Brain-Predictive Reasoning Embedding through Residual Disentanglement

Editors: List of editors' names

Abstract

Conventional brain encoding analysis using language models that feeds whole hidden states can be biased toward shallow lexical cues. Here we present a residual-layer disentangling method that extracts four nearly orthogonal vectors from a language model, respectively containing information corresponding to lexicon, syntax, meaning, and reasoning. We first probe the model to locate the layers where each linguistic feature is maximal, then strip lower-level feature layer-by-layer. Applying bootstrap-ridge encoding to natural-speech ECoG yields three insights: 1) Our residual pipeline isolates a reasoning embedding with unique predictive value, possible only because the latest large language models exhibit emergent reasoning behavior. 2) Apparent high-level predictive performance in conventional analyses is largely attributable to recycled shallow information, rather than genuine deep processing. 3) The reasoning embedding reveals distinct spatiotemporal brain activation patterns, including recruitment of frontal and visual regions beyond classical language areas, suggesting a potential neural substrate for high-level reasoning. Together, our approach removes shallow bias, aligns distinct transformer strata with brain hierarchies, and provides the first brain-relevant representation of reasoning.

Keywords: Disentangled Representations, Encoding Models, Reasoning, Language

1. Introduction

Large language models (LLMs) exhibit strong alignment with human brain activity during language comprehension, as shown by language encoding models that map LLM hidden states to neural responses (Wehbe et al., 2014; Huth et al., 2016; de Heer et al., 2017; Jain and Huth, 2018; Toneva and Wehbe, 2019; Goldstein et al., 2022; Oota et al., 2022; Heilbron et al., 2022; Chen et al., 2023; Li et al., 2023; Antonello et al., 2023; Aw and Toneva, 2023; Chen et al., 2024a,b). These models provide a powerful lens into cortical processing, but most prior work has focused on semantics or low-level phonology (Antonello et al., 2021; Vaidya et al., 2022; Mischler et al., 2024; Caucheteux and King, 2022), leaving open how higher-level reasoning aligns across brain and machine. One barrier is historical—robust reasoning has only recently emerged in LLMs (Ke et al., 2025). Another is methodological—standard encoding models treat hidden states as monolithic, without disentangling lexical, syntactic, semantic, and reasoning features.

Recent evidence (Lampinen et al., 2024) suggests such embeddings are biased toward simpler, linearly extractable features. This raises the possibility that apparent brain alignment largely reflects lexical or syntactic overlap, rather than reasoning.

We address this by introducing residual reasoning embeddings, which isolate the reasoning-specific variance in LLM representations beyond lexical and semantic content. Using these embeddings, we predict human electrocorticographic (ECoG) activity during inference tasks. An overview of our framework is shown in Figure 1. We find that reasoning signals align with distinct spatiotemporal neural patterns—later in time and more anterior cortices—whereas full embeddings remain biased toward low-level features. This disentanglement reveals a shared computational hierarchy between LLMs and the human brain, offering a new perspective on both model interpretability and the neural basis of abstract linguistic reasoning.

2. Methods

Datasets To identify LLM layers specialized for syntactic, meaning, and reasoning features, we use established probing datasets. For syntax, we employ the Benchmark of Linguistic Minimal Pairs (BLiMP) (Warstadt et al., 2020), which tests grammatical competence across 67 controlled paradigms. For meaning and reasoning, we use the Conceptual Minimal Pair Sentences (COMPS) dataset (Misra et al., 2023), which evaluates concept–property associations with increasing levels of inference complexity. These datasets allow us to determine the layers where syntactic, semantic, and reasoning features are most strongly encoded. Illustrative examples are provided in Appendix A.

Language Model We use the Qwen2.5-14B model, a 14.7B-parameter transformer with 48 layers (Team, 2025). Each layer outputs hidden states of dimension 5120. The model supports a maximum context length of 131k tokens, though our experiments only use context sizes of 50 tokens. We use the base model without any instruction tuning or task-specific fine-tuning, in order to examine the model's inherent reasoning capability rather than reasoning behavior induced by supervised alignment.

Minimal Pair Probing We apply minimal pair probing (He et al., 2024a,b) to identify the LLM layers most specialized for syntactic, meaning, and reasoning features. Using probing datasets introduced in Section 2, we train classifiers on layer-wise representations and determine the *saturation layer*, i.e., the earliest layer where feature-specific performance stabilizes. The hidden states from these saturation layers are then used to construct syntactic, meaning, and reasoning embeddings for subsequent brain alignment analysis.

To strengthen this analysis, we additionally validated reasoning probes across multiple tasks (Appendix E) and extended the probing to a broad range of Qwen models (Appendix F), confirming both the robustness of our reasoning measure and the generality of the observed feature emergence order. Full methodological details are provided in Appendix B and C.

Feature-Specific Embeddings Building on the saturation layers identified in Appendix B, we construct four embeddings corresponding to lexical, syntactic, meaning, and reasoning features. Lexical information is taken directly from the input layer, while the other three are obtained by removing lower-level contributions from higher-layer representations. This residualization procedure yields feature-specific embeddings aligned with the linguistic hierarchy. Full methodological details, including regression formulations and dataset setup, are provided in Appendix G.

Encoding Model We assess the neural alignment of our feature-specific embeddings using the Podcast ECoG dataset (Zada et al., 2025), which provides high-gamma intracranial recordings from nine participants with a time-aligned transcript. Ridge regression models are trained to predict neural responses from lexical, syntactic, meaning, and reasoning embeddings, with performance quantified by correlation between predicted and actual signals. To control for generic acoustic effects, we regress out word-rate features. Full details of preprocessing, regression setup, variance partitioning, and statistical baselines are provided in Appendix H.

Hierarchical Variance Partitioning To quantify the unique neural contributions of lexical, syntactic, meaning, and reasoning features, we perform hierarchical variance partitioning. A fixed representational budget ensures comparability across models, and the unique effect of each feature is estimated by measuring the drop in variance explained when that feature is removed from the composite model. Full methodological details are provided in Appendix I.

3. Disentanglement Validation

Mutual Independence Theorem Because linguistic features emerge progressively across LLM layers (syntax early, meaning in mid layers, reasoning later), the residualization procedure yields embeddings that are approximately orthogonal. Intuitively, later layers already contain information from earlier ones, so subtracting out lower-level contributions produces feature-specific embeddings with minimal overlap. We formalize this result in Appendix J.

Mutual Independence via Cosine Similarity We validate that the four feature-specific embeddings encode distinct information by measuring pairwise cosine similarity. Figure 2 shows that while raw hidden states at saturation layers exhibit substantial overlap (especially between meaning and reasoning), the residual embeddings display near-zero off-diagonal similarity. This confirms that our residualization procedure disentangles overlapping features, yielding orthogonal representations of lexical, syntactic, semantic, and reasoning information. Full computational details are provided in Appendix K.

Feature Specificity of Residual Embeddings To confirm that our residual embeddings capture distinct cognitive features rather than generic complexity, we evaluated each embedding (syntax, meaning, reasoning) on all probing tasks. This forms a cross-task matrix where each embedding is expected to perform best on its corresponding task. The results reveal strong diagonal dominance, supporting the specificity of our residualization pipeline and demonstrating that the residual embeddings meaningfully isolate syntactic, semantic, and reasoning information. Full results and analyses are reported in Appendix L.

4. Results

Shallow Features Explain More in the Language Encoding. We assessed the variance explained by each feature-specific embedding when used independently in the encoding model. Across all subjects, syntactic and lexical features consistently accounted for more variance than meaning or reasoning. Encoding models built on shallower features achieved higher peak correlations with neural activity, both at the group and individual level. Moreover, the spatial profile of the full embedding closely resembled that of the syntactic embedding, suggesting that low-level structural information dominates neural predictions. Detailed statistics and figures are reported in Appendix M.

Reasoning Embedding Shows Different Spatiotemporal Pattern Compared to Lexicon, Syntax, and Meaning. We observed clear spatiotemporal distinctions across linguistic features. Syntactic signals exhibited strong correlations both before and after word onset, while lexical signals rose sharply shortly after onset. Meaning signals peaked later, and reasoning signals showed the latest temporal maximum, occurring several hundred milliseconds after onset. Spatially, reasoning signals emerged weakly in temporal regions before progressing anteriorly toward frontal areas, whereas syntactic signals showed widespread alignment across auditory and perisylvian regions throughout the time window. These results highlight distinct temporal and spatial dynamics for reasoning compared to shallower features, suggesting that disentanglement enables isolation of reasoning-specific neural signals. Full statistics and visualizations are provided in Appendix N.

Reasoning Recruits More than Language Areas Compared to Low-level Aspects. We observed a distinct spatial pattern for reasoning representations across cortical regions. Within the superior temporal gyrus (STG), reasoning-related activity shifted anteriorly, in contrast to the posterior bias of lexical and syntactic features. Moreover, reasoning uniquely engaged regions outside classical language areas, including frontal and occipital cortices. These findings suggest that reasoning involves higher-order cognitive regions and may even recruit visual areas during abstract inference. Full statistical details and regional breakdowns are provided in Appendix O.

5. Discussion and Conclusion

By disentangling lexical, syntactic, meaning, and reasoning information in LLMs, we revealed distinct neural alignment patterns that are obscured in full embeddings. Lower-level features such as syntax and lexicon explained more variance and appeared earlier and more broadly, while reasoning emerged later, localized to anterior and frontal regions. These findings suggest that reasoning in LLMs corresponds to higher-order cortical processes beyond classical language areas. Our framework thus provides a novel way to map feature-specific representations in LLMs onto human brain activity, offering new insights into the neural basis of language and reasoning. Further limitations and directions for future work are provided in Appendix P.

References

Richard Antonello, Javier S Turek, Vy Vo, and Alexander Huth. Low-dimensional structure in the space of language representations is reflected in brain responses. *Advances in Neural Information Processing Systems*, 34, 2021.

Richard Antonello, Aditya Vaidya, and Alexander G. Huth. Scaling laws for language encoding models in fmri, 2023.

Khai Loong Aw and Mariya Toneva. Training language models to summarize narratives improves brain alignment, 2023.

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

Catherine Chen, Tom Dupré la Tour, Jack Gallant, Daniel Klein, and Fatma Deniz. The cortical representation of language timescales is shared between reading and listening. *bioRxiv*, pages 2023–01, 2023.

Peili Chen, Linyang He, Li Fu, Lu Fan, Edward F Chang, and Yuanning Li. Do self-supervised speech and language models extract similar representations as human brain? In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2225–2229. IEEE, 2024a.

Peili Chen, Shiji Xiang, Linyang He, Edward F Chang, and Yuanning Li. Convergent representations and spatiotemporal dynamics of speech and language in brain and deep neural networks. *bioRxiv*, pages 2024–12, 2024b.

Wendy A de Heer, Alexander G Huth, Thomas L Griffiths, Jack L Gallant, and Frédéric E Theunissen. The hierarchical cortical organization of human speech processing. *Journal of Neuroscience*, 37(27):6539–6557, 2017.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380, 2022.

Linyang He, Peili Chen, Ercong Nie, Yuanning Li, and Jonathan R Brennan. Decoding probing: Revealing internal linguistic structures in neural language models using minimal pairs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4488–4497, 2024a.

Linyang He, Ercong Nie, Helmut Schmid, Hinrich Schütze, Nima Mesgarani, and Jonathan Brennan. Large language models as neurolinguistic subjects: Identifying internal representations for form and meaning. *arXiv preprint arXiv:2411.07533*, 2024b.

Micha Heilbron, Kristijan Armeni, Jan-Mathijs Schoffelen, Peter Hagoort, and Floris P De Lange. A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, 119(32):e2201968119, 2022.

Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/f471223d1a1614b58a7dc45c9d01df19-Paper.pdf.

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese, et al. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems. *arXiv preprint arXiv:2504.09037*, 2025.

Andrew Kyle Lampinen, Stephanie C. Y. Chan, and Katherine Hermann. Learned feature representations are biased by complexity, learning order, position, and more. *Transactions on Machine Learning Research*, September 2024. URL https://openreview.net/forum?id=aY2nsgE97a.

Yuanning Li, Gopala K Anumanchipalli, Abdelrahman Mohamed, Peili Chen, Laurel H Carney, Junfeng Lu, Jinsong Wu, and Edward F Chang. Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nature Neuroscience*, 26(12):2213–2225, 2023.

Huanru Henry Mao, Shuyang Li, Julian McAuley, and Garrison Cottrell. Speech recognition and multi-speaker diarization of long conversations. *arXiv preprint arXiv:2005.08072*, 2020.

Gavin Mischler, Yinghao Aaron Li, Stephan Bickel, Ashesh D Mehta, and Nima Mesgarani. Contextual feature extraction hierarchies converge in large language models and the brain. *Nature Machine Intelligence*, pages 1–11, 2024.

Kanishka Misra, Julia Rayz, and Allyson Ettinger. Comps: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2920–2941, 2023.

Subba Reddy Oota, Manish Gupta, and Mariya Toneva. Joint processing of linguistic properties in brains and language models. *arXiv preprint arXiv:2212.08094*, 2022.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025. URL https://arxiv.org/abs/2412.15115.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/749a8e6c231831ef7756db230b4359c8-Paper.pdf.

Aditya R Vaidya, Shailee Jain, and Alexander G Huth. Self-supervised models of audio effectively explain human cortical responses to speech. *arXiv preprint arXiv:2205.14252*, 2022.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575, 2014.

Zaid Zada, Samuel A Nastase, Bobbi Aubrey, Itamar Jalon, Sebastian Michelmann, Haocheng Wang, Liat Hasenfratz, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Sasha Devore, Adeen Flinker, Orrin Devinsky, Ariel Goldstein, and Uri Hasson. The "podcast" ecog dataset for modeling neural activity during natural language comprehension. *bioRxiv*, 2025. doi: 10.1101/2025.02.14.638352. URL https://doi.org/10.1101/2025.02.14.638352.

Appendix A. Dataset Examples

Dataset for syntactic probing. To construct feature-specific representations for brain alignment, we first need to identify the LLM layers most specialized for encoding syntactic, meaning, and reasoning features. For syntactic encoding, we use the Benchmark of Linguistic Minimal Pairs (BLiMP) (Warstadt et al., 2020). BLiMP evaluates syntactic knowledge through controlled minimal pairs that differ in grammaticality. By training a classifier to distinguish acceptable from unacceptable sentences based on hidden states at each layer of the LLM, we detect where syntactic competence emerges and peaks. For instance, in a task targeting subject–verb agreement:

a. The cats annoy Tim. (Correct)

b. The cats annoys Tim. (Incorrect)

We apply this diagnostic probing setup across all 67 syntactic paradigms in BLiMP, averaging classifier performance to determine the saturation layer, where syntactic features are stably encoded.

Dataset for meaning and reasoning probing. For meaning and reasoning encoding, we use the Conceptual Minimal Pair Sentences (COMPS) dataset (Misra et al., 2023). Sentence pairs in COMPS differ in whether the subject plausibly inherits a property. By probing whether each layer favors the correct concept–property association, we detect the layers where meaning and reasoning abilities peak.

COMPS-BASE evaluates surface-level understanding without requiring inference. For example, given the property "can heat food":

a. An oven can heat food. (Correct)

b. A refrigerator can heat food. (Incorrect)

COMPS-WUGS-DIST increases reasoning complexity by replacing known concepts with nonsense words (e.g., "wug", "dax") and inserting distractor sentences that introduce interference. This design prevents the model from relying on surface-level co-occurrence or positional cues: distractors can be reordered. The model must infer the identity of the nonsense word from context to generalize property inheritance:

a. A wug is an oven. A dax is a refrigerator. Therefore, a wug can heat food. (Correct)

b. A wug is an oven. A dax is a refrigerator. Therefore, a dax can heat food. (Incorrect)

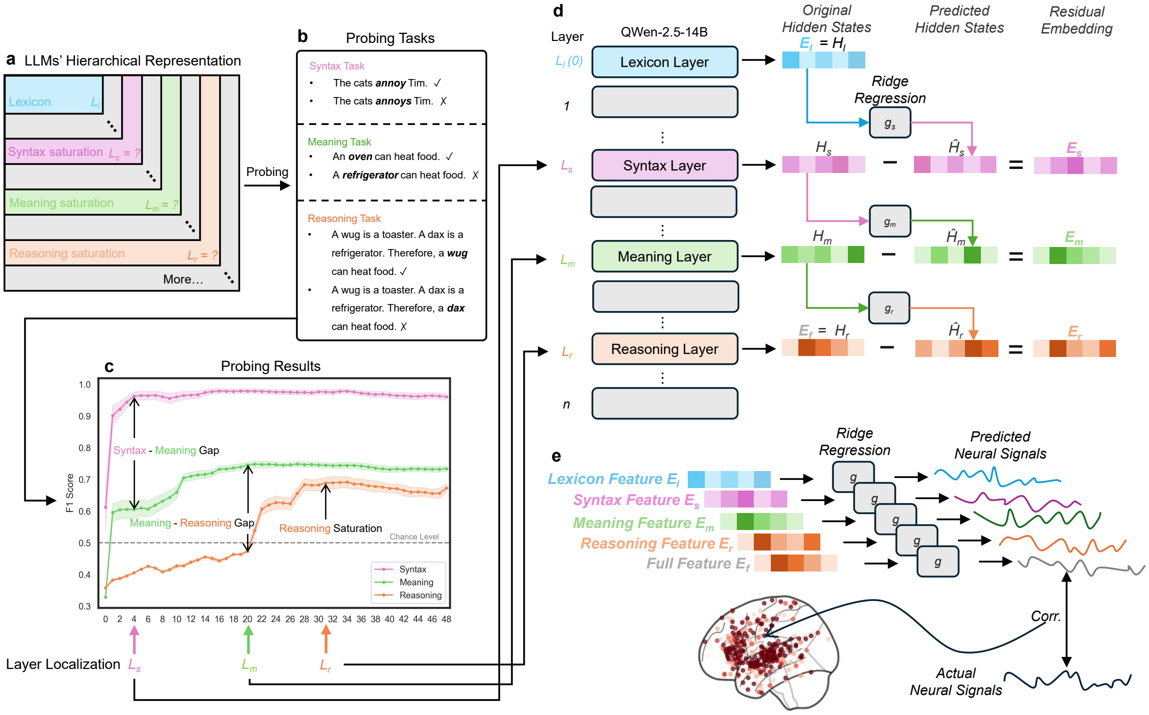

Figure 1: **a)** Hierarchical representations in an LLM. Transformer layers accumulate information in order: lexical features emerge first, followed by syntax, contextual meaning, and eventually higher-order reasoning, with still-richer knowledge continuing in later layers. **b)** Minimal-pair probing tasks. Three diagnostic sentence sets separately test syntax, concept meaning, and multi-premise reasoning. **c)** Layer localization from probing curves. We define L_s – the earliest layer where syntax performance saturates while meaning is still low; L_m – layer where meaning saturates but reasoning has not yet emerged; and L_r – layer where reasoning performance plateaus. These identified layers through probing are used in later analyses. **d)** Feature disentangling across layers. Starting from the localized layers, we iteratively regress lower-level features out of higher ones. Details of residual embedding constructions are in Algorithm 1. Residual disentangling yields four orthogonal embeddings that isolate lexicon, syntax, meaning, and reasoning information. **e)** Brain encoding with purified features. Each residual feature is fed into a ridge encoder to predict high-gamma ECoG responses during podcast listening. Comparing predicted and actual neural signals reveals the spatiotemporal distribution of cortical activity uniquely associated with lexicon, syntax, meaning, and reasoning representations.

Appendix B. Minimal Pair Probing Details

For each probing dataset, we extract sentence representations from each LLM layer by feeding the sentence in isolation and taking the hidden state of its final token. These representations are used to train a logistic regression classifier to predict the correct item in each minimal pair. Model performance at a given layer is measured by the normalized F1 score of this classifier.

Formally, let $H_l \in \mathbb{R}^{n \times d}$ denote the matrix of sentence embeddings at layer l, where each row corresponds to the final-token hidden state of one sentence in the minimal pair dataset. To determine the most specialized layer for a given feature, we define the saturation layer as the earliest layer where performance plateaus:

$$L_x := \min \left\{ l \mid \forall l' > l,\ F_1^{\mathcal{D}_x}(H_{l'}) - F_1^{\mathcal{D}_x}(H_l) < \varepsilon \right\}, \quad x \in \{s, m, r\},$$

where $\mathcal{D}_s$ = BLiMP, $\mathcal{D}_m$ = COMPS-BASE, and $\mathcal{D}_r$ = COMPS-WUGS+COMPS-WUGS-DIST. Here, ε is a small threshold representing tolerance for marginal improvement.

The hidden states at the identified saturation layers—denoted $H_s := H_{L_s}$, $H_m := H_{L_m}$, and $H_r := H_{L_r}$—are used to construct feature-specific embeddings in the next stage of our analysis.

Appendix C. Probing Task Filtering

We employ a Bag-of-Words (BoW) baseline and exclude tasks where the BoW model achieves an accuracy above 0.6, resulting in a final set of 29 BLiMP tasks (out of 67). The conceptual (COMPS-BASE) and reasoning (COMPS-WUGS-DIST) tasks remain unchanged.

This filtering step is motivated by the fact that BoW readily captures tasks solvable via simple lexical cues, indicating a shallow design. By removing these superficially "easy" tasks, we ensure that the retained tasks demand genuine syntactic or semantic understanding rather than mere lexicon-based heuristics.

Appendix D. Probing Results of More Models

In addition to Qwen2.5-14B, we also evaluated Qwen1.5, Qwen2, and Qwen3 models.

Appendix E. Cross-Validation of Reasoning Probes

In the main text, reasoning saturation layers were primarily identified using COMPS-WUGS-DIST. To assess robustness across tasks of varying complexity, we incorporated additional reasoning probes: a 5-hop deductive reasoning task from ProntoQA and the WinoGrande benchmark.

We applied COMPS-WUGS-DIST, ProntoQA, and WinoGrande to 17 Qwen models of varying sizes and training modes. The results demonstrate strong consistency across the three tasks. Specifically, the average difference in reasoning saturation layers between COMPS-WUGS-DIST and ProntoQA was only 0.94, and between COMPS-WUGS-DIST and WinoGrande was 0.53. This small divergence, relative to overall model depth (25–49 layers), supports the robustness of our reasoning probe.

Reasoning Saturation Layer Results. Table 1 presents the layer positions identified by the three tasks across all models. The high degree of agreement confirms that reasoning-related saturation is reliably detected across different reasoning formats, from property inheritance to multi-step deduction and commonsense coreference.

Table 1: Reasoning saturation layer identified by different probes across Qwen models.

Model	**ProntoQA**	**COMPS-WUGS-DIST**	**WinoGrande**
Qwen-1.8B	13	14	14
Qwen-7B	13	15	16
Qwen-14B	18	20	20
Qwen1.5-1.8B	13	14	14
Qwen1.5-7B	14	16	16
Qwen1.5-14B	20	20	22
Qwen2-1.5B	17	17	17
Qwen2-7B	16	18	19
Qwen2.5-1.5B	16	18	17
Qwen2.5-7B	16	19	19
Qwen2.5-14B	28	28	29
Qwen3-1.7B (thinking-off)	19	19	19
Qwen3-8B (thinking-off)	23	23	23
Qwen3-14B (thinking-off)	27	27	26
Qwen3-1.7B (thinking-on)	16	16	16
Qwen3-8B (thinking-on)	20	20	21
Qwen3-14B (thinking-on)	20	21	22

Summary. These findings confirm that our framework's reasoning probe is not limited to COMPS but generalizes across reasoning tasks. The consistency across COMPS-WUGS-DIST, ProntoQA, and WinoGrande strengthens confidence in our identification of reasoning-specific saturation layers.

Appendix F. Probing Across Qwen Model Families

In the main text, we primarily used Qwen2.5-14B. To assess generality across architectures and sizes, we extended our probing pipeline to 17 models across the Qwen family, including base, v1.5, v2, v2.5, and v3 models with and without ”thinking mode.”

F.1. Consistent Emergence Order.

Across nearly all models, we observed the same progression of feature emergence: syntax saturates earliest, followed by meaning, and then reasoning. The only exception was Qwen-1.8B, where syntax and meaning saturated at the same layer. Table 2 reports the saturation layers identified for all models.

Table 2: Saturation layers for syntax, meaning, and reasoning across Qwen models.

Model	**Syntax**	**Meaning**	**Reasoning**
Qwen-1.8B	11	11	14
Qwen-7B	11	13	15
Qwen-14B	9	16	20
Qwen1.5-1.8B	10	11	14
Qwen1.5-7B	9	13	16
Qwen1.5-14B	8	16	20
Qwen2-1.5B	10	14	17
Qwen2-7B	7	14	18
Qwen2.5-1.5B	7	14	18
Qwen2.5-7B	7	15	19
Qwen2.5-14B	6	20	28
Qwen3-1.7B (thinking-off)	9	16	19
Qwen3-8B (thinking-off)	7	20	23
Qwen3-14B (thinking-off)	5	17	27
Qwen3-1.7B (thinking-on)	5	13	16
Qwen3-8B (thinking-on)	6	13	20
Qwen3-14B (thinking-on)	4	14	21

F.2. Relative Depth Analysis.

To compare models with different depths, we define the relative depth of a feature as the fraction of total layers between its saturation point and that of the preceding feature:

$$Depth_x = \frac{L_x - L_{\text{prev}}}{n}, \quad x \in \{s, m, r\},$$

where n is the number of layers in the model. Table 3 reports these values for 14B-scale models.

Table 3: Relative depth of features in 14B-scale Qwen models.

Model	**Syntax Depth**	**Meaning Depth**	**Reasoning Depth**
Qwen-14B	0.225	0.175	0.10
Qwen1.5-14B	0.20	0.20	0.10
Qwen2.5-14B	0.125	0.291	0.167
Qwen3-14B (thinking-off)	0.125	0.300	0.250
Qwen3-14B (thinking-on)	0.100	0.250	0.175

Summary. We observe a clear trend across model generations: newer Qwen models devote proportionally fewer layers to syntax and more to meaning and reasoning. This suggests that as model capabilities improve, shallow linguistic features are encoded more efficiently, freeing up representational capacity for higher-level semantics and reasoning.

F.3. Reasoning Performance Results

In addition to probing saturation layers, we evaluated how well different Qwen models perform on reasoning tasks. We tested three benchmarks—ProntoQA (multi-hop deductive reasoning), COMPS-WUGS-DIST (property inheritance with distractors), and WinoGrande (commonsense coreference). Table 4 reports accuracy on each benchmark, along with the average across tasks.

Table 4: Reasoning task performance across Qwen models. The **Average** column reports the mean accuracy across ProntoQA, COMPS-WUGS-DIST, and WinoGrande.

Model	**ProntoQA**	**COMPS-WUGS-DIST**	**WinoGrande**	**Average**
Qwen-1.8B	0.797	0.522	0.523	0.614
Qwen-7B	0.886	0.641	0.602	0.710
Qwen-14B	0.880	0.695	0.647	0.741
Qwen1.5-1.8B	0.792	0.513	0.523	0.609
Qwen1.5-7B	0.848	0.667	0.603	0.706
Qwen1.5-14B	0.910	0.670	0.653	0.744
Qwen2-1.5B	0.784	0.586	0.552	0.640
Qwen2-7B	0.851	0.636	0.660	0.716
Qwen2.5-1.5B	0.783	0.605	0.566	0.651
Qwen2.5-7B	0.879	0.673	0.664	0.739
Qwen2.5-14B	0.922	0.691	0.698	0.770
Qwen3-1.7B (thinking-off)	0.739	0.500	0.530	0.589
Qwen3-8B (thinking-off)	0.876	0.629	0.651	0.719
Qwen3-14B (thinking-off)	0.912	0.716	0.670	0.766
Qwen3-1.7B (thinking-on)	0.836	0.575	0.562	0.658
Qwen3-8B (thinking-on)	0.972	0.623	0.671	0.755
Qwen3-14B (thinking-on)	0.981	0.674	0.694	0.783

Summary. We find consistent improvements in reasoning performance with larger and newer Qwen generations. In particular, Qwen3 models—especially in "thinking mode"—achieve the highest average scores, indicating stronger reasoning capabilities. These results complement our saturation layer analysis by showing that the deeper allocation of layers to reasoning (Appendix F) also translates into improved task-level performance.

Appendix G. Feature-Specific Embeddings

Lexical embedding. We define the lexical embedding $E_l \in \mathbb{R}^d$ as the hidden state at layer 0 of the LLM. As this layer follows token embeddings directly, it reflects uncontextualized lexical properties.

Residual embeddings for syntax, meaning, and reasoning. For the remaining features—syntactic, meaning, and reasoning—we construct residual embeddings by removing lower-level contributions from higher-layer representations. For instance, to isolate reasoning information, we remove the meaning contribution from the layer where reasoning saturates:

$$E_r := H_r - g_r(H_m), \quad \text{where } g_r = \arg\min_W \|H_r - WH_m\|_F^2 + \alpha \|W\|_F^2 \,,$$

where g_r is a ridge regression trained via 4-fold cross validation on a podcast corpus described below. The same procedure applies to compute E_s and E_m. Specifically:

$$\begin{aligned} E_m &= H_m - g_s(H_s) && \text{(meaning minus syntactic)} \\ E_s &= H_s - g_l(H_l) && \text{(syntax minus lexical)} \end{aligned}$$

This yields feature-specific representations that are aligned with the linguistic hierarchy and minimally confounded by lower-level signals.

Dataset for residual regression training. To extract feature-specific residual embeddings, we train ridge regression models that require a large number of training samples. However, the transcript later used for neural alignment—drawn from a single 30-minute podcast episode—is too short to support stable regression. To address this, we train the models on an expanded corpus of 16 transcribed episodes from the same podcast series, including the episode used in the alignment analysis (Mao et al., 2020). This 160k-token dataset enables regression without PCA, preserving richer structure in the hidden states. We then apply the trained models to the target transcript to extract residuals for encoding analysis.

Algorithm 1: Construction of Feature-Specific Residual Embeddings

```
Input: LLM hidden states $\{H_L\}_{L=0}^{L_{\max}}$ for each token; probing datasets $\mathcal{D}_s$, $\mathcal{D}_m$, $\mathcal{D}_r$
Output: Feature-specific embeddings $E_l$, $E_s$, $E_m$, $E_r$
Perform probing with $\mathcal{D}_s$, $\mathcal{D}_m$, $\mathcal{D}_r$ to find saturation layers:;
    $L_s \leftarrow$ syntax saturation layer from $\mathcal{D}_s$;
    $L_m \leftarrow$ meaning saturation layer from $\mathcal{D}_m$;
    $L_r \leftarrow$ reasoning saturation layer from $\mathcal{D}_r$;
    $L_l \leftarrow 0$;
Define lexical embedding: $E_l \leftarrow H_{L_l}$;
for each $(L_{low}, L_{high})$ in $\{(L_l, L_s), (L_s, L_m), (L_m, L_r)\}$ do
    Train ridge regression $g$ to predict $H_{L_{\text{high}}}$ from $H_{L_{\text{low}}}$;
    Compute residual embedding: $E \leftarrow H_{L_{\text{high}}} - g(H_{L_{\text{low}}})$;
    Assign $E_s$, $E_m$, $E_r$ accordingly;
end
```

Appendix H. Encoding Model Details

ECoG dataset. After constructing feature-specific embeddings, we assess their neural alignment using the Podcast ECoG dataset (Zada et al., 2025). This dataset contains high-gamma band (70–200 Hz) intracranial recordings from nine participants as they listened to a 30-minute narrative podcast. It includes 1,330 electrodes and a time-aligned word-level transcript, making it ideal for testing how lexical, syntactic, meaning, and reasoning embeddings align with neural activity during language comprehension.

As described in Section G, we train ridge regression models on an expanded podcast corpus to isolate feature-specific residuals. We then apply these models to the ECoG-aligned transcript to extract feature-specific embeddings, which are used to predict neural responses. Each embedding is aligned to individual word onsets, and for each word, neural signals are epoched in a ±2s window and downsampled to 32 Hz, yielding $t = 128$ time bins per event. Let $X \in \mathbb{R}^{n\times d}$ be the matrix of input embeddings (lexical, syntactic, meaning, or reasoning) across n word-aligned tokens, and $Y \in \mathbb{R}^{n\times(c\cdot t)}$ the corresponding ECoG response matrix across c electrodes and t time lags. We fit:

$$W^* = \arg\min_{W} \|Y - XW\|_F^2 + \alpha\|W\|_F^2,$$

with α selected via 5-fold cross-validation over a log-spaced grid, and $b = 5$ bootstrap resamples per fold using contiguous chunks of length $l = 32$. Model performance is quantified by Pearson correlation between predicted and actual signals. Temporal profiles are obtained by averaging over channels at each lag; spatial maps visualize per-channel peak correlations on 3D brain coordinates.

Word-rate feature regress out. We controlled for generic acoustic onset responses by adding a two-column word-rate covariate (word onsets and syllable-rate) to every ridge model. All variance-partitioning steps therefore quantify the variance explained beyond that attributable to mere word onsets.

Let R_{full} denote the cross-validated prediction correlation achieved using the combined feature set (embedding + word rate), and let R_{wr} denote the correlation using only the word rate features. Assuming approximate orthogonality between the two feature sets, we estimated the unique contribution of the embedding features as:

$$R_{\text{embed}} = \operatorname{sign}(R_{\text{full}}) \cdot \sqrt{\max(0, R_{\text{full}}^2 - R_{\text{wr}}^2)}.$$

This operation projects the full correlation vector onto the embedding-only axis, removing variance explained by word rate features. The assumption of orthogonality is approximately satisfied due to the preprocessing pipeline, and helps to prevent over-attribution of shared variance.

Shuffle baseline and standardization. Considering that different channels and features have varying signal-to-noise ratios (SNRs), we constructed a subject–electrode–specific null distribution to assess whether a feature block explains neural activity beyond chance and to enable cross-electrode analysis. This was done by shuffling the feature rows 100 times while keeping the word-onset covariates fixed. For every shuffle we refit the ridge model and recorded the peak correlation R. Because Pearson correlations are bounded and skewed near ±1, we applied the standard Fisher z-transform (`atanh`) to all correlations, computed the shuffle mean and s.d. in z-space, and converted the true correlation to a z-score. Electrodes with $z > 1.96$ (two-tailed $\alpha = .05$) were deemed responsive.

Appendix I. Hierarchical Variance Partitioning

To assess the contribution of each linguistic feature to neural responses, we perform hierarchical variance partitioning. To prevent apparent gains in encoding accuracy from being driven simply by larger feature spaces, we fix the total representational budget at 500 dimensions for every cumulative model. Concretely, when the model contains N information blocks (lexicon, syntax, meaning, reasoning, ...), we apply PCA independently to each block, retaining exactly $500/N$ principal components per block. The reduced blocks are then concatenated to form a 500-dimensional matrix that is passed to the ridge encoder.

Using the same ridge regression and evaluation pipeline as in the main encoding analysis, we compute the variance explained (R^2) by the composite model that includes all four feature-specific embeddings. To estimate the unique contribution of each feature, we remove the feature from the composite model and measure the resulting drop in explained variance. Let $R^2_{\text{Composite}}$ denote the variance explained by the composite model, and let $R^2_{lmr}, R^2_{lsr}, R^2_{lsm}$ represent ablated models with the syntactic, meaning, or reasoning component removed, respectively. The contribution of each feature is then:

$$\Delta R^2_{\text{Syntactic}} = R^2_{\text{Composite}} - R^2_{lmr}, \quad \Delta R^2_{\text{Meaning}} = R^2_{\text{Composite}} - R^2_{lsr}, \quad \Delta R^2_{\text{Reasoning}} = R^2_{\text{Composite}} - R^2_{lsm}.$$

Appendix J. Mutual Independence Theorem

We justify the approximate orthogonality of the feature-specific embeddings E_l, E_s, E_m, E_r based on the progressive emergence of linguistic features across LLM layers (syntax peaks early, meaning in mid layers, reasoning in later layers). Once a feature reaches saturation, its F1 score remains stable in deeper layers, indicating that later representations retain earlier features. As a result, representations like H_m already embed information from H_l and H_s, making regression from H_m alone nearly as informative as from all three:

$$g_r(H_m) \approx g'_r([H_l, H_s, H_m]) \quad \Rightarrow \quad E_r \approx H_r - g'_r([H_l, H_s, H_m]) =: E'_r$$

Since E'_r is the residual of a linear projection onto $[H_l, H_s, H_m]$, it is orthogonal to each:

$$E'_r \perp H_l, \quad E'_r \perp H_s, \quad E'_r \perp H_m$$

Each residual embedding $E_i \in \{E_l, E_s, E_m\}$ is a linear combination of earlier hidden states (e.g., $E_m = H_m - g_m(H_s) = H_m - W_m H_s$). By the bilinearity of covariance, we have:

$$\text{Cov}(E_i, E'_r) = \text{Cov}(H_i - W_i H_j, E'_r) = \text{Cov}(H_i, E'_r) - W_i \, \text{Cov}(H_j, E'_r) = 0$$

whenever $E'_r \perp H_i$ and H_j, for appropriate $i, j \in \{l, s, m\}$. This implies:

$$\langle E'_r, E_i \rangle = 0 \quad \forall i \in \{l, s, m\}$$

Applying this proof across all residual stages, we conclude approximate mutual orthogonality:

$$\langle E_i, E_j \rangle \approx 0 \quad \text{for all } i \neq j$$

Appendix K. Mutual Independence via Cosine Similarity

To validate that the four feature-specific embeddings encode distinct information, we assess their mutual independence using pairwise cosine similarity. For each token $i \in \{1, \ldots, N\}$, let $E_l^i, E_s^i, E_m^i, E_r^i$ denote the four feature-specific vectors. We compute a 4×4 cosine similarity matrix C_i, take its absolute value $|C_i|$, and average across samples:

$$[C_i]_{j,k} = \frac{\langle E_j^i, E_k^i \rangle}{\|E_j^i\| \cdot \|E_k^i\|}, \qquad \bar{C} = \frac{1}{N} \sum_{i=1}^{N} |C_i|, \qquad j, k \in \{l, s, m, r\}.$$

In the ideal case of perfect disentanglement, off-diagonal entries of $\bar{C}$ would be zero, indicating orthogonality between different embeddings.

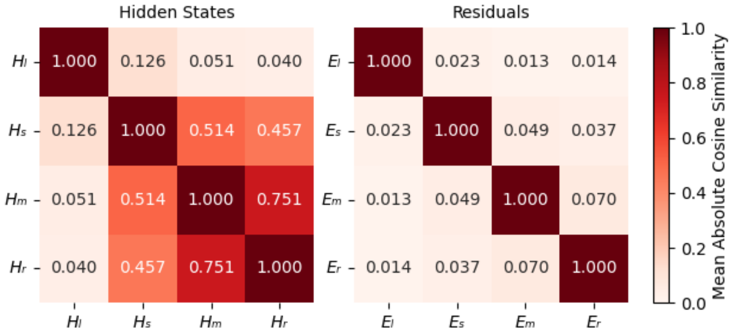

Figure 2: Pairwise cosine similarity among representations before (left) and after (right) residual disentanglement. The hidden states at feature saturation layers (H_l, H_s, H_m, H_r) exhibit substantial overlap. In contrast, the residual embeddings (E_l, E_s, E_m, E_r) show near-zero off-diagonal similarity.

Figure 2 compares the pairwise mean absolute cosine similarity among raw hidden states at the four saturation layers (left) and among the corresponding feature-specific residual embeddings (right). While the hidden states show substantial overlap—especially between meaning and reasoning layers ($\bar{C} = 0.751$)—the residual embeddings exhibit near-zero off-diagonal similarity across all pairs. This sharp drop confirms that our residualization procedure effectively disentangles overlapping features, yielding orthogonal representations of lexical, syntactic, semantic, and reasoning information.

Appendix L. Feature Specificity of Residual Embeddings

To evaluate whether our residual embeddings truly capture syntax, meaning, and reasoning—rather than reflecting general model complexity—we tested them directly on the probing tasks used to derive the saturation layers. Specifically, we trained ridge regression classifiers using each residual embedding and assessed performance on syntax, meaning, and reasoning tasks, forming a 3×3 evaluation matrix.

Baseline performance at saturation layers. Before residualization, embeddings at saturation layers performed well across all tasks, highlighting the need for disentanglement. Table 5 shows normalized scores close to 1.0 for all combinations.

Table 5: Baseline performance at saturation layers (normalized by task peak).

	Syntax Layer	Meaning Layer	Reasoning Layer
Syntax Task	0.947	0.999	0.989
Meaning Task	0.812	0.995	0.995
Reasoning Task	0.617	0.685	0.988

Residual embeddings. After applying residualization, each embedding showed clear specificity, with highest scores on its corresponding task and substantially lower scores on others (Table 6).

Table 6: Performance of residual embeddings across tasks (normalized).

	Syntax Embedding	Meaning Embedding	Reasoning Embedding
Syntax Task	0.882	0.664	0.563
Meaning Task	0.781	0.919	0.802
Reasoning Task	0.648	0.770	1.036

Bias and normalization effects. Two factors explain residual cross-task performance:
1. **COMPS-BASE bias.** Even simple bag-of-words models can exceed chance (0.665 accuracy) on COMPS-BASE. Raw scores confirm that residual syntax and reasoning embeddings fall below this baseline, showing effective disentanglement despite apparent overlap.
2. **Normalization.** Scores were normalized relative to each task's peak performance, unintentionally inflating cross-task values. Raw results (Table 7) clarify the diagonal dominance more clearly.

Table 7: Performance of residual embeddings across tasks (raw scores).

	Syntax Embedding	Meaning Embedding	Reasoning Embedding
Syntax Task	0.863	0.650	0.551
Meaning Task	0.589	0.693	0.605
Reasoning Task	0.448	0.532	0.716

Summary. Together, these results confirm that residual embeddings are feature-specific: syntax embeddings capture syntactic information, meaning embeddings semantic associations, and reasoning embeddings higher-order inference. The disentanglement pipeline thus yields cognitively interpretable and non-overlapping representations.

Appendix M. Shallow Features Explain More in the Language Encoding

We quantified the variance explained by each feature-specific embedding when used independently in the encoding model. Among the four features, syntax accounted for the largest proportion of explained variance at 33.06%, followed by meaning at 25.58%, lexicon at 18.98%, and reasoning at 17.60%. All embeddings were evaluated on the same number of time-aligned samples (1268 word onsets × 128 lags).

Figure 5 illustrates these results. Encoding models built on shallower features consistently achieved higher peak correlations with neural activity. Lexical and syntactic embeddings yielded significantly stronger correlations than meaning or reasoning, both across subjects and at the individual level. Further, the spatial profile of the full embedding closely resembled that of the syntactic embedding, indicating that full model representations are dominated by low-level structural information.

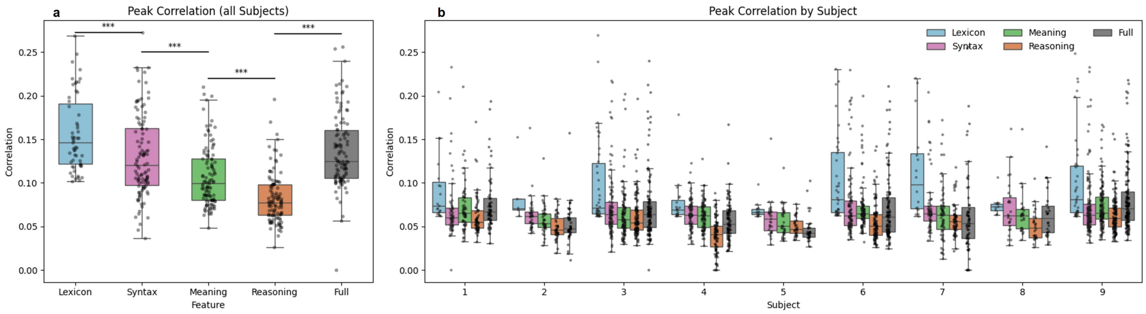

Figure 3: **a)** Peak correlations for each feature across all subjects. Lexical features show the highest correlations. Asterisks indicate significant differences. **b)** Peak correlations by subject show consistent lexical dominance, with variability in other features.

Appendix N. Spatiotemporal Dynamics of Reasoning Embeddings

As shown in Figure 4, syntactic signals exhibit significant correlation with neural activity both before and after word onset. Lexical signals show a sharp increase shortly after onset, followed by meaning signals, which peak later in the post-onset window. Reasoning signals exhibit the latest peak among all features, with a temporal maximum occurring approximately 300–350 ms after word onset.

In addition to temporal dynamics, the residual embeddings exhibit distinct spatiotemporal alignment patterns across the cortex. As shown in Figure 5, the reasoning embedding initially shows weak neural correlation, begins in the superior and middle temporal regions, and progresses anteriorly to the inferior frontal areas. In contrast, syntactic embeddings produce strong and widespread alignment throughout the entire time window, including early auditory and perisylvian regions.

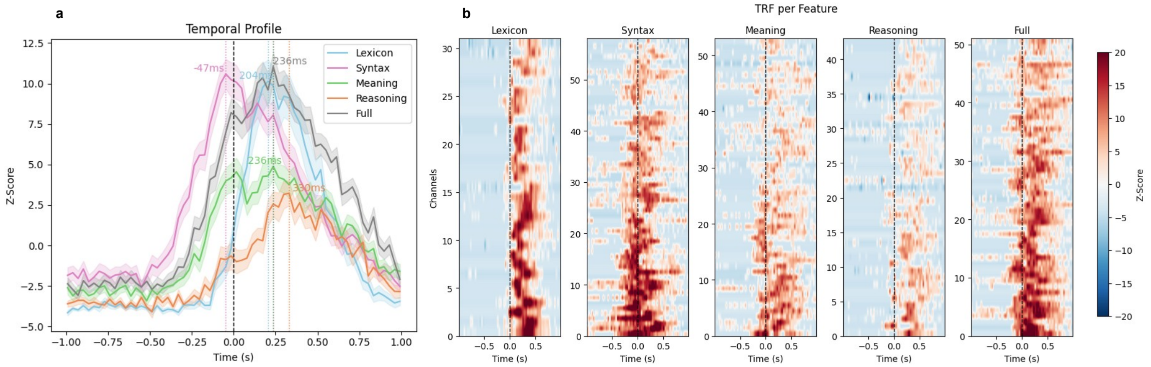

Figure 4: **a)** Temporal profile of different features. The top 10% of electrodes (by peak z-score), among those exceeding 1.96 (p ¡ 0.05), were selected to include highly responsive channels to reveal the temporal pattern. **b)** Temporal receptive field (TRF) across time for electrodes selective to one of the five features: Lexicon, Syntax, Meaning, Reasoning, and Full. Electrodes were selected the same way as in Figure a.

Appendix O. Regional Recruitment of Reasoning Embeddings

As shown in Figure 6, we observed a distinct spatial pattern for reasoning representations across cortical regions.

First, within the superior temporal gyrus (STG), reasoning-related activation increased from posterior to anterior subregions, with the strongest responses in anterior STG. This pattern differs from lower-level features such as Lexicon or Syntax, which were more posteriorly distributed. The anterior bias aligns with the hypothesis that reasoning embeddings capture higher-order cognitive processes.

Second, reasoning uniquely engaged regions outside classical language areas, including the superior frontal gyrus (SFG) and the superior occipital sulcus. These findings suggest that reasoning may involve high-level cognitive regions and potentially recruit visual areas during abstract inference.

Appendix P. Limitations and Future Directions

Despite promising results, several limitations remain:

- **Data constraints.** Our analysis relies solely on ECoG data, which—while offering high temporal precision—provides limited spatial coverage. Frontal regions often implicated in reasoning are under-sampled. Future work could incorporate complementary modalities such as fMRI to improve spatial resolution.

- **Dataset coverage.** Our probing analysis is shaped by the structure of existing datasets: while BLiMP provides broad syntactic coverage, COMPS focuses narrowly on property inheritance. Probing with a wider range of reasoning tasks may reveal

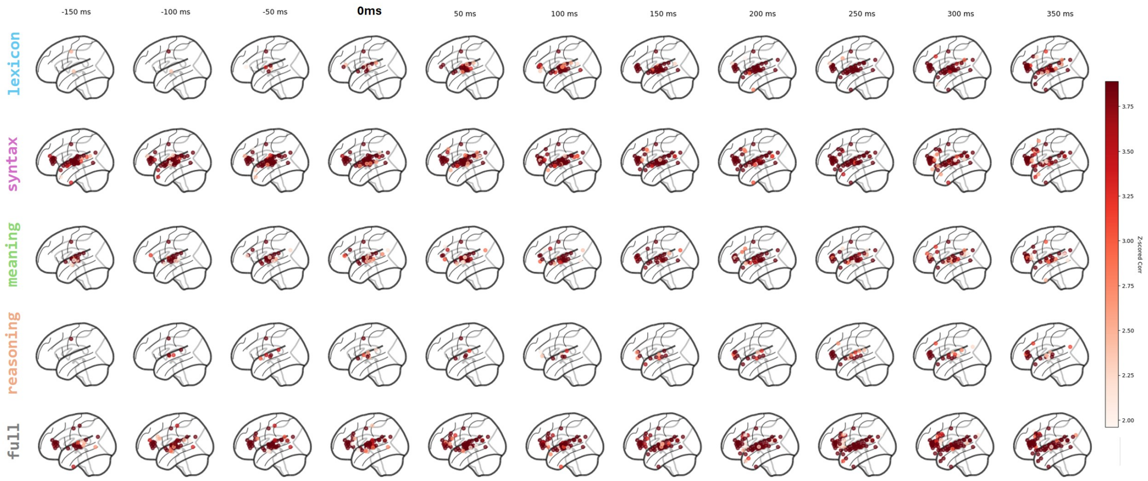

Figure 5: Spatiotemporal distribution of responsive electrodes across language features. Each subplot displays left-hemisphere electrodes with significant encoding performance (Fisher z-scored correlation ¿ 1.96, corresponding to p ¡ 0.05) at a given time point. Colors represent the z-scored correlation strength, capped at 3.89 (p ¡ 1e-4) to enhance visual contrast and prevent high values from masking more moderate responses. For each feature (rows), correlation values are averaged over a ±100 ms window centered at the indicated time points (columns).

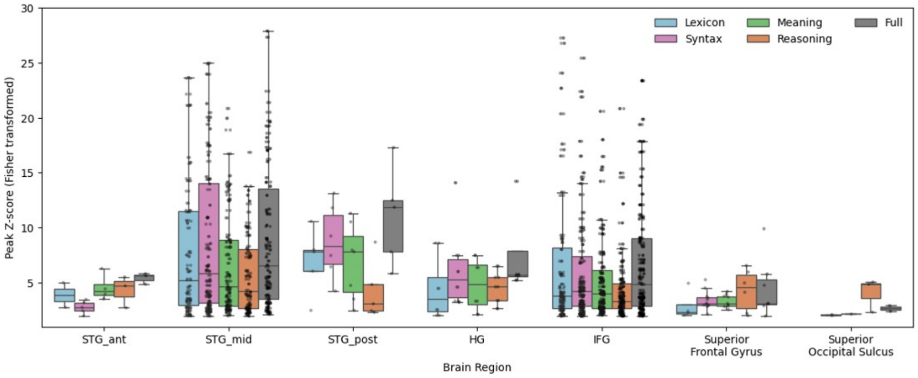

Figure 6: Fisher z-scores across brain regions for each feature. Only significant electrodes with peak $z \geq 1.96$ ($p < 0.05$) were included.

whether distinct reasoning types are differentially encoded in both models and the brain.

- **Model choice.** Due to computational constraints, we focused on a single open-weight model (Qwen2.5-14B). While it balances performance and accessibility, larger models—especially those optimized for multi-step reasoning—may reveal stronger or more differentiated reasoning signals. Extending our framework to such models could shed light on how reasoning representations scale and specialize.

- **Instability in deepest layers.** Beyond the reasoning zone (after layer 30), residualization begins to fail, with regression losses rising dramatically. This suggests a shift in representational structure—possibly toward less compressible, generative transformations. Future work could explore this to better understand the boundary between structured reasoning and open-ended synthesis in LLMs, and how this maps onto neural computation.

Appendix Q. Supplemented Neuroscience Analysis

Figure 7 is a plot of activated channels across different brain areas.

Q.1. Spatial Dissociation of Linguistic Features in the Cortex

As shown in Figure 8, when assigning each electrode to the linguistic feature for which it shows the highest encoding z-score, we observe clear regional differentiation across features. Electrodes dominated by shallow lexical and syntactic features (Lexicon, Syntax) are primarily localized to canonical perisylvian language areas, including the superior temporal gyrus (STG) and the inferior frontal gyrus (IFG), particularly its ventral portion. Semantic feature-selective electrodes (Meaning) extend more dorsally within IFG and are distributed more broadly across the temporal and frontal cortices, consistent with prior accounts of distributed semantic representation. In contrast, reasoning-related electrodes (Reasoning) engage distinct regions, notably including the superior frontal gyrus (SFG) and occipital areas, suggesting recruitment of domain-general and potentially visual-associative mechanisms unique to high-level inferential processes.

Q.2. Temporal Gradient of Linguistic Feature Processing Across Cortex

As shown in Figure 9, we examined the spatial distribution of peak response times for four linguistic features using temporal response function (TRF) analysis. The results revealed a clear spatiotemporal gradient across the cortex:

Superior Temporal Gyrus (STG) exhibited predominantly pre-onset peaks, consistent with its role in early auditory and phonological processing. Within STG, a finer gradient was observed: posterior STG peaked earlier than anterior STG, suggesting a hierarchy of temporal integration along the auditory pathway.

In contrast, Inferior Frontal Gyrus (IFG) showed post-onset peaks, especially for higher-order features such as syntax and reasoning, indicating involvement in late-stage integration and abstraction.

These results suggest a temporal-to-frontal cascade of processing, where early auditory regions process incoming speech in anticipation of the word onset, while frontal areas integrate contextual and abstract information after the word has begun.

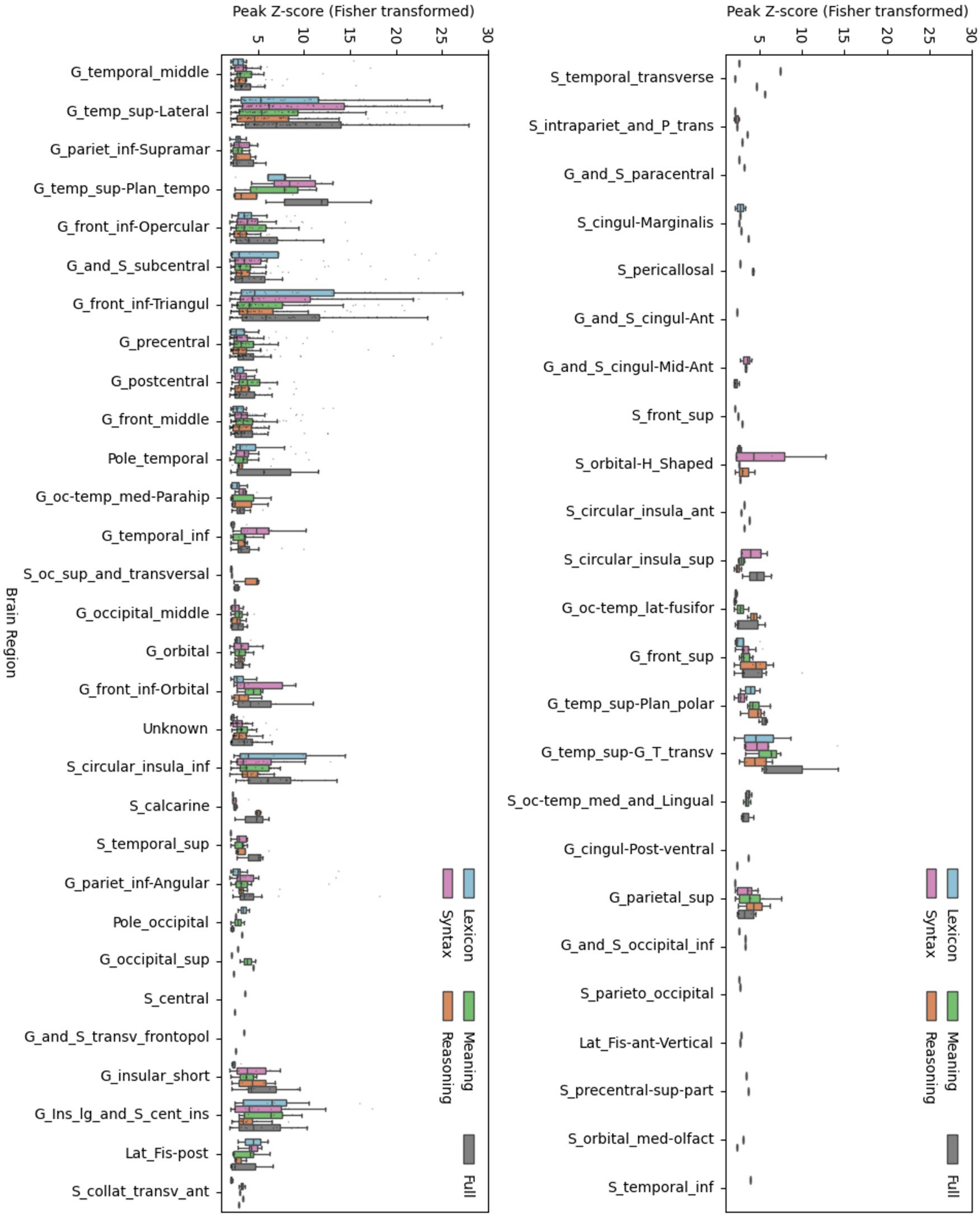

Figure 7: Activated channels across different brain areas.

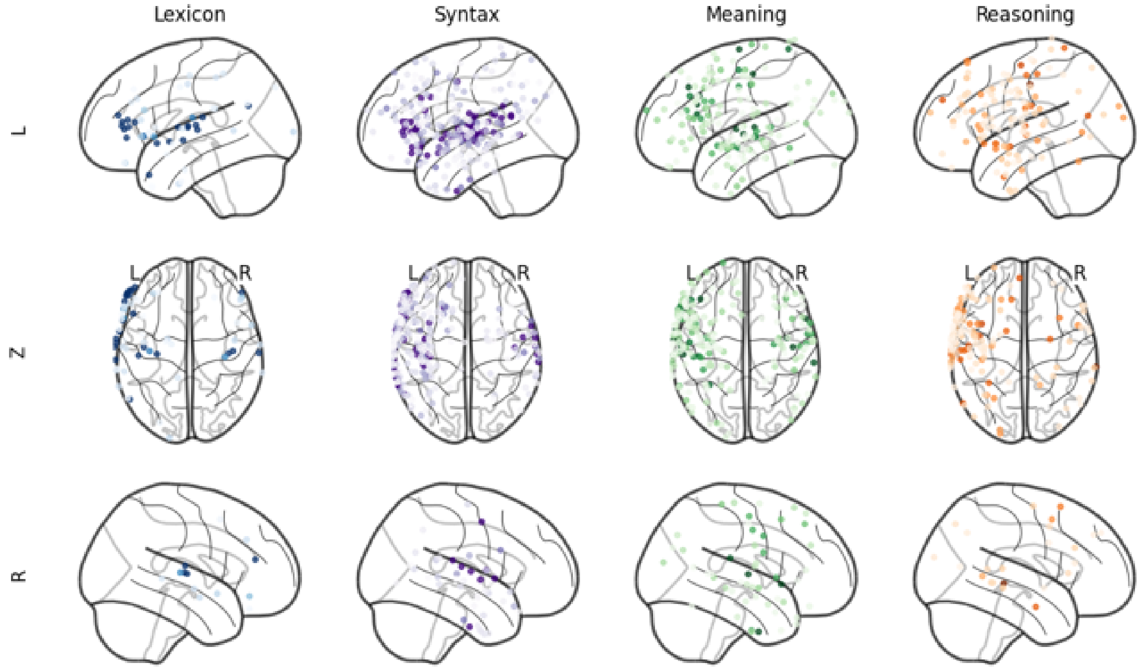

Figure 8: **Brain maps showing the spatial distribution of electrodes dominated by each linguistic feature, across three anatomical views.** Electrodes are colored according to their most selective feature based on peak z-scored encoding correlation values, computed after removing word rate confounds. Each feature (Lexicon, Syntax, Meaning, Reasoning) is shown in a separate column, and each row corresponds to a different cortical view: left lateral, dorsal (axial), and right lateral. Only electrodes exceeding a z-threshold of 2.58 are shown.

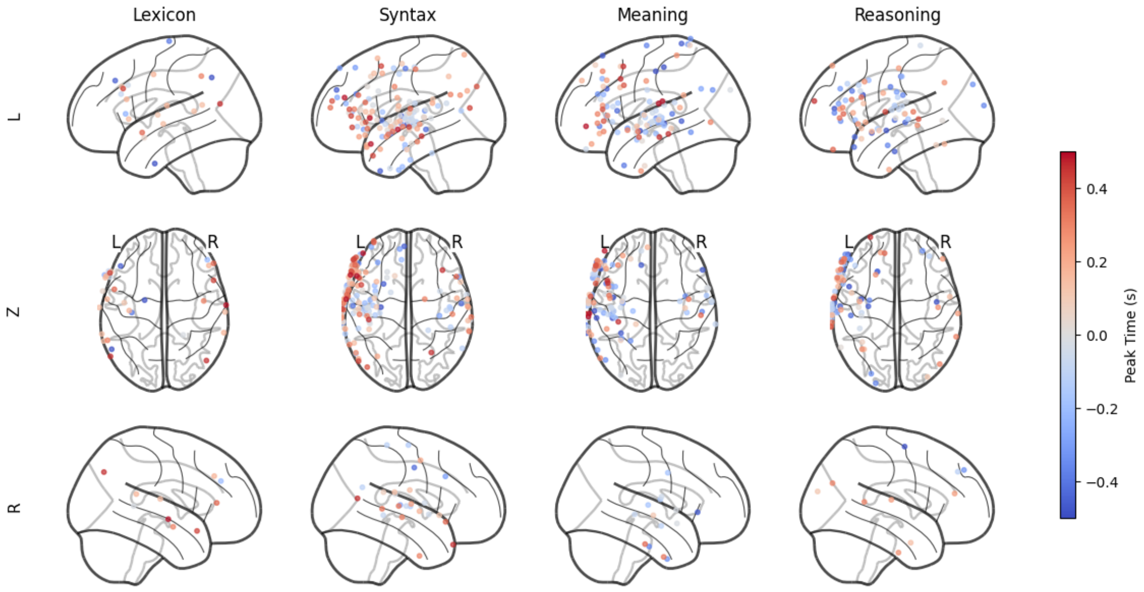

Figure 9: **Spatial distribution of peak response time for each linguistic feature.** Each dot represents an electrode whose neural response is significantly predicted by one of the four feature sets (Lexicon, Syntax, Meaning, Reasoning). The color of each dot encodes the time point (in seconds relative to word onset) at which the encoding model reaches peak performance (z-scored). Blue indicates earlier peaks (pre-onset), while red indicates later peaks (post-onset). Views are shown from the left (L), dorsal (Z), and right (R) perspectives, separately for each feature.