blue), norm score (orange), smoothed norm-debiased score (green), and the fitted component (pale dashed line). Red vertical lines mark the selected layer. Right: selection frequencies over 500 item resamples; the red bar is the layer selected on the full sample. These are selection diagnostics, not independent validation curves.

L20 and L22 yield cross-validated R^2 values within .003 of L21, indicating near-equivalent representational quality across this region. Resampling redistributes the selection within that plateau, on which the choice of layer is close to immaterial.

The right-hand panel of Table B2 makes the same point directly. Across the evaluated Ettin-Encoder layers, the within-condition contextual increment has a median of $+.132$ and ranges from $+.103$ to $+.159$, remaining positive throughout this range.

B.3 Stage 2 applied to Stage 1 models

Stage 2 run in isolation on the three checkpoints where Stage 1 succeeds returns L10 for RoBERTa-base, identical to the Stage 1 selection, and L19 for Ettin-Decoder-1B, within one layer of it. For Ettin-Encoder-1B it returns L18 with a bootstrap stability of 97.2%, against 29.8% for the Stage 1 selection. The stages therefore need not select the same layer even when Stage 1 succeeds. The within-condition layer sweep assesses the extent to which the contextual increment persists across these choices.

C Full results tables

Tables C1 and C2 report the estimator sweep and within-condition comparisons discussed in Section 4.

Table C1: **Aggregate contextual gains depend on the estimator.** Mean out-of-fold R^2 ($\pm$SD over 20 repeats; $n = 204$, 20×5 CV). ΔR^2 compares both blocks with global features alone; p values use the corrected repeated-CV t-test with Holm adjustment across five checkpoints per estimator. The same global features are used across checkpoints.

Estimator	Checkpoint	Distributional	Contextual	Both	ΔR^2 (Holm p)
Ridge (linear)					
	RoBERTa-base	.157 $\pm$.009	.388 $\pm$.010	.421 $\pm$.015	+.264 (<.001)
	Ettin-Encoder-1B	.157 $\pm$.009	.415 $\pm$.160	.500 $\pm$.031	+.342 (.006)
	Ettin-Decoder-1B	.157 $\pm$.009	.370 $\pm$.015	.425 $\pm$.017	+.268 (<.001)
	Qwen2.5-3B	.157 $\pm$.009	.231 $\pm$.017	.292 $\pm$.019	+.135 (.046)
	Qwen2.5-7B	.157 $\pm$.009	.052 $\pm$.032	.178 $\pm$.034	+.021 (.749)
Histogram gradient boosting					
	RoBERTa-base	.352 $\pm$.038	.196 $\pm$.046	.426 $\pm$.024	+.073 (.676)
	Ettin-Encoder-1B	.352 $\pm$.038	.385 $\pm$.032	.525 $\pm$.023	+.172 (.104)
	Ettin-Decoder-1B	.352 $\pm$.038	.217 $\pm$.043	.486 $\pm$.035	+.134 (.167)
	Qwen2.5-3B	.352 $\pm$.038	.143 $\pm$.042	.397 $\pm$.028	+.044 (.721)
	Qwen2.5-7B	.352 $\pm$.038	−.006 $\pm$.034	.360 $\pm$.033	+.008 (.882)
Random forest					
	RoBERTa-base	.469 $\pm$.026	.305 $\pm$.026	.482 $\pm$.018	+.013 (1.000)
	Ettin-Encoder-1B	.469 $\pm$.026	.410 $\pm$.017	.566 $\pm$.022	+.098 (.839)
	Ettin-Decoder-1B	.467 $\pm$.028	.259 $\pm$.035	.507 $\pm$.022	+.041 (1.000)
	Qwen2.5-3B	.469 $\pm$.026	.134 $\pm$.034	.513 $\pm$.026	+.045 (.867)
	Qwen2.5-7B	.469 $\pm$.026	.140 $\pm$.021	.467 $\pm$.024	−.002 (1.000)
Condition index (WordType) alone		.525 / .526 / .526 for Ridge / HistGB / RF			

Table C2: **Separating the two granularities.** Within-condition R^2 is obtained with a random forest under 20×5 repeated cross-validation; intervals are 20,000-replicate item bootstraps. Outcome and feature condition means are estimated on each training fold and applied unchanged to its validation fold. Performance is relative to the out-of-fold error of the training-fold condition-mean predictor. The right-hand panel fits each condition separately ($n = 51$, ridge).

	Within condition (all 204 items)				Per condition, R^2			
Checkpoint	Distrib.	+Contextual	ΔR^2	95% CI	Real	Nonce	NonStem	Nonaffix
RoBERTa-base	.119	.220	+.102	[+.006, +.204]	.173	.360	−.080	−.083
Ettin-Encoder-1B	.119	.263	+.144	[+.019, +.269]	.154	.287	−.058	−.114
Ettin-Decoder-1B	.118	.180	+.062	[−.047, +.180]	.073	.177	−.180	−.087
Qwen2.5-3B	.119	.171	+.053	[−.021, +.139]	.042	.169	−.061	−.094
Qwen2.5-7B	.119	.148	+.029	[−.051, +.119]	.125	.065	−.188	−.091

Table D1: **Contextual geometry discriminates the lexical-attestation contrast.** Pooled out-of-fold AUROC under 20×5 repeated stratified CV, with logistic regression on standardised features and p values from 199 label permutations. Contrast A compares intact with degraded forms; contrast B compares Real with Nonce forms within the intact design. The global block is shared across language models.

	A. Intact vs degraded (n=204)		B. Real vs Nonce (n=102)	
Checkpoint	Contextual	Distributional	Contextual	Distributional
RoBERTa-base	.811*	.909*	.704*	.459
Ettin-Encoder-1B	.830*	.909*	.720*	.459
Ettin-Decoder-1B	.875*	.909*	.645*	.459
Qwen2.5-3B	.687*	.909*	.627*	.459
Qwen2.5-7B	.709*	.909*	.552	.459

* permutation $p < .05$. Chance = .500.

D VIOLATION PROBES

The two feature blocks differ in their ability to discriminate the tested contrasts. Table D1 probes them directly, as balanced classification problems under the same cross-validation protocol.

Contrast A tests whether the intact–degraded distinction is recoverable from the features. Both blocks discriminate it: contextual AUROC ranges from .687 to .875 and distributional AUROC is .909, all with permutation $p = .005$. These multivariate classification results are distinct from the univariate norm contrasts reported below; AUROC is not converted to a standardised mean difference.

Contrast B is the one that separates the blocks. Comparing the two structurally intact conditions that differ in lexical attestation, contextual geometry separates Real from Nonce items with AUROC .552 to .720, significant in four of five checkpoints, whereas the distributional block is not reliably above chance (.459, $p = .51$; no individual distributional feature reaches $|g| > .20$). The contextual block thus discriminates this attestation contrast more readily than the tested global block. Rated transparency also differs between the conditions (Real $M = .599$, Nonce $M = .343$, $g = 1.23$, $p < .001$). This is a complementary diagnostic of feature sensitivity, not a replication of the within-condition regression or evidence that attestation is the only difference affecting the representations.

Aggregate increments and uncertainty. Item bootstraps place the aggregate increment at $+.102$ $[-.024, +.233]$ for Ettin-Encoder-1B and $+.049$ $[-.013, +.111]$ for Qwen2.5-3B, with $P(\Delta > 0) = .94$ in both. These aggregate intervals include zero, whereas the within-condition intervals exclude zero for the two encoder models (Table C2). The estimates refer to different evaluation targets; a small or statistically unresolved aggregate increment alone does not establish absence of within-condition information.

D.1 WITHIN-WINDOW ROBUSTNESS AND SELECTION ABLATION

Positive contextual increments extend beyond the selected layers. Contextual features are extracted from the model-specific target layer identified via the procedure in Section 3. Table B2 reports a three-way ablation of that procedure alongside the robustness of the main result across the candidate window.

Run in isolation, Stage 2 reproduces the Stage 1 selection exactly for RoBERTa-base and to within one layer for Ettin-Decoder-1B; for Ettin-Encoder-1B it selects L18 rather than L21. For the two Qwen checkpoints Stage 1 finds no significantly positive layer on any fold and falls back to a forced selection, while Stage 2 selects L22 with unanimous fold agreement. Stage 2 provides an alternative selection criterion when the positive norm–transparency criterion fails; agreement between the stages is assessed empirically rather than assumed.

Recomputing the within-condition increment at every evaluated layer under 5×5 cross-validation, with condition means estimated inside each training fold, gives 45 positive values out of 49 layer-by-checkpoint cells; the four exceptions, two layers each of Ettin-Decoder-1B and Qwen2.5-7B, lie

between $-.02$ and zero. The selected layer is the window maximum in only one of five checkpoints. Positive increments are therefore not confined to the selected layer, and the selection window itself was fixed in advance from a relative-depth heuristic (Section 3).

Where the L_2 diagnostic fails. The relation between norm and structural condition differs across the five checkpoints. Univariate norm contrasts reveal large effects at the semantic peak layer in models with strong L_2–transparency alignment ($d = 0.75$–1.39), and negligible effects in Qwen models ($|d| \leq 0.21$); these representation-level effects do not fully translate into predictive gains. The effect also attenuates from the selected layer to the final layer in the three checkpoints where Stage 1 succeeds ($g = 0.75$, 1.17 and 1.38 against 0.31, 0.26 and 0.34), showing that this univariate contrast depends on the evaluated depth. In the two Qwen checkpoints the pattern inverts: the selected layer shows no effect ($g = -0.09$ and -0.21) while the final layer shows a large one ($+1.22$ and $+0.93$). The multivariate probe of Table D1 nonetheless recovers the contrast at the selected layer in both (.687 and .709), so morphological information is recoverable at that depth despite its weak expression in the norm feature. Norm-related architectural effects provide one motivation for examining this discrepancy, but the comparison does not isolate its mechanism. The Stage 1 outcome, depth pattern and univariate contrast provide related diagnostics of the observed difference between these checkpoints.

E CHECKPOINT COMPARISONS AND ARCHITECTURAL GROUPING

Checkpoint point estimates show an ordering without establishing a general ranking; grouping by architecture is exploratory. Because the five checkpoints share the same 204 items, differences between them are paired and can be tested by joint item bootstrap. Ettin-Encoder-1B and Ettin-Decoder-1B are drawn from the Ettin suite, paired encoder- and decoder-only models with matched training data and parameter count (Weller et al., 2026), providing a more constrained comparison than unrelated model families. Ordered by within-condition increment (random forest, 5×5 repeated cross-validation at the selected layer, condition means estimated within training folds) the checkpoints run Ettin-Encoder-1B $+.151$ $[+.018, +.282]$, RoBERTa-base $+.098$ $[+.000, +.199]$, Qwen2.5-3B $+.057$ $[-.016, +.139]$, Ettin-Decoder-1B $+.049$ $[-.062, +.166]$ and Qwen2.5-7B $+.006$ $[-.073, +.094]$. Of the ten paired contrasts, one excludes zero, Ettin-Encoder-1B against Qwen2.5-7B ($+.145$ $[+.013, +.266]$); the matched Ettin pair gives $+.102$ $[-.021, +.219]$ in favour of the encoder.

Neither controlled comparison reaches significance: the matched Ettin pair, which compares the paired models under matched training data and recipe, gives $+.102$ $[-.021, +.219]$, and the Qwen scale contrast gives $+.051$ $[-.020, +.124]$.

Grouping by attention pattern yields a positive estimated contrast, with an interval at the boundary of zero. The mean within-condition increment is higher for bidirectional than causal models ($\Delta = +.087$, $[+.000, +.167]$, $P(\Delta > 0) = .975$). At the output level the divergence between softmax distributions for intact and degraded items is large and consistent in the two encoders ($g = -0.94$ and -1.02, both Holm $p < .001$) and inconsistent across the three decoders, which give one null, one negative and one reversed effect. This grouping is reported as exploratory, since two of the three causal models come from the same family and directionality is therefore not fully separable from family.

F PREDICTIVE CONTRIBUTIONS OF INDIVIDUAL CONTEXTUAL FEATURES

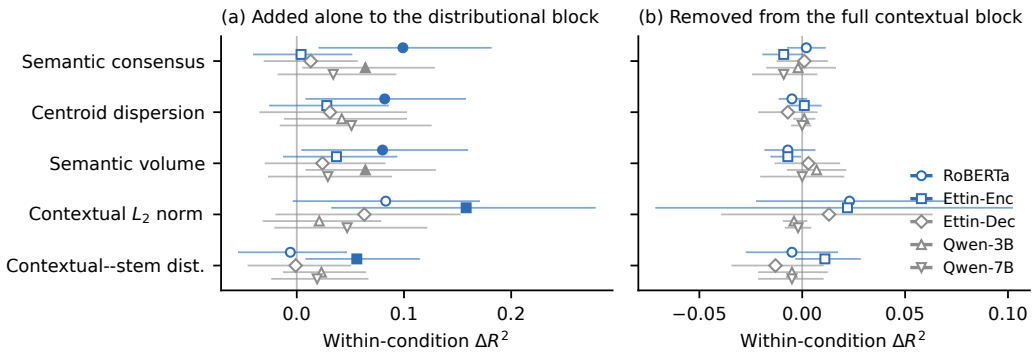

Figure F1: **Predictive contributions vary across features and checkpoints.** (a) Gain from adding one feature to the global block. (b) Loss from removing it from the combined block; negative values favour removal. Whiskers show nominal 95% item-bootstrap intervals, and filled points indicate exclusion of zero on either side ($n = 204$, 20×5 CV). Blue denotes encoders; grey denotes decoders.

Figure F1 compares two diagnostics under the main within-condition protocol: random forests, 20×5 repeated cross-validation, and condition means estimated in each training fold. The add-one gain is $R^2(G + f) - R^2(G)$ for global features G and contextual feature f. The drop-one loss is $R^2(G + C) - R^2(G + C \setminus f)$, where C contains all five contextual features; a negative loss means that removing the feature improves prediction. Intervals use 20,000 item-bootstrap resamples of paired out-of-fold squared errors averaged over repeats. They are nominal 95% intervals, without multiplicity correction across the feature comparisons. The selected layers remain fixed, as in the main analysis.

In RoBERTa-base, semantic consensus adds $+.099$ $[+.021, +.181]$, centroid dispersion $+.082$ $[+.009, +.157]$, and semantic volume $+.080$ $[+.005, +.159]$ when each is added alone. In Ettin-Encoder, norm adds $+.158$ $[+.033, +.278]$, compared with $+.144$ for the full block; contextual–stem distance adds $+.056$ $[+.009, +.114]$. These exploratory results show that more than one summary can supply predictive information. They do not identify a necessary feature or establish the direction predicted by the stability hypothesis, because the estimator can exploit non-linear relations.

The largest estimated drop-one loss is .023 (RoBERTa-base, norm); no positive loss has an interval excluding zero. There is one small negative interval excluding zero: removing semantic volume in Ettin-Encoder gives a loss of $-.007$ $[-.015, -.001]$. Small estimated losses are compatible with overlapping predictive information and estimator interactions, but their uncertainty does not establish that each feature is dispensable. The analysis is a diagnostic of the fitted feature block, not a causal ablation of the language model.

G ROBUSTNESS TO THE NUMBER OF ITEMS

Figure G1 re-estimates the within-condition contextual increment on stratified subsamples of the item set ($n = 80$, 120 and 160, equal numbers per condition, 12 random subsamples each, 2×5 repeated cross-validation per subsample, random forest, condition means estimated inside training folds) and on the full set ($n = 204$, 12 re-runs of the cross-validation). The mean increments are similar across sample sizes: RoBERTa-base $+.12$, $+.12$, $+.11$, $+.12$ and Ettin-Encoder-1B $+.17$, $+.17$, $+.16$, $+.15$ across the four sizes, with every subsample positive from $n = 120$ upward in both; the three decoders stay between $+.02$ and $+.08$ at every size. What n changes is the spread across subsamples, which roughly halves from $n = 80$ to $n = 204$. These comparisons support stability of the mean increment within the present stimulus pool; they do not establish that sample size affects precision alone or determine the effect size for new stems, languages or independently constructed contexts. Intervals at $n = 204$ reflect cross-validation re-runs only and are narrower than the item-bootstrap intervals of Table C2.

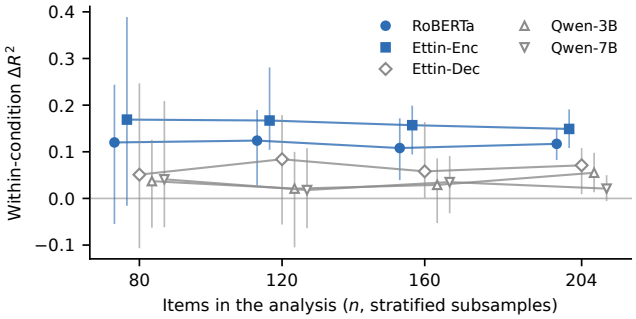

Figure G1: **The within-condition increment is stable in size across item counts.** Mean within-condition ΔR^2 (random forest, fold-internal centring) on stratified subsamples of n items; whiskers span the 2.5th to 97.5th percentile over 12 subsamples ($n < 204$) or 12 cross-validation re-runs ($n = 204$); encoders in blue, decoders in grey.

H AFFIX PRODUCTIVITY

Affix productivity is computed as the logit of the hapax-based productivity ratio, $\log\left(\frac{n_1/N}{1-n_1/N}\right)$, where n_1 is the number of hapax legomena carrying the affix and N is the token count of the affix, with a constant $\varepsilon = 10^{-6}$ guarding the boundary. Values are standardised across the 204 items.

I HUMAN RATINGS AND SENTENCE CONTEXTS

Collection. Ninety-two native English speakers recruited through Prolific each rated the 204 target items and 50 attested derived words serving as fillers (for example *blogger*, *fluency*, *concealable*), one rating per item on a nine-point scale, after a block of practice trials; age, nationality and other languages spoken were recorded. Items were presented in isolation, as bare word forms without sentence context, so the ratings measure a context-free judgement of the form itself. Every item received 90 to 92 ratings. Fillers were rated near the ceiling (mean 8.05), and the target conditions were ordered as designed: Real 6.31, Nonce 4.94, Nonaffix 4.04, NonStem 3.95 on the raw scale.

Normalisation and reliability. The item mean is rescaled linearly to $[0, 1]$ over the full stimulus list including fillers (raw means 3.11 to 8.46), so target items span 0 to .90. Agreement between raters on individual items is modest (ICC(2,1) = .21), which is why many raters were used; the reliability of the item mean over 92 raters is ICC(2,k) = .96, and split-half reliability of item means with Spearman–Brown correction is .978 (95% interval over 1,000 random splits .971 to .983). Within condition, ICC(2,k) is .95 for Real, .94 for Nonce, .81 for Nonaffix and .76 for NonStem, so the within-condition variation that the main analysis targets is reliable, most so in the two intact conditions where the contextual signal is concentrated. Missing rating cells are filled with the item mean for these reliability calculations.

Sentence contexts. Each item has 30 sentences in which it appears once. Examples: *The flagless ship cautiously approached the coast* (Real); *The deckant platform resisted heavy waves* (Nonce); *Most coockive ideas often come from open discussion* (NonStem); *The market sold dried beanlass* (Nonaffix). The sentences were used only to obtain contextual representations; raters never saw them. The sentence database, the ratings and the feature tables will be released on acceptance.

J TOKEN-COUNT SENSITIVITY ANALYSIS

To assess sensitivity to subword fragmentation, each item's piece count is added to both the global-only predictor and its contextual augmentation. Counts are obtained by tokenising the form with a leading space using each checkpoint's tokenizer; they are type-level counts, not counts re-extracted from each of the 30 sentence contexts. The analysis retains the selected layers and the main within-

condition protocol: random forests with 100 trees, 20×5 stratified cross-validation, and condition means estimated within each training fold for the outcome and all predictors, including piece count. Paired out-of-fold squared errors are averaged over repeats; 95% percentile intervals use 20,000 item-bootstrap resamples of these paired errors.

Table J1: Within-condition contextual increments before and after adding piece count to both predictors ($n = 204$). Brackets give 95% paired item-bootstrap intervals.

Checkpoint	Main ΔR^2	With piece count ΔR^2
RoBERTa-base	$+.102$ $[+.006, +.204]$	$+.100$ $[+.005, +.202]$
Ettin-Encoder-1B	$+.144$ $[+.019, +.269]$	$+.122$ $[-.001, +.242]$
Ettin-Decoder-1B	$+.062$ $[-.047, +.180]$	$+.047$ $[-.060, +.162]$
Qwen2.5-3B	$+.053$ $[-.021, +.139]$	$+.054$ $[-.020, +.141]$
Qwen2.5-7B	$+.029$ $[-.051, +.119]$	$+.023$ $[-.054, +.110]$

Point estimates remain positive in all five checkpoints, and only RoBERTa-base's adjusted interval excludes zero. Ettin-Encoder's estimate decreases from $+.144$ to $+.122$, with its adjusted interval extending slightly below zero. The analysis assesses adjustment for piece count, not for every property of segmentation; it does not eliminate differences in subword identity or boundaries. Layer selection remains as described in Appendix B.